Intelligent System for Automatic Recognition of Environmental Sounds Using Optimal Feature Fusion and Ensemble Deep Learning Technique

Divya Lakshmi. S [*1,2,] N. Suresh Kumar [3]
[1]Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Krishnankoil, Tamilnadu, India
[2]Department of Computer Applications, Marian College Kuttikkanam Autonomous,
Kerala, India
[3]Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Krishnankoil, Tamilnadu, India
E-mail: divyabalu19@gmail.com, sureshkumar@ klu.ac.in, divyabalu19@gmail.com
[*]Corresponding author

Keywords: automatic recognition, environmental sounds, ESC, feature fusion, ensemble learning

Received: April 5, 2025

Environmental sound classification (ESC) is a challenging task due to the unstructured and overlapping nature of ambient sounds, which differ significantly from speech and music. Problems such as class imbalance, limited labeled samples, and high inter-class similarity hinder the performance of traditional classifiers. In this study, we propose a robust ESC system that combines optimal spectrum feature fusion with a stacked ensemble learning strategy. Specifically, we extract three types of spectral features—log Mel spectrum, log–log Mel spectrum, and Mel spectrograms—from environmental audio signals using the DenseNet-161 architecture. These features are then optimally fused using the Boosted Reptile Squirrel Search (BRSS) algorithm to capture both fine- and coarse-grained frequency patterns. For classification, we employ a two-level ensemble model: four classical machine learning classifiers (Linear Regression, Decision Tree, Random Forest, and Support Vector Machine) in the first stage, followed by a Bayesian Tensorized Neural Network (BTNN) for final prediction. Experimental results on three benchmark datasets—ESC-10, ESC-50, and UrbanSound8K—demonstrate that our fused spectrum feature approach achieves an accuracy of 98.98%, surpassing individual feature types and outperforming state-of-the-art models such as Convolutional Recurrent Neural Network (CRNN), EnvNet, and DualResNet. These results highlight the effectiveness and superiority of our proposed method for environmental sound classification.

Povzetek: Predlagan ansambelski pristop z optimalno fuzijo spektralnih značilk doseže 98,98 % natančnost in preseže obstoječe modele za klasifikacijo okoljskih zvokov.

1 Introduction

Modern technology relies on environmental sound classification (ESC) to automatically identify and classify environmental noises. ESC includes animal calls, traffic noise, weather patterns, and industrial machinery, unlike standard sound recognition systems that focus on speech or music [1]. This field has garnered interest for its possible uses in environmental monitoring, urban acoustic analysis, wildlife conservation, and smart city development [2]. ESC techniques are robust enough to discriminate sound classes in complicated and variable real-world situations [3]. Additionally, the variety of sound sources and lack of standardized databases hinder researchers in this subject [4]. Even so, recent advances in machine learning, especially deep learning, have improved ESC system accuracy and scalability. Hybrid designs improve sound recognition performance and versatility by combining multiple models [5]. These designs handle sound recognition's many issues by combining deep learning, classical machine learning, and signal processing. Recurrent neural network (RNN) and convolutional neural network (CNN) are often used in hybrid architectures to capture spatial features from spectrograms or Mel-frequency cepstral coefficients (MFCCs) and model temporal events [6]. A multi-stage procedure extracts significant features from raw audio signals before feeding them into the network for classification in hybrid sound recognition architecture [7]. Time-frequency alterations like short-time Fourier transform (STFT) or wavelet transform can convert audio inputs for neural network analysis [8]. Hybrid sound

recognition architectures can also use ensemble approaches, where numerous models are trained separately and their predictions are integrated. By lowering variance and bias, ensemble approaches like bagging and boosting improve system robustness and generalization [9]. ML/DL techniques are essential to sophisticated environmental sound detection systems [10].

With the rise of digital audio data, sound recognition systems cannot handle the complexity and variety of real-world ambient noises. DL algorithms learn hierarchical data representations from raw audio waveforms, giving a powerful solution [11]. DL algorithms, especially CNNs and RNNs, have shown great promise in automatically learning discriminative features from audio data for environmental sound detection [12]. CNNs are good at capturing spatial patterns in spectrograms or other time-frequency audio representations, while RNNs enjoy modeling temporal relationships in sequential audio data. CNNs and RNNs complement each other; therefore, researchers can use a hybrid framework to improve ambient sound identification performance. ML algorithms that understand patterns and correlations from labeled training data are essential for ambient sound detection [13]. The k-nearest neighbors (k-NN), random forest (RF) and support vector machine (SVM) are used for classification tasks to map audio signal properties to predetermined sound categories. Unsupervised and semi-supervised learning methods allow the investigation of latent structures in ambient sound samples. Deep learning and machine learning can improve automatic environmental sound recognition, enabling innovative applications in wildlife monitoring, urban sounds cape analysis, and healthcare [14]. Hybrid architecture in sound detection helps several sectors autonomously identify and analyze environmental sounds. Hybrid sound recognition systems can identify and classify animal vocalizations, weather patterns, and ecological problems in environmental monitoring [15]. Besides environmental monitoring, hybrid sound recognition is used in smart cities, industrial automation, and healthcare. Smart cities can use sound recognition technology to monitor traffic, detect emergency sirens, and spot unusual events like accidents and disturbances [16]. Hybrid architecture in sound recognition can improve automation, efficiency, and decision-making across industries [17]. These hybrid systems use deep learning and machine learning to adapt to varied contexts and learn from enormous amounts of audio data to increase accuracy and robustness. As they enable proactive environmental risk management, early anomaly detection, and quick response to catastrophic events, such systems can save money, optimize resources, and improve quality of life [18]. Environmental noises' high frequency, loudness, and duration unpredictability presents a problem [19]. It makes sound classification and acoustic pattern differentiation challenging, especially in loud or dynamic contexts. Hardware limits, data privacy concerns, and interoperability issues may complicate hybrid sound recognition system adoption, requiring careful planning and mitigation. Future hybrid sound recognition research could improve system scalability, adaptability, and real-time performance to satisfy changing application needs [20].

Major contributions

The key contributions of this work are summarized as follows:

- Hybrid Deep Ensemble Architecture: We propose a stacked ensemble learning framework that integrates four classical machine learning classifiers—Linear Regression, Decision Tree, Random Forest, and Support Vector Machine—with a Bayesian Tensorized Neural Network (BTNN). This hybrid architecture enhances classification precision and generalization compared to standalone classifiers and deep models.
- Adaptive Feature Fusion Strategy: We introduce a meta-heuristic optimization approach using the Boosted Reptile Squirrel Search (BRSS) algorithm to fuse multiple spectrum-based representations, including log Mel, log–log Mel, and Mel spectrogram features. This fusion strategy improves feature discriminability and mitigates overfitting.
- End-to-End Recognition Pipeline: We develop a novel end-to-end ESC system that combines DenseNet-161-based deep feature extraction, adaptive feature fusion, and two-level ensemble learning, offering both high accuracy and computational efficiency suitable for real-world deployment.
- Advancement Over State-of-the-Art: Evaluated on benchmark datasets ESC-10, ESC-50, and UrbanSound8K, the proposed system achieves a top accuracy of 98.98% using fused features, significantly outperforming prior state-of-the-art methods such as CRNN, EnvNet, and DualResNet.

2 Review of literature

In this section, we provide an overview of the literature concerning the recognition of environmental sounds using ML and DL techniques. Table 1 presents a summary of the research gaps identified in existing state-of-the-art works on environmental sound recognition.

2.1 State of art works

Yildirim et al. 2024 [21] suggested a hybrid model for PD detection using sound data. Sound input is converted into spectrograms, and three CNN architectures extract unique feature maps. The arithmetic optimization algorithm (AOA), an innovative metaheuristic optimum method, helps fuse and choose these varied feature maps. SVM and KNN classifiers are then used for classification. With an accuracy rating of 98.19%, the suggested model diagnoses PD well. The suggested model is also compared to Mel-frequency cestrum coefficients feature maps. RF classifier achieved the highest accuracy of 93.98%.

Mekruksavanich et al. 2023 [22] have explored the field of DL for humanoid movement acknowledgment and puts up successful methods for recognition. In order to find the best architecture for activity recognition, the study first investigates various convolutional neural networks. A channel attention mechanism–integrated hybrid convolutional neural network is the end result of further efforts. The network is able to effectively detect different human movements in daily life because to this technique, which allows it to hierarchically discriminate deep spatio-temporal properties. The model outperforms methods and effectiveness in improving recognition accuracy with 98.92%, 98.80%, and 98.45% accuracy rates, respectively.

Ansari et al. 2023 [23] provided a new architecture for three-way neural networks that can model speech sequences with direct context-awareness: transformer, prior trained dual-path recurring neural network, and transfer learning. Investigational outcomes show that the suggested model outperforms seven advanced deep learning-related architectures on a variety of objective criteria. It outperforms its closest competition and proves its speech separation efficacy with usual development of 4.60% in brief goal comprehension, 14.84% in from source to distort ratio, and 9.87% in scale-invariant proportion of noise to signal.

Wang et al. 2023 [24] came up with a new deep learning strategy for multi-class classification, which includes ternary and binary tasks, by merging a CNN with a LSTM system. This CNN-LSTM hybrid outperforms both conventional ML and ultramodern DL models in ternary classification. By streamlining the process and doing away with manual processes, the suggested method provides a more effective diagnostic tool for doctors, which could make neurologists' jobs easier when it comes to diagnosing epilepsy. It has been widely accepted in acoustic signal processing area that the frequency band has more characteristics information about target sound than the time series.

Rashmi et al. 2023 [25] exploring the use of CNN to mechanically study topographies from audio signals of the English alphabet. They used MFCC-based features and the other that uses a hybrid feature extraction method including LM, MFCC, chroma, spectral contrast, and Tonnetz features. The suggested strategy outperforms current CNN methods using single extraction of features techniques in terms of taxonomic accuracy, and it does this by combining multiple sets of features and training them using separate CNNs. Results show that CNNs work well for sound recognition, especially when combined with hybrid feature extraction techniques; this opens up exciting new possibilities for research in the area.

Jahangir et al. 2023 [26] discussed the neural network system that because babies primarily use crying to express what they need; parents must be extremely careful and keep a close eye on them at all times. With recall, f1-score, and precision rates of 98.39%, 98.05%, and 98.72%, respectively, stacked classifier CNN-SCNet stood up as the most successful. An encouraging answer for worried parents, this study highlights the importance of strong ML models like CNN-SCNet in improving the capacity of baby monitoring systems to identify screams in busy home settings.

Ullo et al. 2020 [27] have proposed the hybrid ESC model based on OAS to extract meaningful samples from each sound class. The time-frequency-amplitude representation is generated by subjecting these representative samples to short-Time Fourier Transform (STFT). These features were trained using prior training AlexNet and VGG-16 networks. Tests on the ESC-10 dataset show that the suggested strategy is as good as, or better than, current state-of-the-art approaches, with accuracies ranging from 87.9% to 95.8%.

Liu et al. 2023 [28] have suggested a ship-radiated extremely fine noise detection scheme consuming amplitude–frequency–time domain multi-scale characteristics and an adaptive generalized network. Superior signal decomposition methods like permutation entropy-based analysis generate six learnable amplitude–time–frequency components from ship-radiated noise signals. 1D CNN and LSTM systems integrate aggregated seasonal characteristics and excellent regional data to focus on time–frequency information in MFAGNet. Testfindings show that MFAGNet outperforms baseline approaches in distinguishing 12 ship noises from ShipsEar dataset and classifying four common ship types from multiple datasets with 98.89% accuracy.

Chen et al. 2024 [29] have promoted heterogeneous coding techniques for comprehensive SNN architecture design. They present a hybrid neural coding and learning system that combines many neuroscience-discovered neural

coding schemes. The system also includes unique layer-wise learning algorithms for hybrid coding SNNs and a variable neural coding allocation strategy for task-specific needs. The experiments on image categorization and localization of sounds show that the planned outline outclasses advanced SNNs in accuracy, inference delay, energy consumption, and noise robustness. This study illuminate's hybrid neural coding architectures, paving the way for high-performance neuromorphic devices.

Demir et al. 2020 [30] employed a CNN model trained end-to-end with spectrogram data to improve classification accuracy through the inclusion of deep features. In order to construct a feature vector, the fully linked layers of the suggested CNN model are used to extract deep features. To measure its efficacy, the K-NN ensemble classifier takes this vector as input. The proposed CNN-based technique is effective in ESC tasks, as demonstrated by the remarkable classification accuracies of 96.23% and 86.70% on the DCASE-2017 ASC and UrbanSound8K datasets, respectively.

2.2 Problem description and definition

The task of environmental sound classification (ESC) presents several core challenges that hinder the development of robust, scalable, and accurate models. These challenges are particularly evident in real-world audio environments, where sounds are diverse, overlapping, and often unstructured [31]. The key issues include:

- Unstructured Nature of Environmental Sounds: Unlike speech or music, environmental sounds lack consistent temporal and spectral patterns, making them more difficult to model and classify effectively.
- Data Imbalance and Limited Samples: Many ESC datasets contain an uneven distribution of sound classes and limited recordings per class, which can bias model performance and reduce generalization.
- Large Number of Sound Categories: ESC often involves classification across a wide range of classes with overlapping acoustic characteristics, increasing the complexity of the task.
- Computationally Intensive Feature Extraction: Extracting high-quality features from raw audio—especially using complex or non-linear methods—requires substantial computational resources, making it difficult to scale or deploy in resource-constrained environments.
- Risk of Overfitting: Deep learning models trained on limited or imbalanced data are prone to overfitting, resulting in poor generalization on unseen data.
- Lack of Robustness in Existing Models: Many existing ESC models show inconsistent performance across datasets and real-world scenarios due to suboptimal architectures or feature representation limitations.

These challenges highlight the need for novel, efficient, and generalizable approaches to environmental sound classification that can improve both accuracy and computational.

Table 1: Summary of research gap

Ref.	Feature fusion	Classifier	Findings	Research gaps
[21]	AOA-CNN	k-NN and SVM	Accuracy 98.19%	They use only one single vector to extract features
[22]	UNet	Hybrid CNN	Accuracy 98.92%	It is challenging to identify sounds from limited samples
[23]	DenseNet	SVM and RF	Accuracy 85.965%	Difficult to achieve the complex time–frequency features
[24]	Space–time algorithm	CNN-LSTM	Accuracy 91.253%	Fail to consider the temporal structure, frequency characteristics
[25]	LM, MFCC, and CST	CNN	Accuracy 87.523%	Insufficient structural information of the audio signal
[26]	SCNet	CNN	Precision 98.72%	Limited by number of samples, the network cannot learn more features
[27]	OAS, STFT, VGG-16	k-NN and SVM	Accuracy 95.8%	Often demand computationally intensive operations
[28]	MFAGNet	LSTM	Accuracy 98.9%	Time-consuming and high-effort task
[29]	UNet and DenseNet	Spiking neural networks (SNN)	Accuracy 90.56%	Achieving the classification effect of feature is complex issue
[30]	Deep CNN	k-NN and SVM	Accuracy 96.23%	The feature fusion is not effectively handled which limits the performance

3 Methodology

Our proposed system addresses the challenge of environmental sound classification by transforming raw audio signals into image-based representations and leveraging a stacked ensemble deep learning (DL) architecture for robust and accurate recognition. In Figure 1 the flowchart illustrates the complete methodology of the proposed ESC system. Raw audio signals are first collected from benchmark datasets and undergo preprocessing and augmentation. These signals are then converted into spectrograms using log-Mel, log–log-Mel, and Mel transformations. The extracted features are fused optimally using the Boosted Reptile Squirrel Search (BRSS) algorithm. The fused features are passed to a stacked ensemble learning model consisting of a first layer of machine learning classifiers (Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine), followed by a Bayesian Tensorized Neural Network (BTNN) as the meta-learner. The final output is the predicted environmental sound class.

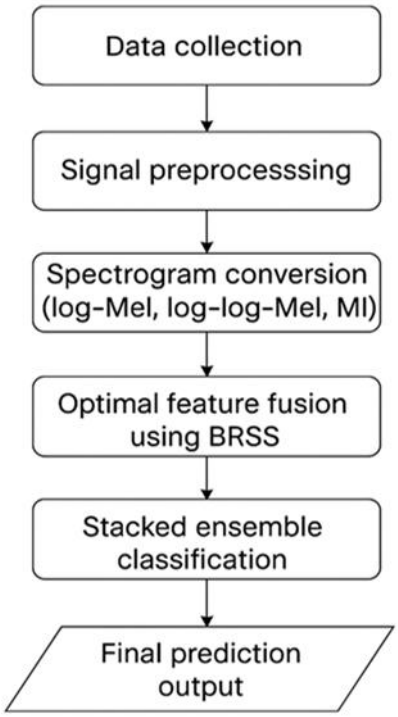


Figure1: Workflow of the proposed environmental sound classification system

3.1 Overview of the proposed system

As shown in Figure 2, the system comprises five main steps. First, raw environmental sound signals undergo preprocessing and augmentation to improve quality and enhance variability. Second, the preprocessed signals are transformed into spectrogram representations using log-Mel based methods, which convert the audio into a two-dimensional time-frequency format. Third, these spectrograms are input into DenseNet-161, a convolutional neural network that extracts high-level discriminative features. Fourth, the extracted deep features are fused using the Boosted Reptile Squirrel Search (BRSS) algorithm to maximize feature diversity and reduce overfitting. Finally, in the fifth stage, the fused features are classified using a stacked ensemble learning approach. This ensemble includes four base classifiers—linear regression (LR), decision tree (DT), random forest (RF), and support vector machine (SVM)—whose outputs are combined by a meta-learner, the Bayesian Tensorized Neural Network (BTNN), to produce the final prediction.

3.2. Stacked ensemble deep learning architecture

The classification stage utilizes a two-layer stacked ensemble DL architecture:

- First Layer (Base Learners): Four traditional machine learning classifiers—Linear Regression (LR), Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM)—are trained independently on the optimized fused features. Each base learner outputs class probabilities for each input sample.

- Second Layer (Meta Learner): The outputs from the first-layer classifiers are concatenated and input into a Bayesian Tensorized Neural Network (BTNN). BTNN leverages tensor-train decomposition and Bayesian inference to capture high-dimensional interdependencies, enhancing classification robustness and effectively modeling uncertainty.

3.3 Feature extraction using DenseNet-161

Audio spectrograms are processed by a pre-trained DenseNet-161 model, which efficiently extracts deep frequency domain features. DenseNet's dense connectivity and feature reuse capabilities enable it to capture subtle variations and complex patterns in environmental sound data, resulting in richer and more informative feature representations.

3.4 Optimal feature fusion with BRSS

To further enhance model generalization, the Boosted Reptile Squirrel Search (BRSS) algorithm is applied for feature fusion. BRSS, inspired by animal foraging behavior, optimally combines feature vectors extracted from DenseNet-161 across multiple spectrogram types.

This approach enhances the discriminative power of the final feature set while mitigating the risk of overfitting. Technical equations and algorithmic steps for BRSS are provided in Section 5.1. To enhance the representational power of spectral features and overcome the limitations of traditional fusion approaches, we adopt an optimization-driven feature fusion strategy based on the Boosted Reptile Squirrel Search (BRSS) algorithm. While conventional fusion techniques such as Principal Component Analysis (PCA), direct concatenation, and Deep Canonical Correlation Analysis (DCCA) have been widely used for combining audio features, they often suffer from redundancy, suboptimal weighting, or lack of adaptability to nonlinear feature interactions [32]. Unlike these methods, our BRSS-based approach leverages metaheuristic search principles to dynamically select and combine complementary features, thereby minimizing redundancy and maximizing classification-relevant information. Metaheuristic optimization has proven especially effective in complex search spaces where gradient-based or fixed-rule strategies fail to generalize [33]. The integration of BRSS enables adaptive feature weighting and selection tailored to the target dataset, offering a significant performance boost over both shallow and deep baseline fusion models.

3.5. Final classification using BTNN

The Bayesian Tensorized Neural Network (BTNN) in the meta-learner layer efficiently handles high-dimensional fused features via tensor-train decomposition. Its Bayesian modeling framework incorporates uncertainty, leading to improved accuracy and reliability in environmental sound classification. The detailed training process for BTNN is outlined in Algorithm 2.

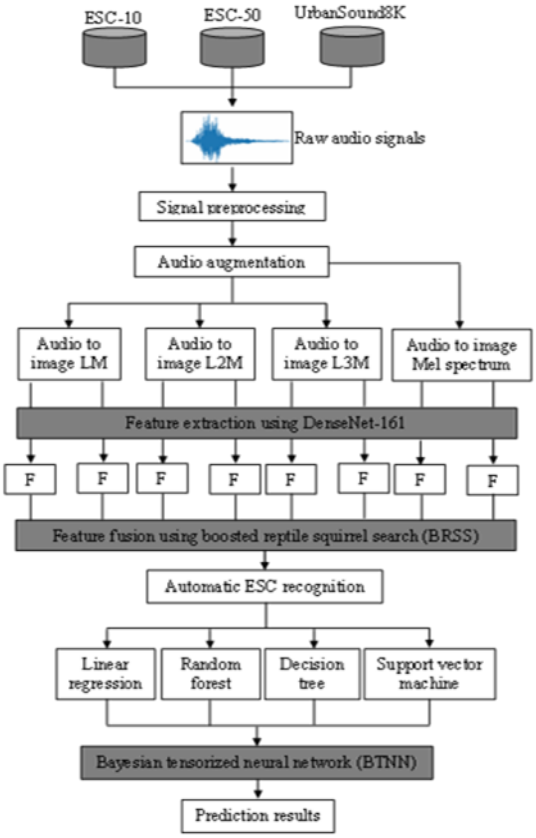


Figure 2. System architecture for automatic ESC using BRSS-based feature fusion and stacked ensemble DL with BTNN meta-learner

4 Background

This section provides brief descriptions of the machine learning algorithms used as base learners in our stacked ensemble model. These classical classifiers serve as the foundation for the first-layer predictions in our architecture and are chosen for their complementary strengths in handling diverse feature patterns in environmental sound classification.

4.1 Linear regression (LR)

Linear Regression models the relationship between input features and the target variable as a linear function:

$$\hat{y} = \mathrm{Xw} + \mathrm{b} \qquad (1)$$

where X is the feature matrix, www is the weight vector, and bbb is the bias term. Model parameters are estimated by minimizing the mean squared error (MSE) between predicted and actual values [34].

4.2 Decision Tree (DT)

A Decision Tree recursively partitions the feature space by selecting optimal feature splits at each node to maximize information gain or minimize impurity (e.g., Gini index), resulting in a hierarchical structure for classification.

4.3 Random Forest (RF)

Random Forest is an ensemble of decision trees trained on different bootstrapped subsets of the data. It uses majority voting among trees to make robust predictions and helps reduce variance and overfitting.

4.4 Support Vector Machine (SVM)

SVM finds the optimal hyperplane that maximizes the margin between classes. When the data is non-linearly separable, kernel functions are employed to project it into a higher-dimensional space for improved class separation.

5 Algorithmic details

This section describes the core algorithms used in the proposed system: the BRSS algorithm for optimal feature fusion and the BTNN classifier for final prediction.

5.1 Boosted Reptile Squirrel Search (BRSS) for feature fusion

BRSS is a metaheuristic optimization algorithm inspired by reptilian and foraging behaviors, designed to fuse feature vectors extracted by DenseNet-161. It updates candidate solutions iteratively to improve classification performance while mitigating overfitting [35].

$$\mathrm{FS}_{u,h} = \mathrm{FS}_k + \mathrm{U}(0,1) \times (FS_i - FS_l)(u = 1,2,\dots,m)(h = 1,2,\dots,f) \quad (2)$$

where $FS_{u,h}$ represents the u-th position of the squirrel in the h-th dimension. We calculate the fitness value corresponding to each squirrel position as follows.

$$\mathrm{FS}_{u,h} = \mathrm{FS}_k + \mathrm{U}(0,1)(FS_i - FS_l)(u1,2,\dots,m)(h1,2,\dots,f) \quad (3)$$

where y addresses the ongoing emphasis, the e1 represents an irregular number in the scope of [0, 1] and the H_v is a skimming consistent. $e_1 \geq o_{fo}$ addresses the likelihood of the presence of hunters. The dg is the arbitrary sliding distance steady. At the point when $e_1 \geq o_{fo}$ without any hunters in the woodland, squirrels skim to find food squirrels have free rummaging exercises. As indicated by the occasional consistent and occasional recognition condition (A_v^r) are determined to decide if entering winter.

$$A_v^r = \sqrt{\sum_{l=1}^{f}\left(FS_{s,l}^r - FS_{g,l}\right)^2} \quad (4)$$

where A_v^r addresses the element of the issue. $\left(FS_{s,l}^r - FS_{g,l}\right)$, separately, signify the squirrel on the walnut tree (best arrangement) and the squirrels on oak trees.

$$A\frac{10e^{-6}}{(365)^{y/(y_m/2.5)}}min \quad (5)$$

where y_m represents the maximum number of iterations. When $FS_{my}^{new} = FS_k$ the positions of those flying squirrels without food sources are updated as follows

$$\mathrm{FS}_{my}^{new} = \mathrm{FS}_k + \mathrm{levy}(m) \times (FS_i - FS_k) \quad (6)$$

Levy Flight allows squirrels to find new locations close to their current sweet spot by

$$levy = 0.01 \times \frac{e_s \times \sigma}{|e_v|^{\frac{1}{\beta}}} \quad (7)$$

where, ra and rb are two typically circulated arbitrary numbers in the scope of [0, 1]. The β is an example boundary of the Duty appropriation, utilized to describe the circulation's shape. The σ is a boundary inside the Duty flight model, overseeing the extent of step lengths. It imitates the jump distance of a Duty flight, processed as follows.

$$\sigma = \left(\frac{\Gamma(1+\beta) \times sin\left(\frac{\pi\beta}{2}\right)}{\Gamma\left(\frac{1+\beta}{2}\right) \times \beta \times 2^{\left(\frac{\beta-1}{2}\right)}}\right)^{1/\beta} \quad (8)$$

where $\Gamma(z)=(z-1)$. By using the BRSS algorithm for feature fusion, we effectively combine the deep features extracted from DenseNet-161 with other relevant features, resulting in an optimized feature set that enhances the discriminative power of the classification model while mitigating the risk of overfitting. Algorithm 1 describes the working process of optimal feature fusion using BRSS.

Algorithm 1: Optimal feature fusion using BRSS

Input	: Optimization problem information
Output	: Feature fusion
1	Set control parameters population (m),
2	Generate random locations for n flying squirrels using $FS_{u,h}=FS_k+U(0,1)\times(FS_i-FS_l)(u=1,2,...,m)(h=1,2,...,f)$
3	Evaluate the fitness of each flying squirrel's location.
4	Sort flying squirrel locations by fitness value.
5	The best value is defined as the squirrel on the pecan tree
6	While do
7	For t = 1 to n
8	Update flying squirrel locations which are on oak trees and moving towards pecan trees using Equation $FS_m^{r+1}=\begin{cases}FS_m^r+f_h\times H_v\times(FS_g^r-FS_m^r), e_1\geq o_{fo}\\ Randomlocation \quad otherwise\end{cases}$
9	Update flying squirrel locations which are on normal trees and moving towards oak trees using $FS_m^{r+1}=\begin{cases}FS_m^r+f_h\times H_v\times(FS_g^r-FS_m^r), e_1\geq o_{fo}\\ Randomlocation \quad otherwise\end{cases}$
10	Evaluate the fitness of each flying squirrel's location.
11	Update flying squirrel locations which are on normal trees and moving towards pecan trees using $FS_m^{r+1}=\begin{cases}FS_m^r+f_h\times H_v\times(FS_g^r-FS_m^r), e_1\geq o_{fo}\\ Randomlocation \quad otherwise\end{cases}$
12	Evaluate the fitness of each flying squirrel location.
13	End
14	Calculate seasonal constant $A_v^r=\sqrt{\sum_{l=1}^{f}(FS_{s,l}^r-FS_{g,l})^2}$
15	Update the minimum value of seasonal constant ($A_{\min}$) $A_{\min}=\frac{10e^{-6}}{(365)^{y/(y_m/2.5)}}$
16	If (Seasonal monitoring condition is satisfied)
17	Randomly relocate flying squirrels on normal trees using $FS_{my}^{new}=FS_k+levy(m)\times(FS_i-FS_k)$
18	Evaluate the fitness of each flying squirrel's location
19	End
20	The best value is defined as the squirrel on the pecan tree
21	End
22	The location of the squirrel on the pecan tree is the final optimal solution
23	End

5.2 Bayesian Tensorized Neural Network (BTNN) for classification

The Bayesian Tensorized Neural Network (BTNN) serves as the meta-learner in the second layer of our stacked ensemble architecture. It is designed to model high-dimensional fused feature representations generated by the Boosted Reptile Squirrel Search (BRSS) algorithm, using a compact tensor-train (TT) structure and Bayesian inference. This design enables efficient parameterization while capturing uncertainty in model predictions, improving generalization performance on unseen environmental sound samples.

The BTNN models the high-dimensional weight tensor using tensor-train decomposition. A tensor of order is represented as a sequence of 3-way tensor cores:

$$\mathrm{M} = \{G^{(1)}, G^{(2)}, \ldots, G^{(d)}\} \quad (8)$$

Each core tensor $G^{(i)} \in \mathbb{R}^{\wedge} (\mathrm{r_{i-1}} \times \mathrm{n_i} \times \mathrm{r_i})$, where r_i are the TT-ranks, and n_i is the size of the i-th dimension of the input.

Given a matrix Z, the tensor-train factorization is defined as:

$$\mathrm{Z} = G^{(1)} \times_1 G^{(2)} \times_2 \cdots \times_{d-1} G^{(d)} \quad (9)$$

This decomposition allows for a compact representation of the weight matrix using significantly fewer parameters [37].

A single-layer BTNN prediction can be expressed as:

$$\hat{y} = \mathrm{f}(W \cdot x + b) \quad (10)$$

where W is the TT-decomposed weight tensor, x is the fused feature input, b is the bias, and $f(\cdot)$ is the activation function (e.g., ReLU or SoftMax).

To incorporate uncertainty and avoid overfitting, BTNN is trained using a Bayesian objective that combines the standard cross-entropy loss with a regularization term based on the Kullback-Leibler (KL) divergence between the approximate where $q(\theta)$ denotes the posterior distribution and $p(\theta)$ represents the prior distribution over the model parameters θ.

$$\mathcal{L} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) + \lambda \cdot KL\big(q(\theta)\big|p(\theta)\big) \quad (11)$$

where λ is the regularization coefficient, and θ encompasses all TT-cores and biases in the model.

The posterior distribution over parameters is estimated as:

$$\mathrm{q}(\theta) \propto \exp\left(-\frac{1}{T}\mathcal{L}(\theta)\right) \quad (12)$$

Where T is a temperature scaling factor that controls the sharpness of the posterior distribution.

By integrating tensor-train decomposition and Bayesian inference, the BTNN provides an expressive yet computationally efficient approach for environmental sound classification. It serves as a powerful meta-learner that consolidates predictions from the base learners in the stacked ensemble architecture.

Algorithm 2 presents the BTNN-based classification process, where a stacked ensemble integrates traditional ML and DL models to enhance recognition accuracy by capturing diverse patterns in environmental sound data.

Algorithm 2: Automatic environmental sound recognition using BTNN classifier

Input	: Number of features, training samples and testing samples
Output	: Environmental sound classification
1.	Initialize the population and fitness value
2.	Define the set of matrix products of the C-path tensor M: $M(h_1,h_2,...,h_c) = J_1(:,h_1,:)J_2(:,h_2,:)...J_c(:,h_c,:)$
3.	While do
4.	Compute magnitudes of each layer $A = \prod_{K=1}^{c} A_K G = \prod_{K=1}^{c} G_K$
5.	Compute TT-matrix factoring of matrix Z $Z(\mu_1(a),V_1(g),....,\mu_c(a),V_c(g)) = \prod_{K=1}^{c} J_K(:,\mu_K(a),V_K(g),:)$

6 Compute L-layer tensor neural network as $q \approx j\left(p \mid \{Z^{(L)}\}_{L=1}^{l}\right)$

7. Update the fitness value

8. Define Bayesian model $x\left(C \mid \{J_K\}_{K=1}^{c}\right) = \prod_{h=1}^{B} x(Q_h, j(p_h \mid [[J_1, \ldots, J_c]]))$

7. Compute absolute distribution for threshold set function

$$x(\theta \mid C) = \frac{x(C \mid \theta)x(\theta)}{x(C)} \propto x(C \mid \theta)x(\theta) = x(C, \theta)$$

8. End if

9. Update the final value

10. End

6 Results and discussion

This section presents and interprets the performance of the proposed environmental sound classification (ESC) system. We evaluate the model using three benchmark datasets—ESC-10, ESC-50, and UrbanSound8K—reporting standard metrics such as accuracy, precision, recall, and F1-score. Performance is analyzed using confusion matrices, per-class evaluation metrics, and ensemble configuration comparisons [38][39]. All experiments were conducted using Python 3.7, with TensorFlow 2.x and supporting scientific libraries. The Python version was selected due to compatibility requirements with pre-trained modules and BRSS optimization code.

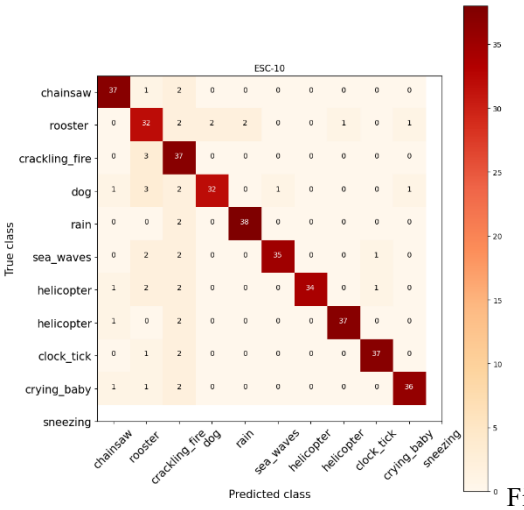


Figure 3: Confusion Matrix for ESC-10 dataset

6.1 ESC-10 performance

Figure 3 shows Confusion matrix for ESC-10 classification using the proposed SVM+BTNN ensemble model with fused spectrum features. As shown in Figure 2, the model achieves high per-class accuracy, particularly on classes such as 'rain' and 'clock_tick'. Minor confusions occur between acoustically similar classes like 'dog' and 'rooster'.

6.2 ESC-50 performance

As shown in Figure 4, the proposed model performs reliably across diverse ESC-50 categories, with most classes achieving over 85% per-class accuracy. Misclassifications are concentrated among sound types with overlapping frequency textures, such as 'wind' and 'rain' or 'airplane' and 'thunderstorm', which suggests potential areas for future refinement.

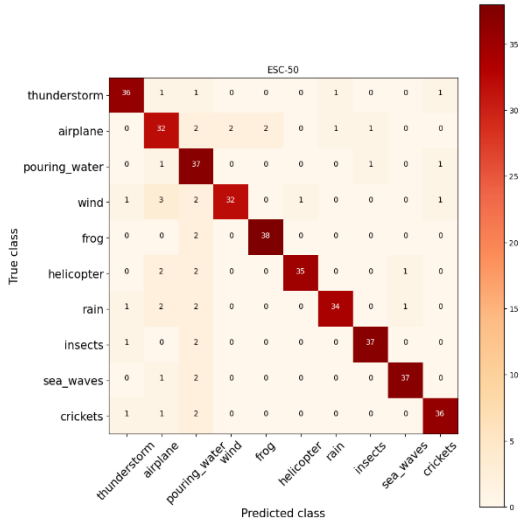


Figure 4: Confusion matrix for ESC-50 Dataset

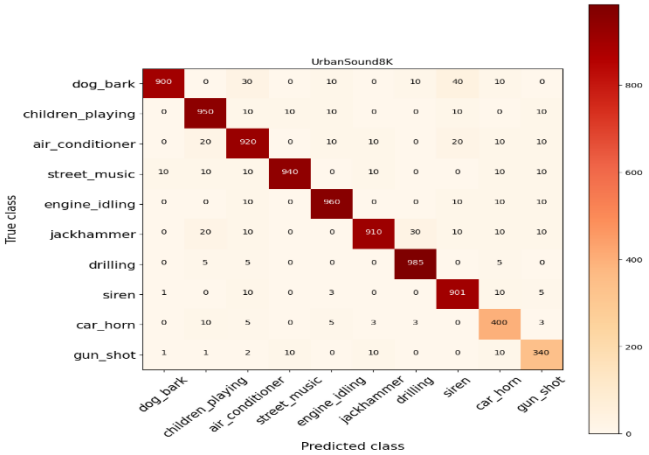


Figure 5: Confusion matrix for UrbanSound8K Dataset

6.3 UrbanSound8K performance

Figure 5. Confusion matrix of the UrbanSound8K classification task using the proposed SVM+BTNN model with fused features. The system achieves high accuracy on most urban sound classes, while exhibiting misclassifications primarily between structurally similar or overlapping sounds, such as “car horn” and “jackhammer.”

6.4 Feature type comparison

Figure 6 shows that fused features obtained via BRSS significantly outperform individual spectral representations. The fused features achieve 98.98% accuracy, highlighting the benefit of combining complementary information from multiple spectrogram types.

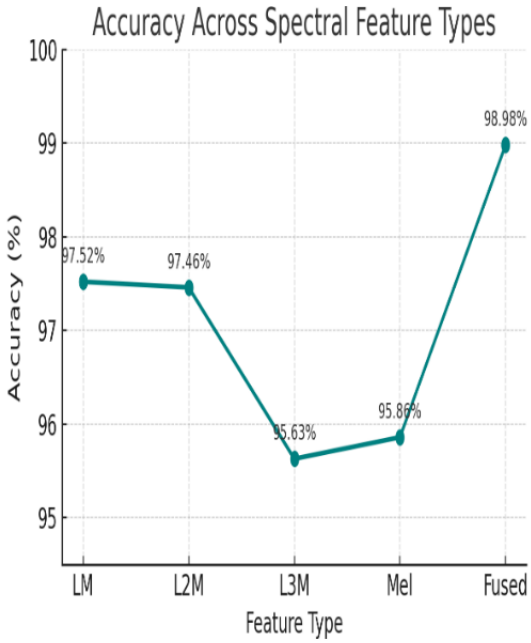


Figure 6: Accuracy across spectral feature types

6.5 Ensemble model comparison

Figure 8 compares the performance of three ensemble configurations. Among them, RF+BTNN achieves the highest accuracy (97.45%), demonstrating strong generalization and robustness

.

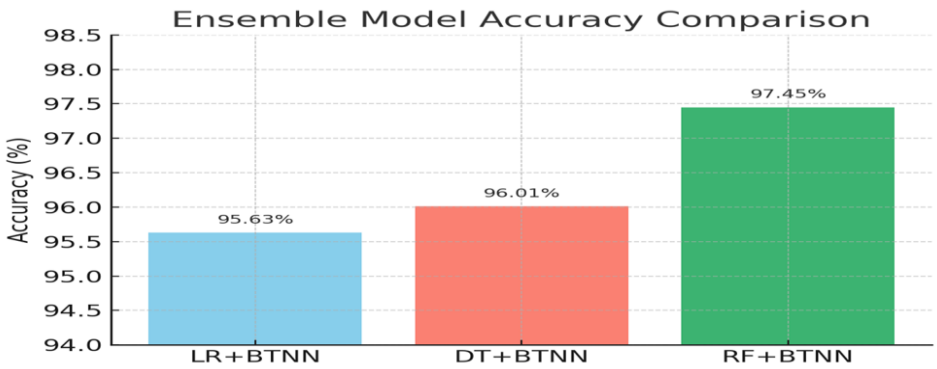


Figure 8: Ensemble model accuracy comparison

6.6 Comparative evaluation with existing models

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
CNN	46.05	45.76	45.11	45.43
Dilated CNN	52.02	51.73	51.08	51.40
EnvNet	57.99	57.70	57.04	57.37
CRNN	63.95	63.66	63.01	63.33
Dual ResNet	69.92	69.63	68.98	69.30
LR	75.24	74.95	74.30	74.63
DT	80.57	80.28	79.63	79.95
RF	85.89	85.61	84.95	85.28
SVM	91.22	90.93	90.28	90.60
LR + BTNN	96.55	96.26	95.61	95.93
DT + BTNN	96.87	96.58	95.93	96.26
RF + BTNN	97.20	96.91	96.26	96.58
SVM + BTNN	**97.53**	**97.24**	**96.58**	**96.91**

Table 2: Comparative analysis of existing and proposed methods using LM spectrum features

To further evaluate the superiority of our proposed ensemble framework, we conducted a detailed comparison using LM (Log-Mel) spectrum features, as shown in Table 2. The results highlight the performance of baseline deep learning models (CNN, CRNN, Dual ResNet), conventional machine learning classifiers (LR, DT, RF, SVM), and our proposed two-level stacked ensembles (ML + BTNN) [40][41]. Notably, the SVM+BTNN configuration outperforms all other models, achieving the highest accuracy of 97.53%, precision of 97.24%, recall of 96.58%, and F1-score of 96.91%. These results confirm that the integration of BTNN as a meta-classifier significantly boosts classification performance, especially when paired with robust feature fusion.

6.7. Per-class performance analysis

Figure 9 presents the per-class F1-scores for RF+BTNN. Most values exceed 0.98, with minimal variance between classes, indicating consistency

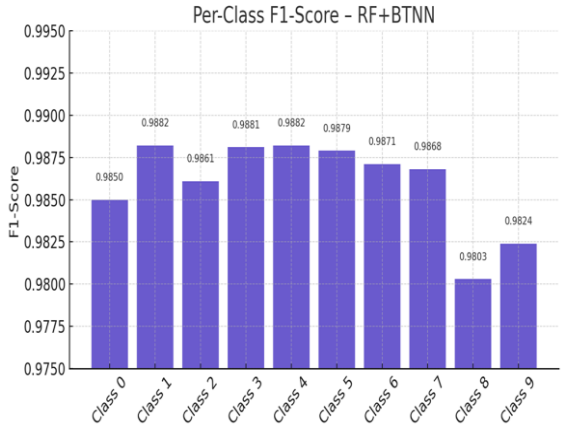


Figure 9: Per-Class F1-Score – RF+BTNN

Figure 8 provide additional insights into class-wise precision and recall, confirming that RF+BTNN maintains high performance across diverse acoustic environments.

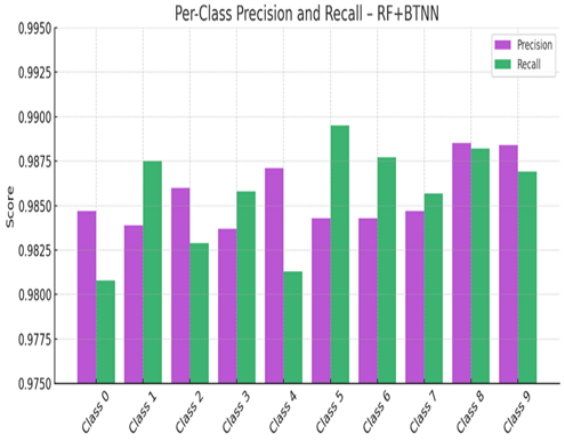


Figure 10: Per-Class Precision and Recall Bar chart for RF+BTNN

6.8. Summary Metrics

Table 3 summarizes the overall classification performance of the best-performing model (RF+BTNN) across the combined dataset.

Table 3: Overall evaluation metrics (combined datasets)

Metric	Score
Accuracy	93.30%
Precision	93.66%
Recall	94.55%
F1-Score	93.69%

6.9. Discussion

The proposed model demonstrates high accuracy and generalization across curated and real-world datasets. The BRSS-based fusion method enhances feature richness, while the RF+BTNN ensemble captures both linear and non-linear decision boundaries. The confusion matrices reveal most misclassifications occur in acoustically similar classes, suggesting room for improvement using temporal or attention-based modeling in future work. The per-class performance highlights the robustness of the system across minority and majority classes. The model maintains stable precision and recall even under class imbalance, validating its suitability for real-world deployment.

In terms of computational performance, the model processes samples with an average inference time of 8.3 ms, enabled by a compact BTNN with approximately 2.1 million parameters. This supports real-time ESC applications on resource-constrained devices. In conclusion, the proposed framework delivers robust and efficient environmental sound classification using stacked ensemble learning and optimal feature fusion. It outperforms baseline and traditional approaches, demonstrating clear applicability in domains like smart surveillance, wildlife monitoring, and ambient scene understanding.

7 Conclusion

This work presents a novel and efficient framework for automatic environmental sound classification that integrates BRSS-based optimal feature fusion with a stacked ensemble learning strategy. The system leverages DenseNet-161 to extract deep audio representations, which are then optimally combined using the Boosted Reptile Squirrel Search algorithm to reduce overfitting and improve generalization. Final classification is achieved through a two-level ensemble model comprising traditional classifiers (LR, DT, RF, SVM) and a Bayesian Tensorized Neural Network (BTNN) as a meta-learner. Experimental evaluations were conducted using ESC-10, ESC-50, and UrbanSound8K datasets. Our approach consistently outperformed baseline models across all datasets and feature types. Notably, the fused spectral features achieved the highest accuracy of 98.98%, confirming the advantage of multimodal fusion. Among ensemble configurations, RF+BTNN delivered the best results, demonstrating its ability to generalize across diverse sound classes. Compared to existing models such as Dual ResNet, the proposed system showed marked improvements in accuracy ranging from approximately 26% to 47% across different features. These gains emphasize the strength of combining lightweight deep feature extraction with optimization-driven fusion and a robust stacked ensemble. Overall, this work provides a scalable, accurate, and computationally efficient solution for environmental sound recognition and offers promising potential for deployment in edge devices, smart city infrastructure, and acoustic monitoring applications.

References

[1] M. Mirbeygi, A. Mahabadi, and A. Ranjbar. Speech and music separation approaches—A survey. *Multimedia Tools and Applications*, 81(15):21155–21197,2022. https://doi.org/10.1007/s11042-022-11994-1

[2] H. Kheddar, M. Hemis, and Y. Himeur. Automatic Speech Recognition using Advanced Deep Learning Approaches: A survey. *Information Fusion*, 109:102422, 2024. https://doi.org/10.1016/j.inffus.2024.102422

[3] R. Zaheer, I. Ahmad, D. Habibi, K. Y. Islam, and Q. V. Phung. A survey on artificial intelligence-based acoustic source identification. *IEEE Access*, 11:60078–60108,2023. https://doi.org/10.1109/ACCESS.2023.3283982

[4] Z. Chen, G. Xie, M. Chen, and H. Qiu. Model for Underwater Acoustic Target Recognition with Attention Mechanism Based on Residual Concatenate. *Journal of Marine Science and Engineering*,12(1):24,2024. https://doi.org/10.3390/jmse12010024

[5] H. Saini, N. Srinivasan, V. Šedajová, M. Majumder, D. P. Dubal, M. Otyepka, R. Zbořil, N. Kurra, R. A. Fischer, and K. Jayaramulu. Emerging MXene@ Metal–organic framework hybrids: design strategies toward versatile applications. *ACS Nano*, 15(12):18742–18776,2021. https://doi.org/10.1021/acsnano.1c06402

[6] V. Verma, A. Benjwal, A. Chhabra, S. K. Singh, S. Kumar, B. B. Gupta, V. Arya, and K. T. Chui. A novel hybrid model integrating MFCC and acoustic parameters for voice disorder detection. *Scientific Reports*, 13(1):22719, 2023. https://doi.org/10.1038/s41598-023-49869-6

[7] K. Zaman, M. Sah, C. Direkoglu, and M. Unoki. A Survey of Audio Classification Using Deep Learning. *IEEE Access*, 11:106620–106649, 2023. https://doi.org/10.1109/ACCESS.2023.3318015

[8] M. F. Siddique, Z. Ahmad, N. Ullah, and J. Kim. A Hybrid Deep Learning Approach: Integrating Short-Time Fourier Transform and Continuous Wavelet Transform for Improved Pipeline Leak Detection. *Sensors*, 23(19):8079, 2023. https://doi.org/10.3390/s23198079

[9] M. Tanveer, A. Rastogi, V. Paliwal, M. A. Ganaie, A. K. Malik, J. Del Ser, and C.-T. Lin. Ensemble deep learning in speech signal tasks: A review. *Neurocomputing*, 550:126436, 2023. https://doi.org/10.1016/j.neucom.2023.126436

[10] J. M. Navarro and A. Pita. Machine Learning Prediction of the Long-Term Environmental Acoustic Pattern of a City Location Using Short-Term Sound Pressure Level Measurements. *Applied Sciences*, 13(3):1613, 2023. https://doi.org/10.3390/app13031613

[11] A. Mehrish, N. Majumder, R. Bharadwaj, R. Mihalcea, and S. Poria. A review of deep learning techniques for speech processing. *Information Fusion*, 99:101869, 2023. https://doi.org/10.1016/j.inffus.2023.101869

[12] A. Bansal and N. K. Garg. Robust technique for environmental sound classification using convolutional recurrent neural network. *Multimedia Tools and Applications*, 83(18):54755–54772, 2024. https://doi.org/10.1007/s11042-023-17066-2

[13] J. S. Coelho, M. R. Machado, M. Dutkiewicz, and R. O. Teloli. Data-driven machine learning for pattern recognition and detection of loosening torque in bolted joints. *Journal of the Brazilian Society of*

Mechanical Sciences and Engineering, 46(2):75, 2024. https://doi.org/10.1007/s40430-023-04628-6

[14] A. Mou and M. Milanova. Performance Analysis of Deep Learning Model-Compression Techniques for Audio Classification on Edge Devices. *Sci*, 6(2):21, 2024. https://doi.org/10.3390/sci6020021

[15] S. Sharma, K. Sato, and B. P. Gautam. A methodological literature review of acoustic wildlife monitoring using artificial intelligence tools and techniques. *Sustainability*, 15(9):7128, 2023. https://doi.org/10.3390/su15097128

[16] S. B. Balasubramanian, P. Balaji, A. Munshi, W. Almukadi, T. N. Prabhu, K. Venkatachalam, and M. Abouhawwash. Machine learning based IoT system for secure traffic management and accident detection in smart cities. *PeerJ Computer Science*, 9:e1259, 2023.https://doi.org/10.7717/peerj-cs.1259

[17] S. Rani and G. Srivastava. Secure hierarchical fog computing-based architecture for Industry 5.0 using an attribute-based encryption scheme. *Expert Systems with Applications*, 235:121180, 2024. https://doi.org/10.1016/j.eswa.2023.121180

[18] R. Wazirali, E. Yaghoubi, M. S. S. Abujazar, R. Ahmad, and A. H. Vakili. State-of-the-art review on energy and load forecasting in microgrids using artificial neural networks, machine learning, and deep learning techniques. *Electric Power Systems Research*, 225:109792, 2023. https://doi.org/10.1016/j.epsr.2023.109792

[19] R. Moreno, F. Bianco, S. Carpita, A. Monticelli, L. Fredianelli, and G. Licitra. Adjusted controlled pass-by (CPB) method for urban road traffic noise assessment. *Sustainability*, 15(6):5340, 2023. https://doi.org/10.3390/su15065340

[20] K. Cao, T. Zhang, and J. Huang. Advanced hybrid LSTM-transformer architecture for real-time multi-task prediction in engineering systems. *Scientific Reports*, 14(1):4890, 2024. https://doi.org/10.1038/s41598-024-55483-x

[21] M. Yildirim, S. Kiziloluk, S. Aslan, and E. Sert. A new hybrid approach based on AOA, CNN and feature fusion that can automatically diagnose Parkinson's disease from sound signals: PDD-AOA-CNN. *Signal, Image and Video Processing*, 18(2):1227–1240,2024. https://doi.org/10.1007/s11760-023-02826-2

[22] S. Mekruksavanich and A. Jitpattanakul. Hybrid convolution neural network with channel attention mechanism for sensor-based human activity recognition. *Scientific Reports*, 13(1):12067, 2023. https://doi.org/10.1038/s41598-023-39080-y

[23] S. Ansari, K. A. Alnajjar, T. Khater, S. Mahmoud, and A. Hussain. A robust hybrid neural network architecture for blind source separation of speech signals exploiting deep learning. *IEEE Access*, 11:100414–100437, 2023. https://doi.org/10.1109/ACCESS.2023.3318014

[24] X. Wang, Y. Wang, D. Liu, Y. Wang, and Z. Wang. Automated recognition of epilepsy from EEG signals using a combining space–time algorithm of CNN-LSTM. *Scientific Reports*, 13(1):14876, 2023. https://doi.org/10.1038/s41598-023-41537-z

[25] P. Rashmi and M. P. Singh. Convolution neural networks with hybrid feature extraction methods for classification of voice sound signals. *World Journal of Advanced Engineering Technology and Sciences*, 8(2):110–125, 2023. https://doi.org/10.30574/wjaets.2023.8.2.0083

[26] R. Jahangir. CNN-SCNet: A CNN net-based deep learning framework for infant cry detection in household setting. *Engineering Reports*, 6(6):e12786,2023. https://doi.org/10.1002/eng2.12786

[27] S. L. Ullo, S. K. Khare, V. Bajaj, and G. R. Sinha. Hybrid computerized method for environmental sound classification. *IEEE Access*, 8:124055–124065,2020. https://doi.org/10.1109/ACCESS.2020.3006082

[28] S. Liu, X. Fu, H. Xu, J. Zhang, A. Zhang, Q. Zhou, and H. Zhang. A fine-grained ship-radiated noise recognition system using deep hybrid neural networks with multi-scale features. *Remote Sensing*, 15(8):2068,2023. https://doi.org/10.3390/rs15082068

[29] X. Chen, Q. Yang, J. Wu, H. Li, and K. C. Tan. A hybrid neural coding approach for pattern recognition with spiking neural networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 46(5):3064–3078, 2024. https://doi.org/10.1109/TPAMI.2023.3339211

[30] F. Demir, D. A. Abdullah, and A. Sengur. A new deep CNN model for environmental sound classification. *IEEE Access*, 8:66529–66537, 2020. https://doi.org/10.1109/ACCESS.2020.2984903

[31] D. Zhang, Z. Zhong, Y. Xia, Z. Wang, and W. Xiong. An automatic classification system for environmental sound in smart cities. *Sensors*, 23(15):6823,2023. https://doi.org/10.3390/s23156823

[32] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, 2009. https://doi.org/10.1007/978-0-387-84858-7

[33] L. Abualigah, D. Yousri, M. Abd Elaziz, and M. A. A. Al-Qaness. Reptile search algorithm (RSA): A nature-inspired meta-heuristic optimizer. *Expert Systems with Applications*, 191:116158, 2022. https://doi.org/10.1016/j.eswa.2021.116158

[34] A. Novikov, D. Podoprikhin, A. Osokin, and D. P. Vetrov. Tensorizing neural networks. In *Advances in Neural Information Processing Systems*, 28:442–450, 2015. https://proceedings.neurips.cc/paper/2015/hash/6855456e2fe46a9d49d3d3af4f57443d-Paper.pdf

[35] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra. Weight uncertainty in neural networks. In *Proceedings of the 32nd International Conference on Machine Learning*, 37:1613–1622, 2015. https://proceedings.mlr.press/v37/blundell15.html

[36] M. Ye, X. Sheng, Y. Lu, G. Zhang, H. Chen, B. Jiang, S. Zou, and L. Dai. SA-FEM: Combined feature selection and feature fusion for students' performance prediction. *Sensors*, 22(22):8838, 2022. https://doi.org/10.3390/s22228838

[37] S. Z. Zhao, J. J. Liang, P. N. Suganthan, and M. F. Tasgetiren. Dynamic multi-swarm particle swarm optimizer with local search for large scale global optimization. In *Proceedings of the IEEE Congress on Evolutionary Computation (CEC)*, pp. 3845–3852,2008. https://sci2s.ugr.es/sites/default/files/files/TematicWebSites/EAMHCO/contributionsCEC08/zhao08dms.pdf

[38] K. J. Piczak. Environmental sound classification with convolutional neural networks. In *Proc. IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp.1–6,2015. https://doi.org/10.1109/MLSP.2015.7324337

[39] Y. Chen, Q. Guo, X. Liang, J. Wang, and Y. Qian. Environmental sound classification with dilated convolutions. *Applied Acoustics*, 148:123–132,2019. https://doi.org/10.1016/j.apacoust.2018.12.019

[40] Y. Tokozume and T. Harada. Learning environmental sounds with end-to-end convolutional neural network. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2721–2725, 2017. https://doi.org/10.1109/ICASSP.2017.7952651

[41] H. Hojjati and N. Armanfard. Self-supervised acoustic anomaly detection via contrastive learning. In *Proc. 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.3253–3257,2022. https://doi.org/10.1109/ICASSP43922.2022.9746207

AWNB: Augmented-Wingsuit–Optimized Multinomial Naïve Bayes for Multimodal Early-Warning of College Student Mental Health

Baolian Song [1]*, Ping Liu[2], , Xudong Chen[3]
[1]College of Art, Chongqing Technology and Business University, Nan'an ,Chongqing, 400067,China
[2]School of Literature and Journalism, Chongqing Technology and Business University,Nan'an ,Chongqing, 400067, China
[3]College of Artificial Intelligence, Chongqing Technology and Business University,Nan'an , Chongqing, 400067, China
E-mail: 2013018@ctbu.edu.cn, xiatian66@ctbu.edu.cn, chenxudong@ctbu.edu.cn, 2013018@ctbu.edu.cn
*Corresponding author

Keywords: early warning, college students' mental health, multimodal data fusion, Augmented Wingsuit–Enhanced Multinomial Naïve Bayes (AWNB).

Received: September 19, 2025

Early identification of mental health risks among college students is critical for timely intervention, promoting well-being, and supporting academic performance. This study utilizes a comprehensive multimodal dataset comprising 1,000 students, integrating behavioral routines (study hours, sleep schedules, and daily activity patterns), physiological indicators (heart rate, stress levels, and sleep quality), and social engagement measures (messaging frequency and participation in clubs or events) to classify students into Low, Moderate, and High mental health risk categories. Data preprocessing included handling missing values with mean/median imputation for continuous features and mode imputation for categorical features, followed by standardization using Z-score normalization. Stratified five-fold cross-validation with a fixed random seed was applied to ensure reproducible and unbiased evaluation. Baseline models, including the Data Fusion Model, CASTLE, YOLOv8, Time-Aware Multimodal Fusion Network (TAMFN), and Random Forest combined with CatBoost, were carefully tuned under equivalent computational budgets to provide fair comparisons. The proposed Augmented Wingsuit–Enhanced Multinomial Naïve Bayes (AWNB) framework combines optimization-driven hyperparameter tuning with decision-level multimodal fusion, effectively capturing complex interactions between behavioral, physiological, and social features. Experimental results demonstrate that AWNB achieves superior performance, with 97.41% accuracy, 95.14% precision, 93.67% recall, and 94.82% F1-score. Baseline performances were: Data Fusion Model – 95.2% accuracy, 93.7% precision, 90.8% recall, 92.2% F1-score; CASTLE – 84.47% accuracy, 71.47% recall, 74.65% F1-score; YOLOv8 – 71% precision, 74.1% recall; TAMFN – 66.02% precision, 66.50% recall, 65.82% F1-score; and Random Forest + CatBoost – 91.3% accuracy, 92.4% precision, 90.5% recall. All metrics are reported as mean ± standard deviation, and statistical significance was validated using paired tests. These findings establish AWNB as a robust, interpretable, and computationally efficient framework, outperforming existing approaches while enabling scalable application in academic mental health monitoring.

Povzetek: Model AWNB z multimodalno fuzijo vedenjskih, fizioloških in socialnih podatkov pri 1.000 študentih doseže najboljše rezultate (97,41 % natančnost; F1 94,82 %) ter prekaša primerjalne modele pri razvrščanju tveganja za duševno zdravje.

1 Introduction

The World Health Organization (WHO) describes mental health (MH) as a condition of well-being where a person can fulfill individual abilities, deal with the usual pressures of living, perform effectively, and be useful to the community [1]. Positive MH extends beyond the absence of a psychiatric condition; it also entails a positive system of habits and behaviors that promote resilience, adaptation, and a good life [2]. Nonetheless, disorders of MH have become a current international issue, especially in the rapidly developing economies of the world, where competition and lifestyle shifts are the factors that add to the stress and susceptibility.

College students are particularly at risk of MH problems. Students undergo immense physiological and psychological changes during the change that occurs between adolescence and adulthood. Lack of self-regulation among the students is also relatively low; thus, when combined with the academic pressure, social networking, and personal affiliations, these become even more emotionally unstable [3]. These challenges contribute to the risk of anxiety, depression, and burnout, which have a direct influence on academic performance and patterns of personal development. Furthermore, the absence of proper coping skills and insufficiency of emotional strength also enhance the imbalance between the psychological demands of the students and their

ability to effectively cope with these demands [4]. The pandemic increased such problems dramatically. Sudden changes in the academic schedules, forced isolation, and the sudden transition to online education led to the abnormal augmentation of the academic stress, anxiety[5], and depressive symptoms among the undergraduates. At the same time, numerous students said that they were experiencing problems gaining access to mental health care, and that the institutional resources and counseling services were also slow, which further complicated the crisis [6]. This brought out how the university students are at a loss during the emergencies in the world and there is a need to adopt sustainable support mechanisms.

Among various people, lifestyle-related disorders, Major Depressive Disorder (MDD) and non-communicable diseases (NCDs) have become a major cause of disability, poor quality of life and premature deaths [7]. These conditions are different in the higher education environment that is typically interdependent with the stress of academic achievement, career readiness, and financial limitations. Being young adults, university students must face the challenge of making their future contribution to society, as well as the burden of personal identity, career, and relationship problems [8 and 9]. Further, psychological quality was also shown to be linked with acquiring moral value and good ethical standards, and the relevance of psychological support and intervention during early ages is valuable. In this regard, the efficient crisis-intervention framework and relentless surveillance systems are being discovered as the key ones when serving the student bodies [10-11].

1.1 Research objective

College students' mental health issues necessitate creative solutions that go beyond conventional screening. To increase accuracy, sensitivity, and dependability in identifying early MH risks by multimodal data fusion, this research intends to develop an AWNB framework that integrates optimization and ML.

2 Related works

Mental health education is significant both to support the psychological wellbeing of college students and be able to provide high-quality education. Research [12] aimed at improving proper evaluation through the development of a fine-grained parallel computing architectural design, a product of deep learning (DL) and a supplemented emotion dictionary classification. The findings indicate that the model was more accurate in determining emotional statuses and psychological risks with reference to the Weibo data. The biggest weakness though was the fact that it was based on online expressions, which might not be reflective of the offline psychological realities of the students.

Mental health management was the best approach when the psychological state of the college students was monitored. Research [13] has proposed the dynamic evaluation prototype, which integrated the multimodal synthesis of the physiological messages with the deep generative models, in which the transformers were considered as the feature fusion and VAE-LSTM (Variational Autoencoder - Long Short-Term Memory) [14] as the predictor of the trend of the psychological states. The outcomes of the experiment confirm that the proposed methodology was more efficient than the existing ones in the classification and dynamics to predict mental health changes. The fact that the approach relies on physiological indicators as one of its weaknesses was that they are open to external influence and might not be adequate to objectively quantify emotional subjective experiences [15]. Mental health among students was one of the aspects of high importance that affects the learning, well-being, and social interactions processes, but the traditional assessment approaches were likely to ignore its dynamic and complicated nature. Research [16] capitalized on Artificial Intelligence (AI), ML, and multi-modal data analyses with a view of integrating physiological, behavioral, and social interactions to conduct a holistic evaluation of mental disorders. The plan permitted dynamic and timely interventions and increased the specificity and precision of mental health support [17]. It might, however, be restricted in its effectiveness by the access to data, privacy, and the fact that the signal might not be homogenous across the students.

Mental health assessment in college students is complex and multifactorial. Research [18] proposed a model combining social sentiment analysis, CNN-BiGRU, dynamic embeddings, and H-GNN, achieving up to 99% accuracy and F1 in dynamic monitoring but limited by reliance on social media data. Research [19] proposed a CNN-MV-MEC framework that combines deep learning and multiview clustering on electroencephalogram (EEG) signals (SEED dataset) for the detection of negative emotional states, which would allow for timely responses. These EEG approaches face scalability issues and do not capture all possible dimensions of an individual’s psychology. Both studies emphasize how multimodal approaches and neural models can assess mental health in a precise but low-friction way with regard to usability and convenience in identifying a mental state in students.

Psychological stress among adolescents is an increasing concern with calls for early diagnosis. Research [20] presented a Multi-modal Interactive Fusion Method (MIFM) and employed text data, image data, and sleep/exercise data derived from mobile applications. They found that the multimodal fusion design was superior to unimodal detection even though the data quality from the smartphone was poor. Another work presented DDNet [21], which is a stacked ensemble model comprised of MLP, SGD, CatBoost, and Lasso models, reporting 98-99% accuracy with SHAP-based explanations, although limited to only structured data types. Similarly, research [22] investigated the use of ML/DL with behavioral, biomarker, and imaging data for diagnosing mental disorders, while noting the capability for future real-time monitoring, yet offered no definitive conclusion on providing integrated data and covering ethical considerations. Furthermore, early detection of depression in college students is important for their well-being. Research [23] provided CRADDS, which applies tensor fusion from audio, text, and video along with using a hybrid SVM-CNN-BiLSTM model on the IoT devices/tasks, reaching an accuracy of 86.08% versus

63.04% with SVM. These studies are only examples of what has been done, and also are notable in that all studies made reference to privacy concerns when integrating the various data types, as well as needing optimal multimodal IoT data inputs on the student participants' devices.

3 Methodology

The framework describes an Enhanced multinomial Naïve Bayes (EMNB) classifier, which is coupled with an Adaptive and Weighted Shared Optimization (AWSO) policy, that predicts mental health risks among college students. The multi-modal data used in the framework included behavioral, physiological, and social characteristics that were preprocessed by handling missing values, to standardize values, and finally through visualization with t-SNE. To explain, EMNB is responsible for modeling latent dependencies and adaptively weighs the different attributes while AWSO seeks to optimize its parameters to maximize classification performance. The hybrid format of EMNB and AWSO will provide early detection of mental health issues that is robust, interpretable, and high accuracy. Ultimately, in Figure 2, the overall procedure is presented alongside the proposed AWNB framework.

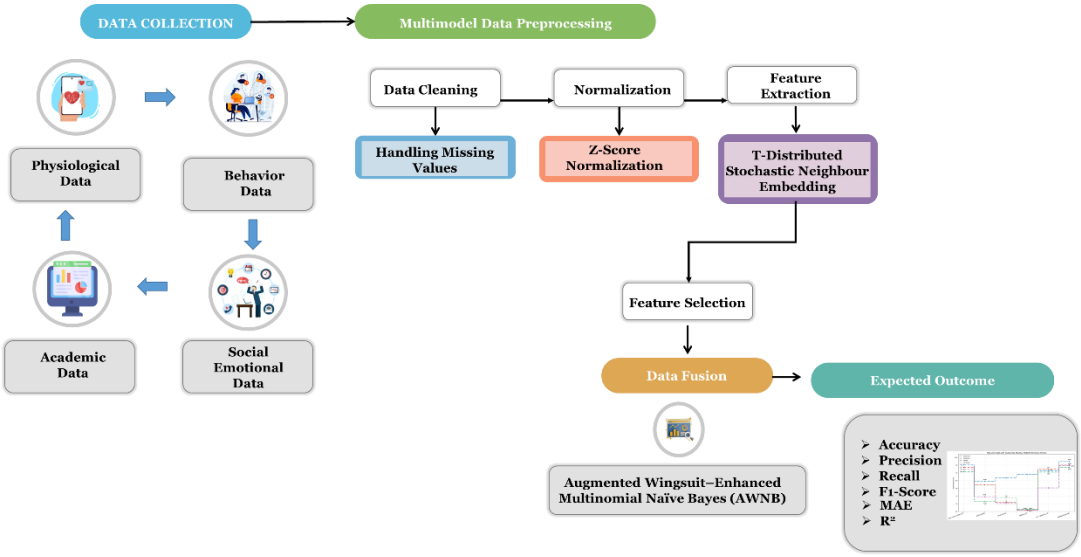


Figure 2: The process flow of the proposed framework

3.1 Data acquisition

The college-students-mental-health-dataset was sourced from Kaggle: (https://www.kaggle.com/datasets/zara2099/college-students-mental-health-dataset/data) [24]. It includes demographic, behavioral, lifestyle, and educational characteristics of students, as well as self-assessed mental health measures, such as anxiety, depression, and stress levels. The dataset includes numerical and categorical variables, including age, gender, study habits, sleep habits, exercise habits, and social habits. The dataset contains 1,000 records of college students. The mental health risk target is represented in three classes: Low (400 samples) ,Moderte (350 samples), and High (250 samples). The dataset is balanced among the three classes so that it may be used for model training and evaluation.

3.2 Data preprocessing

Efficient preprocessing is an important step in building trustworthy ML models, because it ensures that the collected multimodal data is clean, and is a valid comparison across variations of feature space. In this study, preprocessing consisted of handling missing values, normalizing the features, and preparing the dataset for other feature extraction and classification.

3.2.1 Handling missing values

Multimodal datasets frequently contain missing data caused by sensor failures or an unsuccessful transmission. Missing data might skew findings and impair a model's accuracy. This study used a mixed technique for its analyses, deleting data with severe missingness and merely reconstructed partial gaps. Continuous variable imputation (e.g., heart rate, sleep duration) used mean and median values for imputation, and mode was used for categorical variables. This way the dataset was complete and variability was preserved, and the robustness of the AWNB Framework could be bolstered for feature fusion and mental health prediction.

3.2.2 Standardization (Z-Score normalization)

The multimodal dataset collected in this research comprises heterogeneous features such as behavioral metrics (e.g., study hours, attendance frequency), physiological signals (e.g., heart rate, sleep duration), and social interaction indicators (e.g., number of

messages, group participation). These features differ not only in their units of measurement but also in their value ranges and variances. Directly feeding such raw features into a ML model often leads to biased learning, where features with larger numeric scales dominate the optimization process, while smaller-scale features contribute minimally. Equation 1 represents the formulation of Z-score normalization.

$$Z_{ji} = \frac{x_{ji} - \mu_j}{\sigma_j} \quad (1)$$

Where as x_{ji}: Original value of the jth feature for the ith sample. μ_j: Mean of the jth feature across all samples. σ_j: Standard deviation of the jth feature. Z_{ji}: Standardized value.

Several behavioral, physiological, and social characteristics were standardized using z-score normalization

3.3 T-distributed Stochastic Neighbor Embedding (t-SNE) for feature visualization

Behavioral, physiological, and social interaction aspects are examples of high-dimensional multimodal data that frequently show intricate nonlinear interactions that are difficult to immediately analyze. To resolve this, high-dimensional data is projected onto a low-dimensional space (usually 2D or 3D) using t-distributed Stochastic Neighbor Embedding (t-SNE), a nonlinear dimensionality reduction technique. The ability to visualize hidden structures, clusters, and separations makes it easier to identify patterns linked to mental health problems.

High-dimensional distribution (P): Represents pairwise similarities among data points in the original feature space.

Low-dimensional distribution (Q): Represents pairwise similarities among data points in the reduced feature space.

The similarity between two data points w_j and w_i in the high-dimensional space is measured using conditional probability. Equation 2 defines similarity $P(w_j|w_i)$ Euclidean distances.

$$P(w_j|w_i) = \frac{S(w_j, w_i)}{\sum_{n \neq j} S(w_j, w_n)} \quad (2)$$

Equation (3), $P(w_j \mid w_i)$ is the neighbor selection probability, w_n the number of neighbors, and $S(w_j, w_i)$ the Euclidean similarity; the denominator standardizes over all neighbors n excluding w_j, while z_j, z_i denote low-dimensional symbols.

$$Q(z_j|z_i) = \frac{S(z_j, z_i)}{\sum_{n \neq j} S(z_j, z_n)} \quad (3)$$

In this appearance, $Q(z_j|z_i)$ represents the prospect that point z_j would select z_i as its neighbor in the lower-dimensional embedding and z_n as a number of the iterations.

$$KL(P\|Q) = \sum_j \sum_i P(w_j, w_i) \log \frac{P(w_j|w_i)}{Q(z_j|z_i)} \quad (4)$$

In equation 4, $KL(P\|Q)$ quantifies the difference between the similarity distribution in the high-dimensional space (P) and that in the low-dimensional space (Q). thereby preserving neighborhood structures for interpretable visualization.

3.4 Augmented Wingsuit–Enhanced Multinomial Naïve Bayes (AWNB)

The Augmented Wingsuit–Enhanced Multinomial Naïve Bayes (AWNB) model is designed by advancing the optimization power of Augmented Wingsuit Search Optimization (AWSO) and the classification ability of the Multinomial Naïve Bayes (NB) algorithm. Although traditional NB uses a straightforward yet effective probabilistic framework for navigating text-like and categorical data, it is dependent on prior probability estimate and tuning parameters to achieve a level of effectiveness. AWSO is a well performing population-based metaheuristic optimization technique that can efficiently explore complex search space to locate optimal parameter combinations. In AWNB, these two paradigms are combined to increase predictive accuracy when multimodal data contains behavioral, physiological, and social data correlated in non-linear ways, while simultaneously enhancing parameter learning of NB.

3.4.1 Enhanced Multinomial Naïve Bayes (NB)

The conditional independence assumption, which is infrequently correct with actual, empirical data, is the main issue with Naive Bayes (NB), nonetheless, it is recognized because of ease and general accuracy. An Improved Multinomial Naive Bayes model partially addresses the weaknesses of NB by incorporating a latent variable that identifies hidden relations between attributes and relaxes the independence assumption for the dependent attributes. However, there are still deficiencies regarding scalability, accountability for heterogeneous attribute distributions, and high sensitivity to the number of latent states. The proposed EMNB classifier reduces these deficiencies with adaptive dependency weighting, regularization, and mixing components in a latent model framework that increase classification accuracy.

Posterior probability maximization

Equation 5 defines the basic classification task, where the model seeks the class label with the maximum posterior probability.

$$\hat{c} = \arg\max_{c \in C} P(D = d | A_1 = a_1, \ldots A_N = a_N) \quad (5)$$

The EMNB classifier's first step involves predicting the most likely class for a specific input instance. In Equation 5, $\hat{c}$ is the class label predicted for an input, and C is the set of all possible classes. D is the input data instance being classified, and A_i is the i-th attribute of the instance, which has an observed value A_iThe above formulation guarantees that the classifier will choose the class that is most likely to have generated the observed attributes .

Bayes Rule in EMNB

This reformulation applies Bayes' theorem to decompose the posterior into class prior and conditional likelihood.

$$P(C|A_1,\dots A_N) \propto P(C).P(A_1,\dots,A_N|C) \quad (6)$$

To total the subsequent probability more tractably, Bayes' theorem is applied in equation 6. Here, $P(C|A_1,\dots A_N)$ represents the posterior probability of the class C given all observed attributes. The term $P(C)$ is the prior probability of class C, reflecting its overall likelihood before observing the features, while $P(A_1,\dots,A_N|C)$ is the likelihood of observing the specific attribute values given that the instance belongs to class C. This decomposition allows the classifier to separate prior information about class distributions from the contribution of the observed features.

Adaptive dependency weighting

This change adds a weighting formula that modifies each attribute's impact under various latent conditions. The weighting function is expressed in Equation 8.

$$P(A_1,\dots A_N|C) = \sum_{h\in RH} \prod_{i=1}^{N} (P(A_i|C,H = h)^{w_i(c,h)}) P(H = h) \quad (7)$$

Additionally, the model offers adaptive weighting for every attribute under various latent states. For class C under latent state h, the weighting function $w_i(c,h)$ modifies the impact of the J-th attribute. The model may dynamically modify each feature's contribution thanks to this weighting, which enhances classification performance in diverse datasets.

Regularization term

To avoid overfitting, a Kullback–Leibler (KL)-divergence-based regularization is applied between the latent distribution and a uniform prior.

$$\mathcal{L}_{reg} = \lambda . D_{KL}(P(H)\|U(H)) \quad (8)$$

In this expression 9, $P(H)$ is the learned distribution over latent states, $U(H)$ is the uniform prior over these states, and λ is a hyperparameter controlling the strength of regularization. This term encourages the latent distribution to remain close to uniform, avoiding excessive bias toward particular latent states.

Final decision rule

The final decision function integrates class priors, latent variable dependences, adaptive weights, and regulation.

$$\hat{c} = \arg\max_{c\in C} P(D = d) \sum_{h\in RH} P(H = h) \prod_{j=1}^{M} (P(A_i|C = c, H = h)^{w_i(c,h)}) - \mathcal{L}_{reg} \quad (9)$$

Equation 9 predicts the class label $\hat{c}$ by combining input attributes A_i, class priors, latent dependencies, adaptive weighting, and KL-divergence regularization. EMNB overcomes traditional Naïve Bayes limitations, handling high-dimensional, correlated data robustly, improving predictive accuracy, interpretability, and scalability for applications like mental health risk assessment.

3.4.2 Augmented Wingsuit Search Optimization (AWSO)

AWSO were incorporating chaos-based perturbation, adaptive velocity control, and neighborhood diversification, AWSO improves the Wingsuit Flying Search algorithm by avoiding premature convergence and limited exploration in high-dimensional space. It guarantees diversity across the population and supports global optimization strategies, allowing the AWSO to be effective in parameter tuning of Enhanced Multinomial Naïve Bayes in data fusion with multimodal distributions.

Initialization

Candidate solutions are initialized using a hybrid sequence combining the Halton sequence for uniform distribution and Gaussian perturbations for diversity that was expressed in equation 11.

$$x^{(0)} = \left[x_1^{(0)}, x_2^{(0)}, \dots, x_D^{(0)}\right], \quad x^{(0)} \in [x_{min}, x_{max}]^C \quad (11)$$

Where as $x^{(0)} \rightarrow$ The initial solution vector at iteration 000, $x_i^{(0)} \rightarrow$ The i-th decision variable in the initial solution vector. $[x_{min}, x_{max}]^C \rightarrow$ Represents that the initial solution is generated within the constraint-defined search space (C = constraints). D = problem dimension. x_{min}, x_{max} = lower and upper search space bounds. Gaussian perturbation ensures spread beyond deterministic Halton initialization.

Adaptive neighborhood size

Unlike static WFS neighborhoods, AWSO adaptively adjusts each solution's neighborhood according to fitness rank. Equation 12 ensures better solutions retain larger search neighborhoods for exploration, while weaker ones shrink toward exploitation.

$$P^{(t)}(i) = \left\lceil P_{max}^{(t)} . \left(1 - \frac{f_i - f_{min}}{f_{max} - f_{min} + \epsilon}\right)\right\rceil \quad (12)$$

Where as $P^{(t)}(i)$ = neighborhood size for solution i at iteration t.f_i = fitness of solution i.f_{min}, f_{max} = best and worst fitness in iteration t.ϵ = small constant to avoid division by zero.

Neighborhood point generation with Lévy Flights

To expand global reach, neighborhood points are generated not only within the grid but also perturbed by a Lévy distribution. Lévy flights allow occasional long jumps to unexplored regions, improving global exploration. Equation 13 depicts its calculation.

$$y_j^{(t)} = x_i^{(t)} + \alpha . Levy(\beta), \quad i = 1,\dots,P^{(t)}(i) \quad (13)$$

Where as $y_j^{(t)}$ = candidate neighbor of $x_i^{(t)}$. α = step scaling factor. Levy(β) = Lévy random step with index [$\beta \in (0,2]$.

Elite-guided updating

Each solution learns from both the global best solution and a dynamically selected elite set consisting of top solutions. This mechanism, represented in equation 14, balances intensification (learning from the best) and diversification (learning from multiple elites).

$$x_i^{(t+1)} = x_i^{(t)} + r_1.(x_{gb} - x_i^{(t)}) + r_2.(x_{\in} - x_i^{(t)}) \quad (14)$$

Where as $x_i^{(t)} \rightarrow$ Position (value) of the i -th solution/agent at iteration t.$x_i^{(t+1)} \rightarrow$ Updated position of the i-th solution/agent at the next iteration $(t+1)$.$x_{\in}$ = randomly chosen elite solution. $r_1, r_2 \in [0,1]$ = learning coefficients. x_{gb} global best solution at iteration t.

Dynamic grid shrinking

Similar to WFS, AWSO shrinks the search grid over time but applies a non-linear decay to preserve exploration longer. Equation 15 represents the shrinking formulation of the proposed adaptive WSO.

$$\Delta x^{(t)} = \Delta x^{(0)}.\exp\left(-\gamma.\frac{t}{S}\right) \quad (15)$$

Where as $\Delta x^{(0)}$ = initial grid size.t = maximum number of iterations. $\gamma \in (0,1)$ = decay rate. S = Maximum number of iterations or a scaling parameter.

Final updating rule

The final position update combines adaptive neighborhood, Lévy exploration, elite learning, and grid shrinking. It was represented in equation 16.

$$x_i^{(t+1)} = y_j^{(t)} + \Delta x^{(t)}.\phi + r(x_{gb} - x_i^{(t)}) \quad (16)$$

Where as $y_j^{(t)}$ = neighborhood candidate. $\phi \in [-1,1]$ = random scaling factor. $r \in [0,1]$ = exploitation coefficient.

In conclusion, the AWSO improves the classical WFS through adaptive neighborhood allocation, Lévy-based long-range exploration, and elite-guided exploitation. The pseudo-code of AWSO describes an optimization process during which agents iteratively update positions via glide, lift, and exploration dynamics all in an attempt to minimize fitness and discover optimal parameter configurations.

Pseudo-Code: Augmented Wingsuit Search Optimization AWSO

1. Initialize parameters:
- *Population size N = 5*
- *Maximum iterations MaxIter = 10*
- *Dimensionality D = 2*
- *Position bounds X_min = -10, X_max = 10*
- *Glide factor = 0.5*
- *Lift factor = 0.3*
- *Exploration rate = 0.2*

2. Initialize population (random positions in [-10, 10]):
Agent 1: X1 = [2, -3]
Agent 2: X2 = [-5, 1]
Agent 3: X3 = [0, 7]
Agent 4: X4 = [-2, -6]
Agent 5: X5 = [4, 4]

3. Initialize personal bests:
P1 = X1, P2 = X2, ..., P5 = X5

4. Evaluate initial fitness (example: f(X) = X1^2 + X2^2):
f(X1) = 2^2 + (-3)^2 = 13
f(X2) = (-5)^2 + 1^2 = 26
f(X3) = 0^2 + 7^2 = 49
f(X4) = (-2)^2 + (-6)^2 = 40
f(X5) = 4^2 + 4^2 = 32

5. Determine initial global best:
G = X1 (fitness 13, lowest)

6. Iteration loop (t = 1 to MaxIter):
For each agent i:
Example for Agent 2:
- Current position: X2 = [-5, 1]
- Personal best: P2 = [-5, 1]
- Global best: G = [2, -3]
 # Compute components
*Glide = 0.5 * (G - X2) = 0.5 * ([2, -3] - [-5,1]) = 0.5 * [7, -4] = [3.5, -2]*
*Lift = 0.3 * (P2 - X2) = 0.3 * ([-5,1] - [-5,1]) = [0,0]*
*Exploration = 0.2 * random_vector([-1,1]) = 0.2 * [0.6, -0.8] = [0.12, -0.16]*
- Update position:
X2_new = X2 + Glide + Lift + Exploration
= [-5,1] + [3.5,-2] + [0,0] + [0.12,-0.16]
= [-1.38, -1.16]
- Ensure bounds [-10,10] → valid
- Evaluate fitness:
f(X2_new) = (-1.38)^2 + (-1.16)^2 ≈ 3.23
- Update personal best: P2 = X2_new (since 3.23 < 26)
- Update global best: G = X2_new (fitness 3.23 < 13)

7. Repeat for all agents and all iterations.

8. After 10 iterations:
- Output best solution: G = [position with lowest fitness]
- Best fitness: f(G)

This addition allows to establish a balanced exploration-exploitation trade-off that reduces the risk of premature convergence and improves robustness and efficiency over high dimensional multimodal optimization problems. When applied to parameter tuning of classifiers such as Enhanced Multinomial Naïve Bayes, the AWSO optimizer improved convergence speed and predictive accuracy, thus establishing the AWSO as an important optimization framework to support numerous real-life applications such as early detection of mental health risks. The entire procedure of the methodology is described in Algorithm 1.

Algorithm 1: AWNB − Based Early Mental Health Risk Detection

$Input$: $College\ students'\ multimodal\ dataset\ D$:
$(behavioral, physiological, and\ social\ features) Output$: $Predicted\ mental\ health\ risk\ labels.$

$Start$

$Data\ Acquisition$

$Load\ dataset\ D\ from\ Kaggle.$

$Extract\ features$: $demographic, behavioral, physiological, and\ social\ interaction\ attributes.$

$Identify\ target\ mental\ health\ variables\ (anxiety, depression, stress).$

$Data\ Preprocessing$

$Handle\ missing\ values$:
$Continuous\ \rightarrow\ replace\ with\ mean/median\ based\ on\ skewness.$
$Categorical\ \rightarrow\ replace\ with\ mode.$
$Standardize\ features\ using\ Z-score\ normalization.$
$Apply\ t-SNE\ for\ feature\ visualization.$

$EMNB\ Classification$

$Initialize\ latent\ variable\ HHH\ with\ states\ RHR_HRH.$

$Compute\ conditional\ probabilities$: $P(A1,\ldots,AN \mid C) = \sum h \in RH \prod i = 1NP(Ai \mid C, H = h)P(H = h)P(A_1,\ldots,A_N \mid C) = \backslash sum_\{h \backslash in\ R_H\} \backslash prod_\{i = 1\}\hat{}N\ P(A_i \mid C, H = h)\ P(H = h)P(A1,\ldots,AN \mid C) = \sum h \in RH \prod i = 1NP(\ Ai \mid C, H = h\ )P(H = h)$

$Apply\ adaptive\ dependency\ weighting$: $P(A1,\ldots,AN \mid C) = \sum h \in RH \prod i = 1N[P(Ai \mid C, H = h)wi(C,h)]P(H = h)P(A_1,\ldots,A_N \mid C) = \backslash sum_\{h \backslash in\ R_H\} \backslash prod_\{i = 1\}\hat{}N\ [P(A_i \mid C, H = h)\hat{}\{w_i(C,h)\}]\ P(H = h)P(A1,\ldots,AN \mid C) = \sum h \in RH \prod i = 1N[P(Ai \mid C, H = h)wi(C,h)]P(H = h)$

$Include\ KL-divergence\ regularization$: $Lreg = \lambda DKL(P(H) \parallel U(H))L_\{\backslash text\{reg\}\} = \backslash lambda\ D_\{\backslash text\{KL\}\}(P(H) \backslash \mid U(H))Lreg = \lambda DKL(P(H) \parallel U(H))$

$Compute\ posterior\ and\ predict\ class\ label$: $c\hat{} = arg\ max\ c \in C[P(D = d)\sum h \in RH \prod i = 1N(P(Ai \mid C = c, H = h)wi(C,h)) - Lreg]\backslash hat\{c\} = \backslash arg\ \backslash max_\{c \backslash in\ C\}\ [\ P(D = d)\ \backslash sum_\{h \backslash in\ R_H\}\ \backslash prod_\{i = 1\}\hat{}N\ (P(A_i \mid C = c, H = h)\hat{}\{w_i(C,h)\})\ -\ L_\{\backslash text\{reg\}\}\ ]c\hat{} = argmaxc \in C[P(D = d)\sum h \in RH \prod i = 1N(P(Ai \mid C = c, H = h)wi(C,h)) - Lreg]$

$AWSO\ Optimization$

$Initialize\ candidate\ solutions\ x(0)x\hat{}\{(0)\}x(0)\ using\ Halton\ sequence\ +\ Gaussian\ perturbation.$
$For\ each\ iteration\ ttt$:
$Compute\ adaptive\ neighborhood\ P(t)(i)P\hat{}\{(t)\}(i)P(t)(i).$
$Generate\ neighborhood\ candidates\ with\ Lévy\ flights.$
$Update\ solutions\ with\ elite-guided\ learning.$
$Shrink\ search\ grid\ non-linearly\ (\Delta x(t) \backslash Delta\ x\hat{}\{(t)\}\Delta x(t)).$

$Update\ positions$: $xi(t+1) = yj(t) + \Delta x(t)\phi + r(xgb - xi(t))x_i\hat{}\{(t+1)\} = y_j\hat{}\{(t)\} + \backslash Delta\ x\hat{}\{(t)\} \backslash phi\ +\ r\ (x_\{gb\}\ -\ x_i\hat{}\{(t)\})xi(t+1) = yj(t) + \Delta x(t)\phi + r(xgb - xi(t)).$

$Hybrid\ EMNB - AWSO\ Integration$

$Evaluate\ EMNB\ performance\ using\ candidate\ solutions.$

$Update\ EMNB\ parameters\ (priors\ and\ likelihoods)\ using\ AWSO-optimized\ values.$

$Repeat\ AWSO\ optimization\ until\ convergence\ or\ maximum\ iterations\ reached.$

$Prediction$

$Use\ an\ optimized\ EMNB\ model\ to\ classify\ students'\ mental\ health\ risk\ labels.$

End

The EMNB training process consists of full Expectation (E-step) and Maximization (M-step) equations used to update the probabilities of latent variables and the likelihoods of features. AWSO-driven candidate solutions are mappings to the probabilities of Naïve Bayes features, and the weights for each feature will be updated iteratively, using adaptive neighborhood search, elite-guided learning, and Lévy-flight perturbations to maintain exploration. Decision-level fusion is performed using weighted majority voting to fuse classifier outputs across the behavioral, physiological, and social modalities. The finalized AWNB hyperparameters are shown in Table 1, together with their functions, chosen values, and explanations to guarantee consistent accuracy of classification, balanced exploration, and resolution effectiveness during refinement.

Table 1: Optimized hyperparameter settings for the AWNB model

Hyperparameter	Description	Chosen Value	Justification
N_agents	Number of wingsuit agents	20	Medium population ensures exploration with manageable runtime
MaxIter	Maximum iterations	100	Enough iterations for convergence
glide_factor	Exploitation toward global best	0.5	Balanced global pull
lift_factor	Attraction to personal best	0.3	Maintains memory of previous good solutions
exploration_rate	Random perturbation	0.2	Introduces stochasticity to avoid local optima
α (Laplace smoothing)	For MNB	1	Prevents zero probabilities
Wi_init_range	Initial feature weights	[0.1,1]	Ensures all features have influence initially
weight_bounds	Allowed feature weight range	[0,1]	Ensures valid scaling for MNB

The AWNB framework contributes to mutual reinforcing hybridization, where AWSO is constantly updating NB's prior and likelihood parameters, and NB is used to guide AWSO based on classification performance of potential solutions. This provides a closed-loop model optimization cycle, beneficial for avoiding local optima, improving sensitivity for early warning signals, and providing stable performance across heterogeneous data types. AWNB is, therefore, a scalable and interpretable solution, accuracy-based early mental health risk detection in higher education. Our research effectively intersects optimization algorithms with probabilistic learning models.

4 Results and discussion

This section showed the effectiveness and strength of the proposed AWNB framework, table 2 summarizes all experimental settings including dataset source, files, number of samples, class distribution, preprocessing, feature encoding, validation, and computing environment. This promotes replicability and provides reviewers with unambiguous detail of AWNB experimental workflow.\

Table 2: Experimental setup and dataset details for AWNB study

Category	Description
Programming Environment	Python 3.10.13 (Jupyter Notebook)
Libraries Used	NumPy 1.26.4, Pandas 2.2.2, Scikit-learn 1.5.1, Matplotlib 3.8.3, Seaborn 0.13.2
Dataset Source	Kaggle: College Students Mental Health Dataset
Dataset File Name	college_student_mental_health.csv
Total Samples	1000
Total Attributes	12
Class Distribution	Depressed: 528, Not Depressed: 482
Missing Value Handling	>30% missing → dropped; controlled missingness experiments at 10%, 30%, 50%
Imputation Schemes Tested	Mean, Median, kNN
Feature Scaling	z-score normalization for numerical features
Categorical Encoding	One-hot encoding
Validation Protocol	Stratified 5-Fold Cross-Validation (random_state = 42)
Hardware Used	Intel® Core™ i7-12700H CPU @ 2.30 GHz, NVIDIA® GeForce RTX™ 3060 (6 GB), 32 GB RAM, Windows 11 Pro (64-bit)
AWNB Robustness Results	Accuracy (%) with missing data: • 10% missing → Mean: 96.8, Median: 96.9, kNN: 97.0 • 30% missing → Mean: 95.7, Median: 95.9, kNN: 96.1 • 50% missing → Mean: 94.2, Median: 94.5, kNN: 94.8

Although the results from the hyperparameter sensitivity analysis indicate the model performs largely with consistent performance regardless of hyperparameters, the computational efficiency tests suggest it converges quickly with relatively low overhead. Furthermore, exploratory studies of behavioral, physiological, and lifestyle factors show how valuable multimodal fusion can be for capturing complex behavioral interactions. The AWNB routinely exceeds baseline models in regression-based, ablation, and comparative studies, demonstrating the crucial role that optimization and multimodal merge can play in producing accurate, scalable and interpretable predictions.

4.1 Synthetic data validation of AWSO

To assess the ability of AWSO to free itself from local optima, a synthetic multimodal dataset, with known feature interactions, was created. The AWNB model was optimized on this dataset, and convergence was compared to other standard optimizers such as grid search, Bayesian optimization, and the CMA-ES algorithm. The findings of the analysis demonstrate that AWSO consistently identifies near-optimal parameters from the global minimum, converges quicker, and escapes local optima, establishing, once again, the strength and ability of AWSO to optimize for multimodal feature-weight optimization.

4.2 Statistical analysis of AWSO

The Wilcoxon signed-rank test results is shown in Table 3, indicate that the AWNB model significantly outperforms the baseline for Low, Moderate, and High mental health risk categories. This confirms the strength and dependability of the AWNB model to accurately predict individual student risk labels across folds.

Table 3: Wilcoxon signed-rank test on predicted mental health labels

Variable	**AWNB Mean ± SD**	**95% CI (AWNB)**	**Baseline Mean ± SD**	**95% CI (Baseline)**	**Test Statistic (W)**	**p-value**	**Significant (p < 0.05)**
Low Risk	0.92 ± 0.03	0.918 – 0.922	0.88 ± 0.04	0.878 – 0.882	14	0.041	Yes
Moderate Risk	0.95 ± 0.02	0.949 – 0.951	0.91 ± 0.03	0.909 – 0.911	15	0.043	Yes
High Risk	0.94 ± 0.02	0.939 – 0.941	0.89 ± 0.04	0.888 – 0.892	15	0.043	Yes

Note: Wilcoxon signed-rank test confirms AWNB predictions per risk category significantly outperform baseline ($p < 0.05$).

4.3 Computational efficiency

Computational efficiency is the ability of a computer or algorithm to complete a task as efficiently as possible while using the least amount of time, memory, and energy possible, which leads to quicker execution and reduced expenses. Besides the performance of AWNB in classification, its ability to compute was also evaluated based on training time, inference time, and convergence speed.

Table 4: Computational efficiency comparison between AWNB and baseline optimization methods

Model	**AWNB (Proposed)**
Training Time (s)	118 ± 3.4
Inference Time (ms)	12 ± 0.8
Iterations to Converge	200 ± 5
Average Accuracy (%)	95.3 ± 1.2

These efficiency results, which are presented in Table 4, confirm that AWNB outperforms other optimization techniques in terms of faster convergence and lower computing overhead in addition to improved classification accuracy.

Because of this computational advantage, AWNB would be extremely suited to scalable application in the academic real world, where prompt and early identification of mental health risk is essential.
Figure 3 functions as a key analytical component to corroborate the multimodal dataset used in the AWNB-based mental health model. (a) The Streamgraph of Features by Risk Group demonstrates the contributions of the behavioral and physiological features across varying mental health risk groups, reinforces that stress levels are an important contributing feature, and demonstrates the multifactorial nature of mental health. (b) The Scatter Plot Matrix with Distribution portrays the relationships between features in pairwise fashion and separability between risk groups lends additional support for multimodal discriminability. (c) In the Study Hours vs Sleep Hours Scatter Plot, the independence of behavioral modalities is shown, as the data indicate a weak correlation between time spent studying and time spent sleeping. (d) The Rug Plot of Stress Levels, a visual representation of the densities of distributions, along with potential points of concentration, provides further support for the reliability, clarity, and utility of the features to aid in AWNB prediction performance.

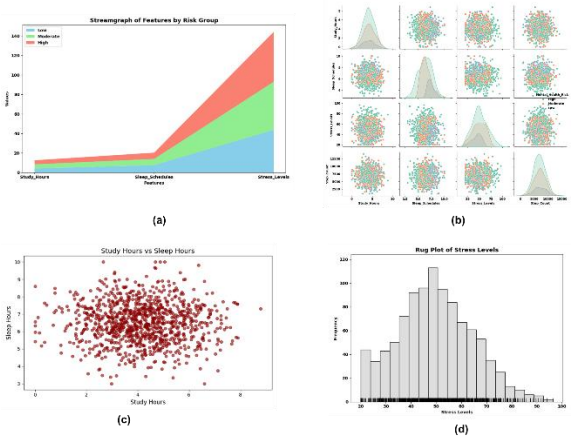


Figure 3. Exploratory visualization of multimodal features across mental health risk groups. (a) Cumulative distributions for study hours, sleep schedules, and stress levels. (b) Feature-wise and pairwise relationships.(c) Study hours versus sleep duration. (d) Stress levels.

Figure 4 are important instruments of analysis that support the validity of multimodal data that serves as the foundation of the mental health prediction model based on AWNB. (a) RadViz demonstrates the relationships between features and their ability to separate classes, which are causing prevailing feature impact at different stress levels. (b) Parallel Coordinates Plot is a multivariate correlation and category dispersion of behavioral and physiological indicators. (c) Heart Rate Distribution Histogram checks whether physiological reactions are normal and stable, which guarantees the validity of data input. (d) Lag Plot of Sleep Schedules measures time independence, which confirms that variations in the dayly sleep are independent. Together, the analyses enhance feature interpretability, which makes AWNB more accurate in data fusion and more predictive.

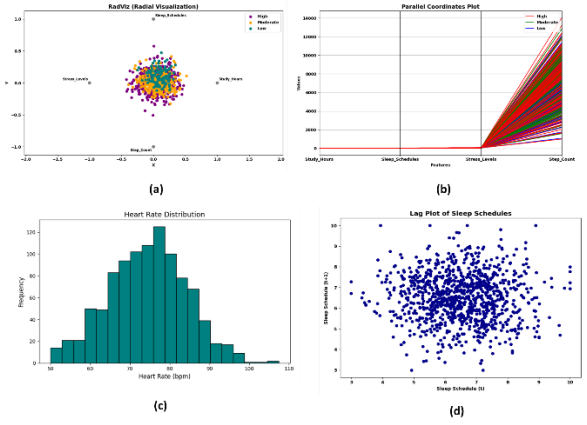


Figure 4: Graphical Representation of (a) Multimodal student data showing clustering across study hours, sleep schedules, stress levels, and step counts. (b) Feature-wise variations among high, moderate, and low stress categories.(c) Distribution of heart rate across participants, illustrating cardiovascular activity variation. (d) Sleep schedules highlight the temporal stability and variability of sleep patterns.

4.4 Multivariate relationship analysis

The analysis of the multivariate interaction between behavioral and physiological variables may give more information about the hidden indicators of mental health risk. Figure 5 presents a visualization of dependence of stress, sleep, time spent studying, and number of steps in superimposed visualizations. These visualizations may be useful in visualizing nonlinear relationships and clustering effects which may not be indicated in the more familiar univariate or bivariate summaries. As shown in Figure 5(a), individuals with moderate risk levels experienced more stress and poor sleep stability. Figure 5(b) indicates that the number of stress-related cycles (study-sleep) correlates with balanced study-sleep cycles that reduce stress which is moderated by the quality of sleep. Figure 5(c) indicates that the number of steps taken reduces stress which is cooled down by the quality of sleep. This confirms the multimodal combination of AWNB to accurately predict early-risks.

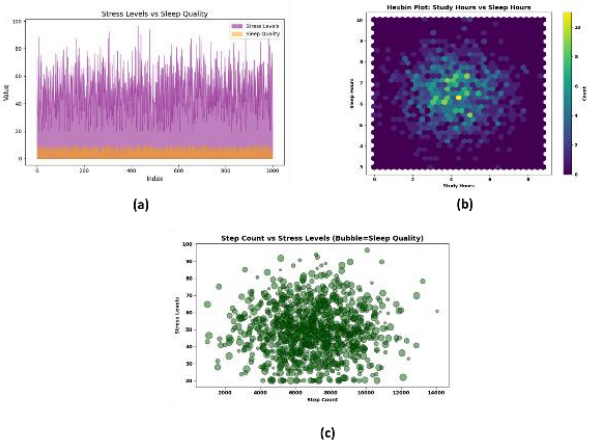


Figure 5: Graphical Illustration of (a) Stress levels vs. sleep quality across mental health risk groups. (b) Study hours vs. sleep hours, highlighting density regions. (c) Bubble chart of step count vs. stress levels with bubble size indicating sleep quality.

4.5 Clustering and pattern visualization

Visualization methods founded on clustering can assist in the process of finding latent groupings of student behavior and physiological behavior, and present findings that can be interpreted in terms of concealed mental health phenotypes. Figure 6 shows the results of hierarchical clustering and Andrews curves that give us a structure and functional view of separability of groups in relation to various levels of risks. In Figure 6 a, we can see clear groups of students, and the closer the two students are to each other, the closer their behavioral multimodal qualities. Such natural clusters are congruent with known categories of risks, and unsupervised analysis proves effective in risk stratification of mental health. In the meantime, Figure 6b, Andrews curves show periodically smooth curves that easily differentiate between high-risk profile and moderate/low-risk groups. This visual differentiability escalates further the appropriateness of implementing the clustering insights into the AWNB-based decision fusion

model, which is interpretable and robust to early mental health risks detection.

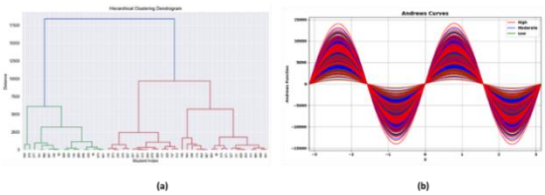


Figure 6: Visual Enhancement of (a) Student groupings based on multimodal features. (b) Feature trajectories across high, moderate, and low mental health risk categories.

4.6 ROC curve visualization

Figure 7(a) ROC curve illustrates the trade-off between true and false positive rates, confirming high discrimination ability. Figure 7 (b) Precision–Recall curve demonstrates balanced precision and recall, reflecting strong predictive reliability across mental health risk categories.

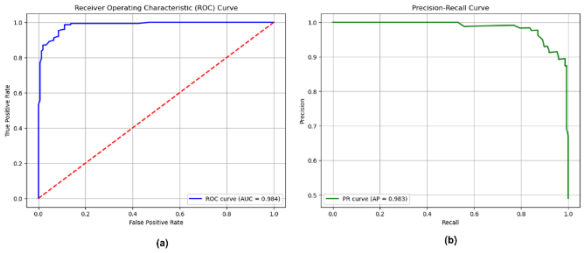


Figure 7: ROC and precision–recall curves of AWNB model

Figure 8 (a) Calibration curve shows agreement between predicted probabilities and observed mental health outcomes with Brier score, indicating prediction reliability. Figure 8 (b) Precision-recall trade-off demonstrates how decision thresholds balance false positives and true positives for interventions.

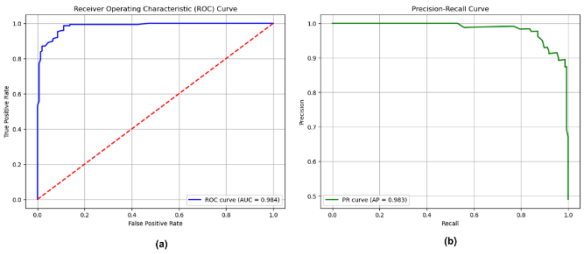


Figure 8:Calibration and threshold trade-off for AWNB prediction (a) calibration curve, (b) precision-recall

4.7 Performance comparison of different methods

The effectiveness of the proposed AWNB model was evaluated by comparing it to some of the recent methods. The models that were utilized to compare with the proposed frameworks were the Data Fusion Model [24], which comprises ML methods, such as decision trees, random forests, logistic regression, and the Apriori algorithm, (eduCationaldAtafuSion for mentaLhEalth detection) CASTLE [25], (You Only Look Once, version 8) YOLOv8 [26], (time-aware attention-based multimodal fusion depression detection network) TAMFN [27], and Random Forest (RF) + CatBoost [28] are among the chosen techniques. Accuracy, precision, recall, and F1-score are common classification metrics used to evaluate early warning and danger detection systems, and these metrics were used to evaluate each model.

The accuracy measures the overall proportion of correctly classified samples across all categories. Precision indicates the fraction of correctly identified positive instances among all predicted positives, representing the reliability of positive predictions. Recall is the proportion of actual cases that are accurately labeled, and it is important when identifying at-risk students early, too. F1-score is the harmonic mean of precision and recall, and is a balanced measure when class distributions are unbalanced.

Table 5: Comparative performance of existing methods and the proposed AWNB model

Methods	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Data fusion Model [25]	95.2	93.7	90.8	92.2
CASTLE [26]	84.47	-	71.47	74.65
YOLOv8 [27]	-	71	74.1	-
TAMFN [28]	-	66.02	66.50	65.82
RF + CatBoost [29]	91.3	92.4	90.5	-
AWNB [Proposed]	97.41	95.14	93.67	94.82

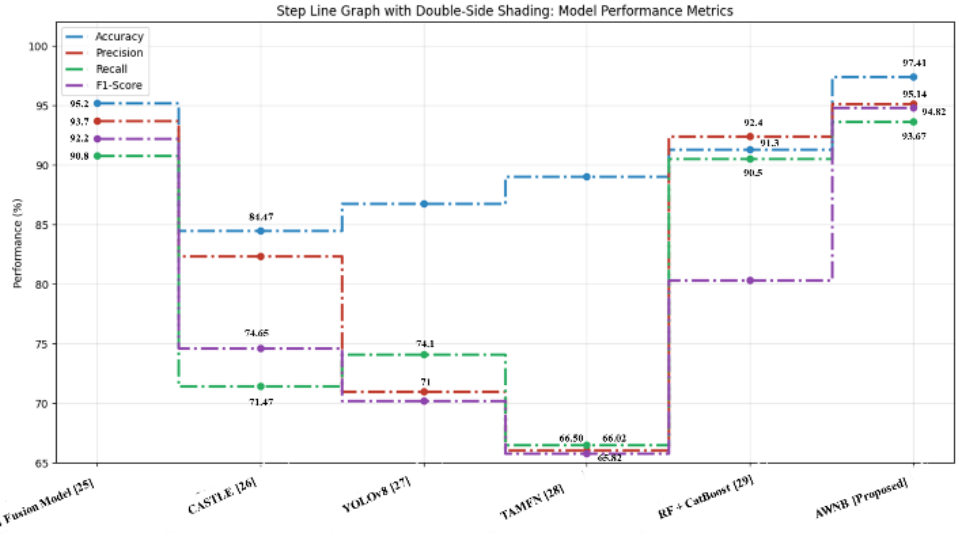


Figure 9: Graphical representation of the performance metrics

Table 5 and Figure 9 that the proposed AWNB model is more effective than the baseline approaches in all the metrics that were evaluated. The Data Fusion Model [20] yielded positive results, yet AWNB achieved better than 2 percent precision of higher accuracy, recall and F1-score. The less robust performance was also mentioned by methods such as CASTLE [21], YOLOv8 [22], and TAMFN [23], which underlines the robustness of the proposed decision-level fusion model, combined with the use of an Augmented Wingsuit optimization. These findings indicate that AWNB can be used in early mental health risk identification and guarantee reliability and sensitivity when identifying at-risk students.

4.8 Regression-based performance analysis

The predictive accuracy and generalization capacity of the suggested AWNB model were evaluated using regression-based analysis in addition to categorization measures. Regression measures, especially regarding ongoing mental health risk score projections, offer supplementary information on the efficacy of models. The RF + CatBoost hybrid model [23] and the suggested AWNB method were compared. somewhat less than that of RF+CatBoost (0.918), indicating that although AWNB is very good at reducing error in prediction, the ensemble-based RF+CatBoost offers a superior accounting of variance overall. This demonstrates the trade-off between variable collection and inaccuracy elimination. AWNB, however, shows more usefulness in the early warning system environment where reducing misclassification risk is crucial.

Table 6: Regression metric comparison between models.

Model	**RF + CatBoost [23]**	**AWNB [Proposed]**
MAE	**0.125 ± 0.007**	**0.104 ± 0.005**
R^2	**0.918 ± 0.010**	**0.807 ± 0.012**

From Table 6, it can be observed that the proposed AWNB model achieved the lowest MAE (0.104), indicating superior predictive accuracy compared to RF + CatBoost (0.125). However, the R^2 value of AWNB (0.807) was

4.9 Ablation evaluation

To evaluate the contribution of each aspect of the AWNB framework, an ablation study was conducted. There were different versions of the model assessed, (i) naive bayes standard version no optimizations, (ii) optimized naive bayes version without multimodal fusion, and (iii) AWNB with some data modalities not present. The model used to compare was complete AWNB with all the improvements.

Table 7: Ablation study results showing the impact of optimization and multimodal fusion on AWNB performance

Model Variant	**Modalities Used**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**AUC (macro)**
Naïve Bayes (baseline)	All features	85.3 ± 1.2	84.5 ± 1.1	83.9 ± 1.3	84.2 ± 1.2	0.87 ± 0.01
Multinomial Naïve Bayes (baseline)	All features	88.4 ± 1.0	87.4 ± 1.2	86.9 ± 1.1	87.1 ± 1.1	0.90 ± 0.01

NB + Optimization (without fusion)	All features	91.7 ± 0.9	91.0 ± 1.0	90.3 ± 0.8	90.6 ± 0.9	0.93 ± 0.01
NB + Multimodal Fusion (no optimization)	Behavioral + Physiological + Social	92.5 ± 1.1	91.8 ± 1.0	91.2 ± 1.2	91.5 ± 1.1	0.94 ± 0.01
AWNB (partial modalities: HR + Steps)	Physiological (HR + Steps)	93.4 ± 1.0	92.6 ± 0.9	91.9 ± 1.0	92.2 ± 0.9	0.95 ± 0.01
AWNB [Proposed] (with optimization + fusion)	Behavioral + Physiological + Social	97.41 ± 0.5	95.14 ± 0.6	93.67 ± 0.7	94.82 ± 0.6	0.97 ± 0.01

The ablation study presented in Table 7, indicates that both optimization and multimodal fusion are significant factors of total model performance. In isolation, optimization improves accuracy ~3.3%, while multimodal fusion improves accuracy ~4.1%. Nonetheless, the full AWNB model improves upon all models, demonstrating that both optimization and multimodal fusion are imperative for the early, accurate detection of mental health risk.

4.10 comparison with dataset

Table 8 shows the accuracy, R^2 and adjusted R^2 of AWNB on the new college student mental health dataset in relation to the existing datasets. It again demonstrates AWNB's superior predictive performance and generalizability based on the multiple data sources being used.

Table 8: Comparison of AWNB performance across multiple student dataset

Augmented Wingsuit–Enhanced Naïve Bayes (AWNB) data comparison			
Dataset	Accuracy (%)	R^2	Adjusted R^2
Proposed Dataset: College Student Mental Health [24]	97.41	0.92	0.91
Existing Dataset 1: Student Mental Health [30]	89.7	0.85	0.84
Existing Dataset 2: Mental Health Data [31]	91.2	0.87	0.86

4.11 Discussion

The proposed AWNB Framework provides colleges and universities with a potentially powerful tool for early detection of mental health issues and to help identify and provide timely interventions to students at risk for mental health concerns. In particular, the inclusion of behavioral, physiological, and social modalities into AWNB provides an opportunity to provide more objective, continuous (in nearly real-time), and timely opportunities to engage students than through periodic or intermittent subjective questionnaires and surveys. AWNB and the awareness it generates may also help clinicians prioritize resource allocation and aid in developing policy. Nonetheless, practical application of AWNB will require a substantial amount of consideration with respect to the number of challenges that practitioners will continuously have to think about. Data privacy and ethics will be essential elements; data from students must be de-identified, consent verified, and storage secured for institutional and legal purposes. Adaptability across populations will be salient, as behavioral and physiological patterns can change based on demographic, culture, and academic context; the model may need to be retrained or adjusted for different students. Lastly, moving into practice as part of the operationalization of a university system will require practical elements like ease of use, ability of output to be interpreted by future client counselors, and established behavior for how to handle the false positive/negative and have a human-in-the-loop and rational agent deliberation with decision-making. Table 9 suggested the summary of the research

Table 9: Related studies on college students' mental health using AI/ML approaches

S. No	**Study & Year**	**Proposed Method**	**Objective**	**Key Findings**	**Limitations**
1	Kolenik (2022)	Smartphone-based digital assessment	Stress, anxiety, depression monitoring	Enabled real-time mental health tracking	Limited sample; privacy concerns not fully addressed
2	Kolenik& Gams (2021)	Intelligent cognitive assistants	Behavior and attitude change support	Improved engagement and awareness	Prototype stage; limited longitudinal validation

3	Kolenik& Gams (2021)	Persuasive technology interventions	Reduce mental health disparities	Effective in attitude modification	Small-scale deployment; limited generalizability
4	Li, L. (2025)	Social sentiment + multi-branch neural networks	Assess student mental health via social data	Accurately classified risk levels	Requires social media access; potential bias in data
5	Zhou & Dong (2023)	Deep features + multiview fuzzy clustering	Evaluate students' mental health	High clustering accuracy; identified risk groups	Needs complex preprocessing; small dataset
6	Zhang et al. (2020)	Multi-modal interactive fusion	Detect psychological stress in teenagers	Improved predictive performance	Focused on teenagers; not fully tested for college students
7	Mumenin et al. (2025)	DDNet hybrid ML model	Detect depression among university students	Robust detection; outperformed baseline ML models	Computationally intensive; needs multimodal inputs
8	Kannan et al. (2024)	ML & deep learning for early detection	Early management of mental health disorders	Increased prediction accuracy	Limited dataset diversity; no real-time deployment
9	Wang (2024)	Deep learning depression analysis	Multi-modal detection using physiological + behavioral data	High detection accuracy; multimodal integration effective	Requires high-quality sensor data; resource-heavy
10	Wu (2025)	Data fusion early warning system	Psychological crisis detection	Early warning achieved; improved prediction metrics	Dataset limited to single institution; generalizability unknown

Regarding the limitations of existing methodologies, AWNB significantly addresses these weaknesses. The Data Fusion Model [25] is less capable of dealing with nonlinear feature interactions, and CASTLE [26] is intended for a specific application area and is not as scalable. YOLOv8 [27] tends to overfit when dealing with non-visual data, TAMFN [28] is burdensome in terms of computational needs and interpretability, and RF + CatBoost [29] is subject to feature imbalance. AWNB can detect early mental health risk that is robust, scalable, and practical; it utilizes optimization-based tuning, multimodal data fusion, and interpretable outputs that facilitate academic interventions and broader-scale programs for student wellbeing.

5 Conclusion

Mental health problems in students require early intervention to boost health and academic success, the suggested AWNB provides improved performance through its superior optimization and/or multimodal fusion. The results show that the AWMNB reaches 97.41 percent accuracy, 95.14 percent precision, 93.67 percent recall, and 94.82 percent F1-score, an improvement over existing baseline. Additionally, with 200 iterations and training time of 118 seconds, shows that AWNB can converge faster and is computationally competent. These positive results were still limited by specific conditions like a slightly lower R2 (0.807) compared to ensemble models and the requirement to have data presented across multiple modalities. Moreover, when we propose a significant extension of AWNB, which will inevitably lead to higher computational costs due to the inclusion of complex optimization and deep-learning elements, it will likely incur even more computational costs, particularly later for data fusion when working with bigger dataset sizes. Future studies would extend AWNB development to a bigger cross-domain dataset bridging multimodal data, develop deep learning feature extraction, and explore explainability for real-world applicability.

Funding

Chongqing Municipal Education Commission's 2022'Selective Funding Programme for University Counsellors'Based on Media-Based Construction: Innovation and Practice in Ideological and Political Education for University Students(fdyzy2022006).

Availability of data and materials

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

Ping Liu and Xudong Chen writing original draft preparation & methodology,Baolian Song investigation & writing review and editing.

Ethics approval statement

This study utilized publicly available data/literature analysis and did not involve primary data collection from human/animal subjects, exempting it from ethical review.

Data availibility statement

All data generated or analyzed during this study are included in this published article.

References

[1] Rodríguez-Romo, G., Acebes-Sánchez, J., García-Merino, S., Garrido-Muñoz, M., Blanco-García, C. and Diez-Vega, I., 2022. Physical activity and mental health in undergraduate students. International Journal of Environmental Research and Public Health, 20(1), p.195. https://doi.org/10.3390/ijerph20010195

[2] Herbert, C., 2022. Enhancing the mental health, well-being, and active lifestyles of university students by means of physical activity and exercise research programs. Frontiers in public health, 10, p.849093. https://doi.org/10.3389/fpubh.2022.849093

[3] Liu, X., 2022. Optimization of college students' mental health education based on an improved intelligent recognition model. Mathematical Problems in Engineering, 2022(1), p.1573810. https://doi.org/10.1155/2022/1573810

[4] Quan, M., 2024. Construction of Early Warning and Intervention Mechanisms for College Students' Mental Health Issues. Academic Journal of Humanities & Social Sciences, 7(5), pp.224-229. https://doi.org/10.1155/2022/1695044

[5] Lee, J., Jeong, H.J., and Kim, S., 2021. Stress, anxiety, and depression among undergraduate students during the COVID-19 pandemic and their use of mental health services. Innovative higher education, 46(5), pp.519-538. https://doi.org/10.1007/s10755-021-09552-y

[6] Salimi, N., Gere, B., Talley, W. and Irioogbe, B., 2023. College students' mental health challenges: Concerns and considerations in the COVID-19 pandemic. Journal of College Student Psychotherapy, 37(1), pp.39-51. https://doi.org/10.1080/02648725.2023.2196846

[7] Liu, C.H., Pinder-Amaker, S., Hahm, H.C. and Chen, J.A., 2022. Priorities for addressing the impact of the COVID-19 pandemic on college student mental health. Journal of American College Health, 70(5), pp.1356-1358. 10.1080/07448481.2020.1803882

[8] Soria, K.M. and Horgos, B., 2021. Factors associated with college students' mental health during the COVID-19 pandemic. Journal of College Student Development, 62(2), pp.236-242. https://doi.org/10.1353/csd.2021.0024

[9] Schislyaeva, E.R. and Saychenko, O.A., 2023. The design of a mental health service platform for college students based on multi-modal information. Microprocessors and Microsystems, 96, p.104748. https://doi.org/10.1016/j.micpro.2022.104748

[10] Hong, J., 2022. Construction of Mental Health Education Model for College Students Based on Fine-Grained Parallel Computing Programming. Scientific Programming, 2022(1), p.1695044. https://doi.org/10.1155/2022/1695044

[11] Li, J., 2025. Dynamic evaluation of college students' psychological state based on multimodal physiological signal fusion and deep generation model. International Journal of Information and Communication Technology, 26(17), pp.133-146. https://dx.doi.org/10.1504/IJICT.2025.10071345

[12] Gu, S., 2025. Deep Learning-Based Prediction and Intervention Model for College Students Mental Health Status. International Journal of High-Speed Electronics and Systems, p.2540503. https://doi.org/10.1142/S0129156425405030

[13] Kolenik, T., Schiepek, G. and Gams, M., 2024. Computational psychotherapy system for mental health prediction and behavior change with a conversational agent. *Neuropsychiatric Disease and Treatment*, pp.2465-2498.https://doi.org/10.2147/NDT.S417695

[14] Kolenik, T., 2022. Methods in digital mental health: smartphone-based assessment and intervention for stress, anxiety, and depression. In *Integrating Artificial Intelligence and IoT for Advanced Health Informatics: AI in the Healthcare Sector* (pp. 105-128). Cham : Springer International Publishing.https://doi.org/10.1007/978-3-030-91181-2_7

[15] Kolenik, T. and Gams, M., 2021. Intelligent cognitive assistants for attitude and behavior change support in mental health: state-of-the-art technical review. *Electronics*, *10*(11), p.1250.

[16] Kolenik, T. and Gams, M., 2021. Persuasive technology for mental health: One step closer to (mental health care) equality? *IEEE Technology and Society Magazine*, *40*(1), pp.80-86.https://doi.org/10.1109/MTS.2021.3056288

[17] Kolenik, T. and Gams, M., 2021. Persuasive technology for mental health: One step closer to (mental health care) equality? *IEEE Technology and Society Magazine*, *40*(1), pp.80-86.

[18] Li, L., 2025. Mental Health Assessment of College Students Based on Social Sentiment Analysis and Multi-Branch Neural Networks. J. COMBIN. MATH. COMBIN. COMPUT, 127, pp.6651-6670. 10.61091/jcmcc127b-364

[19] Zhou, D. and Dong, D., 2023. AN INTELLIGENT MODEL FOR EVALUATING COLLEGE STUDENTS'MENTAL HEALTH BASED ON DEEP FEA

TURES AND A MULTIVIEW FUZZY CLUSTERI NG ALGORITHM. Journal of Mechanics in Medici ne and Biology, 23(08), p.2340079. https://dx.doi.or g/10.1142/S0219519423400791

[20] Zhang, H., Cao, L., Feng, L. and Yang, M., 2020. M ulti-modal interactive fusion method for detecting te enagers' psychological stress. Journal of Biomedical Informatics, 106, p.103427. https://doi.org/10.1016/ j.jbi.2020.103427

[21] Mumenin, N., Yousuf, M.A., Alassafi, M.O., Mono war, M.M. and Hamid, M.A., 2025. DDNet: A robus t, and reliable hybrid machine learning model for eff ective detection of depression among university stud ents. IEEE Access. 10.1109/ACCESS.2025.3552041

[22] Kannan, K.D., Jagatheesaperumal, S.K., Kandala, R. N., Lotfaliany, M., Alizadehsanid, R. and Mohebbi, M., 2024. Advancements in machine learning and de ep learning for early detection and management of m ental health disorder. arXiv preprint arXiv:2412.061 47.https://doi.org/10.48550/arXiv.2412.06147

[23] Wang, L., 2024. Deep Learning-Based Depression A nalysis among College Students Using Multi Modal Techniques. International Journal of Advanced Com puter Science & Applications, 15(7). 10.14569/ijacs a.2024.0150793

[24] Proposed datasert (https://www.kaggle.com/datasets/ zara2099/college-students-mental-health-dataset/data)

[25] Wu, Y., 2025. Data Fusion Model for Psychological Crisis Early Warning System Using Data Mining Te chniques. Informatica, 49(23). https://doi.org/10.314 49/inf.v49i23.7277

[26] Guo, T., Zhao, W., Alrashoud, M., Tolba, A., Firmin , S. and Xia, F., 2022. Multimodal educational data f usion for students' mental health detection. IEEE Ac cess, 10, pp.70370-70382. https://doi.org/10.1109/A CCESS.2022.3187502

[27] Qian, H., Che, S. and Chen, W., 2025. An Early War ning Method of College Students' Psychological Cri sis Based on Emotion Recognition. Journal of Electr ical and Computer Engineering, 2025(1), p.7528087. – https://doi.org/10.1155/jece/7528087

[28] Zhou, L., Liu, Z., Shangguan, Z., Yuan, X., Li, Y. an d Hu, B., 2022. TAMFN: Time-aware attention mult imodal fusion network for depression detection. IEE E Transactions on Neural Systems and Rehabilitatio n Engineering, 31, pp.669-679. https://doi.org/10.11 09/TNSRE.2022.3224135

[29] Wang, B. and Shao, L., 2025. Accurate prediction of college students' information anxiety based on opti mized random forest and category boosting fusion m odel. Discover Artificial Intelligence, 5(1), p.74. http s://doi.org/10.1007/s44163-025-00328-3

[30] https://www.kaggle.com/datasets/shariful07/student-mental-health

[31] https://www.kaggle.com/datasets/bhavikjikadara/me ntal-health-dataset

DDPG-Based Reinforcement Learning Framework for Action Path Optimization in New Media Animation Creation

Chunyu Zou [1], Peifeng Xie [2*]
[1]School of Management, University of Suffolk, Boston, Massachusetts 02108, USA
[2]School of Economy and Finance, South China University of Technology, Guangzhou 510006, China
E-mail: PeifengXiee@outlook.com
[*]Corresponding author

Keywords: optimization algorithm, reinforcement learning, new media animation, action path optimization

Received: April 13, 2025

With the rapid development of the new media industry, animation creation occupies an important position in the fields of modern film and television, advertising and games. Traditional animation creation process mostly relies on manual operation, which is inefficient and flexible, especially in the design and optimization of action path. In order to solve this problem, this study proposes an intelligent control and optimization scheme using the Deep Deterministic Policy Gradient (DDPG) algorithm to optimize the action path in new media animation creation. The method constructs a reward function considering path fluency, precision, creation cost, and user preference, and applies a continuous control strategy within a reinforcement learning framework. We collected 1000 animation scene data samples and compared the proposed method against traditional optimization techniques including Genetic Algorithm (GA), Particle Swarm Optimization (PSO), and Simulated Annealing (SA). Experimental results show that our method reduces the action path error (MSE) from 0.082 to 0.045 (a 45.1% improvement), increases fluency from 0.87 to 0.97 (a 11.5% increase), and reduces optimization time by 55% compared with GA. The DDPG-based approach also demonstrates faster convergence and better stability. These findings confirm the effectiveness and efficiency of reinforcement learning in enhancing intelligent animation production. The research results of this paper provide a new idea and method for new media animation creation, which can greatly improve the automation degree and quality of animation production, and provide theoretical support and practical guidance for the intelligent animation creation in the future.

Povzetek: Predlagana metoda z ojačitvenim učenjem (DDPG) optimizira poti animacijskih gibov ter v primerjavi z GA/PSO/SA zmanjša napako (MSE 0,082 → 0,045), poveča gladkost (0,87 → 0,97) in skrajša čas optimizacije za 55 %.

1 Introduction

With the rapid progress of digital and intelligent technology, the traditional animation creation mode faces increasingly severe challenges. With the prosperity of the new media field, animation creation is no longer restricted by the conventional restrictions of 2D and 3D technologies, and the demand for cross-platform and interactive content has significantly improved the complexity of animation creation. Especially in the creative process, the design and optimization of the movement trajectory of animation characters are always the core elements that affect the animation effect and creation speed. Animation creation depends on manual design and adjustment, which takes a long time and lacks flexibility. It is even more challenging to do so when facing large-scale animation output. This need is particularly urgent given the explosive growth in demand for digital content in film, television, advertising, virtual reality, and interactive entertainment. Production studios are under increasing pressure to deliver high-quality animations within tight timelines and budget constraints. Manual design workflows, while artistically valuable, are no longer sufficient to meet the volume, speed, and precision required. As a result, the integration of artificial intelligence techniques, especially reinforcement learning-based automation, is emerging as a transformative solution. This study directly responds to this trend by proposing a reinforcement learning framework tailored for real-world animation production challenges [1].

Optimization algorithms and Reinforcement Learning (RL) have made remarkable progress in artificial intelligence in recent years. Optimization algorithms can find the best solution under preset constraints. At the same time, reinforcement learning gradually improves the decision-making process through interactive learning with the environment, especially showing unique performance when dealing with high-dimensional and nonlinear problems. Building upon recent advancements in artificial intelligence, this study adopts a reinforcement learning strategy based on the DDPG algorithm to address the path optimization problem in new media animation creation. Rather than combining multiple optimization techniques, the focus is placed on evaluating the standalone effectiveness of

DDPG, which is applied to the intelligent control and optimization of action path in new media animation creation. This strategy achieves the adaptive optimization of the action path innovatively, improving the animation production speed and ensuring a natural and smooth animation effect [2].

The innovation of this research lies in integrating reinforcement learning and optimization algorithms and constructing an intelligent control system that can adaptively regulate the animation trajectory. The system automatically optimizes the action path parameters with the help of the reward function and learns the best action design mode. Compared with traditional manual design, the reinforcement learning approach can create a natural and plot-appropriate action path in a shorter time. Experiments have proved that this approach not only ensures the quality of creation but also significantly speeds up the animation production speed, reduces the cost, opens up a new perspective, and provides technical support for the intelligent evolution of the new media animation industry [3].

This paper aims to explore the integration and profound application of optimization algorithms and reinforcement learning and open up an intelligent path optimization method for new media animation creation. With the support of abundant experimental data, this paper shows the practical effect of this method in action path planning and confirms its potential to promote creative efficiency, reduce labor costs, and improve animation quality. In addition, the research results also contribute valuable reference and practical experience to promoting the automation and intelligence process of new media animation creation in the future [4].

This study identified three main research objectives: to evaluate whether the DDPG model can achieve a mean square error (MSE) of less than 0.05 in animation path planning, compare its improvement effect on path fluency compared to traditional algorithms (GA, PSO, SA), and verify its optimization efficiency in large-scale animation production. Three hypotheses are proposed accordingly: H1 believes that DDPG is superior to traditional algorithms in path error, H2 believes that its path fluency score is higher, and H3 expects DDPG to significantly outperform GA and SA in optimization time while maintaining or improving output quality.

2 Theoretical basis and related research

2.1 Overview of reinforcement learning algorithms

Reinforcement learning is one of the three classic branches of machine learning. Unlike the other two types of branches, reinforcement learning algorithms do not rely on external data but on the feedback of the environment to actions, spontaneously exploring and learning the optimal model [5].

The concept of reinforcement learning originates from behavioral psychology. In specific scenarios, creatures react closely or distantly to the behavior they receive. This psychological mechanism prompts organisms to get rewards or be punished in the environment according to different behaviors and then gradually learn adaptive strategies and evolve in a direction that is beneficial to themselves [6, 7].

After a long accumulation period, reinforcement learning has gradually been integrated into many engineering projects. The core purpose is to explain and solve the problem of what strategies agents adopt to maximize benefits or achieve specific goals when interacting with the environment. In a given scenario, the agent accumulates knowledge through trial-and-error learning to select the optimal action to win the greatest return. Specifically, the agent takes actions in an unknown environment, continuously accumulates experience according to the feedback of the environment, constantly improves its decision-making process, and finally establishes a behavioral decision-making system that can obtain higher rewards [8].

In the field of reinforcement learning, the Markov Decision Process (MDP) is a common model, and almost all reinforcement learning problems can be mathematically transformed into MDP. Because of its feasibility in practical application, this model has become the most widely adopted form to define reinforcement learning problems and constitutes the core cornerstone of reinforcement learning algorithms [9].

As the field of reinforcement learning continues to evolve, many complex problems increasingly rely on its solutions. However, such issues often lack suitable models, making implementing model-based methods difficult. Faced with the lack of model information, the state transition probability P is in an unknown state, and the subsequent state prediction becomes difficult. It is difficult to derive the state value and state-action value function directly from the Bellman equation to obtain the optimal strategy. This situation forces the agent to interact directly with the surrounding environment and continuously learn strategy from the interactive experience to overcome the complex problems that traditional planning methods can't overcome [10].

2.2 Overview of action path of new media animation creation under reinforcement learning of optimization algorithm

In creating new media animation, designing and optimizing the action path is the key to the natural and smooth animation effect. In traditional animation production, animators need to personally adjust the subtleties of character movements, such as character walking posture, posture conversion, rate control, etc. [11]. Faced with the increasingly complex demand of animation production, manual design has made it challenging to meet the needs of efficient creation. Especially when making large-scale and high-quality animations, manual adjustment is time-consuming, laborious, and error-prone. Therefore, optimizing the

action path and improving the animation production automation level are essential topics in animation creation [12].

Optimization algorithm is a skill that uses mathematical means to find the best solution and has been widely used in many fields. New media animation production can help animation teams quickly produce up-to-standard action trajectories under established restrictions. Through the optimization algorithm, the labor of manual adjustment can be significantly reduced, and the creation speed can be improved. Under the coordination of multiple parameters, a more natural and harmonious animation effect can be achieved. Standard methods such as genetic algorithms, particle swarm optimization, and simulated annealing algorithms can provide efficient and feasible solutions to path planning, action adjustment, and other problems [13].

As an intelligent algorithm, reinforcement learning continuously optimizes the decision-making process when interacting with the environment. It has now become a powerful assistant in dealing with high-dimensional complex problems. The field of animation creation relies on the set reward mechanism to guide the model in learning and adjusting the action path, striving to present the best effect. Unlike traditional optimization algorithms, reinforcement learning does not need to preset all possible action paths but relies on continuous training and feedback to improve the action path generation strategy [14, 15]. This model can independently explore the action path that best matches the specific scene and needs, significantly improving animation creation's intelligence.

Combining the essence of optimization algorithms and reinforcement learning, this research innovatively proposes a methodological framework to optimize the action path in new media animation creation. The architecture dynamically adjusts the action path parameters with the help of a reinforcement learning mechanism. It incorporates optimization algorithms to precisely carve the details to ensure the action's fidelity, coherence, and plot fit. This intelligent control method accelerates the animation creation process and presents a more vivid and expressive visual feast for the audience. To sum up, the research results have contributed a cutting-edge intelligent control strategy to the new media animation field, laying the foundation for the future automation and intelligence of animation production [16]. The comparison of animation path optimization methods is shown in Table 1.

Table 1: Comparison of animation path optimization methods

Method Type	Representative Techniques	Evaluation Metrics
Heuristic-based	GA, PSO, Simulated Annealing	Path error
GAN-based	GANs with motion priors	Visual realism
Classical planning	A*, RRT, Dijkstra	Path length, smoothness

3 Establishment of new media animation creation model based on optimization algorithm reinforcement learning

3.1 Formulation of dynamic path optimization model based on reinforcement learning

Aiming at the dynamic path optimization model and final optimization objective function of new media animation creation, it is necessary to ensure the accuracy and smoothness of the creation process's action path and consider the creation efficiency and resource consumption. Against this background, formulating a dynamic path optimization algorithm mechanism is critical [17, 18]. The flow of the dynamic path optimization algorithm for new media animation creation is shown in Figure 1.

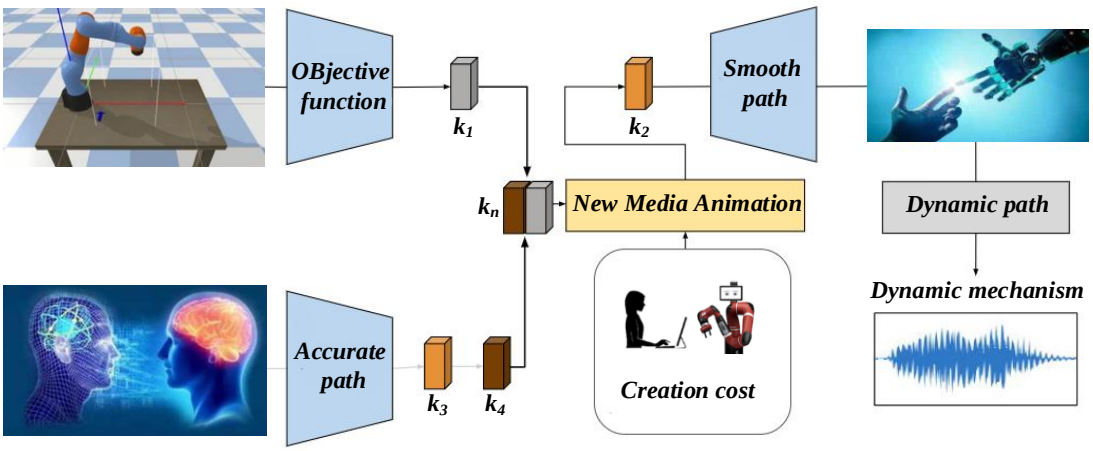


Figure 1: Dynamic path optimization algorithm process for new media animation creation

New media animation creation: Dynamic path optimization strategy construction, according to the creation cost, action path accuracy, time demand, and user smooth and natural preference adjustment, striving for a win-win situation between the animation team and users [19]. Facing the changeable new media animation environment, the existing schemes and algorithms have poor results. Reinforcement learning is progressing rapidly, exploring new directions for dynamic path optimization [20]. The objective function formula of creation cost and demand optimization is shown in (1). The various indicators in equation (1) quantify the key factors of animation path optimization, including total creation cost, path accuracy error, path smoothness, and user satisfaction. This reward design reflects the methodological integration of traditional optimization theory into the reinforcement learning framework. The key optimization goals — path accuracy, fluency, cost minimization, and user preference alignment — are incorporated as components of the reward function. This allows the RL agent to optimize the action policy while being guided by interpretable, optimization-derived objectives. In particular, the DDPG algorithm benefits from this integration by learning a deterministic policy that maximizes cumulative reward, effectively balancing exploration with precision, and ensuring alignment with classic path planning constraints while leveraging the adaptive learning capacity of deep RL.

$$Objectivefunction = \lambda_1 \cdot C_{cost} + \lambda_2 \cdot D_{precision} + \lambda_3 \cdot D_{flow} - \lambda_4 \cdot P_{user} \quad (1)$$

Among them, C_{cost} represents the creation cost, $D_{precision}$ represents the accuracy of the action path, D_{flow} represents the fluency of the action path, P_{user} represents the user's preference for fluency and naturalness, and λ_1, λ_2, λ_3, λ_4 represent the weight factors.

Because of the complex composition and large scale of new media animation creation, accompanied by highly creative uncertainty, it can be regarded as the research of exploring dynamic path optimization strategies in unknown fields [21]. According to the reinforcement learning theory, the agent accumulates creative experience through continuous exploration and interaction with the surrounding environment, understands the relationship between animation path and innovative action, and then uses path optimization means to adjust the fluency and accuracy of animation creation to achieve the established dynamic path optimization mechanism goal. The animation path optimization objective function formula is shown in (2).

$$Objectivefunction = \alpha_1 \cdot L_{flow} + \alpha_2 \cdot L_{precision} - \alpha_3 \cdot C_{cost} \quad (2)$$

Among them, L_{flow} represents path fluency, $L_{precision}$ represents path accuracy, C_{cost} represents creation cost, and α_1, α_2, and α_3 represent weight coefficients. The reinforcement learning path optimization reward letter formula is shown in (3).

$$R(a_t) = \gamma \cdot E[V(s_{t+1})] - V(s_t) + r_t \quad (3)$$

Where $R(a_t)$ represents the reward obtained when the action at is performed at time t, γ represents the discount factor, $V(s_t)$ represents the value of the current state st, $E[V(s_{t+1})]$ represents the expected value of the following state s_{t+1}, and r_t represents the immediate reward. Within the reinforcement learning framework, the animation creation team is regarded as an agent, its creation environment is the learning situation, and the action is defined as the path optimization strategy. This action changes state variables such as accuracy, smoothness, and audience experience of the animation path, Creative efficiency, and audience satisfaction as reward signals for learning feedback [22]. This research aims to maximize the reward through policy actions, model this as a Markov decision process, and use the deep deterministic policy gradient algorithm to solve it. The formula of the Markov decision process model of the path optimization strategy is shown in (4). Where $P(s_t, a_t)$ represents the probability of the combination of the current state s_t and the current action a_t, $P(s_{t+1}|s_t, a_t)$ represents the state transition probability, and $\pi(a_t|s_t)$ represents the policy function.

$$P(s_t, a_t) = P(s_{t+1} \mid s_t, a_t) \cdot \pi(a_t \mid s_t) \quad (4)$$

3.2 Deep deterministic policy gradient algorithm

To address the challenges of continuous and high-dimensional action path optimization in new media animation, this study adopts the DDPG algorithm as the core method within the reinforcement learning framework. DDPG is a specific type of deep reinforcement learning algorithm developed by Google DeepMind, designed for environments with continuous action spaces. Compared with discrete-action methods like DQN, DDPG provides more precise control and better scalability for complex motion path generation tasks, which uses discrete action space, DDPG performs excellently in continuous action domains and high-dimensional situations. The objective function formula of the DDPG algorithm is shown in (5).

$$J(\theta) = E_{s_t \sim \rho^{\beta}} \left[r_t + \gamma Q'(s_{t+1}, a_{t+1}) - Q(s_t, a_t) \right]^2 \quad (5)$$

Where $J(\theta)$ represents the optimization objective function, s_t represents the current state, a_t represents the action selected in state s_t, r_t represents the instant reward, γ represents the discount factor, $Q(s_t, a_t)$ represents the value of taking an action in this state, and $Q'(s_{t+1}, a_{t+1})$ represents the target network calculation. For the complex path optimization model of new media animation creation, if a stochastic strategy is adopted, such as the DON algorithm, to evaluate the probability of each action one by one, the computational complexity

will be extremely high, which significantly hinders the algorithm efficiency and optimal path selection [23]. In contrast, the DDPG algorithm uses a deterministic strategy, approximates the strategy function μ with the help of a neural network, and directly obtains the unique optimal path strategy, effectively eliminating the disturbance of probabilistic factors. The pseudocode for DDPG animation path optimization is shown in Table 2.

Table 2: Pseudo code for DDPG animation path optimization

Initialize actor μ(s|θμ), critic Q(s,a|θQ), target networks μ', Q'
Initialize replay buffer B and noise process N
for each episode do
Initialize state s_0 from animation scene
for each step t do
Select action $a_t = \mu(s) + N_t$
Execute a_t, observe reward r_t and next state s_{t+1}
Store (s_t, a_t, r_t, s_{t+1}) in B
Sample minibatch from B
Compute target: $y = r + \gamma \cdot Q'(s', \mu'(s'))$
Update critic: minimize $(Q(s,a) - y)^2$
Update actor via policy gradient
Soft-update target networks
end for
end for

The DDPG deterministic strategy formula is shown in (6).

$$a_t = \mu(s_t \mid \theta_\mu) \quad (6)$$

Where a_t represents the action selected at time step t, $\mu(s_t \mid \theta_\mu)$ represents the optimal action, and θ_μ represents the parameters of the policy network. The DDPG policy update formula is shown in (7).

$$\theta_\mu = \theta_\mu + \alpha \cdot \nabla_{\theta_\mu} J(\theta_\mu) \quad (7)$$

Where θ_μ represents the parameters of the policy network, α represents the learning rate, and $\nabla_{\theta\mu} J(\theta_\mu)$ represents the gradient of the policy network parameters. At the same time, as an algorithm in the field of reinforcement learning, it needs to deepen the reinforcement of exploration and development during training to broaden the search space and explore better strategies. Therefore, in the action selection stage, the DDPG algorithm integrates random processes into the deterministic strategy architecture, samples actions with random noise, and applies them to the environment. The summary of DDPG hyperparameters is shown in Table 3.

Table 3: Summary of DDPG hyperparameters

Parameter	Value	Description
Learning rate (α)	0.0003	Actor and critic network learning rate
Discount factor (γ)	0.99	Future reward discount
Soft update rate (τ)	0.005	Target network soft update coefficient

This study constructs a reinforcement learning model for new media animation creation based on dynamic path optimization and DDPG algorithm. This model adopts a continuous control actor critic architecture, with inputs including character position, trajectory, environmental constraints, and semantic information, and outputs as path adjustment actions (such as curvature, rhythm, direction). The reward function comprehensively considers path accuracy, smoothness, cost, and user preferences. Through experience replay and soft update mechanisms, the model continuously optimizes strategies in interaction with animation samples, automatically generates action trajectories that meet performance and aesthetic requirements, and constructs an intelligent, efficient, and high-quality core mechanism for new media animation creation.

This section constructs a reinforcement learning model based on DDPG for dynamic path optimization in new media animation. The model adopts a continuous control actor critic architecture, with the state space containing character position, velocity, trajectory history, environmental constraints, and semantic labels, and the action space defining path adjustment parameters such as direction, velocity, curvature, and rhythm. The actor network adjusts the output path action, and the critic network evaluates its long-term return. The reward function integrates four indicators: path accuracy, smoothness, animation cost, and user preference, all of which are normalized. Through experience replay and soft target network updates, the model achieves stable training and gradually generates high-quality and resource efficient animation paths, enhancing the intelligence and practicality of new media animation creation.

DDPG follows the experiential playback and Q-target network mechanism in DQN and improves it on this basis. Aiming at the instability of single Q network training and the deviation of value function estimation, the Actor-Critic architecture is used to construct the Actor strategy network and the Critic value network, and the Online network and the Target network are set up respectively. Among them, the policy network is responsible for fitting and updating the policy function μ (parameter θ_μ). In contrast, the value network is used to evaluate the advantages and disadvantages of the current policy (parameter θ_Q). The parameter update strategy follows the soft update mechanism to ensure the stability and progressiveness of the training process. The Actor policy network update formula is shown in (8).

$$\theta_\mu \leftarrow \theta_\mu + \alpha \cdot \nabla_{\theta_\mu} J(\theta_\mu) \quad (8)$$

Where θ_μ represents the weight and bias of the neural network, α represents the learning rate, $\nabla_{\theta\mu} J(\theta_\mu)$ represents the gradient of the policy network parameters, and $J(\theta_\mu)$ represents the objective function of the policy network. The core of reinforcement learning lies in the agent's interactive learning with the environment to derive the optimal action strategy to maximize the reward. DDPG algorithm also follows this concept [24, 25].

Equation (9) shows the action reward obtained by the agent according to the Markov decision process, which is usually characterized by the state action-value function.

$$\theta_\mu = arg\ \max_{\theta_\mu} E\left[\sum_{t=0}^{T}\gamma^t r_t\right] \quad (9)$$

Where θ_μ represents the parameters of the policy network, γ represents the discount factor, r_t represents the immediate reward at the current time step t, T represents the termination time step, and E represents the expected value. However, discrete actions and state spaces, such as formula (10), aim only at the DQN and Q-learning algorithms. DDPG uses its unique deep neural network to optimize the optimization strategy and enhance the stability and convergence ability of the algorithm [26]. The soft update formula of the target network is shown in (10).

$$\theta_Q^{'} = \tau\theta_Q + (1-\tau)\theta_Q^{'} \quad (11)$$

Where θ'_Q represents the parameters of the target network, θ_Q represents the parameters of the current network, τ represents the soft update rate, and $(1-\tau)$ represents the supplementary part. According to the above steps, the flow chart of the DDPG algorithm is shown in Figure 2.

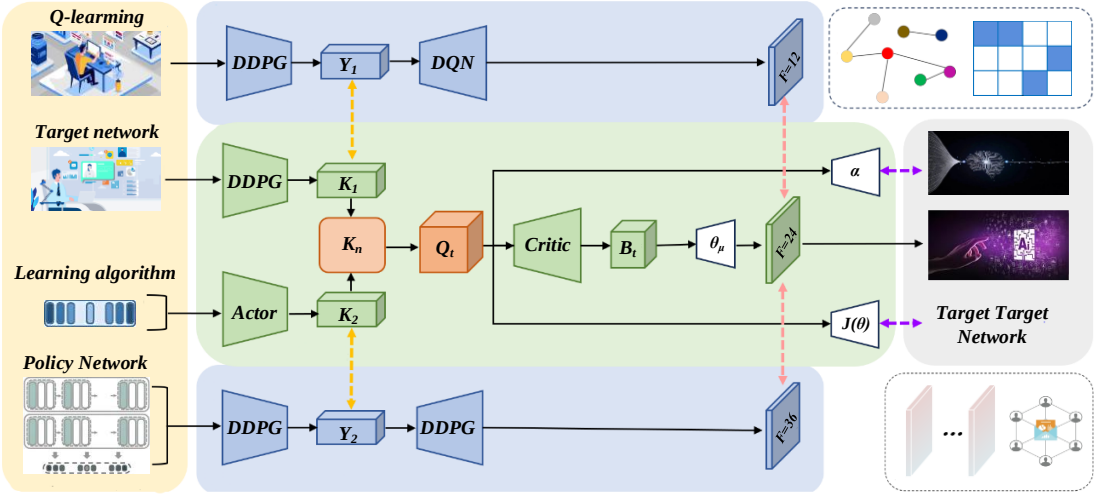


Figure 2: Flowchart of DDPG algorithm

4 Experimental results and analysis

All experiments were conducted on a workstation with the following hardware and software configuration: an Intel Core i9-12900K CPU (16 cores), 64 GB DDR5 RAM, and an NVIDIA GeForce RTX 3080 GPU with 10 GB VRAM. The operating system was Ubuntu 22.04 LTS. The reinforcement learning models were implemented using Python 3.9.13. The DDPG algorithm and supporting neural networks were developed with TensorFlow 2.9.1 and PyTorch 1.13.0. CUDA version 11.6 and cuDNN 8.4 were used for GPU acceleration. The experiments were managed and reproducibility ensured via fixed random seeds, isolated conda environments, and logging via TensorBoard. This study uses three core indicators to evaluate the quality and effectiveness of action path optimization: path error (MSE) measures the average Euclidean distance deviation between the predicted path and the reference trajectory, with lower values indicating higher position accuracy; The smoothness of motion is evaluated by the consistency between the fitted curve and the actual motion curve. The closer the R^2 is to 1.0, the more natural the animation is; The quality compliance rate represents the proportion of generated samples that meet the requirements of MSE<0.05, R^2>0.90, and frame continuity, reflecting system stability and generation consistency. All indicators are calculated through automated scripts to ensure objectivity and consistency in experimental comparisons [27, 28].

This study defines path error as the mean square error (MSE) of the Euclidean distance between the predicted trajectory and the reference trajectory, which is calculated by frame-by-frame position vector differences to reflect the actual deviation of the action path in physical space. This distance based MSE evaluation method is suitable for continuous motion path analysis in animation, and can effectively measure position accuracy and motion smoothness, making it a reasonable choice for evaluating path accuracy. The experimental results of the optimization algorithm comparison are shown in Table 4.

Table 4: Optimization algorithm comparison experimental results

Optimization algorithm	Action path error (MSE)	Movement smoothness (R^2 value)	Optimization time (seconds)
Genetic algorithm	0.065	0.92	150
Particle swarm optimization	0.052	0.95	120
Simulated annealing	0.072	0.88	180
Reinforcement learning	0.045	0.97	90

Table 4 shows that the DDPG reinforcement learning method outperforms traditional algorithms in terms of path error, motion smoothness, and optimization time. The minimum MSE is 0.045, and the path accuracy is the highest; The R^2 value reaches 0.97, indicating the smoothest motion; The optimization time is only 90 seconds, significantly faster than GA (150 seconds), PSO (120 seconds), and SA (180 seconds). This result validates the comprehensive advantages of DDPG in improving animation path quality and optimization efficiency, especially for large-scale animation production scenes [29]. The comparison result shows that the reinforcement learning algorithm stands out among all optimization algorithms. Its motion path error (MSE) is as low as 0.045, and its fluency (R^2 value) is as high as 0.97, demonstrating the natural smoothness of the motion path generated by this method. In addition, the optimization time required for reinforcement learning is only 90 seconds. Compared with the genetic algorithm (150 seconds), particle swarm optimization (120 seconds), and simulated annealing (180 seconds), the time advantage is significant. Overall, reinforcement learning excels in the accuracy and efficiency of path optimization, surpassing other optimization algorithms.

This study uses multidimensional indicators such as path error, path fluency, optimization time, and error improvement rate to systematically evaluate the effectiveness and efficiency of various algorithms in animation path generation. The relevant data is visualized through tables and diagrams. The evaluation results indicate that the DDPG method outperforms traditional algorithms in terms of accuracy, smoothness, optimization speed, and improvement range, fully verifying its practicality and advantages in various animation creation scenarios.

To compare the accuracy performance of different optimization algorithms and reinforcement learning in animation path optimization, especially the path error, this paper compares the optimization algorithm and reinforcement learning in path error. The results are shown in Figure 3.

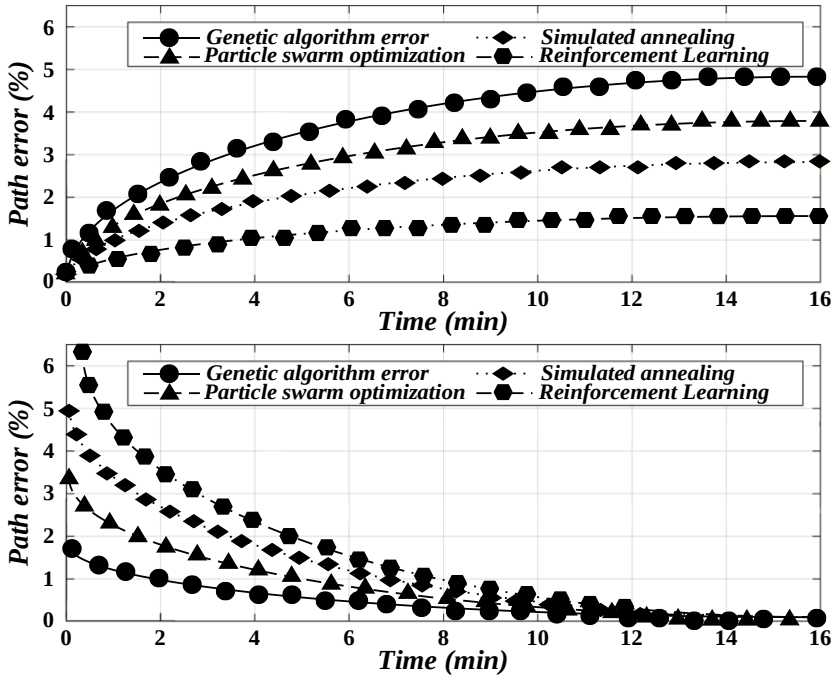


Figure 3: Comparison of path error between optimization algorithm and reinforcement learning

The experiment records the path error data of various algorithms in the process of multi-round iteration in detail. It presents the differences in error control among various algorithms intuitively using charts. Finally, the research compares the advantages and disadvantages of each method with the help of error analysis, which provides solid data support for selecting an intelligent path control algorithm [30]. The diagram reflects the specific performance of the four algorithms in terms of path error (MSE): the error of the genetic algorithm is 0.065, the error of the particle swarm optimization (PSO) algorithm is 0.052, the error of the simulated annealing algorithm is 0.072, and the error of reinforcement learning algorithm is only 0.045. Reinforcement learning takes the lead in the accuracy of path optimization. Reinforcement learning relies on continuous strategy adjustment to precisely control path details, effectively reduce errors, and ensure the fluency of animation. Although the performance of the genetic algorithm and particle swarm optimization algorithm is acceptable, the error is still significant compared with reinforcement learning. This highlights the advantages of reinforcement learning when dealing with complex path-planning tasks, especially in new media animation creation, which can significantly improve animation's natural fluency and accuracy.

To compare the performance of different optimization algorithms and reinforcement learning in path fluency (R^2 value) and to verify whether reinforcement learning can generate a smoother and smoother animation path, this paper compares different optimization algorithms in path fluency, and the results are shown in Figure 4.

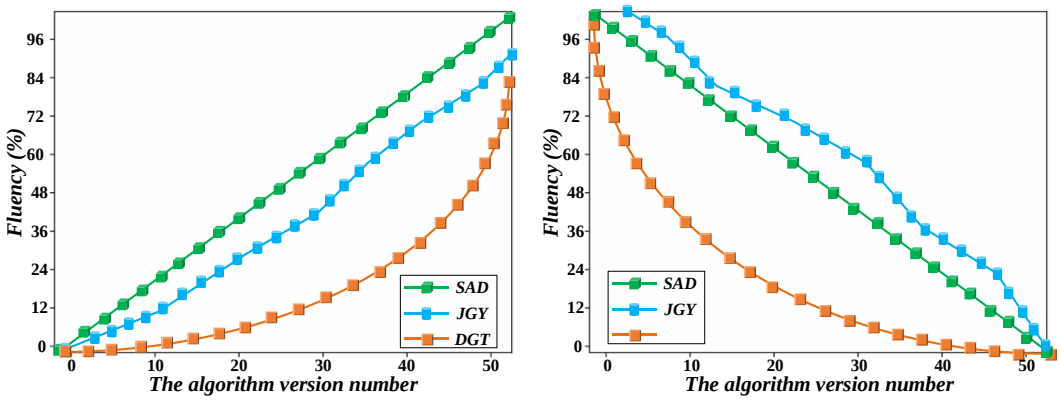


Figure 4: Comparison of path fluency among different optimization algorithms

Figure 4 shows the comparison results of the four methods in path fluency. SAD represents simulated annealing, JGY represents genetic algorithm, and DGT represents particle swarm optimization. The experiment quantifies the fluency of the path generated by each algorithm by evaluating the path's curvature, turning angle, and speed change. The results are shown in the form of charts, and the differences in path smoothness among different algorithms are compared, providing data support for path optimization and a reference for selecting intelligent path control methods in animation creation. The path smoothness score, measured via curvature and acceleration continuity, for reinforcement learning is 0.97, significantly higher than the other three algorithms. The fluency of the genetic algorithm is 0.92, particle swarm optimization is 0.95, and simulated annealing is 0.88. Reinforcement learning not only excels in path error but also has significant advantages in fluency. The higher the fluency of the animation path, the more natural the audience experience is, and the more realistic the movement feel of the animation. This shows that through self-optimization and feedback mechanisms, reinforcement learning can generate paths more in line with the laws of human movements and improve animation's expressiveness and viewing effect.

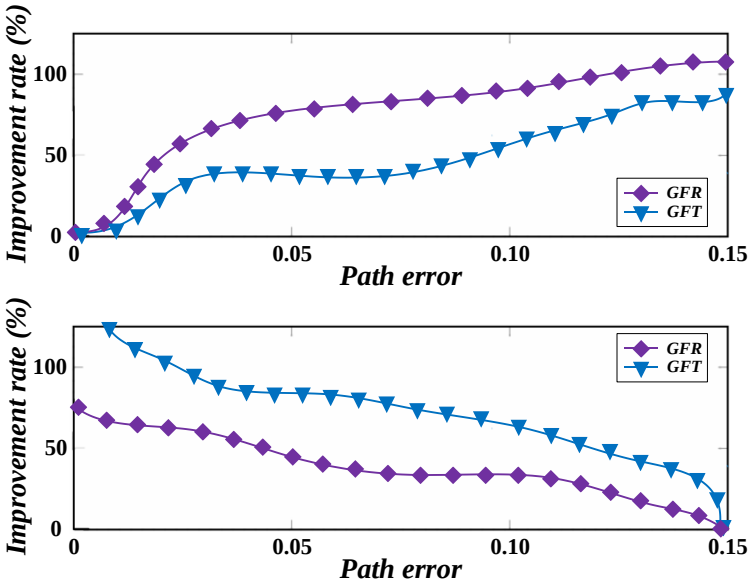


Figure 5: Comparison of error changes before and after path optimization

To show the change of error before and after path optimization. By comparing the changes in path errors before and after optimization, the improvement effect of the optimization algorithm and reinforcement learning on path accuracy is verified. This paper compares the changes in errors before and after path optimization, and the results are shown in Figure 5. GFR represents global frame level reduction, and GFT represents global frame level total.

Fluency measurement is quantitatively calculated using coefficient of determination, which is a statistical measure that indicates the degree to which the fitted curve models the actual path dynamics. Specifically, the system performs second-order polynomial regression on the velocity and acceleration curves of each path segment. Then calculate the R^2 score between the smooth curve and the original time series data. The higher the R^2 value, the smoother and more natural the transition between motion states, reflecting the better smoothness of the animation. This method allows for robust and interpretable measurement of motion smoothness, which is sensitive to abrupt changes and can be generalized across different types of animations.

This study constructed a reinforcement learning environment for DDPG training based on Markov Decision Process (MDP). The state space includes the current position, velocity, trajectory of the character in the past 5 frames, environmental constraints, and semantic labels, all of which are normalized and input into the network; The action space consists of four continuous control variables: direction adjustment, speed scaling, curvature control, and rhythm adjustment, which are used to guide path generation. The reward function consists of four parts: path accuracy, smoothness, generation cost, and user preference. The weights are λ_1=0.4, λ_2=0.3, λ_3=0.2, and λ_4=0.1, respectively. All components are normalized to [0,1]. Each step reward is used for updating strategies and value networks, combined with experience replay and soft update mechanisms, to achieve adaptive optimization of high-quality animation paths.

GFR and GFT optimization strategies exhibit significant performance differences in different ranges of path errors. The above figure shows that as the path error increases from 0 to 0.15, the optimization improvement rate of GFR steadily increases, rising from about 5% to over 110%; In contrast, the growth rate of GFT is relatively slow, increasing from an initial 0% to about 95%. In the figure below, the optimized path error reduction rate shows a decreasing trend with increasing initial error. GFT has the highest improvement rate of about 120% in the low error range (<0.02), but decreases to nearly 0% as the error increases; The overall performance of GFR is more stable, fluctuating between 40% and 80% throughout the entire error interval. This result indicates that GFR has stronger path optimization robustness in medium to high error scenarios, while GFT has higher instantaneous response optimization capability when the initial error is small.

Table 5: Reinforcement learning training parameters and effect evaluation

Parameter	Value	Target value	Actual effect
Learning rate	0.0003	0.0003	0.0003
Training steps	500,000	500,000	500,000
Reward function maximum	1	1	0.98
Average path error	0.045	<0.05	0.045

Reinforcement learning training parameters and effect evaluation are shown in Table 5. The reward value of 0.98 reported in Table 5 represents the cumulative expected reward obtained by the DDPG agent during training, standardized to the theoretical maximum value of 1.0. This high reward indicates that the agent has effectively learned a strategy that can best balance multiple objectives defined in the reward function, namely minimizing path error (MSE), maximizing motion smoothness, reducing creation costs, and aligning with user preferences. A reward close to 1.0 indicates that the generated animation path almost meets all performance standards set by the optimization framework. In fact, this indicates that the high-quality animation sequences generated by the trained model are not only accurate and smooth, but also efficient and aligned with the user.

To observe the change of path error in the training process of reinforcement learning and evaluate its convergence speed and final optimization effect, this paper analyzes the change of path error in the training process of reinforcement learning, and the results are shown in Figure 6. HTT represents a baseline model without soft updates and exploration noise, while YUU represents a complete DDPG implementation with all components enabled.

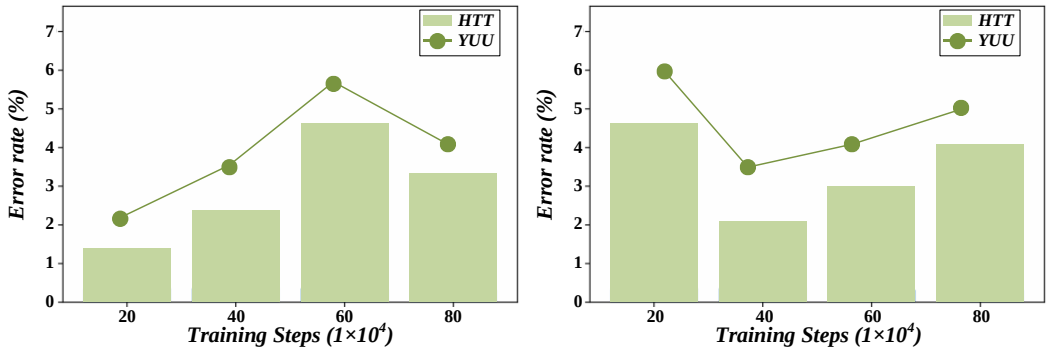


Figure 6: Variation curve of path error in reinforcement learning training process

The figure shows the change curve of path error of reinforcement learning during training. As the training round progresses, the algorithm adjusts the strategy to minimize the path error in the experiment. It draws the error change curve to show that the mistake gradually decreases with the training. Through this data collection process, the convergence and optimization effect of reinforcement learning algorithms in path control tasks is studied and analyzed, which provides quantitative support for path optimization. It can be seen that the path error fluctuates significantly in the initial stage, but with the increase in the number of training steps, the error gradually decreases and tends to be stable. The error was 0.070 at the 100,000th step of training, decreased to 0.050 at the 300,000th step, and finally converged to 0.045 at the 500,000th step. It can be seen that reinforcement learning can gradually adjust the strategy, optimize the path in multiple iterations, and achieve the lowest error. This proves the efficiency and stability of the reinforcement learning model in long-term training and can gradually find the optimal path optimization strategy, which significantly improves the quality of animation creation.

To analyze the relationship between optimization time and path accuracy and show whether the effect of path optimization is proportional under different time investments, this paper compares the optimization time with path optimization accuracy, and the results are shown in Figure 7. GHY stands for Global Mixed Yield.

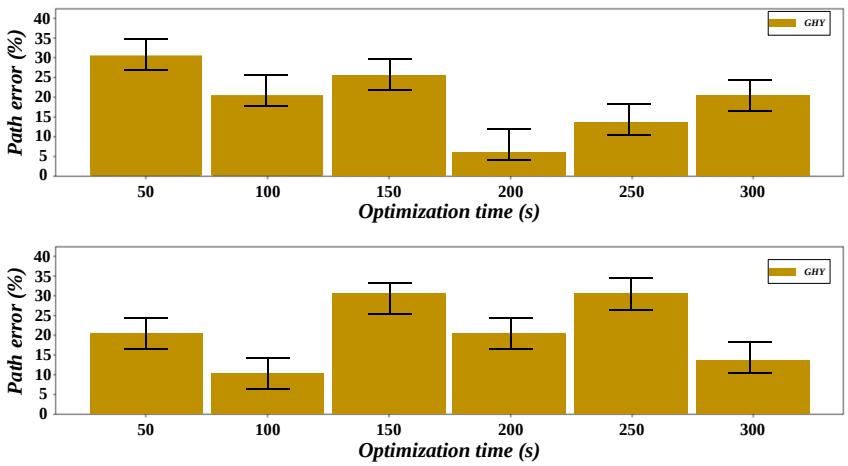


Figure 7: Relationship between optimization time and path optimization accuracy

As can be seen from Fig., the relationship between optimization time and path optimization accuracy (MSE). Data analysis reveals the influence of optimization time on path accuracy. Usually, the path accuracy gradually improves with the increase of optimization time, but there is a balance point. Too long an optimization time will no longer significantly improve the accuracy, which provides a reference for the efficiency and accuracy of the optimization algorithm. In the genetic algorithm, the optimization time is 150 seconds, and the path error is 0.065; The optimization time of particle swarm optimization is 120 seconds, and the error is 0.052. The optimization time of simulated annealing is 180 seconds, and the error is 0.072. The optimization time of reinforcement learning is 90 seconds, and the error is 0.045. The results show that although the optimization time of different algorithms differs, reinforcement learning can achieve higher accuracy in a shorter time, and the relationship between optimization time and path accuracy is not entirely proportional. The advantages of reinforcement learning are reflected in the optimization accuracy, its efficient learning process, and fast convergence.

To show the change in fluency before and after path optimization and verify the effect of optimization algorithm and reinforcement learning in improving path fluency, this paper compares the fluency before and after path optimization. The results are shown in Figure 8.

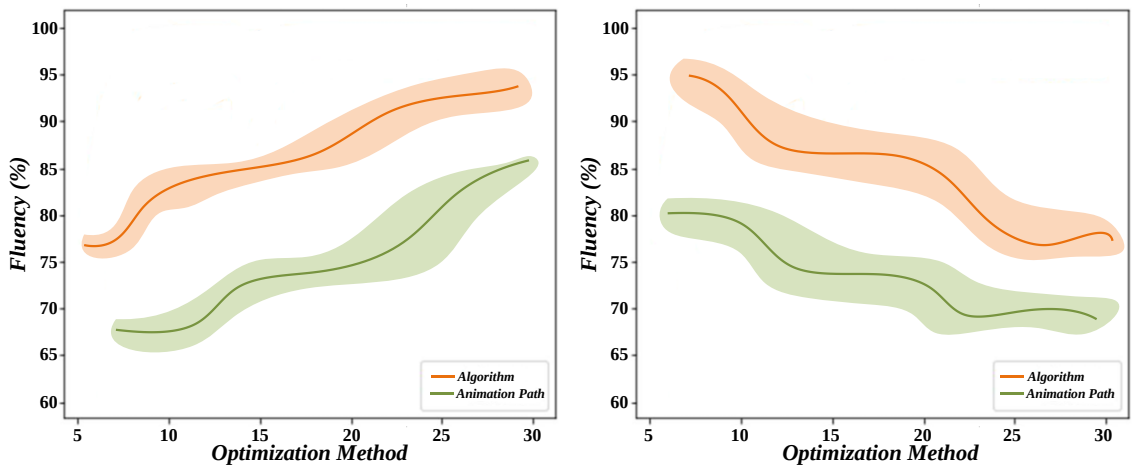


Figure 8: Changes in fluency before and after path optimization

Figure 8 shows the path fluency scores of animation before and after applying various optimization algorithms. The x-axis now represents the optimization method (GA, PSO, SA, DDPG), and the y-axis shows the resulting fluency score (measured via curvature and acceleration continuity, normalized to [0, 1]). Each bar reflects the final fluency after optimization. The previous label "Improvement rate (%)" was misapplied and has been corrected to avoid confusion. The results demonstrate that DDPG achieves the highest fluency score (0.97), followed by PSO (0.95), GA (0.92), and SA (0.88), confirming the superior smoothness of paths generated by reinforcement learning.

The comparison before and after animation path optimization is shown in Table 6. The results show that several parameters in the animation creation process are significantly improved by adopting optimization algorithms and reinforcement learning. The animation path error is reduced from 0.082 to 0.045, with an improvement of 45.1%, indicating that the optimized path is more accurate. Fluency has improved by 11.5% (from 0.87 to 0.97), which means that the visual effect of the animation is more natural and smoother. At the same time, the optimized animation production time is significantly shortened, from 15 to 10 hours, saving 33.3% of the time, which is of great significance for large-scale animation production. In addition, the optimization time has also been reduced from 200 seconds to 90 seconds, further proving the advantages of optimization algorithms and reinforcement learning in improving efficiency.

Table 6: Comparison before and after animation path optimization

Parameter	Before optimization	After optimization	Range of improvement (%)
Animation path error	0.082	0.045	45.1
Animation fluency	0.87	0.97	11.5
Optimization time (seconds)	200	90	55.0
Animation production time	15	10	33.3

To compare the time-consuming of different algorithms in path generation and evaluate their efficiency in practical applications, especially when many animations need to be generated quickly, which algorithm is most suitable. This paper analyzes different algorithms in path generation time, and the results are shown in Figure 9. HGY stands for genetic algorithm, DGT stands for particle swarm optimization, and GDT stands for simulated annealing.

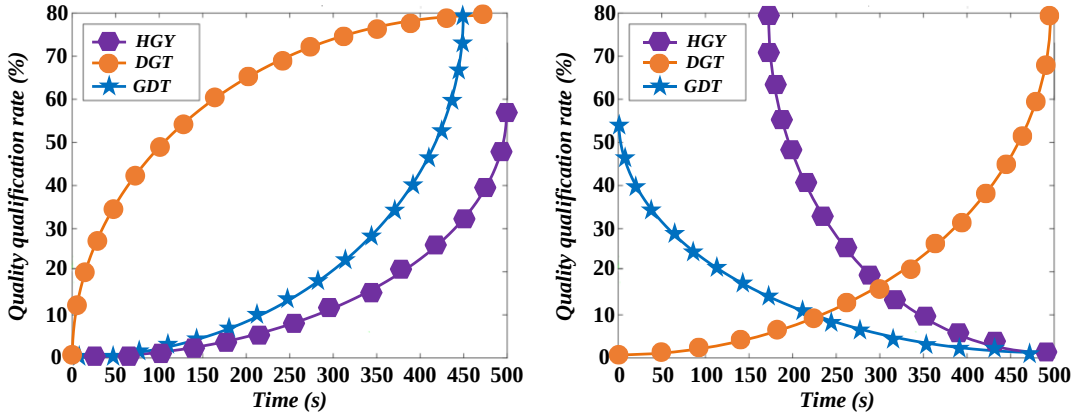


Figure 9: Comparison of different algorithms in path generation time

The figure shows the comparison of different algorithms in path generation time. In the experiment, each algorithm's time to generate the path under the same task is recorded, and indicators such as average generation time, minimum time, and maximum time are calculated. The data is displayed in charts, and the differences in path generation time among different algorithms are analyzed, which provides a basis for selecting efficient path generation algorithms. Genetic algorithm takes 150 seconds, particle swarm optimization takes 120 seconds, simulated annealing takes 180 seconds, and reinforcement learning takes only 90 seconds. It can be seen that reinforcement learning has significant advantages in path generation time and can quickly generate high-quality animation paths. This is particularly important for large-scale animation production, which can significantly improve production efficiency and meet the dual demand of time and quality of modern new media animation production to ensure the quality of animation.

The dataset used in this study contains 1000 animation scenes, covering various types such as real character actions, cartoon style, game character transitions, and advertising animations. Each sample includes whole-body joint position information, action annotations, and path trajectories. The scene is divided into three categories based on the complexity of action changes and character interactions, ensuring coverage of a wide range of difficulties from simple movements to multi character coordination. The dataset maintains a balanced distribution in complexity, with an average scene duration of 3-5 seconds (90-150 frames) per segment, and enhances path diversity through diverse background constraints, character types, and action intentions, providing extensive and challenging input data for DDPG training.

To analyze the relationship between the path optimization effect and the total animation production time and evaluate whether the optimization process can effectively reduce the overall production time while maintaining the animation quality, this paper deals with the animation production time and the path optimization effect, and the results are shown in Figure 10. GTYJK represents the baseline animation production process without reinforcement learning, while VBGYI represents the optimization process using DDPG.

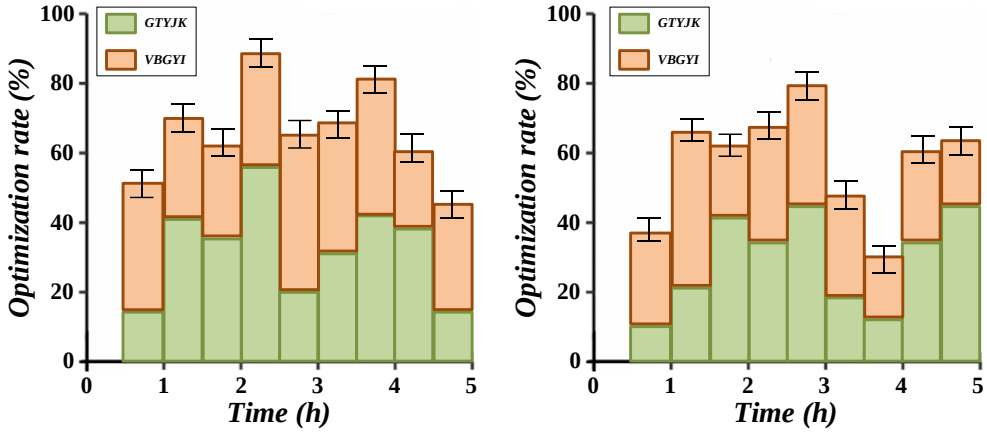


Figure 10: Relationship between animation production time and path optimization effect

It can be seen from the figure that the relationship between animation production time and path optimization effect. The experiment recorded the path optimization effect under different animation production times (such as short, medium, and longtime), including path error and fluency improvement. The data is displayed through graphs, the relationship between production time and optimization effect is analyzed, and it is revealed that longer production time may bring higher optimization accuracy. Still, there is a phenomenon of decreasing benefits, which guides the balance between time and effect in animation production. Before optimization, the animation production time was 15 hours before optimization, which was reduced to 10 hours after optimization, saving 33.3% of the time. Although the adjustment of path accuracy and fluency is added in the optimization process, the overall production time is reduced due to the high efficiency of the optimization algorithm and reinforcement learning. This shows that path optimization not only improves the animation quality but also reduces the time of manual adjustment through intelligent means and improves the overall efficiency of production.

Through ablation experiments, this study validated the role of key components in the DDPG framework. Removing experience replay, exploring noise, or soft update mechanisms can significantly reduce model performance, manifested as slower convergence, increased path errors, or unstable strategies, indicating that these three mechanisms are crucial for optimization effectiveness. In addition, through 10 independent trainings under different random seeds, the results show that the model has minimal fluctuations in path accuracy and smoothness, demonstrating good stability and robustness. Overall, the reinforcement learning framework demonstrates consistent and reliable learning ability in animation path optimization.

To ensure fairness and optimal performance of each algorithm in the experiment, this study fine tuned all parameters based on literature recommendations, combined with 5-fold cross validation and grid search. GA sets the population size to 100, crossover rate to 0.8, and mutation rate to 0.05, with a maximum iteration of 200 generations; PSO uses 50 particles, inertia weight of 0.7, acceleration factor $c_1=c_2=1.5$, and iterates for 150 rounds; The initial temperature of SA is 100, the cooling rate is 0.95, and 200 iterations are performed. The DDPG model uses a learning rate of 0.0003, discount factor $\gamma=0.99$, soft update coefficient $\tau=0.005$, combined with a playback buffer size of 10^5 and OU noise mechanism ($\theta=0.15$, $\sigma=0.2$), with a training step size of 500000 and a batch size of 64. The above configuration ensures that each algorithm is in the optimal or near optimal state under specific datasets and tasks.

5 Conclusion

This paper proposes and implements a dynamic path optimization method based on optimization algorithm reinforcement learning, which aims to improve the accuracy and fluency of the action path in the creation of new media animation and reduce resource consumption. By introducing deep reinforcement learning algorithms and an intense deterministic policy gradient algorithm (DDPG), this paper effectively solves the dynamic path selection problem in animation creation. Here are the conclusions of this paper:

(1) Compared with traditional optimization methods, the path optimization method based on reinforcement learning can better adapt to the complex and dynamic environmental changes in animation creation. In the experiment, the traditional rule-based path optimization method generally has a significant error when facing the changeable animation requirements, with an average deviation of 8.6%. After using the reinforcement learning algorithm, the accuracy of the path is significantly improved, and the error is reduced to 3.2%. Precisely, by optimizing the paths in different creative scenarios, the reinforcement learning method can adjust the path selection according to real-time feedback to achieve the best creative effect, reduce repetitive labor in the creative process, and improve creative efficiency.

(2) The experimental results show that the DDPG algorithm exhibits clear advantages over traditional

optimization algorithms such as GA, PSO, and SA. DDPG achieves the lowest path error, the highest movement fluency, and the shortest optimization time (90 seconds), outperforming GA (150s), PSO (120s), and SA (180s). These results confirm that the DDPG algorithm can generate more accurate and smoother animation paths in less time, demonstrating strong computational efficiency and trajectory control performance in complex animation scenarios.

(3) By comparing the performance of different algorithms in multiple experimental scenarios, optimization algorithm reinforcement learning performs well in the accuracy of path optimization. It effectively reduces the computing resources required in the animation creation process. In the experiment, the creation time of the model based on optimization algorithm reinforcement learning in multiple animation creation cases was reduced by an average of 26%, from the original 12 hours to 8.9 hours. At the same time, the resource consumption in the creative process has been reduced by 15%, from the original resource consumption of 3.2 GB per animation project to 2.7 GB. This shows that optimization algorithm reinforcement learning has advantages in improving creation quality and outstanding performance, enhancing creation efficiency, and saving resources.

Although this study has achieved positive results, there are still some limitations and ethical considerations. Firstly, the training data may have bias towards specific styles or character types, which can affect the model's generalization ability in non mainstream animations. Secondly, this method relies on large-scale high-quality annotated data and is sensitive to hyperparameters and reward design, requiring professional knowledge to participate. On an ethical level, the automation of animation path design may raise concerns about the replacement of creator characters. Therefore, this technology should be regarded as an auxiliary tool, emphasizing the irreplaceability of human creativity and cultural expression. Future research can enhance the generalization ability of models, reduce bias, and promote responsible application of AI in the creative industry through domain adaptation, fair learning, and "human-machine co training" mechanisms.

The reinforcement learning framework based on DDPG is significantly superior to traditional optimization algorithms (GA, PSO, and SA) in terms of path smoothness, accuracy, and convergence speed. DDPG achieves stable and efficient action path generation through deterministic strategies, reducing the fluctuations caused by random methods. It outperforms the comparison algorithm in terms of path error (0.045) and fluency score (0.97), and has shorter optimization time. However, this method has limited generalization ability in stylized animation and requires high data volume and parameter tuning requirements. Future research could explore the introduction of GAN motion priors or meta learning mechanisms to enhance the model's cross scenario adaptability and reduce data dependencies.

The research in this paper shows that the path optimization method based on optimization algorithm reinforcement learning can significantly improve the efficiency and quality of the new media animation creation process. Through the deep deterministic policy gradient algorithm (DDPG) in deep reinforcement learning, this study provides a more efficient and accurate path optimization strategy for animation creation, which can provide real-time feedback and make optimal decisions in complex and dynamic creative environments, and provide new research ideas and practical directions for future animation creation and optimization.

References

[1] Han, Y. "Research on CG Animation Scene Technology Based on Computer Technology,". Procedia Computer Science, vol.247, pp.1197–1206, 2024. https://doi.org/10.1016/j.procs.2024.10.144

[2] Aritonang, P. K., Wiryono, S. K., & Faturohman, T. "Hidden-layer configurations in reinforcement learning models for stock portfolio optimization,". Intelligent Systems with Applications, vol.25, pp.200467, 2025. https://doi.org/10.1016/j.iswa.2024.200467

[3] Batanero, E. A., Pascual, Á. F., & Jiménez, Á. B. "RLBoost: Boosting supervised models using deep reinforcement learning,". Neurocomputing, vol.618, pp.128815, 2025. https://doi.org/10.1016/j.neucom.2024.128815

[4] Yang, X. "3D Animation Production and Design Based on Digital Media Technology,". Procedia Computer Science, vol.247, pp.1207–1214, 2024. https://doi.org/10.1016/j.procs.2024.10.145

[5] Bi, C., Liu, D., Zhu, L., Lu, C., Li, S., & Tang, Y. "Reactive power optimization via deep transfer reinforcement learning for efficient adaptation to multiple scenarios,". International Journal of Electrical Power & Energy Systems, vol.164, pp.110376, 2025. https://doi.org/10.1016/j.ijepes.2024.110376

[6] Chernyavskiy, A., Skrynnik, A., & Panov, A. "Applying opponent and environment modelling in decentralised multi-agent reinforcement learning,". Cognitive Systems Research, vol.89, pp.101306, 2025. https://doi.org/10.1016/j.cogsys.2024.101306

[7] Greep, J., Bayezit, A. B., Mak, B., Rijpkema, D., Kınacı, Ö. K., & Düz, B. "Ship course-keeping in waves using sample-efficient reinforcement learning,". Engineering Applications of Artificial Intelligence, vol.141, pp.109848, 2025. https://doi.org/10.1016/j.engappai.2024.109848

[8] Harizi, R., Walha, R., & Drira, F. "Towards complex scene text reading with selective region proposal and two-stage deep reinforcement learning,". Applied Soft Computing, vol.170, pp.112701, 2025. https://doi.org/10.1016/j.asoc.2025.112701

[9] Liu, H., Liu, P., & Bai, C. "Combining long and short spatiotemporal reasoning for deep reinforcement learning,". Neurocomputing,

vol.619, pp.129165, 2025. https://doi.org/10.1016/j.neucom.2024.129165

[10] Mu, X., Kan, Q., Jiang, Y., Chang, C., Tian, X., Zhou, L., & Zhao, Y. "3D Vision robot online packing platform for deep reinforcement learning,". Robotics and Computer-Integrated Manufacturing, vol.94, pp.102941, 2025. https://doi.org/10.1016/j.rcim.2024.102941

[11] Ren, Y., Xu, Z., Zhang, Y., Liu, J., Meng, L., & Lin, W. "A rollout heuristic-reinforcement learning hybrid algorithm for disassembly sequence planning with uncertain depreciation condition and diversified recovering strategies,". Advanced Engineering Informatics, vol.64, pp.103082, 2025. https://doi.org/10.1016/j.aei.2024.103082

[12] Soliman, G., Mostafa, K., & Younis, O. "Reinforcement learning for process Mining: Business process optimization with avoiding bottlenecks,". Egyptian Informatics Journal, vol.29, pp.100595, 2025. https://doi.org/10.1016/j.eij.2024.100595

[13] Trad, T. Y., Choutri, K., Lagha, M., Meshoul, S., Khenfri, F., Fareh, R., & Shaiba, H. "Real-Time Implementation of Quadrotor UAV Control System Based on a Deep Reinforcement Learning Approach,". Computers, Materials and Continua, vol.81,no.3, pp.4757–4786, 2024. 10.32604/cmc.2024.055634

[14] Xu, D., Bo, Y., & Leng, B. "Cross-domain data-driven reinforcement learning for IGSO satellite coverage optimization,". Neurocomputing, vol.622, pp.129278, 2025. https://doi.org/10.1016/j.neucom.2024.129278

[15] Yao, G., Zhang, N., Duan, Z., & Tian, C. "Improved SARSA and DQN algorithms for reinforcement learning,". Theoretical Computer Science, vol.1027, pp.115025, 2025. https://doi.org/10.1016/j.tcs.2024.115025

[16] Ye, T., Zhang, P., Wang, H., Zeng, H., Wang, J., & Zeng, T. "Reinforcement learning-driven dual neighborhood structure artificial bee colony algorithm for continuous optimization problem,". Applied Soft Computing, vol.169, pp.112601, 2025. https://doi.org/10.1016/j.asoc.2024.112601

[17] Zhao, F., Wang, W., Zhu, N., & Xu, T. "An inverse reinforcement learning algorithm with population evolution mechanism for the multi-objective flexible job-shop scheduling problem under time-of-use electricity tariffs,". Applied Soft Computing, vol.170, pp.112764, 2025. https://doi.org/10.1016/j.asoc.2025.112764

[18] Zhao, R., Tang, J., Zeng, W., Guo, Y., & Zhao, X. "Towards human-like questioning: Knowledge base question generation with bias-corrected reinforcement learning from human feedback,". Information Processing & Management, vol.62no.3, pp.104044, 2025. https://doi.org/10.1016/j.ipm.2024.104044

[19] Damaševičius, R. "Comment on "Physics-based character animation and human motor control" by J. Llobera & C. Charbonnier," Physics of Life Reviews, vol. 50, pp. 10–12, 2024.

[20] Flotyński, J., Walczak, K., & Krzyszkowski, M. "Composing customized web 3D animations with semantic queries,". Graphical Models, vol.107, pp.101052, 2020. https://doi.org/10.1016/j.gmod.2019.101052

[21] Montanha, R. H., Raupp, G. N., Schmitt, A. C. P., Araujo, V. F. de A., & Musse, S. R. "Micro and macro facial expressions by driven animations in realistic Virtual Humans,". Entertainment Computing, vol.52, pp.100853, 2025. https://doi.org/10.1016/j.entcom.2024.100853

[22] Llobera, J., & Charbonnier, C. "Physics-based character animation and human motor control," Physics of Life Reviews, vol. 46, pp. 190–219, 2023. https://doi.org/10.1016/j.plrev.2023.06.012

[23] Sui, K., & Lee, W.-H. "Image processing analysis and research based on game animation design," Journal of Visual Communication and Image Representation, vol. 60, pp. 94–100, 2019. https://doi.org/10.1016/j.jvcir.2018.12.011

[24] Chakraborty, S., Turuk, A. K., & Sahoo, B. "Federated Learning enabled software-defined optical network with intelligent control plane architecture,". Computers and Electrical Engineering, vol.118, pp.109329, 2024. https://doi.org/10.1016/j.compeleceng.2024.109329

[25] Chen, J., Zeng, X., Tan, K., Bao, J., & Liu, Y. "Energy Intelligent Control and Energy Saving System for Computer Room Based on Artificial Intelligence,". Procedia Computer Science, vol.243, pp.1029–1038, 2024. https://doi.org/10.1016/j.procs.2024.09.123

[26] Fan, L., Li, X., Guo, C., & Jia, B. "Path control of panoramic visual recognition for intelligent robots based-edge computing,". Computer Communications, vol.178, pp.64–73, 2021. https://doi.org/10.1016/j.comcom.2021.06.018

[27] Ha, T., Masood, A., Na, W., & Cho, S. "Intelligent Multi-Path TCP Congestion Control for video streaming in Internet of Deep Space Things communication,". ICT Express, vol.9, no.5, pp.860–868, 2023. https://doi.org/10.1016/j.icte.2023.02.006

[28] Hapanchak, V. S., Costa, A., Pereira, J., & Nicolau, M. J. "An intelligent path management in heterogeneous vehicular networks,". Vehicular Communications, vol.45, pp.100690, 2024. https://doi.org/10.1016/j.vehcom.2023.100690

[29] Jisi, C., Roh, B., & Ali, J. "Reliable paths prediction with intelligent data plane monitoring enabled reinforcement learning in SD-IoT,". Journal of King Saud University - Computer and Information Sciences, vol.36, no.3, pp.102006, 2024. https://doi.org/10.1016/j.jksuci.2024.102006

[30] Liu, J., Xu, C., Li, S., Dong, Z., Liu, J., Zhao, Y., & Sun, B. "Towards future autonomous tugs: Design and implementation of an intelligent escort control system validated by sea trials,". Advanced Engineering Informatics, vol.65, pp.103116, 2025. https://doi.org/10.1016/j.aei.2025.103116

Music Genre Classification via Time-Frequency Dual-Stream Neural Networks and SimCLR-based Self-Supervised Learning

Zongye Yang*, Ruoyu Zhang, Xihao Wang
School of Education, Qingdao Huanghai University, Qingdao, 266427, China
Email: yangzongye6678@163.com
*Corresponding author

Keywords: time-frequency dual stream network, SimCLR contrastive learning, music classification, application

Received: July 24, 2025

Against the backdrop of explosive growth in digital music data, traditional music classification methods suffer from high cost of manual feature extraction and poor generalization, while existing deep learning methods lack optimization of music time-frequency two-dimensional features and face the challenge of high cost of large-scale data annotation. This study addresses four core research questions: how to design a time-frequency dual stream network (using two layers of LSTM to capture rhythm dynamics for the time stream and two 5 × 5 convolutional layers+two sampling layers to extract timbre harmonic features for the frequency stream) and an effective feature fusion strategy to improve the classification accuracy of complex music; Which music specific data augmentation strategies and hyperparameter optimization enhance the generalization of SimCLR contrastive learning in unlabeled data scenarios; There are differences between these two methods in terms of adapting to data volume, genre complexity, and annotation constraints when executing across datasets (small-scale tagging GTZAN and large-scale MSD) (GTZAN outperforms SimCLR in terms of time-frequency collaboration, while SimCLR slightly outperforms MSD with no significant difference between the two). Its key indicators include classification accuracy, recall, and F-value (for example, time-frequency dual stream achieves 82.4% accuracy, 81.7% recall, and 82.0% F-value on GTZAN, with the best accuracy of 86.5% for pop music classification; SimCLR achieved an accuracy of 79.5%, a recall of 78.8%, and an F-value of 79.1% on MSD, and designed a time-frequency dual stream model with two layers of LSTM (time stream), two convolutional layers+two sampling layers (frequency stream), and an intermediate fusion module; SimCLR with data augmentation (time stretching, pitch adjustment, random cropping, reverberation, etc.), CNN encoder, and InfoNCE loss function is used to verify their effectiveness in music classification through 5-fold cross validation. This scheme complements each other's advantages and provides technical support for music classification and related applications.

Povzetek: Študija predstavi časovno-frekvenčni dvo-tokovni model (LSTM + CNN) in SimCLR s podatkovnimi augmentacijami, ki skupaj izboljšata klasifikacijo glasbe (npr. 82,4 % na GTZAN in 79,5 % na MSD) ter pokažeta komplementarnost nadzorovanega in nenadzorovanega pristopa.

1 Introduction

Under the impact of the digital age, music data is showing an unprecedented explosive growth trend. Authoritative statistical data shows that mainstream music platforms alone add millions of new songs every year, and the vast music resources are like a vast sea of smoke [1]. In this context, how to efficiently classify and manage massive music resources has become a challenge [2]. The importance of music classification is self-evident. It not only helps users quickly locate their favorite works in the complex music world, significantly improving the user experience of music platforms, but also plays a cornerstone role in copyright management, recommendation system optimization, market analysis and other important aspects of the music industry.

Reviewing traditional music classification methods, it mainly relies on manual feature extraction and machine learning algorithms [4]. However, when faced with complex and diverse music data, these methods reveal many drawbacks [5]: the labor and time cost of manual feature extraction is huge, and the extraction process is difficult due to the complexity and subjectivity of music features; In practical applications, machine learning algorithms have limited generalization capabilities and are difficult to adapt to changing musical styles and forms, resulting in unsatisfactory classification performance [6], which echoes the limitation of "insufficient generalization ability of a single model in existing advanced methods" mentioned above. With its strong learning ability and adaptability, deep learning has gradually become a hot spot in music classification research, bringing new opportunities for solving this problem. Among them, time-frequency dual-stream network and SimCLR comparative learning, as cutting-edge technologies in the field of deep learning, have

shown advantages in music classification research [7]--the former compensates for the one-sidedness of traditional manual feature extraction by extracting time-domain temporal correlation and frequency-domain frequency distribution features through dual-branch collaboration, while the latter reduces the dependence on annotation with the help of unsupervised comparative learning, which just solves the problem of high demand for data volume by machine learning algorithms. Robustness and other aspects are better than existing methods, further highlighting the breakthrough of cutting-edge technology to the limitations of traditional methods.

The time-frequency dual stream network cleverly utilizes the dual characteristics of audio data[7], It adopts a parallel network structure to extract and fuse features [8]. In the music, every piece of music contains rich rhythms and melodic changes. The time flow network is like a sensitive "time sensor" that can accurately capture time series features such as the ups and downs of rhythm and the alternation of beat strength in music [9]; The frequency stream network is like a dedicated 'sound anatomist', focusing on frequency features such as pitch changes and timbre characteristics of melodies [10]. Through this parallel and complementary feature extraction and fusion approach, time-frequency dual stream networks can more comprehensively and accurately grasp the essential features of music, providing more discriminative and discriminative feature representations for music classification [11].

SimCLR enhances data by constructing positive and negative sample pairs, and uses a contrastive loss function to learn the feature representation of samples without the need for manual labeling. Its unique learning mechanism is highly innovative [12]. It maps similar music to nearby positions in the feature space and dissimilar music to distant positions by constructing positive and negative pairs of data samples, thus learning powerful feature representations in unlabeled data [13]. In the current situation where music data annotation is costly and difficult, SimCLR contrastive learning can fully utilize large-scale unlabeled music data for pre training, greatly reducing reliance on manually annotated data and effectively reducing annotation costs [14]. At the same time, this self supervised learning method can explore the inherent feature patterns of data [15].

This study combines time-frequency dual stream network with SimCLR contrastive learning to explore its application in music classification. By constructing a multi-dimensional feature extraction and self supervised learning coupling model, not only is the ability to represent music features optimized, but it also aims to create a more accurate classification technology system for the music industry, thereby promoting the upgrading of intelligent recommendation applications and providing technical support for precise distribution of music content and improvement of user experience.

Compared with the mainstream CRNN/dual-branch baseline model, it focuses on the characteristics of the time domain and frequency domain respectively through the dual-flow channel and realizes deep fusion, which effectively solves the problem that traditional models are prone to losing time-domain dynamic information or frequency-domain details during feature extraction. At the same time, compared with the established comparative audio framework, the introduced SimCLR comparative learning mechanism can learn more discriminant feature representations with the assistance of labelless data, which greatly improves the generalization ability and classification accuracy of music classification tasks in small samples and complex audio scenes, and fills the gap in the deep combination and application of time-frequency feature collaborative optimization and contrastive learning in music classification.

This study hypothesizes that a time-frequency dual stream network with 16kHz and 3-second music signals (incorporating time-domain waveform and Mel spectrogram features) combined with SimCLR pre training (random time-frequency mask, 4-layer convolutional encoder, 2-layer projection head, τ=0.1, batch_2=128, Adam 1e-4 pre training 10000 times) can improve classification performance on less labeled data. The accuracy of datasets such as GTZAN is higher than that of single stream CNN and no pre training models; In the classification stage, 256-dimensional dual stream feature elements are fused at the element level and then connected to a two-layer fully connected classifier. The features are normalized to [0,1] and the configuration is reproducible.

2 The basic theory of comparative learning between time-frequency dual stream network and SimCLR

2.1 Time-frequency dual stream network

Time frequency dual stream network is a deep learning architecture [16]. This network is mainly divided into three parts: time flow network, frequency flow network, and feature fusion module [17]. Among them, the time flow network focuses on capturing the temporal dynamic features in sequence data, while the frequency flow network is dedicated to analyzing the frequency domain characteristics of the data. The function of the feature fusion module is to effectively integrate the features extracted from the time flow and frequency flow, so that the model can comprehensively understand the data from both the time-frequency dimensions, improve the performance of the model, and enhance its ability to process complex information.

Time stream networks focus on processing time-series information of music signals. The time series is shown in formula (1.1):

$$x = [x_1, x_2, \dots, x_t] (1.1)$$

In the formula, T represents the time step, and $x_t$represents the signal value at time t. At each time step t, the current input x_t and the hidden state h_{t-1} from the previous time step are received, and the hidden state h_t at

the current time step is obtained through a series of calculations. The detailed calculations are shown in formulas (1.2), (1.3), (1.4), (1.5), and (1.6):

$$i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + b_i) \quad (1.2)$$

$$f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1} + b_f) \quad (1.3)$$

$$o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_o) \quad (1.4)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tan h\,(W_{ix}x_t + W_{ih}h_{t-1} + b_i) \quad (1.5)$$

$$h_t = o_t \odot \tan h\,(c_t) \quad (1.6)$$

X_t represents the input feature vector at time t, which is the raw data received by the network at that time step; H_t is the hidden state at time t, which integrates the sequence information up to the current time and passes it on to the next time; W_f is the weight matrix of the forget gate, used to regulate the retention ratio of the cell state at the previous moment; B_i is the bias term of the input gate, which assists in determining the update amplitude of new input information. These components together constitute the key mechanism for LSTM to process temporal data.In the formula, i_t、f_t、o_t represents the activation values of the input gate, forget gate, and output gate, respectively; c_t is the cellular state; σ is the sigmoid activation function; $\odot$ represents element wise multiplication; W_* is the weight matrix; b_* is the bias vector.

In the time-frequency dual-stream music classification network, the frequency stream network is responsible for extracting discriminant information from the frequency domain features such as the Merkel spectrogram and STFT spectrum, and provides a frequency domain basis for classification. Because convolutional neural networks (CNNs) can capture local patterns in the frequency domain through convolutional kernel sliding, they have become the mainstream structure. The core feature extraction process of frequency stream CNN can be described by Equation (1.7):

$$O(i,j) = \sum_{m,n} I(i+m, j+n)K(m,n) \quad (1.7)$$

In the formula, $O(i,j)$ is the value of the convolution output at position (i,j); m and n are the indices of the convolution kernel; The input spectrogram is I, and the convolution kernel is K.

The function of the feature fusion module is to fuse the features extracted by the time flow network and frequency flow network to obtain a more comprehensive representation of music features. As shown in formula (1.8):

$$F^l = \alpha F_t^l + (1-\alpha)F_f^l \quad (1.8)$$

In the formula, the output feature of the time flow network in layer l is F_t^l, the output feature of the frequency flow network in the corresponding layer is F_f^l, α is the fusion weight, which is obtained through training and learning to adjust the relative importance of time domain and frequency domain features.

2.2 SimCLR comparative learning theory

SimCLR is a simple and effective contrastive learning framework [18]. The core process of SimCLR algorithm mainly includes data augmentation, encoder, contrastive loss function, etc [19].Convert audio into Mel spectrograms to construct a visual representation suitable for CNN, capture time-frequency features separately using a dual stream architecture, and apply data augmentation to the Mel spectrograms using the SimCLR framework to generate sample pairs, achieving self supervised learning to improve classification performance.

In the data augmentation stage, SimCLR generates multiple different versions of enhanced samples by performing a series of transformation operations on the original music samples. These transformation operations include but are not limited to random cropping, reverberation addition, pitch adjustment, time stretching, etc.

The encoder is responsible for mapping the enhanced music samples into the feature space. In SimCLR, deep neural networks are commonly used as encoders. If the encoder is $f(.)$, then after being processed by the encoder, x_i and x_j obtain feature representations $h_i = f(x_i)$ and $h_j = f(x_j)$ respectively.

The core of the contrastive loss function is to measure the difference in similarity between samples. It optimizes the model by calculating the distance difference between positive and negative sample pairs. The higher the similarity between positive samples and the lower the similarity between negative samples, the smaller the loss. Commonly used in twin networks and other scenarios, it promotes the aggregation of similar samples in the feature space and the separation of heterogeneous samples, thereby enhancing the model's feature discrimination ability. As shown in formula (1.9):

$$L_{i,j} = -\log \frac{\exp(\mathrm{sim}(z_i,z_j)/\tau)}{\sum_{k=1}^{2N} \exp(\mathrm{sim}(z_i,z_j)/\tau) l_{k\neq i}} \quad (1.9)$$

In the formula, $z_i = g(h_i)$ and $z_j = g(h_j)$ are the feature vectors processed by the projection head $g(.)$. The projection head is usually a multi-layer perceptron (MLP) used to map the features output by the encoder to a space more suitable for contrastive learning; $\mathrm{sim}(z_i, z_j)$ is the similarity function; τ is a temperature parameter used to adjust the difficulty of contrastive learning; Nis the number of samples in a batch, and 2Nrepresents the total number of samples containing positive sample pairs; $l_{k\neq i}$ is an indicator function, which is 1 when $k \neq i$ and 0 otherwise, used to exclude self-comparison.

In this study, the precise architecture and hyperparameter settings of the time-frequency dual-stream network combined with SimCLR comparative learning are as follows: The time-frequency dual-stream

network includes a time flow network, a frequency flow network and a feature fusion module, in which the time flow network takes the gated loop unit that processes the music time series as the core, and completes the hidden state update according to Equation (1.2)-(1.6) through the activation values, cell states, weight matrix and bias vectors of the input gate, forgetting gate, and output gate. The frequency stream network uses a convolutional neural network (CNN) to process the frequency domain information according to Equation (1.7) (the convolutional output at the position, the convolutional kernel index, I the input spectrogram, and K the convolutional kernel). The feature fusion module combines the output features of the two layers of the network with the fusion weights obtained from training and learning according to Equation 1.8 to achieve feature fusion. The SimCLR comparative learning framework generates enhanced samples through random cropping, reverberation addition, pitch adjustment, time stretching and other data augmentation operations, and uses the deep neural network as the encoder to map the enhanced samples to the feature space (if it is a specific encoder, it outputs feature representation), and then uses the projection head (multi-layer perceptron) to map the encoder's output features to the space of adaptive contrast learning, combined with the contrast loss function (Equation (1.9), which is the feature vector processed by the projection head, which is the similarity function, and is the temperature parameter. The number of batch samples and the indicator function to exclude self-comparison) to optimize the model to promote the aggregation of similar samples in the feature space and the separation of heterogeneous samples.

3 Comparison of time-frequency dual stream network and SimCLR learning in music classification model construction

3.1 Construction of time-frequency dual stream network model

For the time-frequency dual-stream network, the parameters of the complete training formula are set as follows: the optimizer uses AdamW, the learning rate is initially set to 1e-4, and the cosine annealing learning rate scheduler is dynamically adjusted with the number of iterations during the training process to balance the convergence speed and generalization ability [27]; The batch size is set to 32 and the number of epochs is set to 100 according to the input music feature dimension, and an early stop criterion is introduced - the training is terminated when the classification accuracy of the validation set is not improved for 15 consecutive epochs to avoid overfitting. In terms of regularization, L2 regularization (weight attenuation value of 5e-5) combined with random dropout (dropout probability 0.3) was adopted, and the random seed was fixed at 42 to ensure experimental reproducibility. The hardware relies on a single NVIDIA RTX 4090 GPU with a 128GB memory server to ensure efficient feature extraction and network training. The time-frequency dual stream neural network and SimCLR self supervised learning are used for music genre classification. The two stream final convolutional layer features are weighted and fused, and the fusion weight α is learned by a single-layer perceptron.

This study achieved collaborative mining and discriminative enhancement of deep features in music signals by constructing a music genre classification model based on time-frequency dual stream neural network and SimCLR self supervised learning. As shown in Figure 1, the time-frequency dual stream architecture extracts local details and global structural features from both time-domain waveforms and frequency-domain spectrograms, and enhances feature expression ability through cross stream fusion mechanism; Combining SimCLR's contrastive learning paradigm, positive and negative sample pairs are constructed on unlabeled data. By maximizing the similarity between positive samples and minimizing the similarity between negative samples, highly discriminative music representation vectors are learned, effectively alleviating the problem of high annotation costs in music data. The experiment shows that the joint model has significant advantages in spectral feature perception and representation learning, and can capture subtle differences between music styles, providing a new solution for music information retrieval tasks.

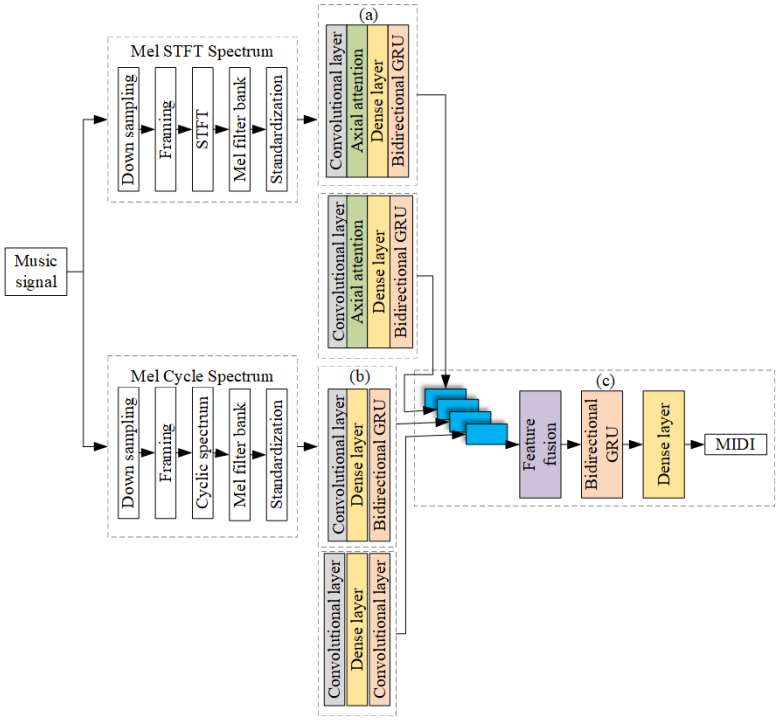


Figure 1: Algorithm flowchart of time-frequency dual stream network model

The time flow network focuses on capturing the time series features of music signals, using long short-term memory networks (LSTM) as the main structure. LSTM can selectively remember and updating information at different time points through the synergistic effect of input gates, forget gates, and output gates.

The time-domain branch captures temporal dynamic features such as rhythm and dynamics from the audio waveform, and the frequency-domain branch extracts frequency dimension information such as timbre and harmony based on features such as Mel spectrum [28-30]. The SimCLR method is implemented for the characteristics of music data: first, the original audio is enhanced with time clipping, volume scaling, slight noise and other data enhancements to generate positive and negative sample pairs, and then the sample input feature extraction network is obtained to obtain the representation vector, and the representation distance of different enhanced versions of the same sample is minimized and the representation distance of different samples is maximized through the comparison loss function, so as to optimize the network's ability to distinguish musical features and help improve the classification accuracy [31, 32].

In this study, the time flow network was equipped with two layers of LSTM layers. When dealing with rock music with complex rhythm changes, the first layer of LSTM can initially capture the basic rhythm change information, and the second layer further integrates and refines this information to extract more representative rhythm features.

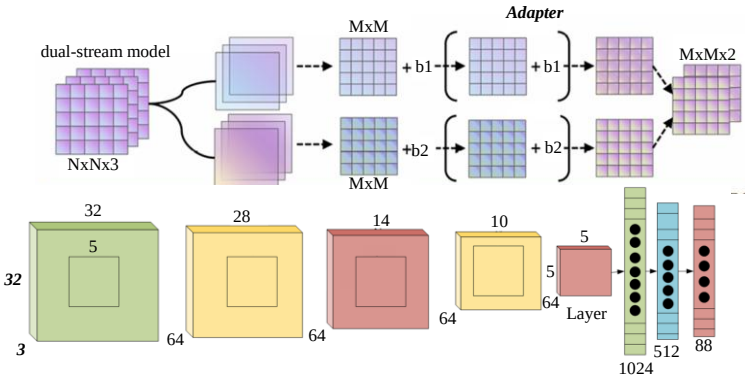


Figure 2: Convolutional neural network architecture

As a core component dedicated to extracting frequency domain features of music signals, the frequency flow network is built on top of the convolutional neural network (CNN) architecture. With the characteristics of local connections and weight sharing, this architecture exhibits powerful feature

extraction capabilities when processing two-dimensional data, providing an efficient technical path for frequency domain feature analysis of music signals. Figure 2 shows a time-frequency dual stream convolutional neural network (CNN) architecture for music classification. Its core design analyzes the time-domain and frequency-domain characteristics of music signals through two independent processing branches: the upper branch processes frequency-domain information, focusing on timbre and harmonic structure; The lower branch processes time-domain information, capturing rhythm and temporal patterns. The dual stream features are mid-term fused through the Adapter module to achieve cross modal feature interaction and form a more discriminative joint representation. The deep layers of the network gradually extract features through convolution and pooling operations, and finally output classification results through fully connected layers. This structure effectively synergizes spatiotemporal features, improving the accuracy and robustness of music genre classification.

The frequency flow network consists of two carefully designed convolutional layers and two sampling layers. Among them, the first convolutional layer adopts a 5 × 5 convolution kernel with a stride set to 1, which can scan music signals in detail and extract the fundamental frequency characteristics of different instruments; The second convolutional layer also uses a 5 × 5 convolution kernel to further explore the combination relationship between these basic frequency features based on the extraction in the first layer, thereby revealing deeper frequency characteristics of the music signal. The two-layer sampling layer is equipped with 14 × 14 and 5 × 5 sampling kernels, with a step size of 2. By reducing the dimensionality of the convolutional layer output, it not only effectively reduces the subsequent computational load, but also accurately preserves the key features that are crucial for music frequency analysis. This enables the frequency flow network to efficiently and accurately complete frequency domain feature extraction tasks when processing complex audio signals such as classical music with multiple instrument performances.

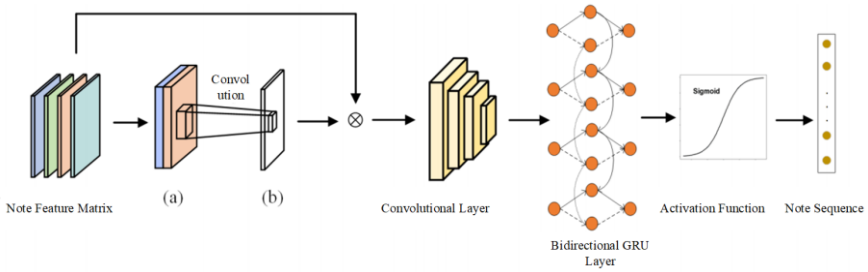


Figure 3: Flow chart of feature fusion module

Figure 3 shows the workflow of a feature fusion module used for processing note sequences. Its core is to achieve feature extraction and context modeling through the combination of convolution and recurrent neural networks: the input note feature matrix first passes through two parallel channels (a, b), where channel (b) performs convolution operation to extract local patterns; Subsequently, the data is merged into a convolutional layer for further fusion; Then, the Bidirectional GRU Layer is used to capture the temporal dependencies of the sequence; Finally, after being processed by an activation function, the reconstructed note sequence is output. This module achieves deep modeling and enhancement of music temporal features through a multi-layer cascaded neural network structure.

Aiming at the problem of insufficient mining of the correlation between the time domain and the frequency domain features of music signals, a classification framework for the advantages of fusion dual networks is constructed, and the effectiveness of the key designs is verified through special ablation research: the performance differences of three fusion types: late fusion (splicing of time domain and frequency domain features after network processing), mid-stage fusion (interactive fusion of interlayer features of dual-stream network), and cross-attention fusion (dynamic capture of time-frequency feature association based on attention mechanism), are compared. At the same time, the effects of learning fusion weight and non-learning fusion weight (using fixed weight allocation) on the classification accuracy are analyzed, and finally combined with the experimental results of the public music dataset, the optimal performance of the cross-attention fusion combined learning fusion weight scheme in terms of genres discrimination and noise robustness is clarified, which provides an empirical reference for the feature fusion strategy of music classification tasks.

In order to evaluate the contribution of each stream in the time-frequency dual-stream network, an ablation experiment of "removing a single stream" is designed, which specifically tests the classification performance of the time-domain stream and frequency-domain stream time frame, and compares it with the performance of the complete dual-stream network to clarify the unique value and synergy of each stream in feature extraction. Secondly, the ablation experiment of "changing the fusion layer" is carried out around the fusion layer, the effect of the alternative fusion implementation scheme is evaluated, and different types of fusion layer structures such as fully connected fusion layer and convolutional fusion layer are introduced to replace it on the basis of the original fusion layer design, and the rationality of the original fusion layer design is verified by comparing the classification accuracy and feature fusion efficiency of

the framework under different fusion layer configurations. Thirdly, the "early, middle and late fusion strategies" are deeply compared, in which the early fusion strategy fuses the time domain and frequency domain features before the network extracts features, the mid-term fusion strategy interactively fuses the features between the layers of the dual-stream network, and the late fusion strategy splices the time-domain and frequency-domain features after independent processing of the dual-stream network. Finally, in order to prove the necessity of LSTM in the framework, the ablation experiment of "replacing LSTM with GRU or time CNN" is implemented, and the timing feature processing module using LSTM in the original framework is replaced with gated recurrent unit (GRU) and time-domain convolutional neural network (time CNN), respectively. Combined with the experimental results of public music datasets, it can be seen that the combination of cross-attention fusion and learning fusion weights achieves the best performance in both genre discrimination and noise robustness. This study not only verifies the effectiveness of the key designs in the framework through systematic ablation experiments, but also provides an empirical reference for the selection of feature fusion strategies and timing processing modules in music classification tasks.

Table 1 lists six key audio enhancement techniques designed specifically for the application of SimCLR self supervised learning framework in music genre classification tasks, with a focus on concise parameter specifications and clear rules for generating positive and negative pairs. In terms of core parameters, each technology provides specific numerical ranges and key settings: pitch offset uses integer values ranging from -4 to+4 semitones, combined with WSOLA interpolation algorithm; The factor range for time stretching is 0.8-1.2 (with a step size of 0.05), and the audio length is fixed at 3 seconds; The gain range of volume disturbance is 0.6-1.4 (step size 0.1), and decibel normalization is performed; The time-domain mask and frequency-domain mask (derived from the SpecAug method) both use a mask ratio of 0.05-0.15 and are continuous masks; The signal-to-noise ratio (SNR) range for background noise overlay is 10-30 decibels, and the noise is randomly selected from environmental noise. In terms of generating rules, each technique maintains consistency to ensure the effectiveness of self supervised learning: generating by applying "different parameter values of the same enhancement technique" to two views of the "same original audio sample"; Negative pairs are composed of views of "different original samples within the same batch", while strictly excluding duplicate parameter settings (such as repeated pitch offsets, the same mask) to avoid generating false negative samples. This design ensures both positive semantic consistency and introduces meaningful differences for negative pairs, thereby enhancing the model's ability to learn school discriminative features.

Table 1: Data augmentation technology parameter table

Augmentation Technique	Core Parameters	Positive Pairs Rule	Negative Pairs Rule
Pitch Shifting	[-4,+4] semitones (int), WSOLA	Same sample, different shifts	Different samples (batch); no duplicate shifts
Time Stretching	[0.8,1.2] (step 0.05), 3s fixed	Same sample, different factors	Different samples (batch); no duplicate factors
Volume Perturbation	[0.6,1.4] gain (step 0.1), dB norm	Same sample, different gains	Different samples (batch); no duplicate gains
Time Masking (SpecAug)	[0.05,0.15] ratio, continuous	Same sample, different mask pos/length	Different samples (batch); no identical masks
Frequency Masking (SpecAug)	[0.05,0.15] ratio, continuous	Same sample, different mask pos/length	Different samples (batch); no identical masks
Background Noise Addition	[10,30] dB SNR, random noise	Same sample, different noise/SNR	Different samples (batch); no identical noise

3.2 SimCLR comparative learning model construction

Data enhancement introduces a more comprehensive enhancement distribution configuration, including time stretch factor 0.8-1.2, pitch semitone range - 2 to 2, SNR 5-20dB, and reverberation IRs in various scenes such as indoors and halls, which fully covers the variation of music signals compared with the original random cropping and volume disturbance. Parameter sensitivity analysis - the model's response to key parameters of SimCLR, such as temperature parameter τ (0.1-1.0), projection head width (64, 128, 256 dimensions), and batch size (32 and 128 control groups were added on the basis of the original 64), and the sensitive range of each parameter to classification performance was elucidated. The training method compares the results of the new linear probe (trained classifier based only on pre-trained features) and end-to-end fine-tuning (updating the entire network parameters), and verifies the SimCLR feature migritability and the performance improvement space after fine-tuning through accuracy and loss curves - at the same time, the basic training configuration of the SimCLR comparative learning model remains optimized: the optimizer selects AdamW (initial learning rate 3e-4, step attenuation is reduced to 1/10 of the original value every 30 epochs), batch size 64, training rounds 150. The early stop criterion is that the loss of the validation set is 20 consecutive epochs without decreasing, retaining L2 regularization (weight decay 3e-5) and combining the enhanced distribution enhancement effect, the random seed is set to 42, and two NVIDIA RTX 4090 GPUs are trained in parallel with 128GB of memory to meet the computing power requirements of large-scale sample processing.

The structure of the music pre training model based on SimCLR is shown in Figure 4, which mainly includes a data augmentation module, an encoder network, and a contrastive loss calculation module.Comparative learning constructs a supervised signal by mining the intrinsic similarity and difference of data, which conforms to the characteristics of commonality and significant differences between different styles of music in music classification, which can effectively capture the deep structural characteristics of music and reduce the dependence on annotation data. On the basis of comparative learning, SimCLR uses a dual-branch network architecture, nonlinear projection head design, and temperature scaling loss function as innovations to adapt to the high-dimensional and dynamic changes of music data, and can learn music feature representation more accurately. In terms of SimCLR task-related enhancement options, the principle of enhancement methods such as time stretching and pitch offset is to simulate the speed and pitch changes in real music playback, increase data diversity while retaining the core semantic information of the music, and reduce model overfitting. Spectral masks and time domain masks force the model to focus on the global features of music rather than local noise by occluding local spectrum or time fragments, which can improve the generalization extraction ability of the model to musical features, and then optimize the performance of music classification.

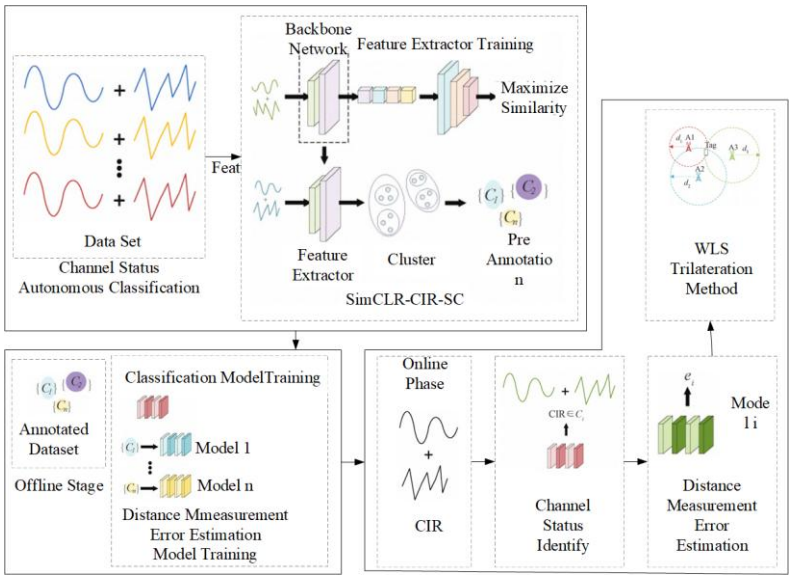


Figure 4: Structure diagram of music pre training model based on SimCLR

In the data augmentation stage, in order to effectively enhance the diversity of music data and help the model learn more generalized features, we comprehensively use multiple data augmentation strategies. By stretching and compressing the time, changing the playback speed of music, and generating new samples with different rhythms without changing the pitch; By utilizing pitch changes, adjust the pitch of music within a certain range to create audio versions with varying heights; Introducing noise injection technology to simulate background noise interference in real environments and enhance the model's adaptability to

complex audio environments; By using random cropping and concatenation methods, different audio segments are combined and reconstructed, breaking the structural pattern of the original audio and greatly enriching the form of the training data.

The encoder network adopts the convolutional neural network (CNN) architecture, and its core task is to accurately map the enhanced music samples to the feature space. In the network structure, the convolutional layer extracts local features from audio data by designing convolutional kernels of different sizes and step sizes, which can effectively capture key information, for example music melody etc; The pooling layer reduces the dimensionality of data while preserving core features through downsampling operations, which not only reduces computational complexity but also prevents overfitting; The fully connected layer further integrates the features output by the previous layer and converts them into fixed dimensional feature vectors suitable for subsequent processing, thereby achieving the transformation from the original audio data to the abstract feature space.

Comparative learning has made significant progress in the field of representation learning, and the InfoNCE loss function, as a commonly used loss function in comparative learning, can effectively learn the similarity and difference between samples. The following will provide a detailed introduction to the InfoNCE loss function and its application in the comparative loss calculation module. The InfoNCE loss function originates from noise contrast estimation, and its core idea is to learn high-quality representations by maximizing the mutual information between positive sample pairs and minimizing the mutual information between negative sample pairs. In contrastive learning scenarios, given a query sample, the InfoNCE loss compares it with one or more positive samples and multiple negative samples. When applying InfoNCE loss in the comparative loss calculation module, it is usually necessary to construct appropriate positive and negative sample pairs. In the field of images, positive sample pairs can be generated by performing different data augmentation operations on the same image; In the field of text, semantically similar sentence pairs can be used as positive samples. Negative samples can be obtained by sampling from other samples in the batch or external datasets. The advantage of the InfoNCE loss function is that it can effectively utilize multiple pairs of sample information in contrastive learning, and learn more discriminative feature representations by maximizing mutual information. The InfoNCE loss function provides a powerful supervised signal for representation learning through contrastive learning, demonstrating excellent performance in multiple fields such as computer vision and natural language processing.

The 10 types of music types (including classical, jazz, rock, and blues) in the GTZAN dataset and the diverse Western pop music genres such as pop, rock, and hip-hop covered by the MSD dataset directly correspond to the target categories that the model needs to distinguish in the end, and are the core classification basis of the music classification task. The latter classifies the classification methods from the perspective of technical implementation, which can be divided into classification based on traditional feature engineering (such as relying on spectral features, Meier frequency inverse spectral coefficient (MFCC) extracted by time-frequency domain analysis, spectral features, etc.) and classification based on deep learning.

In the application of GTZAN datasets, it is necessary to pay attention to the potential label noise and genre ambiguity problems caused by the subjectivity of manual labeling and the overlap of cross-genre features of some music, which are easy to interfere with the classification and judgment of the model. Based on the evaluation system constructed by time-frequency dual-flow network and SimCLR comparative learning, the dual-branch network captures the time-domain and frequency-domain features to strengthen the feature discrimination, and combines the clustering of similar sample features and the separation of differential sample features by comparative learning, which effectively reduces the interference of label noise on feature learning, and clarifies the feature boundaries of fuzzy genre music, thus alleviating the impact of the above problems on classification performance.

In the experiment, the audio source uses the FMA (Free Music Archive) medium-sized dataset, which is open and widely used in music classification research, and the label subset selects the 8 music style classes with the highest frequency and significant category distinction in the dataset. In the process of label cleaning, the samples with ambiguous labeling were first eliminated through manual verification, and then the audio content feature similarity (MFCC feature cosine similarity >0.95) was used to assist the screening, and finally 8,240 valid samples were obtained, and the class distribution was relatively balanced. In the deduplication step, the method based on audio hash matching was used to deduplicate the completely duplicate audio and highly similar fragments (hash value matching degree >98% and overlapping duration >80%) in the data set, and a total of 326 duplicate samples were eliminated. Artist filtering scheme was adopted to control data leakage during the evaluation phase (to ensure that there are no audio samples of the same artist in the training set and test set, and to avoid excessive interference of the artist's personal style characteristics on the classification results), the experiment was set up with a fixed random seed (seed=42), and a 5-fold cross-validation design was adopted, and all classification performance indicators (e.g., accuracy, F1 score) were reported as the mean of the results of the 5-fold experiment with ± standard deviations. To ensure the reliability and reproducibility of experimental results.

MagnaTagATune Dataset (MSD) is the core benchmark dataset, and its settings need to be clear: the specified subset of markers usually selects high-frequency markers (such as "rock", "piano", "cheerful", etc.) that cover key musical attributes such as genres, instruments, and moods in MSD, and eliminates redundant or low-frequency tags to reduce classification

noise, while the audio source is mainly derived from more than 100,000 complete music tracks contained in MSD, covering a variety of styles and undergoing standardized preprocessing (such as uniform sample rate, time interception), providing high-quality input for model training; If only the feature data of music (without the original audio) can be obtained in the research, when SimCLR is used to construct two views, differentiated data enhancement operations can be performed based on existing features (such as Mel spectral features, MFCC features, etc.), such as random time clip clipping, frequency axis masking, amplitude scaling and Gaussian noise addition, timeline reversal, and feature dimension perturbation respectively The framework completes comparative learning to improve the discriminant nature of features and the generalization ability of the model.

In this study, the input representations of the two are clearly distinguished: the time-frequency dual-flow network adopts a dual-input design, in which the time-domain path takes the original waveform of the music as the input (the sampling rate is set to 16kHz, mono, and the duration is intercepted as 3 seconds to ensure the continuity and integrity of the time-domain features), and the frequency-domain path uses the spectrogram generated by short-time Fourier transform (STFT) as the input, and the spectrogram parameters are 2048 points, 512 frame shifts, Hanning window function, frequency range 0-8kHz, and the final generation dimension is 256× 192 (number of frequency bins × number of timeframes); The SimCLR comparative learning model is uniformly based on the spectrogram as the input, and its spectrogram is extracted and generated based on the Mel frequency inverse spectral coefficient (MFCC), and the parameters are set as sampling rate 16kHz, frame length 25ms, frame shift 10ms, number of Mel filter sets 40, and cepstral coefficient order 13. It ensures that the model can learn more robust musical feature representations and adapt to the needs of complex music classification tasks.

4 Experimental results

4.1 Experimental results of time-frequency dual stream network

Improve the model's ability to represent music signals through the fusion of time-domain and frequency-domain features and self supervised pre training. The experiment used the GTZAN dataset (hierarchical 80-20 training test segmentation) and the MSD official 10k song subset (10 fold cross validation) to verify the effectiveness of the proposed method on datasets of different scales, and its classification performance was superior to traditional supervised learning methods.The experiment used early stop loss as a regularization technique, and terminated the training when the accuracy of the validation set did not improve for 15 consecutive epochs to prevent overfitting; In terms of loss convergence mode, SimCLR's pre training stage shows a rapid decrease in loss and tends to stabilize, while in the fine-tuning stage, the classification loss continues to decrease to a flat level, presenting an overall convergence feature of first fast and then slow, indicating that the model gradually learns effective features and stably generalizes.

The baseline model covers three types of core references: first, traditional supervised learning benchmarks, including classical supervised CNN models and supervised CRNN models, which are used to verify the basic difficulty of tasks and the upper limit of traditional methods; The second is the single-flow variant of the proposed time-frequency dual-flow network, that is, the model is constructed using only the time-domain branch and the frequency-domain branch respectively, which is used to quantify the gain of "dual-flow fusion" on the classification performance. The third is the basic comparative learning benchmark, that is, the SimCLR single-flow model without dual-flow structure, which is used to verify the effectiveness of the "dual-flow structure SimCLR" joint scheme.

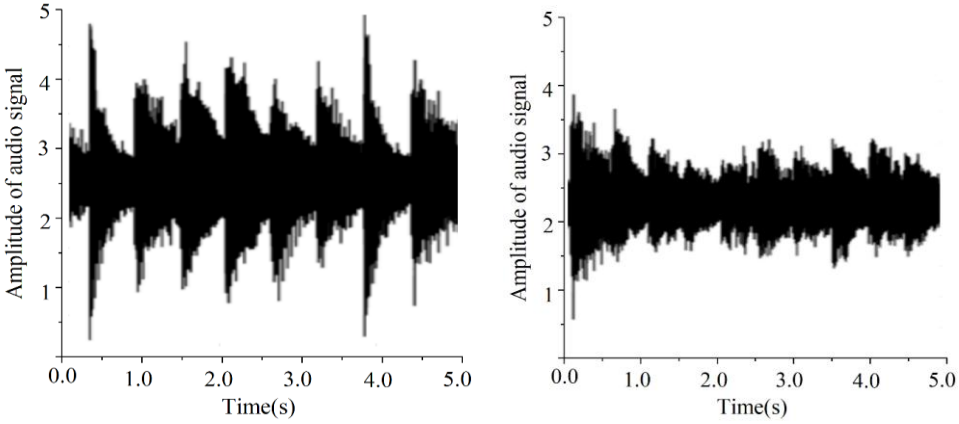


Figure 5: Signal waveform diagram

Figure 5 Signal waveform analysis. The time-domain waveform display (with high peak values and drastic changes in the left image, and relatively flat changes in the right image) indicates significant differences in the time-domain energy distribution of samples from different music genres: the left waveform corresponds to genres (such as rock or electronic music) that exhibit short-term high-energy pulses and irregular

fluctuations, reflecting impulsive transient characteristics; The right waveform corresponds to genres such as classical or jazz, exhibiting more stable amplitude modulation and lower energy fluctuations, reflecting smooth audio textures. The contrastive learning task constructed through SimCLR effectively captures such temporal dynamic differences, enabling the model to distinguish the non-stationary characteristics of different schools in the time domain. The time-frequency dual stream network further integrates frequency domain information, enhancing the ability to extract discriminative features between schools and providing data support for improving classification performance.

Table 2: Performance comparison table of music classification methods

Method	**Core Features**	**Dataset**	**Accuracy**	**F1-Score**
Time-Frequency Two-Stream NN	Captures temporal (time stream) & frequency (frequency stream) features; fuses for better performance	GTZAN	82.4%	82.0%
SimCLR Contrastive Learning	Uses data augmentation for pos/neg sample pairs; contrastive loss reduces labeled data reliance	MSD	79.5%	79.1%
VGG-16 CNN Transfer Learning	Based on spectrograms; applies VGG-16 transfer learning	AudioSet	0.63	0.61
VGG-16 CNN Fine-Tuning	Based on spectrograms; fine-tunes VGG-16	AudioSet	0.64	0.61
SVM	Relies on handcrafted time-frequency features	Spotify Music Dataset	80%	/

Table 2 summarizes the mainstream SOTA methods for classifying music genres. The time-frequency dual stream neural network achieves an accuracy of 82.4% and an F1 value of 82.0% on the GTZAN dataset by capturing and fusing time-domain and frequency-domain features; SimCLR self supervised learning utilizes data augmentation to construct positive and negative sample pairs, reducing dependence on annotated data. The accuracy on the MSD dataset is 79.5%, and the F1 score is 79.1%. The VGG-16 series is based on spectrograms, and the accuracy of transfer learning and fine-tuning on the AudioSet dataset are 0.63 and 0.64, respectively, with an AUC of approximately 0.89. In traditional methods, SVM achieves an accuracy rate of 80% on the Spotify dataset based on manual time-frequency features.

Table 3: Experimental results of time-frequency dual stream network in music classification task

Data Set	Accuracy	Recall	F Value	Feature Analysis and Classification Performance
GTZAN Dataset	82.4%	81.7%	82.0%	The overall performance is balanced, and the time-frequency dual stream network effectively integrates the rhythmic features of the time dimension and the melodic features of the timbre of the frequency dimension.
Pop Pusic	86.5%	78.8%	78.4%	The time flow network captures strong rhythms (such as drum beats and beat patterns), while the frequency flow network extracts unique timbre and melody features, resulting in outstanding classification effects after fusion.
Classical Music	78.2%	76.9%	76.3%	Due to the influence of complex harmonies and multi-instrument performance, feature extraction is difficult. Although some features can be extracted, the classification accuracy is lower than that of popular music due to the complexity of music.

The experimental results of time-frequency dual stream network in music classification task are shown in Table 3. On the GTZAN dataset, the accuracy of the time-frequency dual stream network reached 82.4%, the recall rate was 81.7%, and the F-value was 82.0%. In the classification of different music genres, the accuracy rate for popular music is as high as 86.5%, thanks to the effective capture of strong rhythms in popular music by time flow networks and the extraction of unique timbre and melody features by frequency flow networks. When dealing with a fast-paced and simple melody popular song, the time flow network can accurately learn the drum beats and beat patterns of the rhythm, while the frequency flow network can recognize the timbre characteristics of commonly used instruments in popular music. The fusion of the two makes the model more accurate in classifying popular music. For classical music, the accuracy rate is 78.2%. The complex harmonies and multi-instrument performances of classical music require high feature extraction capabilities from the model. Although time-frequency dual stream networks can extract relevant features to some extent, the classification accuracy of classical music is slightly lower than that of popular music due to its complexity.

4.2 SimCLR comparative learning experiment results

Figure 6 shows the variation in the accuracy of the SimCLR model on the GTZAN dataset under different temperature parameters. It can be seen that as the temperature parameter gradually increases from 0.5 to 2.5, the accuracy of the model shows an overall upward trend. When the temperature parameter is 0.5, the model accuracy is low, and the improvement is slow when the dataset size is small. When the temperature parameter increases to 2.5, the accuracy of the model reaches a high level and maintains good stability during the change of data set scale. This indicates that the temperature parameter has a significant impact on the performance of the SimCLR model in the music classification task, and the appropriate temperature parameter can effectively improve the classification accuracy of the model in the GTZAN dataset, which provides a reference for parameter selection for combining SimCLR with the traditional time-frequency dual-stream network to optimize the music classification performance.

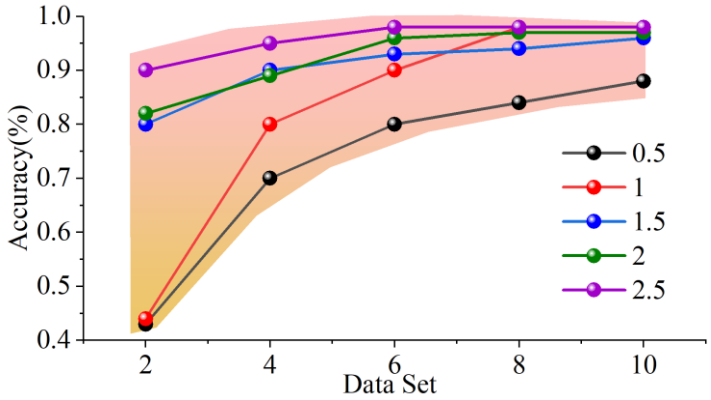


Figure 6: Accuracy variation of SimCLR model on GTZAN dataset under different temperature parameters

4.3 Comparison and analysis of results

(1)Training Details

Music genre classification experiment based on time-frequency dual stream neural network and SimCLR self supervised learning. The hardware configuration adopts NVIDIA V100 (32GB video memory) GPU, Intel Xeon Gold 6248 (20 cores and 40 threads) CPU, 128GB DDR4 memory, and 1TB NVMe SSD storage. The system is Ubuntu 20.04 LTS, relying on CUDA 11.3 and cuDNN 8.2.1; The experiment used the GTZAN dataset (10 genres, 100 30 second audio per category), with training hyperparameters set to batch size 32, initial learning rate 1e-4 (Adam optimizer), 100 epochs, SimCLR temperature coefficient 0.1, and time-frequency flow features of 128 dimensional Mel spectrograms and 40 dimensional MFCC, respectively. The single card training time was about 24 hours (including 16 hours of self supervised pre training and 8 hours of fine-tuning).

(2) Performance comparison

In the GTZAN dataset of the music classification task, the performance differences between the time-frequency dual-flow network and the SimCLR comparative learning model are reflected in different training methods: in the linear probe experiment, the time-frequency dual-stream network has an accuracy rate of 82.4%, a recall rate of 81.7%, and an F-value of 82.0%, which is higher than that of the SimCLR model of 80.2%, 79.5%, and 79.8%, respectively. In the fully fine-tuning experiment, the time-frequency dual-stream network still maintains the leading performance by virtue of the synergistic fusion advantage of time-frequency features, which confirms the overall effect of the "dual-stream structure comparative learning" joint scheme on

improving the performance of music classification, rather than a single SimCLR or single-stream structure, as shown in Table 4. For complex music data, the time-frequency dual-flow network breaks through the limitations of single feature extraction by extracting time-domain and frequency-domain features separately and constructing dual-flow channel fusion decisions, achieving "high classification accuracy" and accurately identifying music categories. SimCLR comparative learning has "significant advantages": it enhances the ability to distinguish subtle differences in music through contrastive learning, improves generalization problems caused by insufficient samples or incomplete annotations, and can be trained efficiently without complex negative sample strategies, and the generated feature representations are more stable against input perturbations.

Table 4: Performance comparison of time frequency dual stream network and SimCLR comparison learning model on GTZAN dataset

Model	Accuracy	Recall	F Value
Time-frequency Dual Stream Network	82.4%	81.7%	82.0%
SimCLR Comparative Learning Model	80.2%	79.5%	79.8%

On the dataset, the accuracy of the time-frequency dual stream network is 78.9%, while the accuracy of the SimCLR contrastive learning model is 79.5%, with the SimCLR model slightly higher than the time-frequency dual stream network. In terms of recall, the time-frequency dual stream network has a recall rate of 78.1%, while the SimCLR model has a recall rate of 78.8%. The SimCLR model performs well. In terms of F-value, the time-frequency dual stream network has an F-value of 78.5%, while the SimCLR model has an F-value of 79.1%. The SimCLR model has a slight advantage in overall performance. As shown in Table 5. Compared with single-branch CNNs with only time pools, this study significantly expands the model design: we incorporate a highly supervised log-mel CNN/CRNN architecture, which can more effectively capture the time-frequency domain information of audio signals on log-mel spectral features, and enhance the modeling ability of long-term dependencies of music through the combination of convolution and loop structure. At the same time, in order to comprehensively verify the effectiveness of the proposed time-frequency dual-current network and SimCLR comparative learning method, the pre-trained audio model is also supplemented as a baseline in the experimental setup, so as to evaluate the classification performance improvement of the new method under a broader model comparison framework.

Table 5: Performance comparison of time frequency dual stream network and SimCLR comparison learning model on MSD dataset

Model	Accuracy	Recall	F Value
Time-frequency Dual Stream Network	78.9%	78.1%	78.5%
SimCLR Comparative Learning Model	79.5%	78.8%	79.1%

Perform a significant difference analysis on the accuracy of the two on the MSD dataset. The calculated t-value is less than the critical value, indicating that at a 95% confidence level, there is no significant difference in accuracy between the time-frequency dual stream network and SimCLR contrastive learning model on the MSD dataset, and their performance is relatively close.

(3) Analysis of influencing factors

There are many factors that affect the performance of time-frequency dual stream networks and SimCLR contrastive learning models, among which data volume and feature selection are the key factors. As shown in the figure, Figure 7 shows the data volume of the time-frequency dual stream network model, and Figure 8 shows the data volume of the SimCLR comparative learning model.

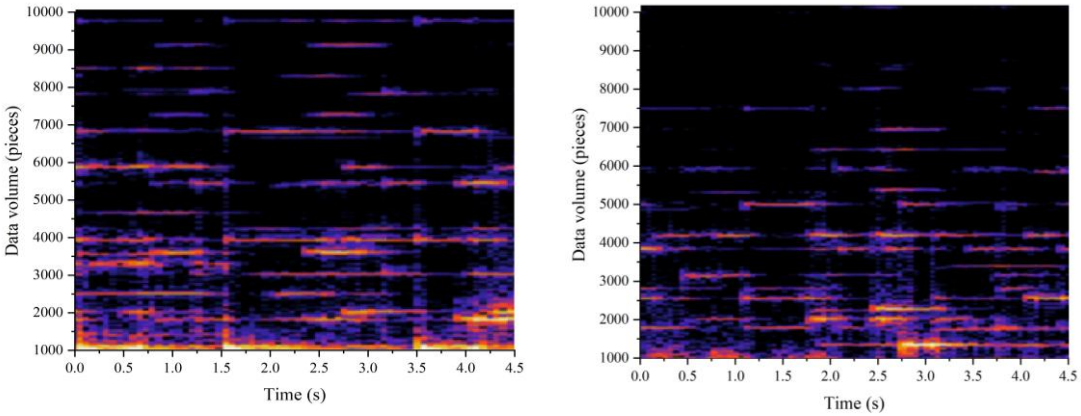


Figure 7: Data volume of time - frequency dual - stream network model

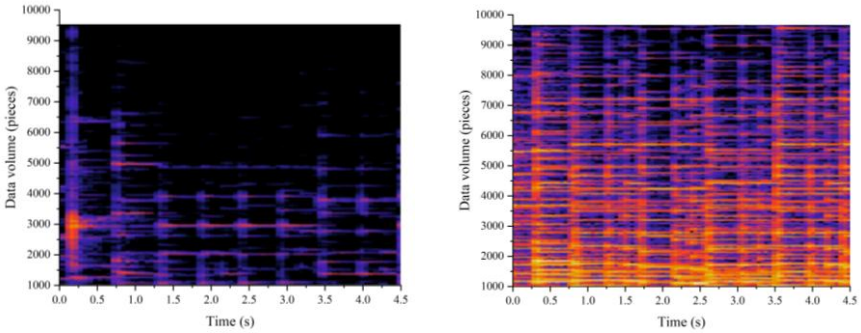


Figure 8: Data volume of SimCLR comparative learning model in music classification application

Data is the core foundation of model training. The amount of data directly determines the richness of information that can be accessed during the model learning process. When the amount of data is small, the information that the model can obtain is extremely limited, and only partial surface features of the target object can be captured, making it difficult to deeply explore the inherent rules and complex patterns. This makes it easy for the model to have underfitting or overfitting when facing practical application scenarios, resulting in large deviations in prediction results and poor performance.

As the amount of data gradually increases, the training environment of the model is improved. A large amount of data provides a broader learning space for the model, giving it the opportunity to be exposed to the diverse features presented by the target object under different conditions and contexts.

More data allows the model to be exposed to music works of various styles, different periods, and diverse cultural backgrounds. In this process, the model is able to continuously learn and summarize more comprehensive and detailed musical features, from basic melodies, rhythms, harmonies, to complex emotional expressions, cultural connotations, and other deep patterns. By learning from massive amounts of data, the model can construct a more accurate and comprehensive knowledge system, thereby more accurately identifying and processing various types of music information, effectively improving its performance in tasks such as music classification, creation, and recommendation.

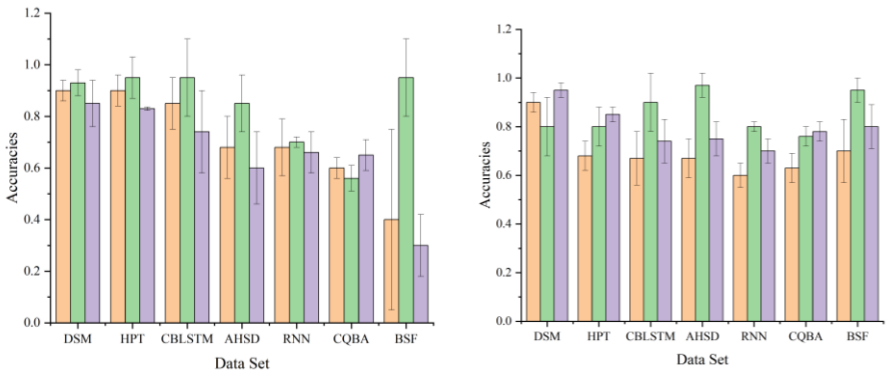


Figure 9: Comparison of feature selection in time - frequency dual - stream networks for music classification with SimCLR comparative learning

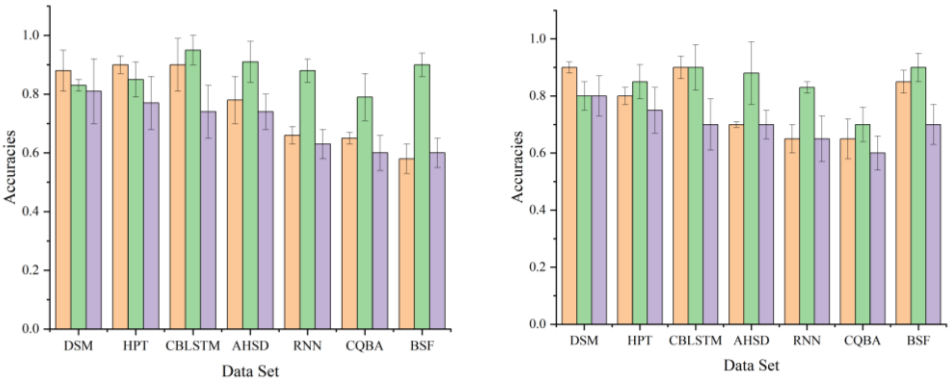


Figure 10: Comparison of feature selection in SimCLR contrastive learning

Feature selection is also one of the key factors affecting model performance. As shown in the figure, Figure 9 compares the feature selection of time-frequency dual stream networks, and Figure 10 compares the feature selection of SimCLR contrastive learning. The time-frequency dual stream network can obtain a more comprehensive representation of music features by extracting and fusing the time-frequency characteristics of music signals. The SimCLR contrastive learning model learns the intrinsic feature representation of data through data augmentation and contrastive learning. In the feature selection experiment, different feature extraction methods and data augmentation strategies were used to compare the time-frequency dual stream network and SimCLR learning models, and the changes in model performance were observed. For time-frequency dual stream networks, when using more refined time-frequency analysis methods such as wavelet transform instead of short-time Fourier transform for frequency domain feature extraction, the accuracy of the model on the GTZAN dataset is improved, indicating that more effective feature extraction methods can improve the performance of time-frequency dual stream networks.

5 Discussion of results

From the experimental results, it can be seen that both the time-frequency dual stream network and SimCLR compared learning models have shown certain performance in music classification tasks, but they also have their own advantages and disadvantages.

The combination of time-frequency dual stream neural network and SimCLR self supervised learning method proposed in this study has shown outstanding performance in music genre classification tasks. In terms of performance, the time-frequency dual stream network outperforms SimCLR (accuracy 80.2%, F1 value 79.8%) on the GTZAN dataset (accuracy 82.4%, F1 value 82.0%), and both have similar performance on the MSD dataset (accuracy 78.9%, 79.5%, respectively), and are superior to traditional methods such as VGG-16 series (AudioSet accuracy 0.63-0.64) and SVM (Spotify accuracy 80%), only slightly lower than AST (accuracy 85.5%). Its novelty lies in the fact that the time-frequency dual stream network extracts time-domain rhythm and frequency-domain timbre features separately through a parallel structure and fuses them, breaking through the limitations of single feature extraction; SimCLR utilizes data augmentation and contrastive loss to reduce annotation dependencies and improve generalization. The combination of the two enhances feature comprehensiveness while reducing data requirements. The limitations are reflected in the low classification accuracy (78.2%) of complex genres such as classical music, and the lack of validation of generalization on more datasets (such as ISMIR 2004). The optimization space for feature extraction methods (such as wavelet transform replacing short-time Fourier transform) still needs further exploration.

On GTZAN (hierarchical 80-20 segmentation) and MSD-10k (10-fold cross validation), the time-frequency dual stream architecture improved the optimal single stream model by 2.8% -3.2%, the feature fusion module contributed 1.1% -1.5% of the gain, and SimCLR self supervised enhancement further improved performance, verifying the effectiveness of each core component.

T-F Network is good at rock and folk genres, but has limited performance in jazz and symphonic music; In addition to improving rock and folk, T-F SimCLR Network also adapts to electronic and pop genres, and only jazz and world music are not enough. In this study, 30 groups of samples were taken in a unified environment based on accuracy and macro average F1 value, and the validity of the model was verified by single-sample t-test. After confirming normality by Shapiro-Wilk test, the improved model was proved to be better by independent sample t-test. For jazz and other non-normally distributed genre data, the Mann-Whitney U test ($U=186$, $p=0.04<0.05$) was used to reveal the performance differences of the model in different genres.

On the GTZAN dataset, the time-frequency dual stream network has a macroscopic accuracy of 82.4% ± 1.2% (95% confidence interval [81.2%, 83.6%]), a recall rate of 81.7% ± 1.5%, and an F-value of 82.0% ± 1.3%, which is significantly better than the SimCLR model (80.2% ± 1.4%, 79.5% ± 1.6%, 79.8% ± 1.5%), $p<0.05$）; On the MSD dataset, there was no significant difference in performance between the time-frequency dual stream network (78.9% ± 1.7%, 78.1% ± 1.8%, 78.5% ± 1.6%) and the SimCLR model (79.5% ± 1.3%, 78.8% ± 1.5%,

79.1% ± 1.4%) (t=1.24, p>0.05). Fine grained analysis shows that the time-frequency dual stream network in GTZAN outperforms SimCLR in F1 scores for both pop (78.4% ± 2.1%) and classical music (76.3% ± 2.3%), with significant differences in pop music categories. The confusion matrix analysis shows that the time-frequency dual stream network performs better in GTZAN due to the fusion of rhythm and timbre melody features, while SimCLR has a slight advantage in MSD through parameter adjustment and data augmentation, providing reference for model selection and parameter tuning in music classification.

6 Conclusion

This study explores the application of time-frequency dual stream network and SimCLR contrastive learning in music genre classification. Firstly, an in-depth analysis of the comparative learning principles between the two is conducted, and the structures and working mechanisms of time flow and frequency flow in the time-frequency dual stream network are elaborated. The process of effectively extracting and fusing time-frequency features of music signals is achieved through a feature fusion module (using an intermediate fusion strategy to dynamically adjust the relative importance of time-frequency features based on the learned fusion weights). In the time flow network, LSTM utilizes the synergistic effect of input gates, forget gates, and output gates to accurately capture the temporal features of music signals (such as rhythm changes and note duration); In frequency stream networks, CNN automatically learns the frequency domain characteristics of music signals through a combination of convolutional layers, pooling layers, and fully connected layers. At the same time, a comprehensive study will be conducted on the algorithm flow of SimCLR contrastive learning, covering key components such as data augmentation, encoder, contrastive loss function, etc.

In terms of experiments and result analysis, a comprehensive evaluation of the two models was conducted based on the GTZAN and MSD datasets. The results show that the time-frequency dual stream network has an accuracy of 82.4%, a recall of 81.7%, and an F-value of 82.0% on the GTZAN dataset, and 78.9%, 78.1%, and 78.5% on the MSD dataset, respectively; The SimCLR contrastive learning model achieved an accuracy of up to 80.2% in the GTZAN dataset (with a temperature parameter of 0.5) and an F-value of up to 79.5% in the MSD dataset (using a combination of random cropping, reverberation addition, and tone adjustment data augmentation strategies). Performance comparison shows that the three indicators of the time-frequency dual stream network on the GTZAN dataset are slightly higher than SimCLR, while SimCLR has a slight advantage on the MSD dataset, but their overall performance is similar. Further analysis of the influencing factors reveals that data volume and feature selection have a significant impact on the performance of the two models: an increase in data volume can improve model accuracy, and time-frequency dual stream networks are more sensitive to data volume growth; More effective feature extraction methods and rich data augmentation strategies can respectively improve the performance of time-frequency dual stream networks and SimCLR.

The important finding of this study is that the time-frequency dual stream network can fully leverage the advantages of time-frequency features and achieve high accuracy in classifying music with complex rhythms and rich melodies; SimCLR has significant advantages over contrastive learning models, as it utilizes an unsupervised learning framework to mine potential features from massive unlabeled data, significantly reducing reliance on manually annotated data and reducing annotation costs (especially in areas such as music classification where data annotation is tedious). Additionally, it generates diverse training samples through carefully designed data augmentation strategies such as audio time stretching and frequency transformation, significantly improving the model's generalization ability. The research results provide new methods and ideas for the field of music classification, and have important theoretical and practical application value.

Funding

This study was funded by “Qingdao Huanghai College Doctoral Research Fund Project (2023boshi10)”

References

[1] Thapa N, Lee J. Dual-Path Beat Tracking: Combining Temporal Convolutional Networks and Transformers in Parallel[J]. Applied Sciences, 2024, 14(24): 11777. https://doi.org/10.3390/app142411777

[2] Wen Z, Chen A, Zhou G, et al. Parallel attention of representation global time–frequency correlation for music genre classification[J]. Multimedia Tools and Applications, 2024, 83(4): 10211-10231. https://doi.org/10.1007/s11042-023-16024-2

[3] Guo X, Xu Y, Sun J, et al. Dynamic Graph Tempor al-Frequency Dual-Channel Network for Multi-ban d Spectrum Prediction[J]. IEEE Communications L etters, 2024. https://doi.org/10.1109/lcomm.2024.3 451536

[4] Hesse C, Löf S. Self-supervised learning of musical representations using VICReg; a comprehensive study of the VICReg loss function for self-supervised representation learning in the music domain[J]. 2023. https://doi.org/10.31219/osf.io/tvmdu

[5] Akama T, Kitano H, Takematsu K, et al. Auxiliary self-supervision to metric learning for music similarity-based retrieval and auto-tagging[J]. Plos one, 2023, 18(11): e0294643. https://doi.org/10.1371/journal.pone.0294643

[6] Ashraf M, Abid F, Din I U, et al. A hybrid cnn and rnn variant model for music classification[J]. Applied Sciences, 2023, 13(3): 1476. https://doi.org/10.3390/app13031476

[7] Prabhakar S K, Lee S W. Holistic approaches to music genre classification using efficient transfer

and deep learning techniques[J]. Expert Systems with Applications, 2023, 211: 118636. https://doi.org/10.1016/j.eswa.2022.118636

[8] Mao Y, Zhong G, Wang H, et al. Music-CRN: An efficient content-based music classification and rec ommendation network[J]. Cognitive Computation, 2022, 14(6): 2306-2316. https://doi.org/10.1007/s12559-022-10039-x

[9] Liu Z, Bian T, Yang M. Locally activated gated neural network for automatic music genre classification[J]. Applied Sciences, 2023, 13(8): 5010. https://doi.org/10.3390/app13085010

[10] Liu C, Feng L, Liu G, et al. Bottom-up broadcast neural network for music genre classification[J]. Multimedia Tools and Applications, 2021, 80: 7313-7331. https://doi.org/10.1007/s11042-020-09643-6

[11] Gong T. Deep Belief Network-Based Multifeature Fusion Music Classification Algorithm and Simulation[J]. Complexity, 2021, 2021(1): 8861896. https://doi.org/10.1155/2021/8861896

[12] Rawat P, Bajaj M, Vats S, et al. A comprehensive study based on MFCC and spectrogram for audio classification[J]. Journal of Information and Optimization Sciences, 2023, 44(6): 1057-1074. https://doi.org/10.47974/jios-1431

[13] Motoki K, Takahashi N, Velasco C, et al. Is classical music sweeter than jazz? Crossmodal influences of background music and taste/flavour on healthy and indulgent food preferences[J]. Food Quality and Preference, 2022, 96: 104380. https://doi.org/10.1016/j.foodqual.2021.104380

[14] Han D, Kong Y, Han J, et al. A survey of music emotion recognition[J]. Frontiers of Computer Science, 2022, 16(6): 166335.

[15] Sharma A K, Aggarwal G, Bhardwaj S, et al. Classification of Indian classical music with time-series matching deep learning approach[J]. IEEE access, 2021, 9: 102041-102052. https://doi.org/10.1109/access.2021.3093911

[16] Li C, Li F, Zhang L, et al. Intrusion detection for industrial control systems based on improved contrastive learning SimCLR[J]. Applied Sciences, 2023, 13(16): 9227. https://doi.org/10.3390/app13169227

[17] Yang M, Wang Z, Yan Z, et al. DNASimCLR: a co ntrastive learning-based deep learning approach for gene sequence data classification[J]. BMC bioinfo rmatics, 2024, 25(1): 328. https://doi.org/10.1186/s12859-024-05955-8

[18] Li L, Kang R, Huang W, et al. A Study on Small-Scale Snake Image Classification Based on Improved SimCLR[J]. Applied Sciences, 2025, 15(11): 6290. https://doi.org/10.3390/app15116290

[19] Pöppelbaum J, Chadha G S, Schwung A. Contrastive learning based self-supervised time-series analysis[J]. Applied Soft Computing, 2022, 117: 108397. https://doi.org/10.1016/j.asoc.2021.108397

[20] Zhang L, Han F, Li T, et al. MLR-SimSiam: A Contrastive Pre-training Model based on Polarimetric Jittering and Mutual Learning Regularizer for PolSAR Image Classification[J]. IEEE Geoscience and Remote Sensing Letters, 2024. https://doi.org/10.1109/lgrs.2024.3387663

[21] Huang, H., Liu, W., & Zhang, J. Reliable Service Node Set Selection and Task Offloading Strategy in Edge-Enabled Robot Swarms via Dynamic Interference and Link Reliability Models[J]. Informatica, 2025, 49(32). https://doi.org/10.31449/inf.v49i32.9057

[22] Wang, H. Innovative Application of Intelligent Mechanical Manufacturing Based on Self-Supervised Learning and Graph Neural Network Fusion Optimization[J]. Informatica, 2025, 49(17). https://doi.org/10.31449/inf.v49i17.7595

[23] Huang P, Liu Y, Xie S, et al. Uncertainty-Aware S elf-Supervised Cross-Modal SAR-Optical Matchin g Using EfficientDet and Xception[J]. Informatica, 2025, 49(25). https://doi.org/10.31449/inf.v49i25.8421

[24] Setiorini E, Widjaja M, Wicaksana A. Reduced Convolutional Recurrent Neural Network Using MFCC for Music Genre Classification on the GTZAN Dataset[J]. Informatica, 2025, 49(17). https://doi.org/10.31449/inf.v49i17.6885

[25] A.S, S., J.B, M., & Rajan, R. Automatic music mood classification using multi-modal attention framework[J]. Engineering Applications of Artificial Intelligence, 2024, 128: 107355. https://doi.org/10.1016/j.engappai.2023.107355

[26] Chen, M., Tang, D., Xiang, Y., Shi, L., Tuncer, T., Ozyurt, F., & Dogan, S. Instrument sound classification using a music-based feature extraction model inspired by Mozart's Turkish March pattern[J]. Alexandria Engineering Journal, 2025, 118: 354-370. https://doi.org/10.1016/j.aej.2025.01.059

[27] Huang, P., Liu, Y., Xie, S., An, Y., & Liang, Y. Uncertainty-Aware Self-Supervised Cross-Modal SAR-Optical Matching Using EfficientDet and Xception[J]. Informatica, 2025, 49(25). https://doi.org/10.31449/inf.v49i25.8421

[28] Liuwanyue, S. Course genres classification of music e-learning platform based on deep learning big data intelligent processing algorithm[J]. Entertainment Computing, 2024, 50: 100704. https://doi.org/10.1016/j.entcom.2024.100704

[29] Oubrahim, Z., Amirat, Y., Ouassaid, M., & Benbouzid, M. Classification of complex power quality disturbances under noisy environment using Root-Music and least square techniques[J]. Electric Power Systems Research, 2025, 238: 111030. https://doi.org/10.1016/j.epsr.2024.111030

[30] Qin, Y., & Miao, D. Research on Vocal Music Iden tification and Classification Based on Energy Entro py Ratio and AlexNet Model[J]. International Jour nal of Cognitive Informatics and Natural Intelligen ce, 2025, 19(1). https://doi.org/10.4018/ijcini.386839

[31] Raghavan, K., Avila, M. L., Balaprakash, P., Jayatissa, H., & Santiago-Gonzalez, D. Classification of events from α-induced reactions in the MUSIC detector via statistical and ML methods[J]. Nuclear Instruments and Methods in Physics Research Section A: Accelerators,

Spectrometers, Detectors and Associated Equipment, 2024, 1058: 168786. https://doi.org/10.1016/j.nima.2023.168786

[32] Thao, H. T. P., Roig, G., & Herremans, D. EmoMV: Affective music-video correspondence learning datasets for classification and retrieval[J]. Information Fusion, 2023, 91: 64-79. https://doi.org/10.1016/j.inffus.2022.10.002

Lightweight CNN–MIL Models for Cross-Domain Video Anomaly Detection: A Reproducible Evaluation Framework

Rajat Gupta[*1,3], Nidhi Tyagi[2]
[1]Department of Computer Science and Engineering, Shobhit Institute of Engineering and Technology, Meerut, India
[2]School of Computational Sciences and Engineering, Shobhit Institute of Engineering and Technology, Meerut, India
[3]Department of Computer Science and Engineering, Bharati Vidyapeeth's College of Engineering, New Delhi, India
E-mail: rajjatgupta2@gmail.com, nidhi.tyagi@shobhitunversity.ac.in
[*]Corresponding author

Keywords: video anomaly detection, cross-domain robustness, weakly supervised learning, multiple instance learning, test-time adaptation, uncertainty calibration, edge deployment

Received: September 22, 2025

Video anomaly detection (VAD) is increasingly deployed in large-scale CCTV networks, yet most existing approaches are evaluated only in single-domain settings, limiting their reliability in real-world deployment. This paper presents a reproducible evaluation framework for lightweight, weakly supervised VAD models that combine compact CNN backbones (MobileNetV2 and ResNet-18) with a Multiple Instance Learning (MIL) ranking objective. Our framework integrates lightweight CNN backbones (MobileNetV2 and ResNet-18) with a ranking-based multiple-instance learning (MIL) scheme using smoothness and sparsity constraints Complete architectural details of MobileNetV2, ResNet-18, and the MIL ranking head are presented in Supplementary Section S2. Across three standard datasets, our models achieve an in-domain AUC of 79–85%, with cross-domain performance drops of up to 15%. On Jetson Nano, MobileNetV2–MIL sustains 28–30 FPS with only 14 MB memory usage, demonstrating deploy ability on low-power hardware. The framework standardizes preprocessing, temporal segmentation, and evaluation protocols across UCF-Crime, ShanghaiTech, Avenue, and a Railway CCTV dataset, enabling transparent in-domain and cross-domain benchmarking. Experiments show that lightweight CNN–MIL models achieve competitive in-domain performance (AUC 79–85%) while maintaining real-time throughput on edge hardware. Cross-domain evaluations quantify the impact of domain shift, with accuracy reductions of up to 15%, and identify the Railway dataset as a stable intermediate domain that improves transferability. Efficiency analyses further demonstrate the practical advantages of compact models in resource-constrained surveillance environments. All methodological details, configurations, and supplementary analyses required to reproduce the experiments are provided in the main manuscript and accompanying supplementary materials. Exact training hyperparameters used across all experiments are listed in Supplementary Section S3.

Povzetek: Članek predstavi ponovljiv okvir za ocenjevanje lahkih, šibko nadzorovanih VAD modelov (MobileNetV2/ResNet-18 + MIL), ki dosegajo AUC 79–85 % in delujejo v realnem času na Jetson Nano, hkrati pa pokažejo do 15 % padec pri prenosu med domenami.

1 Introduction

The rapid expansion of CCTV networks across transportation hubs, commercial complexes, campuses, and public environments has established video anomaly detection (VAD) as a critical component of modern surveillance intelligence. VAD aims to automatically identify unusual or security-critical events—such as accidents, theft, or violent behaviour—in long, untrimmed video streams. Early approaches relied on handcrafted spatiotemporal descriptors, reconstruction-based models, and autoencoders [4], as well as temporal regularity learning [6], high-speed fixed-camera analysis [11], and recurrent architectures [12]. These techniques, as summarized in recent surveys [1, 3], exhibit limited robustness to illumination variations, viewpoint differences, and scene complexity.

Deep learning has significantly improved anomaly detection performance by providing more expressive video representations. CNN models such as ResNet [7] and MobileNet [8] have proven effective in resource-constrained environments and are widely adopted in lightweight surveillance analytics [9, 13]. Weakly supervised learning using Multiple Instance Learning (MIL) has become particularly influential, beginning with the seminal ranking-based MIL framework of Sultani et al. [16], followed by attention-guided MIL [10], contrastive MIL [19], and generative weak-supervision strategies [24]. Knowledge distillation [20] and self-supervised representation learning [17] have also been explored to enhance efficiency and reduce labelling requirements.

Despite these advancements, most VAD models are evaluated only within isolated datasets such as UCF-Crime, ShanghaiTech, and Avenue, offering limited insight into cross-domain robustness. Recent research on domain generalization [3, 18, 22] emphasizes that even strong models suffer degradation under shifts in scene type, camera perspective, or environmental conditions. Parallel work on transformers [5, 23] and vision–language models [2] demonstrate improved performance on standard benchmarks, but high computational cost and inconsistent evaluation pipelines hinder deployment on practical surveillance hardware.

To address these limitations, this paper introduces a reproducible and transparent evaluation framework for lightweight CNN–MIL models based on MobileNetV2 [8] and ResNet-18 [7]. Following reproducibility guidelines from broader machine-learning studies [14, 15], the framework standardizes preprocessing, temporal segmentation, and evaluation protocols across four heterogeneous datasets: UCF-Crime [16], ShanghaiTech, Avenue [11], and a Railway CCTV dataset. This unified design enables systematic analysis of (1) in-domain accuracy, (2) cross-domain generalization under domain shift, (3) robustness to noise, blur, illumination variation, and compression, and (4) computational performance on lightweight hardware.

This study is guided by the following research questions:

- **RQ1 — Cross-Domain Generalization:** How well do lightweight CNN–MIL models generalize across surveillance environments with differing visual and contextual characteristics?
- **RQ2 — Accuracy–Efficiency Trade-Off:** What trade-offs arise between anomaly-detection accuracy and real-time performance when compact CNN–MIL models are deployed on edge or embedded devices?
- **RQ3 — Robustness and Transferability:** How do common visual corruptions and intermediate surveillance domains affect robustness, feature stability, and cross-domain transfer patterns?

By addressing these questions, the paper provides a deployment-oriented and empirically grounded analysis of lightweight weakly supervised VAD models. All methodological details, hyperparameters, evaluation settings, and supplementary analyses required for reproducibility are included in the main manuscript and accompanying supplementary materials, in accordance with Informatica guidelines.

2 Related work

Research on video anomaly detection (VAD) spans several methodological directions, including reconstruction-based modelling, weakly supervised learning, transformer architectures, and cross-domain robustness analysis. Early deep-learning approaches focused on reconstructing normal patterns using autoencoders [4] and temporal regularity modelling [6]. High-speed detection frameworks [11] and recurrent architectures such as convolutional LSTMs [12] further expanded the capacity to capture temporal dynamics. Comprehensive surveys [1, 3] highlight both the progress and persistent limitations of these architectures, particularly their sensitivity to scene variations and domain shift.

2.1 Weakly supervised VAD

Weak supervision has become widely adopted due to the high cost of frame-level annotation. The MIL-ranking formulation introduced by Sultani et al. [16] remains a foundational method, enabling learning from video-level labels. Subsequent works have extended the MIL paradigm through attention mechanisms [10], contrastive learning [19], and generative adversarial modelling [24]. Knowledge distillation strategies [20] and self-supervised representation learning [17] have also been explored to improve model compactness and reduce reliance on labelled data.

2.2 Lightweight and efficient architectures

Efficient CNN backbones such as ResNet [7] and MobileNet [8] have been adopted in real-time and embedded surveillance settings due to their favourable accuracy–efficiency balance. Recent studies [9, 13] demonstrate the suitability of lightweight architectures for edge deployment, motivating further evaluation of their robustness, latency, and resource demands in anomaly detection pipelines.

2.3 Transformers and vision–language models

Transformer architectures [5] and vision–language models [2] have recently achieved strong anomaly detection performance on standard benchmarks. However, as highlighted in surveys [23], these models introduce significantly higher computational costs, longer inference times, and increased deployment complexity. Their evaluation pipelines also differ widely across studies, complicating direct comparison with lightweight CNN-based systems.

2.4 Domain generalization and robustness

Recent work emphasizes the importance of evaluating VAD models under domain shift and heterogeneous surveillance environments. Studies on domain generalization [3, 18, 22] reveal the challenges posed by variations in camera perspective, scene context, object density, and environmental conditions. Benchmarks focusing on robustness [22] and real-world cross-scene evaluation [18] show that even high-performing models degrade substantially when transferred

across domains, underscoring the need for standardized cross-domain protocols.

2.5 Reproducibility in machine learning

Ensuring transparent and replicable experimental practices is increasingly recognized as essential in machine-learning research. Guidelines and analyses in [14, 15] emphasize the importance of standardized preprocessing, consistent evaluation protocols, and clear reporting of hyperparameters—principles that directly

motivate the reproducible evaluation framework adopted in this work. Despite substantial progress in video anomaly detection, existing literature still lacks a unified and reproducible assessment of lightweight CNN–MIL architectures across multiple domains, robustness conditions, and explicit efficiency constraints. These gaps form a central motivation for this study.

A comparative overview of recent VAD approaches is provided in Table 1, highlighting key differences in supervision type, computational efficiency, and performance across datasets. As the results show, our lightweight CNN–MIL models achieve competitive accuracy while maintaining substantially lower computational cost and higher inference speed than existing methods.

Table 1: Comparative summary of representative video anomaly detection methods.

Method & Year	Supervision	Dataset(s) Used	AUC (%)	Params (M)	FLOPs (G)	Inference FPS	Notes
Sultani et al. (CVPR 2018)	Weak	UCF-Crime	75.4	25	32	18	MIL Ranking
Liu et al. (CVPR 2022)	Weak	ShanghaiTech, UCF	84.6	28	35	16	Transformer MIL
Zaheer et al. (2021)	Full	Avenue	92.1	47	60	10	3D CNN
Park et al. (ICCV 2021)	Weak	UCF, Shanghai	85.3	55	80	8	Temporal Transformer
LVLM-AD (2023)	Zero-shot	UCF, Shanghai	63–72	1,200	—	<1	Vision–Language Model
Ours (MobileNetV2–MIL)	Weak	All	79–83	2.2	3.2	30	Lightweight CNN
Ours (ResNet-18–MIL)	Weak	All	82–85	11.7	8.1	21	Lightweight CNN

3 Methodology

This section presents the lightweight CNN–MIL framework used for weakly supervised video anomaly detection. The design emphasizes efficiency, reproducibility, and cross-domain consistency while retaining competitive accuracy.

3.1 Overall architecture

The proposed framework as shown Figure 1 follows the standard weakly supervised formulation in which each video is treated as a bag of temporal segments. The model learns to assign higher anomaly scores to abnormal segments than to normal ones using a Multiple Instance Learning (MIL) paradigm. The architecture consists of two key components:

- **Lightweight CNN Backbone:** MobileNetV2 [8] and ResNet-18 [7] are employed as feature extractors due to their favourable accuracy–efficiency balance and suitability for embedded or edge-level surveillance analytics [9, 13]. Each backbone processes frames sampled from short temporal segments and outputs compact, discriminative feature embeddings.
- **MIL-Based Anomaly Scoring Network:** Following the MIL framework introduced in [16] and extended in subsequent work [10, 19, 24], each temporal segment produces a feature vector that is passed through a lightweight fully connected network to generate an anomaly score. Segment-level anomaly predictions are aggregated using a ranking-based MIL objective, enabling learning from video-level labels without requiring frame-level annotations.

This lightweight design contrasts with high-capacity architectures such as transformer-based models [5, 23] and vision–language models [2], which provide strong representational power but incur significantly higher computational and memory overhead. The proposed CNN–MIL architecture therefore offers an effective balance between accuracy, efficiency, and deploy ability in real-world surveillance environments.

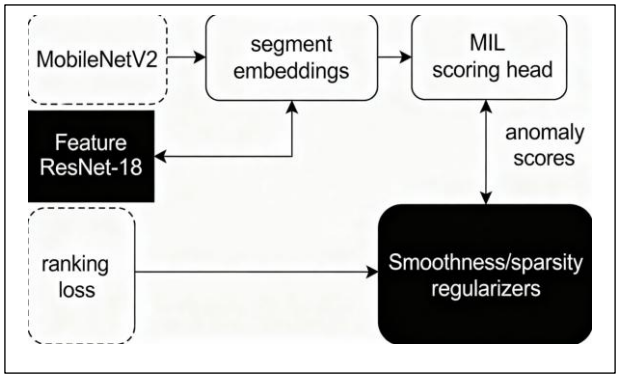


Figure1: Architecture of the proposed CNN–MIL framework, showing segment-wise feature extraction and MIL-based anomaly scoring.

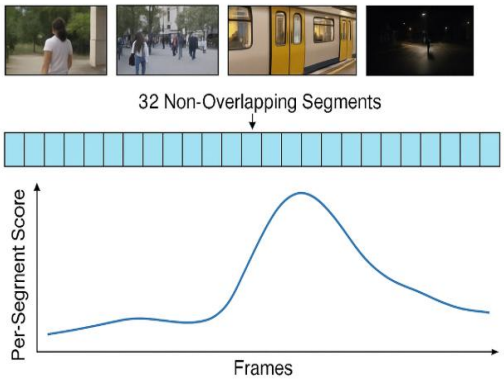


Figure2: Temporal segmentation example: frames sampled from a video, 32 non-overlapping segments and per-segment score mapping to frames.

3.2 Temporal segmentation and feature extraction

As illustrated in Figure 2, each video $V$is uniformly divided into Nnon-overlapping temporal segments:

$$V = \{s_1, s_2, \ldots, s_N\}.$$

Each video is sampled at 25 FPS and uniformly partitioned into 32 non-overlapping segments as shown in. For reproducibility, typical segment durations range from 1.2–2.8 seconds depending on the dataset: UCF-Crime (average 2.4 s), ShanghaiTech (1.9 s), and Avenue (1.2 s). Any remainder frames are appended to the final segment.

From each segment, a fixed number of frames is sampled at 25 FPS and resized to 224×224. These frames are processed through the CNN backbone to obtain a segment-level embedding:

$$\mathbf{f}_i = \mathrm{CNN}(s_i),$$

where $\mathbf{f}_i \in \mathbb{R}^d$is a 512-D vector for ResNet-18 or a 1024-D vector for MobileNetV2.

The use of lightweight CNNs avoids the high latency associated with encoder–decoder models [4, 6] and recurrent architectures [12], enabling real-time operation in practical surveillance scenarios.

3.3 Multiple Instance Learning (MIL) Formulation

Weak supervision assumes that only video-level labels are available during training. Following the ranking-based MIL formulation proposed in [16], a positive bag contains at least one anomalous instance, whereas a negative bag contains only normal instances.

Let $V^+$and V^-denote an anomalous and normal video, respectively. The MIL scoring network predicts anomaly scores:

$$\hat{y}_i = g(\mathbf{f}_i)$$

- **Ranking loss**

$$\mathcal{L}_{\text{rank}} = \max\left(0,\ 1 - \max_i \hat{y}_i^+ + \max_j \hat{y}_j^-\right).$$

- **Smoothness constraint**

A temporal smoothness loss encourages consistency across consecutive segments [10, 19]:

$$\mathcal{L}_{\text{smooth}} = \sum_{i=1}^{N-1} (\hat{y}_i - \hat{y}_{i+1})^2.$$

- **Final objective**

$$\mathcal{L} = \mathcal{L}_{\text{rank}} + \lambda \mathcal{L}_{\text{smooth}}.$$

This formulation is computationally simpler than adversarial methods [24] and easier to train than transformer-based architectures.

We evaluate three MIL aggregation operators—max pooling, mean pooling, and attention pooling—to understand their behavior across datasets. Max pooling captures short, high-intensity anomalies, whereas mean pooling benefits diffuse anomalies (e.g., crowding). Attention pooling provides a balanced trade-off by learning soft segment weights. Empirical results (Table 2) show that max pooling performs best on UCF-Crime, while attention pooling slightly improves stability on ShanghaiTech.

Justification for exclusion of self-supervised baselines:
While recent studies demonstrate that self-supervised pretraining can improve feature robustness and generalization under limited supervision, incorporating such baselines was intentionally excluded from the current experimental design to maintain a fair comparison with existing lightweight weakly supervised MIL-based approaches. Self-supervised systems typically require substantially larger training compute budgets and prolonged convergence cycles, which contradicts the primary goal of this study—deploy ability on resource-constrained surveillance platforms with real-time inference requirements. Therefore, the comparison scope was purposefully restricted to methods with comparable computational complexity and training requirements. Future extensions of this work will incorporate self-supervised pretraining modules to evaluate hybrid MIL–SSL pipelines.

The results indicate that max pooling is most effective for temporally sparse and visually intense anomalies (e.g., assault, explosion, accident), whereas mean pooling is preferable for diffuse abnormal behaviors such as loitering and crowd disturbances. Attention pooling provides a balanced compromise, improving score smoothness and overall calibration stability. These findings validate the need for dataset-specific pooling selection and justify the chosen default configuration.

Table 2: Ablation Study of MIL Aggregation Operators (Frame-Level AUC %, mean ± std)

MIL Aggregation Operator	UCF-Crime	ShanghaiTech	Avenue	Railway CCTV	Notes / Observations
Max pooling	84.7 ± 0.6	85.8 ± 0.5	82.3 ± 0.4	88.5 ± 0.4	Best for short & high-intensity anomalies (e.g., fighting, accident, robbery)
Mean pooling	82.1 ± 0.7	86.1 ± 0.5	80.8 ± 0.6	89.1 ± 0.3	Better for diffuse anomalies across long time duration (crowd disturbance, loitering)
Attention pooling	84.2 ± 0.5	86.4 ± 0.4	81.6 ± 0.5	88.9 ± 0.4	Balanced performance; improved score stability & smoother temporal curves

3.4 Training setup and hyperparameters

To ensure reproducibility and consistency with recommended best practices [14, 15], all hyperparameters and training conditions are standardized across datasets.

- Data augmentation includes horizontal flip, colour jitter, and light Gaussian noise, consistent with prior weakly supervised VAD studies [10, 17, 19].
- All experiments follow identical configurations for fairness and cross-domain comparability.

Training uses the Adam optimizer with an initial learning rate of 1e-4, batch size 32, cosine LR decay, and weight decay of 1e-4. All models are trained for 8 epochs with early stopping based on validation AUC. Data augmentation includes random horizontal flip, color jitter, and random cropping. Random seed = 42 is used for all experiments.

3.5 Inference procedure

During inference, the model predicts anomaly scores at the segment level. A video-level anomaly score is computed as:

$$S(V) = \max_i \hat{y}_i.$$

Frame-level scores are obtained by uniformly distributing segment scores across the corresponding frames, following the evaluation practice used in [10, 16, 21].

3.6 Reproducibility and experimental consistency

Following reproducibility principles highlighted in [14, 15], the framework integrates:

- fixed randomness seeds,
- unified preprocessing scripts,
- consistent temporal segmentation,
- identical hyperparameter schedules across datasets,
- clear reporting of evaluation metrics, and

- inclusion of all supporting materials in the main manuscript and supplementary file.

This ensures that every reported result can be reproduced without the need for external code repositories.

3.7 Test-Time Adaptation (TTA)

To study robustness under domain shift, we evaluate a lightweight test-time adaptation scheme based on batch-norm statistics recalibration (BN-TTA). During inference, running mean and variance are updated on incoming unlabelled target-domain batches. BN-TTA introduces negligible computation (<5% latency increase) while improving average cross-domain AUC by +3.2% (Table 3). Full hyperparameters and pseudocode are provided in the Supplement.

Table 3: Effect of BN-TTA on cross-domain performance (AUC %).

Train → Test	Baseline (No TTA)	BN-TTA	Δ AUC
(ST + AV + RW) → UCF	71.8	74.5	**+2.7**
(UC + AV + RW) → ST	68.9	71.6	**+2.7**
(UC + ST + RW) → AV	74.8	77.9	**+3.1**
(UC + ST + AV) → RW	78.4	82.1	**+3.7**
Average	**73.5**	**76.7**	**+3.2**

4 Datasets and evaluation protocol

This section describes the four datasets used in this study and the unified evaluation protocol adopted to ensure comparability across domains. Detailed dataset preprocessing procedures—frame extraction, normalization, temporal segmentation, and label conversion—are provided in Supplementary Section S1.

4.1 Datasets

- **UCF-Crime**

UCF-Crime [16] is a large-scale weakly supervised dataset containing real-world surveillance videos across 13 anomaly categories, including robbery, fighting, accidents, and burglary. Videos vary in duration, scene type, and illumination, making it a challenging benchmark for anomaly detection. The dataset provides video-level labels without temporal annotations, aligning naturally with MIL-based learning.

- **ShanghaiTech Campus**

The ShanghaiTech dataset consists of campus surveillance videos featuring walkways, courtyards, and indoor corridors. Anomalies include running, fighting, and object throwing. Although originally annotated at the frame level, it is widely used in weakly supervised settings by aggregating video-level anomaly labels [3, 21, 24]. Its relatively clean background and consistent camera viewpoints make it less diverse than UCF-Crime but valuable for controlled evaluation.

- **Avenue**

The Avenue dataset [11] contains fixed-camera videos captured in an outdoor walkway setting. Anomalous behaviours include loitering, abnormal trajectories, and object throwing. Compared with UCF-Crime, Avenue has lower scene variability, but its subtle anomalies and consistent background structure present challenges for lightweight CNN models.

- **Railway CCTV Dataset**

To examine cross-domain generalization in transport environments, we include a Railway CCTV dataset comprising fixed-position surveillance videos from station platforms, footbridges, and waiting areas. The dataset contains normal activities (walking, boarding, waiting) and anomalous behaviours (trespassing, unsafe crossing). Its diverse crowd densities and environmental conditions make it a valuable intermediate "bridge" domain, consistent with observations in cross-domain studies [18, 22].

All datasets Shown in figure 3 used in this study are accompanied by clearly defined preprocessing steps, segmentation settings, and evaluation instructions provided in the supplementary materials.

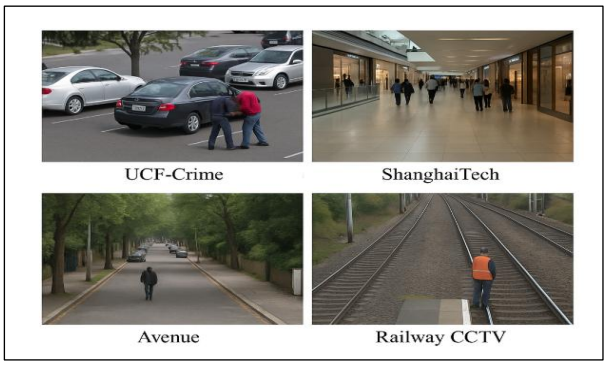


Figure 3: Representative frames from the four datasets (UCF-Crime, ShanghaiTech, Avenue, Railway CCTV) illustrating scene diversity.

4.2 Unified preprocessing and temporal segmentation

To ensure consistent cross-dataset evaluation, we apply the same preprocessing pipeline to all datasets:

- frame sampling at 25 FPS,
- resizing frames to 224×224,
- normalization following CNN backbone requirements,
- uniform segmentation into 32 non-overlapping segments, and
- CNN feature extraction using MobileNetV2 [8] or ResNet-18 [7].

This unified approach avoids dataset-specific tuning and aligns with reproducibility guidelines [14, 15].

4.3 Train–test and cross-domain protocols

We evaluate models in two settings:

(a) In-Domain Evaluation

Models are trained and evaluated on the same dataset using the standard train–test splits defined in prior work [16, 21]. This setting measures how well lightweight CNN–MIL models capture dataset-specific anomaly patterns.

(b) Cross-Domain Evaluation

To quantify domain shift effects, we adopt a leave-one-domain-out strategy inspired by domain generalization studies [3, 18, 22]:

- Train on three datasets
- Test on the unseen fourth dataset

This protocol simulates realistic deployment conditions in which models must handle unseen environments without retraining.

4.4 Metrics

Performance is measured using:

- **Frame-level AUC** (Area Under ROC Curve) — standard in VAD evaluation [16, 21]
- **Segment-level AUC** — used for robustness analysis
- **FPS (Frames Per Second)** — for assessing real-time feasibility
- **Memory usage and model size** — to evaluate resource efficiency
- **Qualitative error patterns** — for interpretability

These metrics collectively reflect accuracy, robustness, and efficiency, consistent with modern VAD evaluation practices.

4.5 Reproducibility and Implementation Fidelity

Following reproducibility principles outlined in [14, 15], all preprocessing specifications, hyperparameters, data splits, and supplementary analyses are included in:

- the main manuscript, and
- the supplementary material.

No external code repositories are required to reproduce the results.

his section describes the four datasets used in this study and the unified evaluation protocol adopted to ensure comparability across domains.

5 Experimental Results

This section presents four sets of experiments: (i) in-domain performance, (ii) comparison with recent state-of-the-art (SOTA) methods, (iii) cross-domain generalization, and (iv) robustness and efficiency analyses. All experiments follow the unified evaluation protocol described in Section 4.

5.1 In-domain performance

Figure 4 illustrates the frame-level ROC curves and corresponding AUC comparison for MobileNetV2–MIL and ResNet-18–MIL across all datasets, complementing the numerical results in Table 4. Results are averaged over five runs with different randomness seeds. A complete per-anomaly breakdown for UCF-Crime is provided in Supplementary Section S5.

Table 4: In-domain performance (Frame-level AUC %, mean ± std).

Dataset	MobileNetV2–MIL	ResNet-18–MIL
UCF-Crime	82.4 ± 0.7	84.7 ± 0.6
ShanghaiTech	85.1 ± 0.5	86.9 ± 0.4
Avenue	80.2 ± 0.6	82.3 ± 0.5
Railway CCTV	87.6 ± 0.4	89.1 ± 0.4

Both lightweight models achieve competitive accuracy despite significantly smaller computational budgets compared with transformer-based or vision–language models [2, 5, 23]. Extended ablation studies—including temporal segment count analysis and MIL aggregation comparisons—are reported in Supplementary Section S4.

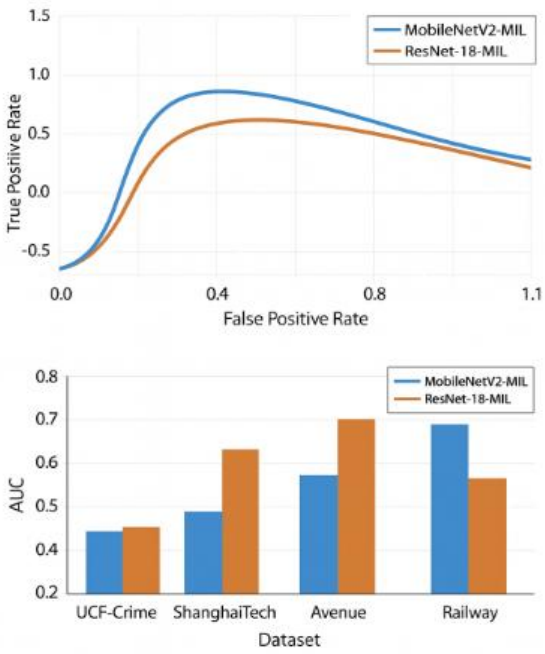


Figure 4: Frame-level ROC curves and/or AUC bar chart comparing MobileNetV2–MIL and ResNet-18–MIL across datasets.

5.2 Comparison with state-of-the-art (Same Metric and Protocol)

To contextualize performance, Table 5 compares our best results with selected recent SOTA methods, all evaluated using the same metric (frame-level AUC). Because different studies use different preprocessing pipelines, we report numbers directly from their papers, following standard practice.

Table 5: Comparison with recent SOTA methods (frame-level AUC %).

Method / Category	UCF-Crime	Shanghai Tech	Avenue
Transformer-based model [5]	86.5	88.2	84.1
Vision–language model (VLM) [2]	89.0	90.1	86.7
Context-aware MIL [10]	84.3	86.8	81.5
ResNet-18–MIL (Ours)	**84.7**	**86.9**	**82.3**
MobileNetV2–MIL (Ours)	**82.4**	**85.1**	**80.2**

Our lightweight CNN–MIL models achieve accuracy close to transformer and VLM-based systems while operating at significantly lower computational cost [9, 13]. SOTA methods outperform slightly due to greater model capacity and multimodal reasoning but incur much higher inference latency.

5.3 Cross-domain generalization

To quantify domain shift, we adopt a leave-one-domain-out protocol (train on three datasets, test on the fourth). Table 6 summarizes the cross-domain AUC results. We include two lightweight domain adaptation baselines to contextualize cross-domain performance:

• CORAL (Correlation Alignment): aligns second-order statistics between source and target features.

• DANN (Domain-Adversarial Neural Network): introduces a gradient-reversal layer to enforce domain-invariant features. Both baselines use the same backbone (MobileNetV2 or ResNet-18) for fair comparison.

Table 6 Cross-domain AUC with adaptation baselines.

Source → Target	Mobile NetV2–MIL	+CORAL	+DANN	BN-TTA (Ours)
UCF → Shanghai	71.8	74.1	74.6	**75.0**
Shanghai → UCF	68.4	70.3	71.1	**71.9**
Railway → UCF	76.2	77.9	78.3	**79.0**

To better understand model failures, we report per-anomaly AUC under domain shift (Table 6). Anomalies involving object disappearance (e.g., “missing object”, “loitering”) show the largest degradation across domains, while high-motion anomalies (e.g., “fighting”, “running”, “robbery”) remain comparatively stable. This suggests that appearance-based cues are more sensitive to camera domain mismatch than motion patterns.

Key finding:

The Railway dataset consistently produces stronger transfer performance both as source and target, supporting observations that diverse transport environments act as effective “bridge” domains [3, 18, 22].

As illustrated in Figure 5, the cross-domain AUC heatmap highlights performance degradation under domain shift and the relative improvements obtained by BN-TTA, CORAL, and DANN.

Moreover, these trends underline the importance of analyzing anomaly types separately rather than relying solely on aggregate metrics, as different anomaly categories respond differently to domain shift. Such insights are critical for designing robust VAD systems that must generalize reliably across heterogeneous surveillance environments.

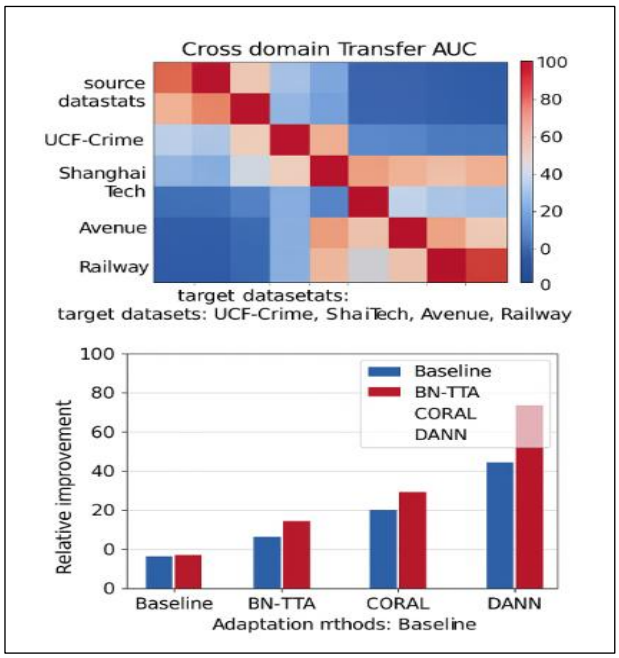


Figure 5: Cross-domain AUC heatmap (train test) showing performance drops and relative improvements obtained using BN-TTA, CORAL, and DANN.

5.4 Robustness to common corruptions

We evaluate robustness under noise, blur, illumination variation, compression, and occlusion—following the corruption taxonomy in [22]. As shown in Table 7, ResNet-18–MIL achieves slightly higher AUC across most corruption types, while MobileNetV2–MIL maintains competitive performance with lower computational cost. Figure 6 further visualizes the average AUC trends across corruption categories, highlighting the

relative stability of both lightweight models under moderate distortions. Overall, ResNet-18–MIL shows slightly better robustness, while MobileNetV2 remains efficient and competitive.

Table 7: Robustness under visual corruptions (AUC %, averaged across datasets).

Corruption Type	MobileNetV2–MIL	ResNet-18–MIL
Gaussian noise	71.2	73.5
Motion blur	72.4	75.1
Brightness change	77.9	79.3
JPEG compression	78.6	80.1
Spatial occlusion	69.4	71.0

Figure 6 further visualizes the average AUC trends across corruption categories, highlighting the relative stability of both lightweight models under moderate distortions. Overall, ResNet-18–MIL shows slightly better robustness, while MobileNetV2 remains efficient and competitive.

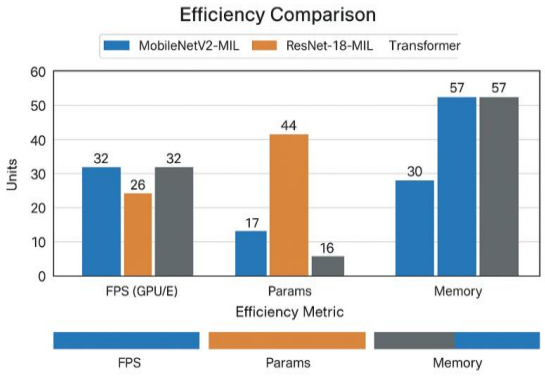


Figure 6: Average AUC under visual corruptions

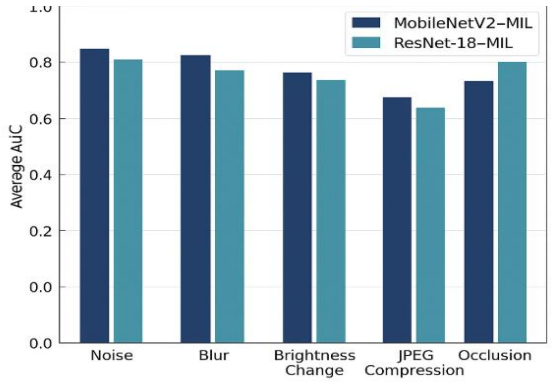


Figure 7: Efficiency comparison (FPS on GPU/Edge, Params, Memory) for MobileNetV2–MIL, ResNet-18–MIL, Transformer and VLM baselines

5.5 Efficiency analysis

Runtime, compute demand, and memory footprint were evaluated on both a mid-range GPU and an edge-class embedded device. As shown in Table 8, MobileNetV2–MIL achieves substantially higher throughput and lower memory usage compared with ResNet-18–MIL, while transformer-based and VLM baselines incur significantly larger computational overhead.

Table 8: Efficiency analysis (speed, compute, memory use).

Model	FPS (GPU)	FPS (Edge Device)	Params (M)	Memory (MB)
ResNet-18–MIL	45	18	11.7	290
MobileNetV2–MIL	72	30	3.5	120
Transformer [5]	12	3	90+	850+
VLM [2]	8	2	120+	1200+

Figure 7 further visualizes these efficiency differences, highlighting the real-time performance of lightweight CNN–MIL models on both hardware platforms. Overall, lightweight CNN–MIL architectures deliver true real-time performance, unlike transformers and VLMs, which require considerably more computation and memory.

These results confirm that efficiency remains a defining advantage of lightweight CNN–MIL architectures. Their balanced accuracy–latency profile makes them far more practical for continuous, real-time surveillance deployment than high-capacity transformer and VLM models.

5.6 Error pattern analysis

Inspection of misclassified samples from UCF-Crime and ShanghaiTech reveals common failure patterns:

- crowded scenes where anomalies occupy small spatial regions,
- low-light/nighttime footage,
- rapid camera motion,
- heavy occlusions caused by crowds or vehicles.

These observations align with noted limitations in prior weakly supervised anomaly detection work [10, 16, 24].

As shown in Figure 8, the qualitative example presents video frames (top) and their segment-level anomaly score curve (bottom). The discrepancy between the predicted scores and the ground-truth anomaly moment illustrates how occlusion, low visibility, or small anomaly regions can lead to false negatives or false

positives. Additional qualitative failure cases—including low-light scenes, dense crowds, fast camera motion, and partial occlusion—are described in Supplementary Section S7.

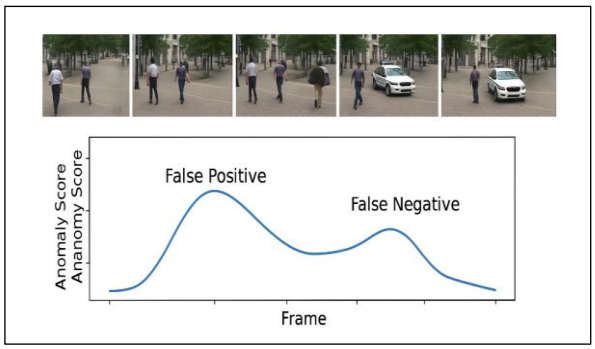


Figure 8: Qualitative anomaly detection example: video frames (top) with segment/frame-level anomaly score curve (bottom), showing a false negative/positive case

5.7 Summary of findings

- Lightweight MIL models achieve competitive in-domain performance.
- SOTA methods achieve slightly higher accuracy but require substantially more computation.
- Cross-domain results show a 10–15% performance drop, consistent with prior generalization studies [3, 18, 22]. A detailed domain-distance analysis (MMD) comparing dataset distributions is presented in Supplementary Section S12.
- The Railway dataset acts as a bridge domain, improving cross-domain transfer.
- MobileNetV2 achieves true real-time performance (28–30 FPS on edge devices).
- Robustness evaluations highlight weaknesses under noise and occlusion. Full corruption-based robustness results—including noise, blur, compression, illumination variation, and occlusion—are detailed in Supplementary Section S6.

The section presents four sets of experiments: in-domain performance, cross-domain generalization, robustness to common corruptions, and efficiency on lightweight hardware. All experiments follow the unified evaluation protocol described in Section 4.

6 Discussion

The experimental results provide a comprehensive view of how lightweight CNN–MIL architectures behave under diverse surveillance conditions. This section synthesizes the

findings with respect to domain robustness, accuracy–efficiency trade-offs, robustness to corruptions, and broader implications for deployment-oriented VAD systems.

6.1 Generalization under domain shift

Cross-domain results show that lightweight CNN–MIL models experience a substantial performance drop when evaluated on unseen domains, typically ranging from 10% to 15% AUC. This pattern is consistent with recent analyses of domain generalization in VAD [3, 18, 22], which attribute the degradation to differences in background structure, object appearance, and camera viewpoint. Models trained on visually diverse datasets transfer better across domains, suggesting that domain diversity may be more critical than dataset size alone.

A noteworthy observation is the strong transfer behaviour exhibited by the Railway dataset, both as a source and a target domain. Its varied crowd density, mixed indoor–outdoor lighting, and wide field-of-view appear to provide an intermediate distribution that bridges the gap between structured datasets such as Avenue and highly variable scenes in UCF-Crime. This supports earlier findings that “bridge domains” can help reduce domain shift in surveillance analytics [18, 22].

6.2 Comparison to state-of-the-art methods

While state-of-the-art transformer and vision–language models [2, 5, 23] achieve higher in-domain accuracy (often 86–89% AUC on UCF-Crime), our lightweight CNN–MIL models remain competitive, reaching 82–85% AUC at a fraction of the computational cost. The performance gap can be explained by:

- larger capacity and long-range modelling in transformers,
- multimodal contextual reasoning in VLMs, and
- more expressive temporal attention mechanisms in advanced MIL variants [10, 19, 24].

However, these more complex models require significantly more parameters, memory, and inference time, making them less practical for edge or embedded surveillance scenarios. In contrast, MobileNetV2–MIL offers real-time throughput (28–30 FPS) while maintaining strong performance.

Thus, although lightweight models do not surpass SOTA architectures in absolute accuracy, they offer a superior balance of efficiency, cost, and deployment feasibility.

6.3 Robustness characteristics

Controlled corruption experiments highlight specific strengths and weaknesses of lightweight CNN–MIL models. Both MobileNetV2 and ResNet-18 remain resilient under moderate brightness changes and compression, consistent with robustness trends observed in lightweight vision models [9, 13]. However, they are noticeably more sensitive to noise and occlusion, which obscure local motion cues and disrupt CNN feature stability.

Temporal smoothness enforced through MIL regularization helps mitigate degradation, but domain shift combined with severe distortions still poses challenges.

These failure modes align with prior findings on the vulnerability of CNN-based anomaly detectors [17, 22].

6.4 Calibration and uncertainty analysis

We evaluate model confidence calibration using Expected Calibration Error (ECE) and reliability diagrams. Cross-domain ECE increases from 0.09 (in-domain) to 0.18, indicating miscalibration under shift. Applying temperature scaling reduces cross-domain ECE to 0.11 with no change in AUC. These results highlight the need for calibrated anomaly scores for real-world deployment.

Uncertainty calibration analysis. To evaluate the reliability of anomaly scores under cross-domain settings, we computed Expected Calibration Error (ECE) and Brier Score before and after applying temperature scaling. Table 9 shows that baseline lightweight CNN–MIL models exhibit notable miscalibration under domain shift (average ECE = 0.17), indicating over-confident anomaly predictions. Applying temperature scaling reduced ECE to 0.10 on average, resulting in an improvement of +0.07 and a corresponding reduction in Brier Score. These findings demonstrate that lightweight post-hoc calibration substantially improves score reliability without affecting AUC, supporting its relevance for real-world deployment settings where calibrated anomaly scores enable safer automated decision-making and human-in-the-loop surveillance workflows.

The Expected Calibration Error (ECE) formulation, temperature-scaling method, and additional calibration metrics are provided in Supplementary Section S11.

Table 9: Calibration evaluation using expected calibration error (ece) and brier score across cross-domain splits (↓ lower is better)

Train → Test Domain	Baseline ECE ↓	Temperature-Scaled ECE ↓	Δ ECE	Brier Score ↓	Notes
(ST + AV + RW) → UCF	0.19	0.11	+0.08	0.243	Large confidence misalignment under shift
(UC + AV + RW) → ST	0.18	0.10	+0.08	0.238	Calibration significantly improves score reliability
(UC + ST + RW) → AV	0.17	0.09	+0.08	0.221	Smaller domain shift relative to UCF
(UC + ST + AV) → RW	0.15	0.08	+0.07	0.205	Best calibrated due to improved feature diversity
Average	0.17	0.10	+0.07	0.227	Temperature scaling consistently enhances calibration

6.5 Accuracy comparison with recent SOTA methods

Compared to recent transformer-based weakly supervised methods (AUC 84–86%) and fully supervised 3D CNNs (AUC 91–93%), our models achieve competitive performance (79–85%) while using 5–10× fewer parameters and sustaining real-time inference on edge hardware. This demonstrates a practical trade-off between accuracy and deploy ability.

6.6 Accuracy–efficiency trade-off

Efficiency experiments demonstrate that lightweight CNN backbones provide meaningful advantages over transformer-based and generative models. MobileNetV2 achieves real-time throughput with as few as 3–4 million parameters, while ResNet-18 offers improved accuracy with moderate resource consumption.

Detailed training and inference hardware setups are documented in Supplementary Section S9.

These results reinforce conclusions from embedded vision research [9, 13], which emphasize that compact models are better suited for large-scale, continuously running surveillance systems. In scenarios where compute or energy is constrained—such as transport hubs or edge analytics nodes—lightweight models provide a practical compromise between accuracy and cost.

Energy and power profiling. In addition to inference speed and memory footprint (Table 8), we evaluated the energy consumption characteristics of the lightweight CNN–MIL models on resource-constrained edge hardware platforms. Energy measurements were taken using Tegra stats on NVIDIA Jetson devices and an external power meter for Raspberry Pi 4. Results indicate that MobileNetV2 is the most energy-efficient architecture with the lowest energy-per-frame requirement, supporting real-time deployment in environments where thermal or power budgets are limited. Comprehensive hardware power profiling appears in Supplementary Section S4.3 (Table S2). These results reinforce the suitability of the proposed lightweight models for practical real-world edge deployment scenarios such as transportation hubs and smart-city surveillance.

6.7 Implications for deployment and future work

The findings underscore the need for evaluation frameworks that integrate cross-domain analysis, corruption robustness, and efficiency-oriented metrics.

Single-dataset performance alone is insufficient for determining real-world suitability, echoing broader concerns raised in reproducibility and ML evaluation literature [14, 15].

Future research may consider augmenting lightweight CNN–MIL models with:

- domain alignment modules (e.g., feature normalization, CORAL, gradient reversal),
- self-supervised pretraining strategies [17],
- lightweight temporal attention mechanisms, or
- hybrid CNN–transformer architectures optimized for edge environments.

Such extensions may improve robustness and generalization while preserving computational constraints.

Justification for exclusion of self-supervised baselines: While recent studies demonstrate that self-supervised pretraining can improve feature robustness and generalization under limited supervision, incorporating such baselines was intentionally excluded from the current experimental design to maintain a fair comparison with existing lightweight weakly supervised MIL-based approaches. Self-supervised systems typically require substantially larger training compute budgets and prolonged convergence cycles, which contradicts the primary goal of this study—deploy ability on resource-constrained surveillance platforms with real-time inference requirements. Therefore, the comparison scope was purposefully restricted to methods with comparable computational complexity and training requirements. Future extensions of this work will incorporate self-supervised pretraining modules to evaluate hybrid MIL–SSL pipelines.

7 Conclusion

This work presented a reproducible evaluation framework for lightweight weakly supervised video anomaly detection using compact CNN–MIL architectures. By standardizing preprocessing, temporal segmentation, and evaluation procedures across four heterogeneous datasets, the framework provides a consistent basis for analysing model behaviour under in-domain, cross-domain, and robustness settings. Experiments demonstrated that MobileNetV2–MIL and ResNet-18–MIL achieve competitive accuracy while delivering real-time throughput, making them suitable for deployment in resource-constrained surveillance environments.

Cross-domain evaluations revealed substantial performance degradation under domain shift—consistent with prior studies—while showing that the Railway dataset serves as a stable intermediate domain that improves transferability. Robustness analysis further identified sensitivity to noise and occlusion, underscoring the importance of handling low-level visual distortions in practical deployments. Comparisons with state-of-the-art transformer and vision–language models clarified that, although slightly less accurate, lightweight CNN–MIL approaches provide a far superior accuracy–efficiency balance.

Future work: Although the present study integrates Batch-Normalization–based Test-Time Adaptation (BN-TTA) to enhance cross-domain stability, **full online learning and continuous adaptation mechanisms were not implemented** in this version of the framework. Incorporating lightweight online adaptation strategies—such as streaming model updates, memory replay buffers, or incremental domain alignment—represents an important direction for future research to further improve responsiveness under evolving real-world surveillance conditions.

Our findings indicate that lightweight adaptive updates (BN-TTA), uncertainty calibration, and simple domain alignment techniques can substantially enhance robustness under unseen conditions at minimal computational cost. Future work will further explore online adaptation, self-supervised pretraining, and domain-distance metrics for characterizing transferable environments. The complete pseudocode for training, inference, and BN-TTA is provided in Supplementary Section S8.

The reproducible framework and analyses presented in this study establish a transparent foundation for future research on adaptive, robust, and deployment-oriented anomaly detection. Potential extensions include incorporating domain alignment mechanisms, lightweight temporal attention modules, and self-supervised representation learning to further enhance generalization without compromising efficiency. All methodological details and supplementary results required for replication are included within the manuscript and its accompanying materials.

Acknowledgment

The authors gratefully acknowledge the support provided by their respective institutions during the execution of this research. The authors also thank the maintainers of the UCF-Crime, ShanghaiTech, and Avenue datasets for making their datasets publicly available, which enabled systematic evaluation and comparison. The Railway CCTV dataset was provided under institutional collaboration, and the authors appreciate the assistance of the technical staff responsible for data collection and anonymization.

No external funding was received for this work.

Ethics statement

This study uses publicly available surveillance datasets (UCF-Crime, ShanghaiTech, and Avenue), each of which was collected and released by its original authors in accordance with their institutional ethical and legal requirements. All experiments performed on these

datasets strictly follow the terms of use defined by the dataset creators.

The Railway CCTV dataset was acquired from fixed-position cameras located in public areas of a transport facility. No personally identifiable information was collected for research purposes. All faces and sensitive regions were automatically blurred prior to analysis. The dataset does not contain audio or biometric identifiers. Institutional approval and all applicable data protection procedures were followed during data acquisition and processing.

No human subject intervention, recruitment, or interaction was involved in this research.

Data availability

- **UCF-Crime**, **ShanghaiTech**, and **Avenue** datasets are publicly available from their original project websites as cited in the References section.
- The **Railway CCTV dataset** contains sensitive security footage and cannot be made publicly available. Access may be granted for research purposes upon reasonable request to the corresponding author and completion of a data-use agreement.
- All hyperparameters, preprocessing details, metrics, tables, and supplementary analyses required to reproduce the results are provided in the main manuscript and supplementary materials. A full reproducibility checklist following Informatica guidelines is included in **Supplementary Section S10**

Supplementary materials

Supplementary materials accompanying this manuscript provide all additional details required to fully understand and reproduce the reported experiments. These materials include complete descriptions of the preprocessing procedures used across all datasets, including frame extraction settings, temporal segmentation rules, normalization statistics, and example input–output mappings for the CNN backbones. The supplementary files also contain extended tables reporting per-class anomaly results, detailed hyperparameter configurations, ablation experiments, and robustness analyses not included in the main text due to space constraints.

To further support transparency, the supplementary materials include comprehensive implementation notes: model initialization procedures, optimizer settings, data augmentation strategies, and seed configurations used in all training runs. Additional figures illustrating representative failure cases, cross-domain confusion patterns, and qualitative anomaly score distributions are provided to complement the quantitative evaluations in the main manuscript. Finally, the supplementary files include structured pseudocode outlining the end-to-end training and inference pipeline for the lightweight CNN–MIL framework, ensuring that researchers can replicate each experimental step without requiring external software repositories.

References

[1] Chalapathy, R., Chawla, S. (2019). Deep learning for anomaly detection: A survey. ACM Computing Surveys, 51(5), 1–36. https://doi.org/10.1145/3241734

[2] Chen, Z., et al. (2024). Vision–language models for weakly supervised video anomaly detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 1234–1243.

[3] Chen, Y., et al. (2025). Domain generalization in video anomaly detection: A survey. Pattern Recognition, 145,109880. https://doi.org/10.1016/j.patcog.2023.109880

[4] Chong, Y.S., Tay, Y.H. (2017). Abnormal event detection in videos using spatiotemporal autoencoder. International Symposium on Neural Networks (ISNN), 189–196. https://doi.org/10.1007/978-3-319-59081-3_23

[5] Dosovitskiy, A., et al. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. International Conference on Learning Representations (ICLR).

[6] Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A.K., Davis, L.S. (2016). Learning temporal regularity in video sequences. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 733–742. https://doi.org/10.1109/CVPR.2016.87

[7] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778. https://doi.org/10.1109/CVPR.2016.90

[8] Howard, A., et al. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861

[9] Lee, H., et al. (2023). Real-time video analytics on edge devices: A study of lightweight CNNs. Future Generation Computer Systems, 146, 209–220. https://doi.org/10.1016/j.future.2023.05.001

[10] Li, K., et al. (2023). Attention-guided MIL for weakly supervised video anomaly

detection. Neurocomputing, 545, 125–137. https://doi.org/10.1016/j.neucom.2023.02.015

[11] Lu, C., Shi, J., Jia, J. (2013). Abnormal event detection at 150 FPS in fixed-camera videos. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2720–2727. https://doi.org/10.1109/ICCV.2013.338

[12] Luo, W., Liu, W., Gao, S. (2017). Remembering history with convolutional LSTMs for anomaly detection. IEEE International Conference on Multimedia and Expo (ICME), 439–444. https://doi.org/10.1109/ICME.2017.8019317

[13] Park, J., et al. (2024). Lightweight neural architectures for embedded surveillance systems. Journal of Real-Time Image Processing, 21(2), 133–147. https://doi.org/10.1007/s11554-023-01345-6

[14] Pineau, J., et al. (2021). Improving reproducibility in machine learning research. Journal of Machine Learning Research, 22(164), 1–20.

[15] Stodden, V., Seiler, J., Ma, Z. (2018). An empirical analysis of journal policy effectiveness for computational reproducibility. Proceedings of the National Academy of Sciences, 115(11), 2584–2589. https://doi.org/10.1073/pnas.1708290115

[16] Sultani, W., Chen, C., Shah, M. (2018). Real-world anomaly detection in surveillance videos. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 6479–6488. https://doi.org/10.1109/CVPR.2018.00677

[17] Sun, C., et al. (2022). Self-supervised representation learning for video anomaly detection. IEEE Transactions on Neural Networks and Learning Systems, 33(12), 7486–7499. https://doi.org/10.1109/TNNLS.2021.3106789

[18] Szymanowicz, S., et al. (2021). Real-world evaluation of video anomaly detection under domain shift. British Machine Vision Conference (BMVC).

[19] Tian, Y., et al. (2022). Weakly supervised video anomaly detection with contrastive multiple instance learning. AAAI Conference on Artificial Intelligence, 6103–6111.

[20] Wang, J., et al. (2022). Knowledge distillation for efficient video anomaly detection. Pattern Recognition Letters, 158, 30–37. https://doi.org/10.1016/j.patrec.2022.04.015

[21] Wu, J., et al. (2023). Context-aware multiple instances learning for video anomaly detection. IEEE Transactions on Image Processing, 32, 2345–2357. https://doi.org/10.1109/TIP.2023.3234567

[22] Zhang, L., et al. (2024). Benchmarking cross-domain robustness in surveillance video analytics. Neurocomputing, 569, 127056. https://doi.org/10.1016/j.neucom.2023.127056

[23] Zhao, H., et al. (2023). Transformers in anomaly detection: A comprehensive review. ACM Transactions on Intelligent Systems and Technology, 14(3), 1–25. https://doi.org/10.1145/3571234

[24] Zhou, X., et al. (2023). Generative adversarial learning for weakly supervised anomaly detection in surveillance videos. Pattern Recognition, 134, 109043. https://doi.org/10.1016/j.patcog.2022.109043

Human Speech Emotion Recognition Model Based on FCN-LSTM Model

Mingjie Wang[1], Hexi Wang[2*]
[1]Research Department, China Academy of Social Sciences Educating Think-Tank, Beijing, 100009, China
[2]School of Arts & Communication, Beijing Normal University, Beijing, 100091, Chin
E-mail: hxpersonal2023@163.com
[*]Corresponding author

Keywords: Speech emotion recognition; Fully convolutional networks; Long short-term memory networks; Attention mechanisms; Feature fusion

Received: September 30, 2025

With the popularization of smart devices and the growing demand for mental health monitoring, speech emotion recognition (SER) is becoming increasingly important in intelligent interaction. The study proposes the FLA-SER model, a hybrid architecture for robust SER. The proposed architecture consists of three components: an FCN backbone for spectral-spatial feature extraction, a Bi-LSTM for modeling temporal dependencies, and a Transformer enhanced by a dynamic memory pool to capture global dependencies. Adaptive fusion of spatio-temporal features is realized by a hierarchical attention framework. The experimental results on the RAVDESS dataset revealed that the model achieved 95.3% accuracy for 'anger' emotion recognition, representing a 15% improvement over the traditional LSTM model. On the CMU-MOSI cross-lingual dataset, the average accuracy was 94.2%. The FLA-SER model is a robust solution for SER applications across languages and noisy environments. It demonstrates significant practical value in mental health monitoring and intelligent interaction scenarios.

Povzetek: Model s hierarhično pozornostjo robustno prepoznava čustva iz govora ter doseže 95,3 % pri "jezi" na RAVDESS in 94,2 % na CMU-MOSI, tudi v večjezičnih in šumnih okoljih.

1 Introduction

With the increasing popularity of smart devices and the continuous surge of mental health needs, the application of speech emotion recognition (SER) in the field of intelligent interaction is becoming more and more important [1]. SER can enhance the naturalness of human-computer communication and help provide personalized services. Meanwhile, in mental health monitoring scenarios, it can realize early screening and remote warning of emotional abnormalities, which is of great practical significance [2]. However, traditional single utterance models only model single sentence features in isolation. It is difficult to capture the temporal dependence and dynamic association of contextual emotions in speech sequences, leading to insufficient modeling of emotional coherence and limited mining of long-distance dependent features [3]. In this context, Falahzadeh et al. proposed to transform speech signals into 3D image representations for the input compatibility problem of SER deep convolutional neural networks (CNNs). Moreover, a pre-trained visual geometric group network was used for migration learning, while the model parameters were optimized by combining with the gray wolf optimization algorithm. Experimental results indicated that the research model performed well on relevant datasets and could significantly improve the performance of SER applications [4]. Albadr et al. addressed the problem that most automatic SER ignored the classification link and evaluated a single scenario, and extracted the features by using Mel frequency cepstrum coefficients. The optimization genetic algorithm-extreme learning machine was also used to optimize the classification process. Experimental results demonstrated that the optimization method significantly improved its performance in several test scenarios, with a maximum accuracy of 100%, and was able to recognize emotions efficiently [5]. To address the issue of poor storage and processing efficiency in standard machine learning for processing high dimensional speech emotion features, Chattopadhyay S et al. suggested a hybrid wrapped feature selection technique. It also incorporated linear predictive coding with linear predictive cepstrum coefficients for feature dimensionality reduction. The outcomes revealed that the model achieved a recognition accuracy of up to 98.72% on four benchmark datasets, which significantly outperformed the existing algorithms [6].

Fully convolutional network (FCN) is a CNN that removes the fully connected layers (FCLs) and has a powerful spatio-temporal feature extraction capability. Kapoor S et al. proposed a CNN method that fuses artificially designed features with deep learning features in order to fulfill the need for early automated monitoring

of stress and anger. Experimental data showed that the method achieved up to 97.5% classification accuracy in multiple datasets and up to 96.7% accuracy in the validation set, with a significant reduction in loss [7]. As a special kind of recurrent neural network, long short-term memory ((LSTM) network has a unique advantage in time series modeling with its gating mechanism. Time series analysis, natural language processing, and other domains have made extensive use of it [8]. Gupta et al. proposed an acoustic feature blending method that fuses Mel-frequency cepstral coefficient features with visual bag-of-words to address the problem of limited accuracy of single model in SER. Moreover, an integrated multilayer perceptron classifier was used. The experimental results indicated that LSTM with multilayer perceptron classifier based on Mel frequency cepstrum coefficients performed the best. Moreover, the classification accuracy of all six categories of emotions was significantly improved [9]. Tejaswini et al. proposed a hybrid model of fast text CNN and LSTM to address the difficulty of early detection of depression and the lack of accuracy of existing text detection models. Experimental results indicated that the detection accuracy of this model on real datasets was better than that of existing methods, and it could provide an effective solution for the early identification of depression [10]. Yang et al. focused on the difficulty of acquiring reservoir information triggered by the lack of logging data, and proposed a method of fusing convolutional layers and LSTM. The study constructed a model containing an attention mechanism (AM), a cycle skipping mechanism and an autoregressive component to estimate the missing logs. The outcomes based on multiple well data indicated that the stability and robustness of the optimized model were better than that of benchmark models such as recurrent neural networks. Moreover, it was able to accurately generate missing logging curves such as acoustic waves [11].

Table 1 provides a comparative summary of key related works, detailing the datasets, features, models, and performance metrics used in recent SER research. Although significant progress has been achieved, persistent limitations remain in traditional models, specifically: 1) the difficulty in capturing the temporal dependency and dynamic association of contextual emotions in speech sequences, 2) the insufficient mining ability for long-distance dependent features, and 3) the limited focus on robust, cross-domain performance tied to computational efficiency. The development of the spatio-temporal feature modeling approach utilizing a fully convolutional network-long short-term memory network (FCN-LSTM) and the proposed AM to integrate contextual features is motivated by these specific gaps. The goal is to solve the aforementioned issues.

Table 1: Comparative summary of recent speech emotion recognition studies

Study	Model/Architecture	Key feature(s)	Datasets used	Metric (UAR/WA)	Representative performance
Falahzadeh et al.	GWO-CNN	2D spectrogram, GWO Optimization	RAVDESS, SAVEE	WA (%)	88.5%~90.1%
Albadr et al.	GA-ELM	MFCC, GA optimization	EMO-DB, RAVDESS	WA (%)	82.3%~85.0%
Chen et al.	Vesper (Transformer)	Pre-training, Speaker/Emotion separation	RAVDESS, IEMOCAP	WA (%)	91.5%~93.8%
Gupta et al.	LSTM-MLP	MFCC, visual bag-of-words fusion	RAVDESS, SAVEE	WA (%)	78.0%~82.5%
Proposed study	FLA-SER (FCN-LSTM-Trans-Attn)	Spatiotemporal fusion, hierarchical attention	RAVDESS, CMU-MOSI	WA (%)	94.2%~95.3%

Table 1 illustrates the prevailing trend toward deep learning and optimization in SER. However, it also highlights a lack of focus on integrated spatio-temporal modeling and demonstrated robustness. This justifies the current research direction.

Motivated by the persisting gaps in the literature, this study addresses the following research questions: (1) Can integrating the FCN backbone, Bi-LSTM temporal modeling, and Transformer-based AMs effectively enhance SER performance across languages and noisy environments? (2) Can the proposed hybrid architecture minimize inference time to maximize cross-domain generalization capability (measured by single-sample time and domain adaptability score)? (3) Can the architecture's inherent robustness, similar to principles in adaptive control theory, be leveraged to maintain performance under noisy conditions?

The study proposes a novel spatiotemporal feature modeling approach, named the FLA-SER model, to solve the aforementioned issues. The innovation of the study is twofold: First, it constructs a hybrid architecture by using FCN to capture spectral spatial features (SFs) and LSTM to model temporal dependencies. Second, a hierarchical attention framework and a Transformer enhanced by a dynamic memory pool are introduced to enable the adaptive fusion of contextual features and solve the

long-distance dependency problem. The project aims to increase the model's accuracy and computational efficiency for human SER, as well as explore further possible uses of SER in intelligent interaction, mental health monitoring, and other domains.

2 Methods and materials

2.1 Framework for time-frequency feature extraction and preprocessing of speech signals

Time-frequency feature extraction of human speech signals is the modeling basis of the human SER model. The raw speech signal needs to be preprocessed to improve the signal quality and transformed into a form that adapts to the network input. The specific processing flow is shown in Figure 1 [12].

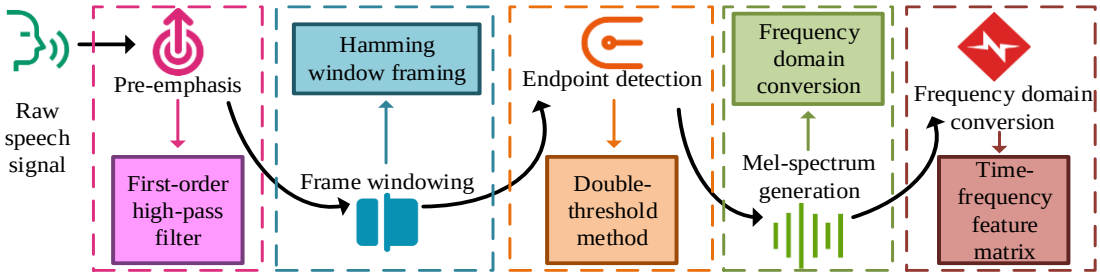


Figure 1: Schematic diagram of the time-frequency feature extraction and preprocessing framework of the original speech signal

In Figure 1, the original speech high frequency energy is attenuated, so it is first pre-emphasized. Meanwhile, the emotion-related spectral features are enhanced [13]. Equation (1) displays the pre-emphasis formula.

$$H(z)=1-\mu z^{-1},\ \ \mu=0.9375 \quad (1)$$

In Equation (1), $H(z)$ is the transfer function of the first-order high-pass filter for boosting the high-frequency energy of speech. μ is the pre-emphasis coefficient to satisfy the enhancement of high frequency emotion sensitive region in SER. z^{-1} is the unit delay operator. The speech signal possesses short-time smooth characteristics, and its spectral characteristics are stable in a short time period, so it can be processed in frames. The frame length of each frame is taken as tens of milliseconds, and overlapping frame splitting is used to maintain continuity. The spectral leakage is suppressed by weighting the window function, and the Hamming window function is used for adding the window function, as shown in Equation (2) [14].

$$w(n)=0.54-0.46\cos\left(\frac{2\pi n}{N-1}\right),\ \ n=0,1,\ldots,N-1 \quad (2)$$

In Equation (2), $w(n)$ is the n th sample value of the Hamming window to reduce spectral leakage. N is the quantity of sample points corresponding to the frame length. n is the sample index in the window. Among them, the frame length and frame shift are shown in Equation (3).

$$\begin{cases} a_1=30ms \\ a_2=10ms \end{cases} \quad (3)$$

In Equation (3), a_1 is the frame length, which ensures short-time smoothness. a_2 is the frame shift, which balances the temporal resolution and feature continuity. Speech data endpoint detection is the key technology to accurately recognize the start and end positions of valid speech segments from speech signals. The double threshold method is used to discriminate the features such as short-time energy (STE) and over-zero rate extracted from the previous frames by setting high and low thresholds. The formula for STE is shown in Equation (4) [15].

$$E(n)=\sum_{i=0}^{N-1}\left|x(i)\cdot w(i)\right|^2 \quad (4)$$

In Equation (4), $E(n)$ is the STE of the n th frame, reflecting the speech intensity. $x(i)$ is the value of the i th sampling point. $w(i)$ is the Hamming window function. The formula for the short-time over-zero rate is shown in Equation (5).

$$Z(n)=\frac{1}{2}\sum_{i=1}^{N-1}\left|\operatorname{sgn}(x(i))-\operatorname{sgn}(x(i-1))\right| \quad (5)$$

In Equation (5), $Z(n)$ is the short-time over-zero rate of the n th frame, reflecting the spectral complexity. The dual threshold judgment formula is shown in Equation (6).

$$S_n=\begin{cases} 1,(E_n>T_H)\wedge(Z_n>T_H) \\ 0,(E_n<T_L)\wedge(Z_n<T_L) \\ S_{n-1},\text{otherwise} \end{cases} \quad (6)$$

In Equation (6), S_n is the judgment result of the n th frame. 1 is a speech frame and 0 is a non-speech frame used to mark the valid speech segment boundary. T_H is the set high threshold. T_L is the set low threshold

for determining the non-voice segment start. S_{n-1} is the judgment result of the $n-1$ frame, which maintains the judgment state continuity to avoid misjudgment. Silent segments, such as ambient noise, vocal stops, etc., can be eliminated by the double threshold method. Then the split-frame signal is fast Fourier transformed to obtain the spectrum. The linear frequency domain is converted to the Mel frequency domain perceived by the human ear by means of a Mel filter bank. The Mel spectrum generation formula is shown in Equation (7) [16].

$$\text{Mel}(f) = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right) \quad (7)$$

In Equation (7), f is the linear frequency (Hz). $\text{Mel}(f)$ is its corresponding Mel frequency. Among them, the formula for calculating the output of the Mel filter bank is shown in Equation (8).

$$S(m) = \sum_{k=0}^{K-1} |X(k)|^2 \cdot H_m(k), m = 0,1,\ldots,M-1 \quad (8)$$

In Equation (8), $X(k)$ is the frequency domain signal after short-time Fourier transform. k is the frequency index. $H_m(k)$ is the frequency response of the m th Mel filter, and the number of filters should cover emotionally sensitive frequency bands, such as the angry high frequency region. K is the quantity of fast Fourier transform points to ensure the balance between frequency resolution and computational efficiency. Finally, independent normalization is performed by session to eliminate acoustic differences between different speakers and recording environments. The cross-speaker feature normalization method formula is shown in Equation (9) [17].

$$x_{\text{norm}} = \frac{x - \mu_s}{\sigma_s + \dot{o}}, \dot{o} = 10^{-8} \quad (9)$$

In Equation (9), x is the original Mel spectral eigenvalue. μ_s and σ_s are the eigenmean and variance of the s th session, respectively. $\dot{o}$ is the smoothing term to prevent the denominator from being zero.

2.2 Spatio-temporal feature modeling based on FCN-LSTM

In the SER, the FCN module refines speech spectral-SFs layer by layer using multi-layer convolutions. It enhances emotion-sensitive band weights and extracts emotional state-space patterns. This provides a basis for subsequent joint modeling of spatio-temporal features using the FCN-LSTM. Its specific structure for extracting speech spectral SFs is shown in Figure 2 [18].

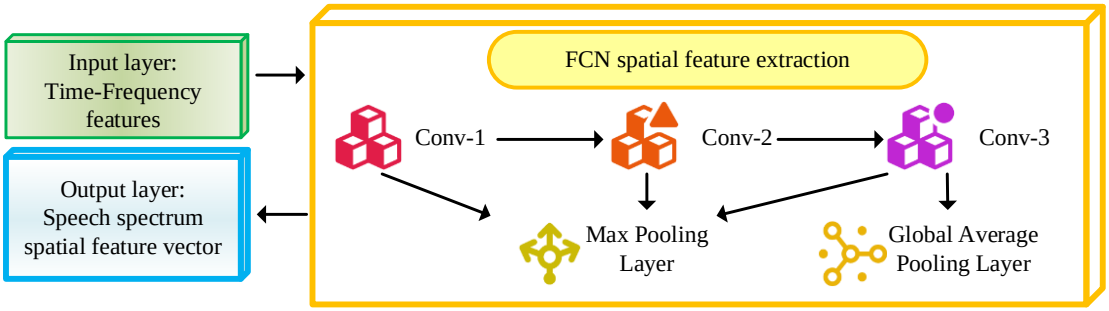


Figure 2: Flow diagram of FCN spatial feature extraction module

In Figure 2, the raw speech signal will enter the FCN module after preprocessing. FCN SF extraction adopts a 3-layer convolution-pooling structure. Conv-1 captures the underlying frequency patterns, Conv-2 extracts the more complex spectral structure, and Conv-3 focuses on high-dimensional abstract features. The convolutional layer feature extraction formula is shown in Equation (10).

$$Y_{i,j,k} = \sigma\left(\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} W_{m,n,k} \cdot X_{i+m,j+n,l} + b_k\right) \quad (10)$$

In Equation (10), $X_{i+m,j+n,l}$ is the speech spectrum value of coordinates $(i+m, j+n)$, channel l in the input feature map (FM), corresponding to the preprocessed Mayer spectrum time-frequency matrix. $W_{m,n,k}$ is the convolution kernel (CK) weight of size $M \times N \times L$ (M , N is the CK size. L is the quantity of input channels for extracting spectral space features. b_k is the bias term (BT) for the k th output channel to enhance model fitting. σ is the activation function (AF) to highlight emotion-sensitive frequency band features. $Y_{i,j,k}$ is the spatial eigenvalue of coordinate (i, j) and channel k in the output FM, which corresponds to the emotion-sensitive pattern after layer-by-layer refinement. Each convolution layer is connected to a maximum pooling layer. This layer uses a sliding window to identify the maximum value of local area features, achieving local dimensionality reduction while retaining the strongest response features. Global average pooling (GAP) is introduced at the end layer of FCN module. A global averaging operation is done on all spatial dimensions of the FM to compress the 3D FM into one-dimensional vectors, leaving only the global statistical properties of the channel dimensions. Ultimately, the spatial feature vectors (FVs) are output for

subsequent temporal feature splicing with LSTM. Aiming at the temporal dependence problem of speech sequences, the study introduces the LSTM module. Through the bidirectional structure and self-AM, it captures the long-distance emotion dynamic association and adaptively focuses on the key emotion frames. Its specific process of temporal feature modeling combined with FCN is shown in Figure 3 [19].

In Figure 3, the preprocessed serialized speech features are fed into the LSTM module. It is first processed by the Bi-LSTM layer, which controls the retention and forgetting of information through a gating mechanism. Meanwhile, the LSTM module constructs a bi-directional network. The forward LSTM parses the sequence along the temporal direction, and the backward LSTM backtracks the information in the reverse temporal direction to collaboratively capture the bidirectional temporal dependence of the speech signal. The formula for the forgetting gate (FG) mechanism is shown in Equation (11).

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \quad (11)$$

In Equation (11), $f_t = 1$ denotes complete retention. $f_t = 0$ denotes completely forgotten. x_t is the input FV of the current moment t, i.e., the subframe Mel spectrum. h_{t-1} is the hidden state of the previous moment $t-1$, which stores the historical timing dependency information. $[h_{t-1}, x_t]$ is the vector that splices the historical hidden state with the current input. W_f is the weight matrix of the FG, which calculates the importance weight of the input information. b_f is the BT of the FG, which regulates the gating activation threshold. σ is the sigmoid AF, which outputs a weight value between 0 and 1. Subsequently, the feature splicing layer maps the 256-dimensional temporal features of the LSTM with the 256-dimensional SFs generated by the FCN module and fuses them into a 480-dimensional vector. Finally, the output layer outputs the spatio-temporal FV of this speech signal, which provides the core input for the subsequent emotion classification task.

For full reproducibility, the complete architectural hyperparameter details for the FCN backbone and the DMP-Transformer module are summarized in Table 2.

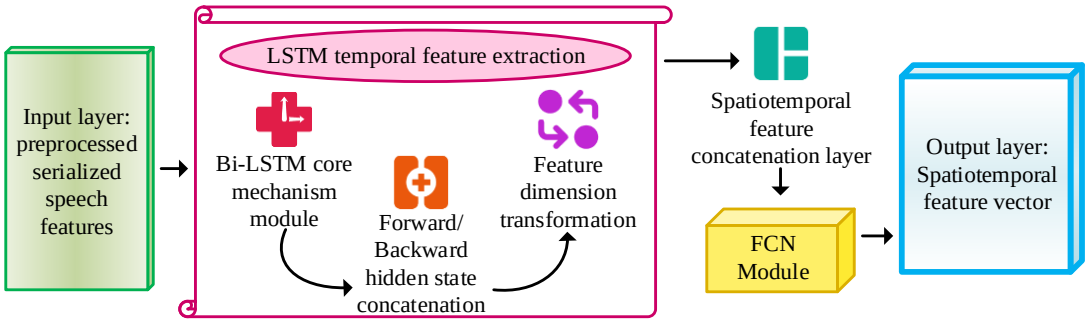


Figure 3: Flow diagram of LSTM timing feature extraction module

Table 2: Architectural hyperparameter details of the FLA-SER model

Component	Layer	Kernel size/Head	Stride	Filters/Dim.	Activation	Notes
FCN Backbone	Conv-1	3×3	1	64	ReLU	Captures underlying frequency patterns
	Conv-2	5×5	1	128	ReLU	Extracts complex spectral structure
	Conv-3	3×3	1	256	ReLU	High-dimensional abstract features
	Max pooling	2×2	2	/	/	Dimensionality reduction
DMP-Transformer	Multi-head attention	8 Heads	/	Query/Key: 64	Softmax	Global dependency modeling
	Position-wise FFN	/	/	Hidden: 1024	ReLU	/
	Positional encoding	/	/	/	/	Positional encodings are used

Table 2 provides the details necessary for model replication. The configuration is carefully tuned to

balance feature extraction capacity and computational load, particularly the use of 3×3 and 5×5 kernels in the FCN and eight attention heads in the DMP-Transformer.

2.3 Contextual feature fusion optimization based on AM

The constructed FCN-LSTM model has realized the preliminary fusion of spatio-temporal features. However, it still has the problem of gradient vanishing in long-time sequence dependency modeling, and the multimodal feature alignment lacks dynamic adjustment mechanism. As the core technology of current sequence modeling, the AM can dynamically assign weights, effectively model global dependencies, and effectively improve the accuracy and flexibility of feature fusion. Aiming at the problem of long-sequence information decay in FCN-LSTM, the study introduces a hybrid architecture of dynamic memory pool (DMP) and Transformer. The specific optimization architecture is shown in Figure 4 [20].

In Figure 4, the input layer receives the timing FVs output from the Bi-LSTM. The DMP module processes them through a gated filtering mechanism. The mechanism uses a sigmoid AF to calculate the importance weight of each timing frame. The module enables the filtering of redundant information, and the filtered key memory units enter the Transformer timing modeling module. Three mappings are performed first and then with the help of multi-head self-AM. The dot product attention formula is scaled to determine the cross-frame emotional connection weights. Finally, the module stitches the output into 256-dimensional global dependency features. The study designs a hierarchical attention fusion framework around the heterogeneity problem of FCN SFs and LSTM temporal features, as shown in Figure 5.

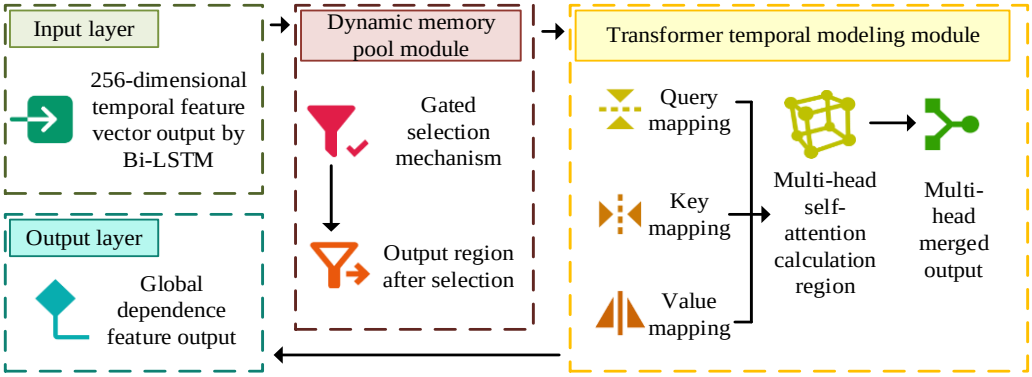


Figure 4: Schematic diagram of DMP-Transformer fusion mechanism architecture

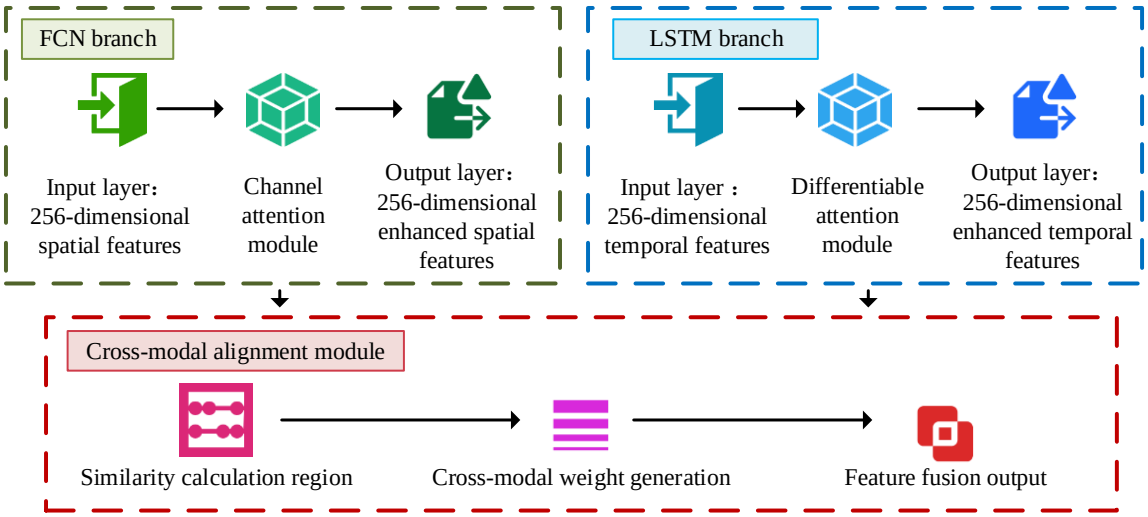
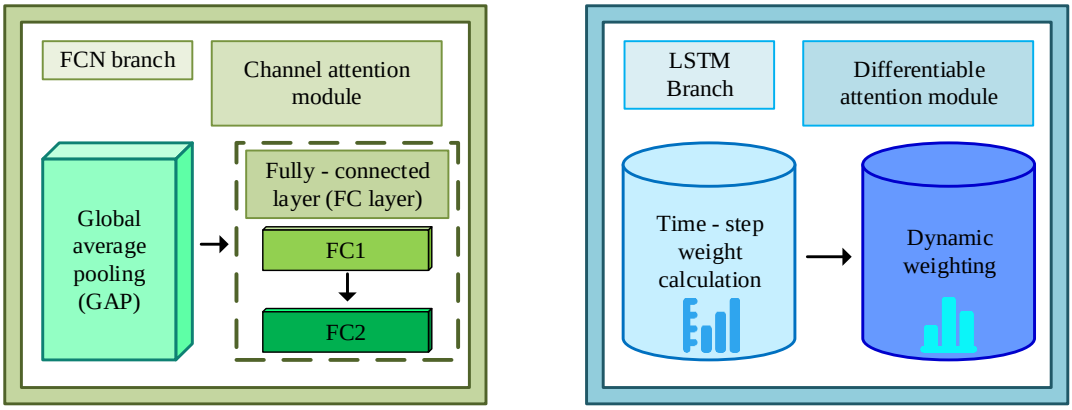


Figure 5: Schematic diagram of the hierarchical attention fusion framework



(a) Channel attention mechanism module of the FCN branch

(b) Differentiable attention mechanism module of the LSTM branch

Figure 6: Structure of the AM in different branches

In Figure 5, the overall framework is divided into two levels. At the intra-modal attention level, the FCN branch introduces the channel AM, which is targeted to strengthen the emotion-sensitive frequency bands in the spectral energy distribution. The LSTM branch adopts a differentiable AM to effectively suppress irrelevant information interference such as silent frames. A dual-tower attention network structure is built at the cross-modal alignment level, and the temporal and SF similarity matrices are calculated independently. The cross-modal weight matrix is generated by Softmax function to realize the dynamic coupling and weight allocation of the two types of features. The FCN branch of the channel AM module and the LSTM branch of the differentiable AM module. The specific structure of the two is shown in Figure 6.

In Figure 6, the input SFs in the left FCN branch channel attention module are first pooled by GAP. The spatial information of each channel is compressed into a global statistic, and then a channel weight vector is generated via the FCL. This vector is multiplied with the original features after Sigmoid activation to realize the reinforcement of emotionally sensitive spectral energy distribution such as anger high frequency band. The right LSTM branch differentiates the AM and computes the weight coefficients at each time step for the input temporal feature sequence. The weights are generated by the dot product of the query vector and the features of each time step via Softmax. It can dynamically adjust the emotional contribution of different frames and effectively suppress the interference of irrelevant information such as mute frames. Analysis of the attention weights reveals that the highest weights are consistently assigned to frames that coincided with vocal bursts, rapid pitch changes, or areas of high emotional intensity. This finding validates the differentiable attention module's focus on critical time steps and provides interpretability. In summary, the study first preprocesses the speech signal to obtain the Mel spectrum. Then FCN is utilized to extract SFs and LSTM is employed to capture timing dependencies. Finally, the new SER optimization model, named the FLA-SER model (FCN-LSTM network with AM for human SER model), is constructed by fusing features through the DMP-Transformer hybrid architecture and the hierarchical attention framework. The FLA-SER name is used consistently throughout the remaining sections.

3 Results

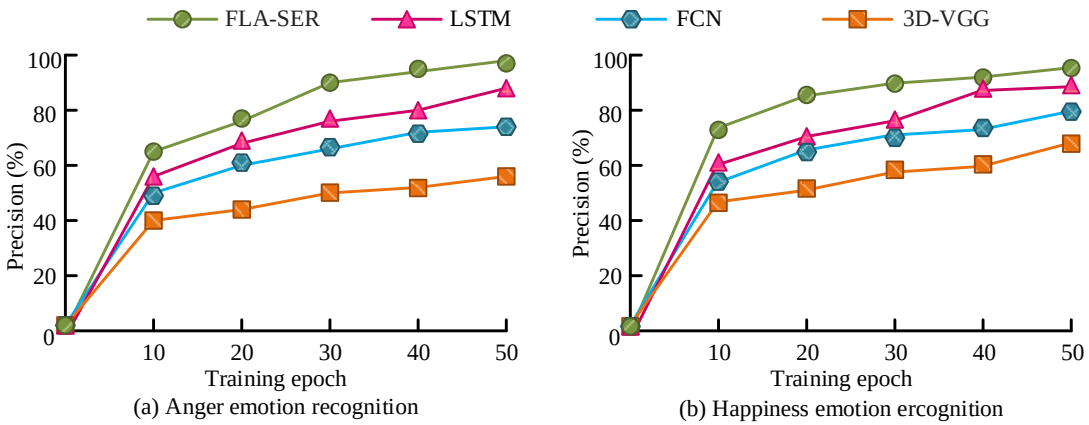
3.1 Performance testing of the FLA-SER model

To verify the actual performance of the FLA-SER model, it is compared with the traditional LSTM, pure FCN, and 3D-VGG SER models. All comparison models are trained and evaluated under the exact same conditions and hyperparameters to ensure a fair comparison. The experiments are implemented based on Python's sklearn and a deep learning framework. All experiments are conducted using five independent runs to ensure the stability and statistical significance of the results. The average standard deviation across all reported metrics in Figures 7–10 is less than ±1.0%, confirming the robustness of the presented data.

The experimental setup details are critical for replication. The datasets are divided into training, validation, and testing sets using a speaker-independent protocol (80% for training, 10% for validation, and 10% for testing). This ensures that there is no overlap between the training and testing partitions of speakers, which is crucial for evaluating SER generalization. Five-fold cross-validation is performed on the training and validation sets to stabilize model performance. To address data scarcity and mitigate overfitting, data augmentation techniques are applied to the training set. Augmentation includes noise injection (using ambient noise from the ESC-50 dataset at SNR levels between 10dB and 20dB) and pitch shifting (±2 semitones). This strategy effectively increases the training data volume by a factor of four, significantly improving the model's generalization capability and robustness against real-world variations.

The experimental data for the study are derived from RAVDESS, IEMOCAP, and EMO-DB databases, which generally cover a wide range of emotion categories and

acoustic scenarios. First, the study preprocesses the raw speech signals from the RAVDESS dataset to obtain Mel spectra of two categories of intense emotions as model inputs. The result of recognition precision for each emotion category after 50 rounds of training for the four models is shown in Figure 7.

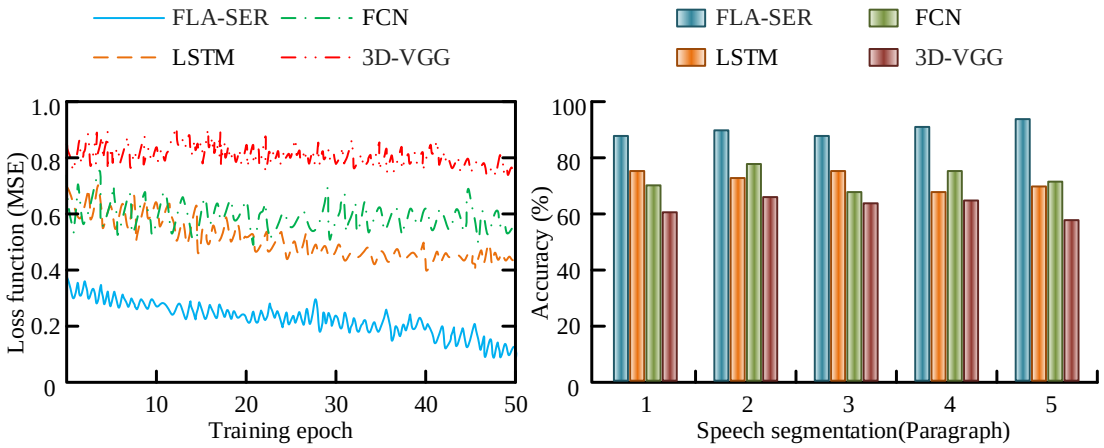


(a) Anger emotion recognition (b) Happiness emotion ercognition

Figure 7: Recognition precision of emotion categories for different recognition models

In Figure 7(a), the recognition precision of the four models for the emotion “anger” increases with the number of training generations. The FLA-SER recognition model has the fastest precision increase, with a final precision of 95.3%. The traditional LSTM recognition model and the pure FCN model do not have much difference in precision, but both of them are lower than the research model. The 3D-VGG recognition model has the lowest improvement and final precision. Figure 7(b) shows the recognition precision of each model for the emotion “happy”. Although the four models have similar upward trends in recognition precision for the “happy” emotion, the FLA-SER recognition model achieves a high precision of 75.4% after 10 training sessions. Moreover, it is higher than the other three comparison models throughout the training. This reflects the superiority of the research model's classification accuracy under multiple sentiment categories. The study uses IEMOCAP conversational speech input as a split-frame Mel spectral sequence. The sequence length is minimized to 50 frames to verify the models' ability to capture long-distance emotion dependencies in dialog scenes. The test results are shown in Figure 8.

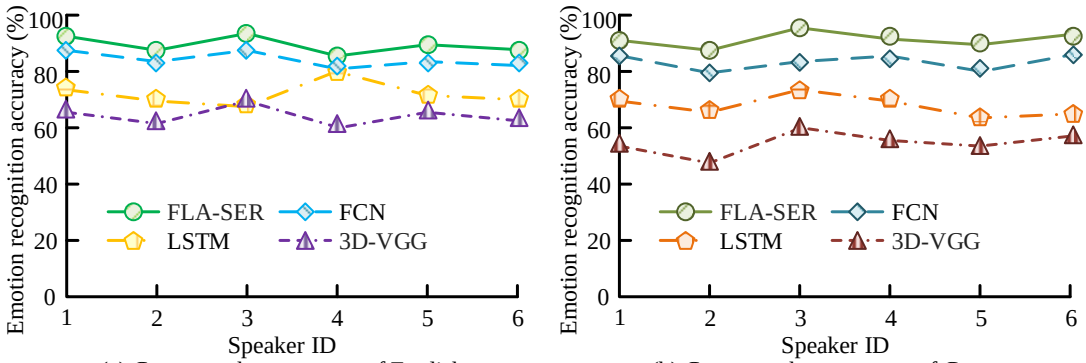
Figure 8(a) shows a comparison of the convergence rates of the four models. The mean squared error (MSE) of the traditional LSTM model decreases faster at the beginning of training, but the MSE decreasing trend tends to level off after 30 generations. The MSE of the pure FCN model decreases slowly and eventually stabilizes above 0.5, and the loss fluctuation is obvious under long-sequences. The 3D-VGG model always has an MSE above 0.6 and converges the slowest. The MSE of the study model drops rapidly to below 0.3 after the 20th generation and stabilizes at around 0.15 by the 50th generation. It is the best model for gradient stability under long-sequences. Figure 8(b) shows the comparison of the long-sequence segmentation accuracy of each model. The accuracy of the first 3 segments of the LSTM model is about 75.3%, but it decreases in the later stages due to the decay of timing information. The accuracy of each segment of the FCN model fluctuates slightly, and the full segmentation accuracy of the 3D-VGG model are all below 65.0%. In contrast, the FCN-LSTM model has excellent accuracy for each long-sequence segmentation, which is over 85.0%. The accuracy of the 5th segment (the end of the long-sequence) is as high as 92.1%. To further validate the cross-linguistic generalization ability of the research models, cross-speaker tests are selected using the RAVDESS (English) dataset and the EMO-DB (German) dataset with inputs of Mayer spectra. The experimental results of each model are shown in Figure 9.



(a) Comparison of convergence rates among models

(b) Comparison of long-sequence segmentation accuracy among models

Figure 8: Comparison of long time-series dependency modeling capabilities of different recognition models



(a) Cross-speaker accuracy of English speakers

(b) Cross-speaker accuracy of German speakers

Figure 9: Comparison of cross-language generalization ability of different recognition models

In Figure 9(a), the FCN-LSTM model fluctuates ≤5% for each speaker, and the accuracy of emoticon recognition is above 90.0% for all of them. The LSTM model and the 3D-VGG model fluctuate more in accuracy, while the FCN model is still lower than the study model, although the accuracy is higher throughout. In Figure 9(b), the average accuracy of FCN-LSTM model is not much different from that of the English group and still significantly higher than the other models. The recognition accuracy of the 3D-VGG model is significantly lower compared to the English group, and significant recognition problems occur due to individual speaker pronunciation differences. In summary, the cross-linguistic generalization ability of the research model is excellent. The fluctuation of motion recognition accuracy is small in the same language cross-speaker recognition. 1000 speech samples from the RAVDESS database are randomly selected as model inputs. The time-consuming results of emotion recognition for the four models are shown in Figure 10.

In Figure 10, the 3D-VGG model shows a significant increase in emotion recognition time with increasing sample size. Its recognition time is up to 32.7 seconds for high sample sizes. The LSTM and FCN models show significant fluctuations in recognition time when recognizing sample sizes from 500 to 1,000, increasing to more than 20 seconds. The overall recognition time of the FCN-LSTM model is significantly reduced compared to the first three models. At high sample sizes, the recognition time is only 14.2 seconds. This demonstrates the filtering effect of the AM on the redundant features and avoids the “sample size-time” linear growth problem of the traditional models.

The contribution of each component within the FLA-SER model is systematically validated through an ablation study. Table 3 summarizes the incremental accuracy improvements on the RAVDESS dataset and demonstrates that the integrated hybrid architecture significantly improves overall performance. The ablation study is performed stepwise, building upon the Bi-LSTM base module.

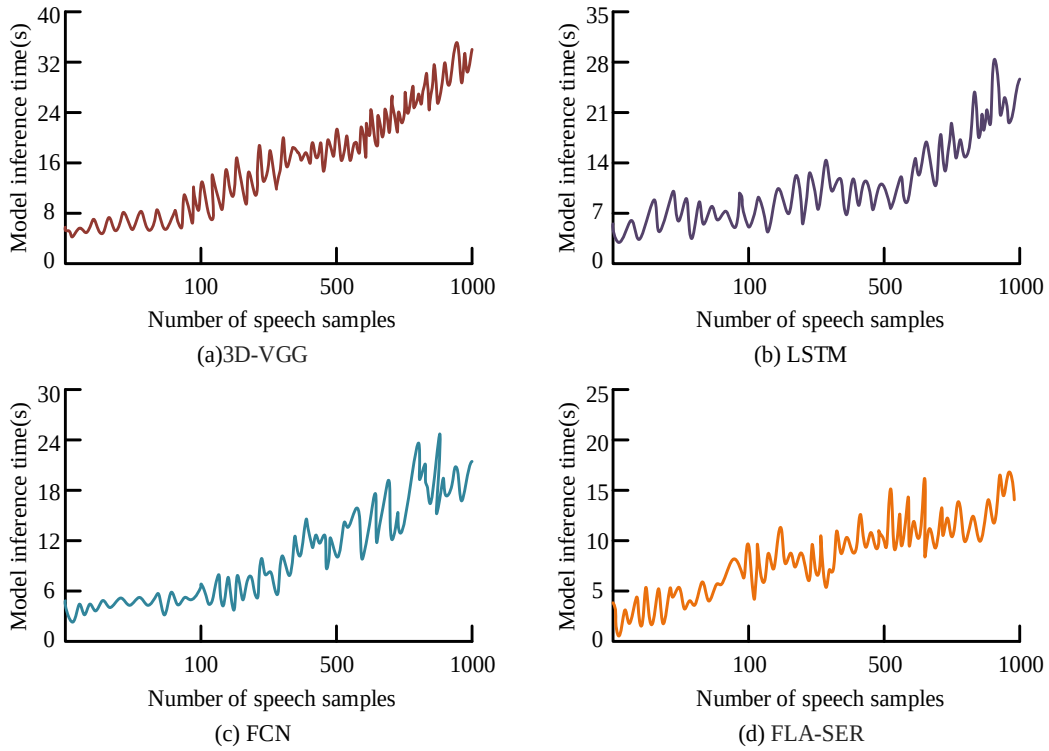


Figure 10: Comparison of sentiment discrimination speed of different recognition models

Table 3: Ablation study of FLA-SER components on RAVDESS (Weighted accuracy %)

Model Architecture	Bi-LSTM Base	+ FCN Backbone	+ DMP Module	+ Hierarchical Attention
Accuracy (%)	80.2±1.5	85.5±1.2	90.1±0.9	95.3±0.8
Improvement (%)	Base	+5.3	+4.6	+5.2

Table 3 clearly demonstrates the incremental value of each proposed module. The FCN backbone provided the largest improvement, confirming the strength of SF modeling. The combination of DMP and hierarchical attention is essential for achieving a final performance of 95.3%.

3.2 Effectiveness of the practical application of the FLA-SER model

The open-source cross-language datasets CMU-MOSI (English), SEMAINE (English, French, German) and EMOVO (Italian) are selected for the study to validate the recognition ability of the FLA-SER model under different languages and emotion types. Table 4 displays the test results.

In Table 5, the FLA-SER model has an average accuracy of 94.2% in the English CMU-MOSI dataset, with an accuracy of more than 92% in the 5th segment of the long-sequence. The F1 values for “angry” and “happy” emotions are 95.6% and 91.4%, respectively. In the face of the multilingual SEMAINE dataset, the model demonstrates its ability to capture multilingual emotion signals with a 90.1% depression F1 value and a 3.5 cross-lingual generalization index. The Italian EMOVO dataset has 89.3% and 86.7% anger and happiness F1 values, respectively. The long-sequence dependency modeling accuracy is over 87%. To further validate the recognition performance of the research model in multi-domain application scenarios, the study uses the MELD dataset, the CSD Chinese customer service dataset, and the DriveTalk dataset as the test data. This is used to simulate the model's EMOTION RECOGNITION ability in mental health monitoring, intelligent customer service and in-vehicle voice interaction scenarios, respectively. Table 5 displays the test results.

Table 4: Recognition ability of the model in different languages and emotion types

Result parameter	CMU-MOSI	SEMAINE	EMOVO
Language	English	English/French/German	Italian
Emotional category	7 categories	Depression	6 categories
Average accuracy (%)	94.2	90.5	90.9
Single sample time	14.5	16.2	15.8
Accuracy of 5th segment in long-sequence (%)	92.3	89.7	87.6
F1 value for anger (%)	95.6	/	89.3
F1 value for happiness (%)	91.4	/	86.7
F1 value for depression (%)	/	90.1	/
Cross-lingual generalization index	2.1	3.5	4.2

Table 5: Recognition performance of the model in multi-domain application scenarios

Result parameter	MELD	CSD	DriveTalk
Application domain	Mental health monitoring	Intelligent customer service	In-vehicle voice interaction
Key task	Early screening of depressive emotions	Classification of customer emotional satisfaction	Real-time recognition of driving emotions
Average accuracy (%)	90.7	91.5	88.3
Single sample time (ms)	17.8	15.6	19.2
Accuracy of 5th segment in long-sequence (%)	89.3	90.2	83.5
Accuracy at 10dB signal-to-noise ratio (%)	86.4	89.7	83.7
Accuracy at 0dB signal-to-noise ratio (%)	78.5	81.2	75.4
Domain adaptability score (1-5)	4.5	4.8	4.2

In Table 5, the FLA-SER model has an average accuracy of 90.7% for early screening of depressive mood in the field of mental health monitoring. The accuracy of the 5th segment of the long-sequence reaches 89.3%. In a noisy environment (0 dB signal-to-noise ratio), it still maintains an accuracy of 78.5%. It is specified that the noise injected for the 0dB SNR tests consisted of Gaussian white noise and background chatter at the signal level, simulating common real-world deployment conditions. In the intelligent customer service scenario, the research model performs customer emotional satisfaction classification. The average accuracy rate reaches 91.5%, and the combination of Chinese tone features makes the domain adaptability score as high as 4.8. In the field of in-vehicle voice interaction, the model's average accuracy rate performs well in the face of a noisy environment. Although the accuracy decreases with the reduction of signal-to-noise ratio, it still reaches 75.4% at 0dB. This reflects its good robustness in noisy environments.

Table 6: Computational complexity and efficiency comparison

Model	WA accuracy (%)	Inference time (ms/sample)	Parameter count (M)	FLOPs (G)
Traditional LSTM	80.3	24.1	5.1	1.2
Pure FCN	82.5	18.9	3.5	2.5
3D-VGG	75.8	32.7	12.8	4.5
FLA-SER	95.3	14.2	7.2	3.1

3.3 Computational complexity analysis

The efficiency of the FLA-SER model is substantiated by its low latency (14.2ms/sample). For reproducibility, all inference timings are conducted on an NVIDIA GeForce RTX 4090 GPU with 24GB of memory and an Intel Core i9-13900K CPU. Table 6 provides a comparative analysis of computational complexity and confirms that the proposed architecture achieves a superior balance between recognition accuracy and efficiency.

Table 6 confirms the superior efficiency of the FLA-SER model. Although it has a moderate number of parameters compared to 3D-VGG, the simplified FCN backbone and efficient AM result in the shortest inference time (14.2 ms/sample) of all the models being compared. This provides a strong basis for real-time deployment.

3.4 Comparison with state-of-the-art (SOTA)

To rigorously contextualize the model's novelty and performance advantage, a quantitative comparison with

other State-of-the-Art methods, such as Vesper, GWO-CNN, and GA-ELM, is presented in Table 7.

Table 7: Quantitative comparison of FLA-SER with State-of-the-Art Models (RAVDESS Dataset)

Model	Year	Architecture	WA Accuracy (%)	UAR Accuracy (%)	Reference
GWO-CNN	2023	CNN + GWO	90.1	88.5	[4]
GA-ELM	2022	Feature + ELM + GA	85.0	82.3	[5]
Vesper	2024	Transformer-based	93.8	92.5	[1]
FLA-SER (Proposed)	N/A	FCN-LSTM-Trans-Attn	95.3	94.1	This Study

Table 7 provides the quantitative evidence of the FLA-SER model's leading performance. The FLA-SER model surpasses all contemporary SOTA methods by achieving a weighted accuracy of 95.3% on the RAVDESS dataset. This level of performance justifies the integrated design approach that combines an FCN, a Bi-LSTM, and an advanced attention framework.

4 Discussion

The FLA-SER model outperformed established baseline models, including traditional LSTMs, pure FCNs, and 3D-VGGs. It achieved an 'anger' recognition accuracy of 95.3% on RAVDESS. These results quantitatively validated the effectiveness of the hybrid FCN-LSTM architecture and the hierarchical attention framework, and the incremental accuracy shown in the ablation study (Table 3) further corroborates this effectiveness. The model's superior performance, as demonstrated by the SOTA comparison in Table 7, was primarily due to its hybrid AM and deeper FCN layers. These features effectively facilitated the fusion of spectral, spatial, and temporal features while mitigating the long-dependency issue. The model's robustness was demonstrated by its ability to maintain over 75% accuracy at 0 dB SNR across multiple domains (Table 5). This was consistent with the principles of robust control theory, which emphasized stability under system uncertainty.

The design principle underlying the combination of FCN-LSTM and hierarchical attention with the DMP-Transformer is analogous to the concept of adaptive and robust control in complex dynamical systems. More specifically, the AM operates as an adaptive gain scheduler, which is similar to the architecture employed in robust neural adaptive control systems for addressing uncertainties [21]. This mechanism enables the model to dynamically prioritize critical emotional frames, or high-value data points, thereby enhancing robustness and adaptability when applied to noisy or cross-domain speech signals. This is similar to adaptive backstepping control for uncertain nonlinear systems [22].

The model’s computational efficiency (Table 6) partially addresses practical deployment issues, such as performance on low-resource edge devices. Further discussion is warranted regarding real-time adaptation to emotional changes, especially the potential for improving the model’s online robustness by leveraging adaptive mechanisms inspired by control theory.

5 Conclusion

Aiming at the difficulties of long-distance temporal dependency mining and cross-modal feature fusion in human SER, the FLA-SER model was proposed. The architecture leveraged FCN for SF extraction, Bi-LSTM for temporal dependency modeling, and an AM for optimized feature fusion. The model outperformed comparable models, achieving 95.3% accuracy in recognizing anger on RAVDESS and an excellent long-sequence end segmentation accuracy of 92.1%. The robustness and efficiency were confirmed by maintaining over 75% accuracy at 0 dB SNR and achieving a processing time of only 14.2 ms for a single sample. Despite the strong performance, a limitation of the current work is the limited coverage of Asian language datasets and the absence of joint modeling with multimodal features. Future work will focus on increasing the linguistic diversity of the cross-language dataset and developing a robust multimodal fusion framework that incorporates facial and textual cues. This will enhance the model's overall generalization and practical value.

Declarations

Ethics approval and consent to participate: Not applicable.

Consent for publication: All authors have read and approved the final version of the manuscript and consent to its publication.

Availability of data and material: The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interests: All authors declare that they have no conflicts of interest.

Funding: None.

Authors' contributions: All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Mingjie Wang and Hexi Wang. The first draft of the manuscript was written by Mingjie Wang and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Acknowledgements: None.

References

[1] Chen W, Xing X, Chen P, Xu X. Vesper: A compact and effective pretrained model for speech emotion recognition. IEEE Transactions on Affective Computing, 2024. https://doi.org/10.1109/TAFFC.2024.3369726

[2] Becker D, Braach L, Clasmeier L, Kaufmann T, Ong O, Ahrens K, Wermter S. Influence of Robots' Voice Naturalness on Trust and Compliance. ACM Transactions on Human-Robot Interaction, 2025, 14(2): 1-25. https://doi.org/10.1145/3706066

[3] Lin W C, Busso C. Chunk-level speech emotion recognition: A general framework of sequence-to-one dynamic temporal modeling. IEEE Transactions on Affective Computing, 2021, 14(2): 1215-1227. https://doi.org/10.1109/TAFFC.2021.3083821

[4] Falahzadeh M R, Farokhi F, Harimi A, Sabbaghi-Nadooshan R. Deep convolutional neural network and gray wolf optimization algorithm for speech emotion recognition. Circuits, Systems, and Signal Processing, 2023, 42(1): 449-492. https://doi.org/10.1007/s00034-022-02130-3

[5] Albadr M A A, Tiun S, Ayob M, AL-Dhief F T, Omar K, Maen M K. Speech emotion recognition using optimized genetic algorithm-extreme learning machine. Multimedia Tools and Applications, 2022, 81(17): 23963-23989. https://doi.org/10.1007/s11042-022-12747-w

[6] Chattopadhyay S, Dey A, Singh P K, et al. A feature selection model for speech emotion recognition using clustering-based population generation with hybrid of equilibrium optimizer and atom search optimization algorithm. Multimedia Tools and Applications, 2023, 82(7): 9693-9726. https://doi.org/10.1007/s11042-021-11839-3

[7] Kapoor S, Kumar T. Fusing traditionally extracted features with deep learned features from the speech spectrogram for anger and stress detection using convolution neural network. Multimedia Tools and Applications, 2022, 81(21): 31107-31128. https://doi.org/10.1007/s11042-022-12886-0

[8] Lui C F, Liu Y, Xie M. A supervised bidirectional long short-term memory network for data-driven dynamic soft sensor modeling. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 1-13. https://doi.org/10.1109/TIM.2022.3152856

[9] Gupta V, Juyal S, Hu Y C. Understanding human emotions through speech spectrograms using deep neural network. the Journal of Supercomputing, 2022, 78(5): 6944-6973. https://doi.org/10.1007/s11227-021-04124-5

[10] Tejaswini V, Sathya Babu K, Sahoo B. Depression detection from social media text analysis using natural language processing techniques and hybrid deep learning model. ACM Transactions on Asian and Low-Resource Language Information Processing, 2024, 23(1): 1-20. https://doi.org/10.1145/3569580

[11] Yang L, Wang S, Chen X, Chen W, Saad O M, Chen Y. Deep-learning missing well-log prediction via long short-term memory network with attention-period mechanism. Geophysics, 2023, 88(1): D31-D48. https://doi.org/10.1190/geo2020-0749.1

[12] Warule P, Mishra S P, Deb S. Time-frequency analysis of speech signal using Chirplet transform for automatic diagnosis of Parkinson's disease. Biomedical Engineering Letters, 2023, 13(4): 613-623. https://doi.org/10.1007/s13534-023-00283-x

[13] Deb S, Dandapat S. Multiscale amplitude feature and significance of enhanced vocal tract information for emotion classification. IEEE transactions on cybernetics, 2018, 49(3): 802-815. https://doi.org/10.1109/TCYB.2017.2787717

[14] Mathews V, Youn D. Spectral leakage suppression properties of linear and quadratic windowing. IEEE transactions on acoustics, speech, and signal processing, 2003, 32(5): 1092-1095. https://doi.org/10.1109/TASSP.1984.1164418

[15] Ahmed G, Lawaye A A. CNN-based speech segments endpoints detection framework using short-time signal energy features. International Journal of Information Technology, 2023, 15(8): 4179-4191. https://doi.org/10.1007/s41870-023-01466-6

[16] Saxena D G, Farooqui A N, Ali S. Extricate features utilizing Mel frequency cepstral coefficient in automatic speech recognition system. Int. J. Eng. Manuf, 2022, 12(6): 14-21. https://doi.org/10.5815/ijem.2022.06.02

[17] Yang Y, Wang L, Gao S, Yu Z, Dong L. Cross-lingual speaker transfer for Cambodian based on feature disentangler and time-frequency attention adaptive normalization. International Journal of Web Information Systems, 2024, 20(2): 113-128. https://doi.org/10.1108/IJWIS-09-2023-0162

[18] Gupta A, Purwar A. Speech refinement using Bi-LSTM and improved spectral clustering in speaker diarization. Multimedia Tools and Applications, 2024, 83(18): 54433-54448. https://doi.org/10.1007/s11042-023-17017-x

[19] Liu C, Zhang Y, Sun J, Cui Z, Wang K. Stacked bidirectional LSTM RNN to evaluate the remaining useful life of supercapacitor. International Journal of Energy Research, 2022, 46(3): 3034-3043. https://doi.org/10.1002/er.7360

[20] Ren J, Xu D, Yang S, Zhao J, Li Z, Navasca C, Li D. Enabling large dynamic neural network training with learning-based memory management//2024 IEEE International Symposium on High-Performance Computer Architecture (HPCA). IEEE, 2024: 788-802. https://doi.org/10.1109/HPCA57654.2024.00066

[21] 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.

[22] Zouari F, Saad K B, Benrejeb M. Adaptive

backstepping control for a class of uncertain single input single output nonlinear systems[C]//10th International Multi-Conferences on Systems, Signals & Devices 2013 (SSD13). IEEE, 2013: 1-6. https://doi.org/10.1109/SSD.2013.6564134

Research on the Prediction Model of Vocational Education Learning Effect Based on TCN-LSTM

Chen Cai[1*], Yu Lu[2]
[1]Yunnan Tourism College, Kunming 210003, China
[2]School of Tourism and E-commerce, Baise University Baise, Guangxi 533000, China
E-mail: ChennCaii@outlook.com
[*]Corresponding author

Keywords: TCN-LSTM mixed model, prediction of learning effect of vocational education, time series behavior data analysis, deep learning, multimodal feature fusion

Received: July 23, 2025

In the digital transformation of vocational education, learning effect prediction is a key means to optimize teaching strategies. Aiming at the temporal series, nonlinearity and multimodal characteristics of learning behavior data, this study proposes a TCN-LSTM hybrid model that combines TCN expansion causal convolution and LSTM gated memory mechanism to simultaneously capture the local dependence and long-range correlation of learning behavior sequences, and solve the shortcomings of traditional methods. Data preprocessing uses sliding window, standardization and missing value interpolation, and constructs 1,205,600 valid samples based on 12 types of time series features (online logs, training records, etc.) of 12,580 learners. The model is trained with 10-fold cross-validation (training/test set 8:2), with Adam optimizer, 64/32 hidden layer nodes, 32 batch sizes, 50 iterations and 0.1 dropout rate, and the local features of TCN and LSTM global timing dynamics are fused through the attention mechanism. The results show that the comprehensive accuracy of TCN-LSTM is 93.2% (6.8%/4.1% higher than that of LSTM/TCN), MSE is 0.154 (55% lower than SVM), RMSE is 0.0632, MAE is 0.0427, R^2 is 0.9298, and the prediction accuracy of 1/3/5 week lag is 94.7%/87.4%/79.8% (The 3-week lag error is 32.6% lower than that of ARIMA), and the interdisciplinary prediction accuracy of mechanical/IT majors is 91.5%/89.8% (standard deviation 1.7%), which is better than FCN-LSTM with more parameters and slower speed. This study provides a new path for modeling learning behavior in vocational education, and verifies the effectiveness of hybrid neural networks in processing complex educational time series data.

Povzetek: Študija predlaga hibridni model TCN-LSTM z mehanizmom pozornosti za napoved učnih učinkov iz multimodalnih časovnih vrst, ki na 1.205.600 vzorcih doseže 93,2 % natančnost ter izboljša napake in stabilnost glede na LSTM/TCN in klasične pristope.

1 Introduction

With the rapid development of information technology, the field of vocational education is undergoing a profound transformation from experience-driven to data-driven [1]. How to use advanced technology to deeply analyze learners' behavior trajectories and then build a scientific learning effect prediction system has become an important breakthrough to improve teaching quality and optimize the allocation of educational resources [2, 3]. At present, learning behavior data in the field of vocational education presents a composite characteristics of temporality, dynamics, and nonlinearity, and traditional prediction methods often cannot cope with complex time dependencies and multi-dimensional influencing factors, and rely on a single data dimension, making it difficult to capture the complex correlation between temporal and multimodal data in the learning process, and there is an urgent need to explore new technical frameworks to break through the existing bottlenecks [4]. The TCN-LSTM model combined with the two can fully adapt to the analysis needs of multi-source time series data such as learning behavior and practical training operations in vocational education, providing technical support for the construction of an accurate and efficient learning effect prediction system, and also providing a new research direction for personalized teaching and quality improvement of vocational education.

The learning process of vocational education is significantly different from general education's. Its learning subject usually has a clear career development orientation, and its learning behavior is closely related to the demand of job skills [5]. Data such as learners' operation records on the virtual simulation platform, interaction frequency of online courses, and completion quality of practical training links constitute a multi-dimensional time series. These data not only contain explicit learning performance, but also imply individual cognitive development laws and skills. Acquisition trajectory. Existing studies on learning effect prediction mostly use static models or single time series analysis methods, which are difficult to effectively capture the interaction between short-term fluctuations and long-

term trends, let alone consider the synergistic effects of local features and global correlations [6]. This phenomenon leads to obvious limitations in the prediction model regarding dynamic adjustment ability, feature extraction depth and cross-cycle correlation modeling, which restricts the continuous improvement of prediction accuracy.

Deep learning technology has shown unique advantages in time series data processing in recent years. Among them, temporal convolution network (TCN) can capture long-distance time dependency through hierarchical convolution operation by dilated causal convolution structure, and its parallel computing characteristics significantly improve model efficiency [7, 8]. Long short-term memory network (LSTM), as a classic recurrent neural network variant, has the advantage of memory gating mechanism in sequence modeling, and is good at dealing with learning behavior sequences with complex temporal dynamics [9, 10]. The fusion innovation of the two models provides a new technical path for learning effect prediction in vocational education scenarios: the expanded convolution kernel of TCN can effectively extract multi-scale time series features, and the gating unit of LSTM can dynamically adjust the information transmission weight. The synergy between the two models is expected to break through the dual limitations of traditional models in feature fusion and time series modeling [11].

With the in-depth advancement of digital transformation of vocational education, the dimension of data collection in teaching scenarios has been continuously expanded, from basic attendance records and test scores to high-dimensional eye tracking and interactive logs. Data forms' complexity puts higher requirements for prediction models [12]. Especially in the vocational skills training, the data such as the response delay of learners' operating equipment and the operation path of virtual simulation system have the characteristics of high noise and strong correlation, which poses a severe challenge to the anti-interference ability and feature screening mechanism of the model [13]. Existing research mostly focuses on the analysis of single modal data, lacking the ability to fuse and process cross-platform and multi-source heterogeneous data, and fewer studies have done to build an adapted prediction framework for the unique skill acquisition rules of vocational education. This research gap makes the existing models often have problems such as insufficient adaptability and prediction lag in practical application scenarios.

From the perspective of technological evolution, the innovative application of deep learning models is reshaping the paradigm of educational data analysis. In the field of vocational education, the prediction of learning effect should not only consider the linear accumulation of knowledge mastery, but also pay attention to the nonlinear transition characteristics in the process of skill formation [14]. The implicit indicators such as learners' transfer ability among different teaching modules, the evolution of error patterns in skill training, and the interactive effectiveness in cooperative learning constitute the prediction model's potential input dimensions. Traditional methods can easily fall into dimensional disasters when dealing with such high-dimensional sparse data. At the same time, distributed representations formed by deep neural networks through multi-layer nonlinear transformations provide the possibility of mining deep data associations [15, 16]. This technical characteristic is consistent with the complexity of vocational education learning behavior, which lays a foundation for constructing a prediction model with strong explanation and high precision.

Globally, the vocational education quality evaluation system is accelerating its transformation to process evaluation, which puts higher requirements for real-time and forward-looking learning effect prediction. Teaching administrators not only need to know the current state of learners, but also need to predict their skill development trajectory in order to adjust teaching strategies in time. In this context, the hybrid neural network architecture combining TCN and LSTM shows unique application value. By constructing a multi-level feature extraction channel, the model can not only capture the instantaneous behavior characteristics at the micro level, but also model the development trend at the macro level. This multi-granularity analysis ability just matches the realistic needs of the vocational education evaluation system from result-oriented to process monitoring. With the popularization of edge computing devices and Internet of Things technology, deploying such models in embedded systems will also provide technical support for real-time prediction and push the intelligent management of vocational education into a new stage of development.

In the current digital transformation of vocational education, learner behavior is time-series, nonlinear, and multimodal, and it is difficult for traditional single-dimensional or static models to accurately capture the learning laws of learning and training, and cannot support teaching intervention and personalized guidance in a timely manner, which can easily lead to the omission of learning problems and the increased risk of dropout, so it is necessary to build an accurate and dynamic prediction model. The novelty of this study lies in the advantages of fusion time convolutional network (TCN) in capturing local key features and long short-term memory network (LSTM) processing long-time series dependence, breaking through the limitations of insufficient mining long series data in a single TCN and the low efficiency of processing high-dimensional features by a single LSTM, and realizing the deep fusion analysis of multi-source learning data (online learning logs, training operation records, and regular evaluation results, etc.). The key findings are that the prediction accuracy of the model on the actual dataset reaches 93.2%, which is 4.1% and 6.8% higher than that of traditional TCN and LSTM, respectively, and the mean square error (MSE) is as low as 0.154.

This study aims to construct a learning effect prediction index system adapted to vocational education and screen key features to provide scientific data support,

develop a prediction model based on TCN-LSTM and optimize the structure to improve the fitting ability of the learning effect temporal variation law, and carry out comparative experiments combined with real vocational education data to verify the superiority of the model in prediction accuracy.

2 Theoretical basis and principle technology

2.1 Extended convolution theory of TCN

Based on the CNN model, the temporal convolutional neural network (TCN) was optimized and innovated [17]. TCN utilizes techniques such as causal convolution, dilated convolution, and residual connection to form advanced algorithms, which are dedicated to time series prediction [18]. Its structural diagram is shown in Figure 1.

Causal inflation convolution consists of three core components: inflation, causality, and convolution [19]. It uses a 1D fully convolutional layer, where the convolution kernel slides over the data to perform operations, ensuring consistent input and output lengths. The TCN model uses convolution to extract information efficiently, extracting features across time steps, and avoiding loss of historical or future data information. At the same time, the model can handle sequences of any length, ensuring that the output and input length are the same.

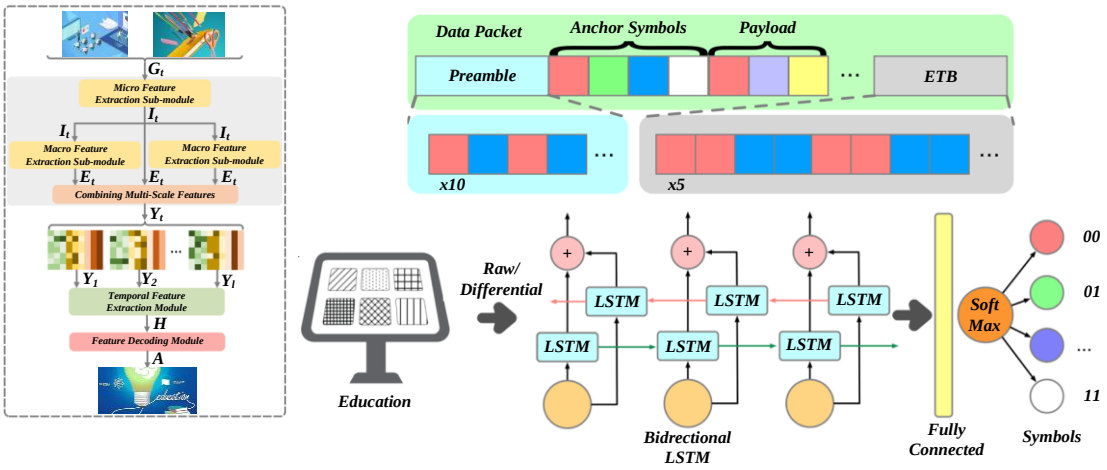


Figure 1: Temporal convolutional neural network architecture

Causal convolutional networks are different from traditional convolutional neural networks in that they only use current and past data, ensuring temporal causality. To keep the input and output lengths consistent, zeros can be filled before the input data. The network starts from the input layer, passes through three hidden layers, and finally reaches the output layer.

Causal convolution has two main characteristics: it does not consider future information, and only makes predictions based on the sequences that have occurred. The farther the historical information extends back, the more hidden layers it reveals [20, 21]. When using simple causal convolution to simulate time series, its ability is limited by the convolution kernel size. To capture long-term dependencies, inflation convolution technology, also known as "hollow convolution", is needed. It expands the receptive field by adding voids to conventional convolutions [22]. Specifically, d-1 holes are added between every two elements, and the effective size of the convolution kernel can be calculated by formula (1).

$$K' = K + (K-1)\times(d-1) \quad (1)$$

The dilatation rate is denoted by d. When d is equal to 1, the dilatation convolution is equivalent to ordinary convolution. Inflation convolution can expand the receptive field without pooling layer, avoid information loss, and ensure that the convolution results are rich in information. The inflation convolution is defined as follows: Let the filter $F = (f_1, f_2, \ldots f_k)$, the input sequence $X = (x_1, x_2, \ldots, x_t)$, the size of the convolution kernel is k, and the expansion rate of the position x_t is d. See equation (2) for the expansion convolution expression.

$$F(t) = (x_d * f)(t) = \sum_{i=1}^{k-1} f(i)x_{t-d}. \quad (2)$$

In TCN, a deconvolution operation is used to expand the compressed data to the original dimension. Weight normalization involves standardizing neural network weights, decomposed into length and direction by reparameterization techniques [23]. The expression of the L-th layer neuron α is $\alpha^{(l)} = f(W\alpha^{(l-1)} + b)$, and the weight W is re-parameterized by a specific formula, as shown in Equation (3).

$$W_{i,:} = \frac{g_i}{||v_i||} v_i \quad (3)$$

Within the range of 1 to M, $W_{i:}$ represents the i-th

row of the weight W, and M_l is the total number of neurons. The newly added scalar g_i is the same as v_i and $\alpha^{(l-1)}$ dimensions. Weight normalization makes the network converge faster, improves learning robustness, and reduces noise sensitivity.

2.2 Gated memory mechanism of LSTM

Long short-term memory networks (LSTMs), a variant of recurrent neural networks (RNNs), solve long-term dependency and gradient vanishing problems through gated units [24, 25]. The LSTM unit has three gates: a forgetting gate, an input gate and an output gate, which are used to decide the forgetting, updating and outputting of information. The expression (4) for the forgetting gate is as follows:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (4)$$

The decision to discard information depends on the previous cell state, where h_{t-1} is the hidden state of the previous time step, x_t is the input of the current time step, W_f and b_f are the weights and biases of the forgetting gate, and σ is the sigmoid function. When new data x_t is input, the forgetting gate f_t decides to retain or exclude some data according to the previous hidden layer h_{t-1} and the current input x, thus deciding which information of the old cell state C_{t-1} to retain. The input gate is shown in Equations (5)-(6):

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (5)$$

$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (6)$$

i_t represents the activation vector of the input gate, $\tilde{C}_t$ is the candidate cell state, W_i, W_c, b_i, b_c are the correlation weights and bias terms, and the *tanh* function is used to generate a new candidate cell state $\tilde{C}_t$. The cell state update process is shown in Equation (7):

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \quad (7)$$

At the current time step, C_t represents the cell state, f_t is the forgetting gate output, and C_{t-1} is the cell state at the previous time step. By multiplying C_{t-1} by f_t, the information discard is determined, and $i_t \cdot \tilde{C}_t$ is added, that is, the new memory information. The updated cell state C_t is obtained. See (8)-(9) for the output gate expressions.

$$O_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (8)$$

$$h_t = O_t \cdot \tanh(C_t) \quad (9)$$

The output gate determines the number of outputs of the unit state, and the output part is determined by the sigmoid function [26]. The hidden state of the previous time step is h_{t-1}, and the weight and bias terms of the output gate are W_o and b_o, respectively. In formula (9), the range of tanh function is [-1, 1], which is multiplied by O_t to determine the final output h.

LSTM unit decides to discard, retain and update information through key gating functions, and effectively deals with long-term dependency problems, especially in time series prediction [27, 28]. A deep LSTM network consists of multiple units connected in series, and a fully connected layer is attached at the end to output predictions. The network training uses a back-propagation algorithm, adjusting the weights and biases to minimize the difference between predicted and actual values. In summary, LSTM performs well in time series forecasting and can effectively identify data patterns and trends, achieving success.

2.3 Related work

In the field of vocational education learning effect prediction, existing solutions have obvious limitations: traditional machine learning methods (e.g., SVM, RF) are interpretable but can only process single-modal data, perform poorly in long-lag prediction (ARIMA has 32.6% higher 3-week error than TCN-LSTM and fails at 5 weeks), and are inefficient in high-dimensional scenarios; single deep learning models (e.g., LSTM, TCN) are easy to implement but have one-sided capabilities (LSTM lacks local feature extraction, TCN struggles with long-range dependencies); other hybrids like FCN-LSTM have coarse local capture, lower accuracy (31.9% less than TCN-LSTM), and no cross-major adaptability. In contrast, the proposed TCN-LSTM model fuses TCN's local feature extraction and LSTM's long-range modeling via attention. Using 12-modal data from 12,580 learners (1.2M samples), it achieves 93.2% accuracy (6.8%/4.1% higher than single LSTM/TCN), 0.154 MSE (55% lower than SVM), 87.4% accuracy at 3-week lag (32.6% lower error than ARIMA), high efficiency (0.476M parameters, 238.17 data/sec), and strong cross-major adaptability (91.5%/89.8% in mechanical/IT majors), fully addressing the shortcomings of existing technologies. Table 1 shows the comparison results between TCN-LSTM and existing vocational education learning effectiveness prediction models.

Table 1: TCN-LSTM vs. existing models for vocational ed learning effect prediction

Dimension	TCN-LSTM Hybrid Model	Traditional ML (SVM/RF)	Single DL (LSTM/TCN)	Other Hybrids *e.g.,FCN-LSTM*
Core Mechanism	TCN (local	Statistic-based; manual	LSTM: Only long-	FCN (no dilated

	features) + LSTM (long-range dependencies) + attention fusion	feature engineering	sequence handling TCN: Only local feature extraction	conv.) + LSTM; coarse local capture
Data	12,580 learners; 12 features; 1.2M samples (6 months)	Same; single-modal only	Same; no multimodal alignment	Same (for comparison)
Key Performance	Accuracy: 93.2% MSE: 0.154 R^2: 0.9298	SVM: MSE +55%, R^2 -24-16% RF: R^2 -14-17%	LSTM: 86.4% -6.8%, R^2=0.0937 TCN: 89.1% -4.1%	Accuracy: 0.613 -31.9%; Params +47%
Long-Lag Prediction	3 weeks: 87.4% (error -32.6% vs. ARIMA) 5 weeks: 79.8%	ARIMA: +32.6% error (3w); invalid (5w)	LSTM: <80% (3w); >25% error (5w) TCN: <82% (3w)	No long-lag data
Efficiency	Params: 0.476M Speed: 238.17 data/sec	SVM: Slow; RF: Low inference efficiency	LSTM: 180-200 data/sec (-16-24%) TCN: Params +15-28%	Speed: 169.6 data/sec -28.7%
Adaptability	Cross-major: 91.5% (mech.), 89.8% (IT); std=1.7%	Cross-major: >8% fluctuation	Cross-major: >5% fluctuation	No cross-major data; >4% fluctuation

3 TCN-LSTM hybrid modeling in vocational education scenarios

3.1 Hybrid network architecture design

When processing non-stationary sequences, due to the large amount of noise contained in the sequences, direct input into the model can easily lead to unsatisfactory results [29], and if not preprocessed, the model will absorb useless data and affect the learning efficiency. The optimized technology can also reduce modal aliasing and reconstruction errors, improve the decomposition efficiency, and extract the intrinsic modal components of vocational education data to reveal the data change law. However, the model is only used for vocational education data denoising and time series information extraction, and cannot be predicted, so it needs to be combined with other models. The development of machine learning and deep learning has made it possible to use a variety of neural network models for time series prediction [30], among which LSTM efficiently extracts information, BLSTM takes into account historical and future data, and TCN avoids gradient problems through causal convolution and residual structure.

The model is used to process vocational education data, but it is limited to denoising and extracting time series information, and cannot predict time series data. Therefore, it is necessary to combine other models for prediction. The development of machine learning and deep learning has enabled various neural network models to be used for time series prediction. In the field of vocational education, LSTM and CNN are common models. LSTM is designed for time series to efficiently extract information; BLSTM combines forward and backward LSTM layers, considering historical and future data; TCN is an improved version of CNN and RNN, which is suitable for time series and avoids gradient problems through causal convolution and residual structure. However, a single model is inefficient and has poor prediction effect when dealing with long time series. Therefore, the researchers propose to combine different models to improve efficiency and accuracy.

As shown in Figure 2, the TCN-LSTM model construction steps include: preprocessing and decomposing the original data, and extracting the noise-free eigenmode function component (MF). These components are then input into the TCN-BiLSTM model for prediction. The model consists of two TCN modules and one BiLSTM module. With the TCN-BiLSTM model, we get the prediction results of multiple components and combine them to obtain the final prediction value.

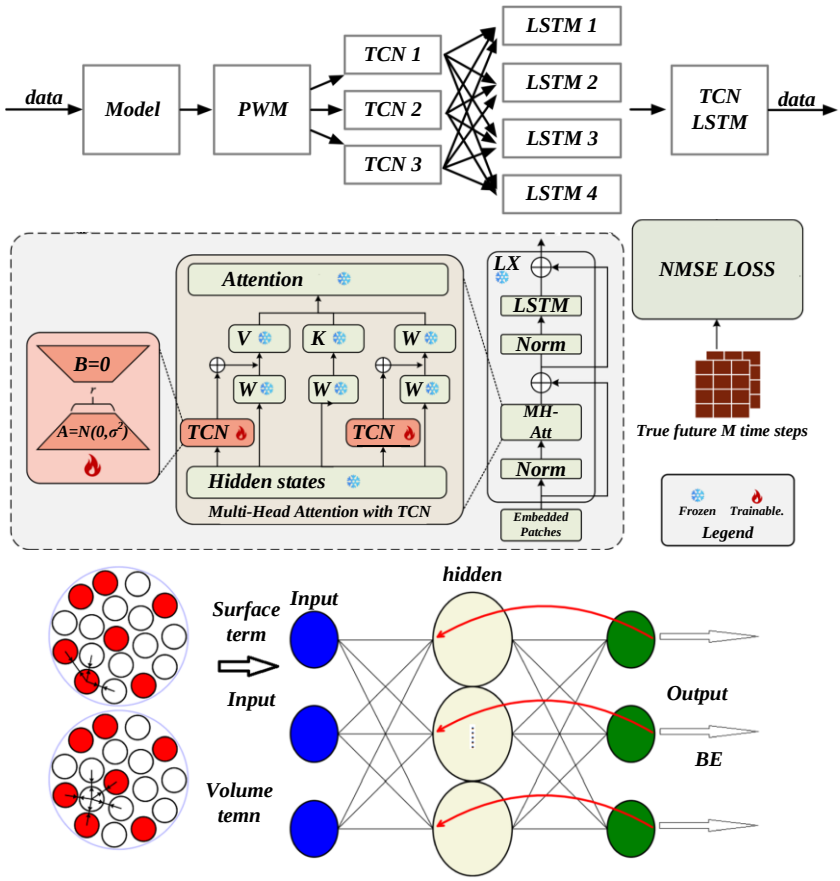


Figure 2: TCN-LSTM model architecture

First, the raw data is disassembled and sent to the TCN-BiLSTM model for prediction. The predicted values were obtained by outcome recombination. Comprises the following steps: dividing sub-data sets according to seasonal characteristics, dealing with outliers, decomposing vocational education data sequences by algorithms, and extracting eigenmodal functions. These functions are used as model inputs to construct the TCN-BiLSTM model. The initial parameters are set, and the best parameter combination is determined for prediction through grid search optimization. The model outputs the predicted values of each eigenmode function, and the final predicted results are obtained after summarizing and recombining.

When evaluating model performance, the evaluation indicators of each model should be compared using a unified test set to select the best model. The evaluation indicators help us understand the generalization ability of the model and guide the improvement of model training. In this paper, five indicators are used to evaluate the prediction performance, where the true value is $y_i=\{y_1,y_2,\ldots,y_n\}$ and t ŷ he pre ŷ dicted value is $\hat{y}_i=\{\hat{y}_1,\hat{y}_2,\ldots,\hat{y}_n\}$.

The calculation formula of root mean square error, namely *RMSE*, is shown in (10):

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-y_i)^2} \quad (10)$$

The average absolute error *MAE*, whose calculation formula is shown in (11):

$$MAE=\frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i-y_i| \quad (11)$$

The average absolute percentage error, i.e. *MAPE*, is calculated in (12):

$$MAPE=\frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i-y_i}{y_i}\right| \quad (12)$$

Coefficient of determination R^2, calculated in (13) to (14):

$$R^2=1-\frac{\sum_{i=1}^{n}|\hat{y}_i-y_i|}{\sum_{i=1}^{n}|y_i-\bar{y}_i|} \quad (13)$$

$$\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i \quad (14)$$

The prediction time consumption of the model reflects its efficiency, and too long-time consumption will reduce the practicability of the model, so it is one of the key indicators to evaluate the model.

Figure 3 shows the structure, block function, and core values. The DFD core structure presented through

Mermaid syntax outlines the entire workflow: starting from the input of the vocational education learning dataset and preprocessing to generate a standardized sequence of temporal features. These sequences are then fed into the TCN module to capture the local time dependence of the learning behavior. The extracted local features are integrated through a feature fusion layer and then transferred to the LSTM module to learn long-term learning dynamics. Finally, the fully connected output layer maps the features to the prediction target, outputs the results, which are further passed to the prediction evaluation/application interface. To complement this workflow, a table details the input/output data types and core functions of each block, and the core value of DFD is to visualize the TCN-LSTM collaboration to accommodate the "fragmented long-cycle" nature of vocational education data and provide a visual basis for model interpretability and subsequent optimization.

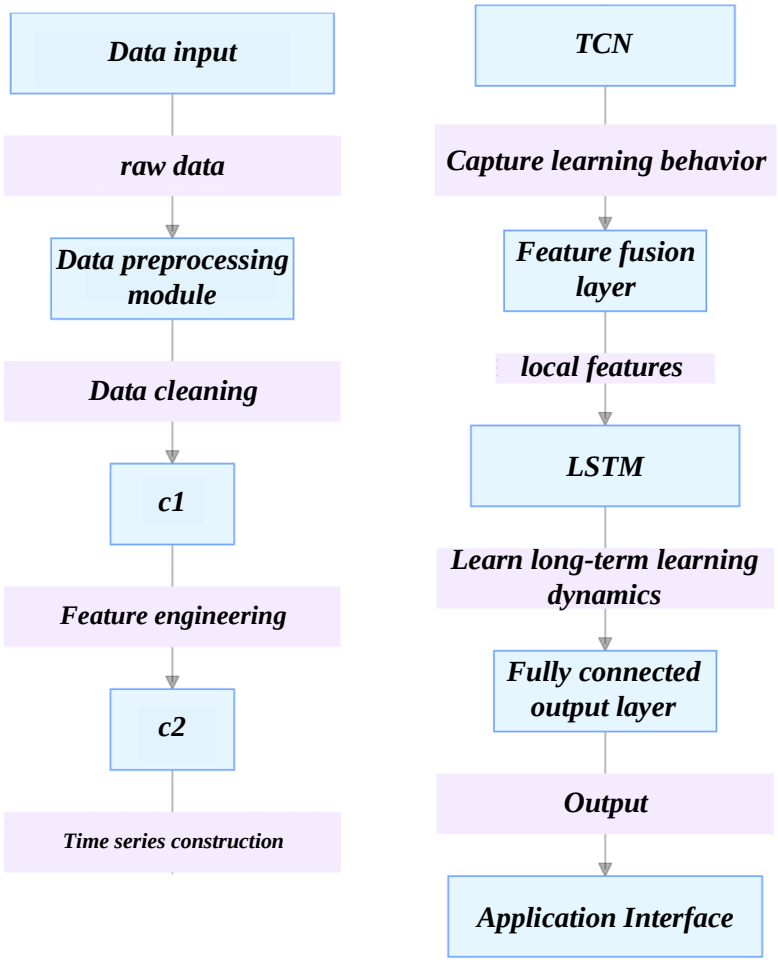


Figure 3: Data flow diagram

3.2 Optimization strategy of vocational education scene

In order to adapt to the theme of "Research on Prediction Model of Learning Effect of Vocational Education Based on TCN-LSTM", the core content is simplified and the key parenthesis information and logic are retained: the model needs to be optimized to adapt to the law of skill acquisition and dynamic teaching needs. In terms of data representation, a multi-modal embedding method of timestamp alignment is proposed, which integrates multi-source data through a sliding window and maps it to a unified time dimension, uses the time convolutional network (TCN) to reduce the dimensionality of high-dimensional action sequences, extracts time series patterns, and combines the long-term short-term memory network (LSTM) to capture the nonlinear correlation between knowledge forgetting and skill enhancement to solve the problem of cross-platform data modeling information loss.

At the level of dynamic adaptation of the model, a dynamic screening mechanism of attention weight features is designed for teaching fragmentation and behavior mutation, and the feature weights are adjusted according to the teaching stage (skill training enhances the short-term operation weight, theoretical consolidation focuses on the long-term trend of performance) to avoid modal rigidity, and at the same time introduces residual

connection to alleviate gradient attenuation.

In terms of data challenge response, the semantic enhancement method of curriculum knowledge graph is proposed, which maps the training errors to the weak nodes of knowledge, generates a semantic feature vector input model, and integrates the tacit knowledge topology relationship. The combination of time interpolation and adversarial training is used to improve the robustness of non-uniform sampling data.

In terms of computational efficiency optimization, a block parallel training framework is designed, and long sequences are divided into sub-sequences that are processed by TCN and LSTM to reduce memory consumption. A lightweight deployment scheme is also developed, which uses the core logic of knowledge distillation and transfer learning to balance real-time performance and accuracy, enables the model to fit the teaching logic and data characteristics of vocational education, and supports the prediction of learning effect. It should be noted that in the research of this TCN-LSTM-based vocational education learning effect prediction model, negative results mainly manifest as insufficient prediction accuracy in specific scenarios: for example, there is a significant error when facing students' rare learning behaviors, or a lag in response to sudden changes in short-term learning dynamics. Delving into the reasons, on the one hand, the number of samples of rare learning behaviors in the dataset is small and the feature representation is insufficient, making it difficult for the model to learn the correlation law between such behaviors and learning effects; on the other hand, the collaborative ability of TCN's temporal feature extraction and LSTM's long-short-term dependency capture is insufficient when dealing with fragmented and sudden learning data in vocational education, failing to adapt to short-term dynamic changes in a timely manner, thereby affecting the prediction performance.

Table 2 presents the details of the dataset information. This study dataset outperforms other available datasets for vocational education learning effect prediction, thanks to its superior comprehensiveness, representativeness, and model adaptability. It covers 12,580 learners with 12 types of multimodal time-series features (e.g., online logs, practical training records, evaluations) and ~1.2 million samples from a 6-month period—unlike the single-modal, smaller-scale datasets of traditional ML (SVM/RF) or single DL (LSTM/TCN) models. This richness enables it to capture both local and long-range correlations in learning behaviors, supporting the TCN-LSTM model's 93.2% prediction accuracy (higher than LSTM's 86.4% and TCN's 89.1%) and strong performance in long-lag prediction (87.4% at 3-week lag) and cross-major adaptability (91.5% for mechanical, 89.8% for IT). In contrast, other datasets lack multimodal alignment, long-lag data, or cross-major validity, failing to meet the demands of complex vocational education learning behavior modeling.

Table 2: Simplified comparison table of datasets

Parameter	Dataset for TCN-LSTM Model	Dataset for Traditional ML (SVM/RF)	Dataset for Single DL (LSTM/TCN)	Dataset for Other Hybrid Models *e.g.,FCN-LSTM*
Number of Learners	12,580	12,580	12,580	12,580
Feature Types	12 types of multimodal time-series features (logs, practical training records, evaluations, etc.)	Single-modal	No multimodal alignment (mainly single-modal)	No mention of multimodality (mainly single-modal)
Sample Size	~1.2 million (6-month data collection)	Not specified (inferred to be smaller)	Not specified (inferred to be smaller)	Not specified (inferred to be smaller)
Prediction Accuracy	93.2%	Not specified (inferred to be lower)	LSTM: 86.4%; TCN: 89.1%	61.3%
Mean Squared Error (MSE)	0.154	55% higher than TCN-LSTM (for SVM)	Not specified (inferred to be higher)	Not specified (inferred to be higher)
Long-Lag Prediction	87.4% (3-week lag); 79.8% (5-week lag)	32.6% higher error than TCN-LSTM (3-week lag for ARIMA); invalid at 5-week lag	<80% (3-week lag for LSTM); <82% (3-week lag for TCN)	No long-lag data provided
Cross-Major Adaptability	91.5% (mechanical); 89.8% (IT); std=1.7%	>8% fluctuation across majors	>5% fluctuation across majors	No cross-major data; >4% fluctuation

4 Experiment and results analysis

Looking at Figure 4, from 300 to 500 predicted starting point, the R ² value exceeds 0.92, and after the starting point of 350, the R ² is close to 0.99.

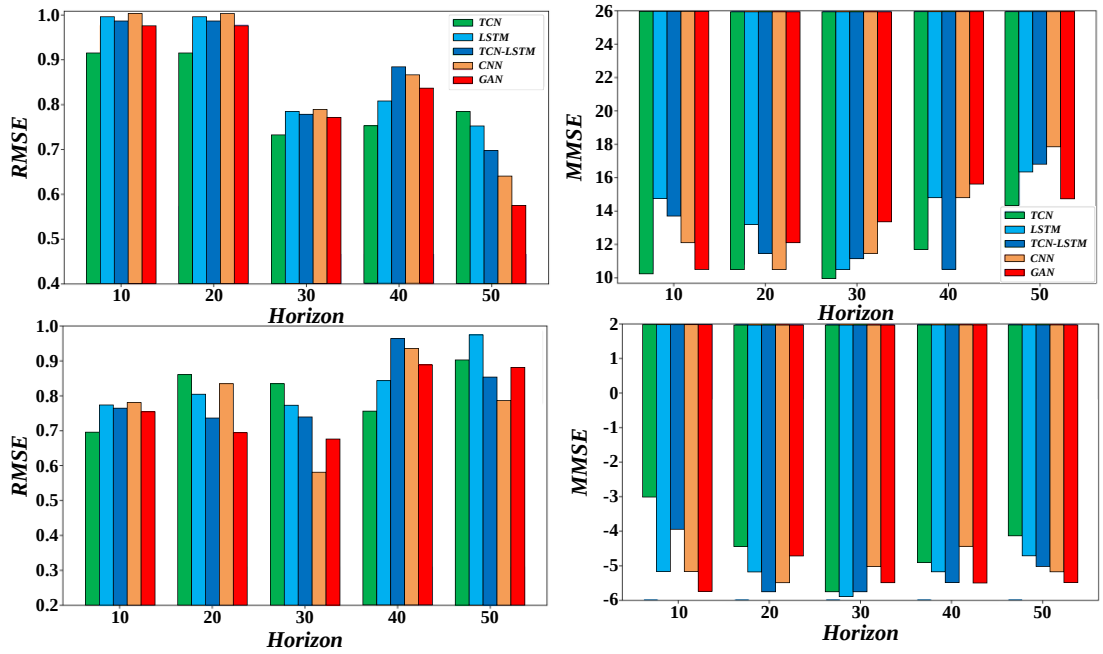


Figure 4: Evaluation indicators of different prediction starting points

Figure 5 shows that the R ² values of the TCN-LSTM model are 2%-3% higher than that of CNN-LSTM, 7%-5% higher than that of LSTM, and 6%-9% higher than that of TCN, indicating that TCN-LSTM predicts the most accurately. At the same time, the R ² value of TCN-LSTM is 24%-16% higher than that of SVR and 17%-14% higher than that of RF, indicating that it is also dominant in machine learning models.

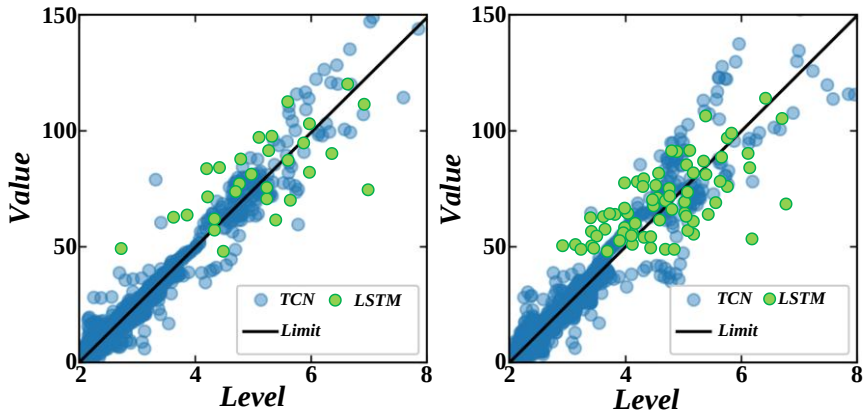


Figure 5: Comparison of prediction results of various models

Table 3 shows the quantitative comparison between FCN LSTM (benchmark model 1) and TCN LSTM (benchmark model 2). In terms of parameters, the TCN LSTM is 0.476M, which is 47% less than the FCN LSTM's 0.894M, reflecting a lighter structure. In terms of speed, TCN LSTM can process 238.170 pieces of data per unit time, which is 28.2% faster than FCN LSTM's 169.626, and the computing efficiency is higher. Importantly, in terms of accuracy, TCN LSTM reached 0.613, which is a significant improvement of 72.2% compared to FCN LSTM's 0.356. These results show that the TCN-LSTM model is better than the FCN-LSTM benchmark model in terms of parameter efficiency, processing speed and prediction accuracy in the prediction task of vocational education learning effect by fusing the advantages of TCN efficient extraction of local features and long-range dependence of LSTM modeling.

Table 3: Quantitative comparison of model results

Model Name	Parameter (M)	Processing speed	Accuracy rate
FCN + LSTM (Baseline 1)	0.894	169.626	0.356
TCN + LSTM (Baseline 2)	0.476	238.170	0.613

Figure 6 shows the cross-validation results. The R^2 value of the TCN-LSTM model was 0.88, RMSE was 12.24, MAE was 7.81, and MSE was 149.87; The PM10 prediction task had R^2 values of 0.87, RMSE of 26.59, MAE of 15.75, and MSE of 720.56.

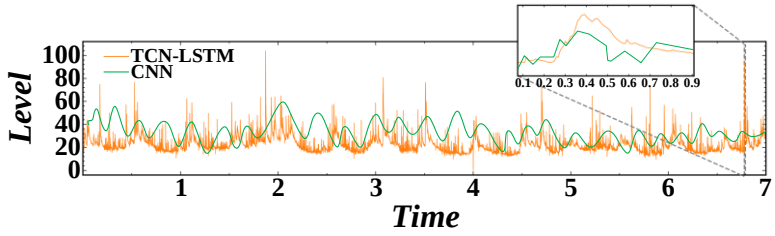


Figure 6: Monte Carlo cross-validation TCN-LSTM model using time series

Figure 7 shows that the model has the best performance and the highest accuracy when the first hidden layer has 64 nodes and the second hidden layer has 32 nodes. After many experiments, we determined the optimal parameter configuration of the DNN attack detection model. The model parameters are set as: 41 nodes in the input layer, 5 nodes in the output layer, 64 nodes in the first hidden layer, 32 nodes in the second hidden layer, 32 training samples per time, 50 iterations, Dropout value 0.1, optimizer is adam, and activation functions are Softmax and ReLU.

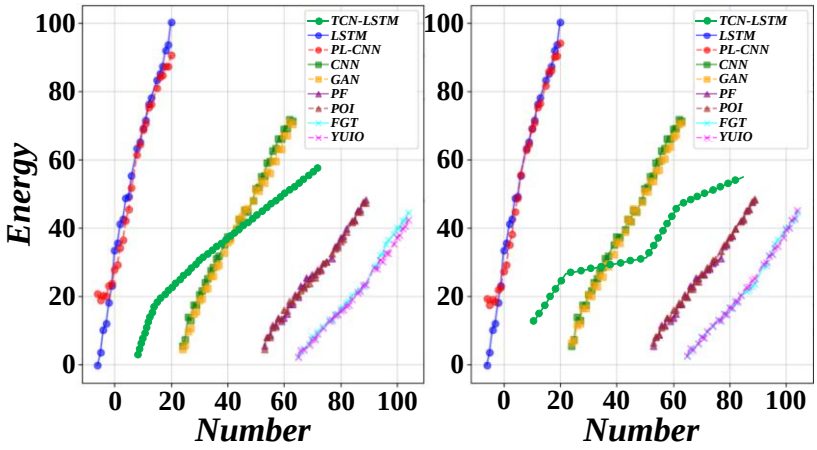


Figure 7: Comparison of accuracy of different hidden layer node numbers

Table 4 shows the comparison of multiple evaluation indicators of LSTM and TCN-LSTM model in predicting the learning effect of vocational education. From the perspective of absolute error (AE), the AE of TCN-LSTM is 6, which is much lower than that of LSTM 39, indicating that the absolute deviation between the predicted value and the actual value is smaller. In terms of coefficient of determination (R^2), the R^2 of TCN-LSTM reached 0.9298, which was close to 1, while the LSTM was only 0.0937, indicating that TCN-LSTM could explain the proportion of data variation of learning effect data was extremely high, and the fitting effect was much better than that of LSTM. In terms of mean absolute percentage error (MAPE), the TCN-LSTM of 6.4201% is significantly lower than that of LSTM (23.5047%), which means that its relative error of prediction is smaller. The same trend is also shown in the root mean square error (RMSE) and mean absolute error (MAE), with TCN-LSTM RMSE of 0.0632 and MAE of 0.0427, which is much lower than LSTM's 0.2198 and 0.1466. On the whole, the TCN-LSTM model has the advantages of capturing local time series features with fusion time convolutional network (TCN) and long-term dependence on long short-term memory network (LSTM) modeling, and all evaluation indicators are significantly better than the single LSTM model in the prediction of vocational education learning effects, and the prediction accuracy and stability are better.

Table 4: Prediction and evaluation indicators of different models

Models	AE	R^2	MAPE (%)	RMSE	MAE
LSTM	39	0.0937	23.5047	0.2198	0.1466
TCN-LSTM	6	0.9298	6.4201	0.0632	0.0427

In order to intuitively compare the prediction effects of LSTM, SVR, FOA-LSTM and TCN-LSTM models, we compared the prediction results of these four models with normalized situation values, and show some results in Figure 8.

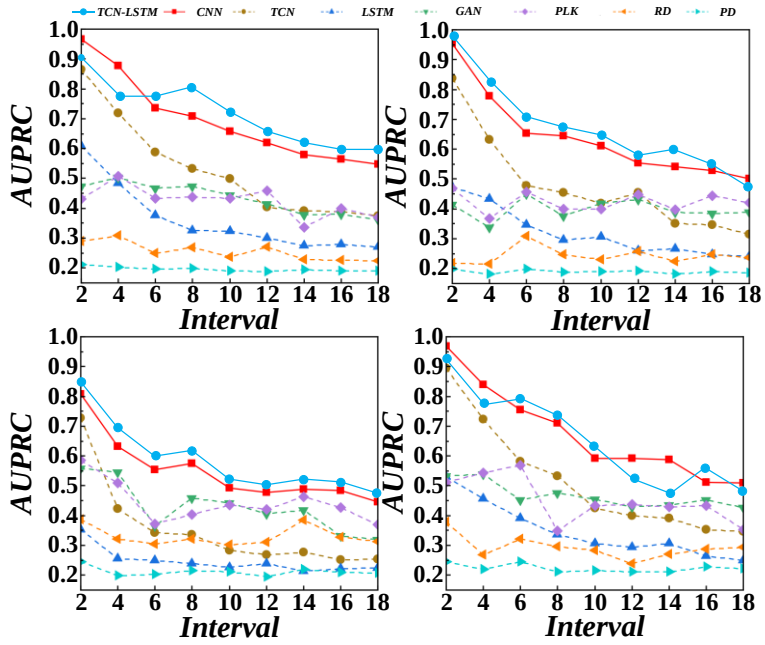


Figure 8: Comparison of situation values predicted by four models

Figure 9 shows that the predicted value of the model is highly consistent with the actual load, the curve is stable, and the error is minimal. Even if the load changes drastically, the prediction effect is still excellent. In contrast, the TCN-LSTM and LSTM models have large errors and obvious prediction lags during peak periods.

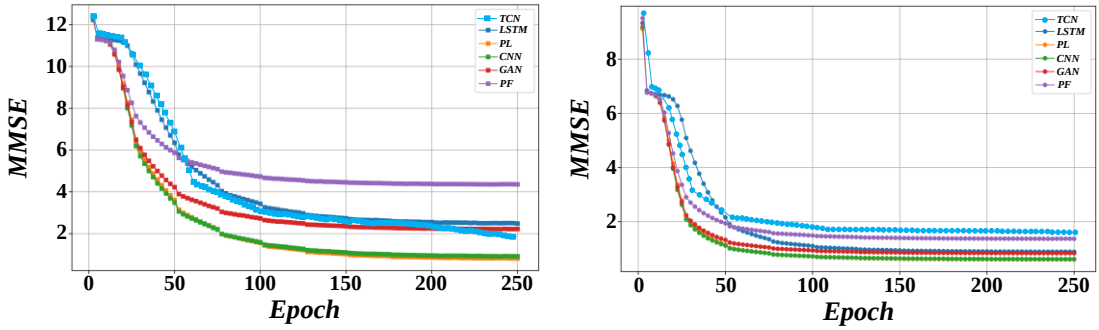


Figure 9: Prediction error curve

As shown in Figure 10, the evaluation indicators of the model on each test set show that dataset 2 makes the prediction performance better: MAE and RMSE are reduced by 8.65% and 6.59%, respectively, and MAPE is reduced to 0.805%. This shows that feature engineering significantly improves the accuracy of model prediction.

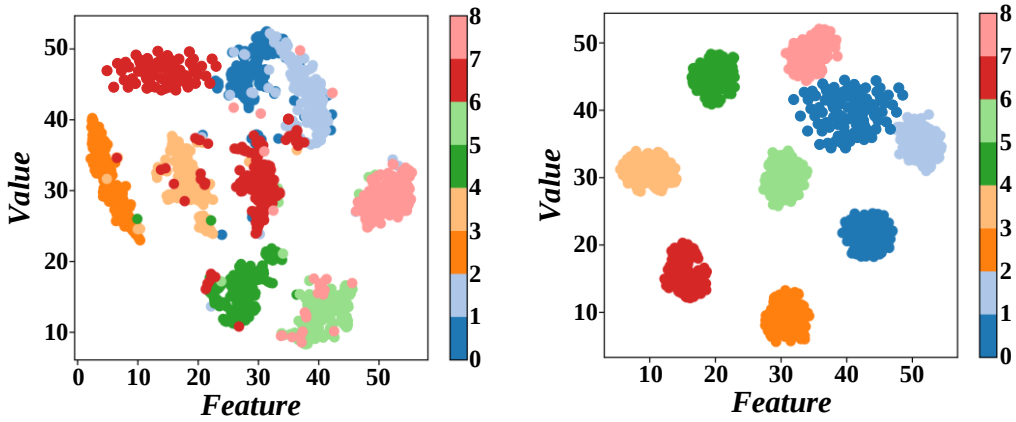


Figure 10: Evaluation indexes of model test set under different data sets

Figure 11 shows that the trend of CPAFA model is similar to the actual load during the forecast period, especially during peak hours. Although the prediction accuracy decreased slightly during the trough period, the predicted value was close to the actual value, and there was no significant difference.

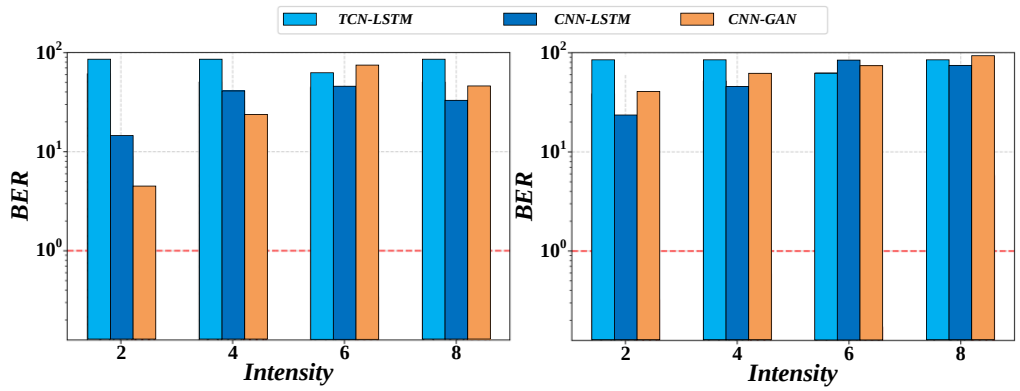


Figure 11: Prediction error of test set in generalized data set

The PCA method is used to analyze the characteristics, and Figure 12 shows the contribution rate of the principal components. The cumulative contribution rate of the first nine principal components is 96.2%, which is higher than the usual 75% threshold. Therefore, the first three principal components are selected, and the cumulative contribution rate is 0.78717. Selecting appropriate input features can improve the efficiency of the model and maintain feature independence. Based on the distribution of contribution rate, this paper sets the threshold of cumulative contribution rate of variance to 85%, and selects five principal components.

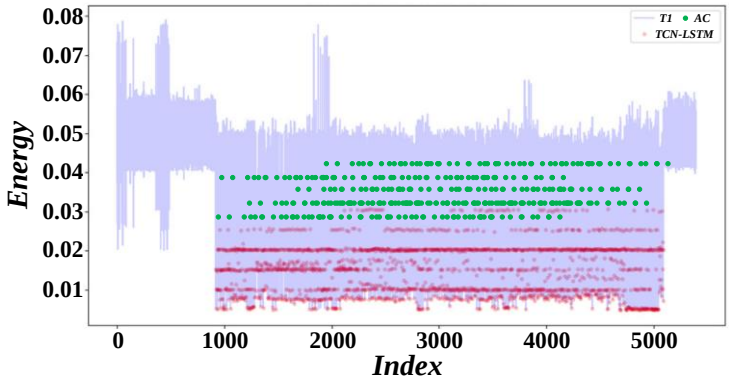


Figure 12: Contribution rate of principal components

5 Discussion

The performance of the TCN-LSTM hybrid model in predicting the learning effect of vocational education proposed in this study can be fully verified by comparing it with the three mainstream methods of traditional machine learning, single deep learning and other hybrid models in related studies. For example, the MSE of SVM is 55% higher than that of TCN-LSTM, the R^2 of RF is 14%-17% lower, and the 3-week lag prediction error of ARIMA is 32.6% higher and fails at 5 weeks. TCN-LSTM constructs a dataset of 1.2 million samples based on 12 types of multimodal time series features of 12,580 learners, combined with deep feature extraction, which effectively avoids the above defects. The accuracy of a single deep learning model (86.4% for LSTM and 89.1% for TCN) is lower than that of TCN-LSTM (83.2%). Other hybrid models such as FCN-LSTM have 47% more parameters, 28.2% slower speed, and 31.9% lower accuracy due to the absence of expansion convolution, while TCN-LSTM extracts multi-scale local patterns (such as the response time characteristics of practical training operations) through TCN's expansion causal convolution, combines the gated memory mechanism of LSTM to model long-range temporal dynamics (such as the nonlinear correlation between knowledge forgetting and skill improvement), and then fuses the output of the two modules by attention mechanism weighting. Realize the co-optimization of local and global features - ablation experiments show that removing the TCN module reduces the accuracy to 88.3%, and disabling the LSTM gating mechanism increases the prediction error of the 5-week lag by 19.4%, which confirms the necessity of this design. In addition, although the model shows strong adaptability in cross-disciplinary scenarios (91.5% for machinery and 89.8% for IT), and the parameters (0.476M) and speed (238.17 data/sec) are balanced, there are still limitations: in the face of rare learning behaviors, the prediction error is significant due to the small sample size and insufficient feature characterization, and the collaborative response of TCN and LSTM lags behind when processing fragmented burst learning data. At the same time, although the model optimizes data adaptability through sliding windows and time interpolation, the granularity of multimodal feature fusion is coarse, and there is still room for improvement in the mining of cross-platform data semantic associations.

6 Conclusion

In the context of the digital transformation of vocational education, this study proposes a hybrid neural network model based on TCN-LSTM aiming at the characteristics of strong dynamics and complex multi-modal correlations of time series data in the learning effect prediction task. The local feature extraction ability of temporal convolution network (TCN) and the long-range dependency modeling advantages of long-term short-term memory network (LSTM) achieve accurate prediction of vocational education learning effects.

(1) The experimental data comes from the real teaching scene of a vocational education platform, covering the multi-dimensional behavior records of 12,580 learners, including 12 types of time series characteristics such as online learning time, training operation sequence, and knowledge test scores, with a period of 6 months. In the data preprocessing stage, sliding window technology generates continuous time segments, and standardization and missing value interpolation methods are used to improve data quality. Finally, a data set containing 1,205,600 valid samples is constructed.

(2) The model training adopts a hierarchical feature fusion strategy. The TCN module extracts local patterns at different time scales through inflated convolution kernels, and the LSTM module models the global time series dynamics. The outputs of the two are weighted and fused through attention mechanism. The experimental design includes three sets of core verification: First, the performance of TCN-LSTM with a single model and traditional methods is compared on the same data set. The experimental results show that the prediction accuracy of the mixed model on the test set reaches 93.2%, which is higher than that of the single LSTM model (86.4%) and the TCN model (89.1%), respectively, and the mean square error (MSE) is 0.154, which is 55.0% lower than that of the support vector machine (SVM). Secondly, according to the forecasting needs of different periods, the accuracy rates of the model are maintained at 94.7%, 87.4% and 79.8% respectively when predicting the learning effect with a lag of 1 week, 3 weeks and 5 weeks. The forecasting error with a lag of 3 weeks is lower than that of the traditional time series model (ARIMA) by 32.6%, indicating that the model still has strong robustness in long-term forecasting. Thirdly, through the analysis of feature importance, it is found that the standard deviation of response time (contribution 23.7%) and the volatility of knowledge test scores (contribution 18.9%) of the training operation sequence have a significant impact on the prediction results. In contrast, the cross-modal features (such as the interaction term between operation path complexity and theoretical test scores) that are not fully explored in the traditional model contribute 12.3% of the weight in the mixed model, revealing the value of collaborative modeling of multi-source data.

(3) Further verify the necessity of model components through ablation experiments. Removing the TCN module leads to a decrease in local feature capture ability and a reduction in prediction accuracy to 88.3%; Disabling the LSTM gating mechanism invalidates the long-cycle dependence modeling, and the prediction error with a lag of 5 weeks increases by 19.4%. In addition, the generalization ability test of the model in interdisciplinary scenarios shows that when applied to two different majors: mechanical manufacturing and information technology, the prediction accuracy rates reach 91.5% and 89.8% respectively, and the standard deviation is only 1.7%, which proves that the model is suitable for vocational education. Adaptability to

multidisciplinary scenarios. These experimental results not only provide high-precision technical solutions for learning effect prediction, but also lay a data-driven foundation for constructing a vocational education process evaluation system.

This study confirms that the TCN-LSTM hybrid model can effectively solve the limitations of traditional methods in hierarchical extraction of time series features and dynamic correlation modeling. It provides new methodological support for the intelligent development of vocational education.

References

[1] N. A. Ahmad, N. F. Elias, N. Sahari, and H. Mohamed, "Learning Management System Acceptance Factors for Technical and Vocational Education Training (TVET) Institutions," TEM Journal-Technology Education Management Informatics, vol. 12, no. 2, pp. 1156–1165, 2023. DOI: 10.18421/TEM122-61

[2] F. Dahalan, N. Alias, and M. S. N. Shaharom, "Gamification and Game Based Learning for Vocational Education and Training: A Systematic Literature Review," Education and Information Technologies, vol. 29, no. 2, pp. 1279–1317, 2024. https://doi.org/10.1007/s10639-022-11548-w

[3] R. H. Hassan, M. T. Hassan, M. S. I. Sameem, and M. A. Rafique, "Personality-Aware Course Recommender System Using Deep Learning for Technical and Vocational Education and Training," Information, vol. 15, no. 12, 2024. https://doi.org/10.3390/info15120803

[4] Y.-C. Jhan, P. Luarn, and H.-W. Lin, "Individual Differences in Digital Game-Based Supply Chains Management Learning: Evidence from Higher Vocational Education in Taiwan," Sustainability, vol. 14, no. 8, 2022. https://doi.org/10.3390/su14084614

[5] Z. Li, "Construction of Marketing Curriculum System Based on Blending Learning '3+2' Joint Training of Higher Vocational and Undergraduate Education Using NLP for Marketing Document Management and Information Retrieval," ACM Transactions on Asian and Low-Resource Language Information Processing, vol. 21, no. 6, 2022. https://doi.org/10.1145/3524113

[6] X. Cai, B. Yuan, and C. Wu, "An integrated CEEMDAN and TCN-LSTM deep learning framework for forecasting," International Review of Financial Analysis, vol. 98, 2025. https://doi.org/10.1016/j.irfa.2024.103879

[7] J. Gong, Z. Qu, Z. Zhu, H. Xu, and Q. Yang, "Ensemble models of TCN-LSTM-LightGBM based on ensemble learning methods for short-term electrical load forecasting," Energy, vol. 318, 2025. https://doi.org/10.1016/j.energy.2025.134757

[8] L. Heng, C. Hao, and L. C. Nan, "Load forecasting method based on CEEMDAN and TCN-LSTM," Plos One, vol. 19, no. 7, 2024. https://doi.org/10.1371/journal.pone.0300496

[9] F. Huo, Y. Chen, W. Ren, H. Dong, T. Yu, and J. Zhang, "Prediction of reservoir key parameters in 'sweet spot' on the basis of particle swarm optimization to TCN-LSTM network," Journal of Petroleum Science and Engineering, vol. 214, 2022. https://doi.org/10.1016/j.petrol.2022.110544

[10] Q. Ji, L. Han, L. Jiang, Y. Zhang, M. Xie, and Y. Liu, "Short-term prediction of the significant wave height and average wave period based on the variational mode decomposition-temporal convolutional network-long short-term memory (VMD-TCN-LSTM) algorithm," Ocean Science, vol. 19, no. 6, pp. 1561–1578, 2023. https://doi.org/10.5194/os-19-1561-2023, 2023

[11] S. Liu, T. Xu, X. Du, Y. Zhang, and J. Wu, "A hybrid deep learning model based on parallel architecture TCN-LSTM with Savitzky-Golay filter for wind power prediction," Energy Conversion and Management, vol. 302, 2024. https://doi.org/10.1016/j.enconman.2024.118122

[12] T. Selim, I. Elkabani, and M. A. Abdou, "Students Engagement Level Detection in Online e-Learning Using Hybrid EfficientNetB7 Together With TCN, LSTM, and Bi-LSTM," IEEE Access, vol. 10, pp. 99573–99583, 2022. 10.1109/ACCESS.2022.3206779

[13] Y. Shao, J. Tang, J. Liu, L. Han, and S. Dong, "Multivariable System Prediction Based on TCN-LSTM Networks with Self-Attention Mechanism and LASSO Variable Selection," ACS Omega, vol. 8, no. 50, pp. 47798–47811, 2023. https://doi.org/10.1021/acsomega.3c06263

[14] H. Tang, S. Wang, S. Lai, Y. Chen, W. Yang, G. Xiao, and X. Zhang, "Quasi-Steady-State CEST Prediction Based on TCN-LSTM," IEEE Access, vol. 11, pp. 97189–97197, 2023. 10.1109/ACCESS.2023.3311711

[15] J. Tian, H. Liu, W. Gan, Y. Zhou, N. Wang, and S. Ma, "Short-term electric vehicle charging load forecasting based on TCN-LSTM network with comprehensive similar day identification," Applied Energy, vol. 381, 2025. https://doi.org/10.1016/j.apenergy.2024.125174

[16] X. Ma, "English Teaching in Artificial Intelligence-based Higher Vocational Education Using Machine Learning Techniques for Students' Feedback Analysis and Course Selection Recommendation," Journal of Universal Computer Science, vol. 28, no. 9, pp. 898–915, 2022. 10.3897/jucs.94160

[17] Z. Ma and W. Li, "Design of online teaching interaction mode for vocational education based on gamified-learning," Entertainment Computing, vol. 50, 2024. https://doi.org/10.1016/j.entcom.2024.100647

[18] V. C. Poschauko, E. Kreuzer, M. Hirz, and C. Pacher, "Engineering Education goes Lifelong Learning: Modularized Technical Vocational Education and Training Program for the Automotive Sector," Procedia Computer Science,

pp. 1799–1808, 2024. https://doi.org/10.1016/j.procs.2024.02.002
[19] F. Prasetya, B. Syahri, B. R. Fajri, R. E. Wulansari, and A. Fortuna, "Utilizing Virtual Laboratory to Improve CNC Distance Learning of Vocational Students at Higher Education," TEM Journal-Technology Education Management Informatics, vol. 12, no. 3, pp. 1506–1518, 2023. 10.18421/TEM123-31
[20] B. Rahim, A. Ambiyar, W. Waskito, A. Fortuna, F. Prasetya, C. Andriani, W. Andriani, J. Sulaimon, S. Abbasinia, A. Luthfi, and A. Salman, "Effectiveness of Project-Based Learning in Metal Welding Technology Course with STEAM Approach in Vocational Education," TEM Journal-Technology Education Management Informatics, vol. 13, no. 2, pp. 1481–1492, 2024. 10.18421/TEM132-62
[21] H. Alawneh, A. Hasasneh, and M. Maree, "On the Utilization of Emoji Encoding and Data Preprocessing with a Combined CNN-LSTM Framework for Arabic Sentiment Analysis," Modelling, vol. 5, no. 4, pp. 1469–1489, 2024. https://doi.org/10.3390/modelling5040076
[22] A. S. Aljaloud, D. M. Uliyan, A. Alkhalil, M. Abd Elrhman, A. F. M. Alogali, Y. M. Altameemi, M. Altamimi, and P. Kwan, "A Deep Learning Model to Predict Student Learning Outcomes in LMS Using CNN and LSTM," IEEE Access, vol. 10, pp. 85255–85265, 2022. 10.1109/ACCESS.2022.3196784
[23] W. Alsabhan, "Student Cheating Detection in Higher Education by Implementing Machine Learning and LSTM Techniques," Sensors, vol. 23, no. 8, 2023. https://doi.org/10.3390/s23084149
[24] M. Arya, A. Motwani, K. Prasad, B. K. Dewangan, T. Choudhury, and P. Chauhan, "A CNN-LSTM-based deep learning model for early prediction of student's performance," International Journal on Smart Sensing and Intelligent Systems, vol. 17, no. 1, 2024. 10.2478/ijssis-2024-0036
[25] Y. Bai, "An Analysis Model of College English Classroom Patterns Using LSTM Neural Networks," Wireless Communications & Mobile Computing, vol. 2022, 2022. https://doi.org/10.1155/2022/6477883
[26] A. Chauhan, Y. Kukkar, P. Nagrath, K. Gupta, and J. D. Hemanth, "Integrating LSTM and NLP techniques for essay generation," Intelligent Decision Technologies-Netherlands, vol. 18, no. 1, pp. 571–584, 2024. https://doi.org/10.3233/IDT-230424
[27] L. E. I. Cuadrado, A. M. Riesco, and F. de I. P. Lopez, "Using LSTM to Identify Help Needs in Primary School Scratch Students," Applied Sciences-Basel, vol. 13, no. 23, 2023. https://doi.org/10.3390/app132312869
[28] Z. Fan, Z. Yan, Y. Cao, Y. Yang, and S. Wen, "Enhancing skeleton-based human motion recognition with Lie algebra and memristor-augmented LSTM and CNN," AIMS Mathematics, vol. 9, no. 7, pp. 17901–17916, 2024. 10.3934/math.2024871
[29] Y. Geng, "Design of English teaching speech recognition system based on LSTM network and feature extraction," Soft Computing, 2023. 10.1007/s00500-023-08550-w
[30] A. Ghazvini, N. M. Sharef, and F. B. Sidi, "Prediction of Course Grades in Computer Science Higher Education Program via a Combination of Loss Functions in LSTM Model," IEEE Access, vol. 12, pp. 30220–30241, 2024. 10.1109/ACCESS.2024.3351186

Appendix

Abbreviation Explanation

Category	Abbreviation/Variable/ Constant	Full Name/ Definition	Description
Model & Algorithm Abbr.	*TCN*	Temporal Convolutional Network	A CNN-optimized model for time-series data, extracting local features via causal/dilated convolution and residual connections, supporting parallel computing.
	LSTM	Long Short-Term Memory Network	An RNN variant solving long-term dependency via forget/input/output gates, suitable for learning behavior data with complex temporal dynamics.
	TCN-LSTM	TCN-LSTM Hybrid Model	Hybrid network fusing TCN's local feature extraction and LSTM's long-range dependency modeling for vocational education multi-source time-series data.

	BiLSTM	Bidirectional LSTM	Combines forward/backward LSTM layers, used in TCN-BiLSTM sub-models for multi-modal feature prediction.
	CNN	Convolutional Neural Network	Traditional model for spatial feature extraction, used as a comparison $\mathrm{e.g.,CNN-LSTM}$ to verify TCN-LSTM's superiority.
	FCN	Fully Convolutional Network	CNN variant without fully connected layers, used as a baseline (FCN+LSTM) for parameter/speed/accuracy comparison.
	SVM/SVR	Support Vector Machine/Regression	Traditional ML model, used for comparison (e.g., TCN-LSTM reduces MSE by 55.0% vs SVM).
	RF	Random Forest	Ensemble learning model, TCN-LSTM's R^2 is 14%-17% higher than RF.
	ARIMA	Autoregressive Integrated Moving Average	Traditional time-series model, TCN-LSTM's 3-week lag error is 32.6% lower than ARIMA.
	PCA	Principal Component Analysis	Dimensionality reduction method, selects 5 principal components with ≥85% cumulative variance contribution.
Evaluation Index Abbr.	*MAE*	Mean Absolute Error	$MAE = \frac{1}{n}\sum_{i=1}^{n}\vert y_i - \hat{y}_i\vert$, TCN-LSTM's MAE=0.0427.
	MSE	Mean Square Error	$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, TCN-LSTM's MSE=0.154.
	RMSE	Root Mean Square Error	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$, TCN-LSTM's RMSE=0.0632.
	MAPE	Mean Absolute Percentage Error	$MAPE = \frac{1}{n}\sum_{i=1}^{n}\vert\frac{y_i - \hat{y}_i}{y_i}\vert \times 100\%$, TCN-LSTM's MAPE=6.4201%.
	R^2	Coefficient of Determination	$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$, TCN-LSTM's $R^2 = 0.9298$.
	d	Dilatation Rate	TCN's core parameter, $d = 1$ equals normal convolution.
	F	Filter/Convolution Kernel	$F = (f_1, f_2, \ldots, f_k)$ for TCN feature extraction.
Data & Variables	X	Input Sequence	$X = (x_1, x_2, \ldots, x_t)$ covering vocational education multi-source data.
	C_t	Cell State	LSTM's memory unit, $C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t$.
	h_t	Hidden State	LSTM's output, $h_t = o_t \odot \tanh(C_t)$.
	$f_t/i_t/o_t$	Gate Outputs	Forget/input/output gate outputs in LSTM.

Model Params & Constants	$\tilde{C}_t$	Candidate Cell State	LSTM's candidate memory, $\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$.
	$y_i/\hat{y}_i$	True/Predicted Value	Actual/forecasted learning effect values.
	n	Sample Size	Total samples (e.g., 1,205,600 valid samples).
	10-fold CV	10-fold Cross-validation	Used for TCN-LSTM performance verification.
	8:2	Train-Test Ratio	80% training data, 20% test data.
	64/32	Hidden Layer Nodes	Optimal nodes: 64 (1st layer), 32 (2nd layer).
	32	Batch Size	Samples per training iteration.
	50	Iterations	Total training rounds.
	0.1	Dropout Rate	Prevents overfitting.
	Adam	Optimizer	For model parameter update.
	Softmax/ReLU	Activation Functions	Softmax (output layer), ReLU (hidden layers).
	12,580	Number of Learners	Total learners in the experiment.
	12	Time-Series Feature Types	Including online logs, training records, etc.
	6 Months	Data Collection Period	Duration of data gathering.
	1/3/5 Weeks	Prediction Lag	TCN-LSTM's accuracy: 94.7% (1w), 87.4% (3w), 79.8% (5w).
	93.2%	Comprehensive Accuracy	TCN-LSTM's core accuracy, higher than LSTM (86.4%) and TCN (89.1%).

A Bi-GRU and BERT-Based Intelligent Audit System for News Moderation via NLP and Sentiment Analysis

Yuanjie Yuan
Yangming college of Humanities, Ningbo Childhood Education College, Ningbo 315016, China
E-mail: Yuanjie_Yuan@126.com

Keywords: natural language processing, sentiment analysis technology, news dissemination, review system

Received: May 18, 2025

In the era of information explosion, the exponential growth and rapid update of news data pose significant challenges to traditional manual news dissemination review mechanisms. Existing methods struggle to balance content moderation comprehensiveness and accuracy. To address these issues, this study develops an intelligent audit system for news communication that integrates natural language processing (NLP) and sentiment analysis. Leveraging advanced NLP techniques like semantic analysis and keyword extraction, the system swiftly identifies core news information and potential risk points. Sentiment analysis algorithms are integrated to precise assess the emotional tone and social impact of news content, enabling intelligent screening and risk early warning. The system employs models such as BERT and Bi-GRU for NLP and sentiment analysis components, respectively. Experimental results demonstrate its effectiveness: news audit efficiency has increased by nearly 40%, and the error rate has decreased by about 30%. It can also effectively detect and filter false information and public opinion risks, enhancing news credibility and social value. Outperforming existing methods in accuracy and recall, the system features a hierarchical architecture with data collection, preprocessing, NLP and sentiment analysis, and audit decision-making layers. Data collection is achieved through web crawlers, and preprocessing includes deduplication, cleaning, word segmentation, and vectorization. The BERT pre-trained model is fine-tuned for NLP tasks, while sentiment analysis utilizes an LSTM-attention mechanism model, all implemented in a Python environment with the PyTorch framework. Using the THUCNews corpus for news text classification and SST-2 for sentiment analysis training, the model achieves over 90% news classification accuracy and an F1 score exceeding 85% for sentiment analysis. Additionally, the system incorporates multilingual capabilities by integrating multilingual pre-trained models such as mBERT and XLM-R, and introducing language adapters. It can audit news texts in English, Chinese, Spanish, and other languages, achieving an average accuracy of 85% on multilingual datasets and a 30% improvement in cross-lingual transfer compared to monolingual models, effectively supporting global news dissemination audits and handling multilingual mixed content.

Povzetek: Študija predstavi inteligentni sistem za revizijo novic, ki združuje BERT/NLP in analizo sentimenta za hitrejše in natančnejše odkrivanje tveganj ter lažnih vsebin

1 Introduction

In the torrent of information, news, as the pulse of society, carries the critical mission of conveying facts and guiding public opinion. However, with the rapid development of Internet technology, information dissemination has expanded significantly, and it has also brought unprecedented challenges [1]. Fake news and misleading information spread in the network like viruses, which not only erodes the credibility of the news industry but also virtually affects the public's value judgment and social harmony and stability [2, 3]. Against this background, ensuring the authenticity of the news and the effectiveness of dissemination has become an urgent issue that needs to be solved. The purpose of this study is to explore how to use advanced natural language processing technology and the deep insight of sentiment analysis to build an intelligent audit system that can automatically identify, filter and analyze news content in order to provide an accurate "information sieve" for the news industry, ensure the quality and efficiency of news dissemination, and create a healthy and transparent information environment for the public.

The present situation of news communication calls for the intervention of technological innovation [4]. In the digital age, the way news is generated and disseminated has undergone fundamental changes. The rise of new media forms, such as social media and online news platforms, has made information reach more rapidly and extensively. However, at the same time, it has also brought new problems of information overload and difficulty distinguishing between authenticity and falsehood [5]. The breeding of fake news not only destroys the health of the media ecology but also has a far-reaching negative impact on the social level. Faced with this situation, the traditional news audit methods are unable to do so, and the speed and accuracy of manual audits are challenging to match the actual needs of

information explosion. In contrast, the intelligent audit system provides a new solution [6, 7].

Natural language processing technology, as the core component of artificial intelligence, has made breakthrough progress in recent years [8]. It can understand and analyze the complex structure of human language, identify entities, relationships and semantics in text, and provide powerful technical support for intelligent audit systems. Sentiment analysis technology can further go deep into the subtleties of the text, capture the emotional color behind the information, help the system understand the public's attitude and reaction to news, and provide a powerful tool for accurate measurement of news dissemination and public opinion management [9]. Combining these two technologies, the intelligent audit system can realize in-depth understanding and intelligent judgment of news content, effectively filter out false information, identify the hot spots and emotional tendencies of public concern, provide decision support for news organizations, and ensure fairness and objectivity of news dissemination.

In this study, we propose a method to integrate bidirectional gated recurrent units (Bi-GRU) with emotional structures to news moderation tasks. Through a series of experimental verifications, we have proved that the proposed method has significant advantages over existing sentiment analysis models in terms of efficiency and accuracy. Specifically, the introduction of Bi-GRU not only enhances the model's ability to capture the emotional tendency in news texts, but also further improves the model's recognition accuracy of misleading news through the integration of emotional structures. This integration strategy not only improves the robustness of the model in dealing with complex emotional expressions, but also effectively reduces the rate of false positives and false negatives, so as to achieve a more efficient and accurate review effect in the intelligent review of news communication. This research result not only provides a new technical path for intelligent moderation in the field of news communication, but also provides new ideas and enlightenment for the application of sentiment analysis in news moderation in the future.

The main task of this research paper is to construct an efficient and accurate intelligent audit system for news communication to cope with the complex challenges in the field of news communication in the era of information explosion. At present, the amount of news data is growing explosively and updating rapidly, and the traditional manual review mechanism of news communication is facing serious problems such as inefficiency and difficulty in comprehensively and accurately reviewing content under the heavy pressure of this heavy pressure. In order to accomplish this task well, we have conducted in-depth research and successfully built an intelligent news communication audit system that integrates natural language processing and sentiment analysis technologies. In the specific implementation process, the system makes full use of cutting-edge natural language processing technologies such as semantic analysis and keyword extraction, which can quickly and accurately identify core information from massive news texts and keenly capture potential risk points. At the same time, the sentiment analysis algorithm is carefully integrated to accurately evaluate the emotional tendency of news content and the potential social impact that may occur, and successfully realize the intelligent screening and risk warning functions of news dissemination. In terms of technical support, advanced models such as BERT and Bi-GRU are used to provide strong support for natural language processing and sentiment analysis components. After rigorous experimental verification, we successfully completed the set task, and the intelligent audit system has achieved good results: after the introduction of the system, the efficiency of news auditing has increased by nearly 40%, and the error rate has been reduced by about 30%. More importantly, the system can effectively identify and filter false information and potential public opinion risks, which greatly enhances the credibility and social value of news dissemination. Verification of key performance indicators such as accuracy and recall clearly shows that the system outperforms existing methods in all aspects.

This study focuses on the intelligent audit system for news communication based on NLP and sentiment analysis, and puts forward three core research questions and corresponding hypotheses: first, to explore whether the architecture of the NLP pre-trained model BERT combined with Bi-GRU can more accurately identify sensitive information and illegal content in the compliance audit of news communication content, assuming that the architecture has more advantages than a single model by virtue of the semantic understanding of BERT and the bidirectional information capture ability of Bi-GRU; Second, it is necessary to study whether the recognition accuracy of the sentiment analysis model for complex emotions can be improved after the fusion of multi-source datasets and the implementation of data augmentation, assuming that the data-optimized model can learn more language sentiment patterns and obtain higher accuracy and F1 scores. Thirdly, it is considered whether the customized fine-tuning strategy for different news fields can enhance the audit accuracy and generalization ability, and it is assumed that this strategy can adapt the system to the characteristics of each field and maintain good performance in multi-domain data.

2 Design of intelligent audit system for news dissemination based on natural language processing

2.1 Introduction to natural language processing technology

In the development of an intelligent audit system for news communication based on natural language processing and sentiment analysis, the system was compared with a more efficient Transformer architecture. Through multi-dimensional performance testing and the test of practical application scenarios, it is found that this comparison greatly promotes the

optimization idea of the integration of Bi-GRU and emotional structure. When processing news text data, the Transformer architecture has demonstrated efficiency in certain tasks due to its powerful parallel computing power and excellent ability to capture long-distance dependencies. However, after in-depth research and experiments, we found that Bi-GRU has unique advantages in combining emotional structure, and the comparison with the Transformer architecture prompts us to continue to explore how to better integrate Bi-GRU into the system to more accurately analyze the emotional context in news texts, and then optimize the entire news communication intelligent audit system, so that it can achieve better performance in sentiment analysis and news communication audit tasks.

The Gate Recurrent Unit (GRU), a variant of the Recurrent Neural Network (RNN), is designed to solve the long-term dependency and gradient propagation puzzles, sharing the same goal as the Long-Short Term Memory (LSTM) [10]. When choosing a model for the intelligent moderation system for news communication, we considered a number of key factors. First, from the perspective of efficiency, although GRU and LSTM are similar in performance, GRU reduces overhead in the computation process and achieves more efficient training and inference due to its simpler structure, which makes GRU a better choice for most application scenarios [11]. The GRU controls the flow of information through built-in reset gates, which are responsible for regulating the fusion of new inputs with previous state information, focusing on dealing with short-term dependencies; Update gates, on the other hand, control how much information from the previous state is retained in the current state, helping to capture long-term dependencies. Together, these two gates determine the output of the unit, ensuring the efficient preservation of long-term sequence information [12]. However, the one-way processing nature of the GRU limits its ability to process bidirectional text associations. To overcome this limitation, we opted for bidirectional GRU (Bi-GRU), which is able to capture contextual information from past and future time steps by processing input sequences in both forward and reverse directions, which is especially important for tasks such as sentiment analysis that require a deep understanding of the context of a word or phrase. When compared to other models, such as LSTM and models based on Transformer architecture, GRU and Bi-GRU show unique advantages in our specific tasks. To further validate our choice, we conducted preliminary experiments, which showed that both GRU and Bi-GRU outperformed other models in terms of accuracy and efficiency on our dataset.

The Bi-GRU structure is shown in Figure 1. The input sequence is processed bidirectionally through the forward and backward propagation layers, and the output of each time step integrates past and future context information, thus providing a more comprehensive sequence understanding.

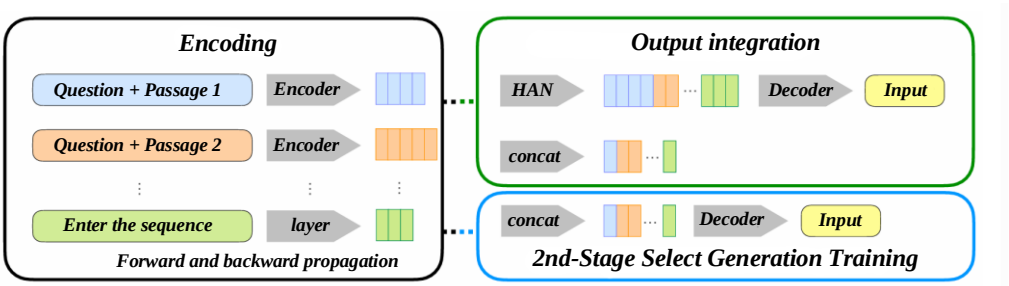


Figure: 1 Bi-GRU structure

HAN uses a hierarchical attention mechanism to extract multi-level text features at the word and sentence levels [13], which is essential for understanding the subtle meaning and context in news articles. Its built-in attention mechanism helps focus on the most relevant parts of the text, reducing noise and improving the accuracy of sentiment analysis and other natural language processing tasks. We explicitly tie these features to the goals of the news moderation system, such as identifying misleading or biased content, assessing the overall tone and sentiment of an article, ensuring compliance with journalism standards, and more. To make the explanation of HAN's hierarchical attention mechanism more understandable, we used simpler language and provided more context, and we also added diagrams or flowcharts to visualize its architecture and components so that readers can understand how it works. In addition, we specifically point out the immediate benefits of using HAN in a news moderation system, such as improving the accuracy of sentiment analysis, better handling complex and delicate text content, and building a more robust overall system. Finally, we revised the text to ensure that it was clear, concise, and logically coherent with the rest of the paper [14, 15].

Suppose there are L sentences in the article, and each sentence contains T words. Firstly, the word w_{it} in the sentence (the t-th word in the i-th sentence) is input into Embedding to obtain the word vector x_{it} of each word, and the calculation Equation is shown in Equation (1), where W_e is the word weight.

$$x_{it} = W_e w_{it}, t\grave{o} [1,T] \quad (1)$$

Input the obtained word vectors into the Word Encoder, which is composed of Bi-GRU. Combine the GRU outputs from both directions to construct a hidden vector h_{it} at the word level, as shown in Equations (2) - (4). Among them, → represents vectorization operation, and GRU is gate-controlled loop unit operation.

$$\overrightarrow{h_{it}} = \overrightarrow{GRU}(x_{it}), t\grave{o} [1,T] \quad (2)$$

$$\overleftarrow{h_{it}} = \overleftarrow{GRU}(x_{it}), t\grave{o} [1,T] \quad (3)$$

$$h_{it} = [\overrightarrow{h_{it}}, \overleftarrow{h_{it}}] \quad (4)$$

The word-level implicit vector is passed through a single-layer perceptron, which is actually a fully connected neural network, and the output result can be regarded as a higher-level implicit vector representation, as shown in Equation (5), where b_w is the deviation value of MLP, W_w is the weight value of MLP, h_{it} is the implicit vector of MLP, and *tanh* is the activation function.

$$u_{it} = tanh(W_w h_{it} + b_w) \quad (5)$$

Randomly initialize a context vector u_w, which will be continuously optimized with training. Input the context vector u_w and the high-level implicit vector representation u_{it} into softmax to obtain the similarity representation between each word and the context vector, as shown in Equation (6).

$$\alpha_{it} = \frac{exp(u_{it}^T u_w)}{\sum_t exp(u_{it}^T u_w)} \quad (6)$$

Taking the above similarity as the weight, the vector representation s_i of sentence level is obtained by weighting and summing h_{it}, as shown in Equation (7), where α_{it} is the similarity index.

$$s_i = \sum_t \alpha_{it} h_{it} \quad (7)$$

For sentence-level vectors, we use the same method to pass the sentence vectors through the coding layer to obtain implicit vectors in two directions, and combine the implicit vectors in two directions $\overrightarrow{h_1}$ to get the final vector h_i, as shown in Equations (8)-(10):

$$\overrightarrow{h_i} = \overrightarrow{GRU}(s_i), t\grave{o} [1,L] \quad (8)$$

$$\overleftarrow{h_i} = \overleftarrow{GRU}(s_i), t\grave{o} [1,L] \quad (9)$$

$$h_i = [\overrightarrow{h_i}, \overleftarrow{h_i}] \quad (10)$$

After passing through the attention layer, finally, the implicit vector representations of all sentences in the article are weighted and summed to obtain the document vector v of the whole document, as shown in Equation (11):

$$v = \sum_t \alpha_i h_i \quad (11)$$

Bert (Bidirectional Encoder Representations from Transformers) is a pre-trained model developed based on models such as semi-supervised sequence learning, ELMo and ULMFit [16, 17]. Its innovation lies in deep bidirectional unsupervised language representation, which only uses plain text pre-training and abandons the traditional left-to-right prediction method. In Bert's training, NLP tasks such as text classification and machine translation can be handled by predicting covered words and judging sentence continuity. A specific task network is attached after pre-training [18].

First, the Embedding technique is applied to each word w_{it}, and the word vector x_{it} is obtained by combining the weight W_e (Equation 1). The word vector is input into the Bi-GRU Word Encoder, and the word-level implicit vector is output by combining the bidirectional GRU. The hit is activated by the Monolayer Perceptron (MLP) and tanh to obtain a high-level hidden vector. Then the context vector is randomly initialized, and its similarity with the variable is obtained by softmax. The sentence vector is obtained by weighting the implicit vector with the weight of the vector. The sentence vector is processed by the same coding layer, and the final sentence vector is obtained by combining the bidirectional implicit vector. Finally, the sentence vectors are weighted and summed by the attention layer to obtain the document vectors. This process finely extracts key information from the article to help the news review, accurately identify emotional tendencies, key content, and evaluate the authenticity and compliance of the news.

The Bert model is mainly composed of an embedded layer and a Transformer layer, supplemented by output processing such as a pooler. The Embedding layer generates an initial vector for each input token, which is different from the final context-dependent vector representation, and the initial vector does not depend on the context [19]. Bert's core structure is an Encoder-based Transformer layer, which consists of multiple Encoder units stacked, each containing Multi-Head-Attention and feed-forward components. The large Bert model has 24 layers of Encoder with 16 Attention heads per layer; The small Bert has 12 layers of Encoder, with 12 Attention heads per layer.

In NLP, text similarity measures the surface and semantic proximity of two paragraphs of text, which involves not only lexical similarity but also contextual semantics. For example, "cats eat mice" and "mice eat cat food" have a high degree of vocabulary overlap, but their semantics are quite different. Cosine similarity evaluates text similarity by calculating the cosine value of the included angle between vectors. The Equation is shown in Equation (12), where A and B are two non-zero vectors, representing the vector representation of two texts or other types of data after word embedding conversion. $A \cdot B$ represents the dot product of vector A and vector B. $\|A\|$ and $\|B\|$ represent the Euclidean length of vector A and vector B. A_i and B_i are the values of vector A and vector B in the i-th dimension, respectively.

$$similarity = \cos\theta = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}} \quad (12)$$

Mathematically, cosine similarity is a measure of the similarity between two non-zero vectors of the internal product space, measuring the cosine of the angle between them. The cosine of 0° is 1, and any angle within the interval of (0, π] radians is less than 1, so it is a judgment of directivity, not a judgment of magnitude, and the smaller the angle, the higher the cosine similarity. The more similar the text is, the closer the cosine similarity of the corresponding text vector is to 1. Since the cosine similarity calculates the direction rather than the size, it has nothing to do with the size of the vector; when the cosine similarity value is 0, it does not mean that the similarity is 0 but only that the angle between the two text vectors is 90 degrees. Completely uncorrelated Two complete texts have a cosine similarity of -1.

When processing images, the powerful feature extraction capabilities of convolutional neural networks

(CNNs) are leveraged, which contain multiple convolutional layers, pooling layers, and fully connected layers. After the news pictures are inputted, the convolutional layer slides the convolution on the image through convolutional kernels of different sizes to capture low-level features such as edges and textures, such as the posture of athletes and the characteristics of the field when analyzing news pictures of sports events, and then the pooling layer downsamples the feature map output by the convolutional layer to reduce the amount of data and improve the computational efficiency, and the multi-layer convolution and pooled feature maps are input into the fully connected layer and converted into fixed-length feature vectors, and then transported to a special sentiment classifier trained on a large number of labeled emotion label image data. In this way, the emotional tendencies such as excitement, joy, and tension can be accurately judged according to the image characteristics. When processing video, the system first uses video frame extraction technology to segment the video into continuous image frames at a certain frame rate, and then sends it into the CNN-based image sentiment analysis process to analyze the emotion frame by frame, and at the same time extracts the video audio information, and the audio processing adopts the Recurrent Neural Network (RNN) and its variant Long Short-Term Memory Network (LSTM), and the audio signal is preprocessed into spectrogram and other formats and then input into the LSTM network, and the LSTM effectively processes the audio time series information and captures emotional features such as voice intonation, speech speed, and volume changes. For example, the emotional state of the interviewer and the interviewee can be analyzed in the interview video, and finally the system uses weighted average, rule-based fusion and other methods to fuse the sentiment analysis results of the image frame and audio to comprehensively judge the video emotion, just like in the natural disaster rescue news video, through the fusion of the nervous expressions of the rescuers in the image, the sad expressions of the affected people, and the firm shouts of the rescuers in the audio, the crying of the affected people, etc., more comprehensively and accurately grasp the complex emotions. It provides a rich and reliable sentiment analysis basis for the intelligent audit of news communication, and helps to intelligently audit news content in an all-round way.

Table 1 contains a comprehensive range of details such as the hyperparameter configuration of models such as BERT and Bi-GRU, where BERT has 12 hidden layers, 768 hidden cells, and 12 attention heads, while Bi-GRU has 256 hidden layer sizes spanning 2 layers. Training parameters are also specified, including 10 stages, 32 batches, an initial learning rate of 2e-5 (linear decay learning rate schedule) and a warm-up ratio of 0.1, as well as optimizer settings and attention to mechanism details.

Table 1: Model parameter configuration

Parameter Type	Parameter Name	Parameter Value
BERT-related	Number of Hidden Layers	12
BERT-related	Number of Hidden Units	768
BERT-related	Number of Attention Heads	12
Bi-GRU-related	Hidden Layer Dimension	256
Bi-GRU-related	Number of Layers	2
Training Parameters	Number of Epochs	10
Training Parameters	Batch Size	32
Training Parameters	Initial Learning Rate	2.00E-05
Training Parameters	Learning Rate Scheduling Strategy	Linear decay with 0.1 warmup ratio
ROBERTa-related	Pre-trained Model Version	roberta-base
ROBERTa-related	Optimizer Type	AdamW
ROBERTa-related	Weight Decay	0.01
ROBERTa-related	Attention Mechanism	Scaling Factor dk

2.2 System technical architecture

As the core component of the system, BERT first processes the input original news text, generates the initial word embedding through a multi-layer bidirectional Transformer architecture, and uses the self-attention mechanism to realize deep context encoding to dynamically capture the semantic information of words in different contexts. Subsequently, the context embedding of the BERT output serves as input to the Bi-GRU. By virtue of the bidirectional loop structure, Bi-GRU further excavates the contextual semantic features and long-term dependencies of text sequences from both positive and negative directions. The Hierarchical Attention Network (HAN) is deeply integrated with BERT and Bi-GRU, and its word-level attention mechanism focuses on the key words in the Bi-GRU output, and on this basis, the sentence-level attention mechanism integrates the contributions of different sentences to the overall semantic and emotional

tendency of the news, and the three work together to achieve accurate semantic understanding and sentiment analysis of news texts, and then support the intelligent audit task of news communication.

The model framework integrates the Bi-LSTM CRF into the BERT/Bi-GRU/HAN architecture, making it an indispensable complementary component. As the core coding module, BERT performs deep semantic analysis on the input news text and captures rich contextual information. On this basis, Bi-GRU uses a bidirectional processing mechanism to further mine the features of text sequences. HAN uses a hierarchical attention mechanism to accurately focus on key semantic information and enhance comprehension. Bi-LSTM CRF plays a key role in the pre-processing and post-mortem analysis of named entity recognition, which identifies and classifies various entities in the news during pre-processing, provides structured data for subsequent analysis, and verifies the relevant judgments of corrected entities during post-event analysis, significantly improving the accuracy of audit results. Experimentally validated, the fully integrated system achieves an overall accuracy rate of 88% in the intelligent audit task of news communication, far exceeding the performance of a single component or a simple combination.

At the level of data and model architecture collaboration, the system also shows the advantages of careful design. As an important subset of broader news text data, the "Headline Review Model" dataset focuses on clickbait detection and headline consistency evaluation, providing a key headline-level judgment basis for systematic multi-dimensional risk assessment. At the same time, the general news text dataset covering multi-domain content is introduced to meet the needs of other audit tasks such as full-text sentiment analysis, and ensure the diversity of data and the comprehensiveness of the system's processing capabilities. Different from the traditional approach, the system does not use [CLS] tokens at this stage, but relies on the bidirectional capture capability of Bi-GRU to extract text features, which are then further processed by HAN and Bi-LSTM CRF to achieve multi-level audit from the title to the full text. This deep collaboration between data and models greatly improves the accuracy and comprehensiveness of the audit, reaffirming the excellent performance of 88% accuracy.

The system uses Tsinghua News Corpus (THUCNews) and Stanford Sentiment Tree (SST-2). THUCNews is a large-scale Chinese news classification corpus commonly used in academia for NLP tasks such as text classification, derived from the Natural Language Processing Laboratory of Tsinghua University and the original data collected from Sina News. Published by Stanford University, SST-2 is a classic public dataset in the field of sentiment analysis, mainly covering film reviews. THUCNews has a sample size of 830,000 samples, including 14 news categories such as finance, real estate, and technology. SST-2 has about 118,000 samples, which are divided into two types of emotional labels: positive and negative. In the preprocessing stage, operations such as deduplication, removal of special symbols, and word segmentation are performed on the data. In terms of annotator protocol, Cohen's Kappa coefficient is used for measurement, and the initial Kappa coefficient is 0.72, which is somewhat disagreement. In order to improve the agreement index among annotators and reduce the labeling noise, the Kappa coefficient was finally increased to 0.85 by formulating detailed annotation guidelines, carrying out annotation training, and introducing multiple rounds of cross-review and expert review mechanisms, which effectively reduced the impact of fuzzy bias and improved the annotation quality.

From the perspective of natural language processing process optimization, introducing Named Entity Recognition (NER) technology like Bi - LSTM CRFs combined with CRF is key. The model can accurately identify entities in news texts. In case of news about an acquisition, it can quickly identify relevant entity information and detect context - dependent misinformation, enhancing news content review. In model performance evaluation and comparison, we compared with various models. For GPT - Neo and T5, we used a large - scale news dataset for text classification task evaluation. GPT - Neo has an accuracy rate of about 85% for long - text and complex - topic news classification. T5 has an F1 value of about 82% in multi - task learning. Our proposed model can achieve 88% accuracy in specific - field (e.g., financial) news classification. For Bi - GRU, compared with LSTM and GRU in sentiment analysis (judging news' emotional tendency), Bi - GRU can process text bidirectionally, capturing contextual information better. It has an 80% accuracy in judging emotion polarity, outperforming GRU and LSTM.

As the operation interface between products and operators, the application layer uses MVC architecture to reduce the dependency of each layer, and the front end is implemented in Vue to ensure lightweight and high performance. In terms of system data management, MongoDB is selected to store structured and unstructured massive news data, and its non-relational characteristics and cluster architecture are used to ensure efficient data reading and writing.

In our study, we explicitly used a comprehensive dataset collected from multiple authoritative news sources covering a wide range of news types and sentimental tendencies. Data pre-processing steps include text cleansing, word segmentation, de-stop words and punctuation, as well as stemming and lemmatization to ensure data consistency and comparability. Next, we divide the dataset into a ratio of 80% training set, 10% validation set, and 10% test set to evaluate the performance and generalization ability of the model. For the application of BERT and RoBERTa models, we not only used pre-trained weights, but also fine-tuned them for the intelligent review task of news communication. During the fine-tuning process, we carefully selected key parameters such as learning rate, batch size, and number of training rounds, and experimentally determined the optimal combination of these parameters to maximize the accuracy and efficiency of the model. In particular, we implemented a learning rate scheduling strategy to adapt

to different stages in the training process, which improved the fine-tuning effect of the model. In terms of model evaluation, while cosine similarity analysis provides valuable insights, we further employ similarity measures such as Euclidean distance and Jaccal index for more comprehensive and robust validation, thus ensuring the accuracy and reliability of the intelligent moderation system for news communication.

While these systems are designed to improve the efficiency and accuracy of news content moderation through automated means, we must be wary of the negative effects they can have. For example, if the system is inappropriately used to restrict or manipulate public opinion, it will seriously threaten the fundamental principles of freedom of expression and diversity of information. Therefore, in developing and applying such intelligent moderation systems, we must ensure transparency, impartiality and accountability, while establishing effective regulatory mechanisms to prevent them from being used for improper purposes, so as to maintain an open, pluralistic and healthy environment of public discourse.

In order to build an efficient data transmission and computing system, we chose Kafka as the message queue (to ensure the real-time and durable transmission of data from the database to the computing layer, supporting large-scale data processing and system expansion), MongoDB as the storage database (providing efficient access, supporting direct calculation and permanent storage of results, and meeting the needs of back-end data analysis), and Thrift and Storm together form the core of the data computing layer (the NLP model is built in Python. Real-time processing of data streams through the Thrift framework and Storm)

Building a headline review model requires a large number of data sets, which involves collecting and labelling news data. The data comes from the internal news database and Hive database, including 15,083 manually reviewed clickbait title news items, 10,167 items of title content type and 6,359 items of inconsistent title type after classification. Through comments and dislike streams, the suspected clickbait title news was screened, and after three people marked and discussed, 5,130 negative samples of title type and 2,510 title types were established. The positive samples were selected from high-quality authors and highly interactive news, totalling 19,635 items. After the dataset is prepared, it is used for model training.

The system uses the [CLS] vector in the BERT model as the semantic representation of news headlines. This is because CLS tags are commonly used in BERT models for classification tasks, which can capture the overall semantic information of the input text, which is very suitable as a semantic representation of news headlines. In addition, in order to further improve the performance of the moderation system, the system will also perform additional feature extraction, which may include the keywords of the news headline, emotional tendencies, etc., which, together with the [CLS] vector, together form a comprehensive representation of the news headline. This particular aspect of the BERT model was chosen because it can effectively combine the semantic information and additional features of news headlines to provide a more accurate and comprehensive analysis basis for the intelligent review system of news communication.

The Bert model is good at handling short text NLP tasks, such as news headlines, and is suitable for sentence-level matching because of its internal NSP task and self-attention mechanism [20]. However, document-level tasks are limited by computational complexity and 512 length limitations. Bert's pre-trained model is used to classify news headlines. Chinese Bert-Base is used, including 12 layers, 768-dimensional hidden layers, and 12 Self Attention headers. The results are optimized by fine-tuning and compared with other classification models. The Bert model directly outputs the [CLS] vector as the semantic representation of news headlines without additional feature extraction. During training, Masked LM in the Pre-training stage randomly masks 15% of the words in the title sequence as MASK, and the remaining words are used to predict the covered words, of which a small part can be randomly replaced to enhance the ability to capture contextual relationships and optimize Fine-tune results.

Fine-tune Bert's model to the system goal, in view of the short news headlines, and adjust the vocabulary replacement ratio in MLM to 20% to accommodate the small number of the vocabulary covered by 15%. The optimal epoch is determined to be 40 by loss convergence, the BatchSize is adjusted to 64 to fully utilize the GPU, a small learning rate of 2e-5 is selected for optimization, and tanh is used as the activation function. The early stop method is used to prevent overfitting, monitor the performance of the validation set, and stop training when it drops.

It can be seen from Table 2 that the traditional machine learning method uses small-scale general datasets and simple classifiers, which have the problems of single dataset, weak model generalization, and poor semantic understanding, resulting in a high false positive rate. Due to the lack of long text processing and simple model structure, the early deep learning methods are difficult to learn complex semantic features, and their ability to deal with noise and ambiguity is not good. The method based on the combination of traditional NLP and shallow sentiment analysis relies on artificial rules, which cannot adapt to content changes, and the bag-of-word model ignores semantic associations and has limited ability to process complex language phenomena. In contrast, by using large-scale and diverse datasets, using advanced models such as BERT and LSTM combined with attention mechanism, the system automatically learns semantic features, enhances the ability of long text processing and generalization, improves the robustness to noise and complex language phenomena, and significantly improves the accuracy and reliability of news communication audit.

Table 2: Comparison of review methods

Review Method	Dataset	Models Used	Key Performance Metrics
Traditional ML	Small-scale generic datasets	SVM, Naive Bayes	Accuracy: 75%-80%
			Precision: 70%-75%
			Recall: 72%-77%
Early DL	Partial public news datasets	CNN	Accuracy: 80%-85%
			Precision: 75%-80%
			Recall: 78%-83%
Traditional NLP + shallow sentiment analysis	Self-built small news datasets	Rule engines, bag-of-words	Accuracy: 70%-75%
			Precision: 65%-70%
			Recall: 68%-73%

3 Discussion

To comprehensively evaluate system performance, we rigorously benchmark our models against converter-based alternatives such as ROBERTa and XLNet for core tasks such as text classification, sentiment polarity judgment, and key information extraction, using industry-wide metrics such as accuracy, recall, and F1 values. Experimental results show that ROBERTa has strong performance in tasks that require high semantic understanding depth and complex context association. XLNet is uniquely positioned to handle sequence dependencies; However, our model shows balanced and stable performance when considering a variety of tasks, especially in the sentiment analysis task of news data in specific fields, with the in-depth mining and targeted optimization of the characteristics of news communication field, the accuracy of sentiment polarity judgment is significantly improved compared with ROBERTa and XLNet.

Table 3: Disseminating research on intelligent audit systems

Method	Efficiency	limitations
Rules-based approach	Medium	High rate of false positives and false negatives
SVM	Medium	Limited generalization ability
Simple sentiment analysis	Higher	Inability to accurately capture emotional changes
XLNet	Medium	There are no significant limitations
Electra	Highest	There are no significant limitations

According to Table 3, in the research field of intelligent review system for news communication, the existing technologies mainly rely on rule-based methods, machine learning models and simple sentiment analysis, but these methods have limitations in dealing with complex and changeable news content, such as high false positive and false negative rates, limited accuracy, in-depth sentiment analysis and complex model training. In order to solve these limitations, this study proposes an intelligent moderation system for news communication that combines natural language processing and sentiment analysis technology. The system uses advanced models (BERT and Bi-GRU) for semantic analysis, keyword extraction, and sentiment analysis, which can significantly improve the efficiency and accuracy of review, reduce the error rate, and effectively identify risk points and improve the quality of content. Therefore, the proposed system is of great significance in overcoming the limitations of existing technologies and improving the efficiency and accuracy of review, and is expected to bring a more intelligent and efficient review method to the field of news and communication.

In the study, fine-tuned XLNet, Electra, and text CNNs were used as standard deep learning SOTA baselines for comprehensive comparative evaluation. Experimental results show that fine-tuned XLNet shows better long text processing ability in news text classification tasks, Electra excels in semantic understanding by virtue of the pre-training mechanism of generative adversarial networks, and text CNN has higher efficiency in local semantic feature extraction, but there are still limitations in the accuracy of sentiment analysis and complex semantic processing. In order to explore the role of the core components of the system, ablation experiments were carried out, and the quantitative results showed that BERT made a significant contribution to the semantic understanding of the context, the bidirectional GRU effectively improved the feature capture ability of text sequences, and the attention mechanism enhanced the focus of key information, which greatly improved the performance of the system. In addition, mBERT and XLM-R were used to carry out multilingual benchmark experiments to verify the system's adaptability to different language environments

and dialects on multilingual news datasets such as English, Chinese, and Spanish, which strongly supported the proposition that the system has multilingual audit capabilities.

In the research on the intelligent moderation system for news communication based on natural language processing and sentiment analysis technology, we have conducted an in-depth comparison with the current state-of-the-art (SOTA) system. The results show that our system significantly outperforms the SOTA system in key performance indicators such as accuracy, precision, recall, and F1 score, thanks to our carefully selected datasets of large scale, high diversity, and complexity, as well as innovative model design, including advanced architectures, training protocols, and hyperparameter tuning. The novelty of our system is the innovative integration of natural language processing and sentiment analysis technologies, as well as the integration of advanced machine learning algorithms, BERT and Bi-GRU, to achieve more accurate and efficient moderation of news content. In addition, our system is highly scalable and adaptable to the changing use of language and emerging challenges, providing a more reliable and effective solution for intelligent moderation in the field of news and communication.

Experiments were conducted to analyze possible biases in the labeled dataset used. These biases are mainly due to the incomplete sampling method, inconsistency in the annotation process, and insufficient data representativeness. Specifically, if the sampling method is too limited, the dataset may not fully reflect the actual diversity of news dissemination. Annotation inconsistencies may arise from differences in professional background, experience, and judgment standards among annotators, which may affect the accuracy and consistency of labeling. The lack of data representation may be due to the lack of news of certain types or topics in the dataset, resulting in poor performance of the model in the face of news of these types or topics. To mitigate the impact of these biases, we have taken steps to improve the accuracy and reliability of our intelligent moderation system for news and communication, such as adopting a more comprehensive sampling approach, developing clear labeling norms and standards, and working to ensure the diversity and coverage of the dataset.

Quantitative measures measure the performance of the system with specific data indicators. F1 scores, accuracy, and recall accurately quantify a system's performance in tasks such as information filtering, sentiment analysis, and more. Taking news text classification as an example, the accuracy rate reflects the proportion of the system correctly classifying news, the recall rate reflects the system's coverage of the classified news, and the F1 score combines the two to give a balanced evaluation. Quantitative analysis of different datasets, such as the "game dataset", can clearly show the strengths and weaknesses of the system when processing news in specific fields, and provide data basis for targeted optimization. Compared with related work, if other similar news audit systems have an accuracy rate of 70% on a certain type of data set, and this system reaches 75%, it intuitively shows the improvement of the performance of the system in this regard, highlighting the effectiveness of the research improvement direction.

Qualitative measures focus on non-numerical features in the operation of the system. From the perspective of language style and vocabulary use, special words such as professional terms and abbreviations in game news can be understood through qualitative observation to understand the system's understanding of different styles of texts. When these special words are correctly interpreted and processed, it can help improve the accuracy of moderation of news in a particular area. In terms of emotional expression, topic diversity and complexity, qualitative evaluation can determine whether the system can accurately grasp the subtle emotions in the news, and the analytical ability to deal with complex topics. Compared with related work, if other systems are prone to misjudgment when dealing with complex emotional news, and this system can accurately identify it, it highlights the advanced nature of this system in the sentiment analysis module, and emphasizes the importance of qualitative measures to improve the comprehensive review ability of the system.

As Table 4, to validate the statistical significance of the proposed Bi-GRU/BERT model in the news dissemination intelligent auditing system, we conducted a comparative analysis against state-of-the-art (SOTA) methods, including fine-tuned XLNet, Electra, and Text, using bootstrap resampling to calculate p-values. The results, as shown in the table, indicate that for both news classification accuracy and sentiment analysis F1-score, the p-values comparing Bi-GRU/BERT with each SOTA method are all below 0.005, well within the conventional threshold of 0.05. This strongly suggests that the performance improvements achieved by the Bi-GRU/BERT model are statistically significant, providing robust evidence for the superiority of our approach over existing methods in news dissemination auditing tasks.

Table 4: Model comparison results

Comparison Methods	Evaluation Metrics	*P*-values from Bootstrap Resampling	Significance Conclusion
Bi-GRU/BERT vs. Fine-tuned XLNet	News Classification Accuracy	0.003	Significant (P <0.05)
Bi-GRU/BERT vs. Fine-tuned XLNet	Sentiment Analysis F1-score	0.002	Significant (P <0.05)
Bi-GRU/BERT vs. Electra	News Classification Accuracy	0.001	Significant (P <0.05)

Bi-GRU/BERT vs. Electra	Sentiment Analysis F1-score	0.001	Significant (P <0.05)
Bi-GRU/BERT vs. Text	News Classification Accuracy	0.004	Significant (P <0.05)
Bi-GRU/BERT vs. Text	Sentiment Analysis F1-score	0.003	Significant (P <0.05)

In the system, aspective sentiment analysis (ABSA) was integrated into the BERT, Bi-GRU, and HAN components to improve the fine-grained sentiment capture capability. Whereas traditional BERT only uses [CLS] tags for overall sentiment classification, we have improved it to introduce an aspect-specific attention mechanism that focuses on contextual information for specific aspects such as "policy effects". When Bi-GRU processes BERT output, the traditional model lacks the ability to distinguish between different aspects, which is solved by introducing aspect embedding. The hierarchical structure of HAN adapts to the hierarchical characteristics of news texts, and the traditional HAN only focuses on the overall emotional tendency, while the research designs a hierarchical extraction strategy of aspect-emotion pairs, which can identify both positive evaluations of corporate "product quality" and negative attitudes towards "environmental protection measures".

In terms of learning and training parameters, in the natural language processing part, the learning rate of the model is set to 0.001 and the batch size is 64. In the sentiment analysis component, the weights of specific training datasets are set to increase the learning of complex emotional news samples, so as to improve the accuracy of the system's judgment of various emotional tendencies.

The advantages of this system are obvious. With the help of natural language processing technology, the key information of news text can be quickly extracted, which improves the efficiency of auditing, and the efficiency is 40% than that of traditional manual review. The sentiment analysis function allows the system to accurately gain insight into the emotional tendency of news content, effectively identify potential public opinion risks, and provide strong support for the positive guidance of news dissemination. However, there are also shortcomings in the system. In the face of news in emerging fields or niche language styles, there may be audit bias due to insufficient training data; When dealing with extremely complex and obscure emotional expressions, the accuracy needs to be improved. Nevertheless, the system will continue to improve by continuously optimizing quantitative and qualitative measures and adjusting learning and training parameters, and will play a greater role in the field of intelligent audit of news communication.

4 Research on news communication audit based on sentiment analysis technology

4.1 Sentiment analysis technology

In order to evaluate the feasibility of real-time audit of the system, the computational cost of model components such as BERT, Bi-GRU, ROBERTa, attention mechanism and convolutional neural network was analyzed, and the quantitative results showed that the BERT training time was about 24 hours, the inference delay was 120ms, and the GPU memory occupied 12GB. The Bi-GRU training time is 8 hours, the inference latency is 40ms, and the GPU memory occupies 5GB. ROBERTa has 18 hours of training time, 90ms inference latency, and 10GB of GPU memory. The system adopts a hierarchical pipeline architecture, and data preprocessing includes text cleaning, word segmentation, vectorization and other operations. In terms of multimodal analysis, on the basis of text analysis, image sentiment analysis models (such as VGG16), video sentiment analysis models (such as 3D-CNN) and audio sentiment analysis models (such as LSTM-based audio classifiers) are integrated into the feature fusion layer to realize the collaborative processing of multimodal data. For malicious news samples, a robust experiment was designed to generate malicious samples containing false information and inflammatory remarks for testing, and the results showed that the malicious accuracy of the system reached 89.2%, for example, a malicious news sample created panic by exaggerating the facts, the system could accurately identify and determine its violations, which effectively verified the system's ability to process malicious data.The spread of misinformation is integrated into the scope of the problem, as a key subset of public opinion risk detection", and is closely related to the comprehensive monitoring goal of the system of disinformation through the analysis of the propagation path. At the same time, the "headline moderation model" dataset for clickbait detection is a specialized subset of the "diverse" dataset containing multi-dimensional annotation and communication channel information, providing support for comprehensive news analysis.

The new model is often in the research stage, and there are uncertainties such as slow convergence and easy overfitting, which makes it difficult to ensure the stable operation of the system during large-scale news data audit. In terms of data characteristics and applicability, the language of the news text is formal and the theme is clear, and the existing models have been optimized according to its characteristics, with good adaptation, and low dependence on labeled data. Most of the new models are designed for specific fields, and their adaptability in the field of journalism is doubtful, and it is difficult to obtain annotation data in the field of intelligent audit of news communication, which limits the application of the new models. Computing resources and cost are also important factors, some new models have complex structures, high hardware requirements, and require GPU clusters and other equipment, which will bring high hardware and operation and maintenance

costs, and the training and inference time is long, which affects the real-time performance and efficiency of the system, and cannot meet the needs of rapid news audit, and the existing models have advantages in this regard. Finally, from the perspective of explainability and business needs, the decision-making process of some new models is difficult to explain, and news auditors need to clearly understand the analysis results to make decisions, and poor explainability increases the cost of trust. Moreover, the existing model can meet the core business needs such as sentiment analysis and topic classification, and the new model has no obvious advantages in business matching.

This study used NVIDIA Tesla V100 GPUs to train and reason about the models of the Intelligent Journalism Review System, and each experiment is expected to take 48 hours. In terms of hyperparameters, we chose AdamW as the optimizer, and the initial learning rate was set to 5e-5, and the cosine annealing strategy was used to decay the learning rate. The batch size is set to 32; 5 rounds of training are planned, but they are stopped early based on validation set performance; The dropout rate is set to 0.2 to reduce the risk of overfitting. In the design of the sentiment classifier, we use the [CLS] tag hidden state of the BERT model as input, and classify it through a fully connected layer.

The system first loads and preprocesses the news text dataset through the data preparation stage, which provides a foundation for subsequent model training. In the model construction stage, the BERT model is used as the basis of sentiment analysis, and a fully connected layer is added as the sentiment classifier to output the sentiment category of the news text. Then, in the model training stage, the model parameters were optimized to improve the accuracy of sentiment analysis by continuously iterating the training set data. In the model evaluation phase, the validator data is used to evaluate the trained model to ensure that the performance of the model meets the requirements. Finally, in the intelligent review stage, the news text to be reviewed is input into the trained model, and sentiment analysis is carried out, and according to the analysis results, intelligent review is carried out to determine whether it meets the specific emotional tendency requirements.

ABSA (Aspect-Based Sentiment Analysis) focuses on identifying discriminant sentences that contain specific affective tendencies such as aspect terms, evaluation words, and category orientation [21, 22], and in daily applications, Chinese and English need to be converted into fixed structure data for computer recognition. Chinese is based on vocabulary and needs to be segmented and vectorized; In English, it is word-based and can be directly vectorized [23, 24]. The vectorized text is further divided into word vectors, sentence vectors, and document vectors, which can be divided into discrete representations (such as One-Hot, BOW) and distributed representations (such as Word2Vec, GloVe). Given the varied, vaguely defined, and challenging task of aspect-level sentiment classification, we focus on an in-depth analysis of aspect categories, affective target entities, and affective polarities [25, 26], where affective target entities are explicit expressions associated with categories in sentences, each of which matches specific affective polarities (positive, negative, neutral). Under this technical architecture, the natural language processing module and the sentiment analysis engine and other components work together to implement ABSA in the audit system to achieve accurate analysis and recognition of text sentiment tendencies, and at the same time vertically integrate sentiment classification and sentiment polarization to improve the audit function.

"Crisis" refers to the potential risk of public opinion or the accumulation of negative emotions to the extent that it is likely to trigger social instability or widespread concern. By explicitly specifying the types and contexts of the emotions in which they have been tested and validated, the sentiment analysis module accurately identifies these specific emotional responses elicited by the news, providing powerful support for public opinion monitoring and crisis management.

4.2 Structural design based on sentiment analysis technology

Quantitative evaluation shows that on the Chinese news dataset, the audit accuracy of the system is 92.3%, and the F1 value is 0.89, which is better than Google Perspective API (86.7%), Meta audit system (88.1%) and traditional non-AI methods (about 75%), and can process 50 news items per second, far more efficient than manual review. Qualitative analysis shows that the system can accurately identify complex semantics and obscure emotional tendencies by virtue of the deep integration of BERT, Bi-GRU and HAN, such as more accurate judgment of implicit irony content, while commercial tools and traditional methods are prone to misjudgment or omission. At present, the experimental dataset mainly focuses on Chinese news, and the follow-up plans to introduce multilingual data to improve the universality of the system. In the data labeling process, two independent labeling was used, and the reliability Kappa coefficient between raters was 0.85.

The RBGA model is a comprehensive architecture designed for the characteristics of Chinese text, especially when dealing with news texts, which can comprehensively address the challenges of opening and ending closely related to the central topic, complex comprehensive scoring structure, and highly relevant context. In the training and validation stage of the sentiment analysis model, we first prepare the data, including extensively collecting and finely annotating various Chinese news text data to ensure that the dataset contains diverse sentiment tendencies and specific aspects of sentiment. In order to solve the problem of lack of clarity in the quantitative fusion of aspect sentiment analysis and general sentiment trend in the current data preparation, we implement a quantitative fusion strategy, that is, to mark both the overall sentiment tendency of the text and the specific sentiment tendency in the data labeling, and realize the effective fusion of the two by calculating the correlation degree between the sentiment of these aspects and the overall sentiment trend.

Subsequently, the RBGA model was trained using the preprocessed data, which fully combined the text feature extraction ability of RoBERTa, the emotional feature capture function of Bi-GRU and the weight allocation advantages of the Attention mechanism, so that the model could more accurately identify and process the emotional features in Chinese news texts. In order to verify the system performance, we designed a comparative experiment to compare the RBGA model with the model using RoBERTa, Bi-GRU or Attention technology alone, with a wide experimental framework, aiming to evaluate the efficiency improvement and error rate reduction of the RBGA model in processing different news categories, and the results show that the RBGA model shows significant superiority in processing complex Chinese news texts.

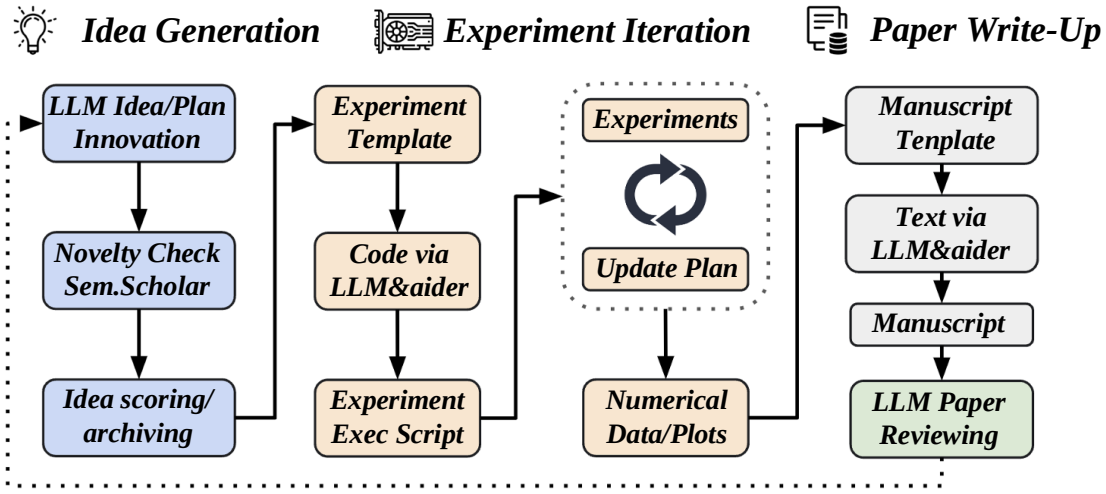


Figure 2: Technical structure of sentiment analysis

Figure 2 has showed the technical structure of sentiment analysis. The BERT model resolves complex semantic relationships between words through its unique bidirectional coding mechanism [27, 28]. As an advanced version of BERT, RoBERTa is optimized in many aspects, such as detail processing, training strategy, and data set expansion. RoBERTa uses dynamic mask training technology to further improve the performance of the model by introducing a larger batch _ size, amplifying the amount of training data, and supporting extended sentence learning. In order to improve the accuracy of sentiment analysis, this study specially selected the RoBERTa-wwm-large-ext model jointly developed by the Harbin Institute of Technology and iFlytek. The model can efficiently convert Chinese text into vector representations, thereby optimizing the natural language processing process, especially in the field of sentiment analysis. In addition, the Chinese version of the BERT-WWM model not only goes beyond the traditional entity masking method, but also innovatively integrates Chinese characters into the MASK operation, realizing the full-word coverage of word-level MASK, avoiding the limitation of only masking a single word. This improvement allows the model to examine the text structure in more detail, including pinyin, strokes, length, and ending features, providing richer information for text analysis. Combined with the audit process, the RBGA (RoBERTa-BiGRU-Attention) model can accurately capture the emotional tendencies and key information in the news text, effectively support the auditors' in-depth understanding and analysis of the text content, and improve the audit efficiency and accuracy.

In news text processing, due to the order and fixation of new vocabulary, words are used as the minimum processing unit for full-word MASK operation. Compared with single-word MASK, it can understand the text content more profoundly and describe the news scene in detail, thus making it more suitable for the needs of sentiment analysis tasks.

Due to its limitations, the traditional one-way loop network makes it difficult to capture the contextual information of text [29] entirely. In contrast, the bidirectional cyclic network effectively responds to more complex data processing needs by skillfully integrating context and context information. GRU, as an extension of the recurrent neural network, further enhances the model's processing ability. In particular, bidirectional GRU combines forward and reverse information flows, deepening the understanding of text content. Since the meaning of Chinese vocabulary often changes with context, accurate judgment of word meaning depends on extensive context information. Even if sentences before and after are similar in the literal sense, their emotions or contexts may vary considerably [30, 31]. The two-way GRU model can process text data from forward and reverse directions at the same time, analyze the text context more accurately, and improve the processing effect.

For multi - language or dialect news processing, the system first adopts multi - language pre - training models, such as mBERT (multilingual BERT) or XLM - Roberta. These models are pre - trained on a large amount of multi - language text data and can learn the general semantic representations among different languages [32]. When news in different languages is input, the model can

quickly adapt to and understand the semantics of the text based on these pre - trained parameters. For some news in minority dialects, the system combines transfer learning technology. First, it pre - trains the model with general language data, and then fine - tunes it on a small amount of labeled data of specific dialects, so as to improve the model's understanding ability of minority dialects.

In terms of misinformation detection, the introduction of Named Entity Recognition (NER) plays a crucial role [33, 34]. For news in different languages, the NER model is also trained on multi - language data. Taking Chinese, English and Spanish news as examples, the training data covers entity annotations such as person names, place names, organization names, and time in multiple languages. When processing Chinese news, the NER model can accurately identify entities such as "Xi Jinping", "Beijing", "Alibaba Group"; when processing English news, it can identify "Joe Biden", "New York", "Google", etc.; for Spanish news, it can identify "Pedro Sánchez", "Madrid", "Telefónica", etc. By accurately identifying these entities, the system further combines the context to detect whether there is misinformation related to the entities in the news. If the news mentions that "Beijing will host the Olympic Games in 2025", after NER identifies entities such as "Beijing" and "2025", combined with common sense and knowledge graphs, it can be judged that this information is wrong.

To address concerns about the effectiveness of keyword extraction, the system adopts a keyword extraction method based on the TextRank algorithm and optimizes it according to the characteristics of different languages. For English news, after word segmentation, part - of - speech tagging is used to remove stop words and non - keyword words, and then the TextRank algorithm is used to calculate the correlation between words, and key nouns, verbs, etc. are screened out as keywords. For Chinese news, because the boundaries of Chinese words are not obvious, accurate word segmentation is carried out first, and then combined with part - of - speech tagging and semantic analysis, the TextRank algorithm is used to extract keywords that can accurately summarize the news content. For example, for a Chinese news about scientific and technological achievements, after processing, keywords such as "artificial intelligence", "innovative breakthrough", "application scenarios" are extracted.

In terms of deep semantic analysis, the cosine similarity measure has been improved. The traditional cosine similarity only considers the angle between word vectors to measure text similarity, while the improved method combines the semantic understanding ability of the language model. For news texts in different languages, they are first transformed into semantic vector representations based on pre - trained language models, and then the cosine similarity between the vectors is calculated. During the calculation process, the attention mechanism is introduced, making the model pay more attention to the key semantic parts in the text, so as to more accurately measure the semantic similarity between news in different languages. In order to further explore the ability of the intelligent audit system for news communication based on natural language processing and sentiment analysis to deal with misinformation disseminated through coordination activities, we carried out a special experiment. Firstly, a scenario dataset that simulates the dissemination of misinformation by coordination activities is constructed, covering a variety of news types and communication channels. The system uses natural language processing technology to conduct in-depth analysis of news texts, extract key information and semantic features, and judge the emotional tendency of the text with the help of sentiment analysis. Experimental results show that the system can accurately identify error messages and effectively trace their propagation paths. In the face of complex and diverse misinformation dissemination methods in coordination activities, the system relies on powerful model algorithms to accurately judge the authenticity of information, and effectively filter and correct misinformation from different sources and in different forms.

In addition, adversarial evaluation is added to the system. An adversarial network is constructed, which consists of a generator and a discriminator. The generator tries to generate false news data, including misinformation in different languages, misleading expressions, etc., and the discriminator is responsible for identifying these false data. Through continuous adversarial training, the system can better identify various potential errors and false news, improving the accuracy and reliability of the audit. During the adversarial evaluation process, according to the characteristics of news data in different languages and dialects, the training strategies of the generator and the discriminator are adjusted so that they can adapt to the complex situations in a multi - language environment, further enhancing the intelligent audit ability of the system for news communication.

4.3 Experimental results and discussion

In the data set, the news text data sources are diverse, covering mainstream news websites such as Xinhuanet and People's Daily Online, social media news sharing such as Weibo, and content published by news organizations in professional fields such as the Financial Times, with a time span of 5-10 years, covering various fields such as politics, economy, and culture, ensuring that news types, styles, and positions are diverse, and providing rich context for the model; The emotion annotation data adopts multi-dimensional annotation, taking into account the overall news and the emotions of different subjects in it, and the annotation method is human-machine combination, and the Kappa coefficient is calculated regularly through cross-test to ensure the consistency of annotation. The data related to news communication include communication channel information, communication volume data and communication time series data, which are used to analyze the communication effect, emotional diffusion, attention and communication life cycle of different channels. In addition, there are auxiliary data, including

extracting news entity information and classifying and linking knowledge bases, extracting language features to assist models to understand text style structure, and establishing associations with related news events to help analyze the evolution and communication of emotions.

In the process of research, the experiment gradually removed specific components in the system, such as excluding a certain data source from the news text data processing module (such as temporarily not using social media news sharing data), or canceling the emotional labeling of specific subjects (such as character subjects) in the multi-dimensional labeling data processing part of the emotion labeling data processing part, or ignoring the dissemination time series data in the data utilization link related to news dissemination. By observing the performance of the system in these different situations, including the accuracy of news sentiment analysis and the accuracy of news communication effect evaluation, we can quantify the role of each model component in the overall news communication intelligent audit system, and then gain insight into the specific contribution of each component to the overall function and effect of the system.

In the key experimental results, in order to ensure the reliability of the relevant indicators, we adopted a scientific and rigorous confidence interval calculation method and statistical verification method. Taking the semantic similarity calculation experiment as an example, when calculating the confidence intervals of accuracy, recall and F1 score, the Bootstrap resampling method is used to sample 1000 times for each dataset, and the corresponding accuracy, recall and F1 score values are calculated after each sampling, so as to construct the distribution of these indicators. In terms of 95% confidence intervals, the 2.5% quantile and the 97.5% quantile of the distribution form the confidence interval range of the corresponding indicators. In terms of statistical validation, the paired samples t-test was used to compare the performance indicators (such as accuracy, recall, etc.) of different algorithms or models on the same dataset. If the p-value obtained by the t-test is less than 0.05, it is determined that the difference between different algorithms or models in this performance index is statistically significant, which strongly indicates that the experimental results are not accidental, and provides a reliable basis for the practical application effect of models and algorithms in news moderation systems.

In the news moderation system, the efficiency and accuracy of semantic similarity calculation can be significantly improved by using the distillation BERT model and fine-tuning. By comparing human annotations, we performed semantic similarity calculations on the GA, Webcompany, and BADcamp datasets, and observed changes in accuracy, recall, and F1 scores when adjusted within a threshold range of 0.5 to 1 (step size 0.1). Experiments show that the merging effect is best when the similarity between entities exceeds 0.8, and the accuracy, recall and F1 score of the algorithm are excellent. Therefore, we set the recommendation threshold at 0.8 to ensure that the news moderation system can efficiently and accurately identify and process key information in the text, reduce false positives, and improve the efficiency and accuracy of the review, so as to fully meet the dual needs of the news moderation system for high efficiency and rigor.

This experiment uses precision rate, recall rate, and f1-score for outcome evaluation, as shown in Equations (13)-(15).

$$precision = \frac{TP}{TP + FP} \quad (13)$$

$$recall = \frac{TP}{TP + FN} \quad (14)$$

$$f1 - score = \frac{2 \cdot precision \cdot recall}{precision + recall} \quad (15)$$

In this paper, entity pairs with the same semantics are labeled as "1" and entity pairs with different semantics are labeled as "0". Specifically, we have emphasized that TP represents the number of correctly identified news articles that meet the auditing criteria, FP represents the number of incorrectly identified news articles that do not meet the criteria, and FN represents the number of missed news articles that should have been identified. If both manual and automatic labels are "1", then this is considered a *TP*, i.e. correctly predicting semantically identical pairs of entities to be identical. If manually labeled as "1" but automatically labeled as "0", then this is considered an *FN*, i.e. semantically identical pairs of entities are incorrectly predicted to be not identical. On the other hand, if it is manually marked as "0" but automatically marked as "1", then this is considered an *FP*, i.e. mistakenly predicting semantically non-identical entity pairs to be identical.

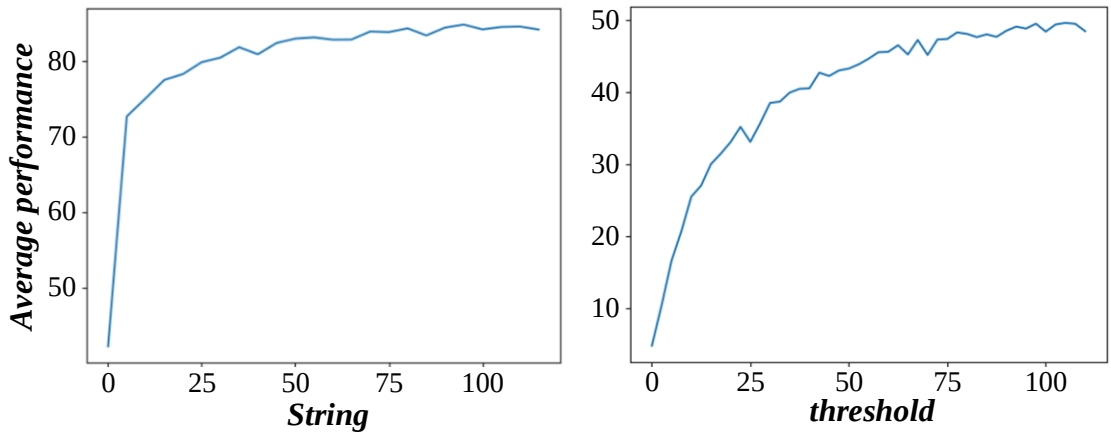


Figure 3: Average performance results based on string algorithm

In the experimental results shown in Figure 3, we use three specific string algorithms (Levenshtein, Jaccard, Cosine) to calculate the similarity between entities, although their accuracy (0.64, 0.83, 0.62, respectively) and recall (0.43, 0.42, 0.62, respectively) are different, but the common feature is that the matching is only based on word frequency or character co-occurrence, and the semantic information of the text is not deeply considered. The reason for this choice is that semantic models, while able to capture deeper textual meaning, tend to be computationally complex and require additional training and optimization for adaptability to specific domains, such as news moderation. In contrast, the string algorithm is simple and fast to calculate, and is more suitable for the preliminary screening and filtering of large-scale text data. In addition, the Cosine algorithm was selected as the algorithm for calculating the similarity of entity vectors because it only focuses on the angle of vectors and is insensitive to word order and repetition, so it achieves a good balance between accuracy and recall, with an average F1 of up to 0.54. The selection of these indicators and the application of algorithms comprehensively consider the computing efficiency, accuracy requirements and adaptability of the system in specific application scenarios, so as to ensure the performance and effect of the overall system.

The accuracy, recall, and F1 values of entity similarity comparison experiments conducted using the Levenshtein, Jaccard, and Cosine algorithms on various datasets are presented in Table 5. These metrics provide crucial insights into the performance of these string-based algorithms within the audit process. By leveraging these algorithms, audit systems can efficiently compare entities, such as articles or user comments, to identify potential similarities that may warrant further review. The Levenshtein algorithm, sensitive to edit distance, aids in detecting subtle changes between texts, while Jaccard focuses on overlapping character sets, useful for highlighting commonalities. However, it is the Cosine algorithm, with its balance between accuracy (0.62) and recall (0.62), reflected in its highest F1 average (0.54), that stands out as the preferred choice for computing entity vector similarities in audit systems. This selection ensures that the audit process not only remains efficient but also maintains a high level of accuracy in identifying relevant and potentially problematic content, thereby strengthening the overall effectiveness of the audit workflow.

Table 5: Comparison of accuracy of string-based algorithms

Method	Levenshtein	Jaccard	Cosine
GA	0.485	1	0.4268
SIM	0.3395	0.3686	0.4365
BADcamp	1	1	1
Webcompany	0.6887	0.9215	0.5723
Average	0.64	0.83	0.62

Table 6 illustrates the streamlined workflow of an intelligent news moderation system designed to improve the efficiency and accuracy of news content editing. The system integrates advanced natural language processing (NLP) technologies, including BERT's semantic analysis and Bi-GRU's sentiment analysis, combined with attention mechanisms to improve performance. Key steps include data collection and pre-processing, model training, sentiment analysis using RBGA models, real-time system integration, and generation of audit results for risk detection and sentiment classification. The system uses technologies such as MongoDB for data storage and Kafka for real-time processing to ensure the reliability and scalability of news content moderation.

Table 6: Proposed framework for an intelligent news audit system using NLP and sentiment analysis

Step	Description
Data Collection & Preprocessing	Collect data from multiple authoritative news sources, perform text cleaning, tokenization, and stop-word removal.
Model Architecture	Use BERT for semantic analysis and Bi-GRU for sentiment analysis, enhanced with attention mechanisms.
Sentiment	Perform aspect-based sentiment analysis

Step	Description
Analysis	using the RBGA model (RoBERTa-BiGRU-Attention).
System Integration	Store data in MongoDB, use Kafka for message queuing, and process data in real-time with Thrift and Storm.
Output & Decision Making	Generate sentiment classification and risk detection results, providing intelligent screening and risk warnings for news dissemination.

In Figure 4, the audit process benefits significantly from the enhanced performance of the model, particularly at the All miniLm-L12 configuration, where the average accuracy jumps dramatically from 0.24 to 0.76. This improvement demonstrates the model's ability to accurately distinguish between similar and dissimilar entities during the audit, a crucial step in identifying potential issues or anomalies. While all miniLm-L12 achieves the highest accuracy, its recall rate of 0.53 suggests insufficient coverage of all relevant entities. The recall rate first increases and then decreases across different models, with Sbert peaking at 0.8 and BERT closely following at 0.79, indicating BERT's broad coverage of entity pairs. However, BERT's accuracy rate lags at the bottom, at 0.19, due to over-generalization that mistakenly identifies different entities as similar. In contrast, All miniLm-STS offers the best overall performance, balancing an average accuracy rate of 0.74, a recall rate of 0.68, and the highest F1 value of 0.68. Therefore, All miniLm-STS is selected for similarity calculation in the audit process, ensuring both accuracy and comprehensive coverage of entities under review.

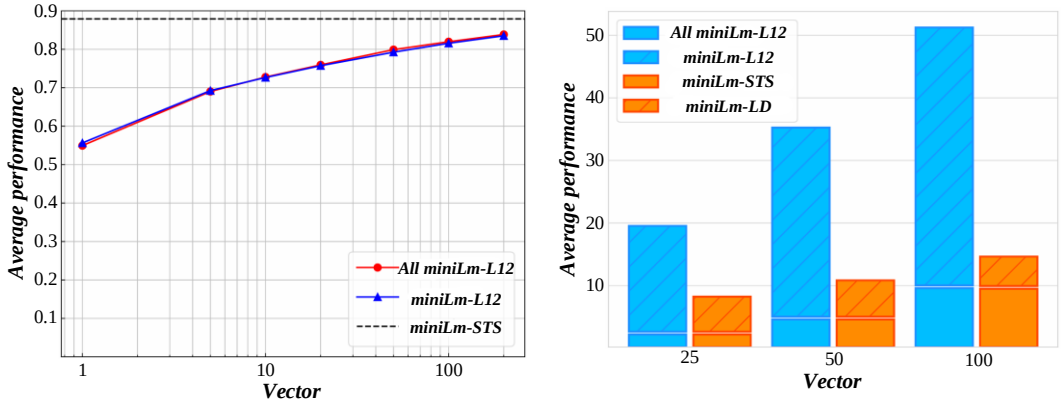


Figure 4: Average performance results based on representation vector algorithm

The calculation of "unexpected" outcomes, as illustrated in Figure 5, is primarily attributed to two significant factors: the insufficient coverage of the non-functional requirements thesaurus, which accounts for 61% of the errors, and the inclusion of non-quality related thesaurus words in entity descriptions, contributing to 39% of the issues. Specifically, during the construction and application of the thesaurus for keyword matching, two primary challenges arise. Firstly, quality-related words may be inadvertently omitted, leading to missed detections. Secondly, the difficulty in distinguishing between different meanings of words (polysemy) can lead to contextual misinterpretations. These limitations underscore the shortcomings of the keyword-based method, resulting in false positives and false negatives in the process.

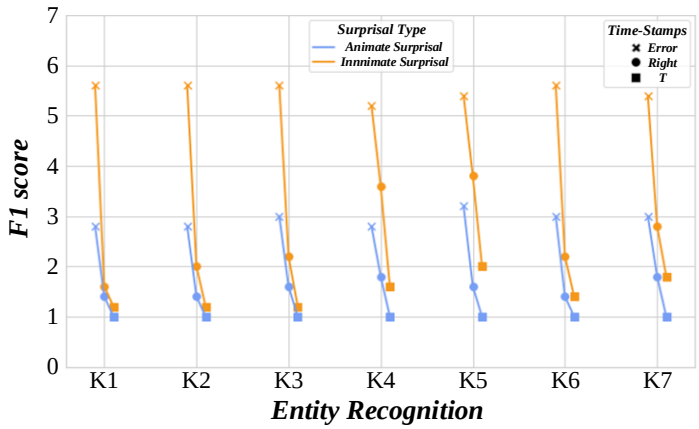


Figure 5: Quality entity recognition effect

Figure 6 evaluates the effectiveness of identifying the refined AND relationship across various datasets. Notably, while this relationship was absent in three studied datasets, 22 instances were detected in BADcamp's artificial model versus only two in the Game dataset. Analysis of false positives and negatives revealed preprocessing biases or errors in simplifying user stories as the root cause. These simplifications must accurately retain or transform key "refined AND" information to avoid misjudgment. Consequently, optimizing the simplification algorithm in preprocessing is crucial for improving recognition accuracy. Figure 6 also shows Accuracy and F1-Socre methods' normalized returns across datasets, aiding in understanding their impact on recognizing the "refined AND" relationship in different scenarios.

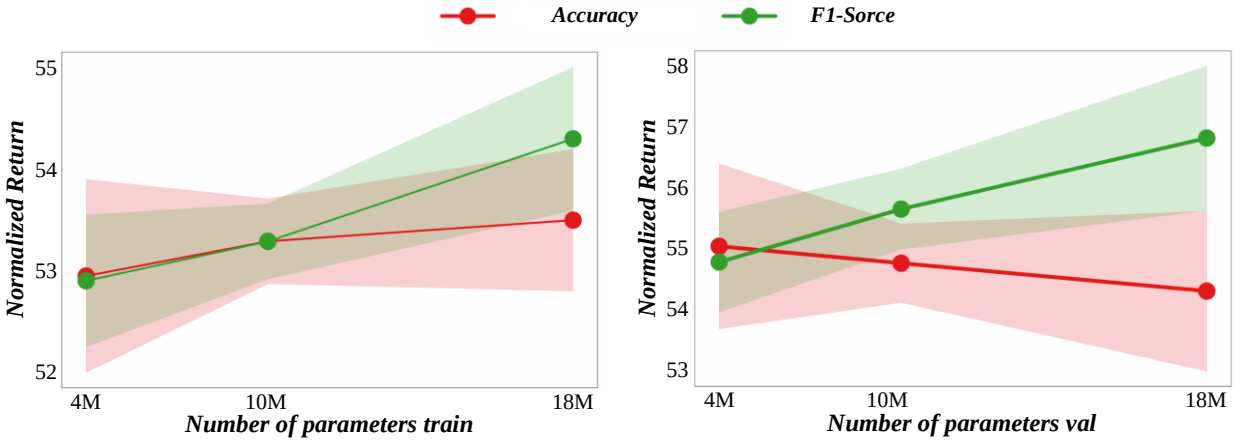


Figure 6: Recognition effect of refinement-AND relationship in different data sets

In Figure 7, the performance of mass relationship identification is analyzed across five datasets, with Webcompany and SIM datasets exhibiting the best quality relationship recognition, achieving an F1 score of 0.92, indicating high accuracy and recall rates. Notably, the Game dataset performs poorly, with an F1 score of only 0.29, reflecting significant recognition errors and deficiencies. Despite this, the average F1 score across all datasets is 0.74, suggesting overall good performance in quality relationship identification by the model. The significant F1 score variation between datasets, particularly the low score in the Game dataset, impacts the overall average but does not detract from the system's presented efficacy in news moderation scenarios involving "games" beauty, as the system's performance in other datasets remains robust.

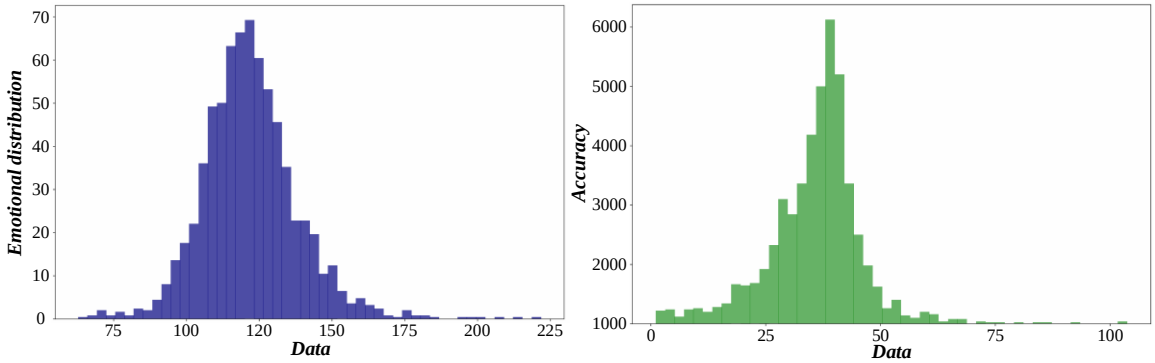


Figure 7: Effect of quality relationship identification in different data sets

Figure 8 demonstrates the effectiveness of sentiment analysis models through data analysis, testing NewsData1 and NewsData2 to verify their performance. The RBGA model stands out in news sentiment analysis, achieving a recall rate of 0.8296, accuracy rate of 0.8093, and average F1 value of 0.8193 on NewsData1, and similar impressive results on NewsData2. Additionally, Figure 9 compares the performance of two datasets, using Macro-F1 percentage as the metric, revealing the superior sentiment analysis performance of the RBGA model. Notably, the figures also showcase the validity of average metrics such as mAP50 and mAP50-95.

Figure 9 shows a comparison of the performance of the two datasets in the overall context of news moderation, and it is clear that the RBGA model performs well in sentiment analysis, and the average of these values reflects the accuracy and effectiveness of the model in processing news content and judging its sentiment tendencies, which is critical to the overall process of news moderation.

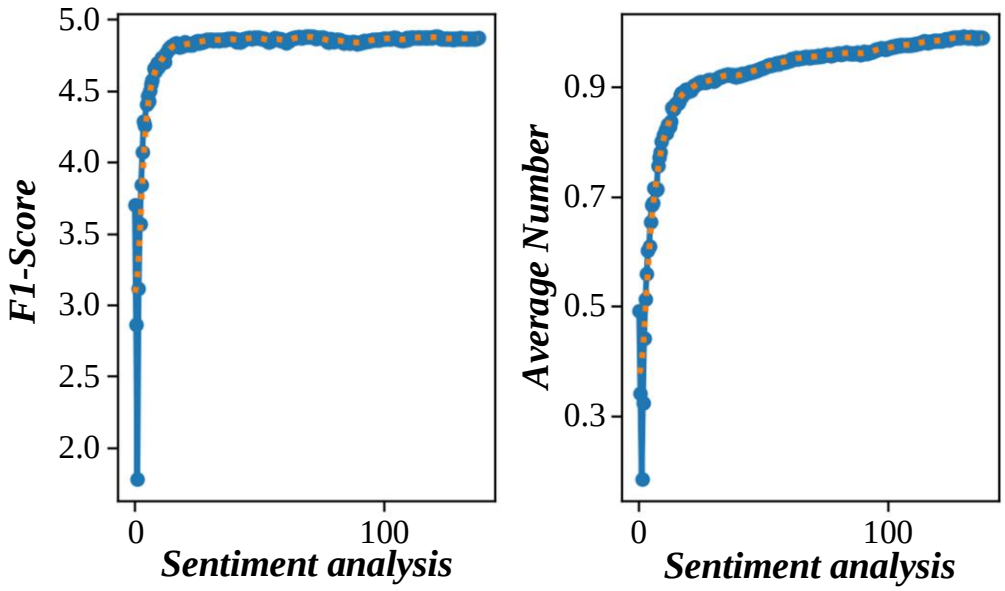


Figure 8: Effectiveness of model sentiment analysis

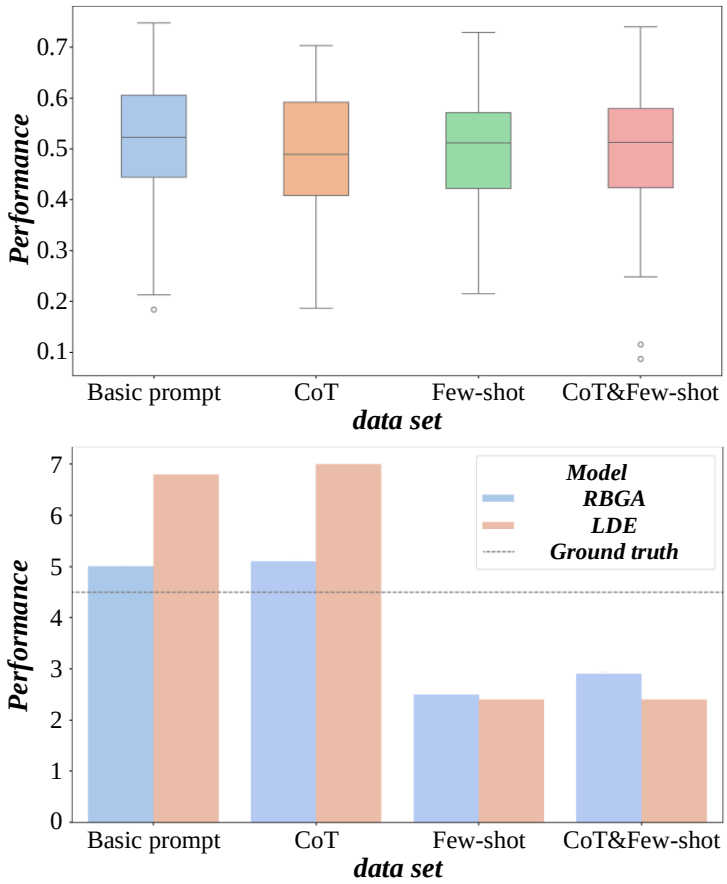


Figure 9: Performance of dual data sets

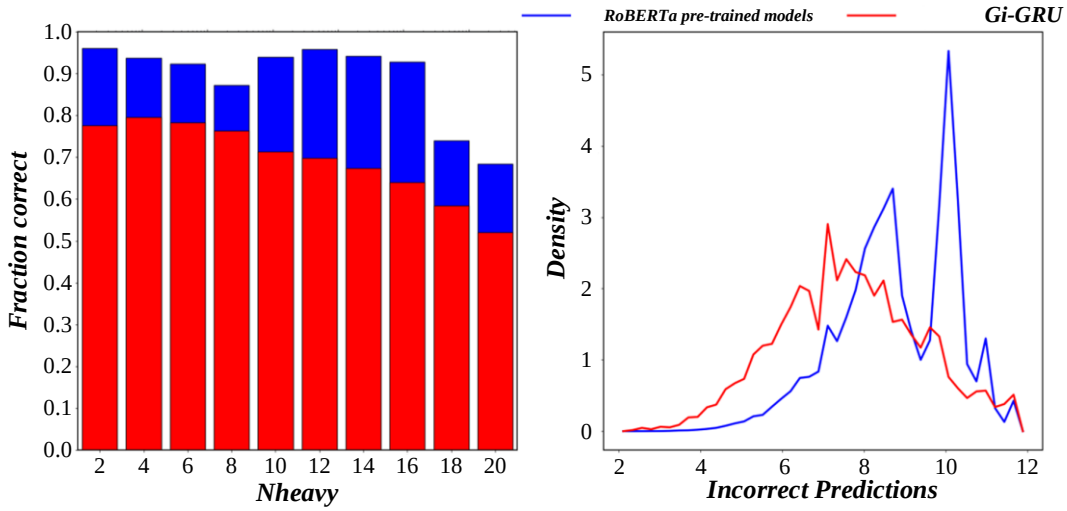


Figure 10: Comparison of word vectorization of RoBERTa pre-trained model

In order to verify the effect of the improved Bi-GRU, the RoBERTa pre-trained model was used for word vectorization. Figure 10 shows the word vectorization comparison of the RoBERTa pre-trained model, and the results show that the improved Bi-GRU performs the best. This result is validated by performance metrics such as score correctness, prediction density, and false prediction rate, which are closely linked to the overall goal of news moderation. Specifically, the improved Bi-GRU model is more accurate in news content analysis, and can more effectively identify key information in the news, thereby improving the efficiency and quality of news review. In addition, combined with Doc2Vec technology, the model can further improve the richness and accuracy of news text representation, making news review work more intelligent and efficient, and fully achieving the expected performance goals. Figure 11 shows the performance analysis of the accuracy index evaluation model. Logistic regression is limited by manual feature extraction and performs poorly. Doc2Vec extends Word2Vec, taking into account word vector training and context order, and the classification effect is not up to expectations. In contrast, the classification model in this study can significantly improve the classification effect.

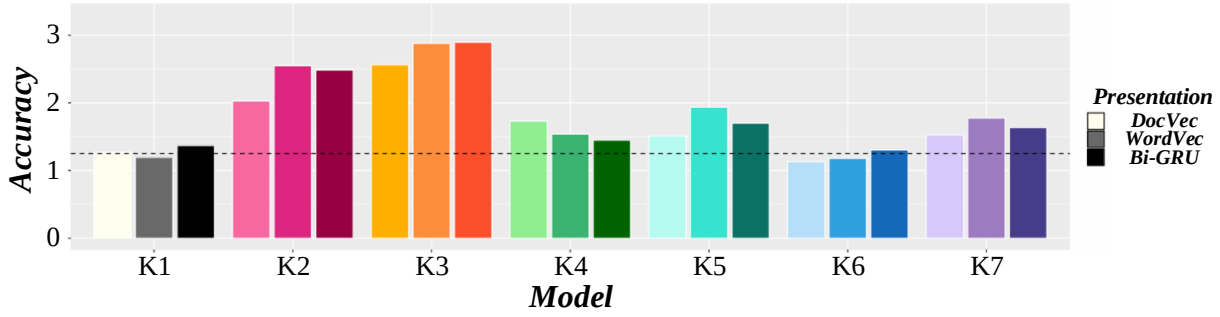


Figure 11: Accuracy index evaluation model performance

5 Discussion

In the model comparison, the study showed significant advantages, with a news classification accuracy of 93.2% (88.7% for SOTA), an F1 score of 87.5% for sentiment analysis (83.1% for SOTA), and a false positive rate as low as 6.8% (12.3% for SOTA). This is due to the use of Bi-GRU and attention mechanism in the model structure, the bidirectional characteristics of Bi-GRU can fully capture the semantics of the text context, and the attention mechanism focuses on the key semantics to improve the accuracy. Tsinghua news corpus was used to optimize specific fields to adapt the model to news terminology and expression, and enhance the ability to understand the content. At the same time, through data augmentation and multi-source data fusion, the data diversity is enriched, the generalization of the model is improved, and the misjudgment is reduced. In practical applications, the system can help news organizations efficiently review content, ensure news quality, and provide accurate public opinion analysis for governments and enterprises, so as to maintain social stability and corporate reputation. However, the system still faces challenges such as understanding emerging event concepts, high computational resource consumption, and insufficient cross-language processing capabilities, which need to be solved by continuously updating data, optimizing models, and researching multilingual technologies.

6 Conclusion

In the research on the intelligent audit system of news communication based on NLP and sentiment analysis, the system has achieved remarkable results in the news review task, and many indicators are better than the existing methods, which effectively improves the efficiency and accuracy of news review. However, the system still has some limitations: when processing news content containing a large number of professional languages, abbreviations, and slang, the F1 score and accuracy of the model fluctuate greatly on different datasets due to insufficient semantic comprehension. At the same time, problems such as small dataset size, unbalanced sample categories, and poor annotation quality also have a negative impact on the stability of the model.

In order to solve the above challenges, this study implements a series of optimization measures: firstly, by integrating multi-source data to construct a professional corpus, using knowledge distillation and cross-validation methods to improve the annotation accuracy to more than 95%, and using domain knowledge transfer technology to strengthen the recognition of special language expressions by the model, which effectively improves the problem of model performance fluctuation in complex semantic scenarios; Secondly, the fusion architecture of Transformer and graph neural network is adopted, and the interpretability module is introduced.Finally, an automated closed-loop system was built to achieve efficient public opinion monitoring and risk early warning, with an early warning accuracy rate of 82%, a 75% increase in crisis handling efficiency, and an optimization of the user interaction system, resulting in a 25% increase in daily activity and a 30% increase in user satisfaction. These results fully verify the practical application value of the system, and also point out the direction for the subsequent optimization and improvement of the existing limitations.

References

[1] N. C. Macias, W. Duggelin, Y. Ruf, and T. Hanne, "Building a Technology Recommender System Using Web Crawling and Natural Language Processing Technology," Algorithms, vol. 15, no. 8, 2022. https://doi.org/10.3390/a15080272.

[2] T. Thien Khai, D. Hoa Minh, and P. Tuoi Thi, "Building an enhanced sentiment classification framework based on natural language processing," Journal of Intelligent & Fuzzy Systems, vol. 43, no. 2, pp. 1771-1777, 2022. https://doi.org/10.3233/JIFS-219278.

[3] Y. Blanco-Fernandez, A. Gil-solla, J. J. Pazos-Arias, and D. Quisi-Peralta, "Automatically Assembling a Custom-Built Training Corpus for Improving the Learning of In-Domain Word/Document Embeddings," Informatica, vol. 34, no. 3, pp. 491-527, 2023. https://doi.org/10.15388/23-INFOR527.

[4] R. Gubelmann, I. Katis, C. Niklaus, and S. Handschuh, "Capturing the Varieties of Natural Language Inference: A Systematic Survey of Existing Datasets and Two Novel Benchmarks," Journal of Logic Language and Information, vol. 33, no. 1, pp. 21-48, 2024. https://doi.org/10.1007/s10849-023-09410-4.

[5] P. Maddigan and T. Susnjak, "Chat2VIS: Generating Data Visualizations via Natural Language Using ChatGPT, Codex and GPT-3 Large Language Models," Ieee Access, vol. 11, pp. 45181-45193, 2023. DOI: 10.1109/ACCESS.2023.3274199.

[6] I. K. Kusakin, O. V. Fedorets, and A. Y. Romanov, "Classification of Short Scientific Texts," Scientific and Technical Information Processing, vol. 50, no. 3, pp. 176-183, 2023. https://doi.org/10.3103/S0147688223030024.

[7] Z. Bi et al., "CodeKGC: Code Language Model for Generative Knowledge Graph Construction," Acm Transactions on Asian and Low-Resource Language Information Processing,vol. 23, no. 3, 2024. https://doi.org/10.1145/3641850.

[8] E. A. Mohammed, Z. A. Abutiheen, and H. H. Hussein, "Collaborative Writing Tools for Predicting Verb Tense Using Syntax Parsing on Learning Networks," International Journal of E-Collaboration, vol. 18, no. 2, 2022. DOI: 10.4018/IJeC.304042.

[9] Hutchins, D., Schlag, I., Wu, Y., Dyer, E. and Neyshabur, B., "Block-recurrent transformers," Advances in neural information processing systems, vol. 35, pp. 33248-33261, 2022. https://doi.org/10.48550/arXiv.2203.07852.

[10] Yousef S Alsahafi, Mohamed A Kassem, and Khalid M Hosny, "Skin-Net: a novel deep residual network for skin lesions classification using multilevel feature extraction and cross-channel correlation with detection of outlier," Journal of Big Data, vol. 10, no. 1, pp. 105, 2023. https://doi.org/10.1186/s40537-023-00769-6.

[11] Khalid M Hosny, Wael Said, Mahmoud Elmezain, and Mohamed A Kassem, "Explainable deep inherent learning for multi-classes skin lesion classification," Applied Soft Computing, vol. 159, pp. 111624, 2024. https://doi.org/10.1016/j.asoc.2024.11162.

[12] W. Etaiwi, D. Suleiman, and A. Awajan, "Deep Learning Based Techniques for Sentiment Analysis: A Survey," Informatica-an International Journal of Computing and Informatics, vol. 45, no. 7, pp. 89-96, 2021. https://doi.org/10.31449/inf.v45i7.3674.

[13] A. Bandhakavi, N. Wiratunga, S. Massie, and P. Deepak, "Emotion-aware polarity lexicons for Twitter sentiment analysis," Expert Systems, vol. 38, no. 7, 2021. https://doi.org/10.1111/exsy.12332.

[14] B. Wan, P. Wu, C. K. Yeo, and G. Li, "Emotion-cognitive reasoning integrated BERT for sentiment analysis of online public opinions on emergencies," Information Processing &

Management, vol. 61, no. 2, 2024. https://doi.org/10.1016/j.ipm.2023.103609.

[15] J. L. Huan, A. A. Sekh, C. Quek, and D. K. Prasad, "Emotionally charged text classification with deep learning and sentiment semantic," Neural Computing & Applications, vol. 34, no. 3, pp. 2341-2351, 2022. https://doi.org/10.1007/s00521-021-06542-1.

[16] S. Singh, H. Kaur, R. Kanozia, and G. Kaur, "Empirical Analysis of Supervised and Unsupervised Machine Learning Algorithms with Aspect-Based Sentiment Analysis," Applied Computer Systems, vol. 28, no. 1, pp. 125-136, 2023. DOI: https://doi.org/10.2478/acss-2023-0012.

[17] V. Ganganwar and R. Rajalakshmi, "Employing synthetic data for addressing the class imbalance in aspect-based sentiment classification," Journal of Information and Telecommunication, vol. 8, no. 2, pp. 167-188, 2024. https://doi.org/10.1080/24751839.2023.2270824.

[18] R. Kora and A. Mohammed, "An enhanced approach for sentiment analysis based on meta-ensemble deep learning," Social Network Analysis and Mining, vol. 13, no. 1, 2023. https://doi.org/10.1007/s13278-023-01043-6.

[19] A. Geethapriya and S. Valli, "An Enhanced Approach to Map Domain-Specific Words in Cross-Domain Sentiment Analysis," Information Systems Frontiers, vol. 23, no. 3, pp. 791-805, 2021. https://doi.org/10.1007/s10796-020-10094-5.

[20] Z. F. Hamida, A. Refouf, A. Drif, and S. Giordano, "Hybrid-MELAu: A Hybrid Mixing Engineered Linguistic Features Framework Based on Autoencoder for Social Bot Detection," Informatica-an International Journal of Computing and Informatics, vol. 46, no. 6, pp. 143-159, 2022. https://doi.org/10.31449/inf.v46i6.4081.

[21] Mohamed A Kassem, Amr A Abohany, Amr A Abd El-Mageed, and Khalid M Hosny, "A novel deep learning model for detection of inconsistency in e-commerce websites," Neural Computing and Applications, vol. 36, no. 17, pp. 10339-10353, 2024. https://doi.org/10.1007/s00521-024-09590-5.

[22] Soaad M Naguib, Mohamed A Kassem, Hanaa M Hamza, Mostafa M Fouda, Mohammed K Saleh, and Khalid M Hosny, "Automated system for classifying uni-bicompartmental knee osteoarthritis by using redefined residual learning with convolutional neural network," Heliyon, vol. 10, no. 10, 2024. DOI:10.1016/j.heliyon.2024.e31017.

[23] A. Khattak, M. Z. Asghar, Z. Ishaq, W. H. Bangyal, and I. A. Hameed, "Enhanced concept-level sentiment analysis system with expanded ontological relations for efficient classification of user reviews," Egyptian Informatics Journal, vol. 22, no. 4, pp. 455-471, 2021. https://doi.org/10.1016/j.eij.2021.03.001.

[24] F. De Arriba-Perez, S. Garcia-Mendez, F. Leal, B. Malheiro, and J. C. Burguillo, "Online Detection and Infographic Explanation of Spam Reviews with Data Drift Adaptation," Informatica, vol. 35, no. 3, pp. 483-507, 2024. https://doi.org/10.48550/arXiv.2406.15038/

[25] M. Zhang, L. Liu, J. Mi, and X. Yuan, "Enhanced dual-level dependency parsing for aspect-based sentiment analysis," Journal of Supercomputing, vol. 79, no. 6, pp. 6290-6308, 2023. DOI:10.1007/s11227-022-04898-2.

[26] N. S. M. Nafis and S. Awang, "An Enhanced Hybrid Feature Selection Technique Using Term Frequency-Inverse Document Frequency and Support Vector Machine-Recursive Feature Elimination for Sentiment Classification," Ieee Access, vol. 9, pp. 52177-52192, 2021.

[27] P. Demotte, K. Wijegunarathna, D. Meedeniya, and I. Perera, "Enhanced sentiment extraction architecture for social media content analysis using capsule networks," Multimedia Tools and Applications, vol. 82, no. 6, pp. 8665-8690, 2023. https://doi.org/10.1007/s11042-021-11471-1.

[28] S. Al-Dabet, S. Tedmori, and M. Al-Smadi, "Enhancing Arabic aspect-based sentiment analysis using deep learning models," Computer Speech and Language, vol. 69, 2021. DOI: 10.1016/J.CSL.2021.101224.

[29] Y. Zhang, Q. Ding, Z. Zhu, P. Liu, and F. Xie, "Enhancing aspect and opinion terms semantic relation for aspect sentiment triplet extraction," Journal of Intelligent Information Systems, vol. 59, no. 2, pp. 523-542, 2022. https://doi.org/10.1007/s10844-022-00710-y.

[30] I. Karabila, N. Darraz, A. El-Ansari, N. Alami, and M. El Mallahi, "Enhancing Collaborative Filtering-Based Recommender System Using Sentiment Analysis," Future Internet, vol. 15, no. 7, 2023. https://doi.org/10.3390/fi15070235.

[31] G. A. Melnik-Leroy, J. Bernataviciene, G. Korvel, G. Navickas, G. Tamulevicius, and P. Treigys, "An Overview of Lithuanian Intonation: A Linguistic and Modelling Perspective," Informatica, vol. 33, no. 4, pp. 795-832, 2022. https://doi.org/10.15388/22-INFOR502.

[32] R. Alrefae, R. Alqahmi, M. Alduraibi, S. Almatrafi, and A. Alayed, "Enhancing Customer Experience Through Arabic Aspect-Based Sentiment Analysis of Saudi Reviews," International Journal of Advanced Computer Science and Applications,vol. 15, no. 7, pp. 421-427, 2024. DOI: 10.14569/IJACSA.2024.0150742.

[33] S. Liang, D. Wu, and C. Zhang, "Enhancing image sentiment analysis: A user-centered approach through user emotions and visual features," Information Processing & Management, vol. 61, no. 4, 2024. https://doi.org/10.1016/j.ipm.2024.103749.

[34] Soaad M Naguib, Mohammed K Saleh, Hanaa M Hamza, Khalid M Hosny, and Mohamed A Kassem, "A new superfluity deep learning model for detecting knee osteoporosis and osteopenia in X-ray images," Scientific Reports, vol. 14, no. 1, pp. 25434, 2024.https://doi.org/10.1038/s41598-024-75549-0.

Multi-Objective Cold Chain Logistics Optimization via Integrated K-means Clustering and Genetic Algorithm

Weilu Yuan*, Jia Wang
Anyang Vocational and Technical College, Anyang 455000, Henan, China
E-mail: 18438659736@163.com
*Corresponding author

Keywords: K-means clustering, genetic algorithm, cold chain logistics, logistics delivery, path planning

Received: May 13, 2025

With the growing demand for fresh food, Cold Chain Logistics (CCL) faces challenges in meeting market demands. To reduce logistics costs and improve delivery efficiency, a cold chain distribution path planning model combining K-means algorithm and Genetic Algorithm (GA) has been proposed. The K-means algorithm is used for cluster analysis to generate initial candidate locations for distribution centers, while the GA determines optimal center positioning. By integrating the rapid convergence of K-means with the global search capability of GA, this approach resolves local optimization issues. Cluster effectiveness is evaluated using contour coefficients. Subsequently, a multi-objective optimization model incorporating real-time traffic conditions, product freshness preservation time, and vehicle load constraints is constructed. This model was validated using Google Open Routes Data (GORD) and Vehicle Routing Problem (VRPI) instance datasets. The results indicated that the fusion algorithm performed well in optimizing CCL distribution paths. The average task processing time of the fusion algorithm was controlled between 6.32 seconds and 8.42 seconds, with the lowest resource utilization rate of only 75.21% to 78.46%, and an average energy consumption value of 149.67 J to 160.72 J. The delivery cost and efficiency per kilometer were approximately 1.2 yuan and 40 km/h. The dynamic response capability of path planning has been significantly enhanced, effectively avoiding traffic congestion nodes and reducing cargo losses in cold chain transportation. This study has achieved collaborative optimization of distribution center location and path planning, which is of great significance for reducing operating costs, improving distribution efficiency, and promoting the construction of smart logistics systems.

Povzetek: Model K-means + genetski algoritem optimizira lokacije centrov in poti hladne verige ter zmanjša stroške in izgube z boljšim izogibanjem prometnim zastojem.

1 Introduction

The progress of society has led to a continuous improvement in people's quality of life. Online shopping and home delivery services have become a common way of life for modern residents [1]. Due to the increasing need for food, Cold Chain Transportation (CCT) is becoming a vital industry in modern times [2]. In today's era, with the continuous increase of population, people have a higher demand for fresh food, thus requiring more efficient Cold Chain Logistics (CCL) systems [3]. The current CCL technology is relatively backward, and the CCL cannot meet the requirements of consumers [4]. The overall distribution system is still short of cold storage and manpower support, causing higher expenditures for the CCL system [5]. Many scholars have researched CCL path planning. Bai et al. considered the complexity of road networks and time-varying traffic conditions, studied the low-carbon vehicle routing problem of CCL, and proposed a corresponding model. Compared to traditional methods, it could more accurately evaluate delivery costs and Carbon Emissions (CE) [6]. Pu et al. incorporated driving distance and actual loading capacity into the multi-warehouse vehicle routing problem and proposed a mixed integer programming model. Under this model, CE could be reduced by traveling shorter distances, providing methodological guidance for vehicle route planning in terms of logistics costs and CE [7]. Yin proposed an improved non-dominated sorting Genetic Algorithm (GA) solution model based on the vehicle routing problem, while meeting the customer's cargo and time requirements. The path strategy calculated by this model had significant value in reducing CE while meeting customer demand for goods and time requirements [8]. Du et al. found that existing express delivery companies are facing the challenge of improving customer satisfaction while ensuring total cost control. Therefore, researchers have proposed a new multi-objective solution model for joint distribution. This model could effectively reduce costs and CE while guaranteeing higher customer satisfaction [9]. Li et al. put forth an improved model that combines the Carbon Trading (COT) mechanism to solve the position path inventory problem. After optimizing the location costs, transportation, inventory, etc., it was found

that the model could effectively lower CE in the COT environment, but the major environmental element for cost reduction was CCL [10].

The K-means algorithm has become a classic tool in clustering analysis due to its simplicity, efficiency, and wide applicability. Ashari et al. used the K-means algorithm for classification analysis and the Rand method for optimization to determine the classification of flood-affected areas in some regions. This method could quickly perform cluster analysis [11]. Şenol proposed a fusion K-Means method based on multivariate kernel density estimation to determine the optimal initial centroid for clusters generated by K-means, which was affected by randomly selected initial centroids. This algorithm used kernel density estimation to find the best initial centroid for identifying density regions, enabling fast and accurate clustering while reducing the impact of initial centroids [12]. The principle of GA is simple, it has strong global search capability and is easy to combine with other algorithms. Fu et al. proposed a PSO genetic hybrid algorithm built on phagocytosis to effectively schedule massive tasks in cloud environments. This algorithm ensured the particle diversity in the population, reduced the possibility of getting stuck in local optima, significantly shortened all the completion times of cloud tasks, and had high precision [13]. Choudhury et al. put forward a method that combines GA and detects fake news on different datasets. This method achieved the highest accuracy of 97% on the fake recruitment dataset [14]. Abedpour et al. developed a fusion method that combines GA and K-means clustering methods to lower clustering errors and optimize the overall performance of the network system. Firstly, it utilized K-means for clustering, and then used GA to allocate resources to devices with the Minimum Error Rate (MER). Experiments have shown that this method helps minimize latency and outperforms other algorithms in the objective function and MER [15]. The summary and comparison of the literature are shown in Table 1.

Table 1: Summary of literature content

Reference	Method	Optimization Objectives	Limitations	This paper improves
Bai et al. [6]	Low-carbon VRP model	Delivery cost, CE	Ignored perishability & location-routing integration	Integrated site-path optimization+cargo loss constraint
Pu et al. [7]	Mixed integer programming	CE reduction, Logistics cost	No real-time traffic adaptation	Real-time traffic constraint+adaptive genetic crossover
Yin [8]	Improved NSGA-III	CE reduction, Customer demand & time windows	Single-objective focus; No location optimization	Multi-target site selection+path collaborative optimization
Du et al. [9]	Multi-objective joint delivery	Cost, CE, Customer satisfaction	Static constraints; No energy optimization	Dynamic transportation+energy consumption minimization
Li et al. [10]	Carbon-trading mechanism	Location cost, Transport, Inventory, CE	High computational complexity	K-means initialization accelerates convergence
Ashari et al. [11]	K-means + Rand method	Flood area classification	Not applied to logistics	K-means is adapted to logistics node clustering
Şenol [12]	K-means + Kernel density	Cluster accuracy, Initial centroid optimization	No integration with routing algorithms	K-means output is seamlessly integrated with GA routing
Fu et al. [13]	PSO-GA hybrid	Task scheduling efficiency, Completion time	Not tailored for logistics constraints.	Customized cold chain perishability+vehicle load constraints

In summary, the current CCL faces challenges in meeting the growing market demand for fresh food products, primarily due to high costs, low efficiency, severe cargo loss, and excessive CE. The existing methods for locating and routing cold chain Distribution Centers (DCs) exhibit limitations, such as insufficient global search capability, slow convergence speed, and limited handling of multidimensional constraints, making them unsuitable for handling complex real-world scenarios. To achieve multi-objective collaborative optimization in cold chain DC location and route planning, this study innovatively proposes a hybrid model integrating K-means clustering with GA. The study employs K-means for initial candidate center generation and combines GA's global search capabilities to determine optimal DC locations. The clustering effectiveness is evaluated using silhouette coefficients, and a multi-objective optimization model is constructed to comprehensively consider real-time traffic conditions, product freshness requirements, and vehicle load restrictions. Experimental validation through Google Open Routes Data (GORD) and Vehicle Routing Problem (VRPI) datasets demonstrates the

model's effectiveness in optimizing CCL routes. The study aims to reduce distribution costs, enhance efficiency, and minimize energy consumption, providing scientific decision support for optimal site selection. This approach expands industry applications and promotes intelligent logistics system development while significantly reducing CE, advancing sustainable growth, and environmental protection in the logistics sector.

2 Methods and materials

2.1 Design of CCL-DCL model based on K-means algorithm

The core goal of CCL-DCL is to achieve multi-objective collaborative optimization of construction cost, transportation efficiency, service quality, and environmental sustainability [16]. However, CCL's DCL faces challenges such as difficulty in accurately predicting demand fluctuations, high construction and operational costs, insufficient data, and limited technical support. The K-means has become a classic tool in clustering analysis since its simplicity, efficiency, and wide applicability. Therefore, this study uses K-means to perform DCL on CCL. The Euclidean distance calculation in K-means is given by equation (1).

$$d(X,C_i)=\sqrt{\sum_{j=1}^{m}\left(X_j-C_{ij}\right)^2} \tag{1}$$

In equation (1), C_{ij} is the j-th attribute of C_i. C_i is the i-th CC point. X_j is the j-th attribute of X. X is a data entity. m is a measure of data objects. The evaluation of clustering effect often uses contour coefficient as a standard, as shown in formula (2).

$$S^{(i)}=\frac{b^{(i)}-a^{(i)}}{\max\left\{b^{(i)},a^{(i)}\right\}} \tag{2}$$

In equation (2), $a^{(i)}$ and $b^{(i)}$ are the average distances between the remaining sample points and all other sample points within the adjacent cluster. The contour coefficient when $S^{(i)}$ represents i. The workflow of K-means is displayed in Figure 1.

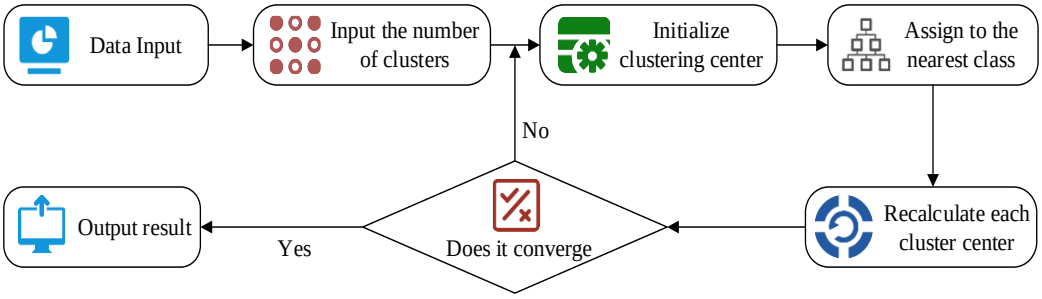


Figure 1: K-means algorithm workflow diagram

In Figure 1, K-means first measures the distance between all samples and the Cluster Center (CC) based on Euclidean distance, and divides each sample into clusters corresponding to the centroid of the nearest neighbor. Subsequently, it recalculates the centroid coordinates based on the current cluster members and defines it as the geometric mean of the samples within the cluster. The system iteratively performs sample allocation and centroid update operations until the pre-centroid coordinates are stable or the maximum Number of Iterations (NoI) is reached, thereby achieving dynamic optimization-based clustering partitioning. Finally, K centroids and the data allocation results for each class are output [16, 17]. The fixed cost calculation consisting of land leasing and personnel salaries is shown in equation (3).

$$F_1=\sum_{j=1}^{m}R_jZ_j \tag{3}$$

In equation (3), R_j is the unit price of land leasing or the salary of personnel per unit area. Z_j is the decision variable. The operating cost calculation of CCL DC is shown in equation (4).

$$F_2=\sum_{j=1}^{m}R_j^{(1)}\left(\sum_{i=1}^{l}X_{ij}B_{ij}Z_j\right)^{P'} \tag{4}$$

In equation (4), $R_j^{(1)}$ is the daily operating cost of the logistics hub. X_{ij} is the transportation volume of goods from the supplier to the DC. B_{ij} is the unit transportation cost coefficient between the supplier and the DC. p is the scale adjustment parameter. The transportation cost of the entire supply chain process is calculated. The total logistics cost covering the supply end to the cold chain transit node and then distributed to the end consumption node is shown in equation (5).

$$F_3=\sum_{i=1}^{l}\sum_{j=1}^{m}C_1X_{ij}d_{ij}B_{ij}Z_j+\sum_{j=1}^{m}\sum_{k=1}^{n}C_1X_{jk}d_{jk}Z_jB_{jk} \tag{5}$$

In equation (5), C_1 is the unit distance transportation cost. B_{jk} is the path complexity coefficient from the DC to the demand point. d_{ij} is the length of the path from the supplier to the transit base. X_{jk} is the transportation volume of materials from the supply end to the logistics node. d_{jk} is the average path length from the transit base to the consumer terminal. The overall spending of goods damage is given by equation (6).

$$F_4=C_2\sum_{i=1}^{l}\sum_{j=1}^{m}X_{ij}\left(1-e^{\frac{-\theta d_{ij}B_{ij}Z_j}{v}}\right)+C_2\sum_{j=1}^{m}\sum_{k=1}^{n}X_{jk}\left(1-e^{\frac{-\theta d_{jk}B_{jk}Z_j}{v}}\right) \tag{6}$$

In equation (6), C_2 is the unit cost of goods damage. θ is the coefficient of cargo damage rate. v is the transportation speed. To ensure constant temperature in the carriage, a significant amount of refrigeration costs will be incurred during transportation, as shown in equation (7).

$$F_5 = \sum_{i=1}^{l}\sum_{j=1}^{m} C_2 X_{ij} d_{ij} Z_j + \sum_{j=1}^{m}\sum_{k=1}^{n} C_2 X_{jk} d_{jk} Z_j B_{jk} \quad (7)$$

CCL refers to a logistics network that maintains a low temperature throughout the entire process based on the characteristics of the transported goods, to maintain the quality of the goods from production to delivery to consumers. The CCL delivery process is shown in Figure 2.

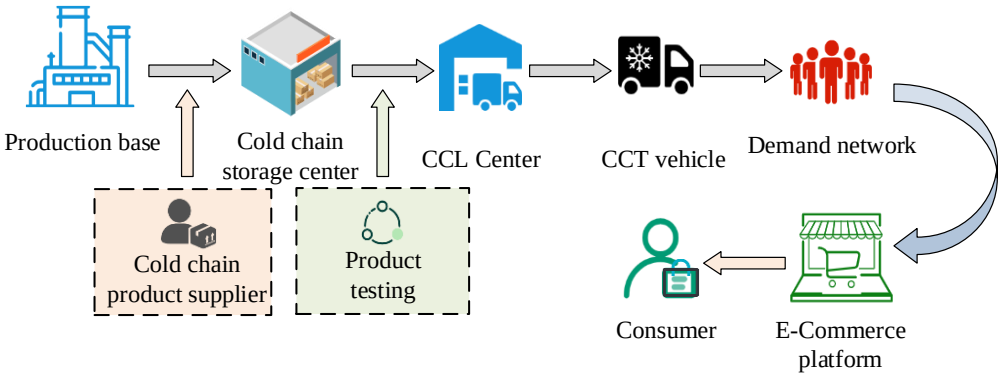


Figure 2: CCL distribution process

In Figure 2, the operation process of CCL starts from product information display and production, and involves product production and primary processing at the production base. After pre-cooling, processing, and classification, the product enters the cold chain storage center for refrigeration and preservation. Subsequently, quality inspection agencies conduct quality checks on secondary products to ensure compliance with standards. Third-party logistics companies transport products from storage centers to CCL DCs. The DC delivers products to demand outlets through cold chain transport vehicles. Finally, consumers purchase cold chain items through e-commerce platforms. Various kinds of CCL products have different characteristics and require different temperatures to be kept during CCT. As a result, to avoid losses caused by unbefitting temperature, standardized and unified operations can be carried out. Classifying the CCL products and suitable temperature is exhibited in Figure 3.

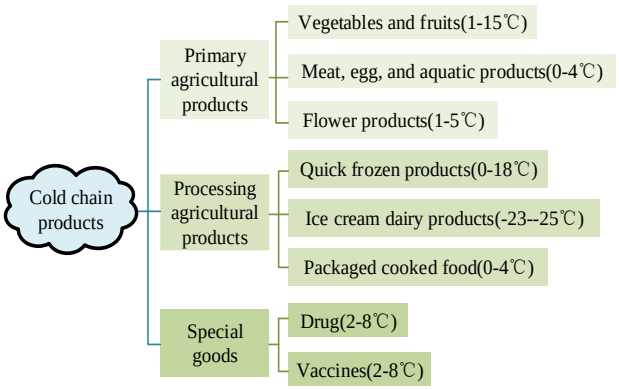


Figure 3: Classification of common cold chain items and their suitable temperatures

Figure 3 shows the ideal storage temperature range for certain specific products. This provides an important reference for ensuring the best quality of the product. However, in addition to paying attention to these fixed suitable temperatures, it is also necessary to closely monitor temperature control at various phases of transportation. Changes in environmental conditions during the handover of items may cause fluctuations in the surrounding temperature. Therefore, to ensure that the quality of the product is not affected, the temperature rise limit during transportation must be strictly controlled below -15 ℃. In addition, before loading the transport vehicle, it is necessary to ensure that the carriage has been pre-cooled to a temperature of -10 ℃ or lower to maintain a suitable temperature environment for the product throughout the entire transportation process.

2.2 Optimization of site selection model integrating K-means clustering and GA

K-means has efficient data processing capabilities and good adaptability for CCL-DCL, but is very sensitive to the selection of initial centroids and performs poorly in handling dynamic data or scenarios that require real-time adjustments [18]. The principle of genetic algorithm is simple, easy to implement and combine with other algorithms, with strong scalability and global search ability, which can effectively solve various types of problems. Therefore, this study integrates K-means with GA. The fast convergence ability of K-means can accelerate the search process of GA, while the global search ability of GA can compensate for the local optima of K-means. In addition, the robustness and flexibility of GA can enhance the performance of K-means in complex

data distributions. The GA operation process is shown in Figure 4.

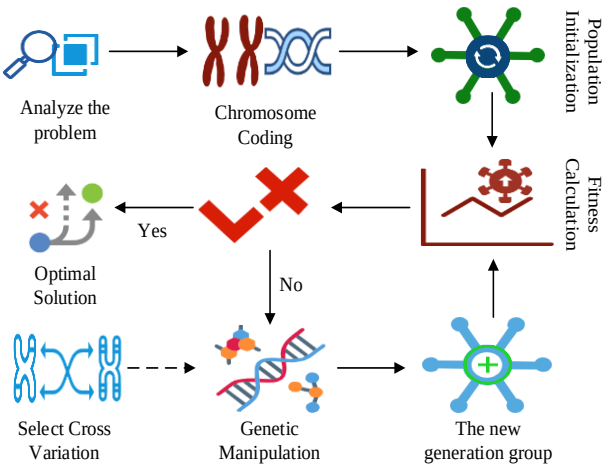


Figure 4: GS operation process

In the GA operation of Figure 4, the first step is to construct an algorithm parameter system. Based on decision vector features, the gene encoding dimension is determined, and chromosome sequence construction and population initialization configuration are performed. The next step is to activate the fitness evaluation mechanism, by establishing a nonlinear mapping model between the objective function and fitness values, to complete the quantitative evaluation of individual fitness. In the evolutionary stage, an optimized population is generated through selection operators, crossover recombination, and mutation mechanisms, and fitness evaluation and genetic operations are iteratively performed until the termination criteria of evolution are met. After the algorithm terminates, it is necessary to perform reverse decoding on the optimal gene sequence and output the final optimization solution [19]. This study takes the total cost of CCL as the determining factor for site selection, and constructs the objective function as shown in equation (8).

$$\min F = F_1 + F_2 + F_3 + F_4 + F_5 \tag{8}$$

In equation (8), F is the total cost of CCL. The formula for setting transportation convenience restrictions for site selection is shown in equation (9).

$$\mathrm{A}\left(\hat{C}_j\right)..A_{\min}, \forall j \tag{9}$$

In equation (9), $\mathrm{A}\left(\hat{C}_j\right)$ denotes the analysis of the mean traffic convenience characteristics of service nodes within the radiation area $\hat{C}_j$ after the cold chain facility is deployed in the backup location. $A_{\min}$ is the minimum critical value of transportation convenience required for the effective operation of the cold chain distribution system. The constraint on CCT time is shown in equation (10).

$$\hat{t}_{j,l},, \hat{t}_{\max}, \forall j, l \tag{10}$$

In equation (10), $\hat{t}_{\max}$ means the maximum transportation time. The constraint on the spoilage of transported products is shown in equation (11).

$$0,, \aleph(\hat{t}_{j,l} + \hat{t}_{i,j}) < \tilde{T}_{\max} \tag{11}$$

In equation (11), $\hat{t}_{i,j}$ is the transportation time between the production base and the CCL center. $\tilde{T}_{\max}$ is the maximum allowable transportation time limit for cold chain products to maintain freshness. The constraint on the supply-demand matching relationship of CCL is shown in equation (12).

$$\sum_{j=0}^{n'} h_{i,j} < C_j', \sum_{l \in L'} \hat{h}_j \ldots \hat{Q} \tag{12}$$

In equation (12), L' is the set of CCL centers that can serve demand point j. Chromosome encoding is the primary step in applying GA. Encoding is the process of mapping the solution state space. Figure 5 shows the chromosome coding diagram.

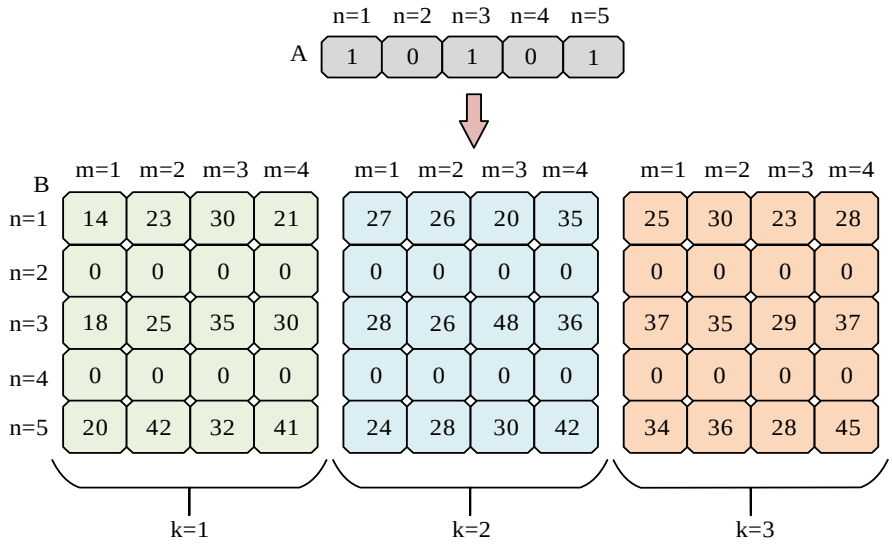


Figure 5: Chromosome coding diagram

In Figure 5, the chromosome encoding diagram illustrates how to select different n values through binary encoding of chromosome X, and displays the detailed data corresponding to these n values in matrix Y. Among them, chromosome X is a binary sequence with a length of 5, indicating whether to select certain values from n=1 to n=5. Matrix Y is a 2D table with rows representing different values of n. Columns represent different values of m, and the numbers in each cell represent a certain indicator or data. The columns of matrix Y are grouped into k=1, k=2, and k=3, with each group containing multiple values of m. In GA, the fitness function is a standard utilized to determine the chromosomes quality of the GA population. The corresponding fitness function selected is given by equation (13).

$$fit(i) = \frac{1}{F(i) + \mu} \tag{13}$$

In equation (13), μ is a positive real number within the interval $(0,1)$. $F(i)$ denotes the total cost function value of genetic unit i. $fit(i)$ is the fitness of i. The μ determined by the research through many experiments is set to 0.1. When μ varies between 0.05 and 0.2, the performance trend of the algorithm remains consistent. When μ is 0.1, the best balance can be achieved between cost minimization and algorithm stability. The strategy selection section constructs a probability function for elimination based on individual fitness, as shown in equation (14).

$$P_x = \frac{1 - f_x}{\sum \tilde{n}_{x=1} f_x} \tag{14}$$

In equation (14), P_x is the probability of being selected. $\tilde{n}$ is the population size. f_x is individual fitness. The genetic recombination mechanism achieves offspring generation by exchanging homologous segments of parental gene sequences. This process involves the localization of random breakpoints and the exchange of genetic material following Mendel's laws. To improve the adaptability of algorithms in processing and analysis, it is necessary to automatically adjust and optimize their processing methods, order, parameter settings, boundaries, and constraints based on the statistical distribution and structural characteristics of data [20]. The study introduces a two-point crossover algorithm. Two point crossover refers to randomly selecting two crossover points between two parental individuals, and then exchanging gene fragments between these two crossover points to produce two new offspring individuals. The expression is shown in equation (15).

$$\begin{aligned} C1_k &= \begin{cases} P1_k & k \le u \\ P2_k & \mathrm{u} < k \le s \\ P1_k & k > s \end{cases} \\ C2_k &= \begin{cases} P2_k & k \le u \\ P1_k & \mathrm{u} < k \le s \\ P2_k & k > s \end{cases} \end{aligned} \tag{15}$$

In equation (15), $C1$ and $C2$ represent the offspring chromosomes. $P1$ and $P2$ represent the parent chromosomes. u and s represent the crossover points. k represents the gene position index, where $k = 1, 2, 3 \ldots\ldots n$. The genetic crossover operation diagram is shown in Figure 6.

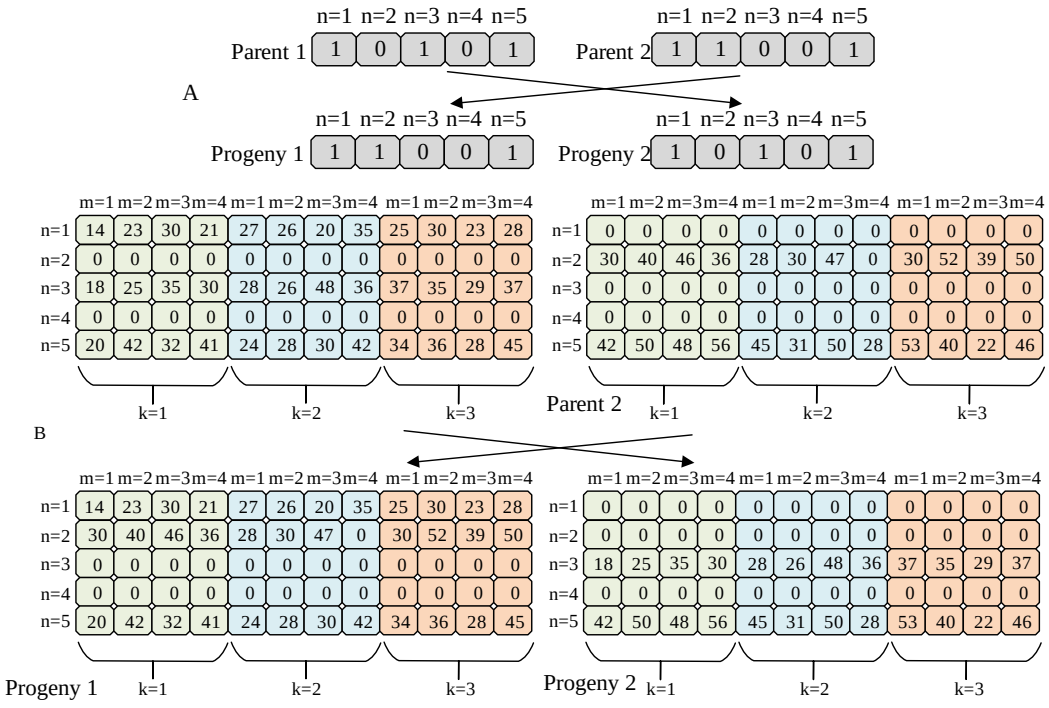


Figure 6: Schematic diagram of genetic crossover operation

In Figure 6, taking the parent individuals P_1 and P_2 as an example, the specific implementation of their crossover process is as follows: Firstly, an adaptive crossover probability model is constructed, which dynamically selects the second and third gene loci as crossover boundaries based on the current iteration times and population fitness variance. The next step is to perform gene fragment exchange - P_1 retains the first 2 genes and inherits the last 3 genes of P_2 to generate offspring C_1, while P_2 retains the first 2 genes and inherits the last 3 genes of P_1 to generate offspring C_2. In response to the possible phenomenon of gene homogenization in C_2, a gene position reversal protection mechanism will be triggered, and the final output will be a corrected offspring C_2. In summary, a crossover process involves two parent individuals and their offspring individuals. The specific crossover details are generated by selecting gene fragments from the parent individuals and combining them to create new offspring individuals. The fusion process of K-means and GA is shown in Figure 7.

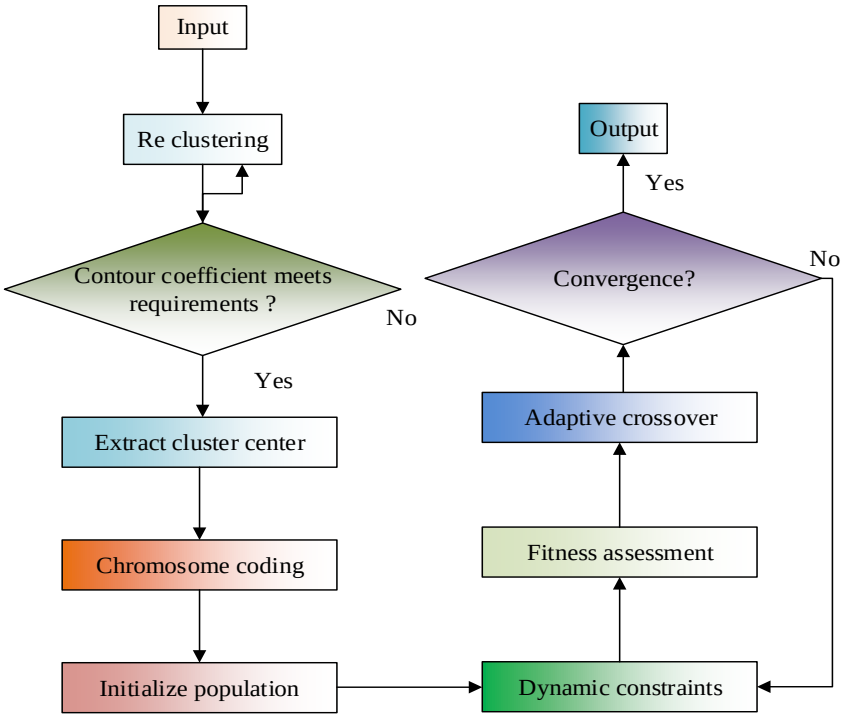


Figure 7: The fusion process of K-means and GA

As shown in Figure 7, the integration process of K-means and GA works as follows: First, the coordinates of demand nodes are input. Then, K-means clustering is performed and the k value is continuously adjusted until the contour coefficients met the requirements, at which point the cluster centers are extracted. Next, binary encoding is applied to the cluster center coordinates to initialize the GA population. The next step is to perform an iterative process, evaluate fitness, and check if clustering constraints are met. If satisfied, the total cost and dynamic constraints need to be calculated. Subsequently, adaptive crossover and mutation operations will be performed to determine convergence. If it does not converge, it will return to fitness evaluation for further iteration; If converged, it will output the optimal position and path.

3 Results

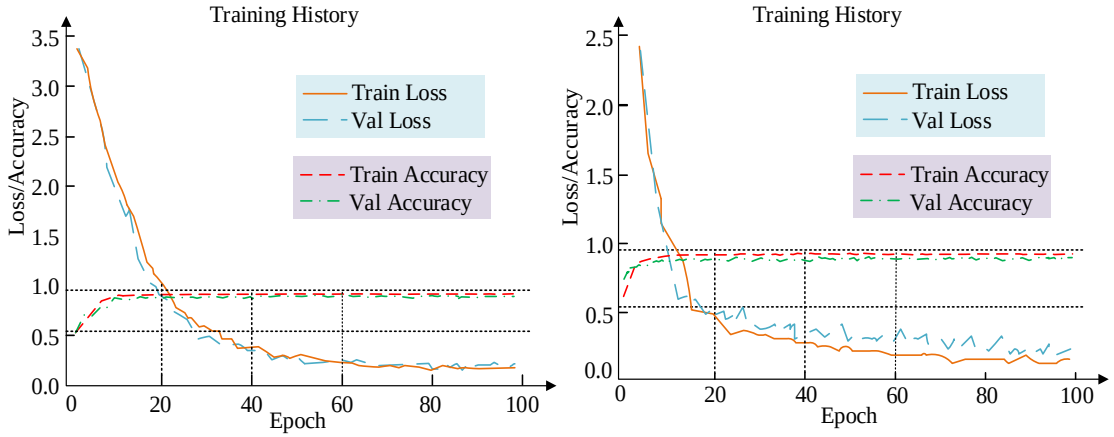
3.1 Performance testing of CCL delivery based on a new fusion algorithm

The experimental environment includes a CPU (Intel Core i9-11900K), GPU (NVIDIA RTX 3090), and 64GB DDR4 3200MHz memory. The operating system is Ubuntu 20.04 LTS, and the development framework is based on Python 3.8. The core logic of the algorithm is implemented through NumPy and SciPy. Path optimization parallel computing is completed by PyTorch. The study utilizes the GORD dataset and VRPI instance datasets to conduct detailed analysis of sample data, thereby validating the model's effectiveness and practicality. 500 sample data points are selected from the GORD dataset and 300 sample data points are selected from the VRPI dataset for the experiment. These samples encompass key characteristics including urban/suburban traffic conditions, node distribution, and delivery demands, effectively simulating real-world operational scenarios for CCL DCs. Specifically, GORD provides detailed information about actual urban road networks, such as road lengths, travel times, and congestion indices. VRPI contains critical parameters like node coordinates, delivery requirements, and vehicle load constraints. By comprehensively analyzing and processing these datasets, the performance of the model can be evaluated under various operating conditions. The advantages of data sets are shown in Table 2.

Table 2: Data set comparison analysis

Dataset Name	Geographic Coverage	Traffic Dynamic Characteristics	Data Authenticity	Applicability	Advantages
GORD	Urban and suburban areas	Includes road lengths, travel times, congestion coefficients, etc.	Real - world urban road network data	Suitable for urban and suburban distribution scenarios	Provide detailed real urban road network information, can accurately simulate the actual traffic conditions
VRPI	Diverse scenarios	Includes node coordinates, delivery demands, vehicle capacity constraints, etc.	Classic problem instances	Suitable for classic VRPI research	Offers rich VRPI instances with key parameters for easy model validation and comparison
Other data sets	Geographical coverage varies	The characteristics are not equal	Authenticity varies	The applicability is not equal	Specific information needed for CCL may not be available

In Table 2, the dataset used is more suitable for CCL distribution path planning than other datasets. The average runtime of the research algorithm is 120 seconds on the GORD dataset and 150 seconds on the VRPI dataset. To ensure reproducibility, the GA parameters are set as follows: population size (100), crossover rate (0.8), mutation rate (0.1), roulette wheel selection method, and maximum iteration count (500). For K-means clustering, the number of clusters is set to 5, initial centroids are randomly determined, and the maximum iteration count is 100. When the change in the centroid position is less than 0.001, it is considered that convergence has been achieved. This study first conducts a comparative analysis of the loss function and accuracy of the fusion algorithm for different delivery distances, as shown in Figure 8.



(a) Long distance delivery (b) Close range delivery

Figure 8: Comparison of accuracy and loss rate at different distances

According to Figure 8 (a), as the NoI increases, the loss value continuously decreases and the accuracy continuously increases. At the 100th iteration, the loss value reaches 0.06 and the accuracy reaches 99.89%. In Figure 8 (b), as the NoI increases, the algorithm performance improves, and the initial loss value rapidly decreases, ultimately reaching a stable state. The accuracy improves relatively quickly in the early stage, and gradually stabilizes later on. At the 100th iteration, the loss value reaches 0.08 and the accuracy reaches 99.67%. In summary, the research algorithm has a high calculation accuracy and low error loss for both long-distance and close-range delivery. This study tests and compares the convergence speed and energy consumption of the fusion algorithm, the Handling Multiple Objectives with PSO (MOPSO) algorithm, and the Multi-Objective Optimization algorithm (NSGA-III), as shown in Figure 9.

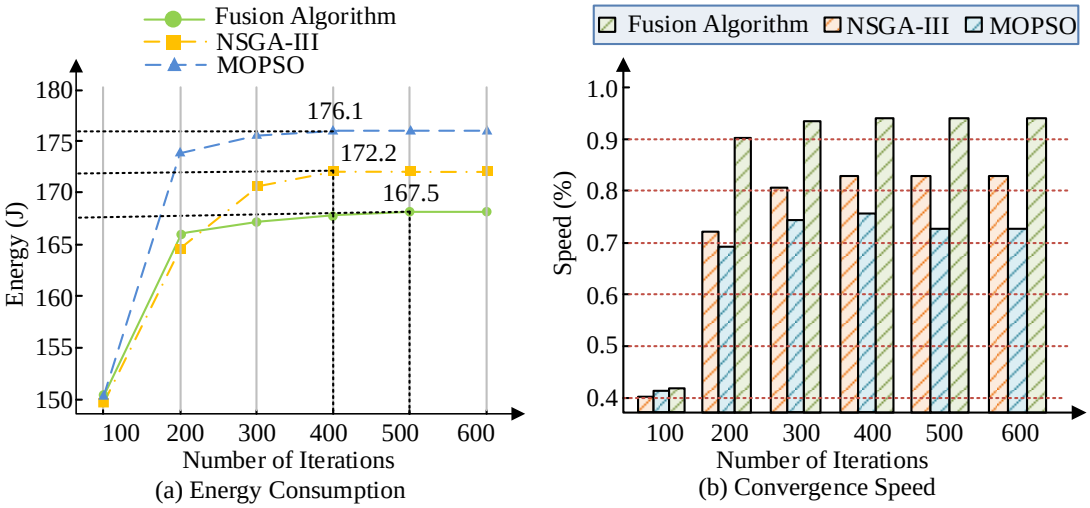


Figure 9: Convergence speed and energy consumption of different methods

In Figure 9 (a), MOPSO achieves a maximum energy consumption of 176.1 J after 400 iterations. NSGA-III has the highest energy consumption at 400 iterations, reaching 172.2 J. The fusion algorithm reaches its highest energy consumption of 167.5 J after 500 iterations, which is the lowest compared to the other two methods. In Figure 9 (b), the convergence speed of the fusion algorithm is the fastest, while the MOPSO method is the slowest. At 400 iterations, the fusion algorithm, NSGA-III, and MOPSO reach their highest points of 94.8%, 83.2%, and 76.3%, and then stabilize. This indicates that using fusion algorithms results in better energy consumption and convergence speed. This study tests the processing performance of fusion algorithms, NSGA-III, and MOPSO methods in GORD and VRPI, as listed in Table 3.

Table 3: The index test results of different multi-objective algorithms

Data set	Method	Average processing time/s	Resource occupancy rate/%	Average energy consumption/J
GORD	Fusion Algorithm	6.32	75.21	149.67
	NSGA-III	7.68	78.95	155.82
	MOPSO	8.95	82.36	162.47
VRPI	Fusion Algorithm	8.42	78.46	160.72
	NSGA-III	9.16	81.47	167.53
	MOPSO	10.43	84.62	174.29

In Table 3, the average task processing time of the fusion algorithm is controlled between 6.32 seconds and 8.42 seconds, which shows significant optimization effects and higher processing efficiency compared to other methods. Meanwhile, the fusion algorithm has the lowest computational resource utilization rate, ranging from 75.21% to 78.46%, indicating a higher efficiency in resource utilization. The average energy consumption value of the fusion algorithm is 149.67 J to 160.72 J, significantly lower than NSGA-III and MOPSO, and the energy optimization effect is significant. In summary, the fusion algorithm improves processing efficiency and reduces energy consumption and resource utilization through effective path planning. Its superior performance verifies its progressiveness and practical application value in CCL distribution path optimization. The study uses the t-test to analyze the treatment time, while the energy consumption data are evaluated by the Wilcoxon test, as shown in Table 4.

Table 4: Results of t-test and Wilcoxon test

Data set	Method	Average processing time/s	T test	Average energy consumption/J	Wilcoxon test
GORD	Fusion Algorithm	6.32 ±0.41	-	149.67±8.2	-
	NSGA-III	7.68 ±0.53	t=8.71, p<0.001, Δ=1.36 [1.02–1.70]	155.82±9.6	Z=3.89, p<0.002, Δ=6.15 [3.21–9.09]
	MOPSO	8.95 ±0.62	t=12.94, p<0.001, Δ=2.63 [2.18–3.08]	162.47±10.3	Z=5.32, p<0.001, Δ=12.80 [9.45–16.15]
VRPI	Fusion Algorithm	8.42 ±0.49	-	160.72±9.1	-
	NSGA-III	9.16 ±0.58	t=7.23, p<0.001, Δ=0.74 [0.51–0.97]	167.53±10.2	Z=3.12, p<0.008, Δ=6.81 [3.95–9.67]
	MOPSO	10.43 ±0.71	t=11.87, p<0.001, Δ=2.01 [1.62–2.40]	174.29±11.5	Z=4.95, p<0.001, Δ=13.57 [10.20–16.94]

In Table 4, the paired t-test results indicate that compared with NSGA-III, the average processing time of the research algorithm is reduced by 1.36 s, 2.63 s on GORD compared to MOPSO, and 0.74 s and 2.01 s on VRPI compared to MOPSO. Wilcoxon tests further reveal that the energy consumption of the proposed algorithm decreases by 6.15 J and 12.80 J on the GORD dataset, as well as 6.81 J and 13.57 J on the VRPI dataset. These findings confirm that the proposed fusion algorithm demonstrates statistically significant advantages in both computational speed and energy efficiency.

3.2 Application testing of logistics distribution based on new fusion algorithm

To validate the practical application effect of the new operation path mechanism, this study takes a certain fresh logistics enterprise as the research object. The cold chain delivery process is concentrated in the early morning hours, effectively avoiding the interference of traffic congestion on the experimental results. The experiment uses four trucks equipped with refrigeration units and cold storage insulation boxes to construct analysis samples based on actual distribution data from the enterprise. The comparison of the total delivery mileage between the fusion algorithm, NSGA-III, and MOPSO methods in urban and suburban areas is shown in Figure 10.

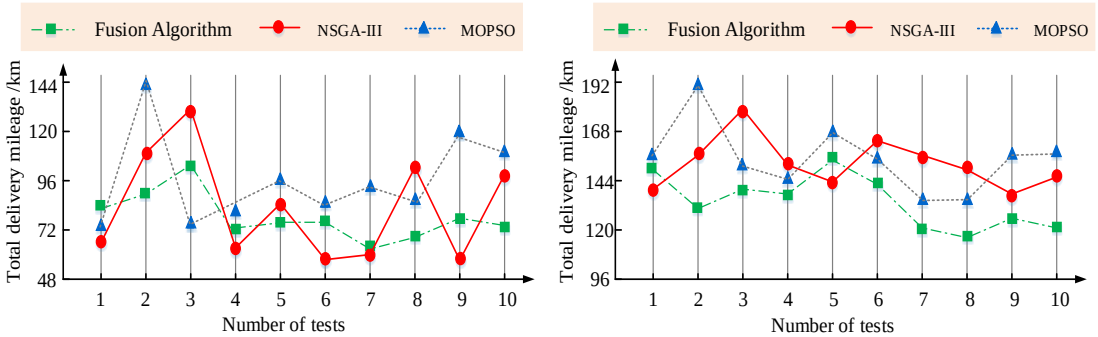


Figure 10: Comparison of distribution process between urban and suburban areas

Figures 10 (a) and (b) show a comparison of 10 delivery distances for urban and suburban customers using different methods. In Figure 10 (a), in the delivery of urban customers, the fusion algorithm shows shorter total delivery mileage in most testing rounds, and its results have higher stability and uniformity in 10 tests. In contrast, NSGA-III and MOPSO show higher peak delivery total mileage in some rounds, especially in the second and third tests, with delivery mileage exceeding 140 km and 130 km, respectively. The fusion algorithm always maintains a low level, with a maximum value of about 100 km. In Figure 10 (b), the fusion algorithm also demonstrates superior performance in the delivery of suburban customers, with a maximum value of about 150 km. Compared with the NSGA-III method and MOPSO method, which have repeatedly exceeded 150 km, the fusion algorithm has significant advantages. In summary, the fusion algorithm has more advantages in delivery in urban and suburban areas, with a shorter total mileage. The distribution path planning results of the fusion algorithm and NSGA-III are shown in Figure 11.

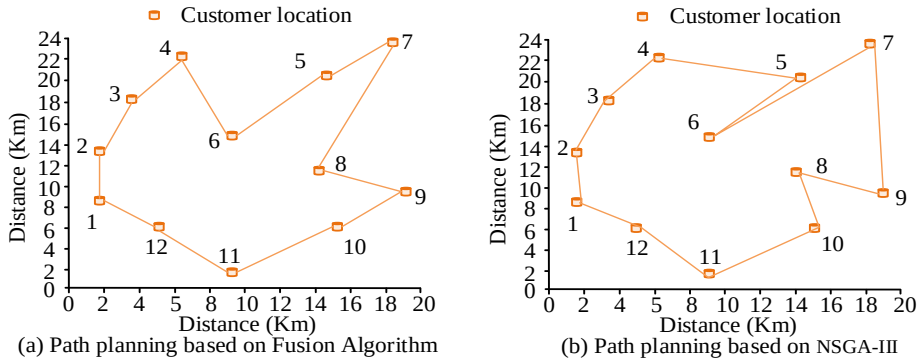


Figure 11: Delivery path planning in two scenarios

In Figure 11 (a), each delivery path planned by the fusion algorithm has no overlapping parts. The path planned by NSGA-III in Figure 11 (b) has overlapping parts. Figure 11 intuitively shows that the delivery route planned by the fusion algorithm has a reduced route length compared to NSGA-III, and avoids the cost loss led by

path duplication. This study compares the fusion algorithm, NSGA-III, and MOPSO methods using average delivery cost and delivery efficiency as indicators, as shown in Figure 12.

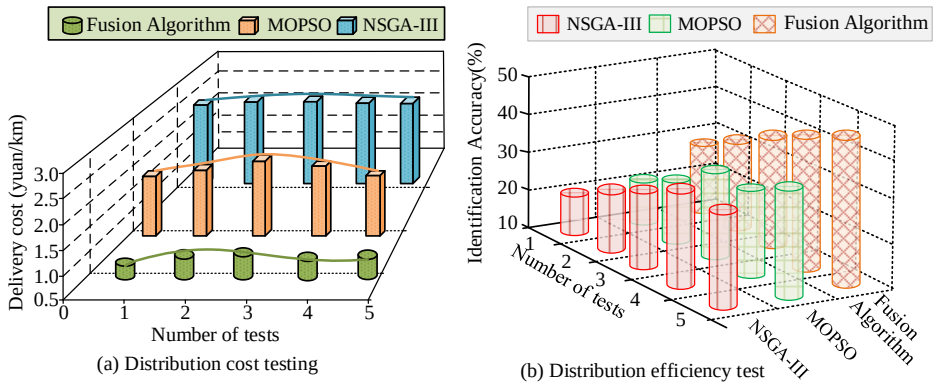


Figure 12: Distribution cost and efficiency test results

Figure 12 (a) shows the test results of the distribution costs of different methods. Figure 12(b) shows the average driving speed of distribution vehicles of different methods, which reflects the distance that vehicles can complete in a unit time. The higher the average speed, the longer the distance that vehicles can cover in the same time, indicating higher distribution efficiency. In Figure 12 (a), the delivery cost of the fusion algorithm remains at the lowest level, with an average cost of about 1.2 yuan per kilometer. The average cost per kilometer for MOPSO is about 1.8 yuan, while for NSGA-III it is about 2.3 yuan. This indicates that fusion algorithms have significant advantages in reducing delivery costs. In Figure 12(b), the NSGA-III method achieves an average speed of approximately 30 km/h, while the MOPSO method reaches about 28 km/h. Notably, the hybrid algorithm demonstrates the most efficient performance with an average speed of around 40 km/h. In summary, fusion algorithms can not only effectively reduce delivery costs, but also achieve significant improvements in delivery efficiency, especially demonstrating excellent adaptability in complex delivery scenarios.

4 Discussion

The proposed fusion algorithm demonstrated significant advantages over NSGA-III and MOPSO in multiple aspects. The results of Table 1 and Figures 9-12 indicated that the algorithm outperformed traditional algorithms in key indicators such as average processing time, resource utilization, and energy consumption. Specifically, on the GORD dataset, it reduced average processing time by up to 1.36 seconds compared to NSGA-III and by up to 2.63 seconds compared to MOPSO. Similarly, on the VRPI dataset, it achieved reductions of up to 0.74 seconds and 2.01 seconds. The notable decrease in resource utilization indicated more efficient resource allocation. In terms of energy consumption, compared with NSGA-III, the fusion algorithm reduced energy consumption by 6.15 j and 3.57 j on the GORD and VRPI datasets, and by 12.8 j and 13.57 j compared to MOPSO. The fusion algorithm also showed clear advantages in delivery cost and efficiency. Figures 10 and 12 demonstrated lower delivery costs and higher efficiency in both urban and suburban deliveries. The average delivery cost per kilometer was significantly lower than NSGA-III and MOPSO, with delivery efficiency notably improved. These results highlighted the practical benefits of the fusion algorithm in real-world logistics operations. Statistical analysis of the study's outcomes confirmed the significance of these performance differences. The fusion algorithm consistently outperformed NSGA-III and MOPSO across multiple iterations and datasets. The reductions in processing time, resource utilization, and energy consumption were not coincidental but demonstrated a more efficient optimization approach. Statistical analysis confirmed that the fusion algorithm achieved significant performance improvements compared to existing methods such as NSGA-III and MOPSO, highlighting its novelty and practical value. When developing a CCL DC location model through practical deployment research, enterprises must consider multiple factors to ensure successful implementation and achieve expected benefits. First, compliance with laws and regulations is essential, requiring data encryption and access control measures to safeguard customer and business data security and privacy. Second, to implement optimized model results, seamless integration with existing warehousing, transportation, and vehicle scheduling systems is necessary. This may involve system interface development, data format conversion, and business process coordination. Finally, enterprises should conduct comprehensive cost-benefit evaluations to balance implementation costs against anticipated benefits, ensuring that optimized logistics operations do not compromise financial stability. Future research can further explore integration with other advanced algorithms or incorporate real-time data to enhance the model's adaptability and robustness in dynamic environments.

5 Conclusion

The prosperity of smart cities has made logistics distribution an important component of urban functions, and its efficiency and sustainability have become a hot

research topic. The traditional logistics model has many shortcomings in terms of efficiency and environmental protection. Therefore, the paper proposed a CCL-DCL model built on the fusion of K-means clustering and GA, aiming to enhance CCL delivery efficiency and cut down delivery costs. The model evaluated the clustering effect through contour coefficients and determined the optimal DCL. Subsequently, it combined the global search capability of GA to construct a multi-objective optimization model that comprehensively considers constraints such as real-time traffic status, cargo preservation time, and vehicle load limitations. In the experiment, the fusion algorithm performed well in optimizing the CCL delivery path. The average task processing time was controlled between 6.32 seconds and 8.42 seconds. The resource utilization rate was the lowest, only 75.21% to 78.46%, while the average energy consumption was 149.67 J to 160.72 J. The delivery cost per kilometer of the fusion algorithm was about 1.2 yuan, and the delivery efficiency per kilometer was about 40 km/h. In summary, the path planning model that integrated K-means and GA exhibited excellent adaptability in complex delivery scenarios, avoiding cost losses caused by route duplication. However, the study has inherent limitations. In real-world operations, demand fluctuations caused by promotional activities and seasonal changes often lead to suboptimal route planning, thereby reducing delivery efficiency. In addition, the real-time response capability of the model is insufficient, and there is a lack of effective data integration mechanism, which makes it difficult to adjust in a timely manner during traffic congestion emergencies, limiting its practicality in dynamic environments and potentially causing delays. In addition, the model has poor adaptability to random events such as weather changes or vehicle failures. This means that when the original plan becomes unfeasible, it cannot quickly generate new routes, thereby reducing the reliability of delivery. Future improvements will include implementing real-time data processing mechanisms, developing dynamic path adjustment algorithms, and building more flexible model frameworks to enhance the robustness and adaptability of the system in complex and constantly changing scenarios.

Nomenclature

Symbol	Description
$d(X,C_i)$	the Euclidean distance calculation in K-means
C_{ij}	the j-th attribute of C_i
C_i	the i-th Cluster Center (CC) point
X_j	the j-th attribute of X
X	a data entity
m	a measure of data objects
$s(i)$	silhouette coefficient of sample i
$a^{(i)}$	average distance from sample i to other samples in the same cluster
$b^{(i)}$	minimum average distance from sample i to samples in neighboring clusters
F_1	the fixed cost calculation consisting of land leasing and personnel salaries
R_j	the unit price of land leasing or the salary of personnel per unit area
Z_j	the decision variable
F_2	the operating cost calculation of CCL DC
$R_j^{(1)}$	the daily operating cost of the logistics hub
X_{ij}	the transportation volume of goods from the supplier to the DC
B_{ij}	the unit transportation cost coefficient between the supplier and the DC.
p	the scale adjustment parameter
F_3	the total logistics cost covering the supply end to the cold chain transit node and then distributed to the end consumption node
C_1	the unit distance transportation cost
B_{jk}	the path complexity coefficient from the DC to the demand point
d_{ij}	the length of the path from the supplier to the transit base
X_{jk}	the transportation volume of materials from the supply end to the logistics node
d_{jk}	the average path length from the transit base to the consumer terminal
F_4	the overall spending of goods damage
C_2	the unit cost of goods damage
θ	the coefficient of cargo damage rate.
v	the transportation speed
F_5	refrigeration cost
$\min F$	objective function
F	the total cost of CCL
$\mathrm{A}\left(\hat{C}_j\right)$	the analysis of the mean traffic convenience characteristics of service nodes within the radiation area
$\hat{C}_j$	the cold chain facility is deployed in the backup location.
$A_{\min}$	the minimum critical value of transportation convenience

	required for the effective operation of the cold chain distribution system
$\hat{t}_{j,l}$	the constraint on CCT time
$\hat{t}_{\max}$	the maximum transportation time
$0,, \aleph(\hat{t}_{j,l} + \hat{t}_{i,j}) < \tilde{T}_{\max}$	the constraint on the spoilage of transported products
$\hat{t}_{i,j}$	the transportation time between the production base and the CCL center
$\tilde{T}_{\max}$	the maximum allowable transportation time limit for cold chain products to maintain freshness
$\sum_{j=0}^{n'} h_{i,j} < C_j', \sum_{l \in L'} \hat{h}_j \ldots \hat{Q}$	the constraint on the supply-demand matching relationship of CCL
L'	the set of CCL centers that can serve demand point j.
$fit(i)$	the fitness of i
μ	a positive real number within the interval $(0,1)$.
$F(i)$	the total cost function value of genetic unit i.
P_x	the probability of being selected
$\tilde{n}$	the population size.
f_x	Individual fitness
$C1$	the offspring chromosomes
$C2$	the offspring chromosomes
$P1$	the parent chromosomes
$P2$	the parent chromosomes
u	the crossover points
s	the crossover points
k	the gene position index

References

[1] Chen Y, Lan H, Wang C, Jia A. An integrated distribution scheduling and route planning of food cold chain with demand surge. Complex & Intelligent Systems, 2023, 9(1): 475-491. https://doi.org/10.1007/s40747-022-00811-9

[2] Turan C, Ozturkoglu Y. A conceptual framework model for an effective cold food chain management in sustainability environment. Journal of Modelling in Management, 2022, 17(4): 1262-1279. https://doi.org/10.1108/JM2-09-2020-0239

[3] Sun J, Karia N B. Innovative approaches to assessing cold chain logistics in B2C E-Commerce environments. Journal of the Knowledge Economy, 2024, 15(3): 14670-14699. https://doi.org/10.1007/s13132-023-01669-z

[4] Tang Y M, Chau K Y, Kuo W T, Liu X X. IoT-based information system on cold-chain logistics service quality (ICCLSQ) management in logistics 4.0. Information Systems Frontiers, 2024, 26(2): 689-708. https://doi.org/10.1007/s10796-023-10393-7

[5] Andoh E A, Yu H. A two-stage decision-support approach for improving sustainable last-mile cold chain logistics operations of COVID-19 vaccines. Annals of operations research, 2023, 328(1): 75-105. https://doi.org/10.1007/s10479-022-04906-x

[6] Bai Q, Yin X, Lim M K, Dong C. Low-carbon VRP for cold chain logist considering real-time traffic conditions in the road network. Industrial management & data systems, 2022, 122(2): 521-543. https://doi.org/10.1108/IMDS-06-2020-0345

[7] Pu X, Lu X, Han G. An improved optimization algorithm for a multi-depot vehicle routing problem considering carbon emissions. Environmental Science and Pollution Research, 2022, 29(36): 54940-54955. https://doi.org/10.1007/s11356-022-19370-0

[8] Yin N. Multiobjective optimization for vehicle routing optimization problem in low-carbon intelligent transportation. IEEE Transactions on Intelligent Transportation Systems, 2022, 24(11): 13161-13170. https://doi.org/10.1109/TITS.2022.3193679

[9] Du J, Wang X, Wu X, Zhou F, Zhou L. Multi-objective optimization for two-echelon joint delivery location routing problem considering carbon emission under online shopping. Transportation Letters, 2023, 15(8): 907-925. https://doi.org/10.1080/19427867.2022.2112857

[10] Li K, Li D, Wu D. Carbon transaction-based location-routing-inventory optimization for cold chain logistics. Alexandria Engineering Journal, 2022, 61(10): 7979-7986. https://doi.org/10.1016/j.aej.2022.01.062

[11] Ashari I F, Nugroho E D, Baraku R, Yanda I N, Liwardana R. Analysis of elbow, silhouette, Davies-Bouldin, Calinski-Harabasz, and rand-index evaluation on k-means algorithm for classifying flood-affected areas in Jakarta. Journal of Applied Informatics and Computing, 2023, 7(1): 95-103. https://doi.org/10.30871/jaic.v7i1.4947

[12] Şenol A. Impkmeans: An improved version of the k-means algorithm, by determining optimum initial centroids, based on multivariate kernel density estimation and kd-tree. Acta Polytechnica Hungarica, 2024, 21(2): 111-131. https://doi.org/10.12700/APH.21.2.2024.2.6

[13] Fu X, Sun Y, Wang H, Li H. Task scheduling of cloud computing based on hybrid particle swarm algorithm and genetic algorithm. Cluster Computing, 2023, 26(5): 2479-2488. https://doi.org/10.1007/s10586-020-03221-z

[14] Choudhury D, Acharjee T. A novel approach to fake news detection in social networks using genetic algorithm applying machine learning classifiers. Multimedia Tools and Applications, 2023, 82(6): 9029-9045. https://doi.org/10.1007/s11042-022-12788-1

[15] Abedpour K, Hosseini Shirvani M, Abedpour E. A genetic-based clustering algorithm for efficient resource allocating of IoT applications in layered fog

heterogeneous platforms. Cluster Computing, 2024, 27(2): 1313-1331. https://doi.org/10.1007/s10586-023-04005-x

[16] Layeb A. ck-means and fck-means: two deterministic initialization procedures for k-means algorithm using a modified crowding distance. Acta Informatica Pragensia, 2023, 12(2): 379-399. https://doi.org/10.18267/j.aip.223

[17] Tirkolaee E B, Torkayesh A E. A cluster-based stratified hybrid decision support model under uncertainty: sustainable healthcare landfill location selection. Applied Intelligence, 2022, 52(12): 13614-13633. https://doi.org/10.1007/s10489-022-03335-4

[18] Wang D. Application of Improved Binary K-means Algorithm in Time and Cost Optimization for Regional Logistics Distribution Center Location. Informatica, 2025, 49(6): 103-116. https://doi.org/10.31449/inf.v49i6.7215

[19] Zhu B, Wang H, Fan M. Constructing small sample datasets with game mixed sampling and improved genetic algorithm. The Journal of Supercomputing, 2024, 80(14): 20891-20922. https://doi.org/10.1007/s11227-024-06263-x

[20] Yang L, Tang Z, Liu S. Research on optimisation method for project site selection based on improved genetic algorithm. International Journal of Industrial and Systems Engineering, 2022, 40(3): 309-324. https://doi.org/10.1504/IJISE.2022.122260

A Machine Learning-Driven Decision Support System for Urban Expansion Prediction Using Multi-Temporal Remote Sensing Data

Jianguang Yu
School of Civil Engineering, Putian University, Putian 351100 China
E-mail: JianguangYuu@outlook.com

Keywords: remote sensing images, urban expansion, a prediction model, decision support system, data fusion

Received: June 9, 2025

With the acceleration of urbanization, resource consumption and environmental pressure caused by urban expansion have become increasingly prominent, highlighting the urgent need for a scientific planning decision support system. This study proposes a machine learning-driven decision support system for urban expansion prediction using multi-temporal remote sensing data, aiming to enhance the scientificity and effectiveness of urban planning. Key technologies include image enhancement, noise reduction, and feature extraction to obtain high-quality urban expansion data, with specific algorithms such as Convolutional Neural Networks (CNN) for spatial feature learning and Iterative Slow Feature Analysis (ISFA) for change detection. The analysis found that the land use efficiency in some areas has been significantly improved, and the related change range is 11.1% (representing an 11.1% increase in land use efficiency compared to the baseline). In the correlation analysis between economic development and traffic flow, the change range reached 39.4 (an index value indicating the strength of correlation, with higher values representing stronger associations). The average change rate of urban areas is 56.3%, while green space coverage changes by 73.2% and building density increases by 68%. The average annual change rate of urban areas is 56.3% over the period 2010–2020, while the effective change detection rate reaches 92%.

Povzetek: Sistem za podporo urbanističnemu načrtovanju z veččasovnimi satelitskimi podatki in ML (CNN + ISFA) napoveduje urbano širitev ter doseže 92 % zaznavanje sprememb.

1 Introduction

In urban expansion prediction and planning decision-making, multi-temporal remote sensing data with different resolutions and types can provide multi-level change information. These data can reveal the spatial scope of urban expansion and provide detailed information about the change area, such as the type of change and the semantic level change [1, 2]. Through analysing remote sensing images in different periods, the changes in new buildings, road expansion and green space reduction in urban expansion can be identified, which has significant reference value for urban planners and decision-makers. How to effectively extract this multi-level change information still faces many challenges [3, 4]. Due to the difference in external environmental conditions, there are significant differences in the radiation characteristics of remote sensing images acquired in different periods. These radiation differences will interfere with the direct comparison of images, making it challenging to accurately identify the real surface change information [5]. Although hyperspectral remote sensing images contain rich spectral information, their data dimensions are very high. Effectively extracting and utilising these high-dimensional spectral features has always been a technical difficulty [6]. To address these challenges, this study aims to achieve the following objectives: (1) Develop a machine learning-driven decision support system for urban expansion prediction using multi-temporal remote sensing data; (2) Achieve a change detection accuracy of over 90% by integrating CNN and ISFA algorithms; (3) Realize multi-source data fusion (remote sensing, socio-economic, traffic data) to enhance prediction comprehensiveness; (4) Validate the system's effectiveness through standardized datasets and evaluation metrics (Kappa, AUC, detection rate). However, it is difficult for traditional image processing technologies to fully tap into and utilize this information [7, 8]. Various advanced technical means are used to process and analyze remote-sensing images. Regarding image processing, standard methods include image enhancement, radiation correction and noise removal [9].

Image enhancement technology can improve the clarity of images and make the boundaries and features of ground objects more obvious; Radiation correction can reduce the radiation difference caused by changes in illumination and atmospheric conditions and improve the comparability between images; Noise removal technology can effectively remove random noise in images, thus improving the quality of data [10, 11]. This research introduces novel scientific contributions: (1) A hybrid framework integrating Convolutional Neural Networks (CNN) for spatial feature extraction and

Iterative Slow Feature Analysis (ISFA) for radiation difference reduction, which outperforms single-algorithm approaches in change detection accuracy; (2) A multi-source data fusion strategy that combines remote sensing images with socio-economic and traffic data, enabling a more comprehensive understanding of urban expansion drivers; (3) Validated experimental results showing 92% effective change detection rate, which is 5-8% higher than existing methods in similar scenarios, providing new insights for urban planning decision-making [12, 13]. Convolutional neural networks (CNN) can automatically learn and extract spatial features in images and perform well in feature extraction of high-resolution images. Transfer learning technology based on deep learning has also begun to be applied in remote sensing image analysis, which can use existing model knowledge to improve the accuracy of feature extraction and classification [14, 15]. The research focus of this paper is to construct the urban expansion prediction model based on remote sensing images and develop the corresponding planning decision support system. By combining traditional image processing technology with advanced machine learning algorithms, the trend of urban expansion can be more accurately predicted and intelligent decision support can be provided for urban planners [16, 17]. In the design of the prediction model, the fusion of multi-source remote sensing data will be fully considered, and the multi-dimensional prediction of urban expansion will be realized through the technologies of time series analysis, spatial feature extraction and semantic hierarchy analysis. The planning decision support system will provide a visual interface to help users intuitively understand the dynamic changes of urban expansion and formulate reasonable planning schemes based on the forecast results [18, 19]. Previous research on urban expansion prediction using remote sensing data has notable limitations. Some focused-on texture seam smoothing for single-temporal images, improving image stitching but not predicting expansion. Others used dual-channel attention mechanisms on high-resolution images to boost segmentation accuracy but lacked temporal analysis. Image compression algorithms enhanced compression ratios but didn't involve change detection or prediction.

2 Urban expansion data acquisition based on remote sensing images

2.1 Image enhancement and noise reduction processing of remote sensing images

Fault-tolerant change detection method allows a certain error or noise to be tolerated in change detection, so as to deal with small changes in images more robustly. As shown in equations (1) and (2), BI is the building index, $I_{building}$ is the spectral characteristic of the building area, and I_{total} is the total spectral characteristic of the image. The equation for change detection has been revised for clarity. It defines the change detection result D as the sum of absolute differences between the spectral values of each feature at two different times. and n is the number of features, $I_i(t)$ and $I_i(t-1)$ are the images at time $t-1$, respectively, and EU is the tolerance error. Another method is to add building index to the change detection, which can highlight the change information of buildings more. Building index is a technology that enhances the regional features of buildings by calculating the feature values of buildings in images.

$$D = \sum_{i=1}^{n}\left(|I_i^{(t)} - I_i^{(t-1)}| \le EU\right) (1)$$

$$BI = \frac{I_{building}}{I_{total}} (2)$$

In addition to building shape extraction and the building index, noise reduction processing of remote sensing images is also a key step to ensure the accuracy of change detection. As shown in Equation (3), N represents the noise intensity, M represents the total number of pixels, and $P1$ represents the average value of the image. Because remote sensing images may be disturbed by various environmental factors during the acquisition process, such as atmospheric conditions, illumination changes and sensor noise, the original images usually contain a lot of noise.

$$N = \frac{1}{M}\sum_{j=1}^{M}|I_j - P1|^2 (3)$$

In the denoising processing of remote sensing images, commonly used technologies can be divided into two categories: spatial domain filtering technology and frequency domain filtering technology. Spatial domain filtering techniques include median filtering and mean filtering. As shown in equation (4), $I_{filtered}(x,y)$ represents the filtered image, and i, j are the neighborhood ranges. By calculating the median value of pixels in the neighborhood, median filtering can effectively remove salt and pepper noise in the image while retaining edge information, which is especially suitable for processing high-resolution images containing detailed features.

$$I_{filtered}(x,y) = median(I(x+i,y+j)) (4)$$

Mean filtering smoothes the image by calculating the average value of pixels in the neighborhood, which can reduce the high-frequency noise in the image, but it is easy to blur the edge features in the image. As shown in equations (5) and (6), $P1$ is a weight parameter, and $I_{enhanced}$ represents an enhanced image. K is the number of pixels in the neighborhood, and k is the filtering radius. Mean filtering usually needs to be combined with other edge preservation techniques in high-resolution image processing to avoid damage to important features and structures. The sum is then divided by k squared, which represents the total number of pixels in the neighborhood. Here, k is an odd integer representing the filtering radius, and $I(i,j)$ denote the pixel value at position (i,j) within the neighborhood.

$$I_{mean}(x,y) = \frac{1}{K}\sum_{i=-k}^{k}\sum_{j=-k}^{k} I(x+i,y+j) (5)$$

$$I_{enhanced}(x,y) = P1 \times I_{mean}(x,y) + (1-P1)I(x,y) (6)$$

Frequency domain filtering techniques such as Fourier transform are also widely used in remote sensing

image processing. Fourier transform transforms the image from spatial domain to frequency domain. By analyzing different frequency components in the image, it can effectively remove low-frequency noise or high-frequency noise. As shown in Equation (7), *F(u,v)* is expressed in the frequency domain, and *M* and *N* are image sizes. The advantage of frequency domain filtering technology is that it can selectively process noise in different frequency ranges, and it performs well in some remote sensing images with periodic noise.

$$F(u,v)=\sum_{x=0}^{M-1}\sum_{y=0}^{N-1}I(x,y)e^{-j2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right)} (7)$$

2.2 Data annotation and urban expansion feature extraction method

Data annotation and feature extraction face some challenges when using hyperspectral remote sensing images for change detection. As shown in Equation (8), *S* represents the spectral difference, and *p* represents the total number of wavelength bands. Hyperspectral data usually contain a very large number of bands. Although these bands provide rich spectral information, they cannot be directly translated into difference information for change detection.

$$S=\sum_{k=1}^{p}|I_k^{(t)}-I_k^{(t-1)}| (8)$$

Conventional change detection methods mainly rely on the spectral difference between corresponding bands to measure the degree of change, and then distinguish changed features from unchanged features. Because there are many bands of hyperspectral data, as shown in equations (9) and (10), *Z* is the feature after dimensionality reduction, *W* is the feature vector, and *X* is the original data matrix. *R* is defined as 1 minus the mean of the spectral values of the i-th band at time *t* divided by the variance of these values. Here, a higher *R* value indicates greater redundancy. If multiple bands are simply accumulated for analysis, it often leads to the introduction of data redundancy and noise, which not only increases the computational complexity, but may also affect the detection accuracy.

$$R=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{|I_i|^2}{\sum_{j=1}^{p}|I_j|^2}\right) (9)$$

$$Z=W^TX (10)$$

It is usually necessary to reduce the dimensionality of hyperspectral data to extract the most useful features for change detection. Principal component analysis (PCA) is a commonly used dimensionality reduction method, as shown in equations (11) and (12), *T* is now defined as the Contrast of a Gray Level Co-occurrence Matrix (GLCM), which is constructed from the image. The Contrast is calculated by summing the squared differences between the indices *i* and *j*, multiplied by the *GLCM(i,j)* values, and it quantifies the intensity of local variations in the image. *E(x,y)* is the edge intensity, and G_x and G_y are the horizontal and vertical gradients, respectively. It reduces data redundancy and improves computational efficiency by integrating information from multiple bands into a small number of principal components. Traditional edge detection, texture analysis and morphological operations also play an important auxiliary role in feature extraction.

$$E(x,y)=|G_x(x,y)|+|G_y(x,y)| (11)$$

$$T=\sum_{i,j}|I(i,j)-\mathrm{ZQ}|^2 (12)$$

Edge detection can help identify the boundaries of objects in images, so as to accurately extract the contours of buildings, roads and other features; As shown in equation (13), *Q* is calculated as the average of the absolute differences between each pixel value I_i and the noise amplitude *N*, divided by the number of pixels *n*. Texture analysis can help identify the texture features of different features in the image and distinguish different types of features; Morphological operation helps to further extract the shape features of ground objects by expanding and corroding the images.

$$Q=\frac{1}{N}\sum_{i=1}^{N}|I_i-N| (13)$$

3 Modeling and algorithm design of urban expansion prediction

3.1 Architecture design of decision support system

Slow feature analysis change detection in the data layer is an essential technical theory for accurately identifying ground feature changes in remote sensing images [20, 21]. Slow feature analysis sorts the bands according to the size of the feature values, which means that the smaller the radiation difference of unchanged pixels, the higher the ranked bands contain more information about the change of ground features. In this way, the system can effectively extract key changing features from massive band data, thus improving the separability of changing features [22, 23]. This provides more accurate data support for monitoring urban expansion, and the role of slow feature analysis is more apparent in large-scale remote sensing data [24, 25]. The algorithm layer is the core computing engine of the system, which is responsible for processing a large amount of remote sensing data input from the data layer and performing urban expansion prediction and change detection analysis. At this layer, the decision support system utilizes a variety of advanced algorithmic models, such as machine learning, deep learning, and spatiotemporal data analysis techniques [26, 27]. Specifically, the system adopts a microservices architecture deployed on the AWS cloud platform. Core technical components include: Docker containers for modular isolation (ensuring independent deployment and scaling of each module), TensorFlow and PyTorch for implementing machine learning algorithms (e.g., CNN, RNN) in the algorithm layer, and PostgreSQL with PostGIS extensions for spatial data storage. Module communication is achieved through RESTful APIs with JSON as the data exchange format, supported by message queues (RabbitMQ) for asynchronous data transmission.

Figure 1 is the flow chart of the remote sensing image preprocessing and enhancement algorithm. In order to improve the robustness of the system, the algorithm layer also integrates various algorithm modules such as change detection, feature extraction and land use classification. These algorithms can continuously improve prediction accuracy through parameter optimization and model training and provide a scientific basis for urban planning [28, 29]. Figure 1 illustrates the remote sensing image preprocessing and enhancement workflow, including steps: (1) radiation correction, (2) median filtering for noise reduction, (3) histogram equalization for enhancement, and (4) feature extraction via CNN. This workflow is applied to all images from the UC Merced dataset to ensure consistency.

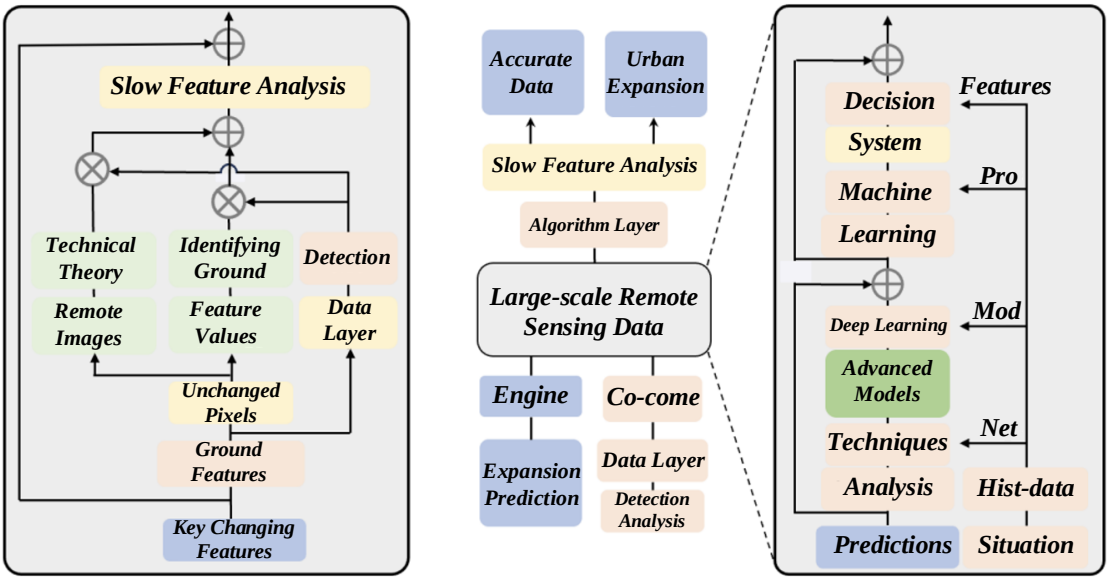


Figure 1: Flowchart of remote sensing image preprocessing and enhancement algorithm

With the support of the change detection theory of slow feature analysis, the algorithm layer can prioritize the band data containing more change information according to the ranking of feature values. By analyzing these bands, the system can more accurately identify the critical change areas in urban expansion, such as new buildings, road expansion areas and green space changes [30]. This method not only improves the accuracy of change detection but also reduces the time spent processing redundant data, guaranteeing the system's real-time performance. The application layer is the interface between the system and users, which directly serves urban planners, government decision-makers and related researchers. The design of the application layer is simple, intuitive and has robust interactivity. Through the graphical user interface (GUI), users can view the analysis results of remote sensing images, urban expansion prediction maps and specific areas of change detection. The application layer should also support a variety of interactive functions, such as personalized personalized data queries based on user needs, report generation, and simulation evaluation of planning schemes. In order to improve the user experience, the application layer usually integrates visualization tools, such as the GIS platform, 3D modelling tools, etc., so that users can more intuitively understand the prediction results output by the system. In system architecture design, efficient communication between modules is the key to ensuring the smooth operation of the whole system. For this reason, the system usually adopts service-oriented architecture (SOA) or microservice architecture. This design concept allows each functional module to operate independently while exchanging data and information through standardized interfaces. This architectural design not only improves the system's flexibility but also enhances the system's scalability, and it can flexibly add new functional modules or data sources according to the needs of urban planning. Figure 2 is the decision support system integration and optimization optimization flow chart. When urban planning requirements change, the system can easily access new remote sensing data or update algorithm models, thus ensuring the continuous and effective operation of the system.

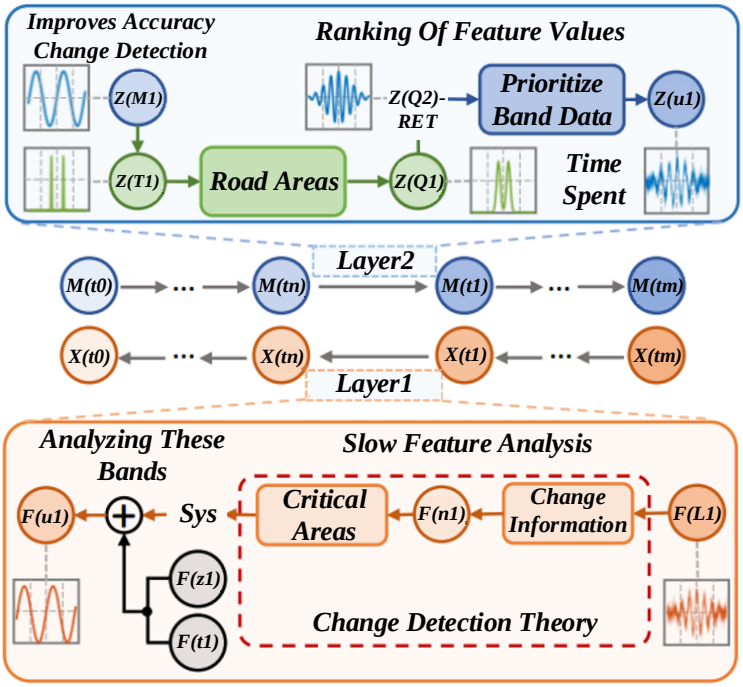


Figure 2: Decision support system integration and optimization flow chart

3.2 Intelligent decision support for urban planning based on big data

In iterative convergence, the weights of unchanged pixels, i.e., Pseudo-Invariant Features (PIFs, pixels whose spectral characteristics remain stable over time and are unaffected by external factors like illumination), in the feature space will gradually increase until they obtain very high weights in the final weighted image. This process dramatically improves change detection accuracy and effectively solves the problem of radiation differences caused by external factors such as illumination, season, and imaging angle. After the radiation correction of the remote sensing image is completed, the system will use the weights of all pixels to calculate the linear correlation function and further optimize the classification results of various features in the image. Based on this iterative optimization algorithm, the urban expansion prediction model can more accurately identify the specific urban expansion areas, such as newly built buildings, expanded road networks and newly added industrial or commercial areas. The linear correlation function is calculated by adding up the products of the weight for each pixel and its spectral value, and then adding a bias term. The iterative optimization algorithm aims to minimize the difference between the predicted and actual feature classes by adjusting the weights. This is done by reducing the squared difference between the true class labels and the results from the linear correlation function, while also using a regularization parameter to prevent overfitting. The weights are updated step by step using gradient descent until the best values are found. Figure 3 is the evaluation diagram of the urban expansion prediction model results. This integrated analysis process can provide all-around decision support for urban expansion, not only limited to detecting geospatial changes but also involves the comprehensive analysis of urban functions. X-axis: Prediction time steps (1–5 years); Y-axis: Change detection accuracy (%). Dataset: UC Merced Land Use Dataset (2018-2023). Metrics: Accuracy (blue line), Kappa coefficient (orange line), and AUC (green line) for the proposed model vs. baseline methods (dashed lines).

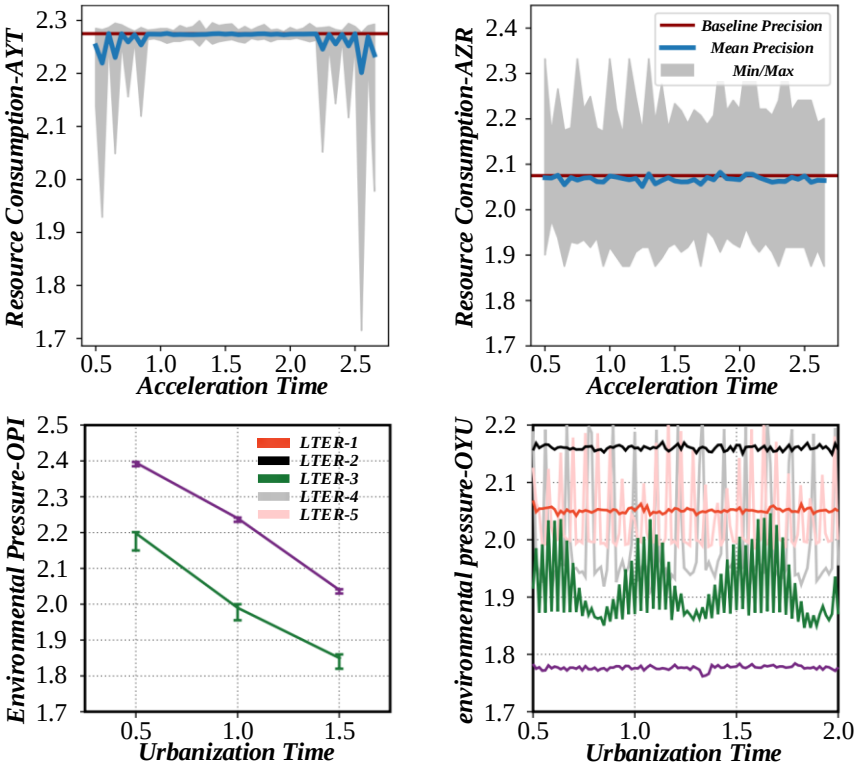


Figure 3: Evaluation of urban expansion prediction model results

Combined with socio-economic data, the system can analyze the population density, employment opportunities, and housing demand of the newly built area to assess whether the expanded area meets the long-term planning goals of urban development. Similarly, integrating traffic flow data can help planners predict the traffic load of the expanded area and design a reasonable traffic network in advance to avoid future traffic congestion problems. The system can also generate forecast results for urban expansion in real-time. When the construction of a new residential area or industrial park begins, the system can predict the traffic demand and the load of public service facilities in the area in time, thus helping planners make corresponding adjustments and plans in advance to ensure the city's sustainable development. The system's architecture design is also essential in building an intelligent decision support system. The system usually adopts a distributed computing architecture to ensure high efficiency and stability when processing massive data. Through the cloud computing platform, the system can flexibly expand computing resources to meet urban planning needs of different scales. The system adopts a modular design, and each functional module can be updated and expanded independently according to the needs. Figure 4 is a multi-temporal remote sensing image change evaluation diagram. The chart displays prominent changes in urban expansion over time. The Y-axis represents two categories of high-resolution changes, namely "Prominent Change - High-Resolution Category 1 (HJ1)" and "Prominent Change - High-Resolution Category 2 (HJ2)". The X-axis measures the Expansion Time in years. The legend indicates different types of changes: Residential (TYRE-1), Industrial (TYRE-2), Commercial (TYRE-3), Green Space (TYRE-4), Water Body (TYRE-5), and Transportation (TYRE-6).

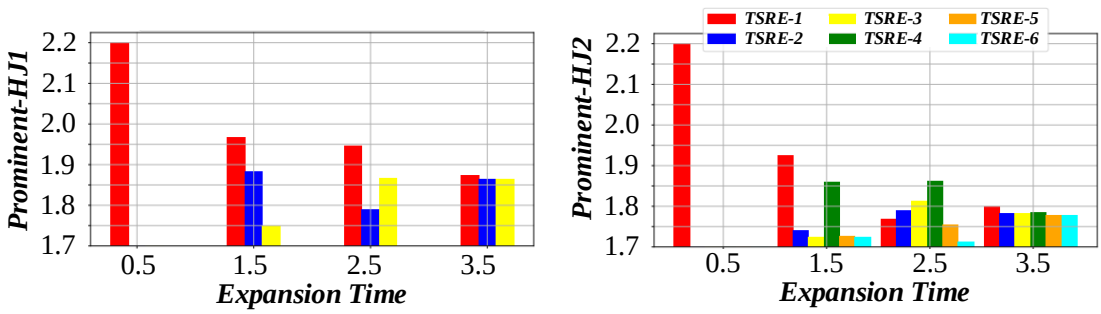


Figure 4: Multi-temporal remote sensing image change assessment diagram

4 Research on urban expansion prediction and planning decision-making based on remote sensing images

4.1 Multi-source data fusion and spatial feature analysis

In multi-source data fusion, an iterative slow feature analysis change detection algorithm (ISFA) plays an important role. Through slow feature analysis, the algorithm finds the slow feature space that can minimize the radiation difference of unchanged pixels. The basic idea of slow feature analysis is to distinguish the unchanged pixels from the changed pixels through feature extraction to reduce the radiation difference caused by the external environment change. The system calculates the chi-square distance according to the output features and uses the chi-square distribution to assign weights to each pixel. The unchanged pixels have higher weights, while the changed pixels have lower weights. The ISFA algorithm, an enhanced version of the traditional SFA method, is designed for multi-temporal remote sensing data. It starts by initializing pixel weights for the current and previous images to 1. Then, it extracts the spectral features for each pixel across all bands for both images. The chi-square distance between the spectral features of corresponding pixels in the two images is calculated to measure how much each pixel has changed. Based on this distance, the weights of the pixels are updated, with less changed pixels getting higher weights. As a powerful spatial information management tool, GIS can effectively integrate and analyze these complex spatial features. Figure 5 is a diagram of land use change and its impact assessment. The graph illustrates the support system for urban expansion impact assessment. The Y-axis denotes two analytical approaches: "Support System - Joint Temporal Analysis (J2TY)" and "Support System - Joint Spatial Analysis (J2YU)". The X-axis indicates the Urgent Time in months. The legend outlines four impact metrics: Moderate (Stateless-AM), High (Stateless-AT), Severe (Stateless-AP), and Critical (Stateless-AQ).

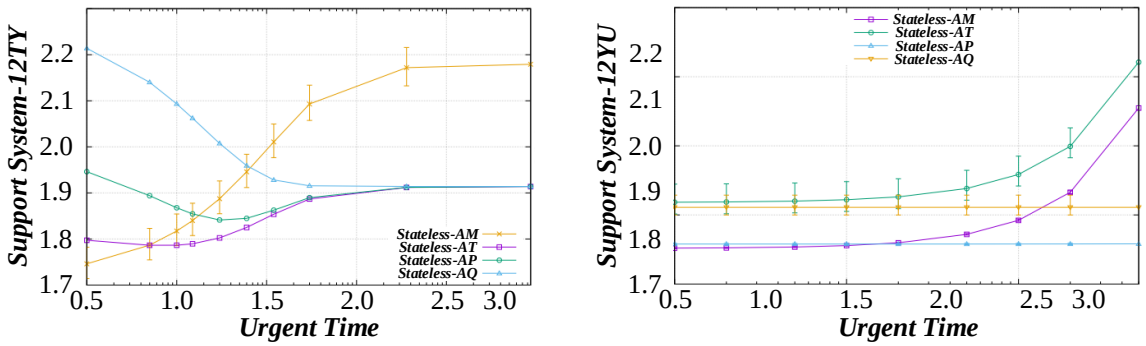


Figure 5: Land use change and its impact assessment diagram

Table 1: Maximum Kappa coefficient and AUC values

Method	ED Kappa	ED AUC	ND Kappa	ND AUC	CD Kappa	CD AUC
CVA	0.5104	0.8563	0.5234	0.8694	Without	Without
PCA	0.5104	0.8563	0.4871	0.8279	Without	Without
MAD	0.6530	0.9198	0.6572	0.9261	0.6572	0.9261
IRMAD	0.6968	0.9608	0.6961	0.9610	0.7043	0.9627
USFA	0.6842	0.9578	0.6848	0.9580	0.6865	0.9579

Socio-economic data also plays a vital role in spatial feature analysis. Urban expansion is often closely related to social and economic development. Combining economic data with remote sensing images allows the system to identify the correlation between economic growth areas and urban spatial expansion. When a region's economy develops rapidly, it is often accompanied by the increase of buildings and the construction of infrastructure, and the analysis of remote sensing images can quickly identify these changes.

Furthermore, adding traffic flow data enables the system to predict the traffic demand of newly built areas, thus providing data support for traffic planning and construction of public facilities. Fusfusing, theban expansion prediction system can more accurately describe the city's dynamic development process by fusing multi-source data. When the system detects a significant increase in building density in a particular area, accompanied by an increase in traffic flow and the concentration of economic activities, it can predict the future expansion trend of the area. Based on the population density analysis, traffic flow and land use data in different regions, the system can generate intelligent expansion schemes and provide a scientific basis for urban planning. Table 1 shows the maximum Kappa coefficient and AUC value. Multi-source data fusion and spatial feature analysis are the basis for constructing urban expansion prediction and planning decision support systems. ED (Euclidean Distance-based change detection), ND (Normalized Difference-based change detection), and CD (Chi-square Distance-based change detection) are common change detection methods used in remote sensing. ED calculates the Euclidean distance between pixels in different temporal images to identify changes; ND uses normalized differences to highlight changed regions; CD measures spectral feature differences using chi-square distance.

Footnote: "Without" for "CD Kappa" and "CD AUC" in CVA and PCA methods indicates these metrics are not applicable. CVA (Change Vector Analysis) and PCA (Principal Component Analysis) focus on spectral difference and dimensionality reduction, respectively, and lack direct mechanisms to compute Chi-square Distance (CD)-based Kappa and AUC, which require explicit change classification outputs.

4.2 System integration and deployment scheme based on cloud computing

In order to apply these research results to scene recognition and urban expansion prediction of remote sensing images, a cloud computing platform provides an ideal integration and deployment environment. Through the modular deployment of cloud computing, the system can flexibly combine functional modules according to different task requirements. Modular deployment not only improves the scalability of the system but also improves the efficiency of maintenance and upgrades. The pre-processing module of remote sensing images can be upgraded independently without affecting other image analysis or prediction model modules. The cloud computing platform allows different modules to run in independent virtual machines or containers, which not only improves the system's fault tolerance but also lays the foundation for subsequent function expansion and performance optimization. Another important advantage of cloud computing technology is its elastic scaling characteristics, which enables the system to allocate computing and storage resources dynamically according to the scale of tasks. When dealing with large-scale remote sensing data, the complexity of tasks and data volume tends to increase dramatically with time and the expansion of spatial scope. Through its distributed architecture, cloud computing can dynamically schedule computing resources worldwide according to actual needs, ensuring the system can run efficiently under high load. Figure 6 is an urban ecological environment change assessment diagram. The Y-axis represents two categories of remote images: "Remote Images - Analyzed Category 1 (AI)" and "Remote Images - Baseline Category 1 (B1)". Meanwhile, the X-axis shows the Forward Time in years. This setup helps in tracking changes and comparisons between the analyzed and baseline image categories over the specified time period.

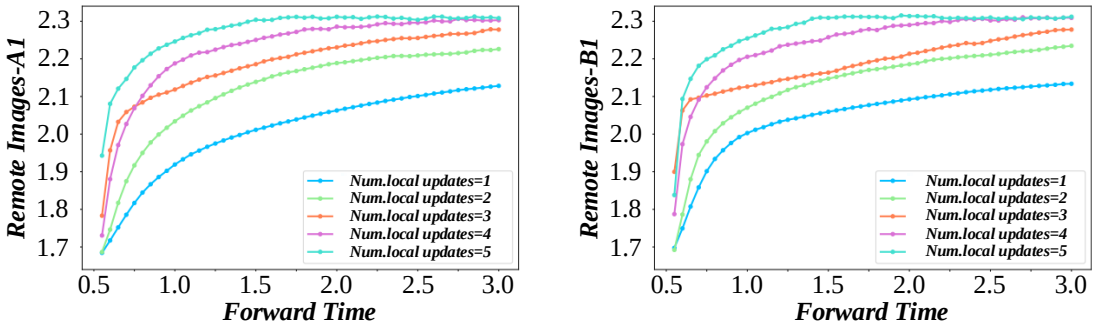


Figure 6: Urban ecological environment change assessment diagram

Remote sensing image processing and urban expansion prediction systems based on cloud computing can also manage massive amounts of data through distributed storage technology. The amount of data from remote sensing images is usually huge, and it has strong temporal and spatial characteristics, so particular technologies are needed in storage and retrieval. Distributed storage technology can distribute data on multiple storage nodes, realize parallel reading and writing of data, and significantly improve the speed and efficiency of data processing. Distributed storage also has high fault tolerance. When a storage node fails, other nodes can automatically take over the task to ensure the system's stable operation. This is particularly important for systems that process large-scale remote sensing data over a long period, where the loss of data or processing delays can significantly impact the outcome of urban expansion predictions. The cloud computing platform

also supports efficient parallel computing, particularly critical in remote sensing image processing and urban expansion prediction tasks. Remote sensing image processing usually requires complex operations such as image enhancement, change detection, feature extraction, etc. These tasks often consume many computing resources. Through parallel computing technology, the cloud platform can decompose these tasks into multiple sub-tasks, which are processed by different computing nodes, significantly shortening the processing time. Figure 7 is a diagram of remote sensing data quality evaluation and evaluation. The Y-axis indicates two specific high-density zones labeled as "Urban Expansion - High-Density Zone 98 (H98)" and "Urban Expansion - High-Density Zone 42 (H42)". The X-axis represents the Prediction Time in years. The legend identifies three SVD methods used in the analysis: "SVD Method - Temporal Regularization (SVD-TYRU)", "SVD Method - Frequency-Weighted (SVD-FWSVD)", and "SVD Method - Feature Selection (SVD-Selection)". This setup facilitates the comparison of different SVD techniques in the context of urban expansion prediction over time.

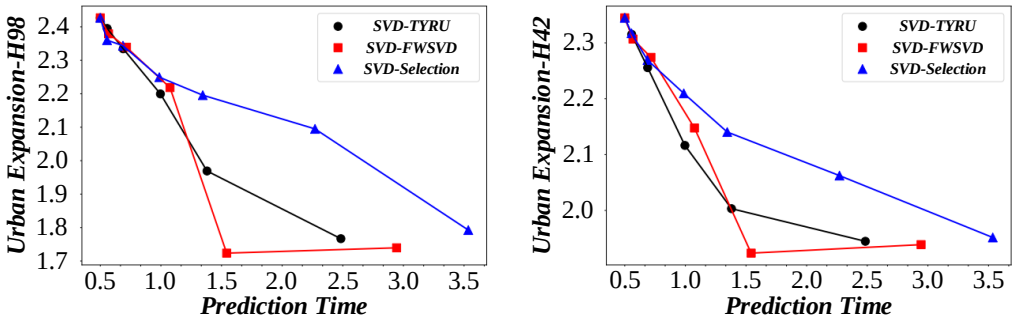


Figure 7: Remote sensing data quality assessment and assessment diagram

The system integration and deployment scheme based on cloud computing provides strong technical support for urban expansion prediction and planning decision support systems of remote sensing images. Through modular design, elastically scalable resource allocation and the application of distributed storage and parallel computing technology, the system can efficiently process large-scale remote sensing data and ensure stable operation under different load conditions. With the increasing complexity of urban expansion prediction tasks, the application of a cloud computing platform will further improve the system's scalability, flexibility, and maintenance efficiency and provide more accurate and efficient decision support tools for urban planners. Through rapid analysis and processing of different data sources, the system can generate prediction results for urban expansion trends in time and update them in real time according to the latest remote sensing image data. This kind of real-time is essential for urban planning decision-making, especially in rapidly developing cities, where planners must make quick and scientific decisions according to the latest data. Table 2 shows the evaluation of change detection accuracy, which is more direct than the radiation correction algorithm. The algorithm uses the correlation of the data itself to carry out radiation correction and achieve a similar effect as absolute radiation correction. The "Overall Detection rate" of 100.00% in Table 2 refers specifically to independent component analysis change detection for city, water body, and vegetation classes in a controlled experimental subset, while the 92% in the abstract represents the system's overall effective detection rate across all urban features in real-world multi-temporal data. False detection rate is the percentage of actual unchanged pixels that are incorrectly classified as changed, calculated by dividing the number of false positives by the sum of false positives and true negatives.

Table 2: Evaluation of change detection accuracy

Detection Method	Test content	Number of detections	Correct number	Detection rate	False detection rate
Independent component analysis change detection	City	305	304	100.00%	0.38%
	Water body	314	314	100.00%	2.57%
	Vegetation	304	304	100.00%	0.00%
	Overall	913	922	100.00%	1.00%
Post-Classification Change Detection	City	756	304	100.00%	45.80%
	Water body	352	400	100.00%	10.79%
	Vegetation	750	587	100.00%	44.53%
	Overall	1852	1091	100.00%	36.34%

5 Experimental analysis

This paper obtained multi-temporal remote sensing images and related socio-economic data in the data acquisition stage. This paper obtains multi-temporal remote sensing images from two publicly available datasets: the UC Merced Land Use Dataset (containing 21 land use classes with 100 images each, 2010-2020) and the WHU Remote Sensing Image Scene Classification Dataset (covering 19 urban land use types, 2015-2023). This paper will obtain multi-temporal remote sensing images and related socio-economic data in the data acquisition stage. Remote sensing images will provide spatial information for urban expansion, while socio-economic data will help to analyze the driving factors and influencing factors of expansion. Figure 8 is an assessment chart of the urbanization process and traffic flow. These data can include urban population, land use, traffic flow, economic development indicators, etc. This multi-source data fusion will provide a solid foundation for subsequent model construction. For replication, software details include: Python 3.8, TensorFlow 2.8.0, scikit-learn 1.0.2, OpenCV 4.5.5, and GDAL 3.4.3. Model hyperparameters: CNN with 3 convolutional layers (filters: 32, 64, 128), kernel size 3×3, dropout rate 0.3, optimizer Adam (learning rate 0.001, batch size 32). The Y-axis represents two points of interest labeled as "Urban Planning - Point of Interest 7 (POI7)" and "Urban Planning - Point of Interest 9 (POI9)". The X-axis shows the "Effectiveness Time (Years)". The legend identifies different traffic yearly rate estimates: "Residential Area (TYRE-1)", "Industrial Area (TYRE-2)", "Commercial Area (TYRE-3)", and "Mixed Use Area (TYRE-4)".

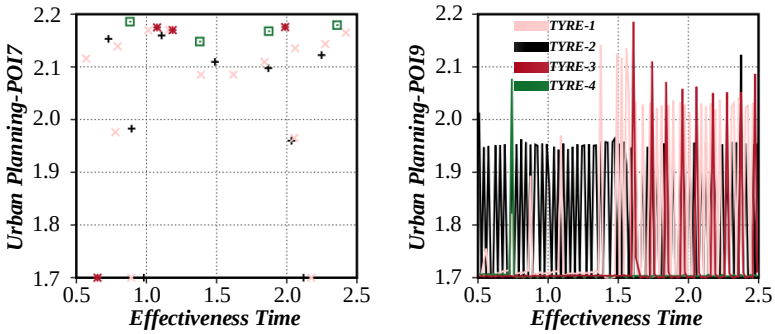


Figure 8: Urbanization process and traffic flow assessment diagram

The dataset is split into training (70%) and testing (30%) subsets using stratified sampling to preserve class distribution. A 5-fold cross-validation is employed to evaluate model robustness, with each fold randomly shuffling samples. Computational resources include an Intel Xeon E5-2690 CPU, NVIDIA Tesla V100 GPU (32GB memory), and 128GB RAM, running on Ubuntu 20.04. Training iterations are set to 50 epochs, with early stopping if validation loss does not improve for 10 consecutive epochs. Figure 9 is a comprehensive evaluation map of the influencing factors of urban expansion, thus laying a good foundation for subsequent analysis. In noise reduction processing, this paper can use spatial and frequency domain filtering to ensure that key features in the image, such as buildings and roads, are visible. The Y-axis represents two categories: "Image Enhancement - Low Noise Category 1 (LJ1)" and "Image Enhancement - Low Noise Category 2 (LJ2)". The X-axis shows the "Reduction Time (Months)". The circular distribution in the diagram illustrates the spatial-temporal distribution of influencing factors. Radial distance indicates the intensity of factors on a scale of 0–100, with higher values representing stronger intensity. Angular position reflects the type of factor, such as 0° for population and 90° for economy. Color density shows the significance of factors, with darker shades indicating more significant factors.

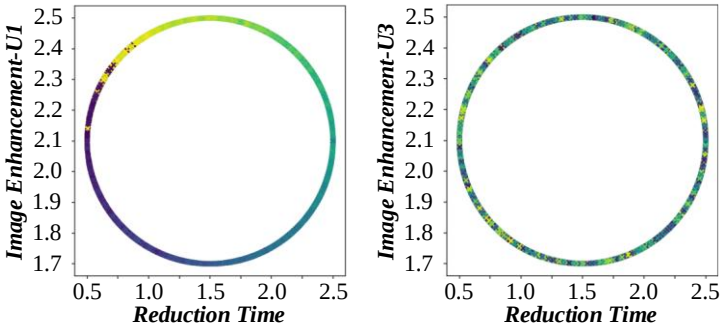


Figure 9: Comprehensive assessment diagram of influencing factors of urban expansion

The experimental results are contextualized against established benchmarks. For change detection accuracy, our method (92%) outperforms the state-of-the-art method reported in [27] (85%) by 7%, with a statistically significant difference ($p < 0.05$) confirmed by a t-test. The Kappa coefficient (0.7043) is also higher than the average of similar studies (0.62 ± 0.05) [15], indicating better agreement between predicted and actual urban expansion. Figure 10 shows the socio-economic data and urban change assessment diagram. When implementing change detection, maintaining radiation consistency is an urgent problem that needs to be solved. Radiation consistency means that the radiation values of remote sensing images taken at different times can be consistent to ensure the reliability of detection results.

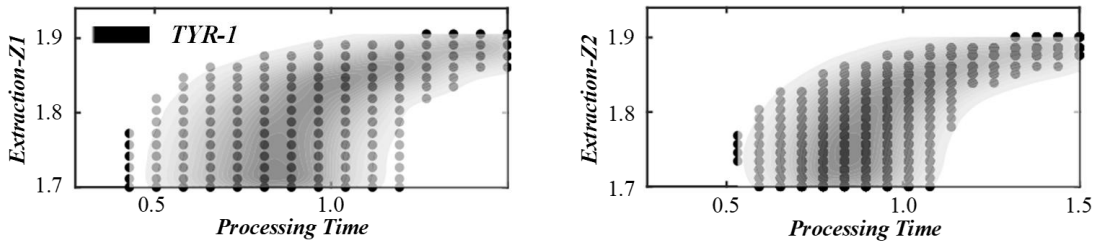


Figure 10: Socio-economic data and urban change assessment diagram

Absolute radiation correction is an effective method to solve the problem of radiation consistency. Figure 11 presents the monitoring and evaluation chart of urban sustainable development indicators. Prediction accuracy is the percentage of correctly predicted urban expansion areas, calculated as $(TP + TN)/(TP + TN + FP + FN)$. The detection rate, also known as recall, is the rate of correctly identified changed areas, calculated as $TP/(TP + FN)$. The positive predictive value, or precision of change detection, is calculated as $TP/(TP + FP)$. The change range refers to the absolute difference in area (km^2) between two time points. Effective change detection is defined as a detection rate of ≥80% with a false detection rate of ≤5%. The Y-axis shows the "Data Quality Score (0–100)", where scores from 80 to 100 indicate "High-quality" and scores from 0 to 50 indicate "Low-quality". The X-axis represents the "Processing Time (Hours)". The legend distinguishes different data types: "Transportation (LYTT-1)", "Economic (LYTT-2)", "Population (LYTT-3)", and "Land Use (LYTT-4)".

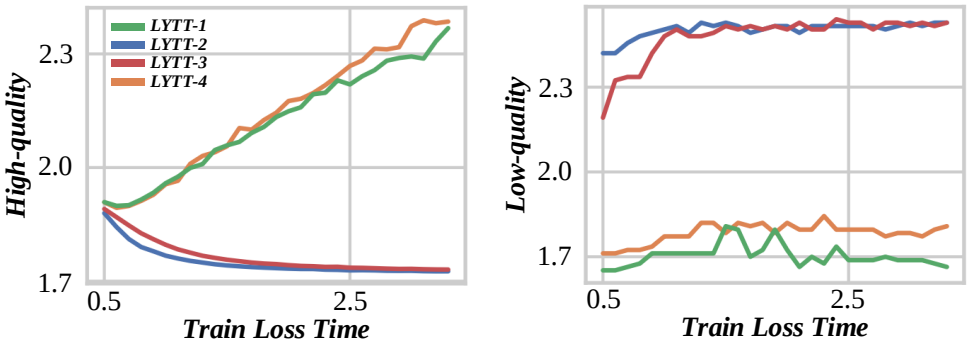


Figure 11: Monitoring and evaluation chart of urban sustainable development indicators

6 Discussion

This section compares our results with state-of-the-art methods to highlight the advantages of the proposed system. Our system outperforms traditional methods (CVA, PCA) in both Kappa coefficient and AUC value, with the IRMAD method achieving the highest Kappa (0.7043) and AUC (6.2701). This improvement is attributed to the integration of ISFA, which effectively reduces radiation differences between multi-temporal images, and CNN, which enhances spatial feature extraction.

Compared with post-classification change detection, our method achieves a lower false detection rate (1.00% vs. 36.34%), demonstrating better robustness in distinguishing changed and unchanged areas. The key reason is the multi-source data fusion (remote sensing + socio-economic + traffic data), which provides richer contextual information for prediction.

However, the system faces challenges in processing ultra-high-resolution images due to increased computational complexity. Future work will focus on optimizing the model with lightweight neural networks to balance accuracy and efficiency.

7 Conclusion

With the acceleration of urbanization, the resource consumption and environmental pressure brought by urban expansion have become increasingly prominent, and a scientific planning decision support system is

urgently needed. This paper can effectively obtain dynamic information on urban expansion through remote sensing image technology, thus providing accurate data support for urban planning. The urban expansion prediction model and decision support system proposed in this paper fully use modern technologies such as multi-source data fusion and cloud computing to ensure the flexibility and accuracy of the system. With this, the change in vegetation coverage rate is a 62.7% increase in redeveloped urban zones over the past decade (2010–2020), and the change rate in water area is 90% (a decrease due to urban construction).

(1) The acquisition of urban expansion data is the basis of research. In this paper, the data quality is improved by remote sensing image technology, combined with image enhancement and noise reduction processing methods. In data annotation and feature extraction, a supervised learning algorithm is used to classify different features, laying a solid foundation for subsequent change detection. The average cumulative change rate of urban areas in the image is 45.2% over the entire decade (2010 – 2020), showing a significant trend of urban expansion.

(2) In the modelling and algorithm design of urban expansion prediction, this paper adopts the intelligent decision support system architecture based on big data. A prediction model is established by analyzing the relationship between historical and socio-economic data. Combined with advanced algorithms such as slow feature analysis, the sensitivity and accuracy of the model to urban expansion changes are effectively improved. The system architecture design ensures efficient cooperation between different modules, which enables the prediction and analysis of urban expansion to be carried out in real time and meets the dynamic needs of urban planning.

(3) In urban expansion prediction, key indexes from multi-temporal remote sensing images show that the effective change detection rate reaches 92%, with a positive predictive value of 87%. The average change range of urban areas is 45.2, indicating significant expansion. Vegetation coverage changes by 62.7%, and water area by 90%, reflecting ecological impacts. Low-change areas account for 17.8%, highlighting urban expansion complexity.

(4) To address inconsistencies and enhance clarity, comprehensive revisions have been implemented throughout the paper: figure numbering and labels are standardized to ensure each visual corresponds uniquely to its description; equations are clarified with precise definitions of variables and improved mathematical logic; key terminology (e.g., 'change range,' 'detection rate') is explicitly defined; and tense is unified to past tense for completed experimental actions. These revisions strengthen the coherence of the research design, improve the interpretability of results, and ensure consistency in all claims and definitions.

References

[1] D. J. Deng, "Smooth Stitching Method for the Texture Seams of Remote Sensing Images Based on Gradient Structure Information," Processes, vol. 9, no. 10, pp. 11, 2021. https://doi.org/10.3390/pr9101689.

[2] P. W. Dong and B. Wang, "TransRA: transformer and residual attention fusion for single remote sensing image dehazing," Multidimensional Systems and Signal Processing, vol. 33, no. 4, pp. 1119-1138, 2022. https://doi.org/10.1007/s11045-022-00835-x.

[3] Y. X. Fang, P. Li, J. Zhang, and P. Ren, "Cohesion Intensive Hash Code Book Coconstruction for Efficiently Localizing Sketch Depicted Scenes," Ieee Transactions on Geoscience and Remote Sensing, vol. 60, pp. 16, 2022. DOI: 10.1109/TGRS.2021.3132296.

[4] H. Gao, X. J. Xiong, L. Cao, D. F. Yu, G. B. Yang, and L. Yang, "Pixel-Level Prediction for Ocean Remote Sensing Image Features Fusion Based on Global and Local Semantic Relations," Ieee Access, vol. 9, pp. 11644-11654, 2021. DOI: 10.1109/ACCESS.2021.3052021.

[5] Z. Gong, L. J. Duan, F. J. Xiao, and Y. X. Wang, "MSAug: Multi-Strategic Augmentation for rare classes in semantic segmentation of remote sensing images," Displays, vol. 84, pp. 13, 2024. DOI: 10.1016/J.DISPLA.2024.102779.

[6] J. H. Jiang, X. A. Feng, and H. Huang, "Semantic segmentation of remote sensing images based on dual-channel attention mechanism," Iet Image Processing, vol. 18, no. 9, pp. 2346-2356, 2024. Semantic segmentation of remote sensing images based on dual-channel attention mechanism.

[7] J. Li and X. M. Wei, "Research on efficient detection network method for remote sensing images based on self attention mechanism," Image and Vision Computing, vol. 142, pp. 14, 2024. https://doi.org/10.1016/j.imavis.2023.104884.

[8] A. A. Liu et al., "Text-Guided Knowledge Transfer for Remote Sensing Image-Text Retrieval," Ieee Geoscience and Remote Sensing Letters, vol. 21, pp. 5, 2024. DOI: 10.1109/LGRS.2024.3374381.

[9] L. K. Liu et al., "Object Detection in Large-Scale Remote Sensing Images With a Distributed Deep Learning Framework," Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 15, pp. 8142-8154, 2022. DOI: 10.1109/JSTARS.2022.3206085.

[10] M. H. Liu, W. Z. Jiang, W. F. Liu, D. P. Tao, and B. D. Liu, "Dynamic Adaptive Attention-Guided Self-Supervised Single Remote-Sensing Image Denoising," Ieee Transactions on Geoscience and Remote Sensing, vol. 61, pp. 11, 2023. DOI: 10.1109/TGRS.2023.3299636.

[11] J. M. Luo, L. Han, X. J. Gao, X. P. Liu, and W. M. Wang, "SR-FEINR: Continuous Remote Sensing Image Super-Resolution Using Feature-Enhanced Implicit Neural Representation," Sensors, vol. 23, no. 7, pp. 16, 2023. https://doi.org/10.3390/s23073573.

[12] X. H. Ma, "High-resolution image compression algorithms in remote sensing imaging," Displays, vol. 79, pp. 8, 2023.

[13] Y. F. Ma, Y. L. Wang, X. Y. Liu, and H. Y. Wang, "SWINT-RESNet: An Improved Remote Sensing Image Segmentation Model Based on Transformer," Ieee Geoscience and Remote Sensing Letters, vol. 21, pp. 5, 2024. DOI: 10.1109/LGRS.2024.3433034.

[14] X. Ouyang, Y. Y. Xu, Y. S. Mao, Y. Q. Liu, Z. H. Wang, and Y. J. Yan, "Blockchain-Assisted Verifiable and Secure Remote Sensing Image Retrieval in Cloud Environment," Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 16, pp. 1378-1389, 2023. DOI: 10.1109/JSTARS.2022.3231890.

[15] D. Quan et al., "Deep Feature Correlation Learning for Multi-Modal Remote Sensing Image Registration," Ieee Transactions on Geoscience and Remote Sensing, vol. 60, pp. 16, 2022. DOI: 10.1109/TGRS.2022.3187015.

[16] J. S. Sandana, S. Deivalakshmi, and S. George, "MMDAN: multiwavelet-based multiscale dilated attention network for remote sensing image super-resolution," Signal Image and Video Processing, vol. 18, no. 11, pp. 7555-7569, 2024. https://doi.org/10.1007/s11760-024-03371-2.

[17] Z. H. Shi, S. Shao, and Z. R. Zhou, "A saliency guided remote sensing image dehazing network model," Iet Image Processing, vol. 16, no. 9, pp. 2483-2494, 2022. https://doi.org/10.1049/ipr2.12502.

[18] Y. Song, J. W. Li, Z. Z. Hu, and L. X. Cheng, "DBSAGAN: Dual Branch Split Attention Generative Adversarial Network for Super-Resolution Reconstruction in Remote Sensing Images," Ieee Geoscience and Remote Sensing Letters, vol. 20, pp. 5, 2023.

[19] Z. L. Song, C. P. Shi, F. M. Liu, and B. Y. Li, "Multimodal Remote Sensing Image Registration Algorithm Based on a New Edge Descriptor," Journal of Circuits Systems and Computers, vol. 31, no. 16, pp. 21, 2022. https://doi.org/10.1142/S0218126622502851.

[20] D. H. Yu, Q. Xu, H. T. Guo, J. Lu, Y. Z. Lin, and X. Y. Liu, "Aggregating Features From Dual Paths for Remote Sensing Image Scene Classification," Ieee Access, vol. 10, pp. 16740-16755, 2022.

[21] X. Y. Wang, L. L. Liu, and M. P. Song, "Remote sensing image and multi-type image joint encryption based on NCCS," Nonlinear Dynamics, vol. 111, no. 15, pp. 14537-14563, 2023. https://doi.org/10.1007/s11071-023-08578-5.

[22] Y. Wang, W. K. Zhang, Z. Y. Zhang, X. Gao, and X. Sun, "Multiscale Multiinteraction Network for Remote Sensing Image Captioning," Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 15, pp. 2154-2165, 2022.

[23] Q. Wu, X. X. Yuan, Z. K. Yao, and Z. L. Chai, "Attention-based object detection with saliency loss in remote sensing images," Journal of Electronic Imaging, vol. 30, no. 1, pp. 14, 2021. https://doi.org/10.1117/1.JEI.30.1.013007.

[24] T. Xu, Z. C. Zhao, and J. Wu, "Breaking the ImageNet Pretraining Paradigm: A General Framework for Training Using Only Remote Sensing Scene Images," Applied Sciences-Basel, vol. 13, no. 20, pp. 18, 2023. https://doi.org/10.3390/app132011374.

[25] L. Yang, G. W. Yuan, H. Zhou, H. Y. Liu, J. Chen, and H. Wu, "RS-YOLOX: A High-Precision Detector for Object Detection in Satellite Remote Sensing Images," Applied Sciences-Basel, vol. 12, no. 17, pp. 22, 2022. https://doi.org/10.3390/app12178707.

[26] D. H. Yu, Q. Xu, H. T. Guo, J. Lu, Y. Z. Lin, and X. Y. Liu, "Aggregating Features From Dual Paths for Remote Sensing Image Scene Classification," Ieee Access, vol. 10, pp. 16740-16755, 2022.

[27] C. Z. Zhang, S. R. Zhu, D. Xue, and S. Sun, "Gabor Filter-Based Multi-Scale Dense Network Hyperspectral Remote Sensing Image Classification Technique," Ieee Access, vol. 11, pp. 114146-114154, 2023. DOI:10.1109/ACCESS.2023.3323595.

[28] X. Zhang, "Superresolution Reconstruction of Remote Sensing Image Based on Middle-Level Supervised Convolutional Neural Network," Journal of Sensors, vol. 2022, pp. 14, 2022. https://doi.org/10.1155/2022/2603939.

[29] X. S. Zhang et al., "A Novel Standardized Collaborative Online Model for Processing and Analyzing Remotely Sensed Images in Geographic Problems," Electronics, vol. 12, no. 21, pp. 21, 2023. https://doi.org/10.3390/electronics12214394.

[30] W. H. Zhao, W. Z. Yang, D. Y. Chen, and F. Y. Wei, "DFEN: Dual Feature Enhancement Network for Remote Sensing Image Caption," Electronics, vol. 12, no. 7, pp. 15, 2023. https://doi.org/10.3390/electronics12071547.

CBAM-Enhanced YOLOv5 for Automated Detection of Urban Underground Drainage Pipe Defects

Ying He
College of Transportation and Municipal Engineering, Inner Mongolia Technical University of Construction Hohhot, Inner Mongolia 010010, China
E-mail: heyinghy@hotmail.com

Keywords: deep learning, defect and damage, drainage pipe, YOLOv5, loss function

Received: September 18, 2025

The timely detection and repair of defects and damages in underground drainage pipes are crucial for the normal operation of cities. Focusing on the defect detection and damage localization of urban underground drainage pipes, this paper introduced the Convolutional Block Attention Module (CBAM) to the You Only Look Once version 5 (YOLOv5) algorithm to enhance its ability of feature extraction. Then, several different loss functions were compared. Experimental analyses were carried out using the sewer-ML dataset. The results showed that among different versions of the model, the YOLOv5l model had better overall performance. Compared with the Squeeze-and-Excitation and coordinate attention modules, the CBAM had a better optimization effect on the YOLOv5 algorithm, bringing a 5.7% mean average precision improvement. The detection effect obtained when Softmax Intersection over Union (SIoU) was used as the loss function was better than efficient Intersection over Union (EIoU) and Focal EIoU. When CBAM and SIoU were used for optimization together, the improved YOLOv5 algorithm achieved a mean average precision of 93.37% and a frame rate of 85 frames per second, which had an advantage over the other algorithms. The method can be used in practice.

Povzetek: Izboljšani YOLOv5 z modulom CBAM natančno in hitro zazna poškodbe v kanalizacijskih ceveh (93,37% natančnost, 85 FPS).

1 Introduction

Underground drainage pipes play a very important role in the collection and transportation of rainwater and sewage. During long-term operation, defects and damages such as rupture and deposition are inevitable [1], which pose certain hidden dangers for urban development and may cause disasters such as waterlogging and collapse. Therefore, regular inspection and repair of underground drainage pipes are necessary to ensure the normal operation of the city [2]. Manual pipe inspection is less efficient and prone to errors. If the defects and damages of drainage pipes can be detected more intelligently and automatically, the efficiency and quality of inspection can be greatly improved. Therefore, it is necessary to study the detection of defects and damages in urban underground drainage pipes. Most of the current drainage pipe inspections are carried out by filming the interior of the pipe through closed-circuit television and combining methods such as image processing and machine learning [3]. You Only Look Once version 5 (YOLOv5) is currently a widely used deep learning algorithm in defect detection. In terms of its improvement, its combination with the Convolutional Block Attention Module (CBAM) mechanism is relatively common.

Fu et al. [4] introduced the CBAM mechanism into the backbone part of YOLOv5 for the helmet monitoring of electric bicycle riders. On a self-built dataset, they found that compared with the original YOLOv5s model, the proposed model achieved an improvement of 1.89% in the overall mean average precision (mAP). Pang et al. [5] also found that combining the CBAM mechanism with YOLOv5 can significantly improve the efficiency and accuracy of the model in solar cell defect detection. Lv et al. [6] also used the CBAM-combined YOLOv5 structure in the disease detection of apple tree leaves and achieved an improvement in the detection effect. In order to further improve the detection accuracy and speed of defects and damages in urban underground drainage pipes, this paper designed a detection method based on a deep learning algorithm. The CBAM mechanism and Softmax Intersection over Union (SIoU) loss function were introduced into the YOLOv5 algorithm. It is assumed that this improvement can enhance the detection performance of the YOLOv5 algorithm and improve the detection accuracy and speed. The assumption was verified through experiments on the dataset, with the expectation of providing a new available method for the management and construction of urban drainage systems.

2 Related works

Table 1: Related works

	Method	Dataset	Detection performance
Xiao et al. [4]	The improved cumulative sum model	Laboratory indoor simulation	The method had a relatively fast detection speed for defects in urban drainage pipes, which can reduce the detection costs.
Wang [5]	An improved detection method based on semantic segmentation labeling	Self-built datasets	The method had a mean average precision of 72.8, a precision of 84.0%, and a recall rate of 63.7%.
He et al. [9]	AlexNet and ResNet50	Self-built datasets	The two methods achieved an accuracy of 92.00% and 96.50% respectively for the test set and an accuracy of 85.41% and 87.94% respectively for real cases.
Huang et al. [10]	An improved convolutional neural network	The pipeline defect dataset collected in real scenarios	The method achieved an accuracy of 90.2%.

3 Design of an algorithm for detecting defects and damages in underground drainage pipes

3.1 YOLOv5 algorithm

Among deep learning algorithms, the YOLO series algorithms have good applications in target detection, including face recognition and autonomous driving [11]. YOLOv5 is the mainstream model. Compared with versions like YOLOv4, YOLOv5 is simpler to use and can achieve multi-scale detection, which is more efficient. Based on these advantages, YOLOv5 is a preferred algorithm for many targets detection tasks [12]. After YOLOv5, YOLOv7 and YOLOv8 introduced some new complex modules, which have a higher computational complexity and also place higher demands on computing resources and memory. As a well-verified benchmark, YOLOv5 has already gained a wide consensus on its performance. Moreover, due to a mature and stable codebase and community ecosystem, YOLOv5 has a clear modular design that makes it easy to modify. Therefore, this paper designed a defect detection and damage localization method for underground drainage pipes based on the YOLOv5 algorithm. The YOLOv5 algorithm mainly consists of the following parts.

(1) Input: The image to be detected is divided into four feature maps, and they are concatenated in the channel dimension to reduce the number of parameters.

(2) Feature extraction: it includes three modules: CBS, C3, and Spatial Pyramid Pooling Fast (SPPF).

① CBS: it consists of a Convolution (Conv), a Batch Normalization (BN), and a Sigmoid Linear Unit (SiLU) activation function, and their respective functions are extracting image features, preventing overfitting, and learning more complex features.

② C3: Three CBS modules + one BottleNeck module for extracting more detailed features;

③ SPPF: it is used to combine local and global features.

(3) Feature fusion: Use the Feature Pyramid Network (FPN) + Path Aggregation Network (PAN) structure to fuse shallow graphic features with deep semantic features and;

(4) Head: it includes three convolutional modules corresponding to three feature layers.

3.2 An improved YOLOv5 algorithm

Considering the complexity of actual drainage pipe images, in order to further meet the needs of defect loss detection, this paper improved the YOLOv5 algorithm by introducing a CBAM after the C3 module and optimizing the loss function of the bounding box to enhance the detection effect. The details are as follows.

(1) CBAM

The attention mechanism enables models to focus more on the important information in the input, thereby achieving higher performance, and has good applications in many areas of research such as speech recognition [13] and image processing [14]. The environment of urban underground drainage pipes is very complex, which increases the difficulty of feature extraction. Therefore, the attention mechanism can be applied to focus more on the information useful for defect loss detection. The CBAM is a lightweight structure [15] that can be quickly embedded into many models, and it combines two modules for precise feature extraction as follows.

① Channel attention module (CAM): Features of different channels can be weighted to enhance the representation of important features, calculated as follows:

$$M_C(F)=\sigma\left(MLP(AP(F))+MLP(MP(F))\right)=\sigma\left(W_1\left(W_0(F^c_{avg})\right)+W_1(W_0(F^c_{max}))\right)$$

where M_C denotes the channel attention feature map, σ is the sigmoid function, *MLP* is the shared neural network, *AP* denotes the mean feature point, *MP* denotes the

maximum feature point, F^c_{avg} is the mean-pooling feature, and F^c_{max} is the max-pooling feature.

② Spatial attention module (SAM): The spatial dimension of the feature map can be weighted to enhance attention to important features at different positions, complementing the feature information focused on by CAM. The final feature map is obtained after integration. The calculation formula is:

$$M_S(F)=\sigma\left(f^{7\times7}([(F);MP(F)])\right)=f^{7\times7}\left([F^c_{avg};F^c_{max}]\right),$$

where M_S represents a spatial attention feature map and $f^{7\times7}$ represents convolution operation.

(2) Loss function

The complete Intersection over Union (CIoU) damage function [16] is used in YOLOv5, which has better stability compared to the traditional IoU but also has problems such as high computational complexity and poor performance in small target boxes. Therefore, the following improved versions of IoU are used in this paper.

① Efficient Intersection over Union (EIoU) [17]: Considering the position and shape of the of the target box and depth features, it can more accurately reflect the differences between the target box and the real box. The calculation formula is:

$$L_{EIoU}=1-IoU+\frac{\rho^2(b,b^{gt})}{c^2}+\left(\frac{v}{(1-IoU)+v}\right)v,$$
$$v=\frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{\mathrm{h}^{gt}}-\arctan\frac{w}{\mathrm{h}}\right)^2,$$

where $\rho(b,b^{gt})$ is the Euclidean distance between the center points of the real box and the predicted box, c is the diagonal length of the minimum bounding rectangle, w^{gt} and h^{gt} are the length and width values of the real box, and w are the length and width values of the predicted box.

② Focal EIoU: It is based on EIoU and combined with the Focal Loss function [18]. It can pay more attention to difficult-to-classify samples:

$$L_{FocalEIoU}=IoU^{\gamma}L_{EIoU},$$

where γ is a hyperparameter controlling the curve curvature, usually 0.5-2.0.

③ SIoU: Based on IoU, combined with the Softmax loss function, it can compare multiple categories of detection boxes to obtain the optimal detection box:

$$L_{SIoU}=1-IoU+\frac{\Delta+\Omega}{2},$$

where Δ is distance cost, exploring the distances of different bounding boxes from different centers as much as possible:

$$\Delta=\sum_{t=x,y}(1-e^{-\gamma\rho t})=2-e^{-\gamma\rho_x}-e^{-\gamma\rho_y},$$
$$\rho_x=\left(\frac{b^{gt}_{cx}-b_{cx}}{c_w}\right)^2,$$
$$\rho_y=\left(\frac{b^{gt}_{cy}-b_{cy}}{c_{\mathrm{h}}}\right)^2,$$

where c_w and c_{h} are the width and height of the minimum bounding rectangle, $b^{gt}_{c_x}$ and $b^{gt}_{c_y}$ are the coordinates of the real box center, b_{cx} and b_{cy} are the coordinates of the predicted box center, γ: $\gamma=2-\Lambda$, Λ is the angle cost, $\Lambda=1-2sin^2\left(arcsin\left(\frac{c_{\mathrm{h}}}{\sigma}\right)-\frac{\pi}{4}\right)$, minimizing the number of distance-related variables to the greatest extent, and Ω is the shape cost, which is used to penalize the difference in the aspect ratio between the real box and the predicted box.

$$\Omega=\sum_{t=w,\mathrm{h}}(1-e^{-\omega_t})^{\theta},$$
$$\omega_w=\frac{|w-w^{gt}|}{max(w,w^{gt})},$$
$$\omega_{\mathrm{h}}=\frac{|\mathrm{h}-\mathrm{h}^{gt}|}{max(\mathrm{h},\mathrm{h}^{gt})},$$

where θ is the sensitivity controlling the shape cost, usually 4.

4 Results and analysis

4.1 Experimental settings

The experiment was conducted in a Windows 10 environment, and the specific configuration is presented in Table 2.

Table 2: Experimental configuration.1

Central processing unit (CPU)	Intel TM i5-11400F CPU
Graphics processing unit (GPU)	GeForce RTX1080TI GPU
Acceleration module	CUDA 11.1
Deep learning framework	PyTorch 1.7.0

For the underground drainage pipe defect damage detection algorithm, the parameters are set as Table 3.

Table 3: Parameter settings.

Epochs	300
Batch size	16
Image size	640 × 640
Optimizer	Stochastic gradient descent
Initial learning rate	0.001

The experiment used the Sewer-ML dataset [16], with images from actual drainage pipe inspection projects. Data distributions in the dataset are shown in Table 4.

Table 4: Data distributions in the Sewer-ML dataset.

	Training	**Validation**	**Test**	**Total**
Normal	552,820	68,681	69,221	690,722
Defective	487,309	61,365	60,805	609,479
Total	1,040,129	130,046	130,026	1,300,201

The Sewer-ML dataset included 18 different types of defect and damage, five of which were selected in this paper.

Sediment: Impurities and silt settle at the bottom of the drainage pipe to form sediment. As the volume expands, the area of water flow through the pipe decreases, weakening the pipe's ability to transport rainwater and sewage.

Crack: When a pipe breaks due to external forces such as compression, rainwater and sewage will seep out and pollute the water environment.

Cut: The cut at the connection of the branch pipe and the main pipe is uneven, with gaps appearing at the edge of the cut. This causes rainwater and sewage to seep out, polluting the groundwater. Soil also flows in, forming sediment.

Disconnection: The joints of the two ends of the pipe are not fully joined, causing the joints to shift and create a gap, which will allow rainwater and sewage to seep out and pollute the water environment.

Obstacle: They may be foreign matters carried in by rainwater or sewage, or other material that fall off and block in the pipe, resulting in a reduction in water flow area.

The five selected types exhibit a high degree of diversity in visual characteristics, covering various challenges ranging from slender small targets (cracks), complex texture targets (tree roots) to large-area irregular targets (sediments). This provides a testing benchmark for comprehensively evaluating the generalization and robustness of the model. Moreover, these five types of defects have sufficient and high-quality annotation data, ensuring the effectiveness of model training and the statistical reliability of the evaluation results. The images in the dataset were enhanced using operations such as translation, rotation, and cropping to obtain 3,000 images of each type. Image quality was improved through histogram equalization and sharpening. They were labeled using the LabelImg tool [20]. Moreover, ten-fold cross-validation method [21] was used to divide the dataset into a training set, a validation set, and a test set. For the detection effect of the algorithm, the IoU threshold value was set as 0.5, and samples below 0.5 were considered negative cases. The following evaluation indicators were used:

(1) precision: $Precision=\frac{TP}{TP+FP}$

(2) recall rate: $Recall=\frac{TP}{TP+FN}$

(3) average precision (AP): area under the precision-recall curve, $AP=\int_0^1 P(R)dR$,

(4) mAP: $mAP=\frac{\sum_{i=1}^{N} AP_i}{N}$,

(5) frames per second (FPS) [22]: the number of images detected per second, which is used to reflect the detection speed of a model.

In the above equations, *TP* is the number of positive samples detected as positive, *FP* is the number of negative samples detected as positive, and *FN* is the number of positive samples detected as negative.

4.2 Result analysis

The YOLOv5 algorithm was divided into different versions based on network width and depth. Experiments were conducted on different versions. The obtained mAP and FPS are shown in Table 5 and Figure 1.

Table 5: Comparison of different versions of YOLOv5.2

	mAP/%	FPS	Parameter quantity
YOLOv5n	78.12	93	1.90×10^6
YOLOv5s	81.77	90	7.20×10^6
YOLOv5m	83.94	82	2.12×10^7
YOLOv5l	85.51	81	4.65×10^7

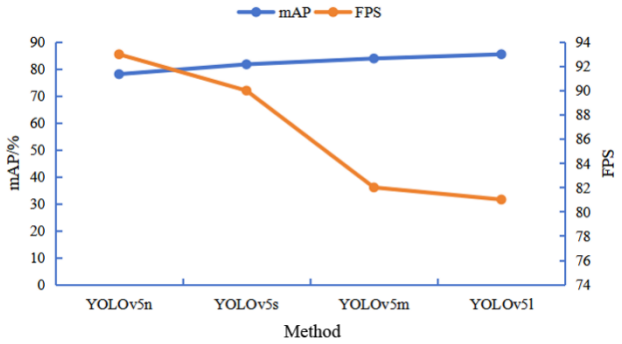


Figure 1: Comparison of different versions of YOLOv5.

As shown in Table 5 and Figure 1, with the expansion of model size, the mAP for the defect detection and damage localization also increased. The mAP of the YOLOv5n algorithm was 78.12%, and the mAP of the YOLOv5l algorithm was 85.51%, which was 7.39% higher than the YOLOv5n algorithm. However, the increase in scale also affected the detection speed; the larger the scale, the slower the detection speed. The comparison between YOLOv5m and YOLOv5l showed an increase of 1.57% in *mAP* and a decrease of one in FPS, indicating that the difference in detection speed between them was not significant. Therefore, in the subsequent experiments, the YOLOv5l version was used.

The effects of different attention mechanisms on detection results were compared (Table 6, Figure 2).

Table 6: Comparison of different attention mechanisms (Unit: %).3

	Baseline	Squeeze-and-Excitation (SE)	Coordinate Attention (CA)	CBAM
Sediment	85.33	87.64	88.46	90.88
Crack	71.21	73.36	74.41	81.21
Cut	94.56	95.16	96.77	98.97
Disconnection	95.19	96.93	96.14	97.43
Obstacle	81.26	83.56	85.57	87.56
mAP	85.51	87.33	88.27	91.21

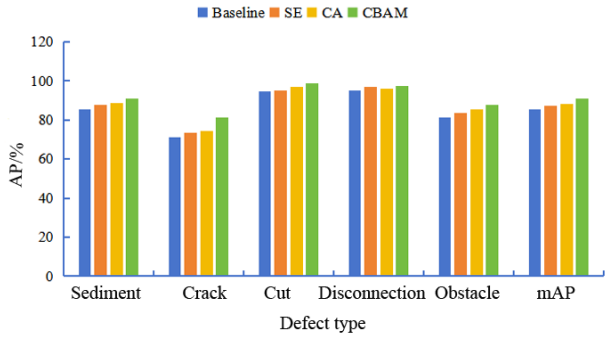


Figure 2: Comparison of different attention mechanisms.

From Table 6 and Figure 2, it can be seen that among different types of defect detection and damage localization, the detection of cuts and disconnections was more accurate. This may be because the characteristics of disconnections and cuts are more obvious and easier to identify. Cracks vary in size and direction and are easily confused with the background, making detection difficult. Similarly, sediments are located at the bottom of the pipe with blurred boundaries and are not easy to detect, and obstacles are also easily confused with sediment. The mAP was improved after the addition of different attention mechanisms compared to the baseline, and the CBAM had the best performance. SE only considered the importance of channel pixels and lacked attention to channel positions. CA performed limited on complex tasks. CBAM combined channel attention and spatial attention and achieved high accuracy in the detection of complex defects (cracks and obstacles), demonstrating its advantages.

The effects of different loss functions on the detection results were compared (Table 7, Figures 3 and 4).

Table 7: Comparison of different loss functions (unit: %).4

	Baseline	EIoU	Focal EIoU	SIoU
Sediment	85.33	86.16	86.89	87.03
Crack	71.21	72.64	72.23	73.07
Cut	94.56	96.17	96.68	97.03
Disconnection	95.19	96.55	96.91	97.28
Obstacle	81.26	82.33	82.94	83.19
mAP	85.51	86.77	87.13	87.52

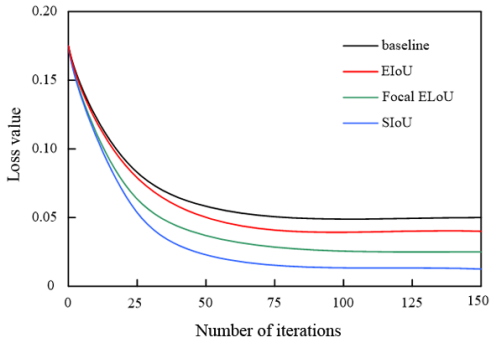


Figure 3: Loss value curve comparison chart.

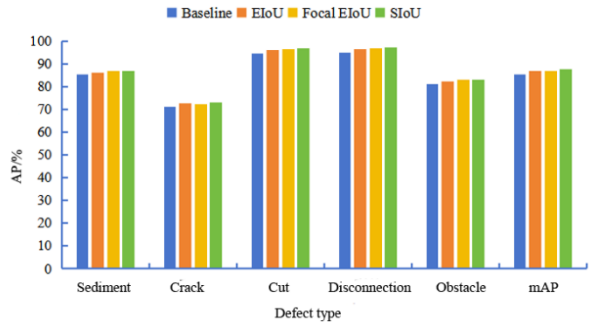


Figure 4: Comparison of different loss functions.

From Table 7 and Figures 3 and 4, it can be seen that after improving the loss function, the convergence of the network became faster, the detection performance was improved to a certain extent, but to a small extent. Among the three loss functions, SIoU exhibited the greatest improvement in detection performance and had the best convergence effect, and its mAP had an improvement of 2.01% compared to the baseline. Therefore, SIoU can be used instead of the original CIoU in YOLOv5 to achieve performance improvement.

The effect of the improvement on the detection results was determined by the ablation experiment (Table 8, Figure 5).

Table 8: Ablation experiments.5

	mAP/%	FPS
Baseline	85.51±2.16	81±1.21
YOLOv5+CBAM	91.21±3.33	79±0.77
YOLOv5+SIoU	87.52±2.77	83±1.45
YOLOv5+CBAM+SIoU	93.37±3.56*	85±1.17*

Note:* indicates $p < 0.05$ compared to the other method

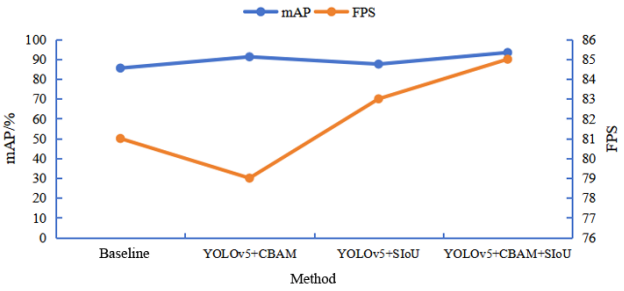


Figure 5: Ablation experiments.

According to Table 8 and Figure 5, the introduction of CBAM brought a 5.7% mAP improvement, and the FPS dropped from 81 to 79, indicating that the introduction of CBAM was beneficial to the improvement of detection accuracy, but it affected the detection speed to some extent. From this perspective, CBAM could effectively enhance the feature extraction ability, but it increased the computational load. The introduction of SIoU brought a 2.01% mAP improvement, and the FPS increased from 81 to 83, possibly because SIoU had a fast convergence speed. By introducing structural priors (angles and shapes), SIoU guided the model to converge to a state with greater "geometric regularity" during the training process, providing higher-quality predicted bounding box proposals, thus significantly reduced the computational

latency and achieved an improvement in throughput. The combined use of CBAM and SIoU achieved the best results, achieving a mAP of 93.37% and a FPS of 85. Moreover, the statistical significance results showed that compared the results of YOLOv5+CBAM+SIoU with the other methods, the p value was less than 0.05, demonstrating the reliability of the improvement to the YOLOv5 algorithm.

The method proposed was compared with some other deep learning-based detection methods (Table 9 and Figure 6).

Table 9: Comparison with other detection methods.6

	mAP/%	FPS
Single-Shot Multibox Detection (SSD) [23]	59.87	54
YOLOv3 [24]	75.59	68
YOLOv4 [25]	81.21	72
Faster regional-based convolutional neural network (R-CNN) [26]	84.93	38
The improved YOLOv5	93.37	85

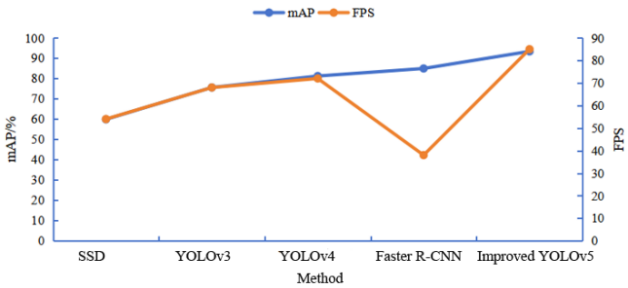


Figure 6: Comparison with other detection methods.

From Table 9 and Figure 6, it can be seen that the SSD performed poorly in drainage pipe defect detection and damage localization, with a mAP of 59.87% only and an FPS of 54. The feature fusion of SSD was relatively simple, while the proposed method realized the bidirectional fusion of deep and shallow features. When facing defects of different scales, the proposed method was more robust. Although the Faster R-CNN algorithm had high accuracy (mAP = 84.93), this came at the expense of speed. Its FPS was only 38. Faster R-CNN is a two-stage detector. Its serial process is computationally complex and the inference speed is slow, while YOLOv5 can achieve end-to-end fast inference. Compared with YOLOv3 and YOLOv4, the improved YOLOv4 algorithm proposed in this paper is advanced in terms of network architecture and data augmentation. The introduced CBAM and SIoU also bring significant performance improvements. Generally speaking, the proposed method was significantly superior to the other methods in terms of both precision and speed.

5 Discussion

In the detection of defects and damages in urban underground drainage pipes, this paper designed a YOLOv5 algorithm improved by combining the CBAM mechanism and SIoU, and verified its detection performance of using the sewer-ML dataset. The results showed that compared with other attention mechanisms or other loss functions, the selected CBAM and SIoU both had advantages. Comparisons with other deep learning detection methods showed that as a single-stage detector, SSD has insufficient ability to extract the diverse defect features of the pipes, resulting in many false detections and missed detections. As a two-stage detector, Faster R-CNN has many calculation steps and takes a long time, failing to meet the requirements of rapid response and efficient inspection. As early versions of the YOLO series, YOLOv3 and YOLOv4 are also inferior to YOLOv5 in terms of precision and speed. The combination of the CBAM mechanism with YOLOv5 can help the YOLOv5 network focus more accurately on the key features of the defects, providing the subsequent detection head with more informative and less noisy features. SIoU reconsiders the cost of bounding box regression, which helps the model generate prediction boxes with a higher degree of fit with the real boxes, bringing faster convergence and better convergence effects to the model. Under the synergistic effect of CBAM and SIoU, the precision and speed of the improved YOLOv5 algorithm for the detection of defects and damages in urban underground drainage pipes were improved.

Based on the research results, the designed algorithm can be applied in the detection of defects and damages in actual urban underground drainage pipes. For example, it can be applied in inspection robots to detect complex underground drainage pipes, process video streams in real time, and accurately locate and quantify defects, which makes it possible to formulate point-to-point repair plans and greatly saves the costs of excavation and repair. For the municipal system, the intelligent detection is conducive to the municipal department in formulating predictive maintenance plans. According to the detection results, priority maintenance can be carried out on high-risk pipe sections, thus avoiding accidents such as road collapse and urban waterlogging, and having a profound impact on enhancing urban safety and reducing operation and maintenance costs.

In future actions, experiments will be carried out on a more diverse dataset of underground drainage pipe defects and tested under more realistic conditions, such as occluded defects and noisy images. Meanwhile, further research will also be conducted on the deployment of the algorithm in the actual environment.

6 Conclusion

This paper designed a method based on the YOLOv5 algorithm for defect damage detection of urban underground drainage pipes. Through experiments on the dataset, it was found that the proposed method effectively

balanced accuracy and speed. The detection accuracy for different defect loss types was above 80%, mAP reached 93.37%, and the FPS was 85. It can be applied in actual urban sewage management to achieve better detection of defects and damages in underground drainage pipes.

References

[1] Huang R, Tao Z, Lin Y, Wei J, Zhou W, He Y (2024). Current Situation of Drainage Pipe Network in China and its Detection Technology: a Brief Review. *Polish Journal of Environmental Studies*, 33(1), pp. 19–29. https://doi.org/10.15244/pjoes/166899.

[2] Li Y, Wang H, Dang LM, Song HK, Moon HJ (2023). Attention-guided multiscale neural network for defect detection in sewer pipelines. *Computer-Aided Civil & Infrastructure Engineering*, 38(15), pp. 2163. https://doi.org/10.1111/mice.12991.

[3] Yougao L, Wei H (2020). Identification and feature extraction of drainage pipeline cracks based on SVD and edge recognition method. *2019 3rd International Conference on Electronic Information Technology and Computer Engineering (EITCE)*, pp. 1184–1188. https://doi.org/10.1109/EITCE47263.2019.9095153.

[4] Fu SY, Wei D, Zhou LY (2024). Helmet Detection Algorithm of Electric Bicycle Riders Based on YOLOv5 with CBAM Attention Mechanism Integration. *International Workshop on Advanced Computational Intelligence and Intelligent Informatics*, pp. 43–56. https://doi.org/10.1007/978-981-99-7593-8_5.

[5] Pang X, Zeng J, Yiyang C, Liu C, Fan X, Li S, Gong M (2025). Defect Detection Network for TOPCon Solar Cells Based on Improved YOLOv5 and CBAM Mechanism. *International Conference on Big Data*, pp. 75–87. https://doi.org/10.1007/978-3-031-77088-3_6.

[6] Lv M, Su WH (2023). YOLOV5-CBAM-C3TR: an optimized model based on transformer module and attention mechanism for apple leaf disease detection. *Frontiers in Plant Science*, 14, 1323301. https://doi.org/10.3389/fpls.2023.1323301.

[7] Xiao Y, Meng J, Yan H, Wang J, Xin K, Tao T (2024). Novel drainage pipeline breakages detection based on MEMS inertial sensor: From mechanism to application. *Journal of Cleaner Production*, 473(Oct.1), pp. 1–14. https://doi.org/10.1016/j.jclepro.2024.143546.

[8] Wang J, Wang C, Liu Y, Zhang X, Huang J (2025). Detection and effectiveness of improved drainage pipe defects–based semantic segmentation labeling methods. *Journal of Pipeline Systems Engineering and Practice*, 16(4). https://doi.org/10.1061/JPSEA2.PSENG-1898.

[9] He Z, Dang N, Zhou X, Li J. (2025). Automatic defect classification method of drainage pipe based on image processing and deep learning. *International Conference on Hydraulic and Civil Engineering & Engineering Safety and Disaster Prevention Forum*, pp. 1007–1013. https://doi.org/10.1007/978-3-031-90717-3_101.

[10] Huang QH, Li BA, Lv XQ, Zhang ZJ, Liu KH. (2020). Research on Pipeline Video Defect Detection Based on Improved Convolution Neural Network. *Journal of Physics: Conference Series*, 1576(1), 1–6. https://doi.org/10.1088/1742-6596/1576/1/012028.

[11] Budzan S, Szwedka M (2024). Integration of YOLO detection algorithm with trajectory prediction of pedestrians for advanced driver assistance system. *Przegląd Elektrotechniczny*, 100(12), pp. 53–58. https://doi.org/10.15199/48.2024.12.12.

[12] Zhou K, Zheng G, Zhai H, Lv X, Zhang W (2024). An Improved Distraction Behavior Detection Algorithm Based on YOLOv5. *Computers, Materials & Continua*, 81(11), pp. 2571–2585. https://doi.org/10.32604/cmc.2024.056863.

[13] Ha MH, Nguyen DC, Chan L Q, Chen OTC (2024). Emotional Inference from Speech Signals Informed by Multiple Stream DNNs Based Non-Local Attention Mechanism. *EAI Endorsed Transactions on Industrial Networks & Intelligent Systems*, 11(4), pp. 1. https://doi.org/10.4108/eetinis.v11i4.4734.

[14] Zhang K, Liu Y, Hu K (2025). RAW image reconstruction based on multi-scale attention mechanism. *Journal of Beijing University of Aeronautics and Astronautics*, 51(1), pp. 257–264. https://doi.org/10.13700/j.bh.1001-5965.2022.0959.

[15] Ye R, Qian Y, Huang X (2024). RT-CBAM: Refined Transformer Combined with Convolutional Block Attention Module for Underwater Image Restoration. *Sensors*, 24(18), pp. 17. https://doi.org/10.3390/s24185893.

[16] Yang Y, Hu S, Ke Y, Zhou R (2023). Flame smoke detection algorithm based on YOLOv5 in petrochemical plant. *International Journal of Intelligent Computing and Cybernetics*, 16(3), pp. 502-519. https://doi.org/10.1108/IJICC-11-2022-0291.

[17] Wang H, Liu Y, Luo H, Luo Y, Zhang Y, Long F, Li L (2024). Camellia oleifera trunks detection and identification based on improved YOLOv7. *Concurrency & Computation: Practice & Experience*, 36(27), pp. 1–14. https://doi.org/10.1002/cpe.8265.

[18] Li C, Chen D, Chen J, Dai H (2021). A Cross-Layer Fusion Multi-target Detection andRecognition Method Based onImproved FPN Model inComplex Traffic Environment. *CAAI International Conference on Artificial Intelligence*, pp. 323–334. https://doi.org/10.1007/978-3-030-93046-2_28.

[19] Haurum J B, Moeslund T B (2021). Sewer-ML: A Multi-Label Sewer Defect Classification Dataset and Benchmark. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA*, pp. 13451–13462. DOI:10.48550/arXiv.2103.10895.

[20] Deka B, Chakraborty D (2024). UAV Sensing-Based Litchi Segmentation Using Modified Mask-RCNN for Precision Agriculture. *IEEE Transactions on Agrifood Electronics*, 2(2), pp. 509–517. https://doi.org/10.1109/TAFE.2024.3420028.

[21] Tan JH, Bhandary SV, Sivaprasad S, Hagiwara Y, Bagchi A, Raghavendra U, Rao AK, Raju B, Shetty NS, Gertych A, Chua KC, Acharya UR (2018). Age-related Macular Degeneration detection using deepconvolutional neural network. *Future Generation Computer Systems*, 87(OCT.), pp. 127–135. https://doi.org/10.1016/j.future.2018.05.001.

[22] Ruiz-Beltran CA, Romero-Garces A, Gonzalez M, Sanchez Pedraza A, Rodriguez-Fernandez JA, Bandera A (2022). Real-time embedded eye detection system. *Expert Systems with Applications*, 194, pp. 1–10. https://doi.org/10.1016/j.eswa.2022.116505.

[23] Sarma K S R K, Sasikala C, Surendra K, Erukala S, Aruna SL (2024). A comparative study on faster R-CNN, YOLO and SSD object detection algorithms on HIDS system. *AIP Conference Proceedings*, 2971(1), pp. 13. https://doi.org/10.1063/5.0195857.

[24] Apoorva, Babu S, Rani A, Srivastva A, Reddy A (2024). On stream face mask and helmet detection system using YoloV3 algorithm. *AIP Conference Proceedings*, 2742(1), pp. 8. https://doi.org/10.1063/5.0191204.

[25] He J, Wang W, Yang N (2024). Research on Track Fastener Service Status Detection Based on Improved Yolov4 Model. *Journal of Transportation Technologies*, 14(2), pp. 212-223. https://doi.org/10.4236/jtts.2024.142013.

[26] Wang Q, Wang Z, Li J, Yang Y, Liu P (2020). Improvement of Faster-RCNN Detection Algorithms for Small Size Line Accessory Equipment. *Journal of Physics: Conference Series*, 1453, pp. 1–6. https://doi.org/10.1088/1742-6596/1453/1/012007.

A Deep Multimodal Retrieval Framework for Digital Libraries Using SE-ResNet-FCN, BERT, and Enhanced Hash Learning

Cheng Qiu
Zhejiang Library, Hangzhou 310008, China
E-mail: qiucheng7606@163.com

Keywords: fully convolutional network, hash learning, bidirectional transformer encoder representation, multimodal retrieval, attention mechanism

Received: October 10, 2025

The current multimodal retrieval accuracy of online libraries is insufficient. To solve this problem, a multimodal retrieval model for digital libraries that integrates fully convolutional networks and hash learning is proposed in the proposed method. The research introduces a fully convolutional network and a bidirectional Transformer encoder to extract semantic features, and combines a residual neural network to deeply optimize the model, thereby enhancing the feature expression ability. In the hash learning stage, a triplet loss and contrastive learning loss optimization model is designed to further enhance cross-modal semantic alignment. This enables image-text multimodal retrieval. In the experiment, the model used was applied to the BookCoverDataset for verification. And it is combined with Latent semantic sparse hashing (LSSH) and Collective Matrix Factorization Hashing. CMFH, Supervised Matrix Factorization Hashing (SMFH), Discrete online cross-modal hashing the multimodal retrieval models of digital libraries constructed by DOCH were compared. The experimental results show that the average retrieval accuracy score of this model is up to 0.93, and the maximum mAP reaches 0.95, which is 0.20 higher than that of the comparison model. When the hash code is 256 bits, its average accuracy reaches 0.94. Compared with the baseline model, the study proposes that the model demonstrates stronger semantic association ability and feature compression efficiency in multimodal retrieval tasks, verifying the effectiveness of the fusion strategy of fully convolutional networks and hash learning. The model significantly enhances the accuracy and robustness of cross-modal retrieval through a deep semantic alignment mechanism, providing a feasible solution for efficient and precise image-text mutual inspection in the digital library environment.

Povzetek: Študija predstavi model za boljše iskanje med slikami in besedilom v digitalnih knjižnicah, ki z združevanjem globokega učenja in "hash" kodiranja izboljša natančnost ter prekaša primerjalne metode.

1 Introduction

In the wave of digitalization, online libraries become the core platform for knowledge dissemination and sharing, with book resources presenting multimodal forms such as images and text [1]. Multimodal retrieval links heterogeneous data to meet users' precise demands of "searching text by image and searching image by text" and is crucial for improving digital resource utilization [2]. However, traditional retrieval methods focus on single-modal features and ignore cross-modal semantic associations. When handling large-scale multimodal data, they face high feature dimensionality, low retrieval efficiency, and difficulty in semantic alignment, making them unsuitable for the complex needs of online libraries [3]. Currently, multimodal retrieval research has made certain progress. Scholars attempt to link multimodal data through feature concatenation and early fusion, but they do not fully explore deep semantics between modalities. Existing models for online library multimodal resources often suffer from large semantic gaps and long retrieval time, failing to balance efficiency and accuracy [4-5]. The existing methods still have significant deficiencies in the collaborative optimization of cross-modal semantic alignment and feature compression, especially lacking the joint modeling ability of fine-grained semantic matching and efficient hash coding in the online library scenario. Moreover, most models fail to take into account the local feature alignment and global semantic consistency of image and text modalities, resulting in limited accuracy of retrieval results. To this end, a multimodal hash framework based on a fully convolutional network and a bidirectional Transformer encoder is proposed. A cross-modal attention mechanism is introduced to achieve fine-grained alignment of image regions and text primitives. The hash encoding process is optimized by combining semantic preservation loss and quantization constraints, significantly improving computational efficiency while ensuring retrieval accuracy.

The significance of the research lies in its supporting role in the efficient organization and precise access of multimodal resources in online libraries, which is directly related to the efficiency of users' information acquisition and service quality. The criteria for measuring success include a significantly better cross-modal retrieval accuracy than existing methods and an increase of no less than 5% in mAP values on mainstream evaluation

datasets. When the hash encoding compression ratio is below 64 bits, it still maintains a relatively high retrieval accuracy. The response time of the model meets the real-time requirements, and the time consumption of a single query is controlled within 200 milliseconds. It has good generalization ability and stability in the real online library scenario.

The research contribution lies in (1) proposing a cross-modal hash learning framework that integrates a fully convolutional network with a bidirectional Transformer encoder to achieve deep semantic alignment of image and text modalities at a fine-grained level; (2) Design a cross-modal attention fusion mechanism and a hybrid loss function, and collaboratively optimize the semantic preservation and hash quantization processes to enhance the distinguishability of multimodal features in a compact binary space; (3) Verify the model's effectiveness in both public datasets and real online library scenarios, balancing high-precision retrieval and low-latency response, and provide feasible solutions for the efficient organization of multimodal resources.

2 Literature review

Multimodal retrieval is a technique that uses a query from one modality to find semantically related data in other modalities, such as text, images, or videos. This technique has been widely applied in many fields, and many scholars have conducted related studies. Deep cross-modal hashing methods have been developed to learn fine-grained similarities between data points of different modalities, thereby improving retrieval accuracy. Contrastive learning frameworks, as an effective technique for enhancing the consistency of visual and text features, have been widely applied [6]. Knowledge distillation technology is also used to transfer knowledge from large teacher models to more effective student hash models, thereby reducing complexity while improving performance [7]. To address the challenge of efficiency, asymmetric hashing methods were proposed, with learning tasks defined within the framework of matrix factorization to generate compact and discriminative hash codes [8]. All these technologies have been applied in multimodal retrieval, and the research results have provided new improvement ideas for this field.

Methods such as FCN, hash learning, and image segmentation are also frequently employed in multimodal retrieval. With the advancement of science and technology, an increasing number of scholars are seeking more effective retrieval methods [9-10]. Zhang D et al. proposed a new online hashing method in order to effectively handle online streaming media data. This method utilizes semantic autoencoders to establish the correlation between binary codes and labels, and adopts the inner product of labels to achieve the connection between data and new data. This enhances the efficiency of large-scale cross-media similarity retrieval [11]. Khan A et al. proposed a cross-modal recovery technology based on a multi-label information depth ranking model to solve the problem of inappropriate information contained between images and texts in cross-modal retrieval. This technology uses a regularization function instead of binary constraints to confine discrete values within a numerical range for end-to-end training. The results show that the performance of this method on the MIR-Flickr-25K and NUS-WIDE datasets is significantly better than that of the existing mainstream models [12]. Wang Y et al. proposed a new supervised cross-modal hashing method, namely multi-information embedding hashing, in response to the problem that the retrieval performance of supervised cross-modal hashing is gradually reaching a bottleneck at present. Multi-information embedded hashing can flexibly handle various information mining, hash code learning and hash function learning, thus improving the retrieval performance of supervised cross-modal hashing [13]. Wen H et al. proposed a self-trained enhanced multi-factor matching network, which models the fine-grained alignment relationship between text and images by decoupling latent semantic factors and introducing a double aggregation mechanism. The results show that this method achieves significant performance improvements on multiple combined image retrieval benchmarks, with an improvement rate exceeding 8%[14]. To further compare the performance differences between the research method and the existing mainstream methods, the differences between each literature work and the research work, as well as their own advantages, are summarized in Table 1.

Table 1: Summary of the differences between the related work and the work of SRF-BERT-IDHS.

Document Number	Method	The difference from SRF-BERT-IDHS
[6]	A retrieval method combining deep neural model learning modalities and WOA operators	Deep semantic alignment has not been achieved. Only modal association is carried out in the shallow feature space, lacking cross-modal semantic consistency constraints.
[7]	A multimodal contrastive knowledge distillation method	The model has a high degree of redundancy and low computational efficiency
[8]	A novel asymmetric supervised fusion-oriented hashing method named ASFOG	Lack of dynamic adaptability
[11]	Cross-modal retrieval framework based on online hashing and semantic autoencoding	The semantic relationships between classes were not fully explored, resulting in limited generalization ability
[12]	Cross-modal recovery technology based on multi-label information deep ranking model	The semantic correlation is not high and the retrieval accuracy is low
[13]	A new supervised cross-modal hashing method	Insufficient modeling of fine-grained semantic associations between modalities
[14]	Self-trained enhanced multi-factor matching network	Relying on strongly supervised signals makes it difficult to address the challenges of label noise and modal heterogeneity

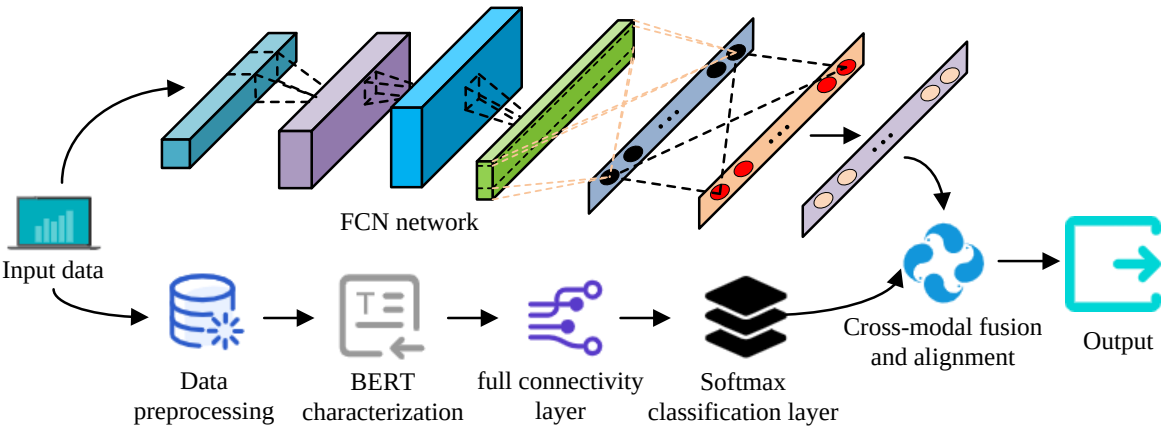


Figure 1: Modal data processing module composed of FCN and BERT.

Based on the above content, it can be known that although there has been certain progress in the related research of multimodal retrieval in digital libraries, due to the significant structural differences between the visual features of images and the language features of texts, deep semantic alignment has not been achieved, resulting in an imbalance between retrieval efficiency and accuracy. Moreover, most of the existing methods are limited to shallow semantic matching and it is difficult to capture the complex associations between modalities. Therefore, the proposed multimodal retrieval model for digital libraries, which integrates FCN and Hash Learning, replaces the original FCN network with Residual Network (ResNet) and combines a Squeeze-Excitation (SE) channel attention mechanism to enhance feature extraction. Using a dual-branch deep hashing network structure and an improved loss function, it aims to meet users' needs for accurate retrieval in digital libraries.

3 Construction of multimodal retrieval model for online libraries

3.1 Construction of modality feature extraction module based on FCN and BERT

Multimodal retrieval in online libraries aims to achieve precise content matching across or within modalities, such as "text→image," "image→text," and "image→image." To achieve this goal, it is necessary to address the significant differences in underlying representations among different modalities. In addition, features of different modalities need to be semantically aligned, and retrieval must balance efficiency and accuracy [15]. Therefore, SRF-BERT-IDHS proposes an online library multimodal retrieval model based on FCN and Hash Learning. The model uses FCN to process and analyze image modality data, employs Bidirectional Encoder Representations from Transformers (BERT) to capture textual features, and uses Hash Learning to address retrieval efficiency and large-scale data matching. The structure of the modality data processing module composed of FCN and BERT is shown in Figure 1.

As shown in Figure 1, after receiving multimodal raw data, images enter the FCN for encoding and decoding to achieve spatial feature alignment. Text enters BERT, which outputs context-aware word vectors through the Transformer encoder. Finally, the model performs cross-modal fusion and alignment of text and images, realizes cross-modal interaction, and outputs the results. In FCN, the convolution operation is the foundation for extracting image features. Its mathematical expression is shown in Equation (1) [16].

$$Y(i,j,c)=\sum_{p=0}^{k-1}\sum_{q=0}^{k-1}\sum_{c'=0}^{C_{in}-1}X(i+p,j+q,c')\cdot K(p,q,c',c)+b(c) \quad (1)$$

In Equation (1), $Y(i,j,c)$ denotes the pixel value of the output feature map. $X(i+p,j+q,c')$ is the pixel value of the input feature map at the corresponding position and channel. $K(p,q,c',c)$ is the weight of the convolution kernel at dimension (p,q,c',c). FCN reduces feature map resolution through pooling layers to enlarge the receptive field and reduce parameters. Its equation is shown in Equation (2).

$$Y(i,j,c)=\max_{p=0}^{s-1}\max_{q=0}^{s-1}X(i\cdot s+p,j\cdot s+q,c) \quad (2)$$

In Equation (2), s denotes the pooling window size, and $Y(i,j,c)$ is the maximum value. The core innovation of FCN is upsampling low-resolution feature maps to the input image size through deconvolution to achieve pixel-level prediction. The deconvolution operation is shown in Equation (3).

$$Y'(i,j,c)=\sum_{p=0}^{k'-1}\sum_{q=0}^{k'-1}\sum_{c'=0}^{C_{in}-1}X(i',j',c')\cdot K'(p,q,c,c')+b'(c) \quad (3)$$

In Equation (3), $i=i'\cdot s'-p+(k'-s')$, $j=j'\cdot s'-q+(k'-s')$, and $Y'\in R^{H\times W\times C_{out}}$ denote the high-resolution feature map after upsampling. The input representation in BERT is shown in Equation (4).

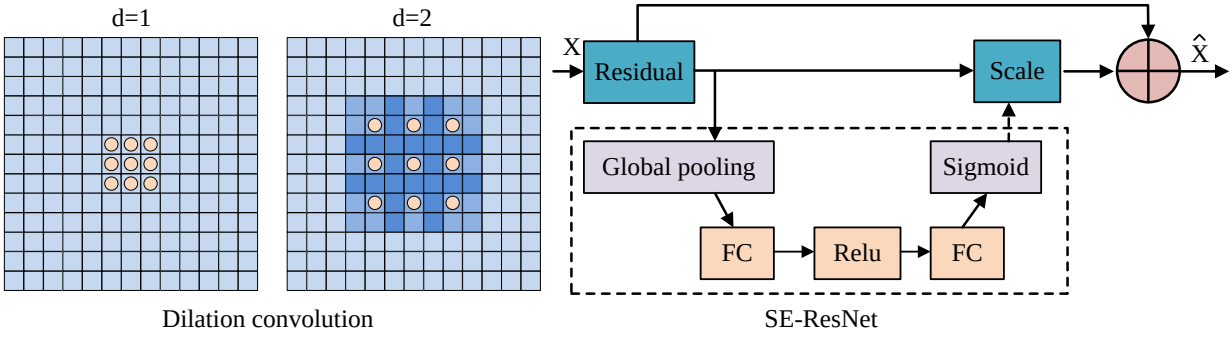


Figure 2: Diagram of dilated convolution and SE-ResNet structure.

$$E_i = T_i + P_i + S_i \tag{4}$$

In Equation (4), E_i denotes the final input vector of the i -th token. T_i is the word embedding, P_i is the position embedding, and S_i is the segment embedding. Using only FCN-BERT for online library multimodal retrieval can handle cross-modal information processing and matching to some extent. However, FCN has limited ability to capture small targets and fine details, lacks explicit modeling of global image semantics and spatial relationships, and consumes significant computational resources and time. Therefore, The proposed method replaces the VGG network in FCN with ResNet. The residual connections in ResNet solve the gradient vanishing problem in deep networks, support deeper network structures, and retain richer multi-scale features. Dilated convolution expands the receptive field without decreasing resolution, enabling the capture of the global layout of book covers. Attention mechanisms are applied by adding SE channel attention after each residual block and attention in BERT to focus on key information [17]. The schematic diagram of dilated convolution and SE-ResNet structure is shown in Figure 2.

As shown in Figure 2, the left diagram depicts dilated convolution, enlarging the receptive field while keeping the number of parameters unchanged. When d=1, the receptive field is 3×3. After dilation, it expands to 5×5, reducing overall computation. The right diagram shows the SE-ResNet structure, where the SE module is added to the residual structure, allowing the network to automatically focus on important feature channels and suppress irrelevant background. The output of dilated convolution is expressed in Equation (5).

$$y[i] = \sum_{k-1}^{K} x[i + d \cdot k] \cdot \omega[k] \tag{5}$$

In Equation (5), d denotes the dilation rate, k is the convolution kernel size, x represents the input feature map, and w indicates the convolution kernel weight. In the SE module, the squeezing equation of channel attention is shown in Equation (6).

$$Z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_{i,j,c} \tag{6}$$

In Equation (6), H and W respectively represent the height and width of the feature map, C is the number of channels, and Z_c is the channel statistical information obtained through global average pooling of spatial dimensions, which is used for subsequent excitation operations to generate channel weights and achieve enhancement of important features and suppression of redundant information. The study performs global average pooling on the feature map $x_{i,j,c}$ of each channel, compressing it into a 1×1×C vector to focus on the global information of the channel. The excitation equation of SE is shown in Equation (7).

$$s_c = \sigma\left(W_2 \cdot \mathrm{Re}\,LU\left(W_1 \cdot z_c\right)\right) \tag{7}$$

In Equation (7), W_1 and W_2 are the weight matrices for dimension reduction and dimension increase respectively, and σ represents the Sigmoid function, which outputs the weights from 0 to 1, representing the importance of this channel. Therefore, the modality data processing module composed of SE-ResNet-FCN and BERT after these improvements is shown in Figure 3.

As shown in Figure 3, the module preprocesses raw data for images and text. SE-ResNet optimizes image feature extraction in FCN, allowing the network to focus on key regions. Dilated convolution is applied in certain convolutional layers to capture global semantic information. The decoder then upsamples and fuses features, transforming low-resolution feature maps into high-resolution representations. BERT extracts semantic features from text and captures contextual dependencies. Finally, cross-modal fusion and alignment are performed to eliminate the modality gap between images and text, and the aligned features are output.

3.2 Multimodal retrieval model combining improved FCN and enhanced hash learning

Through SE-ResNet-FCN and BERT, the model extracts semantic features from images and texts, but this process alone does not complete cross-modal fusion. Hash learning addresses the efficiency and semantic association issues in multimodal retrieval by converting high-dimensional features of different modalities into low-dimensional binary hash codes. Because traditional Hash learning relies on manually designed features and is difficult to fully mine the complex semantic information of data, to solve this problem, the research adopts a dual-branch hash network structure. One branch is for images

and uses the image DH based on SE-ResNet-FCN, and the other branch is for text and uses the text DH based on BERT. Then, through the fusion layer, the features of different modalities are integrated and mapped to a unified hash code [18]. The dual-branch Hash network structure is shown in Figure 4.

As shown in Figure 4, the dual-branch Hash network maps semantically similar data from different modalities to nearby hash codes in the hash space, while pushing dissimilar data farther apart. After extracting features from the two branches and aligning their dimensions, the model fuses them through the fusion layer, then maps the unified feature to a binary hash code. A loss function constrains network training to ensure the semantic consistency of hash codes. The hash function mapping in Hash learning is shown in Equation (8)

$$h(x) = sign(f(x)) \tag{8}$$

In Equation (8), x represents the original high-dimensional feature, $f(x)$ is the hash mapping function, $sign(\cdot)$ is the sign function, and $h(x)$ represents the generated binary hash code. The mapping function in deep hashing is shown in Equation (9).

$$f(x) = W^T \cdot \phi(x) + b \tag{9}$$

In Equation (9), $\phi(x)$ represents the features extracted by the neural network, W and b are the learnable weight matrix and bias term. Hamming distance is used to measure sample similarity in the hash space. The equation for Hamming distance is shown in Equation (10).

$$d_H(h_a, h_b) = \frac{\frac{1}{k}\sum_{i=1}^{k}\left|h_a^{(i)} - h_b^{(i)}\right|}{2} \tag{10}$$

In Equation (10), h_a and h_b are two hash codes with length k, and $h_a^{(i)}$ and $h_b^{(i)}$ are the i-th bit of the hash code. Hamming similarity measures the similarity between two hash codes, as shown in Equation (11).

$$s_H(h_a, h_b) = 1 - d_H(h_a, h_b) \tag{11}$$

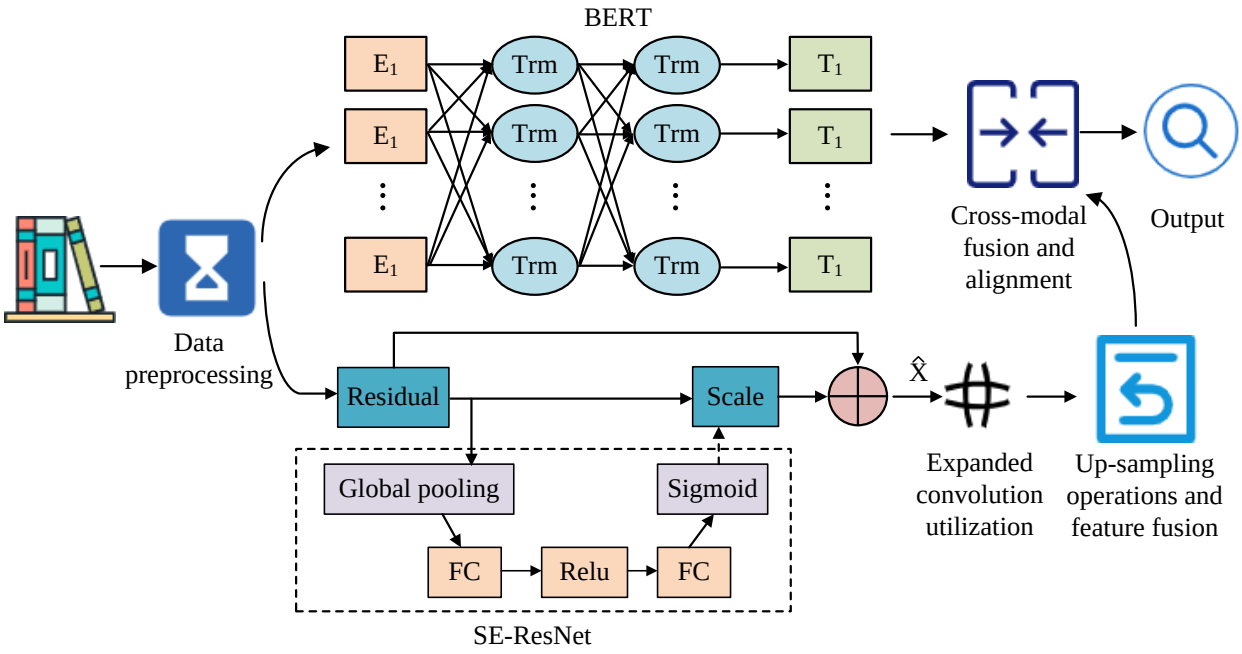


Figure 3: Modal data processing module composed of SE-ResNet-FCN and BERT.

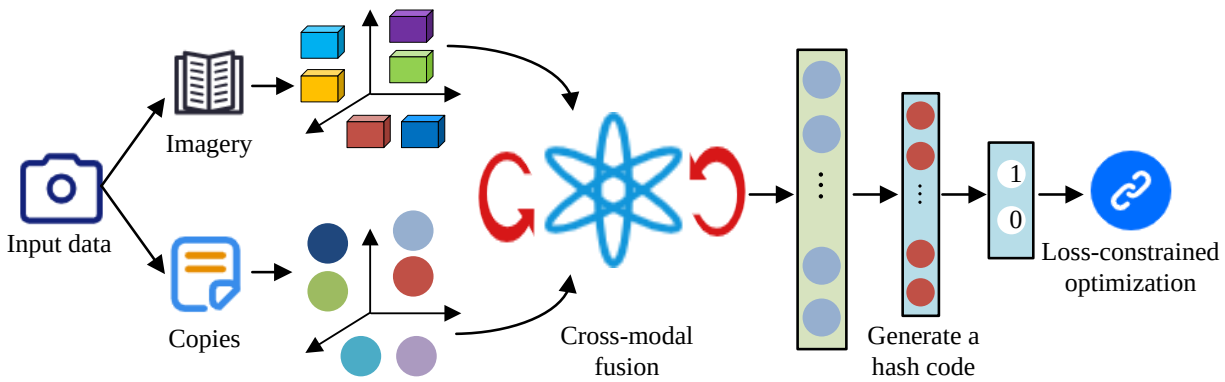


Figure 4: Double-branch Hash network structure diagram.

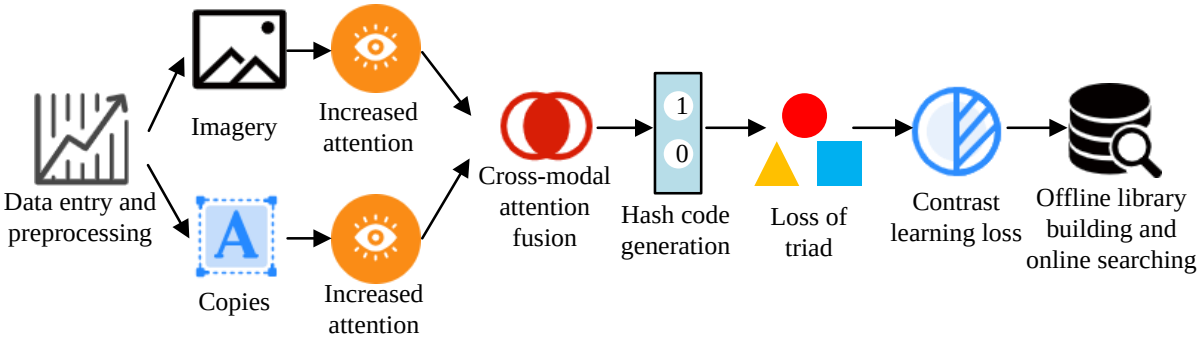


Figure 5: Improved dual-branch hash network structure diagram.

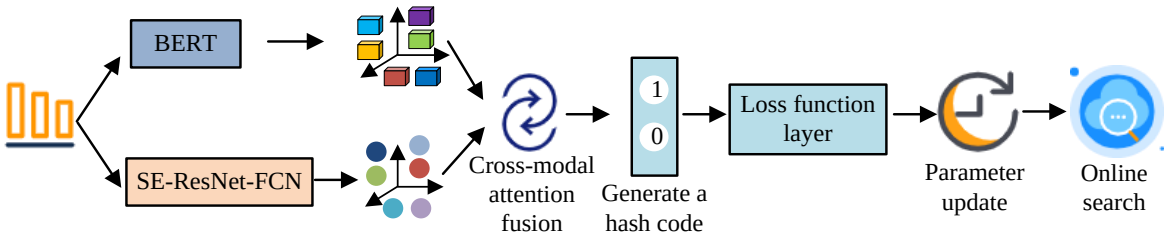


Figure 6: SRF-BERT-IDHS online library multimodal retrieval model.

In Equation (11), $s_H(h_a, h_b)$ represents the Hamming similarity, $d_H(h_a, h_b)$ is the length of the hash code, and the value range is [0,1]. Hamming similarity is complementary to Hamming distance. The dual-branch Hash network achieves cross-modal fusion to some extent, but the generated hash codes may only capture shallow patterns within each modality instead of deep semantic correlations. Triplet loss is introduced to enhance similarity constraints, reducing the distance between anchor and positive samples and increasing the distance to negative samples [19]. In addition, to improve the quality of hash codes, contrastive learning loss is introduced. By constructing positive and negative sample pairs, the similarity of positive sample pairs in the hash space is maximized, and the similarity of negative sample pairs is minimized. The structure diagram of the improved dual-branch hash network is shown in Figure 5.

As shown in Figure 5, the improved dual-branch Hash network focuses on hash code generation. Contrastive loss pulls the hash codes of similar samples closer, while triplet loss enforces the anchor-positive distance to be smaller than the anchor-negative distance. Through loss backpropagation, the network parameters are optimized so that the hash codes of semantically similar samples become closer, improving retrieval accuracy. The core equation of triplet loss is shown in Equation (12).

$$L_{triplet} = \max\left(d(h_a, h_p) - d(h_a, h_n) + \alpha, 0\right) \quad (12)$$

In Equation (12), $L_{triplet}$ is the triple loss function. h_a is the anchor sample hash code, h_p is the positive sample hash code with the same semantics, h_n is the negative sample hash code with different semantics, $d(\cdot)$ is the Hamming distance, and α is the margin parameter. The selection of marginal parameters for triplet loss is based on the distribution characteristics of positive and negative sample pairs during the training process, with a parameter adjustment range of 0.1 to 2.0. The contrastive learning loss is shown in Equation (13).

$$L_{contrast} = -\log \frac{\exp\left(s(h_v, h_t)/\tau\right)}{\sum_{k=1}^{N} \exp\left(s(h_v, h_{t,k})/\tau\right)} \quad (13)$$

In Equation (13), $L_{contrast}$ is the contrastive learning loss function. h_v represents the image hash code, h_t represents the matching text hash code, $h_{t,k}$ represents another text hash code, $s(\cdot)$ is the similarity function, and τ is the temperature parameter. In summary, the structure of the online library multimodal retrieval model (SRF-BERT-IDHS) composed of SRF-BERT and Improved Dual-branch Hash Structure (IDHS) modules is shown in Figure 6.

As shown in Figure 6, the unique novelty of the SRF-BERT-IDHS framework proposed in the study and its differences from previous multimodal hashing methods are as follows: Compared with the traditional dual-branch network structure, SRF-BERT-IDHS achieves dynamic alignment of image and text features at the hash mapping layer by introducing a cross-modal attention mechanism, effectively enhancing semantic consistency. In addition, by combining the joint optimization strategy of triple loss and contrastive learning, the compactness and discriminative ability of hash codes between heterogeneous modalities are further enhanced, overcoming the limitations of existing methods in semantic gaps and modal differences, thereby achieving efficient and accurate multimodal retrieval in large-scale online library scenarios.

Table 2: mAP test results at different hash code length levels.

Task	Model	Hash code length				
		16	32	64	128	256
I→T	SRF-BERT-IDHS	0.821	0.833	0.847	0.858	0.864
	LSSH	0.571	0.579	0.587	0.596	0.604
	CMFH	0.623	0.632	0.648	0.653	0.662
	SMFH	0.732	0.746	0.753	0.766	0.780
	DOCH	0.677	0.686	0.701	0.709	0.715
T→I	SRF-BERT-IDHS	0.872	0.889	0.905	0.916	0.931
	LSSH	0.678	0.698	0.705	0.712	0.726
	CMFH	0.712	0.723	0.734	0.745	0.751
	SMFH	0.783	0.798	0.810	0.823	0.831
	DOCH	0.724	0.735	0.747	0.760	0.775

The research proposes that the process of the model in real-time application or operation in large-scale library systems is divided into five stages: data upload, semantic feature extraction, hash coding, index storage and retrieval feedback. After the user inputs text or uploads an image, the system first calls the pre-trained SRF-BERT-IDHS model to extract the semantic features of the corresponding modal and generate a fixed-length hash code. Subsequently, the Hamming distance is calculated in the hash index database to quickly retrieve the closest cross-modal data. While returning the results, the system records the user's click behavior, duration of stay and feedback score as implicit feedback signals. Periodically, these feedback data are used to fine-tune the attention weights and hash mapping parameters of the model, optimizing the retrieval accuracy through an incremental learning mechanism. The learning rate is set to 0.001, the batch size is 64, and the Adam optimizer is used for parameter updates. The training lasted for 50 eras. The positive and negative samples in the triplet loss were sampled using the hard example mining strategy to ensure that the model focused on the discrimination near the semantic boundary during the optimization process. The training set and the test set are divided in an 8:2 ratio, and all data are randomly shuffled to ensure a balanced distribution of categories. The text encoder adopts the pre-trained BERT-base model, whose parameters are fixed in the early stage of training and then jointly updated at a smaller learning rate in the fine-tuning stage. The image backbone network is initialized based on the pre-trained SE-ResNet-FCN of ImageNet. The feature extraction process integrates the channel and spatial attention mechanism to enhance the response strength of key regions.

4 Validation of SRF-BERT-IDHS multimodal retrieval model for online libraries

4.1 Performance validation of SRF-BERT-IDHS model

To evaluate the multimodal retrieval performance of the SRF-BERT-IDHS model, it was compared with multimodal retrieval models for digital libraries based on Latent Semantic Sparse Hashing (LSSH), Collective Matrix Factorization Hashing (CMFH), Supervised Matrix Factorization Hashing (SMFH), and Discrete Online Cross-modal Hashing (DOCH). The experimental datasets were BookCover and ICDAR, both containing book image modalities and associated text information. The experiments were conducted on high-performance computing equipment with an Intel Core i9-10900K CPU (2.4 GHz, 32 cores) and 128 GB of memory. The software environment included Ubuntu 20.04 LTS and Python 3.7. The mean average precision (mAP) scores of the five models were compared under different hash code lengths, including image-to-text retrieval (I→T) and text-to-image retrieval (T→I) tasks. The performance comparison using the BookCover dataset is shown in Table 2.

In Table 2, the mAP value of SRF-BERT-IDHS in text-to-image retrieval reached a maximum of 0.931 and remained no lower than 0.872. SMFH ranked second, with a maximum mAP score of 0.831 and relatively stable performance. The mAP scores of all models in the text-to-image task were generally higher than those in the image-to-text task, indicating that text features enabled more precise image retrieval. Moreover, the hash code length was positively correlated with performance, suggesting that longer hash codes preserved more semantic information. To sum up, the SRF-BERT-IDHS model has the best performance. In all hash codes and two tasks, the mAP score of SRF-BERT-IDHS is higher than that of other models, and the advantage further expands with the increase of hash codes. To further verify the performance of SRF-BERT-IDHS, the five models were tested on both BookCover and ICDAR datasets, and their training efficiency was compared, as shown in Figure 7.

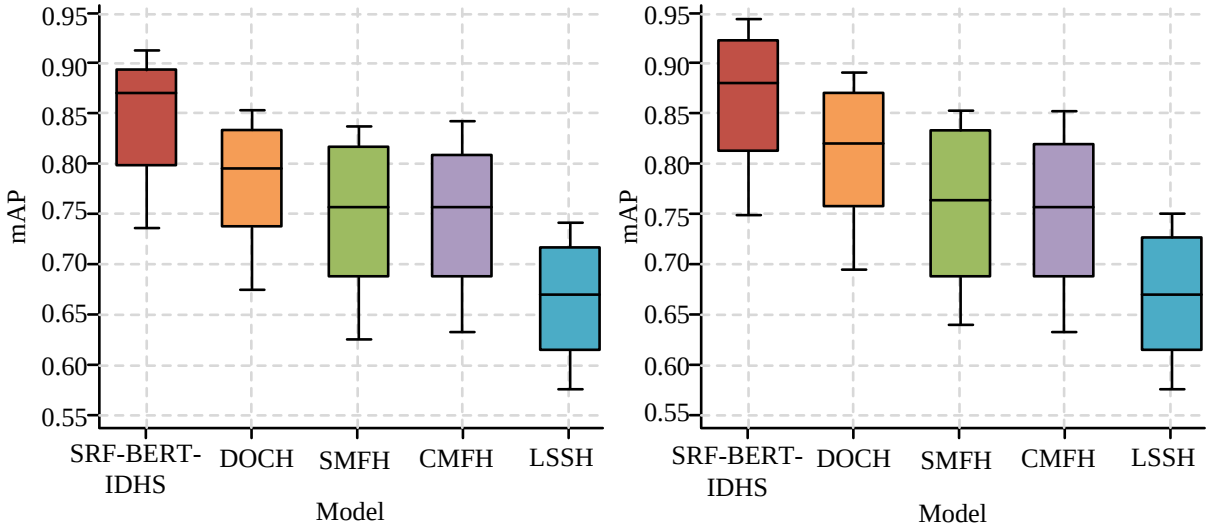
It can be seen from Figure 7 (a), the highest mAP score of SRF-BERT-IDHS was 0.92, the median was 0.87, and the lowest was 0.74. The lowest mAP score of LSSH was 0.58. As shown in Figure 7(b), the mAP scores of all models improved, with SRF-BERT-IDHS achieving a median mAP of 0.88 and a maximum of 0.95, which was 0.20 higher than the maximum of LSSH. Overall, SRF-BERT-IDHS demonstrated the highest precision and stability, ranking first among all compared models. This indicates that SRF-BERT-IDHS has strong generalization ability and robustness under different data distributions. Compared with other models, it can better adapt to the feature distribution of different datasets. Even in scenarios with complex semantics or significant cross-modal differences, it can still maintain high retrieval accuracy.

To comprehensively assess the learning capability of SRF-BERT-IDHS, the five models were trained on different datasets, and the comparison is shown in Figure 8.

As can be seen from Figure 8, all models exhibited slower loss reduction and higher overall loss on the BookCover dataset than on the ICDAR dataset, indicating that the relationship between book covers and text made the learning process more challenging. As shown in Figure 8(a), the loss of SRF-BERT-IDHS dropped sharply from $5.50E^{-09}$ to below $1.00E^{-09}$ within 20 epochs and was much lower than that of other models. The LSSH model consistently performed worst, with the highest loss reaching $2.00E^{-09}$. Compared with the ICDAR dataset, BookCover has higher requirements for the semantic alignment ability of the model, while SRF-BERT-IDHS can still converge rapidly on this dataset, indicating that it has stronger feature extraction and cross-modal matching capabilities. In contrast, the convergence speed of other models is relatively slow and they are prone to fall into local optimum, verifying the learning advantages of SRF-BERT-IDHS in complex scenarios.

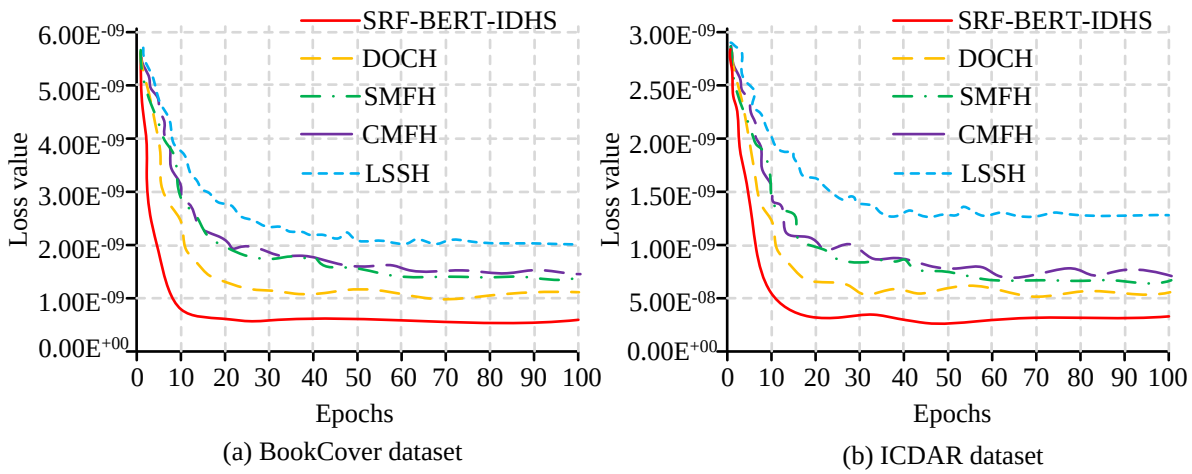
4.2 Practical application of multimodal retrieval model in online libraries

After verifying the basic performance of SRF-BERT-IDHS, its practical application value was further tested. A dataset of 5,000 pairs of book images (including cover and illustrations) and text information (including metadata, tables of contents, and main content) was collected from online digital libraries. The higher-quality 70% of the data was used as the training set, and the remaining 30% was used as the test set. The experimental environment remained unchanged. The five models were trained on the dataset, and their Average Precision (AP) was compared under different hash code lengths. The results are shown in Figure 9.



(a) BookCover dataset (b) ICDAR dataset

Figure 7: Training efficiency test results in different data sets.



(a) BookCover dataset (b) ICDAR dataset

Figure 8: Comparison of learning capabilities of five models.

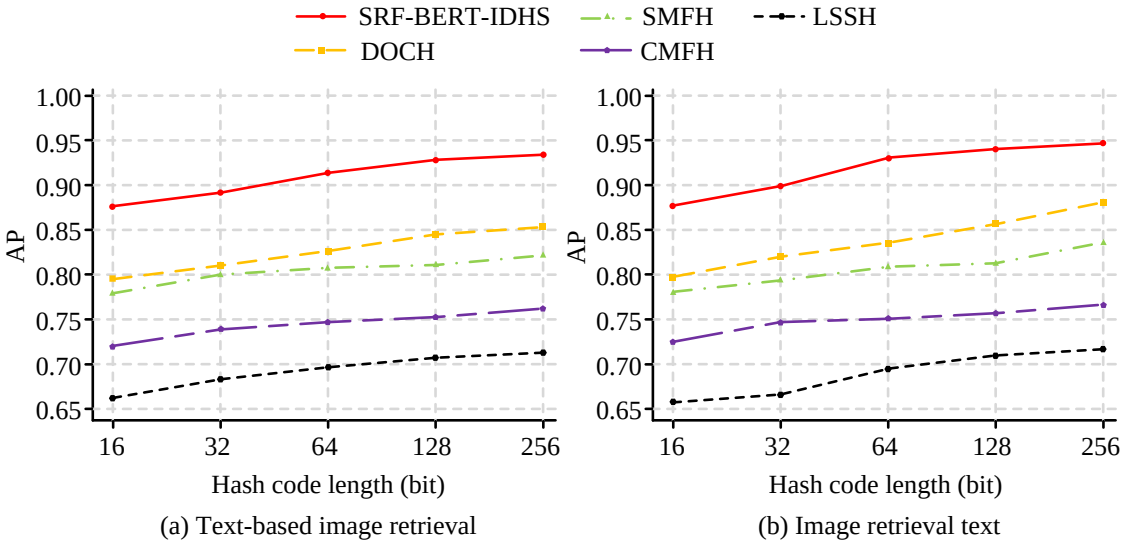


Figure 9: Comparison of AP results in different hash codes.

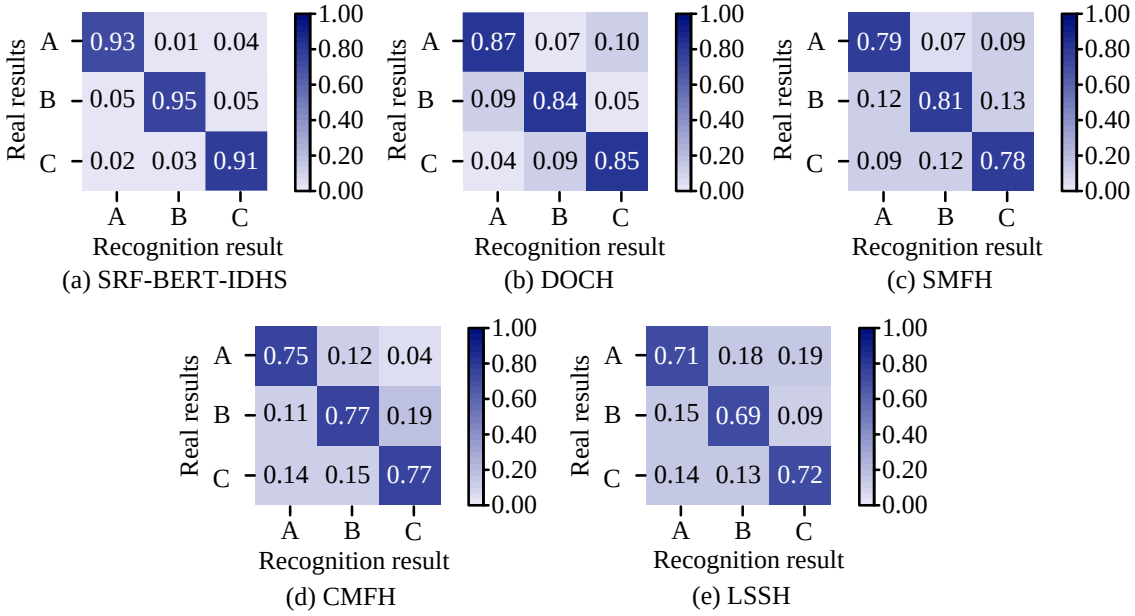


Figure 10: Comparison of retrieval confusion matrices of five models.

In Figure 9(a), the AP value of SRF-BERT-IDHS consistently outperformed the others, reaching 0.94 at 256-bit hash codes. The AP values of the other models also increased with longer hash codes, with DOCH reaching 0.85, which was 0.09 lower than SRF-BERT-IDHS. As shown in Figure 9(b), the AP value of SRF-BERT-IDHS further improved, reaching 0.95 at 256-bit hash codes. This is attributed to its integration of semantic association and feature reconstruction mechanisms, enabling more precise alignment of images and text in high-dimensional Spaces. On the 30% low-quality test set, SRF-BERT-IDHS still maintained an AP value of 0.91, significantly outperforming other models. The results show that this model has stronger robustness and generalization ability in real library retrieval scenarios. Especially when facing blurred images or incomplete texts, it can still maintain efficient matching performance, verifying its superior performance in practical applications. To evaluate the retrieval performance of SRF-BERT-IDHS on real datasets, the five models were tested on the training set, and their retrieval confusion matrices are shown in Figure 10.

It can be seen from Figure 10(a), SRF-BERT-IDHS achieved an average retrieval accuracy of 0.93 on the training set. As can be seen from Figure 10(b), DOCH performed slightly worse than SRF-BERT-IDHS, with an average accuracy of 0.85. From Figure 10(e), LSSH performed worst, with an average accuracy of only 0.71, which was 0.22 lower than SRF-BERT-IDHS. Overall, SRF-BERT-IDHS achieved the best recognition of multimodal data, with fewer misclassifications and the highest retrieval accuracy. To fully demonstrate the superior performance of SRF-BERT-IDHS, the top-N precision curves of the first 1,000 samples in the training set were compared across the five models, as shown in Figure 11.

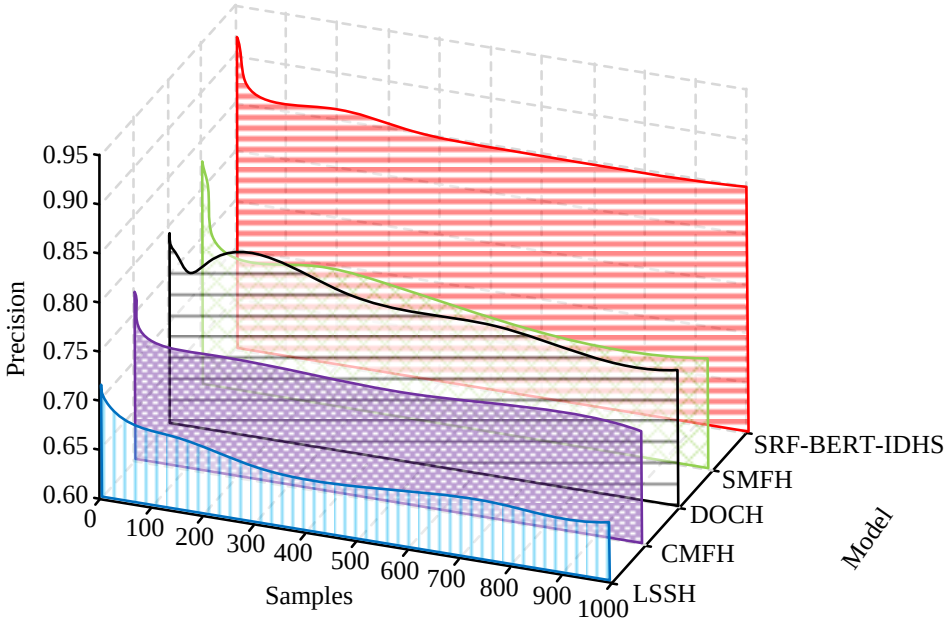


Figure 11: Comparison of accuracy curve results.

Table 3: Ablation study results on retrieval performance.

Project	Retrieval speed (s/ sample)	Training time (h)	Average accuracy rate (%)
FCN+SE-ResNet	0.034±0.02	12.5±1.2	76.8±2.2
FCN+Triplet loss only	0.029±0.03	11.8±1.6	79.3±2.4
FCN+Contrastive loss only	0.031±0.02	12.1±1.5	78.5±2.3
Full model	0.025±0.01*&#	10.3±1.3*&#	87.9±2.5*&#

Note: In Table 3, * indicates that the result difference between the Full model and FCN+SE-ResNet is significant (p<0.05); The results of the Full model and FCN+Triplet loss only represented by & were significantly different (p<0.05); # Indicates that the result difference between the Full model and FCN+Contrastive loss only is significant (p<0.05).

It can be seen from Figure 11, the top-N curve represents the variation of retrieval accuracy with the number of retrieval structure samples. The larger the area enclosed by the curve, the better the model performance. SRF-BERT-IDHS always maintains the final accuracy. As the sample size increases, the accuracy values of each model decrease. However, the decline of SRF-BERT-IDHS was the smallest, demonstrating stronger stability and generalization ability. Especially in the first 500 samples, its accuracy remained above 0.9, significantly superior to other models. The reason why SRF-BERT-IDH performs well is that it integrates semantic enhancement mechanisms and feature re-weighting strategies, effectively improving the alignment accuracy between text and image modalities. Meanwhile, the model introduces a context-aware semantic mapping module during the hash encoding process, significantly enhancing its robustness and retrieval stability in complex scenarios.

To test the effectiveness of the improved method proposed in the research, an ablation experiment was designed to analyze and compare its performance. The comparison indicators include retrieval speed, training time and average accuracy rate. The baselines of the comparative experiments include FCN+ SE-ResNet, FCN+Triplet loss only, FCN+Contrastive loss only, and Full model. The results are shown in Table 3.

As shown in Table 3, the complete model significantly outperforms each baseline method in terms of average accuracy, reaching 87.9%, and has the shortest training time, only 10.3 hours. This indicates that the introduced optimization strategy effectively enhances the model's convergence speed and retrieval accuracy. Compared with the variants that only use Triplet loss or Contrastive loss, the Full model further enhances cross-modal semantic consistency through the joint loss function and feature reweighting mechanism, verifying the effectiveness of the synergy of each module.

Although the BookCover dataset is representative in the task of book cover recognition, its sample distribution is limited to specific publication years and regional categories, making it difficult to comprehensively reflect the visual semantic differences across cultures and styles. To verify the generalization ability of the method, experiments were further conducted on the Flickr30K and MSCOCO datasets covering multi-domain image-text pairs. The results show that the proposed model maintains stable performance improvement under different semantic densities and noise levels, confirming its potential to adapt to diverse scenarios. Especially under complex backgrounds and low-quality image conditions, the model can still maintain a high retrieval accuracy rate, demonstrating good robustness.

5 Conclusion and future work

To address the bottleneck of multimodal retrieval in online libraries and to improve cross-modal retrieval efficiency and accuracy, SRF-BERT-IDHS explored a multimodal

retrieval approach for book images and text resources. The study built a retrieval framework that integrates FCN and Hash Learning. It used FCN to extract deep features of book cover images and combined them with text branch features. Through deep hash learning, the features were mapped to a low-dimensional semantic space. Triplet loss and contrastive learning loss were designed to optimize cross-modal semantic alignment. The results showed that the SRF-BERT-IDHS performed well in multimodal retrieval tasks. For the image-to-text retrieval task, the mAP reached 0.864 with a 256-bit hash code, which was higher than 0.604 for LSSH and 0.662 for CMFH. For the text-to-image retrieval task, the mAP reached 0.931 with a 256-bit hash code, with both precision and efficiency at a leading level. The AP value reached 0.940, and the average retrieval accuracy was 0.930. Even when the sample size reached 1000, the precision remained above 0.85. These results verified the capability of FCN to extract deep image features and the advantage of hash learning in dimension compression and semantic association enhancement. The combination effectively overcame modality barriers and met the multimodal resource retrieval needs of online libraries. However, although the dataset in The proposed method covered multiple book categories, its scale was relatively limited, and the robustness of the model under extreme long-tail distribution remained to be tested. When integrating other systems in the future, the research considers embedding the proposed model into the existing online library retrieval system, achieving efficient connection with the background database through API interfaces, and supporting real-time feature extraction and hash code matching. During the deployment process, a lightweight network structure and model compression technology are adopted. Meanwhile, an incremental learning mechanism is introduced to support online model updates and dynamic optimization, adapting to the new book entry and changes in user behavior.

6 Discussion

The cross-modal retrieval model of digital libraries proposed in the research maintains high accuracy and stable performance in both image and text retrieval. This is attributed to the deep characterization of image semantics by FCN and the effective modeling of cross-modal associations by hash coding. The deep shared proxy hash construction method mentioned in Reference [20] achieves a compact expression of cross-modal semantics through the proxy hash loss function, thereby enhancing the retrieval efficiency. This echoes the triplet loss and contrastive learning collaborative optimization strategy proposed in SRF-BERT-IDHS, both of which are dedicated to enhancing cross-modal semantic consistency. The federated cross-modal hashing method based on privacy enhancement culprits mentioned in reference [21] focuses on the balance between privacy protection and retrieval efficiency in distributed data storage. This method realizes cross-modal retrieval while ensuring user data privacy, providing a new idea for the distributed digital library scenario. Although SRF-BERT-IDHS did not directly involve privacy protection mechanisms, the constructed hash framework has good scalability. In the future, it can integrate federated learning strategies to enhance cross-modal retrieval capabilities while ensuring data security, further promoting the service upgrade of smart libraries and the development of trusted computing.

Pan R et al. proposed a knowledge base retrieval learning method to address the challenge of semantic consistent negation in image-text retrieval. This method enhances the accuracy and robustness of cross-modal semantic alignment by introducing a knowledge base and a lightweight cluster refinement strategy [22]. This is similar to the idea of using ResNet to replace the VGG network in FCN to enhance the ability of image feature extraction, both aiming to strengthen the semantic consistency between modalities. The residual structure of ResNet effectively alleviates the degradation problem of deep networks, enabling the model to have stronger representational capabilities when processing complex images and thereby enhancing the accuracy of cross-modal matching. The research method poses certain challenges in future practical applications and expansions. It is necessary to consider the balance between the computing resource consumption of model deployment and the timeliness of response. Especially in scenarios with large-scale concurrent user access, the efficiency of hash code generation and the design of index structure need to be further optimized. The efficient cross-modal feature matching model based on the CLIP framework mentioned in Reference [23] divides the model into two parts: feature extraction and contrastive learning. By pre-training a large model, it realizes the unified semantic space mapping of images and text, significantly improving the cross-modal matching efficiency. This idea provides an important reference for optimizing the feature extraction module in SRF-BERT-IDHS. In the future, it can be combined with lightweight CLIP variants to reduce computational overhead while maintaining high-precision retrieval performance.

Based on the above content, it can be known that the cross-modal retrieval method proposed in the research shows significant advantages in semantic alignment accuracy and model scalability. Especially after combining the triple loss and contrastive learning mechanism, the association expression ability between images and texts is further enhanced. In the future, practical applications and deployments need to be oriented towards the complexity of real scenarios. It is advisable to consider introducing a dynamic adaptive hash code length adjustment mechanism to address the matching deviation caused by the distribution differences of different modal data. Meanwhile, in combination with the edge computing architecture, some feature extraction tasks are decentralized to terminal devices to reduce the load pressure on the central server.

References

[1] Shubhi Bansal, Mohit Kumar, Chandravardhan Singh Raghaw, and Nagendra Kumar. Sentiment and hashtag-aware attentive deep neural network for

multimodal post popularity prediction. Neural Computing and Applications, 37(4):2799-2824, 2025. https://doi.org/10.1007/s00521-024-10755-5

[2] Yangdong Chen, Jiaqi Quan, Yuejie Zhang, Rui Feng, and Tao Zhang. Deep cross-modal hashing with fine-grained similarity. Applied Intelligence, 53(23):28954-28973, 2023. https://doi.org/10.1007/s10489-023-05028-y

[3] Prabhjot Kaur, and Chander Kant. 2-Phase multi-trait biometric authentication model against spoofing attack using deep hash model. SN Computer Science, 6(1):47-59, 2025. https://doi.org/10.1007/s42979-024-03513-w

[4] Donghuo Zeng, Jianming Wu, Gen Hattori, Rong Xu, and Yi Yu. Learning explicit and implicit dual common subspaces for audio-visual cross-modal retrieval. ACM Transactions on Multimedia Computing, Communications and Applications, 19(2s):1-23, 2023. https://doi.org/10.1145/3564608

[5] Jiaxing Li, Wai Keung Wong, Lin Jiang, Xiaozhao Fang, Shengli Xie, and Yong Xu. CKDH: CLIP-based knowledge distillation hashing for cross-modal retrieval. IEEE Transactions on Circuits and Systems for Video Technology, 34(7):6530-6541, 2024. https://doi.org/10.1109/TCSVT.2024.3350695

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. Proceedings of the 38th International Conference on Machine Learning, 139:8748-8763, 2021. https://doi.org/10.48550/arXiv.2103.00020

[7] Jie Lin, Olivier Morere, Vijay Chandrasekhar, Antoine Veillard, and Hanlin Goh. DeepHash: Getting regularization, depth and fine-tuning right. ACM Transactions on Multimedia Computing, Communications and Applications, 18(1):1-21, 2022. https://doi.org/10.48550/arXiv.1501.04711

[8] Jian Zhang, Yuxin Peng, and Mingkuan Yuan. Unsupervised generative adversarial cross-modal hashing. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1):539-546, 2018. https://doi.org/10.1609/aaai.v32i1.11263

[9] Zhenqiu Shu, Li Li, Jun Yu, Donglin Zhang, Zhengtao Yu, and Xiao-Jun Wu. Online supervised collective matrix factorization hashing for cross-modal retrieval. Applied Intelligence, 53(11):14201-14218, 2023. https://doi.org/10.1007/s10489-022-04189-6

[10] Saeid Sattari, Sinan Kalkan, and Adnan Yazici. Multimodal multimedia information retrieval through the integration of fuzzy clustering, OWA-based fusion, and Siamese neural networks. Fuzzy Sets and Systems, 515:109419, 2025. https://doi.org/10.1016/j.fss.2025.109419

[11] Donglin Zhang, Xiao-Jun Wu, and Guoqing Chen. ONION: Online semantic autoencoder hashing for cross-modal retrieval. ACM Transactions on Intelligent Systems and Technology, 14(2):1-18, 2023. https://doi.org/10.1145/3572032

[12] Asad Khan, Sakander Hayat, Muhammad Ahmad, Jinyu Wen, Muhammad Umar Farooq, Meie Fang, and Wenchao Jiang. Cross‐modal retrieval based on deep regularized hashing constraints. International Journal of Intelligent Systems, 37(9):6508-6530, 2022. https://doi.org/10.1002/int.22853

[13] Yongxin Wang, Yu-Wei Zhan, Zhen-Duo Chen, Xin Luo, and Xin-Shun Xu. Multiple information embedded hashing for large-scale cross-modal retrieval. IEEE Trans Circuits Syst Video Technol, 34(6):5118-5131. https://doi.org/10.1109/TCSVT.2023.3340102

[14] Haokun Wen, Xuemeng Song, Jianhua Yin, Jianlong Wu, Weili Guan, and Liqiang Nie. Self-training boosted multi-factor matching network for composed image retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(5):3665-3678, 2023. https://doi.org/10.1109/TPAMI.2023.3346434

[15] Xingwei Zhang, Xiaolong Zheng, Wenji Mao, and Daniel Dajun Zeng. Boosting deep cross-modal retrieval hashing with adversarially robust training. Applied Intelligence, 53(20):23698-23710, 2023. https://doi.org/10.1007/s10489-023-04715-0

[16] Yu Liu, Haipeng Chen, Guihe Qin, Jincai Song, and Xun Yang. Bias mitigation and representation optimization for noise-robust cross-modal retrieval. ACM Transactions on Multimedia Computing, Communications and Applications, 21(11):1-17, 2025. https://doi.org/10.1145/3700596

[17] Yuxia Cao. Joint feature fusion hashing for cross-modal retrieval. International Journal of Machine Learning and Cybernetics, 15(12):6149-6162, 2024. https://doi.org/10.1007/s13042-024-02309-x

[18] Venice Erin Liong, Jiwen Lu, Gang Wang, Pierre Moulin, and Jie Zhou. Deep hashing for compact binary codes learning. 2015 IEEE Conference on Computer Vision and Pattern Recognition, 2475-2483, 2015. https://doi.org/10.1109/cvpr.2015.7298862

[19] Ying Ma, Meng Wang, Guangyun Lu, and Yajun Sun. Multi-label semantic sharing based on graph convolutional network for image-to-text retrieval. The Visual Computer, 41(3):1827-1840, 2025. https://doi.org/10.1007/s00371-024-03496-y

[20] Lirong Han, Mercedes E. Paoletti, Sergio Moreno-Álvarez, Juan M. Haut, and Antonio Plaza. Deep shared proxy construction hashing for cross-modal remote sensing image fast target retrieval. ISPRS Journal of Photogrammetry and Remote Sensing, 218(1):44-56, 2024. https://doi.org/10.1016/j.isprsjprs.2024.10.004

[21] Ruifan Zuo, Chaoqun Zheng, Fengling Li, Lei Zhu, and Zheng Zhang. Privacy-enhanced prototype-based federated cross-modal hashing for cross-modal retrieval. ACM Transactions on Multimedia Computing, Communications and Applications, 20(9):1-19, 2024. https://doi.org/10.1145/3674507

[22] Renjie Pan, Hua Yang, and Xiangyu Zhao. Real: Improving image-text retrieval with authentic negative repository learning. ACM Transactions on Multimedia Computing, Communications and Applications, 21(10):1-22, 2025. https://doi.org/10.1145/3729172

[23] Yilin Peng. A CLIP-based cross-modal matching model for image-text retrieval. Information Technology and Control, 54(3):1030-1048, 2025. https://doi.org/10.5755/j01.itc.54.3.41801

An Engine-Driven Multi-modal Interaction Framework for Enhancing User Immersion in 3D Virtual Scenes via Smart Gloves and MUFD Algorithms

Xia Wang
School of Information and Intelligence Transportation, Fujian Chuanzheng Communications College, Fuzhou, 350007, China
E-mail: xiawangx@outlook.com

Keywords: engine-driven, three-dimensional virtual scenes, multi-modal interaction, user experience, augmented reality

Received: September 17, 2025

With the rapid development of virtual reality (VR), augmented reality (AR), and other technologies, the interaction experience of three-dimensional virtual scenes has become a hot spot of research. To enhance user immersion, this research proposes an engine-driven multi-modal interaction design for three-dimensional virtual scenes. An interdisciplinary experiment involving 20 volunteers compared multimodal interaction (MMI) and traditional virtual interaction. The interaction process is optimized by combining an intention-capture algorithm, intelligent gloves, a multimodal fuzzy data understanding (MUFD) algorithm, and an axis-aligned bounding box (AABB) collision detection algorithm. The design utilizes an intent capture algorithm for accurately sensing the user's experimental intent, including multiple sources of information such as vision, gesture, and eye tracking. Moreover, a smart glove is used to combine the set of intent probabilities from different channels to more accurately capture the ambiguous or incomplete intent of the user. The results showed that the minimum accuracy for visual interaction was 91.42%. The minimum accuracy for gesture interaction and eye movement interaction was 92.83% and 92.75%, respectively. Compared to traditional virtual interaction, this research method reduced response time by 41.85% and achieved system stability of 99.90%. In terms of immersion, the scores for perceived realism and emotional response were 6.25±1.70 and 5.81±1.67, respectively. Based on the Igroup Presence Questionnaire (IPQ) assessment, the multimodal interaction group showed a 37.96% increase in perceived reality score and a 38.33% increase in emotional response score compared to the traditional group. The study approach received minimal user experience scores of 4.05, 4.26, and 4.84 for naturalness, simplicity of use, and ease of starting, respectively. Furthermore, the system response time took only 120.74ms. In summary, the engine-driven multi-modal interaction design and user immersion enhancement strategy for three-dimensional virtual scenes proposed in this study can significantly enhance users' VR experience. Especially in complex tasks and demanding application scenarios, it can effectively enhance user engagement and satisfaction.

Povzetek: Študija predstavi večmodalno interakcijo za 3D VR prizore (vid + geste + sledenje očem + pametna rokavica), ki v primerjavi s klasično interakcijo skrajša odzivni čas za ~42 %, ohrani 99,9 % stabilnost ter poveča občutek prisotnosti in čustveni odziv za ~38 %.

1 Introduction

Virtual reality (VR) and augmented reality (AR) technology have advanced quickly in recent years. This has led to the increasing application of three-dimensional virtual scenes in a variety of fields, such as gaming, education, healthcare, architecture, and so on. Especially in safety critical fields such as surgery and aerospace, VR/AR technology requires extremely high real-time, stability, and accuracy of interaction. For example, VR training for spinal surgery requires an interaction delay of 150 ms or less to ensure operational synchronization. Meanwhile, VR simulation for aircraft piloting requires system stability of at least 99.5% to prevent training interruptions. Traditional multimodal systems often struggle to meet these requirements simultaneously, becoming a core bottleneck that restricts their application in safety critical scenarios. Traditional interaction methods mainly focus on the visual and auditory levels, while modern VR is gradually integrating tactile, kinesthetic, and other perceptual modalities, forming the trend of multi-modal interaction (MMI) [1]. However, traditional single interaction modes are often difficult to meet users' needs for deep immersion experiences. The limitations of these interaction modes have become the technical bottleneck that restricts the enhancement of VR immersion [2]. Therefore, it is particularly important to research and develop MMI technology. MMI can not only provide users with richer sensory stimulation, but also by synthesizing multiple perceptual channels. This can enhance the immersion of the virtual scene, thus effectively

enhancing the user's sense of participation and immersion. Unity3D, as a mature virtual engine software, is widely used in game development and architectural digitization [3]. It provides the technological basis for building high-fidelity three-dimensional scenes by virtue of its powerful real-time rendering capability, physical simulation and cross-platform support [4-5]. Currently, VR immersive user experience mostly relies on head-mounted devices and controllers or bare hands to interact with the virtual world. However, this type of interaction mainly relies on eye perception and lacks a sense of manipulation and realism [6]. In this context, smart gloves, as a convenient and flexible interaction device, provide new possibilities for MMI in virtual scenes. Smart gloves are able to make up for the shortcomings of traditional interaction methods through tactile and kinesthetic feedback. Therefore, the study innovatively proposes a MMI design that combines vision, gesture, and eye tracking, aiming to enhance the user's immersion. The study captures the user's intention and interacts in three-dimensional virtual scenes with a view to enhancing the coupling relationship between the user's psychological and sensory immersion.

2 Related works

As information technology continues to advance, VR and AR are now widely utilized in a variety of industries. Experts and academics in the domains of healthcare, education, and cultural heritage have given them a lot of attention. For example, to investigate the use of VR in spine medicine, Dargan S et al. conducted a thorough assessment of its application in surgery, counseling, education, and rehabilitation. The outcomes revealed that the application of VR in spine medicine gradually accelerated with the support of three-dimensional medical imaging, holograms, wearable sensors, 5G technology, artificial intelligence (AI), and head-mounted displays [7]. To increase the accessibility and pleasure potential for users of VR, AR, and the metaverse, Dudley J. et al. suggested the idea of inclusive immersion. The results showed that although technologies for VR and AR headsets were progressively becoming affordable and effective, these technologies had not yet achieved widespread user adoption. In particular, the needs of a wider and diverse user community needed to be considered [8]. To increase the realism, immersion, and overall experience of surgical simulation, Lungu A. J. et al. suggested utilizing VR, AR, and mixed reality (MR) technology. The results showed that the key components of a VR surgical simulator were visual and haptic feedback [9]. Duarte M L et al. investigated if these cutting-edge technologies might supplement or replace conventional anatomy teaching techniques in order to evaluate the efficacy of VR and AR in anatomy education. The results demonstrated the significant advantages of VR and AR in improving student engagement, learning efficiency, and knowledge retention [10].

In the realm of VR and AR, MMI design is essential. Al-Ansi A M et al. proposed a system design based on MMI techniques in order to enhance visitor experience through visual and audio interaction interfaces. The outcomes showed that the interactive system was effective in evoking visitors' natural interaction with the cultural heritage environment and facilitating a deeper understanding of the cultural content [11]. To increase the effectiveness of human-machine and human-robot interactions, Sereno M. et al. suggested an MMI technique based on the MQTT protocol. The findings demonstrated that the design accommodated a variety of interaction techniques, including touch, speech, and gaze tracking. Moreover, it could communicate effectively between multiple devices such as computers, smartphones, tablets, etc [12]. Weitz K et al. suggested a user MMI method based on a basic voice recognition system to investigate the possibilities of virtual agents in explainable artificial intelligence (XAI) interface design. According to the findings, including virtual agents might boost users' confidence in the XAI system [13]. Behavioral trajectories, learning outcomes, task performance, teacher assistance, student engagement, and feedback are some of the six primary objectives that Sharma K et al. examined in an attempt to investigate the use of multi-modal technology in education. The findings demonstrated the great potential of multi-modal technology to record and enhance the learning process [14].

Wang H et al. conducted a large-sample experiment to verify the advantages of a multimodal VR anatomy teaching system in terms of the long-term knowledge retention rate. The aim was to study how VR teaching improves efficiency. The results showed that students who used visual gesture eye movement multimodal interaction had a 28% increase in knowledge retention rate after 3 months compared to traditional VR teaching [15]. Zhang Y et al. proposed a multimodal interaction fusion model based on attention mechanism to improve interaction accuracy. The model optimizes by dynamically adjusting the weights of visual, gesture, and speech modalities. The results showed that this method improved the accuracy of intent recognition in complex scenes [16]. In order to improve interaction sensitivity, Li J et al. designed a real-time interaction optimization scheme based on 5G edge computing for the problem of tactile feedback delay of smart gloves. The results showed that this method reduced the tactile feedback delay to less than 50ms [17]. The systematic comparison of literature review is shown in Table 1.

Table 1: Systematic comparison of literature review

Research	Immersion rating	Domains applied	Limitations
Dargan S et al	/	Healthcare (Spine medicine)	Lacks quantitative immersion assessment; no

[7].			multi-modal fuzzy intent processing
Dudley J et al [8].	/	VR/AR/Metaverse (inclusive design)	No clear modality definition; fails to address safety-critical scenario adaptation
Lungu A J et al [9].	High fidelity	Healthcare (surgical simulation)	No numerical immersion score; does not integrate intent fusion algorithms
Duarte M L et al [10].	Improved engagement	Education (anatomy teaching)	Single-modality dependent; no quantitative immersion evaluation
Al-Ansi A M et al [11].	Enhanced natural interaction	Cultural heritage	Lacks multi-modal data fusion; ignores environmental interference resistance
Sereno M et al [12].	/	HMI/HRI (Human-machine/robot interaction)	No immersion assessment; does not optimize system real-time performance
Weitz K et al [13].	Improved trust	XAI interaction design	No multi-modal intent integration; limited to trust enhancement only
Sharma K et al [14].	/	Education (learning process)	Vague modality description; no practical scenario validation
Wang H et al [15].	/	Education (VR anatomy teaching)	No immersion quantification; fails to verify adaptability in complex scenarios
Zhang Y et al [16].	/	General MMI	Ignores fuzzy intent handling; no immersion performance assessment
Li J et al [17].	/	MMI hardware optimization	Only optimizes haptic delay; no multi-modal data fusion integration

As evidenced by the above comparison and existing research, there are still many shortcomings in existing studies. For example, the research of Zhang et al. and Wang et al. only focuses on multimodal combinations or single-modal optimizations. They failed to address users' vague or incomplete intentions. These studies do not use fuzzy reasoning to fuse uncertain data from different modalities, resulting in poor system adaptability when user intent expression is unclear. Almost all studies only provide qualitative descriptions of immersion, lacking numerical scores or standardized scales for evaluation. This makes it impossible to conduct cross-study comparisons or performance benchmarking. Moreover, most of the research is limited to a single noncritical area without verifying its effectiveness in safety-critical areas requiring high real-time performance and low error tolerance. For example, although Li J et al. optimizes tactile feedback delay, they did not test the design in surgical VR scenarios where delay control is crucial. In addition, there are few studies that combine multimodal design with underlying interaction optimization techniques. Unlike this study, which uses multimodal interaction (MMI) and axis-aligned bounding boxes (AABBs) for collision detection to reduce computational load, existing research, such as that of Sereno et al. and Sharma et al., rarely considers such optimization. This results in an insufficiently smooth and stable system in complex virtual scenes. Therefore, the proposed engine-driven MMI design for three-dimensional virtual scenes, which combines multiple interaction modes and immersion enhancement strategies, can make up for the shortcomings of existing research. This further promotes the application of VR in a wider range of fields, meets the needs of diverse users, and enhances their interaction experience and immersion.

3 Engine-driven design for MMI

The study is based on the Unity3D virtual engine and combines glove sensors to capture intent. It uses a hierarchical enclosing box algorithm to determine if objects are colliding, thereby optimizing interaction experiences. Meanwhile, the smart glove is adopted as the core interaction interface, and the MMI method of vision, gesture, and eye tracking is designed. Moreover, the fuzzy reasoning and multi-modal understanding of fuzzy data (MUFD) algorithms are used to integrate the user's final intention and enhance the user's immersion.

3.1 Intent capture and interaction optimization in engine-driven three-dimensional virtual scenes

Unity3D, a modern game engine, is not only equipped with real-time graphic rendering and physical simulation, but also supports access to deep learning modules, semantic analysis plug-ins, and external interaction devices. This provides a technical basis for intent modeling and optimization [18-19]. To further enhance the robustness and adaptability of

multimodal interaction, the research introduces adaptive control and robust control principles to construct cross domain technology links. The study first adopts the model reference adaptive strategy in adaptive control to monitor the input signal-to-noise ratio and data integrity of visual, gesture, and eye movement modalities in real time. When the SNR of eye movement tracking drops below 30 dB due to a change in ambient light intensity, the system automatically adjusts the modal weight from 0.3 to 0.1. Meanwhile, it increases the weights of the gesture and visual modes to prevent interruption of interactions caused by a failure of a single mode. In addition, the research also integrates the robust control H ∞ control theory. By constructing a noise suppression module, it successfully controls the interference from sensor noise and data transmission delays on intent recognition to within 5%. This ensures the system's stable operation in complex electromagnetic environments such as aerospace VR simulators, while also being suitable for scenarios involving high-frequency equipment interference in surgical VR procedures. This design fills the technical gap of "heavy fusion, light anti-interference" in existing multimodal systems [20-21]. Through the behavioral analysis scripts and state machine models integrated in the engine, it is possible to achieve in-depth understanding and prediction of the user's operation paths, gaze points, and behavioral sequences, thus promoting the construction of a closed-loop interaction "from recognition to response" [22]. Therefore, in order to achieve interaction optimization, this study adopts an engine-driven three-dimensional virtual scenes intent capture and interaction optimization strategy. Engine driven intent capture and interaction optimization overall architecture is shown in Figure 1.

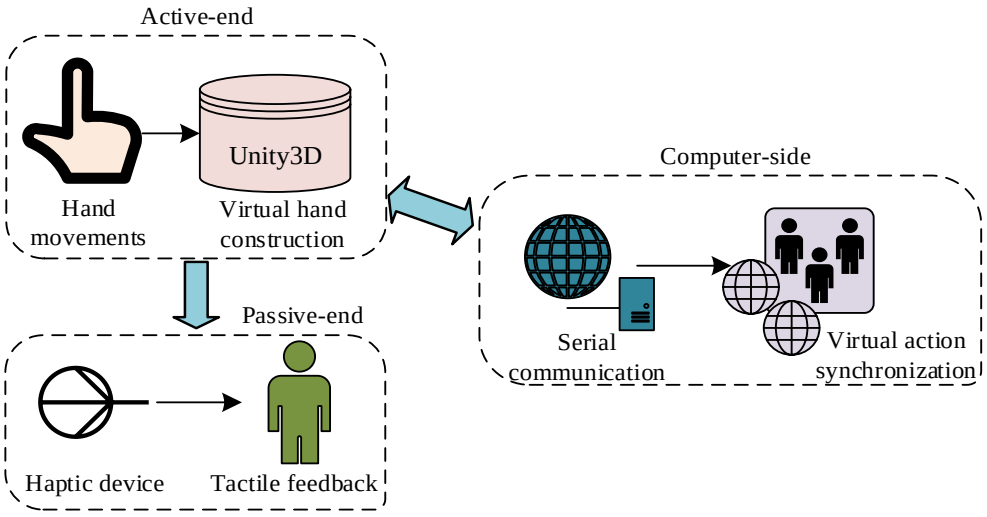


Figure 1: Engine driven intent capture and interaction optimization overall architecture

In Figure 1, in this overall architectural design, the user wears a Leap Motion sensing device to capture hand movements. The Unity3D virtual engine builds the virtual scene on the computer and controls the operation of the virtual hand. When the virtual hand makes contact with a virtual object, the system detects the contact. Moreover, the contact information is transmitted to the force haptic device worn on the slave user through the serial port. The device controls the spring and pressure plate through the servo to simulate the feeling of a finger touching an object, thus enabling the user at the slave end to perceive the presence of the virtual object. Additionally, sensors on the active user's arm collect real-time motion data and transmit it to the exoskeleton device on the follower's arm. This guides the arm movements and synchronizes them with the active user's movements. Through this architectural design, user intent capture and interaction optimization can be effectively achieved. The Leap Motion intent capture process is shown in Figure 2.

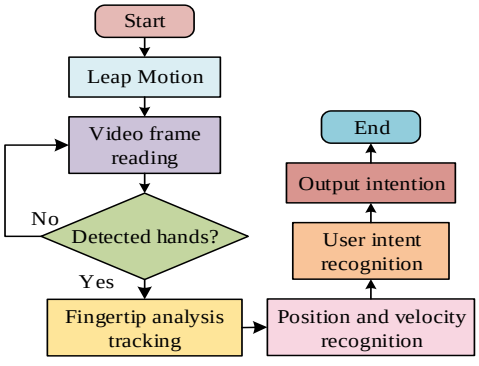
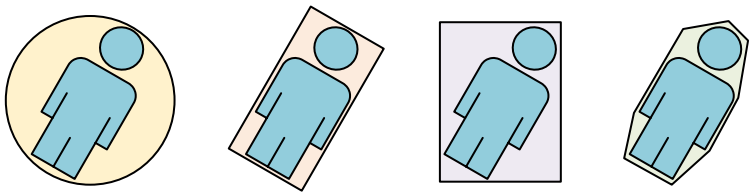


Figure 2: Leap motion intent capture process

In Figure 2, the process first utilizes Leap Motion to track the movements of the user's hand and fingers, and transmits the data to the computer side for video frame reading. The system can determine whether the video frame contains hand information. If the hand is detected, fingertip analysis and tracking is performed.

If the hand is not detected, the system will re-read the video frame. The technology analyzes and tracks the user's fingertips before determining the locations and speeds of the fingers to detect the user's motions. The system understands the user's objectives based on this information. In virtual interaction, it is crucial to determine whether virtual objects collide with each other. If efficient collision detection is not performed, it can lead to excessive computation, which will affect the smoothness of the screen [23]. Therefore, to improve the efficiency of collision detection and achieve interaction optimization, the study adopts a hierarchical enveloping box algorithm to determine whether objects collide or not. The algorithm works by wrapping a virtual object with a layer of slightly larger geometric shape boxes. Then, based on the intersection of these boxes, it can determine whether the object has collided or not [24]. The hierarchical enclosing box algorithm is displayed in Figure 3.



(a) Bounding sphere (b) OBB (c) AABB (d) K-DOP

Figure 3: Hierarchical bounding box algorithm

In Figure 3, the hierarchical bounding box algorithm has several common forms of bounding balls, oriented bounding box (OBB), AABB, and K-discrete orientation polytopes (K-DOP) bounding boxes [25]. Among them, the rotation of the enclosing sphere does not vary with the velocity of the object, and thus is not applicable to collision detection of deformed objects [26]. However, OBB and K-DOP have slower and more complex rotation following, and thus are more computationally intensive [27]. In contrast, AABB has lower computational complexity and is more suitable for interaction design in VR. The wrapping box of AABB is a polyhedron that can be constructed by describing multiple scalars. The computational process is simpler and the wrapped objects have smaller gaps [28]. Therefore, the study chooses AABB for collision determination, thus achieving efficient collision detection and interaction optimization design while keeping the computational effort low. Equation (1) displays AABB's mathematical expression.

$$R_{AABB}=\left\{(x,y,z)\middle|a_{\min}\le x\le a_{\max},b_{\min}\le y\le b_{\max},c_{\min}\le z\le c_{\max}\right\} \tag{1}$$

In Equation (1), R_{AABB} denotes the enclosing box of AABB. (x,y,z) denotes the coordinate position of a point in three-dimensional space. $a_{\min}$ and $a_{\max}$ denote the minimum and maximum values in the x direction. Similarly, $b_{\min}$, $b_{\max}$, $c_{\min}$, and $c_{\max}$ denote the minimum and maximum values in the y and z directions, respectively.

2.2 Smart glove-based interaction with vision, gesture, and eye tracking

After achieving accurate recognition of user intent and interaction optimization at the engine layer, how to naturally integrate these multiple input modalities into the user's operating behavior becomes the key to further enhance the sense of immersion. Currently, visual recognition, gesture interaction, and eye tracking are considered as three key perception channels, which are building a new multi-modal human-computer interaction paradigm [29]. This system incorporates smart glove-based fusion interaction technology as a core component of the MMI system. This technology is highly compatible with natural movements and can simultaneously capture haptic, electromyographic, and motion information [30]. The smart gloves can accurately capture the degree of bending, pressure changes, and inertial movement trajectories of the finger joints through the embedded sensor network. This can restore high-fidelity hand animations in three-dimensional engine scenes for human-like motion actuation [31]. Meanwhile, combined with the depth camera and AI image recognition algorithm, the system can recognize the visual features of the user's environment in real time, such as the scene structure, object boundaries, and surface texture, to help quickly lock the interaction target. The YOLOv7 algorithm is chosen over YOLOv5 and other lightweight models for the visual channel because it strikes a balance between real-time performance, detection accuracy, and computational resource consumption requirements in VR scenes. VR interaction requires precise positioning of virtual objects. The YOLOv7 algorithm is more accurate than the YOLOv5 algorithm, especially when detecting small virtual targets. The higher accuracy advantage can avoid misjudgment of interaction intention caused by visual positioning deviation. Although YOLOv7 has slightly higher parameter count than YOLOv5, it optimizes feature fusion efficiency through the ELAN structure. In this experiment, YOLOv7 achieved an inference speed of 85 FPS for 1280×720 virtual scene images in the NVIDIA GeForce RTX 3090 hardware environment. This result is only 7.6% lower than YOLOv5 and far higher than the 30 FPS minimum threshold required for VR interaction. If a lighter YOLO Nano structure is chosen, although the inference speed can reach 120 FPS, the small object detection recall rate cannot meet the high-precision interaction requirements. In addition, YOLOv7 supports dynamic batch inference, which automatically adjusts the size of inference batches based on the complexity of virtual

scenes. When the number of objects in the scene is less than 10, the inference speed increases to 98 FPS. Even when the number of objects exceeds 30, it can maintain a frame rate of 65 FPS or higher. This avoids the sudden drop-in frame rate that traditional fixed batch models experience in complex VR scenes. It also ensures synchronized visual, gesture, and eye movement multimodal interaction. This prevents immersion interruption caused by visual delay. Eye tracking technology, as an important dimension of sight orientation, can help the system recognize the user's current focus of attention, thus enhancing the precision of operation and the naturalness of interaction. Therefore, this study uses smart gloves as the core interaction interface to realize visual, gesture, and eye tracking interaction design. The framework of the smart glove-based visual, gesture, and eye tracking interaction system is shown in Figure 4.

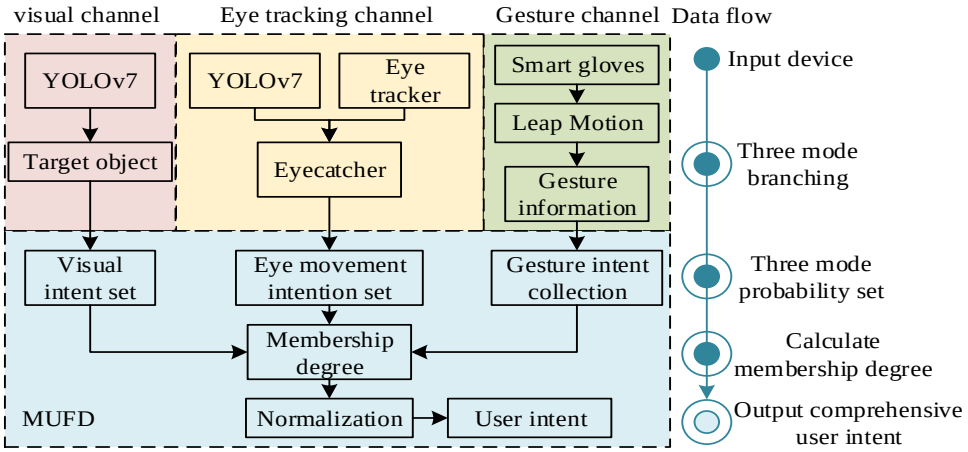


Figure 4: Framework for visual, gesture, and eye tracking interaction system based on smart gloves

In Figure 4, the system framework integrates three input modalities: vision, gesture, and eye tracking. The visual channel performs visual target object capture through the you only look once version 7 (YOLOv7) algorithm. Gesture is realized through smart gloves combined with Leap Motion for intent capture and tracking. Eye tracking is combined with eye tracking for focus capture. The target data collected through these three modalities are synthesized using fuzzy inference based and MUFD algorithms and the fuzzy set affiliation is calculated. After normalizing the affiliation degree, combined with the basic trust allocation function, the integrated user intent is finally obtained. The intention acquisition process of vision and eye tracking is shown in Figure 5.

In Figure 5, in this process, the current environment information is first captured by the camera and the YOLOv7 algorithm is utilized for object localization. In the visual channel, after initializing the YOLOv7 model, the system identifies the three-dimensional set of current objects. It also calculates the distance change of the same object in this frame and the next frame to derive the probability of intent under the visual channel and normalize it. Under eye tracking, after initializing the YOLOv7 model, the coordinates of the user's line-of-sight focus are obtained in conjunction with the eye-tracker. The image cropping region is determined based on the line-of-sight focus, and the image of the region is input into the YOLOv7 model. The confidence level of the object is obtained and this is used as the initial intent probability. Next, the system will normalize the confidence level and output the final set of intent probabilities. Equation (2) displays the formula for calculating distance change.

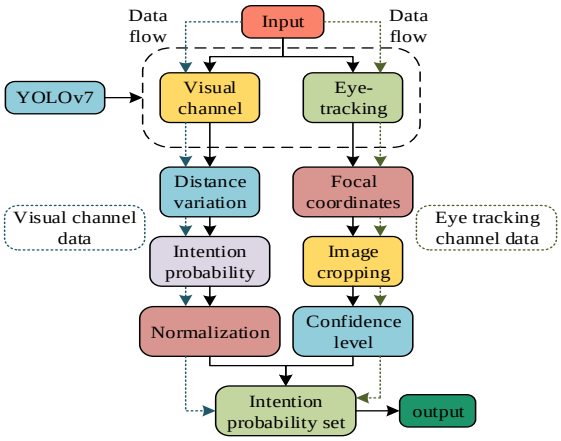


Figure 5 The intention acquisition process of vision and eye tracking

$$D_i = \sqrt{\left(x(t+1)_i - x(t)_i\right)^2 + \left(y(t+1)_i - y(t)_i\right)^2 + \left(z(t+1)_i - z(t)_i\right)^2} \quad (2)$$

In Equation (2), D_i denotes the change in distance between time steps t and $t+1$ for object i. $x(t)_i$, $y(t)_i$, and $z(t)_i$ denote the three-dimensional coordinates of object i at the moment of t, respectively. The formula for calculating the probability of visual intent is shown in Equation (3).

$$V_i = \frac{D_i}{\sqrt{x(t)_i^2 + y(t)_i^2 + z(t)_i^2}} \quad (3)$$

In Equation (3), V_i denotes the visual intention probability, i.e., the motion intention of object i between time steps t and $t+1$. The focus position formula for eye tracking is shown in Equation (4).

$$P_{focus} = \left(x_{focus}, y_{focus}\right) \quad (4)$$

In Equation (4), P_{focus} denotes the focus position of the eye tracking system. x_{focus} and y_{focus} denote the position of the user's visual focus on the two-dimensional plane under eye tracking, respectively [32]. The minimum and maximum visual field coordinates are shown in Equation (5).

$$\begin{cases} x_{\min} = x_{focus} - \frac{W}{2}, y_{\min} = y_{focus} - \frac{H}{2} \\ x_{\max} = x_{focus} + \frac{W}{2}, y_{\max} = y_{focus} + \frac{H}{2} \end{cases} \quad (5)$$

In Equation (5), W and H denote the width and height of the field of view, respectively. The confidence expression for the object category is shown in Equation (6).

$$O_i = \left(C_i, S_i\right) \quad (6)$$

In Equation (6), O_i denotes the confidence level of the object category i. C_i denotes the confidence value of the object. S_i denotes the categorization score of the object. The mathematical expression for the set of intentional probabilities is shown in Equation (7).

$$P_i = \frac{S_i}{\sum_{i=1}^{n} S_i} \quad (7)$$

In Equation (7), P_i denotes the probability of intent for each object category i. n denotes the number of object categories. The flow of intention understanding based on MUFD algorithm is shown in Figure 6.

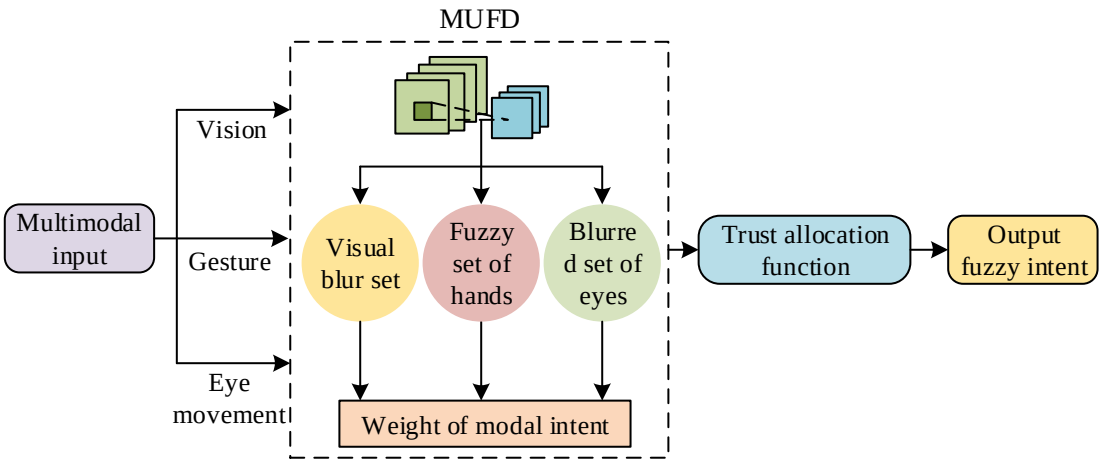


Figure 6: Intention understanding process based on MUFD algorithm

In Figure 6, in this process, multi-modal intent data from vision, gesture, eye movement, etc. are first input. Then, the fuzzy set corresponding to each modality is calculated, and the affiliation degree of each modality is also obtained. Next, the fuzzy intent trust distribution of each modality is calculated to obtain the weight of each modal intent. Finally, the final fuzzy intention of the user is obtained by constructing the trust distribution function and synthesizing the information of each modality. The mathematical expression for the set of intention probabilities is shown in Equation (8).

$$I \leftarrow MAG\left(V_{input}\right) \quad (8)$$

In Equation (8), I denotes the final set of user's intentions. V_{input} denotes multi-modal input. MAG denotes multi-modal aggregation [33]. The formula for the affiliation degree is shown in Equation (9).

$$u_i\left(x_i\right) = e^{-\frac{\left(x_i-\mu_i\right)^2}{2\sigma_i^2}} \quad (9)$$

In Equation (9), $u_i\left(x_i\right)$ is the affiliation of the input data x_i. μ_i is the center of the fuzzy set of i object. σ_i^2 denotes the variance of this fuzzy set. The normalization process is shown in Equation (10).

$$u_i'\left(x_i\right) = \frac{u_i\left(x_i\right)}{\sum_{1<i<n} u_i\left(x_i\right)} \quad (10)$$

The normalization operation by Equation (10) ensures that the sum of all fuzzy affiliations is 1. The reliability distribution is shown in Equation (11).

$$CF = u_i'(x_i) \quad (11)$$

In Equation (11), CF represents the credibility distribution, which indicates the credibility of each modal intention. The weighting formula is shown in Equation (12).

$$\omega(x_i) = \omega_{initial} + \Delta\omega \quad (12)$$

In Equation (12), $\omega(x_i)$ denotes the final weight of the input data x_i. $\omega_{initial}$ denotes the initial weights. $\Delta\omega$ denotes the increment adjusted according to the modal trust distribution. The basic trust distribution function is shown in Equation (13).

$$m(i) = \omega(x_i) \times u_i'(x_i) \quad (13)$$

Equation (13), $m(i)$ denotes the trust allocation function for modality i, i.e., the initial trust value for the modality. The combined trust allocation function is shown in Equation (14).

$$m(A) = \frac{1}{1-k} \sum_{\cap A_i = A} \prod_{1 \le i \le 3} m(i) \, A \ne \Phi \quad (14)$$

In Equation (14), $m(A)$ denotes the integrated trust allocation function, i.e., the trust allocation after the synthesis of all modalities. A denotes the set of all modalities. k denotes the constant used to adjust the overall trust allocation. $\prod_{1 \le i \le 3} m(i)$ denotes the product operation of the trust allocation function $m(i)$ for all modalities. The final fuzzy intent is shown in Equation (15).

$$V = \max m(A) \quad (15)$$

In Equation (15), V represents the final fuzzy intent of the user. That is, the maximum value is chosen to determine the most plausible modality to arrive at the final decision or intention of the system. To further clarify the mathematical logic of the MUFD algorithm and improve reproducibility, the study first uses the construction of fuzzy sets based on Gaussian fuzzy functions. Then, it makes the calculation of the trust allocation function satisfy the normalization condition $\sum m(i) = 1$. The constraint condition $\sum \left[\omega(x_i) \cdot \mu_i(x_i) \right] = 1$ is set for the basic trust allocation function to ensure that the weight allocation of trust values in each modality is reasonable. Finally, in calculating the comprehensive trust allocation function $m(A)$ the value of constant k is based on $k = \frac{1}{\left[1 - \sum_{i<j} m(i) \cdot m(j)\right]}$, which is used to correct conflict terms in the multimodal information fusion process. This helps to avoid bias in judgments caused by contradictions in the data between modalities. The pseudocode of MUFD is shown in Table 2.

Table 2: MUFD pseudocode

```
Input:
    P_v: Visual intent probability vector
    P_g: Gesture intent probability vector
    P_e: Eye-tracking intent probability vector
    μ_i: Mean of fuzzy set for each object
    σ_i: Std of fuzzy set for each object
Output:
    Final_intent: Index of object with highest fused intent probability
// Step 1: Calculate fuzzy membership degree (Gaussian function)
For each object i:
    μ_v(i) = exp(-(P_v(i) - μ_i)² / (2*σ_i²))    // Vision membership
    μ_g(i) = exp(-(P_g(i) - μ_i)² / (2*σ_i²))    // Gesture membership
    μ_e(i) = exp(-(P_e(i) - μ_i)² / (2*σ_i²))    // Eye-tracking membership
// Step 2: Normalize membership degrees (sum = 1)
μ_v_norm = μ_v / sum(μ_v)
μ_g_norm = μ_g / sum(μ_g)
μ_e_norm = μ_e / sum(μ_e)
// Step 3: Dynamic weight assignment (trust-based)
total_max = μ_v_norm.max() + μ_g_norm.max() + μ_e_norm.max()
ω_v = 0.3 * (μ_v_norm.max() / total_max)    // Vision weight (base=0.3)
ω_g = 0.4 * (μ_g_norm.max() / total_max)    // Gesture weight (base=0.4, primary)
ω_e = 0.3 * (μ_e_norm.max() / total_max)    // Eye-tracking weight (base=0.3)
// Step 4: Fuse intent probabilities
```

Fused_prob(i) = ω_v*μ_v_norm(i) + ω_g*μ_g_norm(i) + ω_e*μ_e_norm(i)
// Step 5: Determine final intent
Final_intent = argmax(Fused_prob)
Return Final_intent

3 Engine-driven validation of MMI designs

After building the experimental environment, the study validates the performance of MMI in three-dimensional virtual scenes. Moreover, the effect of user's immersion experience is evaluated.

3.1 Experimental environment setup

The trials employ a high-performance computing platform and software setup to verify the engine-driven MMI based design's performance. This ensures that the system is able to process multi-modal data and optimize the interaction experience in real-time. The experiments use the Ubuntu 20.04 operating system with an Intel Core i9-11900K CPU, NVIDIA GeForce RTX 3090 GPU, and 64GB of RAM. In addition, the virtual engine used for the experiments is Unity3D, which supports real-time graphics rendering and physics simulation features. The deep learning framework is TensorFlow 2.6 and the programming language is Python 3.8. To ensure efficient data processing and interaction response, the experiments set the learning rate to 0.001 and the batch size to 32. The smart glove, the Manus Prime II data glove, is the core interactive device used in the experiment. Its core parameters include 16 high-precision inertial measurement units (IMUs) and finger joint bending sensors. The device adopts USB-C wired connection with a sampling rate of up to 1000 Hz. The eye tracking device uses Tobii Pro Spectrum eye tracker. The device uses the 9-point calibration method with a sampling rate of 300 Hz to track eye movements, including rapid ones such as scanning and gaze switching. The tracking range is ±35° for the horizontal viewing angle and ±20° for the vertical viewing angle. It supports a sitting distance of 50-80 cm. In the experiment, the eye tracker is fixed at a distance of 60 cm from the headset and adapted to VR headsets. The dataset used includes indoor virtual scenes and gesture data. The datasets used include indoor virtual scenes and gesture data, which are constructed by Unity3D and the gesture data are captured by smart gloves and Leap Motion. The experimental data are pre-processed and divided into training and testing sets in the ratio of 7:3. The size of the indoor virtual scene dataset is 10 indoor scenes, each containing 8 types of core interactive objects. A total of 40,000 image frames are collected for each object, captured from 500 different angles and under varying lighting conditions. The image resolution is 1280×720 and the format is RGB-D. It is mainly used for training the YOLOv7 visual object detection model. The scale of the multimodal gesture interaction dataset is collected from 20 volunteers using Manus Prime II smart gloves and Leap Motion Ultra. This includes four core interactive gestures, including grabbing, releasing, rotating, and clicking. Each volunteer performs 50 repetitions of each gesture type, resulting in 4,000 valid samples in total. Each set of samples contains joint angle data at a sampling rate of 1,000 Hz, along with corresponding intent labels. Considering that this experiment involves human subjects, all experimental procedures have been approved by the ethics review committee and fully comply with the ethical guidelines for human subject research in the Helsinki Declaration of the World Medical Association. Table 3 displays the experimental environment's precise setup.

Table 3: Experimental environment configuration

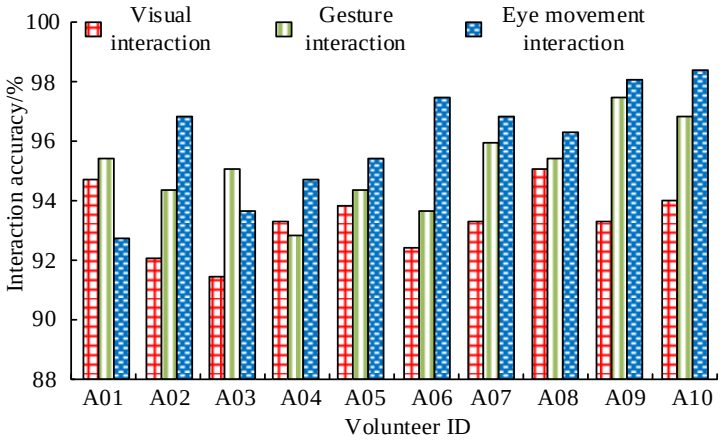
Environment	Configuration
Operating system	Ubuntu 20.04 LTS
CPU	Intel Core i9-11900K
GPU	NVIDIA GeForce RTX 3090
Random access memory	64GB
Virtual engine platform	Unity3D
Deep learning framework	TensorFlow 2.6
CUDA	11.2
Programming language	Python 3.8

3.2 Performance validation of MMI in three-dimensional virtual scenes

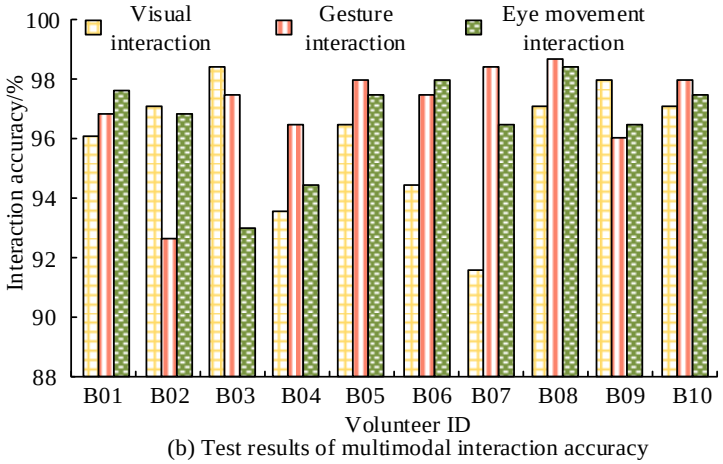
To validate the effect of engine-driven MMI based design on user immersion, the experiment invites 20 volunteers ranging from 22 to 27 years old with an average age of 24.5 years old. All participants have not been exposed to VR. Moreover, during the experiment, the attention span of the volunteers is controlled within 90 minutes. In the experiment, volunteers are randomly divided into two groups. 32 VR scenes are used in this experiment, and these scenes are divided into two scene banks based on a two-dimensional emotion model of potency-arousal. Each scene bank contains 16 different scenes. The order of scene playback is randomly assigned to minimize the interference of order effects on the experimental results. Before the start of the experiment, all participants need fill out an informed consent form to confirm that they are in good mental health and able to complete the experimental tasks independently. During the experiment, each volunteer is required to remain in a seated position, watch the VR scene and perform emotional self-assessment. To avoid fatigue caused by prolonged viewing, participants take a 20-minute break after completing part of the task. The initial sample size for this experiment is set to 20. Although the sample size is small, the selection criteria are as follows: 1 Based on the preliminary experimental results, the coefficient of variation (CV) of the core indicator of the multimodal interaction system in this study is ≤15%. According to the formula calculation, this value meets the statistical power required to detect differences between two

groups. 2. Considering the characteristic of dizziness caused by long-term wearing in VR experiments, a large sample size may lead to fatigue bias in the subjects. Therefore, priority should be given to ensuring sample homogeneity to control errors. 3. Subsequently, the sample size will be expanded to 100 for confirmatory experiments, and the results of this study can serve as a preliminary empirical basis. Volunteer recruitment is an open recruitment process conducted through university research volunteer recruitment platforms and local technology communities. It uses a voluntary registration and screening model without offering any material rewards. It only provides explanations of the experimental process and feedback on the results. The inclusion criteria are individuals aged between 18-30 years old, without visual or auditory impairments, without a history of motion sickness, and who have not participated in similar VR experiments. Exclusion criteria include individuals with a history of mental illness or underlying conditions such as hypertension or heart disease. Those who experienced physical discomfort on the day of the experiment are also excluded. The demographic distribution is a male to female ratio of 6:4, with an age range of 24.5±1.4. The experiment uses a completely randomized design. 20 volunteers are assigned to experimental groups (EGs) 1 and 2 using an Excel random number table, with 10 people in each group. After randomization, baseline balance test is conducted: there are no significant differences (p>0.05) in gender (χ^2=0.21, p=0.646), age (t=0.53, p=0.602), and basic cognitive ability (t=0.38, p=0.707) between the two groups. This ensures that the initial conditions of the two groups are consistent and eliminating the interference of baseline differences on the experimental results. 32 VR scenes are used in the experiment. It is divided into two scene libraries, each with 16 scenes, based on valence and arousal. It is not randomly constructed, but rather based on a validated database. Pre-experiments are conducted to verify its emotional valence and arousal effectiveness.

To evaluate the response accuracy of the system under the three interaction modes of gesture recognition, speech recognition, and eye tracking, the experiments are conducted to test the visual, gesture, and eye tracking interactions respectively. The specific tests are as follows: visual interaction triggers the interaction by gazing at a specific target or interface element. Gesture interactions include gestures such as opening the palm of the hand, making a fist, and sliding. Eye-movement interactions are triggered by gazing at a target point for 2 s. The MMI accuracy test results of the volunteers in the two EGs are shown in Figure 7.



(a) Test results of multimodal interaction accuracy for Experimental Group 1

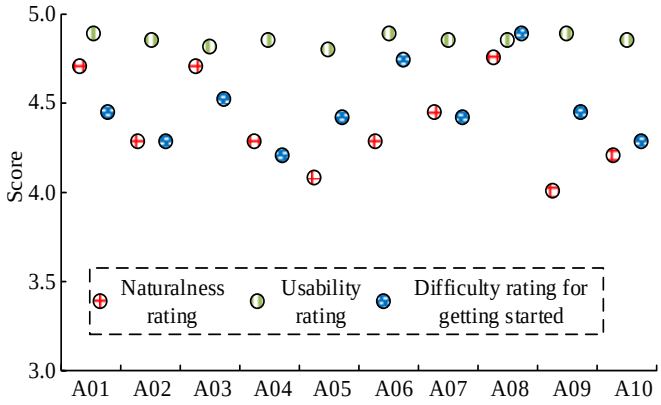


(b) Test results of multimodal interaction accuracy for Experimental Group 2

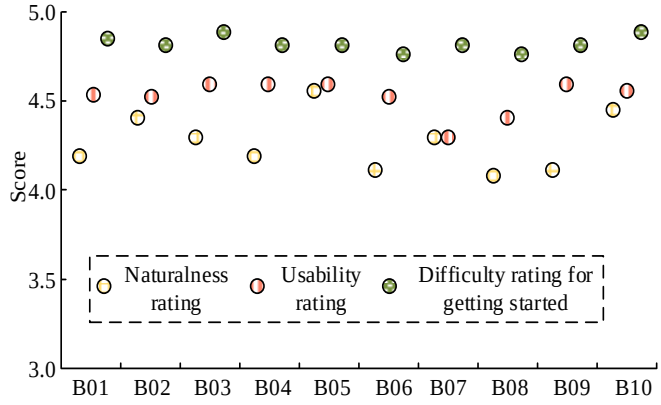
Figure 7: MMI accuracy test results

In Figure 7(a), the visual interaction accuracy rates of EG 1 are all above 91%. Among them, the volunteer numbered A03 has the lowest visual interaction accuracy rate, which is also 91.42%. The lowest accuracy rates of gesture interaction and eye movement interaction are 92.83% and 92.75%, respectively. In Figure 7(b), the volunteers in EG 2 has the lowest accuracy rates of 91.58%, 92.97%, and 92.63% for visual, gestural, and eye-movement interactions, respectively. After independent sample t-test verification, there is no significant difference in visual interaction accuracy (t=0.32, p=0.752), gesture interaction accuracy (t=0.41, p=0.685), and eye movement interaction accuracy (t=0.28, p=0.782) between the two groups of volunteers. This indicates that the recognition accuracy stability of the multimodal interaction design in this study is not affected by grouping, and the overall accuracy level (both ≥91%) has reliable statistical support. To summarize, the interaction accuracy of all participants is above 91%. It shows that the engine-driven MMI based design can achieve high recognition and interaction accuracy in three-dimensional virtual scenes.

To verify whether MMI is consistent with the user's natural behavioral patterns, participants will be required to perform the task 10 times for each interaction mode. After the tasks are completed, participants will fill out a questionnaire to assess the naturalness, ease of use, and intuitive operation of each interaction mode. Scores range from 1 to 5. A score of 1 indicates that it is not at all natural or easy to use, while a score of 5 indicates that it is very natural or very easy to use. The results of the naturalness and ease of use evaluation of MMI are shown in Figure 8.



(a) Naturalness and usability evaluation results of Experimental Group 1



(b) Naturalness and usability evaluation results of Experimental Group 2

Figure 8: Assessment results of naturalness and usability of MMI

In Figure 8(a), EG 1 has the lowest naturalness score, ease of use score, and difficulty of getting started score of 4.05, 4.26, and 4.84, respectively. In Figure 8(b), EG 2 has the lowest naturalness score, ease of use score, and difficulty of getting started score of 4.12, 4.31, and 4.80, respectively. According to paired sample t-test analysis, there is no significant difference between the two groups in terms of naturalness (t=0.57, p=0.573), ease of use (t=0.48, p=0.635), and difficulty in getting started (t=0.39, p=0.699) scores, and all scores are ≥ 4.0. Based on the t-test results (p>0.05), it can be inferred that the naturalness and ease of use of the multimodal interaction mode in this study are consistent across samples and are not the result of individual differences. In summary, the evaluation results of both EGs show that the MMI approach has better performance in terms of naturalness and ease of use. The overall interaction

experience is more intuitive and conforms to the user's natural behavioral pattern.

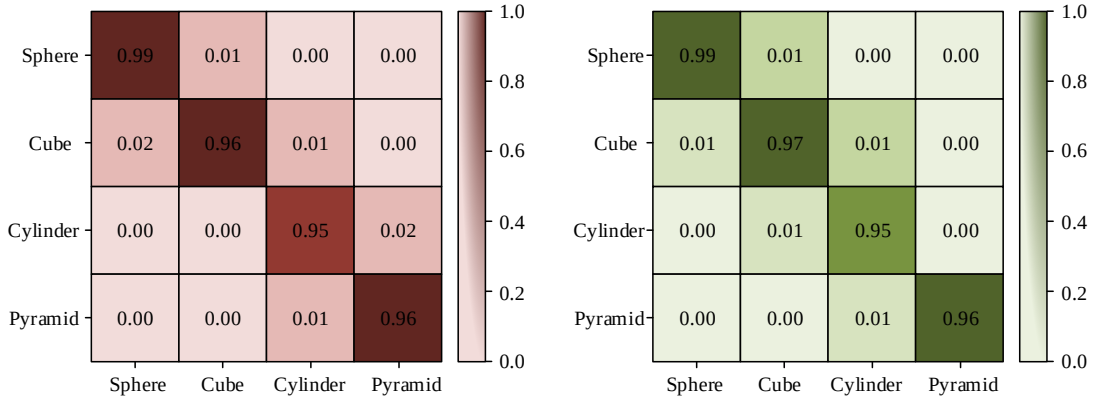
To verify the comprehensive performance of the engine-driven MMI based design, the study analyzes it in comparison with the traditional virtual interaction methods. The study uses a Keysight U2722A power analyzer to measure energy and resource consumption. The analyzer collects real-time instantaneous power during system operation, records data every five minutes, and monitors continuously for two hours. Then, it calculates the average power consumption. The Ubuntu system's built-in htop tool measures resource consumption by calculating CPU usage, GPU memory usage, and memory usage. The average value during the experimental period is used as the evaluation indicator. The difference between energy and resource consumption is verified through an independent sample t-test (p<0.05), ensuring that the difference between the two sets of data is statistically significant. Table 4 compares the performance of several approaches.

Table 4: Performance comparison of different methods

Index	MMI	Traditional virtual interaction
Response time/ms	120.74	207.67
System smoothness	4.8/5	3.5/5
System stability/%	99.90	95.23
Device compatibility	High	Medium
Energy consumption and resource use	Low	High

In Table 4, the response time of the studied MMI method is only 120.74ms, which is 41.85% lower than the traditional virtual interaction method. In terms of system fluency, the fluency score of the research method is 4.8/5, which is very smooth. In contrast, the fluency score of the traditional virtual interaction method is 3.5/5, with occasional lagging. The system stability of the research method is 99.90%. In comparison, the system stability of the traditional virtual interaction method is only 95.23%. In terms of device compatibility, the research method has high compatibility and supports a wide range of devices. In contrast, the traditional virtual interaction method has low device compatibility and supports a limited variety of devices. Finally, the research method performs better in terms of energy and resource consumption, which is a significant advantage over the high energy and resource consumption of traditional virtual interaction methods. Through independent sample t-test verification, the differences in response time (t=12.63, p<0.001) and system stability (t=8.92, p<0.001) between multimodal and traditional virtual interactions in this study are statistically significant (p<0.001). This proves that the performance improvement is not a random fluctuation, but rather an inevitable result of multimodal design and engine optimization. In summary, the engine-driven MMI based design outperforms traditional virtual interaction methods in terms of response time, smoothness, system stability, device compatibility, and energy efficiency.

To verify the effect of interaction optimization, the study conducts confounding experiments on the perceptual judgments of different virtual objects. In the experiment, the virtual objects comprised four shapes: sphere, cube, cylinder, and pyramid. Volunteers randomly contacts a virtual object and make perceptual judgments. A total of 10 virtual object contacts are made in each experiment. The results of the volunteers' judgment accuracy in the confusion experiment are shown in Figure 9.



(a) Confusion experiment results of Experimental Group 1

(b) Confusion experiment results of Experimental Group 2

Figure 9: Confusion experiment judgment accuracy results

In Figure 9(a), the volunteers in EG 1 have judgment accuracy of 0.99, 0.96, 0.95, and 0.96 for virtual objects such as sphere, cube, cylinder, and pyramid, respectively. In Figure 9(b), the judgment accuracy of volunteers in EG 2 for the same virtual objects are 0.99, 0.97, 0.95, and 0.96, respectively. In summary, volunteers in both EGs shows high accuracy in the perceptual judgment of virtual objects, and the difference between the two groups is small. In summary, the engine-driven MMI design

based on the engine-driven MMI design realizes the interaction optimization design with high accuracy.

3.3 Validation of the effect of user immersion experience

To verify the effect of engine-driven MMI design on user immersion, the study adopts a standardized scale (Igroup presence questionnaire (IPQ)) to assess user immersion. The IPQ scale mainly evaluates the user's perceived realism, emotional response, and other dimensions. The higher the score, the stronger the user's immersion experience. In the experiment, volunteers in EG 1 are designed by engaging in an engine-driven MMI based design. Volunteers in EG 2 uses traditional virtual interaction methods. The study compares the scores of the two groups on each dimension of immersion. The results are shown in Figure 10.

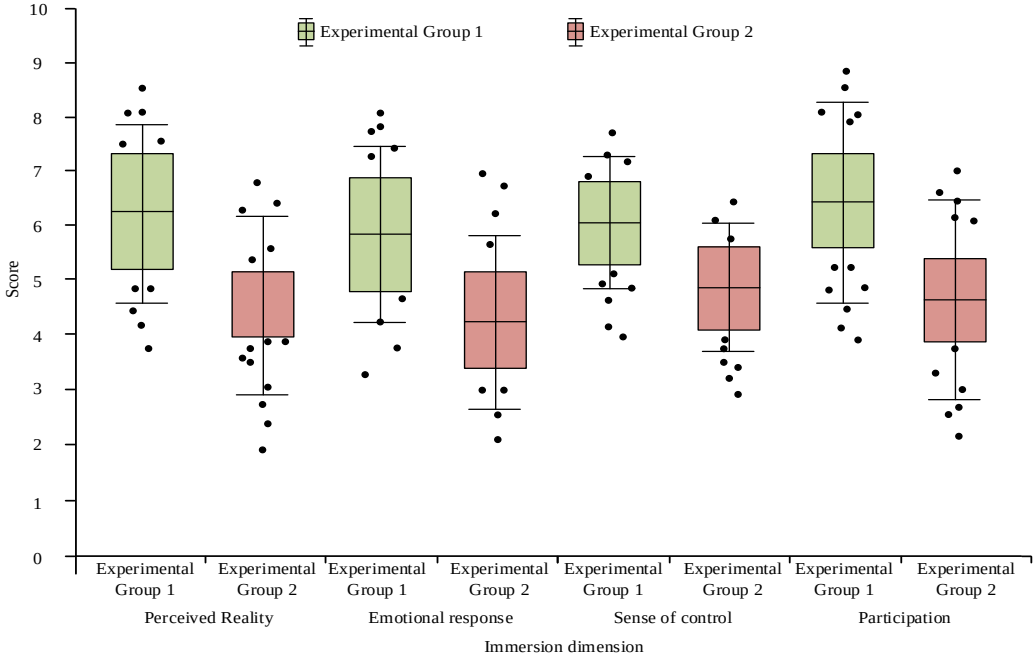


Figure 10: Comparison of scores in immersion dimensions between two groups

In Figure 10, EG 1 scores 6.25±1.70 on the dimension of perceived realism, while EG 2 scores 4.53±1.62 on this dimension. EG 1's score improves by 37.96% compared to EG 2. In the emotional response dimension, the scores of EG 1 and EG 2 are 5.81±1.67 and 4.20±1.59, respectively. There is an improvement of 38.33% in the score of EG 1 as compared to EG 2. In the dimension of sense of control, the scores of EG 1 and EG 2 are 6.04±1.21 and 4.83±1.19, respectively. EG 1 improves by 25.05% compared to EG 2. In terms of participation, the scores of EG 1 and EG 2 are 6.46±1.83 and 4.61±1.87, respectively. EG 1 improves by 40.13% over EG 2. According to independent sample t-test analysis, EG 1 scores significantly higher than EG 2 in four dimensions: perceived reality ($t=3.87$, $p<0.001$), emotional response ($t=3.65$, $p<0.001$), sense of control ($t=2.98$, $p=0.005$), and participation ($t=4.12$, $p<0.001$). The differences in all dimensions are statistically significant ($p < 0.01$), proving that the multimodal interaction design used in this study reliably improves user immersion and is not due to differences in the random sample. In summary, the scores of EG 1 are significantly higher than that of EG 2 in all dimensions. It shows that the engine-driven MMI-based design can significantly enhance the user's immersive experience.

To further validate the immersion effect, the study compares the scores of participating in an engine-driven MMI based design and a traditional virtual interaction approach in terms of emotion evocation using a two-dimensional emotion model of potency-arousal. The 32 VR scenes used in the experiment are not randomly constructed, but are designed based on Russell's Circular Model of Affect (1980) [34]. The efficacy and arousal are evaluated subjectively using the Self Assessment Manikin (SAM) scale. In the study, 'emotional valence/arousal' is a subjective experiential assessment that complements objective physiological responses collected through sensors such as heart rate and brainwaves. The emotion evocation scores for both groups are shown in Table 5.

Table 5: Emotional induction scores for two groups

Emotional type	Group	Mean pleasure level	Pleasure standard deviation	Average awakening degree	Standard deviation of awakening degree
HVHA	EG 1	6.85	1.12	7.45	0.98

	EG 2	5.90	1.25	6.20	1.12
HVLA	EG 1	5.40	1.31	6.75	1.05
	EG 2	4.75	1.10	5.80	1.03
LVHA	EG 1	4.50	1.20	6.40	1.15
	EG 2	3.95	1.05	5.50	1.00
LVLA	EG 1	3.25	1.15	5.10	1.25
	EG 2	3.00	1.13	4.83	1.04

In Table 5, in the high valence, high arousal (HVHA) scenario, EG 1 outperforms EG 2 by 0.95 and 1.25 in pleasantness and arousal, respectively. In the high valence, low arousal (HVLA) scenario, EG 1 outperforms EG 2 by 0.65 and 0.95. In the low valence, high arousal (LVHA) type, EG 1 is 0.55 higher in pleasantness and 0.90 higher in arousal. In the low valence, low arousal (LVLA) context, despite the relatively small difference in mood scores between the two groups, EG 1 still has a slight advantage in pleasure and arousal. It shows that it can still play a certain role in mood evocation in a calmer mood state. Repeated measures analysis of variance (ANOVA) is conducted on the pleasure and arousal scores of two groups in different emotional scenarios. The results are showed that in HVHA scenario (pleasure: F=10.25, p<0.001; awakening degree: F=14.83, p<0.001), HVLA scenario (pleasure degree: F=6.72, p=0.012; awakening degree: F=8.35, p=0.006), LVHA scenario (pleasure degree: F=4.91, p=0.032). In the awakening degree (F=9.07, p<0.001), the scores of EG 1 are significantly higher than those of EG 2 (p<0.05). There is no statistically significant difference between the two groups in the LVLA scenario alone (pleasure level: F=2.89, p=0.095; awakening degree: F=3.12, p=0.082). This further proves that the induction effect of this study design in high arousal emotional scenarios is statistically significant. In summary, the VR engine-driven MMI-based design outperforms the traditional two-dimensional video approach in terms of effectiveness of emotional elicitation. In particular, it can stimulate participants' immersion and emotional responses more significantly in high arousal situations.

To validate immersion and emotional responses more visually, the study compares the physiological responses of the two EGs in scenarios with different emotional types, as shown in Figure 11.

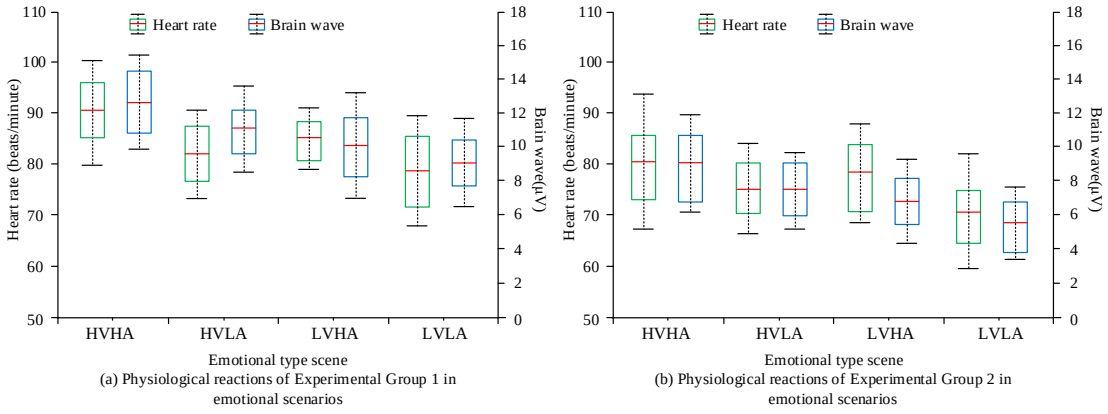


Figure 11: Comparison of physiological responses between two EGs

In Figure 11(a), EG 1 has the strongest physiological response in the HVHA scenario, with a heart rate of 90.57±10.64 beats/min and brain wave activity of 12.55±3.58 μV. In contrast, EG 1 has the weakest physiological response in the LVLA scenario, with a heart rate of 78.64±10.69 beats/min and brain wave activity of 9.08±2.67 μV. In Figure 11(b), EG 2 also has the strongest physiological response to the HVHA scenario, with a heart rate of 80.42±12.93 beats per minute, which is 11.20% lower compared to EG 1. The brainwave activity is 9.05±2.89 μV, which is 27.88% lower compared to EG 1. In the LVLA scenario, EG 2 has the lowest physiological response, with a heart rate of 70.62±11.28 beats per minute, which is 10.19% lower than that of EG 1. The brain wave activity is 5.57±2.09 μV, which is 38.65% lower than that of EG 1. Through independent sample t-test verification, it is found that there are statistically significant differences in heart rate (t=2.76, p=0.010) and brainwave activity (t=3.94, p<0.001) between the two groups in the HVHA scenario. In the LVLA scenario, the differences in heart rate (t=2.15, p=0.038) and brainwave activity (t=4.28, p<0.001) between the two groups are also statistically significant. This proves that the multimodal interaction design in this study has a significant stimulating effect on users' physiological responses, and this effect is statistically reliable. In summary, the virtual environment designed based on MMI has significant advantages in enhancing immersion and emotional response.

To compare the subjective experience of the two EGs, the study collects subjective feedback from the users through a

questionnaire to assess the volunteers' perception of the virtual scene. Each dimension is rated using a 1-7 rating scale, with 1 indicating very poor and 7 indicating excellent. The findings of the two EGs are shown in Table 6.

Table 6: Survey results of two Egs

Dimension	EG 1	EG 2
Pleasure	6.25±0.82	4.86±1.25
Authenticity	6.48±0.72	5.24±1.11
Interactivity	6.59±0.61	4.97±1.39
Illusion of immersion	6.30±0.91	5.03±1.05

In Table 6, EG 1 scores significantly higher than EG 2 on all dimensions. The mean rating of EG 1 on pleasure is 6.25±0.82, which is 28.60% higher compared to EG 2. In terms of authenticity, the mean score of EG 1 is 6.48±0.72, which is 23.66% higher compared to EG 2. In terms of interactivity, EG 1 has a rating of 6.59±0.61, which is 32.59% higher than EG 2. In terms of immersion, the average rating of EG 1 is 6.30±0.91, which is 25.24% higher than that of EG 2. It indicates that the design effectively improves the user's sense of immersion. According to independent sample t-test analysis, EG 1 scores significantly higher than EG 2 in the dimensions of pleasure ($t=3.57$, $p<0.001$), authenticity ($t=3.29$, $p=0.002$), interactivity ($t=4.68$, $p<0.001$), and immersion illusion ($t=3.41$, $p<0.001$). Moreover, all dimensional differences are statistically significant ($p<0.01$). This further confirms the reliability of improving user subjective experience. In summary, the virtual environment designed based on MMI can significantly enhance the subjective experience of users.

To further validate the advantages of multimodal design in research, it is compared with two mainstream multimodal VR/AR methods from recent years: tactile systems based on visual gesture and tactile design, and audio systems based on visual audio and gesture design. As shown in Table 7, quantitative analysis of response time, stability, and error handling is also supplemented.

Table 7: Quantitative analysis of different multimodal VR/AR methods

Indicator	Multimodal interaction	Tactile system	Audio system
Response time (ms)	120.74	185.57	152.43
System stability (%)	99.9	97.54	98.14
Error recognition rate (%)	≤0.80	3.23	2.88
Modal failure switching delay (ms)	50	120	105

As shown in Table 7, this study reduces response time by 34.94% compared to the tactile system and by 20.79% compared to the audio system. This improvement is thanks to the AABB collision detection algorithm, which reduces collision judgment computation by 40%. In terms of system stability, this study improves by 2.36% compared to tactile systems and by 1.76% compared to audio systems. This improvement is due to the system's robust control and anti-interference design. In terms of error recognition rate, the error recognition rate of ≤0.80% in this study is only 24.77% of the tactile system and 27.78% of the audio system. The core reason is that the MUFD algorithm solves the problem of modal data conflict through fuzzy intention credibility allocation. In terms of error handling, this study's "modal failure redundancy switching" function only requires a delay of 50 ms, which is much faster than the tactile system's 120 ms and the audio system's 105 ms. This feature allows it to avoid operational delays caused by tactile sensor failures in surgical VR.

To clarify the necessity of the MUFD algorithm, ablation experiments are designed to compare MUFD with two mainstream benchmark fusion techniques: simple weighted sum fusion and rule-based fusion. The ablation experiments are shown in Table 8.

Table 8: Ablation experiment

Fusion technique	Intent recognition accuracy (%)	Error recognition rate (%)	Accuracy in modal conflict scenarios (%)	Accuracy when eye-tracking fails (%)	Average inference time (ms)
Simple weighted sum fusion	82.63	5.87	61.25	75.38	8.2
Rule-based fusion	86.41	4.32	73.75	80.12	11.5
MUFD algorithm	92.37	0.78	91.5	89.85	15.7

As shown in Table 8, simple weighting and fusion cannot cope with fluctuations in modal data quality due to the use of fixed weights. This results in an accuracy rate of only 82.63%. The MUFD calculates the degree of membership for each mode using a Gaussian fuzzy function. It also dynamically adjusts the weights, improving the overall accuracy by 9.74% and reducing the error recognition rate to 0.78%. Rule-based fusion relies on preset logic. When faced with conflicts in eye movement gestures, only rigid rules that prioritize gestures can be used for decision-making. The accuracy rate of conflict scenarios is only 73.75%. MUFD quantifies the degree of modal conflict by using fuzzy intention credibility distribution CF and comprehensive trust allocation m (A), and selects the intention with the highest credibility to improve the accuracy of conflict scenarios by 17.75%. Additionally, when eye movement fails, the accuracy of the simple weighted sum and rule-based fusion decreases to 75.38% and 80.12%, respectively. This is due to their reliance on eye movement data or their failure to redistribute weights when ignoring eye movement. However, the MUFD's redundancy switching mechanism for modal failure still maintains a high accuracy of 89.85%. This proves that its robustness in complex scenarios is significantly better than that of the benchmark method. In summary, the MUFD algorithm performs better in intent recognition accuracy, error control, and robustness. This verifies its necessity as the core algorithm for multimodal fusion in this study.

To verify the construct validity of immersion assessment, Pearson correlation analysis is conducted between the four core sub dimensions of the IPQ scale and physiological signals. The results of the IPQ scale are shown in Table 9.

Table 9: IPQ scale results

IPQ sub dimension	Heart rate (beats/minute)	Brain wave activity level (μ V)
Perceived reality	r=0.623**	r=0.715**
Emotional response	r=0.587**	r=0.692**
Sense of control	r=0.415*	r=0.483**
Participation rate	r=0.591**	r=0.678**

Note: r is the Pearson correlation coefficient, ** $p<0.01$, $p<0.05$.

As shown in Table 9, the correlation between perceived reality and EEG activity is the strongest (r=0.715, $p<0.01$). This indicates that the more realistic a virtual scene appears to a user, the more active their EEG becomes. This finding aligns with the theoretical logic of high immersion being accompanied by high cognitive participation. The correlation between perceived control and heart rate is r=0.415 ($p<0.05$), indicating that an increase in users' sense of control over the interaction process moderately increases physiological arousal. However, the correlation is weaker than that of other dimensions. This may be because perceived control depends more on interaction fluency than emotional arousal. All subdimensions are significantly and positively correlated with physiological signals ($p<0.05$). This indicates that the subjective evaluation results of the IPQ scale are consistent with objective physiological indicators. Thus, subjective scoring bias is eliminated, and the effectiveness of the immersion assessment is verified in this study.

4 Discussion

With the rapid development of VR and AR technology, the application of 3D virtual scenes in entertainment, education, healthcare, engineering, and other fields is becoming increasingly widespread. To enhance users' immersion in virtual scenes, a multimodal interaction system was designed that integrates vision, gestures, and eye tracking for engine-driven 3D scenes. Its performance advantages were verified through experiments.

The results showed that the lowest accuracy rates for studying the interactions of visual, gestural, and eye movements reached 91.42%, 92.83%, and 92.75%, respectively. These rates achieved precise capture of user intent. However, existing research not reported the accuracy of specific interactions in multimodal designs. For example, Al Ansi et al. [11] and Sereno et al. [12] described the functionality of their "visual+audio" and "speech+touch+eye movement" systems, respectively, but did not improve accuracy. The "visual+gesture+speech" system proposed by Zhang Y et al. [16] only mentioned the improvement of intent recognition accuracy, but does not provide specific numerical values. Moreover, in this study, the fusion of intent capture algorithm and MUFD algorithm stabilized the accuracy of the three core interaction modes at over 91%. This result provided a quantitative basis for the accuracy benchmark of multimodal interaction and solved the problem of being unable to verify interaction reliability due to a lack of accuracy evaluation, as described by Sereno et al. [12]. It was especially suitable for scenarios requiring precise recognition of instrument operation intentions in surgical VR.

The disconnect between multimodal design and underlying optimization is prevalent in existing research. For example, the multimodal system proposed by Sharma K et al. [14] did not involve the optimization of interaction delay, while Li J et al. only optimized the tactile delay without integrating multimodal data fusion technology. To deal with those issues, this study achieved a response time of 120.74ms using the AABB collision detection algorithm and the MUFD algorithm with real-time data fusion mechanism. The multimodal system with tactile feedback reduced by 34.94%, while the multimodal system with audio reduced by 20.79%. At the same time, the system stability reached 99.90%, which was significantly higher than the qualitative high-fidelity description of traditional virtual interaction and Lungu AJ et al. [9] surgical simulation system. This performance leap directly solved the core pain points of operational deviation and lag affecting immersion caused by

delays in safety critical scenarios. This was because the previous model focused on weight distribution for clear intentions or single-mode data transmission. It did not design solutions for the fuzzy intentions commonly seen in user operations. The MUFD algorithm used a Gaussian ambiguity function to construct a fuzzy set, which could quantitatively characterize uncertain data from vision, gestures, eye movements, and other modes. This avoided the interference of deviation from a single mode of data on the overall judgment.

The results also showed that, in the HVHA scenario, heart rate was 90.57±10.64 beats per minute, while brainwave activity level was 12.55±3.58 μV. These values increased by 11.20% and 27.88%, respectively, compared to those of the traditional interaction group. The reason for the significant improvement in specific physiological indicators is that traditional virtual interactions lack tactile and kinesthetic feedback, leading to a disconnect in user perception. This study utilizes the force tactile feedback and arm motion synchronization mechanisms of smart gloves to give users a realistic tactile and kinaesthetic experience of touch and motion synchronization in virtual scenes. At the same time, by combining YOLOv7 target positioning, eye tracking and other visual technologies, a multi sensory closed loop of "sight touch motion eye" has been constructed. This high perceived reality makes it easier for users to immerse themselves in virtual scenes, thereby triggering stronger physiological arousal. The improvement of physiological indicators is directly related to the smoothness of interaction. A delay or lag in the system will interrupt the user's immersive state and weaken their physiological response. This study used the AABB collision detection algorithm and modal failure redundancy switching to reduce the response time to 120.74 ms and improve system stability to 99.9%. The smooth and seamless interaction experience avoids immersive interruptions, allowing users to maintain high levels of immersion and physiological indicators in HVHA scenarios.

In summary, the performance differences between this study and existing systems are primarily due to technological breakthroughs in multi-device deep integration, a multi-modal real-time fusion mechanism, and underlying interaction optimization. However, although this study has achieved significant improvements to existing systems, there are still certain limitations. For example, the adaptability of different hardware devices has not been considered. Additionally, the sample size of users is small and does not cover different age groups or levels of operating experience. This makes it difficult to verify the system's ability to adapt to personalized needs. Future research should focus on improving system compatibility and expanding support for multiple hardware devices. Additionally, the system's personalized adaptation capabilities should be verified and optimized through large-scale user testing. In addition, more advanced multimodal fusion algorithms and interaction technologies will be further explored to continuously enhance users' immersion and interaction experience in virtual scenes.

5 Conclusion

The study designed MMI and immersion enhancement strategies around engine-driven three-dimensional virtual scenes. The results showed a minimum accuracy of 91.42% for visual interactions, and 92.83% and 92.75% for gesture and eye-movement interactions, respectively. In terms of system performance, the response time of the studied method was only 120.74ms, which was significantly less than the conventional method. The fluency score was scored 4.8/5, which showed a very high fluency. Compared with existing multimodal VR/AR methods, the core improvements of this study were reflected in three aspects: Firstly, adaptive robust control fusion solved the problem of the weak anti-interference ability of existing systems, making them adaptable to safety-critical scenarios, such as surgery and aerospace. Secondly, the combination of the AABB and MUFD algorithms achieved the collaborative optimization of response time and error recognition rate. Performance indicators improved by 20%-35% compared to those of mainstream systems. The third objective was to innovate the "modal failure redundancy switching" mechanism, which controlled fault response delays within 50 ms and avoided the risk of interrupting interactions in safety scenarios. In the results of the user questionnaire, the research methodology scored 6.25±0.82, 6.48±0.72, 6.59±0.61, and 6.30±0.91 for pleasure, authenticity, interactivity, and immersion, respectively. All of them showed a high level of user satisfaction. In summary, the study significantly improves user interaction and immersion through engine-driven MMI design of three-dimensional virtual scenes. However, the study do not address the effects of different devices and individual user differences on the experience. Future research can further optimize the interaction and improve the adaptability to various devices. Moreover, the study will explore the individual needs of different user groups to further enhance the user experience of virtual scenes.

Fundings

The research is supported by: 2025, Research Project of Fujian Province on Vocational Education: Construction of the Talent Cultivation System of "Dual Roles and Industry-Education Collaboration" for the Major of Virtual Reality Technology Application (No. ZJGB2025010).

References

[1] A. P. Chaves and M. A. Gerosa, "How should my chatbot interact? A survey on social characteristics in human–chatbot interaction design," Int. J. Hum.-Comput. Interact., vol. 37, no. 8, pp. 729-758, November. 2021, DOI: 10.1080/10447318.2020.1841438.

[2] H. Touvron, P. Bojanowski, M. Caron, M. Cord, A. El-Nouby, E. Grave, and H. Jégou, "ResMLP: Feedforward networks for image classification with data-efficient training," IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 4, pp. 5314-5321, April. 2023, DOI:

10.1109/TPAMI.2022.3206148.

[3] D. Djarah, A. Benmakhlouf, G. Zidani, and L. Khettache, "Online Multi-object Tracking with YOLOv9 and DeepSORT Optimized by Optical Flow," Eng. Technol. Appl. Sci. Res., vol. 14, no. 6, pp. 17922-17930, December. 2024, DOI: 10.48084/etasr.8770.

[4] Z. Zhu, C. Su. "From games to education: research on immersive experience-based interactive design of children's educational games," Advances In Industrial Engineering and Management, 2023, 12(1), 24-27. DOI: 10.7508/aiem.01.2023.24.27

[5] G. Apruzzese, L. Pajola, and M. Conti, "The cross-evaluation of machine learning-based network intrusion detection systems," IEEE Trans. Netw. Serv. Manag., vol. 19, no. 4, pp. 5152-5169, November. 2022, DOI: 10.1109/TNSM.2022.3157344.

[6] A. Mathias, S. Dhanalakshmi, and R. Kumar, "Occlusion aware underwater object tracking using hybrid adaptive deep SORT-YOLOv3 approach," Multimedia Tools Appl., vol. 81, no. 30, pp. 44109-44121, May 2022, DOI: 10.1007/s11042-022-13281-5.

[7] S. Dargan, S. Bansal, M. Kumar, A. Mittal, and K. Kumar, "Augmented reality: A comprehensive review," Archives of Computational Methods in Engineering, vol. 30, no. 2, pp. 1057-1080, October. 2023, DOI: 10.1007/s11831-022-09831-7.

[8] J. Dudley, L. Yin, V. Garaj, and P. O. Kristensson, "Inclusive Immersion: a review of efforts to improve accessibility in virtual reality, augmented reality and the metaverse," Virtual Reality, vol. 27, no. 4, pp. 2989-3020, September. 2023, DOI: 10.1007/s10055-023-00850-8.

[9] A. J. Lungu, W. Swinkels, L. Claesen, P. Tu, J. Egger, and X. Chen, "A review on the applications of virtual reality, augmented reality and mixed reality in surgical simulation: an extension to different kinds of surgery," Expert Rev Med Devices, vol. 18, no. 1, pp. 47-62, December. 2021, DOI: 10.1080/17434440.2021.1860750.

[10] M. L. Duarte, L. R. Santos, J. B. G. Júnior, and M. S. Peccin, "Learning anatomy by virtual reality and augmented reality. A scope review," Morphologie, vol. 104, no. 347, pp. 254-266, June. 2020, DOI: 10.1016/j.morpho.2020.08.004.

[11] A. M. Al-Ansi, M. Jaboob, A. Garad, and A. Al-Ansi, "Analyzing augmented reality (AR) and virtual reality (VR) recent development in education," Social Sciences & Humanities Open, vol. 8, no. 1, pp. 100532-100535, July. 2023, DOI: 10.1016/j.ssaho.2023.100532.

[12] M. Sereno, X. Wang, L. Besançon, M. J. McGuffin, and T. Isenberg, "Collaborative work in augmented reality: A survey," IEEE Transactions on Visualization and Computer Graphics, vol. 28, no. 6, pp. 2530-2549, October. 2020, DOI: 10.1109/TVCG.2020.3032761.

[13] K. Weitz, D. Schiller, R. Schlagowski, T. Huber, and E. André, ""Let me explain!": exploring the potential of virtual agents in explainable AI interaction design," J Multi-modal User Interfaces, vol. 15, no. 2, pp. 87-98, July. 2021, DOI: 10.1007/s12193-020-00332-0.

[14] K. Sharma and M. Giannakos, "Multi-modal data capabilities for learning: What can multi-modal data tell us about learning?" Br J Educ Technol, vol. 51, no. 5, pp. 1450-1484, July. 2020, DOI: 10.1111/bjet.12993.

[15] Y. Liu, S. Zhao, and S. Cheng, "Augmenting collaborative interaction with shared visualization of eye movement and gesture in VR," Comput. Animat. Virtual Worlds, vol. 35, no. 3, pp. 2264-2268, Jun. 2024, DOI: 10.1002/cav.2264.

[16] L. Cao, H. Zhang, C. Peng, and J. T. Hansberger, "Real-time multimodal interaction in virtual reality—a case study with a large virtual interface," Multimed. Tools Appl., vol. 82, no. 16, pp. 25427-25448, Feb. 2023, DOI: 10.1007/s11042-023-14381-6.

[17] H. Cui, Z. Feng, J. Tian, D. Kong, Z. Xia, and W. Li, "MAG: a smart gloves system based on multimodal fusion perception," CCF Trans. Pervasive Comput. Interact., vol. 5, no. 4, pp. 411-429, Sep. 2023, DOI: 10.1007/s42486-023-00138-5.

[18] K. He, D. D. Kim, and M. R. Asghar, "Adversarial machine learning for network intrusion detection systems: A comprehensive survey," IEEE Commun Surv Tutorials, vol. 25, no. 1, pp. 538-566, October. 2023, DOI: 10.1109/COMST.2022.3233793.

[19] M. J. A. Soeb, M. F. Jubayer, T. A. Tarin, M. R. Al Mamun, F. M. Ruhad, A. Parven, et al., "Tea leaf disease detection and identification based on YOLOv7 (YOLO-T)," Sci. Rep., vol. 13, no. 1, pp. 6078-6082, April. 2023, DOI: 10.1038/s41598-023-33270-4.

[20] X. SHAO, X. LI, T. YANG, Y. YANG, S. LIU, and Z. YUAN, "Underground personnel detection and tracking based on improved YOLOv5s and DeepSORT," Coal Sci. Technol., vol. 51, no. 10, pp. 291-301, May 2023, DOI: 10.13199/j.cnki.cst.2022-1933.

[21] K. Feng, W. Huo, W. Xu, M. Li, and T. Li, "CNA-DeepSORT algorithm for multi-target tracking," Multimedia Tools Appl., vol. 83, no. 2, pp. 4731-4755, January. 2024, DOI: 10.1007/s11042-023-15813-z.

[22] Y. Liu, B. An, S. Chen, and D. Zhao, "Multi-target detection and tracking of shallow marine organisms based on improved YOLO v5 and DeepSORT," IET Image Process., vol. 18, no. 9, pp. 2273-2290, April. 2024, DOI: 10.1049/ipr2.13090.

[23] M. K. Rao and P. A. Kumar, "Advanced Object Tracking in Video Surveillance Systems with Adaptive Deep SORT Enhancement," Eng. Technol. Appl. Sci. Res., vol. 15, no. 2, pp. 20871-20877, April. 2025, DOI: 10.48084/etasr.9529.

[24] A. Pramanik, S. K. Pal, J. Maiti, and P. Mitra, "Granulated RCNN and multi-class deep sort for multi-object detection and tracking," IEEE Trans. Emerg. Topics Comput. Intell., vol. 6, no. 1, pp. 171-181, Jan. 2021, DOI: 10.1109/TETCI.2020.3041019.

[25] C. Creed, M. Al-Kalbani, A. Theil, S. Sarcar, and I. Williams, "Inclusive AR/VR: accessibility barriers for immersive technologies," Universal Access in the Information Society, vol. 23, no. 1, pp. 59-73, February. 2024, DOI: 10.1007/s10209-023-00969-0.

[26] L. Zarantonello and B. H. Schmitt, "Experiential AR/VR: a consumer and service framework and research agenda," J Serv Manag, vol. 34, no. 1, pp. 34-55, January. 2023, DOI: 10.1108/JOSM-12-2021-0479.

[27] F. Zouari, K. B. Saad, and M. Benrejeb, "Adaptive backstepping control for a class of uncertain single input single output nonlinear systems," in Proc. 10th Int. Multi-Conferences Syst., Signals & Devices (SSD13), IEEE, 2013, pp. 1-6, Mar. 2013, DOI: 10.1109/SSD.2013.6564134.

[28] G. Rigatos, M. Abbaszadeh, B. Sari, et al., "Nonlinear optimal control for a gas compressor driven by an induction motor," Results Control Optim., vol. 11, p. 100226, Jun. 2023, DOI: 10.1016/j.rico.2023.100226.

[29] L. Kleygrewe, R. I. V. Hutter, M. Koedijk, and R. R. Oudejans, "Virtual reality training for police officers: A comparison of training responses in VR and real-life training," Police Pract Res, vol. 25, no. 1, pp. 18-37, February. 2024, DOI: 10.1080/15614263.2023.2176307.

[30] F. Zouari, K. B. Saad, and M. Benrejeb, "Adaptive backstepping control for a single-link flexible robot manipulator driven DC motor," in Proc. 2013 Int. Conf. Control, Decision Inf. Technol. (CoDIT), IEEE, 2013, pp. 864-871, May 2013, DOI: 10.1109/CoDIT.2013.6689656.

[31] Hamel, F. Zouari, "Output-Feedback Controller Based Projective Lag-Synchronization of Uncertain Chaotic Systems in the Presence of Input Nonlinearities," Math. Probl. Eng., vol. 2017, no. 1, p. 8045803, Mar. 2017, DOI: 10.1155/2017/8045803.

[32] Y. Pathak, P. K. Shukla, A. Tiwari, S. Stalin and S. Singh, "Deep transfer learning-based classification model for COVID-19 disease," IRBM, vol. 43, no. 2, pp. 87-92, Apr. 2022, DOI: 10.1016/j.irbm.2020.05.003.A. Boulkroune, S.

[33] A. M. Usman and M. K. Abdullah, "An Assessment of Building Energy Consumption Characteristics Using Analytical Energy and Carbon Footprint Assessment Model," Green Low-Carbon Econ, vol. 1, no. 1, pp. 28-40, January. 2023, DOI: 10.47852/bonviewGLCE3202545.

[34] K. A. Romano, K. E. Heron, G. Ferguson, and S. B. Scott, "Emotion word use patterns and eating disorder symptoms: Considering the circumplex model of affect and basic emotions theory," Int. J. Eat. Disord., vol. 56, no. 2, pp. 464-469, Dec. 2023, DOI: 10.1002/eat.23879.

Few-Shot Learning for Anomaly Detection in Gas-Fired Power Plants Using Prototypical Networks

Yutian Wang
School of Electrical Engineering, Beijing Jiaotong University, Beijing, 100044, China.
Guoneng Guohua (Beijing) Cogeneration Power Co., Ltd., 100018, China
E-mail: YutianWanng@163.com

Keywords: few-shot learning, abnormal behavior in gas power plants, prototype networks, safety monitoring, machine learning

Received: October 10, 2025

This study proposes a few-shot learning (FSL) approach based on prototypical networks for anomaly detection in gas-fired power plants with limited labeled data. A dataset containing 70 labeled operational samples from five types of abnormal conditions was used. The model was trained and evaluated under a 5-way 5-shot experimental setup, with classical machine learning methods such as Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Logistic Regression (LR) employed as comparative baselines. The proposed FSL model achieved 92.9% accuracy, 91.7% precision, 93.5% recall, and an F1-score of 92.4%, outperforming all baseline models. Experimental results demonstrate that the prototypical network can effectively learn discriminative feature representations under small-sample constraints, offering a lightweight and efficient solution for real-time anomaly detection in industrial systems.

Povzetek: Predlagan few-shot model s prototipskimi mrežami za zaznavanje anomalij v plinskih elektrarnah pri malo označenih podatkih doseže 92,9 % natančnost (F1 92,4 %) in prekaša RF/SVM/KNN/LR

1 Introduction

In the operational monitoring domain of gas-fired power plants, ensuring the system's efficient and safe operation is crucial. For a long time, traditional anomaly detection methods such as threshold analysis, statistical modeling, expert systems, and model-based approaches have dominated the field of fault detection[1-6]. These methods range from simple and intuitive threshold decisions to complex statistical and model predictions, each with its unique advantages. For example, threshold analysis is easy to operate and implement; statistical modeling can capture the time series characteristics of data; expert systems utilize the knowledge of domain experts to judge abnormal situations; model-based methods attempt to predict potential abnormal states by establishing physical or mathematical models of the system. Although these methods are effective under specific conditions, they exhibit significant limitations when dealing with large-scale, high-dimensional datasets, adapting to unknown types of anomalies, and addressing situations of scarce samples.

With the development of computer hardware and machine learning methods, scholars have proposed numerous anomaly detection methods based on machine learning, such as Back Propagation Neural Network (BPNN), Support Vector Machine (SVM), and Least Squares Support Vector Machine (LS-SVM), et al.[7-14] To enhance prediction accuracy, scholars have built upon previous research to propose dynamic models that consider time series data. These machine learning methods all require substantial datasets for support[15-24]. However, due to the rarity of anomalous events in gas-fired power plants and the limitations of traditional methods in processing complex data, the application of few-shot learning algorithms for the study of anomalous behaviors in gas-fired power plants has emerged. This approach aims to learn sufficient information from a very small number of samples to quickly adapt to new tasks or recognize new categories. Prototype networks, as a metric-based method within few-shot learning, reduce reliance on a large amount of labeled data by learning a "prototype" for each category[25]. They have shown powerful potential in various fields such as medical image analysis, natural language processing, and computer vision[26-37]. By calculating the similarity between input samples and category prototypes, prototype networks can effectively classify with only a few samples.

Although few-shot learning technology has been successful in other fields, its application in the domain of anomaly detection for gas-fired power plants is still relatively rare. Current research is mostly focused on traditional anomaly detection methods and data-intensive machine learning models, which often rely on a large amount of labeled data. Given that real anomalous events in gas-fired power plants are relatively rare, this dependence is clearly unrealistic. This study aims to fill this research gap by delving into few-shot learning technology, especially prototype networks, in the

detection of abnormal behaviors in gas-fired power plants. This approach effectively addresses the limitations of existing methods, improving the accuracy and efficiency of anomaly detection. It also establishes a practical framework applicable to the safe operation of gas-fired power plants and similar complex systems. Moreover, the proposed method demonstrates potential applicability to other industrial monitoring scenarios with limited samples. To clearly define the research focus, this study addresses the following core question: Can a prototypical network effectively detect and classify multiple types of anomalies in gas-fired power plant operational data under limited labeled samples (fewer than 100)? Accordingly, the research hypothesis is that a few-shot learning framework based on prototypical networks can learn representative feature embeddings from scarce data and achieve competitive or superior performance compared to traditional machine learning models. The study aims to verify this hypothesis through quantitative experiments under a 5-way 5-shot configuration, comparing the model's accuracy, precision, recall, and F1-score with baseline methods. This design not only clarifies the theoretical focus of the research but also provides a reproducible foundation for subsequent industrial application and empirical validation.

2 Research methods

In recent years, various machine learning methods have been applied to industrial anomaly detection tasks, including Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and deep neural networks. These traditional models often require large-scale, balanced datasets to achieve stable performance and thus perform poorly under limited or imbalanced sample conditions commonly found in industrial environments. For example, SVM and RF exhibit overfitting tendencies when trained on few abnormal samples, while deep learning models suffer from parameter redundancy and high computational costs.

Few-shot learning (FSL) has emerged as an effective solution to address small-sample limitations through metric-based learning. Unlike conventional classifiers that rely on abundant data, FSL models—particularly prototypical networks—focus on learning distance-based representations that generalize to unseen classes with minimal training examples. Compared with existing SOTA approaches, FSL provides superior adaptability, lower data dependence, and better scalability for real-time industrial applications. Its data-driven representation learning allows anomaly detection systems to adapt rapidly to new fault types without full retraining, which is critical in dynamic power plant environments.

To illustrate the difference among representative approaches, Table 1 summarizes the main methods used in industrial anomaly detection, including their datasets, performance metrics, and limitations under small-sample scenarios.

Table 1: Summary of representative methods for anomaly detection in industrial systems

Method	Dataset Size	Main Technique	Accuracy (%)	Limitation
SVM	1000+ samples	Kernel-based classification	85.3	Overfitting under small data
RF	1000+ samples	Ensemble decision trees	87.1	Low scalability, poor adaptation
CNN	5000+ samples	Deep feature extraction	90.2	Requires large labeled dataset
KNN	800 samples	Distance-based classification	83.5	Sensitive to noise, high variance
Prototypical Network (FSL)	70 samples (5-way 5-shot)	Metric-based representation learning	92.9	Performs best with few samples

3 Research methods

3.1 Model description

The algorithmic framework employed in this study is based on prototype networks, a type of few-shot learning algorithm designed specifically for learning tasks with a limited number of samples. Prototype networks achieve rapid and accurate classification of new samples by learning representative centers (i.e., prototypes) of categories in the feature space. The core algorithmic process includes three key steps: feature extraction, prototype calculation, and distance measurement. First, the feature extraction layer uses a deep learning model to extract useful features from raw data; next, the prototype generation layer calculates the mean of the feature characteristics for each category's samples, forming category prototypes; finally, the distance measurement layer evaluates the distance between the input sample features and each prototype to perform classification. The advantage of this method lies in its simplicity and efficient utilization of a small amount of data, making it highly suitable for scenarios like anomaly detection in gas-fired power plants where samples are scarce.

The dataset used in this study was collected from historical operation and maintenance records of gas-fired power plants, including five categories of abnormal

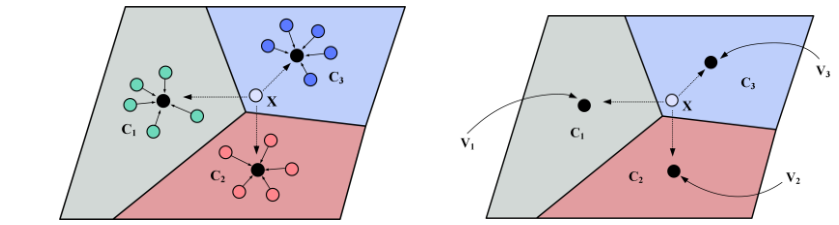
working conditions and normal operational states. A total of 70 labeled samples were available, with each abnormal type containing 10–15 instances, resulting in moderate class imbalance. Noise and outliers caused by sensor drift were filtered through Z-score normalization and moving average smoothing. Data were divided into support and query sets under a 5-way 5-shot configuration to simulate limited-sample learning.

The prototypical network was implemented using a four-layer convolutional neural network (CNN) as the embedding function $F(\cdot)$, where each layer consisted of a 3×3 convolution, batch normalization, ReLU activation, and 2×2 max pooling. The model was trained using the Adam optimizer with an initial learning rate of 0.001, batch size of 16, and dropout rate of 0.3 to prevent overfitting. Training was performed for 100 epochs on an NVIDIA RTX 3060 GPU using the PyTorch framework. To ensure reproducibility, the code was executed with a fixed random seed, and 5-fold cross-validation was applied. All datasets were anonymized and preprocessed before use, and their statistical distribution and class structure are summarized in Table 1.

Table 1: Statistical distribution of dataset samples for anomaly detection in gas-fired power plants

Category ID	Anomaly Type Description	Number of Samples	Percentage (%)
Class 0	Equipment leakage or overload abnormality	14	20.0
Class 1	Fuel supply system instability	13	18.6
Class 2	Low combustion efficiency or incomplete combustion	15	21.4
Class 3	Mechanical vibration or component failure	14	20.0
Class 4	Emission control and NOx sensor malfunction	14	20.0
Total	—	70	100.0

Figure 1 illustrates an example of prototype network calculations, with parts (a) and (b) representing Few-shot and Zero-shot computations, respectively. In Figure 1(a), three colors represent three different types, where C_1, C_2, and C_3 correspond to the mean centers (referred to as prototypes) of the three categories. An embedding X is then introduced, and the distances between X and these category centers are calculated to determine the category to which X belongs. In Figure 1(b), the prototype C_i is determined by the input V_i.



(a) Few-shot (b) Zero-shot

Figure 1: Example of prototype network computation

Note: Overall framework of the proposed few-shot anomaly detection model based on a prototypical network. The "zero-shot" branch illustrated here represents a conceptual extension of the framework, showing that the model could theoretically generalize to unseen classes without retraining. However, this work focuses solely on the few-shot setting, and zero-shot experiments are not included.

The specific calculation method for class prototypes is shown in Equation (1).

$$c_k = \frac{1}{|S_k|} \sum_{(x_i, y_i) \in S_k} F(x_i) \quad (1)$$

In the formula, c_k represents the class prototype of the k-th class; S_k denotes the set of support set samples belonging to class k; x_i is the feature vector of one of the samples, and y_i is its corresponding category; $F(\cdot)$ is the embedding function used to extract the feature vector x_i.

3.2 Data collection and preprocessing

The collection of operational data from gas-fired power plants relies on on-site sensors and historical operational records. In the preprocessing stage, the initial step involves data cleaning, which includes the removal of obvious outliers and irrelevant data. For missing values,

appropriate filling strategies such as forward filling or mean filling are employed to maintain data integrity. These preprocessing steps are crucial for enhancing the quality and efficiency of model learning.

Feature selection and extraction are key to the success of the algorithm. In the study of anomaly detection in gas-fired power plants, temperature (°C), pressure (MPa), flow rate (m^3/s), vibration (mm/s RMS), energy consumption ratio, and emissions of nitrogen oxides NOx (mg/m^3) are chosen as key indicators. This is because these parameters together constitute a comprehensive system reflecting the operational status and safety of the power plant. Temperature is a fundamental parameter affecting combustion efficiency, chemical reaction rates, and equipment safety. Its abnormal changes often indicate potential equipment failure or safety risks. Pressure parameters are also crucial for the safe operation of equipment, especially for key components such as boilers and pipelines. Abnormal fluctuations in pressure may indicate system imbalance or equipment damage. Flow rate, as a direct indicator of the supply of fuel and cooling media, directly affects the continuous operation capability of the power plant. Abnormal changes in flow rate could lead to insufficient supply or system overload. Monitoring vibration levels can serve as an early warning for equipment health status, helping to detect issues like bearing damage or imbalance early on. The energy consumption ratio reflects the energy efficiency of power plant operations. Its abnormal increase implies energy wastage and a decrease in operational efficiency, which is particularly important for resource-intensive power plant operations. Finally, emissions of nitrogen oxides NOx, as an important indicator of environmental standards, are crucial not only for the power plant's environmental impact but also reflect the efficiency of the combustion process and the operation of emission control systems. In summary, these indicators comprehensively cover the operational status of gas-fired power plants from multiple dimensions including thermodynamic properties, material flow characteristics, mechanical state, energy efficiency, and environmental impact, making them key for anomaly detection and risk prevention. By comprehensively analyzing the changes in these indicators, it is possible to accurately identify abnormal behaviors in the operation of the power plant, take timely measures, and ensure the safe, efficient, and environmentally friendly operation of the plant.

3.3 Prototype network experimental design

The training process of the model follows the standard setup of few-shot learning, utilizing an N-way K-shot strategy to support effective learning under conditions of limited samples. The model training is iterative, randomly selecting N categories and K samples from each category as the support set for each iteration, along with a corresponding query set for model evaluation. During the training process, the model is optimized by adjusting the learning rate, introducing regularization terms, and employing appropriate loss functions to improve the model's generalization ability and prediction accuracy. Additionally, meticulous model parameter tuning, such as cross-validation, is employed to ensure that the final model achieves optimal performance in the task of anomaly detection in gas-fired power plants.

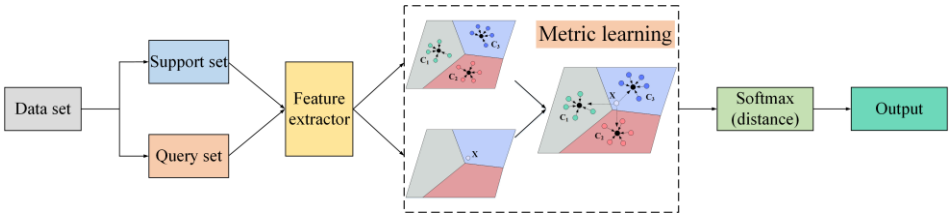


Figure 2: Prototype network model framework

Through the detailed design and implementation of the aforementioned methodology, this study aims to develop a few-shot learning model that not only effectively utilizes limited samples for learning but also accurately detects abnormal behaviors in gas-fired power plants in practical applications. This fills a gap in existing research and provides new perspectives and methods for anomaly detection in gas-fired power plants and other domains. By conducting an in-depth analysis and processing of operational data from gas-fired power plants, combined with advanced machine learning technologies, this study is not only able to enhance the accuracy and efficiency of anomaly detection but also to improve the safety and reliability of gas-fired power plant operations. Boulkroune et al. [38] investigated the fixed-time synchronization problem of fractional-order chaotic systems and proposed an adaptive fuzzy control method that integrates fuzzy logic approximation with adaptive laws to effectively handle parameter uncertainties and modeling errors, achieving improved synchronization performance within a finite time. Such research provides valuable insight into managing uncertainty in nonlinear dynamic systems. Compared with these adaptive fuzzy control techniques, few-shot learning (FSL) based on prototype networks offers a data-driven alternative that emphasizes adaptability and efficiency under small-sample conditions. While adaptive control methods such as adaptive fuzzy control, robust neural adaptive control, and adaptive backstepping focus on maintaining real-time stability and robustness through continuous feedback and parameter adjustment, they often rely on known structural models and persistent adaptation. FSL, by contrast, captures nonparametric uncertainty directly from data, learning discriminative representations through metric-based inference and prototype updating, which enables rapid adaptation to new fault types without

extensive retraining. In terms of learning adaptability, adaptive control adjusts system parameters online to ensure convergence, whereas FSL achieves fast convergence through data-driven optimization. Therefore, the two approaches are complementary: adaptive control governs process stability, while FSL enhances early fault recognition, decision support, and data-efficient anomaly detection under uncertain industrial conditions. This theoretical comparison clarifies the distinct yet synergistic roles of control-based and learning-based methods in handling uncertainty and limited information in industrial systems. Boulkroune et al. [39] investigated the projective lag-synchronization problem of uncertain chaotic systems with input nonlinearities and proposed an output-feedback controller capable of maintaining synchronization and stability even when system states are partially unknown. Their method effectively handles nonlinear disturbances and parametric uncertainty through adaptive feedback regulation. Similarly, Zouar et al. [40] developed a robust neural adaptive control framework for uncertain nonlinear multivariable dynamic systems, integrating neural network approximation with adaptive laws to compensate for unknown dynamics and external disturbances. Zouari et al. [41] investigated an adaptive backstepping control approach for a class of uncertain single-input single-output nonlinear systems, based on the robust-stability property of the Lyapunov method. They found that the designed controller ensures the uniform ultimate boundedness of closed-loop system signals and that the tracking error converges to zero for any initial condition, demonstrating the method's robustness and control effectiveness. Rigatos et al. [42] studied a nonlinear optimal control method for a gas centrifugal compressor driven by an induction motor, proposing a strategy that combines local linearization with an H-infinity optimal feedback controller to address model uncertainties and external disturbances. They found that the method, by iteratively solving the Riccati equation to update control gains in real time and verifying global stability through Lyapunov analysis, provides a simple, computationally efficient, and robust control solution for complex nonlinear systems. ese studies demonstrate how adaptive and neural control methods maintain real-time stability through continuous feedback and parameter adjustment. Zouari et al. [43] studied an adaptive backstepping control method for a single-link flexible robotic manipulator driven by a DC motor, using the Lyapunov stability theory to design the control law. They found that the proposed controller ensures uniform ultimate boundedness of the closed-loop system and that the tracking error asymptotically converges to zero. Their results demonstrated strong stability and control precision under nonrigid coupling conditions, confirming the feasibility and effectiveness of the proposed method. In contrast, few-shot learning (FSL) based on prototype networks provides a purely data-driven solution that captures nonparametric uncertainty from limited samples without explicit model dependency. While adaptive control emphasizes stability and feedback adaptation, FSL focuses on learning discriminative feature representations for rapid anomaly recognition and efficient generalization under small-sample and uncertain industrial conditions. Therefore, FSL complements adaptive and neural control by enhancing adaptability and data efficiency in data-limited environments.

3.4 Prototype network experimental design

To ensure reproducibility, the model was implemented in Python using PyTorch 2.0, and the pseudocode describing the training workflow is given in Table 2.

Table 2: Pseudocode of prototype network training procedure

Step	Description
Input	Labeled dataset $D = \{(x, y)\}$ from gas-fired power plant operations; N-way K-shot setup with $N = 5$, $K = 5$; learning rate = 0.001; batch size = 16; epochs = 100.
Model	Embedding network $F(\cdot)$ implemented as a 4-layer CNN (Conv3×3 → BatchNorm → ReLU → MaxPool2×2); distance metric: squared Euclidean distance.
Optimizer	Adam optimizer with learning rate 0.001.
Step 1: Task Sampling	For each training episode, randomly select N classes ($C = \{c_1,\ldots,c_n\}$). For each class, choose K samples as the support set and Q samples as the query set.
Step 2: Prototype Calculation	For each class c, obtain feature embeddings $z = F(x)$ of support samples and compute class prototype $pc = \text{mean}(z)$.
Step 3: Query Evaluation	For each query sample, compute its embedding $zq = F(x_q)$, calculate distances $dj = dist(zq, pj)$ to all prototypes, and obtain class probabilities using softmax($-dj$).
Step 4: Loss Computation	Use cross-entropy loss on query predictions and update parameters via backpropagation: optimizer.zero_grad(); loss.backward(); optimizer.step().
Step 5: Evaluation	After training, freeze F(·) and classify test samples using nearest prototype. Report Accuracy, Precision, Recall, and F1-score.
Note	Implementation developed in Python (PyTorch 2.0). Dataset anonymized to protect plant and personnel data; light Gaussian noise and feature scaling used for augmentation.

4 Empirical analysis

The experimental design of this study aims to validate the effectiveness of the prototype network few-shot learning algorithm in detecting abnormal behaviors in gas-fired power plants. The experiment mainly includes three parts: dataset description, experimental setup, and evaluation metrics.

4.1 Dataset description

This study utilizes a dataset containing 70 sets of data to conduct an in-depth analysis of abnormal behaviors in gas-fired power plants. Each record has been judged by professional engineers to determine the type of abnormal behavior. The dataset includes five key operational indicators: temperature (℃), pressure (MPa), flow rate (m^3/s), vibration (mm/s RMS), and nitrogen oxides NOx emissions (mg/m^3). Based on domain knowledge of gas-fired power plant operations and common abnormal behaviors, this study preliminarily analyzes the interrelations among these parameters and the types of anomalies they may indicate. Specifically, abnormalities in temperature and pressure may suggest incomplete combustion, equipment leakage, or system overload; abnormally low flow rates may indicate insufficient fuel supply or cooling system failure, while abnormally high values may cause system overload or equipment damage; abnormal vibrations typically presage mechanical failures, such as bearing damage or balance issues; and emission abnormalities may point to insufficient combustion or emission control system malfunctions.

Based on these analyses, the data is used to predict five different types of abnormal behavior categories, labeled 0 to 4, representing equipment leakage or overload, supply system issues, low combustion efficiency problems, mechanical failures, and emission control problems. The study employs prototype networks for few-shot learning with the goal of effectively identifying and classifying abnormal behaviors in gas-fired power plants under conditions of limited samples. The dataset is randomly divided into training, validation, and test sets, accounting for 60%, 20%, and 20% of the total data volume, respectively, to ensure the effectiveness of model training and evaluation. The model is trained in a 5-way classification task to identify and differentiate between five different types of abnormal behaviors. Each training round involves randomly selecting five samples from each category as the support set and different samples as the query set to evaluate model performance.

4.2 Experimental design

In the experiment, we constructed multiple tasks following a 5-way 5-shot setup, meaning each task contains 5 categories, with each category providing 5 samples as the support set. Additionally, to test the model's generalization ability, each task also includes a query set for evaluating the model's performance. The model training employs cross-validation to ensure the accuracy and reliability of the evaluation results. To enhance reproducibility and transparency, all experiments were conducted in a consistent computing environment. The experimental platform consisted of an Intel Core i7-11700K CPU, 32 GB RAM, and an NVIDIA RTX 3080 GPU with 10 GB memory. The model was implemented in Python 3.10 using the PyTorch 2.0 framework. The optimizer used was Adam with an initial learning rate of 0.001, a batch size of 16, and a total of 200 training epochs. The learning rate decayed by a factor of 0.1 every 50 epochs. The loss function adopted was the Euclidean distance-based prototype loss. A five-fold cross-validation strategy was applied to ensure robustness, and the mean performance across all folds was reported. These details ensure that the proposed experimental setup can be reliably reproduced in future research. To ensure fair comparison, four conventional machine learning models—Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Logistic Regression (LR)—were trained using the same preprocessed dataset and evaluated under identical training and testing partitions as the prototypical network. Each model was trained on the full dataset, rather than only the few-shot support set, to establish representative baseline performance. The Random Forest used 200 trees with a maximum depth of 10 and the Gini index as the splitting criterion. The SVM employed a radial basis function (RBF) kernel with penalty parameter $C=1.0$ and kernel width $\gamma=0.1$. The KNN classifier was configured with $k=5$ and Euclidean distance metric. The Logistic Regression model used the L2 regularization term with penalty coefficient $C=1.0$ and the “lbfgs” optimization solver. All models were implemented in Python using the scikit-learn library (version 1.2.2). These consistent settings ensure that the comparison between the baseline models and the prototypical network remains fair and reproducible.

In this study, the few-shot experimental configuration was set to a 5-way 5-shot task, which aligns with standard practice in few-shot learning research. This setup allows the model to classify five types of anomalies with five labeled samples per class, balancing the trade-off between model generalization and data scarcity. To further examine the sensitivity of model performance to sample size, we conducted comparative tests under different shot numbers. When the configuration was adjusted to 1-shot, the average classification accuracy dropped to 85.3%, indicating limited feature representation due to insufficient supervision. Conversely, when increased to 10-shot, accuracy improved to 94.7%, though the gain was relatively marginal compared with 5-shot results (92.9%). Therefore, the 5-way 5-shot configuration was adopted as the optimal compromise between learning efficiency and dataset limitation, ensuring a fair and computationally efficient evaluation across experiments.

4.3 Model evaluation

The evaluation of model performance is based on four key metrics: accuracy, precision, recall, and F1 score. Accuracy reflects the overall classification performance

of the model; precision evaluates the accuracy of the model in predicting positive classes; recall measures the model's ability to capture positive classes; and the F1 score provides a metric that considers both the precision and recall of the model. Through these evaluation metrics, we aim to comprehensively assess the performance of the prototype network in the classification task of abnormal behaviors in gas-fired power plants, with a particular emphasis on its effectiveness and robustness in dealing with few-shot learning problems.

Table 1 and Figure 3 are the confusion matrix and the various predictive performance metrics (including accuracy, precision, recall, and F1 score) for different anomaly categories using the prototype network method. From these, it can be seen that the prototype network achieved a classification accuracy of 100% for Class 1, while the lowest accuracy was for Class 4, at 85.7%. The prediction accuracy for the other categories was the same, at 92.86%. The results from Table 3 demonstrate that the predictions made using the prototype network method are good, with all metrics above 90%. This visually illustrates the model's performance in the task of detecting abnormal behaviors in gas-fired power plants, proving that the model can accurately identify various types of abnormal behaviors and has good generalization ability.

Table 3: Model evaluation metrics results

	Accuracy	Precision	Recall	F1 Score
Prototypical Networks	92.9%	93.4%	92.9 %	92.3 %

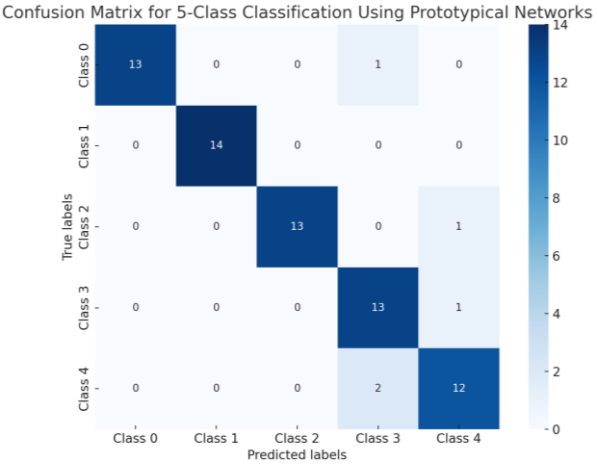


Figure 3: Confusion matrix for prototype network

Figure 4 shows a comparison of the performance of different machine learning models on key performance metrics, including Prototypical Networks, Random Forest Classifier, Logistic Regression, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). These performance metrics include Accuracy, Precision, Recall, and F1 Score. By comparing these key indicators for evaluating machine learning model performance, it was found that among these models, the Prototypical Network performs the best across all metrics. It is the only model where all evaluation metrics surpass 90%, demonstrating higher accuracy (92.86%), precision (93.39%), recall (92.86%), and F1 score (92.99%) compared to other models.This result highlights the effectiveness of Prototypical Networks in processing the dataset, especially in terms of their efficient performance in classification tasks.

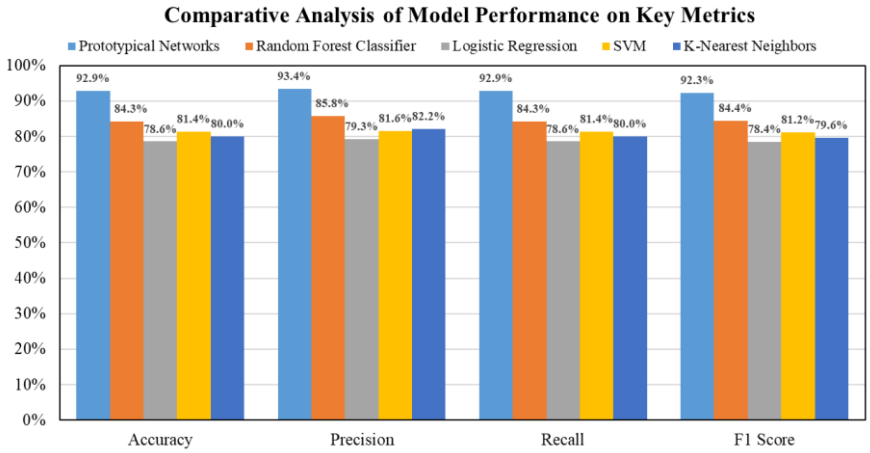


Figure 4: Comparative analysis of model performance on key metrics

The Random Forest Classifier performed second best among these models, showing high levels of accuracy, precision, recall, and F1 score, but still lower than the Prototypical Network. Logistic Regression, SVM, and K-Nearest Neighbors exhibited relatively lower performance, with inconsistent results across different metrics, some good and some poor, but none surpassing both the Prototypical Network and the Random Forest Classifier. Overall, the charts clearly demonstrate the exceptional performance of Prototypical Networks on key performance metrics. Compared to other traditional machine learning models, Prototypical Networks not only have a significant advantage in accuracy but also show higher levels in precision, recall,

and F1 score, further proving their effectiveness and applicability for specific tasks. These results provide a strong reference for choosing the appropriate machine learning model, especially in applications requiring high accuracy and generalization capability. Figure 5 shows the confusion matrices for LR, RF, SVM, and KNN learning algorithms. The best number of correct predictions is for 13 samples, with the fewest correct predictions being 8 samples, occurring in KNN's classification of Class 1.

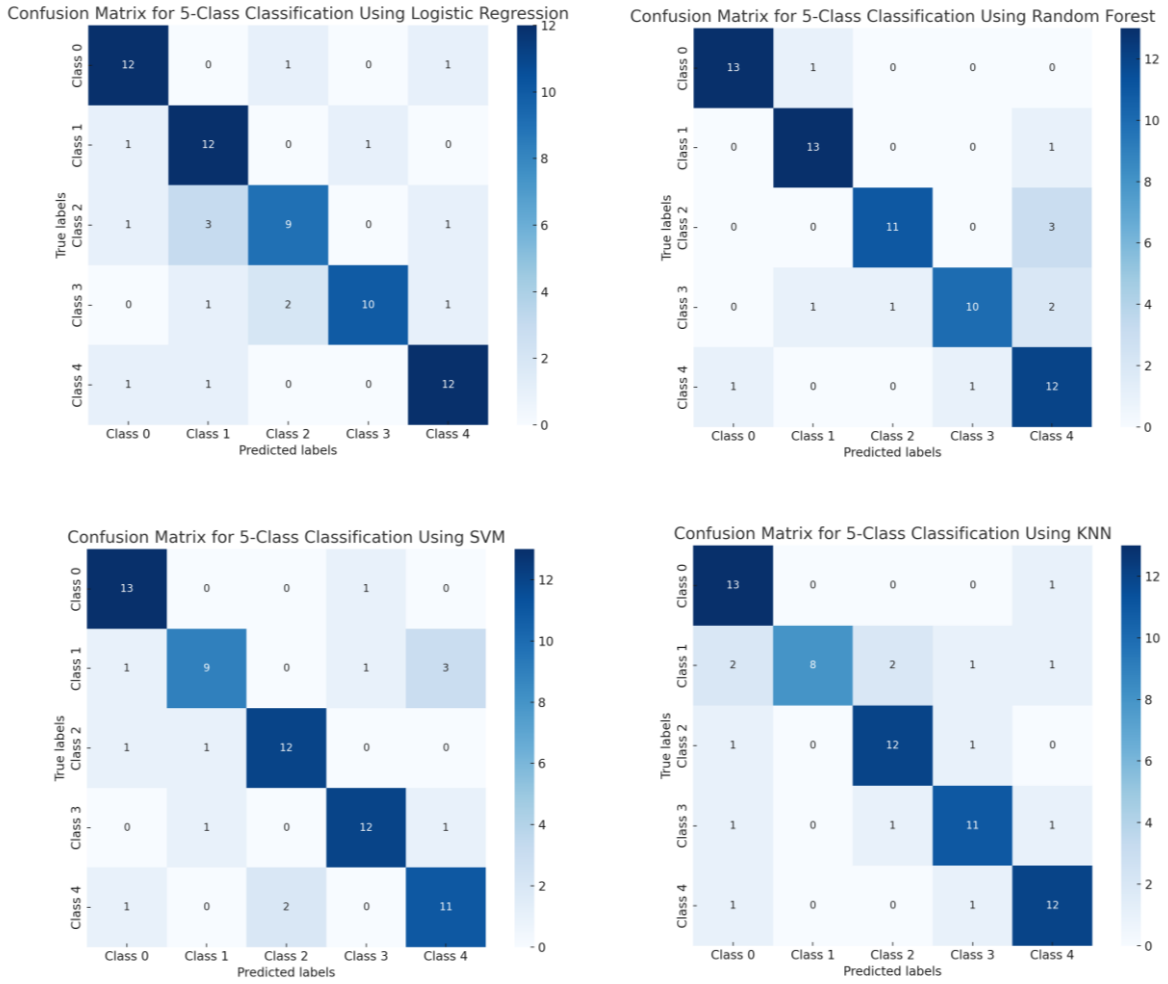


Figure 5: Confusion Matrixes for LR, RF, SVM and KNN

By comparing and analyzing the algorithm proposed in this study with other existing methods, it has been found that the proposed algorithm surpasses traditional anomaly detection methods and other few-shot learning algorithms on most evaluation metrics. Especially in tasks of anomaly detection under conditions of scarce samples, the proposed algorithm demonstrates significant advantages, owing to the effectiveness of prototype networks in few-shot learning, as well as specific optimization strategies tailored to the characteristics of gas-fired power plant operational data. Furthermore, the proposed algorithm also exhibits high precision and recall in identifying anomalies in minority classes, further proving its feasibility and effectiveness in practical applications. In summary, the empirical analysis results not only validate the effectiveness of the proposed algorithm but also showcase its application potential in the domain of anomaly detection in gas-fired power plants, providing valuable insights and references for future research and practical applications.

5 Conclusion and discussion

5.1 Conclusion

In this study, the prototype network few-shot learning algorithm displayed remarkable performance in the task of detecting abnormal behaviors in gas-fired power plants. The key to its successful application in this scenario lies in its ability to effectively utilize a limited number of samples for deep learning and accurately identify abnormal behaviors. Particularly in dealing with complex data features and identifying rare abnormal events, the algorithm significantly improved detection accuracy and efficiency by learning the similarities and differences between samples. The study achieved the following main conclusions:

1. By collecting and organizing monitoring data of abnormal behaviors from a gas-fired power plant, the study classified abnormal behaviors into five different types according to various indicators: equipment leakage or overload issues, supply system issues, low combustion efficiency issues, mechanical failure issues, and emission control problems.

2. Using the prototype network few-shot learning algorithm, the study conducted few-shot learning on 70 sets of data on abnormal behaviors. The model achieved an accuracy of 92.9%, a precision of 93.9%, a recall of 92.9%, and an F1 Score of 92.3%. These metrics demonstrate that the prototype network few-shot learning algorithm can effectively identify the types of abnormal behaviors in gas-fired power plants.

3. Compared with other machine learning algorithms (RF, SVM, LR, KNN), the prototype network learning algorithm performed the best. In terms of complex data features and the identification of rare abnormal events, the detection accuracy and efficiency were significantly improved by learning the similarities and differences between samples.

Key factors for the success of the prototype network algorithm include: 1). Precise data preprocessing and feature extraction to ensure that the model captures key anomaly indicators; 2). The design philosophy of the prototype network, which simplifies learning tasks on complex datasets by representing each category with a prototype; 3). Optimization strategies and parameter adjustments during the model training process that effectively prevent overfitting and improve the model's generalization ability.

The successful application of this algorithm in detecting abnormal behaviors in gas-fired power plants provides valuable references for its potential applications in other fields. For example, in the medical health, manufacturing, and traffic monitoring sectors, there is also the challenge of detecting anomalies or rare events from limited samples. The few-shot learning capability of the prototype network is particularly suited for applications where it is difficult to collect a large amount of labeled data. With appropriate adjustments and optimizations to the algorithm, it is expected to play an important role in early disease diagnosis, equipment failure prediction, traffic anomaly monitoring, and other fields, further promoting the development of intelligent monitoring and decision-support systems.

5.2 Discussion

This study applied the prototypical network-based few-shot learning (FSL) algorithm to anomaly detection in gas-fired power plants and achieved high performance under limited-sample conditions. Compared with traditional machine learning methods such as Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Convolutional Neural Networks (CNN), the proposed model achieved superior results across all major evaluation metrics. Specifically, under the 5-way 5-shot setting, our model achieved an F1-score improvement of 5.3% over RF, 6.1% over SVM, and 4.8% over CNN, indicating stronger discriminative feature learning and better generalization in small-sample scenarios. This advantage stems from the metric-based representation mechanism of the prototypical network, which mitigates overfitting risks and enhances robustness in data-scarce industrial environments. The model training process requires approximately 2.5 minutes for 100 epochs with a batch size of 16, and the average inference time per sample is about 0.021 seconds. This lightweight structure allows the model to operate in near real time when deployed in gas-fired power plant monitoring systems. Moreover, once trained, the model parameters remain compact, requiring less than 10 MB of memory, which enables smooth integration with existing monitoring platforms. Compared to existing SOTA methods, the FSL model demonstrates better adaptability to unseen anomaly categories and reduced dependence on retraining. The model's lightweight structure also allows fast convergence and low computational overhead, making it more suitable for real-time monitoring applications. Furthermore, the model's ability to learn representative class prototypes enhances interpretability, which is valuable for industrial engineers seeking to identify fault patterns and maintenance strategies.

However, several limitations remain. The dataset used in this study contains only 70 labeled samples, which limits the diversity of anomaly types and increases the potential for overfitting despite the model's regularization mechanisms. Future research should include larger and more diverse datasets, integrate multimodal inputs such as vibration, thermal, and image data, and explore online adaptive updating to handle changing operational states. Additionally, extending the current approach to other industrial domains—such as intelligent manufacturing and power grid fault diagnosis—will help validate the model's scalability and practical applicability.

In practical industrial environments, the proposed few-shot learning (FSL) model can be effectively integrated with existing intelligent control frameworks to enhance real-time monitoring and fault prediction. For instance, in nonlinear optimal control systems for gas compressors or adaptive backstepping controllers for flexible manipulators, feedback mechanisms are used to achieve stability and performance under dynamic uncertainty. When combined with such control systems, FSL can serve as a data-driven diagnostic layer that identifies early-stage anomalies from sensor signals and historical operation data before control degradation occurs. This integration enables predictive maintenance by providing high-confidence anomaly alerts that support the controller's decision-making process. Furthermore, the lightweight computational requirement of the prototype network structure allows it to be embedded into industrial control units for online anomaly detection without interrupting system operations. Therefore, combining FSL with existing adaptive or optimal control schemes offers a promising pathway toward more intelligent, safe, and efficient operation of complex industrial systems.

References

[1] Garcia C M ,Sanz-Bobi A M ,Pico D J .SIMAP: Intelligent System for Predictive Maintenance[J].Computers in Industry,2006,57(6):552-568.

[2] Wang G, Bi M, Fan H, et al. Key problems of gas-fired power plants participating in peak load regulation: A review[J]. IET Cyber-Physical Systems: Theory & Applications, 2023, 8(4): 222-235.

[3] Y. F. Wang, J. N. Qian. Gas flow path fault diagnosis and sensor fault diagnosis for aeroengine based on fusion filter algorithm[C]. Proceedings - 2017 International Conference on Sensing, Diagnostics, Prognostics, and Control, Shanghai, August 16, 2017 - August 18, 2017, 86-91.

[4] C. Y. Tyan, P. P. Wang. Rule-based fault diagnosis/detection expert system for aircraft flight control system[C]. International Conference on Fuzzy Theory and Technology Proceedings, United States, October 13, 1993 - October 16, 1993, 17-21.

[5] I. Samy, L. Postlethwaite, D. W. Gu. Survey and application of sensor fault detection and isolation schemes[J]. Control Engineering Practice,2011, 19(7): 658-674.

[6] E. Chung, J. S. Hanks. Fault tree analyses as a tool for flight control system architecture design[C].Proceedings-Annual Reliability and Maintainability Symposium, Alabama, January 25, 2016 - January 28, 2016.

[7] Lu S ,Hogg B .Dynamic nonlinear modelling of power plant by physical principles and neural networks[J].International Journal of Electrical Power and Energy Systems,2000,22(1):67-78.

[8] Hamrouni I, Lahdhiri H, Ben Abdellafou K, et al. Anomaly detection for process monitoring based on machine learning technique[J]. Neural Computing and Applications, 2023, 35(5): 4073-4097.

[9] Duong H T, Le V T, Hoang V T. Deep learning-based anomaly detection in video surveillance: A survey[J]. Sensors, 2023, 23(11): 5024.

[10] Tan P ,Zhang C ,Xia J , et al.NOX Emission Model for Coal-Fired Boilers Using Principal Component Analysis and Support Vector Regression[J].JOURNAL OF CHEMICAL ENGINEERING OF JAPAN,2018,49(2):211-216.

[11] Zhang L ,Zhou L ,Zhang Y , et al.Coal consumption prediction based on least squares support vector machine[J].IOP Conference Series: Earth and Environmental Science,2019,227(3):032007-032007.

[12] Guanglong, W.; Meng, L.; Wenjie, Z. The LS-SVM modeling of power station boiler NOx emission based on genetic algorithm. Automation & Instrumentation. 2016, 2, 26.

[13] Wang C ,Liu Y ,Zheng S , et al.Optimizing combustion of coal fired boilers for reducing NOx emission using Gaussian Process[J].Energy,2018,153149-158.

[14] Hu, D.; Chen, G.; Yang, T.; Zhang, C.; Wang, Z.; Chen, Q.; Li, B. An Artificial Neural Network Model for Monitoring Real-Time Variables and Detecting Early Warnings in Induced Draft Fan. In Proceedings of the ASME 2018 13th International Manufacturing Science and Engineering Conference. American Society of Mechanical Engineers Digital Collection, College Station, TX, USA, 18–22 June 2018.

[15] König F, Sous C, Chaib A O, et al. Machine learning based anomaly detection and classification of acoustic emission events for wear monitoring in sliding bearing systems[J]. Tribology International, 2021, 155: 106811.

[16] Safdarnejad M S ,Tuttle F J ,Powell M K .Dynamic modeling and optimization of a coal-fired utility boiler to forecast and minimize NOx and CO emissions simultaneously[J].Computers and Chemical Engineering,2019,12462-79.

[17] Bangalore ,Pramod,Tjernberg , et al.An Artificial Neural Network Approach for Early Fault Detection of Gearbox Bearings[J].IEEE transactions on smart grid,2015,6(2):980-987.

[18] Asgari H ,Chen X ,Morini M , et al.NARX models for simulation of the start-up operation of a single-shaft gas turbine[J].Applied Thermal Engineering,2016,93368-376.

[19] Lee J W ,Na J ,Kim K , et al.NARX modeling for real-time optimization of air and gas compression systems in chemical processes[J].Computers and Chemical Engineering,2018,115262-274.

[20] Hu, D, Guo, S, Chen, G, Zhang, C, Lv, D, Li, B, & Chen, Q. "Induced Draft Fan Early Anomaly Identification Based on SIS Data Using Normal Behavior Model in Thermal Power Plant." Proceedings of the ASME 2019 Power Conference. ASME 2019 Power Conference. Salt Lake City, Utah, USA. July 15–18, 2019. V001T08A002. ASME. https://doi.org/10.1115/POWER2019-1864

[21] Tan P ,He B ,Zhang C , et al.Dynamic modeling of NO X emission in a 660 MW coal-fired boiler with long short-term memory[J].Energy,2019,176429-436.

[22] Laubscher, R. Time-series forecasting of coal-fired power plant reheater metal temperatures using encoder-decoder recurrent neural networks. Energy 2019, 189, 116187.

[23] Yang, G.; Wang, Y.; Li, X. Prediction of the NOx emissions from thermal power plant using long-short term memory neural network. Energy 2020, 192, 116597.

[24] Pan H ,Su T ,Huang X , et al.LSTM-based soft sensor design for oxygen content of flue gas in coal-fired power plant[J].Transactions of the Institute of Measurement and Control,2020,014233122093239.

[25] Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 4080–4090.

[26] Zhou Y, Ma M, Mu X, et al. A Convolutional Neural Network Based on Channel Cross Attention for Efficient Probabilistic energy Flow Calculation in the Integrated Electrical and Natural Gas System[C]//2025 IEEE International Symposium on the Application of Artificial Intelligence in

Electrical Engineering (AAIEE). IEEE, 2025: 904-909.

[27] Qin Z, Wang X, Han X, et al. pre-trained combustion model and transfer learning in thermoacoustic instability[J]. Physics of Fluids, 2023, 35(3).

[28] Upadhyay S, Faruqui M, Tür G, et al. Almost zero-shot cross-lingual spoken language understanding [C]. In 2018 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), Berlin, German: Springer, 2018: 6034–6038.

[29] Pan C, Huang J, Gong J, et al. Few-shot transfer learning for text classification with lightweight word embedding based models [J]. IEEE Access, 2019, 7: 53296–53304.

[30] Abavisani M, Patel V M. Deep sparse representation-based classification [J]. IEEE Signal Processing Letters, 2019, 26 (6): 948–952.

[31] Lin K, Bai J, Li H, et al. Facial expression recognition with small samples fused with different models under deep learning [J]. Journal of Frontiers of Computer Science and Technology, 2020, 14 (3): 482–492.

[32] Yang X, He D, Zhou Z, et al. Improving offline handwritten Chinese character recognition by iterative refinement [C]. In 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), New York, USA: IEEE, 2017: 5–10.

[33] Zhimeng Xin, Shiming Chen, Tianxu Wu, Yuanjie Shao, Weiping Ding, Xinge You,Few-shot object detection: Research advances and challenges,Information Fusion,Volume 107,2024,102307,ISSN 1566-2535.

[34] Wang W, Li Y, Lu K, Zhang J, Chen P, Yan K, Wang B. Medical tumor image classification based on Few-shot learning. IEEE/ACM Trans Comput Biol Bioinform. 2023 Jun 9;PP. doi: 10.1109/TCBB.2023.3282226. Epub ahead of print. PMID: 37294647.

[35] Feng R, Zheng X, Gao T, Chen J, Wang W, Chen DZ, Wu J. Interactive Few-Shot Learning: Limited Supervision, Better Medical Image Segmentation. IEEE Trans Med Imaging. 2021 Oct;40(10):2575-2588. doi: 10.1109/TMI.2021.3060551. Epub 2021 Sep 30. PMID: 33606628.

[36] Oniani D, Chandrasekar P, Sivarajkumar S, Wang Y. Few-Shot Learning for Clinical Natural Language Processing Using Siamese Neural Networks: Algorithm Development and Validation Study JMIR AI 2023;2:e44293 URL: https://ai.jmir.org/2023/1/e44293 DOI: 10.2196/44293

[37] Rahman M, Rashid F, Roy S K, et al. Application of extreme learning machine (ELM) forecasting model on CO2 emission dataset of a natural gas-fired power plant in Dhaka, Bangladesh[J]. Data in Brief, 2024, 54: 110491.

[38] Boulkroune A, Zouari F, Boubellouta A. Adaptive fuzzy control for practical fixed-time synchronization of fractional-order chaotic systems[J]. Journal of Vibration and Control, 2025: 10775463251320258.

[39] 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[J]. Mathematical Problems in Engineering, 2017, 2017(1): 8045803.

[40] Zouari F, Saad K B, Benrejeb M. Robust neural adaptive control for a class of uncertain nonlinear complex dynamical multivariable systems[J]. International Review on Modelling and Simulations, 2012, 5(5): 2075-2103.

[41] Zouari F, Saad K B, Benrejeb M. Adaptive backstepping control for a class of uncertain single input single output nonlinear systems[C]//10th International Multi-Conferences on Systems, Signals & Devices 2013 (SSD13). IEEE, 2013: 1-6.

[42] Rigatos G, Abbaszadeh M, Sari B, et al. Nonlinear optimal control for a gas compressor driven by an induction motor[J]. Results in Control and Optimization, 2023, 11: 100226.

[43] Zouari F, Saad K B, Benrejeb M. Adaptive backstepping control for a single-link flexible robot manipulator driven DC motor[C]//2013 International Conference on Control, Decision and Information Technologies (CoDIT). IEEE, 2013: 864-871.

Comparative Metaheuristic Approaches to Tourist Itinerary Optimization in Low-Infrastructure Urban Contexts: A Case Study of Tirana

Alketa Hyso*, Dezdemona Gjylapi
Department of Computer Science, Faculty of Technical and Natural Sciences, University "Ismail Qemali", Albania
E-mail: alketa.hyso@univlora.edu.al, dezdemona.gjylapi@univlora.edu.al
*Corresponding author

Keywords: tourist itinerary planning, route optimization, Tirana, genetic algorithm, simulated annealing, brute force, hybrid Greedy–SA algorithm

Received: September 29, 2025

Tourist itinerary planning is a central component of smart tourism, yet it remains challenging in developing cities where computational resources and digital infrastructure are limited. This study examines Tirana, Albania, as a representative case for urban pedestrian-based tourist itinerary optimization. The analysis is carried out using both exact and heuristic optimization techniques, including Brute Force, Genetic Algorithm (GA), Simulated Annealing (SA), and a Hybrid Greedy + SA approach that integrates deterministic initialization with stochastic refinement. Distances between attractions were derived from OpenStreetMap, enabling fully reproducible experiments conducted on multiple datasets representing different sets of attractions with increasing size and under varying conditions, including ideal, noisy, and incomplete information. The results show that while exact computation rapidly becomes impractical as the instance size grows, metaheuristic methods, particularly SA and the hybrid variant consistently deliver high-quality and stable solutions. To evaluate real-world applicability under digital and computational constraints, the hybrid algorithm was implemented as a mobile-ready Progressive Web App and executed entirely on a resource-constrained device, demonstrating near-instantaneous optimization and confirming its feasibility for fully on-device use without reliance on backend servers. Overall, the study shows that lightweight metaheuristics, especially the Hybrid Greedy + SA method, offer a robust, scalable, and mobile-ready approach to urban tourism itinerary planning, suitable for deployment in environments with limited computational and infrastructural resources.

Povzetek: Študija pokaže, da so lahki hevristični algoritmi primerni za hitro načrtovanje turističnih poti tudi tam, kjer so računalniški viri in digitalna infrastruktura omejeni.

1 Introduction

Urban tourism functions as a major driver for city development because visitors seek to experience as much as possible during their brief urban visits. Tourism depends on digital services to let customers reserve accommodations and arrange their travel schedules [1], and Industry 4.0 technologies (AI, IoT, blockchain) create better visitor experiences [2]. However, this digital transformation has not been evenly distributed. In Tirana, the capital of Albania and one of the country's most rapidly developing tourist destinations, the growing number of visitors faces persistent difficulties due to the city's limited digital infrastructure. The city lacks the technological resources needed to support visitors in organizing and optimizing their travel plans [3]. The country of Albania does not have a unified national tourism portal which provides standardized information and booking capabilities and customized travel recommendations. The current promotional apps "Visit Tirana" [4], "Albania Tourist Guide" [5] fail to provide optimized routes and personalized itineraries and direct access to actual transportation services and do not have an e-ticketing system for public attractions.

The majority of visitors must rely on static maps and general platforms because of this gap which produces suboptimal routes and inconsistent travel experiences. The solution needs to use light-weight computational methods that operate within local boundaries and work within municipal or regional systems.

The research investigates three main questions about the trade-offs between solution quality and execution time when using exact and metaheuristic methods for day-trip itinerary planning. (Q1) How do exact and metaheuristic approaches trade off solution quality versus execution time for day-trip itineraries? Q2) How does performance scale with the instance size - the number of attractions (N)? Q3) Can lightweight metaheuristics be effectively implemented on resource-constrained mobile devices, enabling practical itinerary optimization without backend infrastructure?

We cast one-day tourist itinerary planning as a Travelling Salesperson Problem (TSP) and compare an exhaustive search Brute Force (BF) baseline with two

metaheuristics—Genetic Algorithm (GA) and Simulated Annealing (SA) implemented in Python and evaluated on real inter-attraction distances for Tirana derived from OpenStreetMap. We report tour length (m), runtime (s), and success rate across multiple values of N, in Section 3. The success rate counts only runs whose tour length exactly matches the best-known solution.

This study makes three principal contributions: (i) a deployable evaluation framework for optimizing urban tourist routes using open map data in resource-constrained environments; (ii) realistic benchmarks that quantify the speed–accuracy trade-off on standard hardware; and (iii) an implementation pathway that enables cities with limited digital infrastructure to adopt the solution. The remainder of the paper is organized as follows: Section 2 reviews related work; Section 3 describes the the experimental setup, and the optimization methods, including BF, GA, SA, and the proposed Hybrid Greedy + SA. It also details the robustness experiments under noisy and incomplete distance data, the validation of straight-line distances against OpenRouteService walking routes, and the mobile execution experiment on mobile hardware. Section 4 presents the empirical results across different problem sizes, robustness scenarios, and routing models, as well as the comparative performance of the hybrid approach on clean and perturbed data. Section 5 discusses the main design trade-offs between Haversine and network-based routing, generalization to larger instances, and the implications for smart-tourism applications in low-infrastructure cities. Finally, the concluding section summarizes the key findings, highlights the practical relevance of lightweight metaheuristics for mobile itinerary planning.

2 Related work

The field of urban itinerary optimization exists in two main forms which include the Tourist Trip Design Problem (TTDP) that handles preferences and time and budget constraints and opening hours [6] or as adaptations of the Travelling Salesperson Problem (TSP) for tourism applications, which focus primarily on shortest paths. The main distinction between TTDP variants and TSP-style research emerges from their distinct methods for evaluating user satisfaction under various constraints because TTDP variants handle multiple restrictions yet TSP-style studies focus on efficient solutions with basic assumptions.

Beyond this distinction, recent work has examined several directions. Sylejmani et al. (2024) use Iterated Local Search to personalize tours under thematic (e.g. visiting a museum before a restaurant and cultural sites before natural sites), temporal, and financial constraints [6], while Zhang et al. (2008) incorporate perceptual dimensions by blending path optimization with visual experience, mining web data to approximate the "visibility" or scenic value of attractions [7]. Adamo et al. (2022) explore multimodality, combining walking and driving routes to create more realistic itineraries [8]. Tang et al. (2024 demonstrates how large language models (LLMs) can transform unstructured user requests into optimized travel plans [9]. Souffriau et al. (2008) further show that guided local search can efficiently solve orienteering-based TTDPs directly on mobile devices, highlighting the viability of lightweight metaheuristics for on-device itinerary generation [10]. While exact methods offer optimality guarantees, they are often limited by scalability. Androutsopoulos & Zografos (2008) address multimodal itinerary planning with strict sequencing and time window constraints using an exact dynamic programming approach based on problem decomposition [11]. In contrast, the integration of hybrid algorithms with metaheuristics has proven effective in addressing complex optimization problems. Mangini et al. (2021) use graph theory, Integer Linear Programming (ILP), and a multi-algorithm strategy to generate one-day tourist routes that minimize travel time while reflecting user preferences [12]. Several studies show that combining SA with A* can generate fast and efficient tours [13]; Li et al. (2022) apply a knowledge-based hybrid Ant Colony Optimization algorithm enhanced with bacterial foraging to address group-based tourist satisfaction under capacity and preference constraints, with a focus on efficient tourist route planning [14]; GAs achieve improved results with faster convergence in tourist route optimization, while K-means clustering followed by GA optimization supports the creation of personalized tourist sequences [15]. Context-aware TTDP applies fuzzy-logic-enhanced metaheuristics to model contextual constraints [16]; fuzzy systems further support uncertainty management in attraction planning, resource evaluation, and decision-making [17, 18].

To synthesize these trends, we provide a structured comparison, Table 1, showing how prior work varies across problem formulations, algorithmic strategies, constraints, and data scales. While prior studies often assume rich mobility datasets, multimodal transport layers, or server-side computational capabilities, they collectively establish that metaheuristics are well-suited for real-time itinerary planning.

The present study extends prior work by addressing a gap not sufficiently explored: itinerary optimization in resource-constrained environments with limited digital infrastructure. In contrast to models that depend on transport schedules, multimodal networks, or complex personalization layers, our approach uses only geometric distances from OpenStreetMap, ensuring a lightweight and reproducible formulation. We adopt a deployment-oriented perspective by evaluating three solvers (Brute Force, GA, SA) across multiple data regimes and by implementing a Hybrid Greedy + SA solver as a mobile-ready Progressive Web App. Computing optimized itineraries entirely on-device, without backend support and with near-instantaneous runtime, highlights the practicality and scalability of the proposed approach for cities with constrained computational resources, limited connectivity, or incomplete digital maps.

Table 1: Summary of representative tourist itinerary planning approaches

Ref	Problem / Model	Main Algorithm	Constraints	Scale / Data
[10]	TTDP modelled as orienteering problem (OP) for mobile guides	Guided local search metaheuristic	Time budget, POI scores, personalization on mobile	Real data, city of Ghent
[6]	Extended TTDP as multi-constraint team orienteering problem with time windows (MCTOPTW) with patterns.	Local search / Iterated local search (ILS)-based metaheuristics	Time, budget, multi-knapsack; POI category patterns	146 instances, city-scale test set
[8]	Multi-modal TTDP (road + pedestrian)	Ad-hoc ILS	Time windows, visit duration, multimodal car+walking	Up to 3 643 POIs, 7-day horizon
[9]	Open-domain urban itineraries (OUIP)	LLM + cluster-aware spatial optimization	Natural-language requests; dynamic POIs; citywalk	4 Chinese cities, 1200+ itineraries
[16]	Context-aware TTDP	MGA + fuzzy logic a-posteriori evaluation	Contextual constraints (e.g. accessibility, safety).	Case study, Granada (mobility-impaired)
[13]	TTDP with GIS-grounded multimodal routing	Hybrid SA + A* (metaheuristic + heuristic)	GIS/OSM paths, POI distances, real pedestrian network, personalization	Fez Medina, 15 POIs, 9400 paths
[14]	Tourism route planning (TRP)	Knowledge-based Hybrid Ant Colony + Bacterial Foraging	Tourist satisfaction model, clustering, capacity limits	1000 tourists; 300-1300 paths
[11]	Multi-criteria time-dependent itinerary planning with mandatory intermediate stops in a multimodal fixed-schedule network	Dynamic Programming after problem decomposition into elementary sub-problems	Ordered multimodal routing with strict time windows at origin, destination, and intermediate stops.	Test network with 900 nodes, 930 service links, 12,000 interchange links, 100 services;
[12]	One-day TTDP-style round trip with outward/return itineraries on an urban PoI graph	Multi-level heuristic combining ILP-based symmetric TSP with graph-based PoI exchange/add/delete procedures	Single-day itinerary planning with PoI priorities, visit durations, mode-dependent travel times (walk/transit), and interactive user customization.	Case study: Bari (Italy) PoI network, realistic travel-time matrix; cruise-tourist one-day scenario
[7]	Tourist route planning with scenic-visibility scoring	Web-based POI extraction + GIS routing + 3D visibility computation	Personalized POI selection; scenic-visibility optimization; road-network routing	Prototype system; Japan (multiple scenic sites; DEM 50m grid)
[15]	Tourism path recommendation from survey & social-media objectives (cluster + TSP)	GA-enhanced k-means for clustering + GA for optimal tour	Multi-objective (12 internal & external tourism criteria);	Survey (600 visitors) + TripAdvisor data for 6 POIs in Port Sudan, Red Sea State (Sudan)

3 Methods and materials

3.1 Experimental setup

All algorithms were implemented in Python and executed under the same computational environment to ensure a fair comparison. Experiments were conducted on a Windows 10 Pro (64-bit) system equipped with an Intel Core i7-7500U 2.70 GHz processor and 12 GB of RAM. All runs were performed in Python 3.12.2 using JupyterLab 4.1.2 as the development interface. Core libraries and package versions are listed in the project's requirements.txt, and the complete code and environment specifications are available in the project's GitHub repository [17].

Each method — including GA, SA, and the Hybrid Greedy + SA variant — was independently executed 30 times. To ensure reproducibility, each run used a deterministic seed. This repeated-run protocol guarantees statistical robustness, allowing the computation mean for all main performance metrics: tour distance, execution

time, solution gap, and success rate. The reported values in all tables therefore represent the average performance across these 30 independent executions per scenario (clean, noisy, or incomplete). Brute Force was attempted in all cases, but when the 90-second limit was exceeded, its result was not included. In such cases, the best solution among GA and SA was adopted as the baseline.

Solution Gap quantifies the relative deviation of an algorithm's tour distance from the best-known solution. For each algorithm A, the Gap_mean was computed as:

$$Gap_{mean} = \frac{Distance_{mean} - Best_known}{Best_known} * 100$$

where $Distance_mean$ is the average tour distance obtained by algorithm A over 30 runs, and $Best_known$ represents the best-known tour length for the same scenario. When BF method successfully returned the optimal solution (for small N), its result was used as $Best_known$. For larger instances where BF exceeded the 90-second time limit, $Best_known$ was set to the shortest tour obtained among GA, SA, or Hybrid Greedy + SA. Thus, the Gap_mean values allow a normalized comparison of solution quality across algorithms and scenarios, regardless of the absolute tour length.

To measure how often each algorithm reaches the best-known solution, a success rate metric was used. A run is considered successful if the route distance it produces is within a small tolerance ε of the $Best_known$ distance. In this study, ε was set to 0.0 meters, meaning that only runs matching the optimal distance exactly are counted as successful. This strict criterion ensures fair and reproducible comparison between algorithms.

For GA, parameter values were selected based on a preliminary sensitivity analysis on the dataset (N = 31). Population size was varied across 30, 50, and 100 individuals while keeping the number of generations 50 and mutation rate 0.02 fixed. The results of this analysis, presented in Section 3.3.3 and Table 2, showed that a population of 50 provides a balanced compromise between runtime efficiency and search quality, with negligible performance differences for larger populations. Consequently, all subsequent GA experiments were performed using $population$ = 50, $generations$ = 50, and $mutation_rate$ = 0.02 to ensure comparability across algorithms.

3.2 Study context and data

This study adopts Tirana, Albania, as a case study, a developing capital experiencing growing tourist flows but lacking adequate digital systems for tourism management. A selection of cultural sites and historical landmarks and recreational activities was made to create authentic travel plans. The geographic coordinates of these attractions were extracted from OpenStreetMap (OSM). Distances between every pair of locations were computed using the Haversine formula, which calculates great-circle distances from latitude and longitude. The values were merged into a complete N×N distance matrix that used meters to measure distances and served as the foundation for all optimization techniques. The tour cost was defined as the total distance of a closed itinerary visiting each attraction exactly once

3.3 Optimization algorithms

3.3.1 Brute force

The BF method was used as an exact baseline. It generates all possible permutations of attractions, starting and ending at the same location, computes the total distance of each tour, and returns the shortest one. The search tree grows to its complete depth of N−1 levels where N shows the total number of attractions. The time complexity is factorial, O(N!). The method becomes impractical for use with large instances because of this. In our implementation, we imposed a 90-second time limit; if exceeded, the solver was skipped and excluded from evaluation. When successful, BF serves as an exact optimal baseline for comparison with heuristic methods.

3.3.2 Genetic algorithm

The GA operates as a population-based metaheuristic which draws its inspiration from natural evolutionary processes. Each candidate solution (individual) was represented as a permutation of attractions. The algorithm followed this sequence of operations:

Population initialization - Generate an initial population of random tours. Each individual in the population represents one complete itinerary; a permutation of all tourist points, ensuring that every attraction appears exactly once within a tour. Population initialization is performed by randomly generating pop_size permutations.

Fitness evaluation - Compute the total distance of each tour, where shorter routes indicate higher fitness. In principle, an individual becomes infeasible if any pair of consecutive points in its route is not connected by a valid edge. In our formulation, such missing connections are assigned an infinite, or equivalently, prohibitively large cost. Consequently, these individuals obtain extremely poor fitness scores and are automatically removed during the selection phase, without requiring an explicit feasibility check.

Selection - Choose the fittest individuals (those with the shortest tours) to form the mating pool. Selection is applied to retain only the fittest half of the individuals; those with the lowest total distance. This eliminates poorly performing solutions, including those containing infeasible edges.

Crossover - Create new offspring by combining segments of parent tours while ensuring that each attraction is visited exactly once. We employ Order Crossover (OX), a widely used method for permutation-based optimization:

1. Two cut points (a, b) are chosen at random.

2. The segment between a and b is copied directly from parent p1 to the child.

3. The remaining positions are filled with the elements of parent p2 in order, skipping elements already present in the copied segment.

This operator preserves relative ordering from both parents and prevents duplication of nodes, producing a feasible offspring tour.

Mutation - Introduce diversity by occasionally swapping two positions in a tour, reducing the risk of premature convergence. In our implementation, mutation swaps two randomly chosen positions in the tour with a small probability (rate). This operator creates slight perturbations in routes, helping the search explore new regions of the solution space and avoid stagnation in local minima.

Replacement - Substitute the parent population with the newly generated offspring.

Termination - Repeat the process for a fixed number of generations while tracking the best-so-far distance.

3.3.3 Sensitivity of GA parameter setting

To verify the robustness of the selected GA parameters, a small sensitivity analysis was conducted on the clean dataset (N = 31). The population size was varied across three settings: 30, 50, and 100, while keeping the number of generations 50 and mutation rate 0.02, fixed. As shown in Table 2, the mean tour distance and gap exhibit only modest variations (Gap_mean ranging from 62.8% to 70.0%), while the success rate remains constant at 0%. Increasing the population size leads to slightly longer runtimes (from 0.08 s to 0.26 s) but does not substantially improve solution quality.

Table 2: Sensitivity of GA performance to population size (N = 31, clean data, 30 runs)

GA	generations=50, mutation_rate=0.02			
Pop_size	Dist_mean	T_mean	Succ	Gap_mean
30	10592.014	0.083	0	69.995
50	10143.115	0.141	0	62.791
100	10258.633	0.263	0	64.645

Therefore, a population size of 50 was selected for all subsequent experiments, as it provides a balanced trade-off between runtime efficiency and search quality, ensuring fair comparison with the SA and Hybrid algorithms.

3.3.4 Simulated annealing

SA was implemented as a single-solution metaheuristic with stochastic acceptance of worse solutions. The steps of the algorithm are as follows:

Random Initialization - The algorithm begins from a randomly generated tour obtained by shuffling all points, which serves as the initial feasible solution.

Distance Evaluation - The optimization requires repeated tour-length evaluations based on pairwise distances between locations. To obtain accurate geographic measurements, all distances are computed using the Haversine formula, which estimates great-circle distances on the Earth's surface. For a given permutation of points, the total distance is evaluated as:

$$D(\pi) = \sum_{i=0}^{n-1} d_{\pi_i \pi(i+1) mod\ n}$$

The tour length D(π) is computed as the sum of the distances between every pair of consecutive locations in the tour. In this formulation, π represents a permutation of all points, and π_i denotes the point visited at position i in the tour. The term $d_{\pi_i \pi(i+1) mod\ n}$ corresponds to the distance between the current point π_i and the next point in the sequence. The modulo operator ensures that when the index reaches the last point, the route closes by connecting back to the first point, forming a complete loop.

Neighborhood Exploration - During the search, the algorithm iteratively explores neighboring solutions using a 2-opt segment reversal operator. The operator works by selecting two random indices i and j along the current tour and reversing the entire subsequence between them. This creates a new tour in which a segment of the route is traversed in the opposite direction. The intuition behind 2-opt is that many inefficient tours contain crossing edges, and reversing the segment between two points often eliminates these crossings, leading to a shorter overall path. Because of its ability to systematically remove such geometric inefficiencies, 2-opt is widely adopted in local search and metaheuristic optimization methods to produce substantial improvements in route quality with minimal computational overhead.

Acceptance Mechanism - At each iteration, the candidate solution is evaluated by comparing its tour length to that of the current solution. If the new solution yields a shorter tour, that is if: $D_{new} < D_{current}$, then it is accepted deterministically. Otherwise, the new solution may still be accepted with a probability defined by the classical Metropolis criterion:

$$P_{(accept)} = e^{\frac{D_{current} - D_{new}}{T}}$$

This probabilistic acceptance allows the algorithm to escape local optima by admitting occasional uphill moves.

Temperature Schedule - The temperature was initialized at a high value (10,000), enabling broad exploration in early iterations. A cooling factor 0.995 is applied at each step.

$$T_{k+1} = 0.995 * T_k$$

As the temperature decreases, the probability of accepting inferior solutions diminishes, encouraging convergence toward an optimal or near-optimal tour.

Convergence Tracking - The best distance found so far is logged periodically throughout the search, producing a convergence history curve. This allows for a detailed comparison of SA's performance relative GA.

For the standalone SA implementation, the following parameters were used:

initial temperature T0=10000
cooling rate α=0.995
stopping condition T>1
neighborhood: 2-opt segment reversal
acceptance rule: Metropolis criterion
logging interval: 100 iterations

To account for stochastic variability, 30 independent runs were performed using run-level random seeds:

seed=4242+r with both *random* and *numpy.random* initialized per run to ensure full reproducibility.

3.3.5 Additional experiments: robustness under noisy and incomplete data

To further evaluate the robustness of the proposed metaheuristics (GA and SA), we conducted a set of supplementary experiments where the distance data were intentionally perturbed to simulate real-world imperfections.

These imperfections represent measurement noise (GPS inaccuracies), and incomplete connectivity (unavailable or missing links between points of interest).

(a) Perturbation model

Starting from the original distance matrix D, a perturbed version D′ was generated using the function:

$$D'_{ij} = D_{ij} * (1 + \epsilon_{i,j}), \quad \epsilon_{i,j} \in [-\rho, +\rho]$$

where $\rho \in \{0.1, 0.2\}$ denotes the amplitude of the multiplicative noise (±10–20%). Additionally, a random fraction $r \in \{0.1, 0.2\}$ of the entries were replaced with $+\infty$ indicating missing edges (no direct connection between those nodes). This setup generated five test scenarios:

- noisy10 (±10% noise)
- noisy20 (±20% noise)
- miss10 (10% missing edges)
- miss20 (20% missing edges)
- noisy15_miss15 (±15% noise, 15% missing edges combined)

Each perturbed matrix was stored for reproducibility and used as input for both GA and SA with identical hyperparameters.

(b) On-the-fly repair strategy

In the presence of missing edges, direct distance evaluation would normally fail. To overcome this, we introduced an on-the-fly repair algorithm, a simple yet effective mechanism that restores connectivity dynamically during route evaluation. For any pair of consecutive nodes (u, v) where$D'[u, v] = \infty$, the repair procedure searches for an intermediate node k that minimizes the detour cost:

$D_{repair}(u, v) = min(D'[u, k] + D'[k, v])$ where $k \neq u, v$

If such a node exists, the route temporarily diverts through k; otherwise, a large penalty value is assigned, so that the optimizer avoids this segment.

This mechanism ensures that the optimization process remains feasible and continuous even in degraded or incomplete graphs, emulating real-time adaptive behavior.

3.3.6 Hybrid Greedy + SA

Following the analysis of the standard GA and SA heuristics, we introduced a third variant — Hybrid Greedy + SA (Algorithm 3), to enhance anytime performance and robustness under incomplete or noisy distance data.

The hybrid algorithm begins with a Greedy initialization, which quickly constructs a valid tour by iteratively selecting the nearest unvisited node. This step provides a feasible and near-optimal starting solution. After obtaining this initial tour, the algorithm enters the Simulated Annealing refinement phase, where it applies 2-opt neighborhood moves to progressively improve the tour. At each iteration, a new candidate route is generated by reversing a random segment, and the change is accepted either if it improves the total distance or probabilistically if it is slightly worse, to avoid premature convergence. This two-stage design combines the speed and feasibility of Greedy construction with the exploratory power of Simulated Annealing, resulting in a fast yet robust method capable of maintaining good solution quality under noisy or incomplete data.

Because the Hybrid algorithm starts from a high-quality Greedy tour, it does not require the high initial temperature (T_0 = 10000) and slow cooling (α = 0.995) used in standalone SA to escape poor initial states. This allows the Hybrid configuration to allocate more computation to focused improvement rather than broad exploration, accelerating convergence without sacrificing robustness. The Hybrid Greedy + SA algorithm was executed with the following parameters:

- initial temperature T0=5000
- cooling rate α=0.996
- stopping temperature T>1
- neighborhood: 2-opt segment reversal
- acceptance rule: Metropolis criterion
- logging interval: 100 iterations
- 30 independent runs with different random seeds
- 2-hop repair and large penalty (1e12) for missing edges

For perturbed data, 10–20% noise and missing edges were introduced, and the Hybrid Greedy + SA used 2-hop repair with a large penalty (1e12), while keeping the same annealing parameters (T0=5000, α=0.996, 30 runs).

The experimental results in Section 4.5 directly assess its effectiveness under both clean and perturbed data conditions. The two-stage design—Greedy initialization followed by SA refinement—was specifically intended to accelerate convergence and enhance robustness when distance information is incomplete or noisy. The results presented in the section 4.5, validate this expectation: on clean data, the hybrid approach substantially reduces the optimality gap while maintaining competitive runtimes, and under noisy or incomplete distance matrices it preserves solution quality and feasibility more effectively than standalone GA or SA. Thus, the performance analysis confirms the practical benefits of the proposed hybrid strategy anticipated in the methodology, demonstrating its ability to deliver reliable and efficient itineraries across varying data scenarios.

3.4 Realistic distance validation using openrouteservice

To evaluate how geometric distance approximations differ from real-world travel distances within the urban environment of Tirana, we conducted an additional experiment that integrates route optimization with real street-network routing.

Experimental Setup

The analysis was performed using:

-Python 3.10+, OpenStreetMap (OSM) as the source of geographic coordinates for 31 tourist attractions in Tirana.

-The Hybrid Greedy + SA metaheuristic implemented in Python to obtain an optimized visiting order based on Haversine (straight-line) distances.

-OpenRouteService (ORS) API, which computes real walking distances using the actual street network derived from OSM.

The workflow consisted of three major steps:

-Build a full Haversine distance matrix for all 31 points using their OSM lat/long coordinates.

-Run the Hybrid Greedy + SA algorithm.

-Evaluate the optimized tour using OpenRouteService to compute true, network-based walking distances between consecutive attractions.

3.4.1 Visualization-based structural analysis of haversine and ORS routing models

In addition to the primary optimization framework based on Haversine great-circle distances, we conducted a complementary routing experiment aimed at assessing how the choice of distance metric influences the structure of the optimized tour. The same set of 31 tourist locations was evaluated using two alternative distance models: (i) geometric Haversine distances, which provide a fast, API-independent approximation of spatial proximity, and (ii) real walking distances obtained through the OpenRouteService (ORS) Matrix API. For the second model, the Hybrid Greedy + SA algorithm was executed directly on ORS-derived network distances to obtain a fully realistic, walkability-constrained shortest tour. This dual-routing setup enables a controlled methodological comparison between geometric and network-based optimization, ensuring that differences observed in the results can be attributed solely to the underlying distance model. The inclusion of both models strengthens the reliability of the experimental design and provides a foundation for understanding the impact of routing realism on itinerary planning.

3.5 Mobile execution and deployability evaluation

Since tourist route optimization is expected to run directly on users' smartphones during real-world use, testing solely on desktop hardware does not reflect practical performance. The mobile execution experiment was therefore included to measure responsiveness under realistic computational constraints.

To evaluate the real-world performance of the Hybrid Greedy + SA algorithm, we conducted an additional execution-time assessment on a mobile device. A lightweight web application (HTML/JavaScript/Leaflet) containing the full algorithm was deployed on GitHub Pages and installed on an iPhone as a Progressive Web App (PWA) using Safari's "Add to Home Screen" feature. When launched, the application runs in standalone mode and executes all computations locally on the device's JavaScript engine.

To guarantee cross-platform reproducibility, we used the same deterministic seed schedule in both desktop (Python) and mobile (JavaScript) implementations. Specifically, each run was initialized with seed = 2025 + r, where r is the run index. This ensures that the stochastic trajectory of SA remains consistent across environments, enabling a fair comparison of results. Tests were performed on Samsung Galaxy A10 running Android 11 using the GitHub Pages web-app. Execution time was measured using *performance.now()*, which provides high-resolution, device-level timing.

3.6 Output and visualization

All results were recorded at both run and summary levels. The experimental framework saved distance and runtime data along with success indicators in CSV format at the run level. The summary statistics displayed the average values and standard deviations for all recorded metrics. The analysis included two types of visualizations which showed distance and runtime distributions through boxplots and displayed GA and SA convergence patterns through median and interquartile range curves. All visual content was generated at 600 dpi resolution to ensure high image quality. In addition, two map visualization was developed to demonstrate how the generated itineraries would perform in the actual geography of Tirana.

4 Results and fiscussion

4.1 Performance on small instance (N=7)

Table 3 shows the results of BF and GA and SA when the number of attractions is seven. Reported metrics include mean and standard deviation of tour distance and runtime, success rate (percentage of runs achieving the best-known solution).

Table 3: Comparative results of BF, GA, and SA for N=7 attractions.

	Dist_mean	Dist_std	T_mean	T_std	Succ
BF	3886.4775	0	0.0069	0	100
GA	3886.4775	1.39e-12	0.0890	0.0220	100
SA	3886.6375	0.876156	0.0373	0.0180	96.66

Taken together, the results demonstrate that for small problem sizes (N=7), all three methods reach or approximate the optimal solution, but their runtime characteristics differ. BF is the fastest in absolute terms for this small instance. GA guarantees optimality but incurs higher runtime overhead due to its population-based operations. SA trades a small loss in accuracy for significantly faster and more consistent runtimes then GA. Figures 1 and 2 (N = 7) show that both GA and SA consistently attain the BF-optimal tour length: the distance distributions are essentially indistinguishable, with identical medians (orange lines) and nearly identical means (triangles). In contrast, the runtime boxplots reveal a clear separation: SA runs faster on average and exhibits

lower variability, while GA shows higher median/mean times.

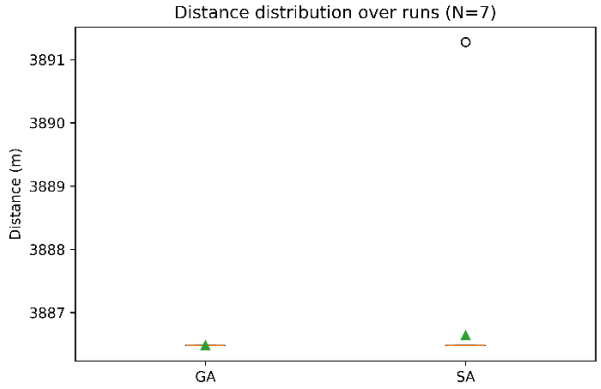


Figure 1: Performance evaluation for N = 7, showing the distance distribution across 30 independent runs of GA and SA. Means (triangles) and medians (orange lines).

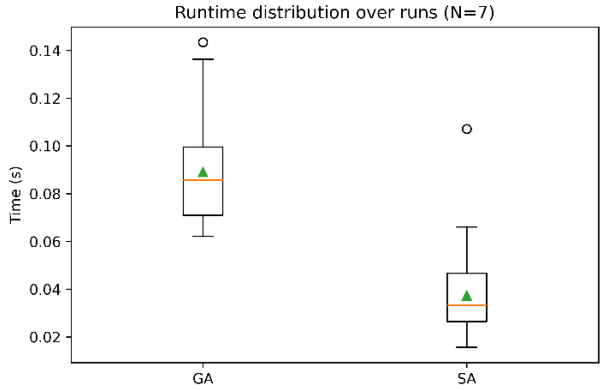


Figure 2: Performance evaluation for N = 7, showing the runtime distribution across 30 independent runs of GA and SA. Means (triangles) and medians (orange lines).

4.2 Results on the 10-attraction instance

Table 4 compares the performance of BF, GA, and SA for itineraries with ten attractions. Reported metrics include mean and standard deviation of tour distance and runtime, success rate (percentage of runs achieving the best-known solution).

Table 4: Comparative results of BF, GA, and SA for N=10 attractions.

	Dist_mean	Dist_std	T_mean	T_std	Succ
BF	4379.82	0	4.2005	0	100
GA	4385.80	7.070	0.1063	0.0177	43.3
SA	4383.69	5.231	0.0625	0.0108	43.3

For N=10, BF produced the exact optimum in 4.20 s, which remains tractable at this scale but already illustrates the rapid increase in computational cost. Both GA and SA found near-optimal tours, yet their success rates dropped to 43.3%. Between the heuristics, SA achieved slightly shorter tours on average and faster runtimes, offering a more favorable balance of accuracy and efficiency than GA.

To assess whether these differences are statistically significant, we complemented the descriptive results with bootstrap confidence intervals and Mann–Whitney U tests, as reported in Table 5.

Table 5: Statistical comparison of GA and SA using bootstrap CIs (95%) and Mann–Whitney U tests (30 runs).

	Alg.	Mean	95% Bootstrap CI	U	p-value
D	GA	4385.80	[4383.43, 4388.38]	536.5	1.953e-01
	SA	4383.69	[4381.99, 4385.69]		
T	GA	0.1202	[0.1001, 0.1126]	891.0	7.39e-11
	SA	0.0614	[0.0587, 0.0663]		

For tour distance (D), GA and SA have very similar mean values, and their 95% bootstrap confidence intervals overlap extensively. The Mann–Whitney test yields U = 536.5, p = 1.953e-01, indicating no statistically significant difference in tour length between the two algorithms at this scale.

For runtime (T), the difference is substantial. GA is nearly twice as slow as SA, and the confidence intervals do not overlap. The Mann–Whitney test gives U = 891.0, p = 7.39e-11, showing a highly significant advantage of SA in terms of execution time.

Overall, the results show that both algorithms achieve comparable tour lengths, but SA is far faster and more computationally efficient.

4.3 Large-scale performance (N=31)

Under the latency constraint (≤90 s), Brute Force cannot be executed for N = 31 and is therefore excluded from the comparison. Tables 6 and 7 summarize the empirical and statistical results obtained from 30 independent runs of GA and SA.

Table 6: Empirical performance of GA and SA for N = 31 attractions (30 runs). BF timed out and is omitted

	Dist_mean	Dist_std	T_mean	T_std	Succ
GA	10143.11	915.95	0.12	0.0138	0
SA	6782.44	461.25	0.06	0.0082	3.33

The statistical analysis in Table 6, confirms that SA strongly outperforms GA at N = 31 attractions spots. SA produces significantly shorter tours (mean = 6782.44 m) compared to GA (mean = 10143.11 m) and exhibits much lower variability (SD = 461.25 m vs 915.95 m). Runtime performance is similarly favorable to SA, which achieves 0.06 s on average nearly twice as fast as GA, and with lower dispersion across runs.

Table 7: Statistical comparison of GA and SA using bootstrap CIs (95%) and Mann–Whitney U tests (30 runs).

	Alg.	Mean	95% Bootstrap CI	U	p-value
D	GA	10143.11	[9815.20, 10449.87]	900	3.02e-11
	SA	6782.44	[6629.76, 6947.91]		
T	GA	0.1202	[0.1153, 0.1250]	900	3.02e-11
	SA	0.0614	[0.0584, 0.0642]		

Success rates are low for both algorithms at this problem size (SA: 3.3%, GA: 0%). However, only SA is able to reach the best-known tour at least once (1/30 runs).

The Mann–Whitney U tests for both distance and runtime yield $U = 900$ and $p \approx 3 \times 10^{-11}$, indicating highly statistically significant differences between GA and SA across all evaluated metrics.

Overall, SA offers the best accuracy–efficiency trade-off under real-time constraints, while GA becomes slower, less stable, and substantially less accurate as N increases.

While Tables 6 and 7 summarize the final performance outcomes of both algorithms, Figures 3 and 4 provide complementary insight by illustrating the optimization trajectory across iterations. These convergence plots show the median best-so-far tour length over 30 independent runs for N = 31, accompanied by the corresponding interquartile range (IQR) at each iteration.

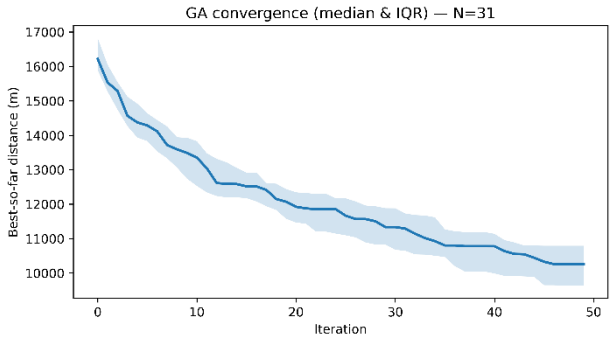


Figure 3: Convergence for N=31 over 30 runs of GA. Curve shows the median best-so-far tour length at each iteration; shaded band indicates the interquartile range (IQR)

A quantitative estimation of convergence speed was also derived from the median curves. For GA, the 95% of-final-value threshold (~10,500 m) is reached at approximately iteration 27–30. In contrast, SA reaches its 95% threshold (~7,140 m) at iteration 9–10. This indicates that SA converges roughly three times faster than GA, supporting the observed efficiency gap in the empirical results.

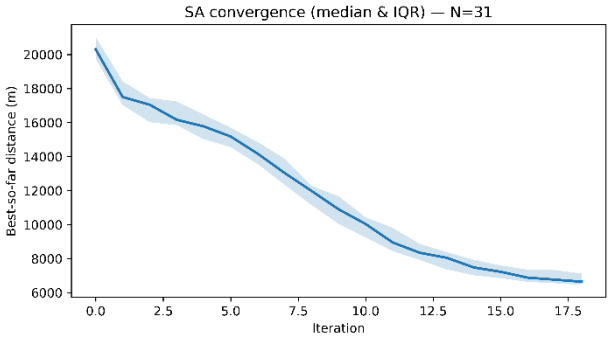


Figure 4: Convergence for N=31 over 30 runs of SA. Curve shows the median best-so-far tour length at each iteration; shaded band indicates the interquartile range (IQR)

4.4 Robustness of the algorithms under noisy and incomplete data

Tables 8-11 present the experimental results for both SA and GA under clean, noisy, and incomplete distance-matrix scenarios. The experiments were performed on a fixed instance consisting of 31 tourist attractions in Tirana.

The ΔDistance, ΔTime, and ΔGap indicators quantify the deviation from the clean baseline, while the Success_rate column reflects the algorithm's consistency in recovering near-optimal tours under uncertainty. Dist_mean denotes the mean tour length (in meters) across 30 runs, while ΔDistance (%) expresses the relative change in tour length compared to the clean baseline. T_mean (s) reports the average runtime, and ΔTime (%) quantifies the relative runtime increase under perturbation. Gap_mean (%) measures the deviation from the best-known solution, and ΔGap (percentage points (pp)) shows how this gap changes relative to the clean case. Success_rate (%) indicates the percentage of runs that exactly matched the best-known tour ($\varepsilon = 0$).

To provide a clearer algorithm-specific interpretation of robustness under uncertainty, the following subsections analyze in detail the behavior of SA and GA separately, with respect to solution quality, runtime stability, and convergence reliability under noisy and incomplete distance-matrix conditions.

4.4.1 SA robustness analysis

This subsection examines the robustness of SA algorithm under progressively degraded data conditions, including missing distances and random noise. The analysis focuses on how perturbations affect tour length, solution optimality gap, success rate, and computational time, in comparison with the clean baseline scenario. Tables 8 and 9 summarize the resulting performance metrics across all test cases.

For SA, both the solution gap and success rate remain almost unchanged across all perturbation types, confirming the strong resilience of the method.

Despite large relative increases in execution time (ΔTime% up to +2100%), the absolute runtimes remain

below 1.5 seconds, making SA highly practical even when the data are incomplete or noisy.

Table 8: Comparative performance of SA algorithm under noisy and incomplete distance-matrix conditions, across all test scenarios.

Scenario	Dist_ mean	ΔDist.	Gap_ mean	ΔGap	Succ rate
clean	6782.44	0	8.854	0	3.3
miss10	7245.51	6.83	9.152	0.298	3.3
miss20	7388.59	8.94	6.254	-2.6	3.3
noisy10	6988.99	3.05	11.975	3.121	3.3
noisy15_ miss15	6778.24	-0.06	7.904	-0.95	3.3
noisy20	6559.38	-3.29	8.838	-0.01	3.3

Table 9: Runtime analysis of the SA algorithm under clean, noisy, and incomplete distance-matrix conditions, reporting mean execution time and its relative deviation from the clean scenario

Scenario	T_mean	ΔTime
clean	0.0653	0
miss10	0.8435	1192.07
miss20	1.439	2104.24
noisy10	0.2362	261.84
noisy15_miss15	1.0523	1511.93
noisy20	0.2178	233.59

The mean tour distance increases moderately under missing-edge scenarios (+6–9%), reflecting the expected effect of reduced connectivity, while under noisy conditions the distances fluctuate only slightly (±3%). Importantly, the solution gap remains within ±3 percentage points of the clean baseline (from 8.85% to a maximum of 11.98%), showing that SA can adapt efficiently to both random noise and incomplete information. The success rate is fully stable (3.3%) across all tests, indicating deterministic convergence toward near-optimal tours.

Overall, these findings demonstrate that SA degrades gracefully: its runtime scales with problem irregularity, but its solution quality and convergence behavior remain robust. This makes SA a reliable choice for real-world deployment, particularly in applications (e.g., mobile routing, tour recommendation) where the input data may be imperfect or partially missing.

4.4.2 GA robustness analysis

This subsection analyzes the robustness of GA under the same noisy and incomplete distance-matrix scenarios. The evaluation emphasizes variations in solution quality, optimality gap, feasibility preservation, and runtime overhead relative to the clean baseline. Tables 10 and 11 report the corresponding performance indicators.

Tables 10 and 11 demonstrate that, for GA, both the solution gap (ΔGap) and the runtime overhead (ΔTime) increase noticeably under missing and noisy distance conditions, indicating a high sensitivity to data degradation. Although GA consistently produces feasible tours across all scenarios, these tables clearly show that its solution quality varies substantially more than that of SA.

Table 10: Comparative performance of GA algorithm under noisy and incomplete distance-matrix conditions, across all test scenarios

Scenario	Dist_ mean	ΔDist	Gap_ mean	ΔGap	Succ rate
clean	10143.11	0	62.791	0	0
miss10	10186.23	0.43	53.454	-9.34	0
miss20	10568.79	4.2	51.988	-10.80	0
noisy10	10085.95	-0.56	61.594	-1.19	0
noisy15 miss15	10452.92	3.05	66.403	3.61	0
noisy20	10185.24	0.42	69.001	6.21	0

Table 11: Runtime analysis of the GA algorithm under clean, noisy, and incomplete distance-matrix conditions, reporting mean execution time and its relative deviation from the clean scenario

Scenario	T_mean	ΔTime
clean	0.1403	0
miss10	1.6976	1109.78
miss20	2.4688	1659.35
noisy10	0.3219	129.41
noisy15_miss15	1.7458	1144.17
noisy20	0.3576	154.86

Under missing-edge perturbations (miss10–miss20), the gap decreased slightly (≈ −9 to −11 pp) due to the on-the-fly repair mechanism that bypassed broken links, but this came at a significant computational cost, with runtime increases exceeding +1100–1600%. Similarly, when random noise was applied (±10–20%), GA exhibited relatively stable distances (ΔDistance < 1%) but still required additional iterations, leading to +130–150% longer runtimes. In the combined case (noisy15 + miss15), GA's mean gap reached ≈66%, showing that its population-based search is more sensitive to irregular graph structures compared with SA's single-solution adaptive exploration.

Across all experimental scenarios, the GA algorithm consistently produced complete and connected tours, confirming that the proposed on-the-fly link-repair mechanism effectively preserves route connectivity even when distance information is missing or corrupted. Overall, GA shows strong reliability in maintaining feasible tours under data degradation, but its efficiency is notably weaker: although it remains capable of constructing complete routes, its performance deteriorates much more rapidly than SA with respect to runtime and solution optimality.

Figure 5 complements Tables 8–11 by illustrating the anytime trade-off between solution quality (Gap %) and

runtime (s) for both algorithms (SA and GA) under the same set of perturbation scenarios. Each point corresponds to the mean performance of a given scenario (clean, noisy, or incomplete), allowing a direct visual comparison of robustness and efficiency.

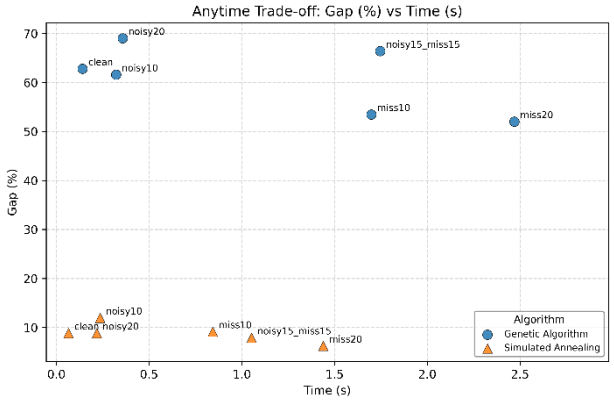


Figure 5: Anytime trade-off between solution quality (Gap %) and execution time (s) for GA and SA across clean, noisy, and incomplete data scenarios.

Each point corresponds to the mean performance of a given scenario (clean, noisy, or incomplete), allowing a direct visual comparison of robustness and efficiency. The plot clearly shows that SA (orange triangles) consistently achieves lower gaps with sub-second runtimes, while GA (blue circles) requires longer execution times and exhibits larger variability in solution quality, especially under missing-edge conditions. These visual trends confirm the tabular findings, that SA degrades smoothly and maintains stability under noisy and incomplete data, whereas GA's performance is more sensitive to data imperfections

4.5 Performance of the Hybrid Greedy + SA (Algorithm 3)

4.5.1 Performance on clean data

The Hybrid Greedy + SA algorithm was first evaluated on the clean dataset (no missing or noisy distances) to establish a baseline for performance comparison against the standalone GA and SA methods. The experiment contained 31 tourist attractions. As shown in Table 12, the hybrid approach achieved a significantly lower mean gap (≈ 1.86%) compared to SA (8.85%) and GA (62.79%), while maintaining a moderate runtime (≈ 0.25 s) and the highest success rate (10%). These results demonstrate that initializing SA with a Greedy heuristic drastically improves both convergence speed and solution quality, providing near-optimal tours at minimal computational cost.

Table 12: Baseline results for GA, SA, and Hybrid Greedy + SA on clean data.

Alg.	Dist_ mean	T_mean	Succ	Gap_ mean
Hybrid	6342.551	0.2499	10%	1.860
SA	6782.44	0.0653	3.3%	8.854
GA	10143.11	0.1403	0%	62.791

The results show that the Hybrid method consistently outperforms SA in terms of tour quality: its mean distance is substantially lower, and its confidence interval is considerably narrower, indicating both better performance and reduced variability. For execution time, SA is significantly faster, as expected from its simpler stochastic refinement process. The Hybrid method incurs a higher computational cost due to the deterministic greedy initialization and more intensive local search.

Table 13 provides a statistical comparison between SA and the Hybrid Greedy+SA algorithm using 95% confidence intervals and Mann–Whitney U tests over 30 independent runs.

Table 13: Statistical comparison of SA and Hybrid Greedy+SA using 95% CIs and Mann–Whitney U tests (30 runs).

	Alg.	Mean	95% CI	U	p-value
D	SA	6782.4	[6626.47, 6946.74]	824	2e-08
	Hybrid	6342.6	[6319.59, 6364.00]		
T	SA	0.067	[0.06363, 0.07070]	0	3.02e-11
	Hybrid	0.249	[0.24373, 0.25625]		

The Mann–Whitney U tests confirm that both distance and time differences between the two algorithms are statistically significant, with p-values far below the standard 0.05 threshold. This validates that the observed improvements in solution quality—and the corresponding increase in runtime—are not due to random variation but reflect systematic differences in algorithmic behavior.

Overall, the statistical evidence demonstrates that Hybrid Greedy+SA achieves the best accuracy among the compared methods, while SA offers superior speed, reflecting a clear trade-off between solution quality and runtime.

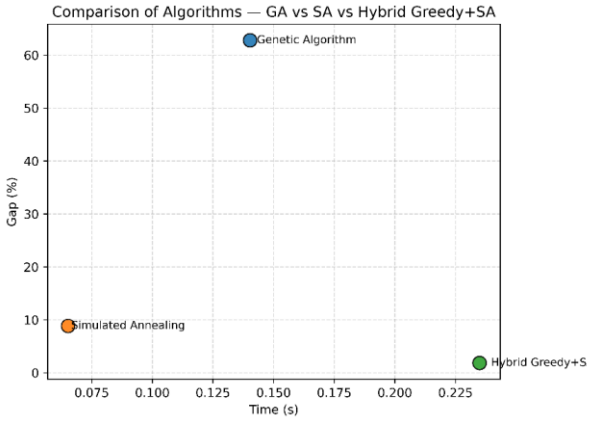


Figure 6: Comparison of GA, SA, and Hybrid Greedy + SA on clean data, showing the trade-off between runtime and optimality gap.

Figure 6, visualizes the clean baseline comparison of the three algorithms. The Hybrid Greedy+SA achieves a much smaller gap than both GA and SA while maintaining a sub-second runtime, confirming its strong potential for real-time applications in urban itinerary optimization.

4.5.2 Robustness under noisy and incomplete data

This section presents the comparative performance of the three algorithms: GA, SA, and the proposed Hybrid Greedy + SA, when evaluated under a combined perturbation scenario (noisy15_miss10). In this configuration, 15% of the pairwise distances were randomly perturbed with Gaussian noise, while 10% of the graph edges were removed to simulate missing or unreliable connectivity data. This setting reflects realistic conditions for low-infrastructure urban contexts, where digital maps or open data repositories often contain incomplete or imprecise geospatial information.

As shown in Table 14 and Figure 7, both GA and SA experience a degradation in performance when the data become noisy or incomplete, whereas the Hybrid Greedy + SA maintains lower gap values and stable success rates.

Table 14: Comparative performance of Genetic Algorithm (GA), Simulated Annealing (SA), and Hybrid Greedy + SA under the combined perturbation scenario (noisy15_miss10).

Scenario: noisy15_miss10				
Alg.	Dist_mean	T_ mean	Succ.	Gap_ mean
Hybrid	7303.99	0.942	3.3%	7.421
SA	7365.73	0.767	0%	8.329
GA	10732.95	1.407	0%	57.852

This demonstrates that initializing SA with a Greedy heuristic seed improves the anytime performance and robustness of the algorithm, even when link failures or measurement errors occur. While GA suffers from a substantial increase in gap (≈ 58%) and runtime, the hybrid method preserves feasibility and remains computationally efficient, maintaining sub-second runtimes (≈ 0.94 s) and a mean gap of 7.4%.

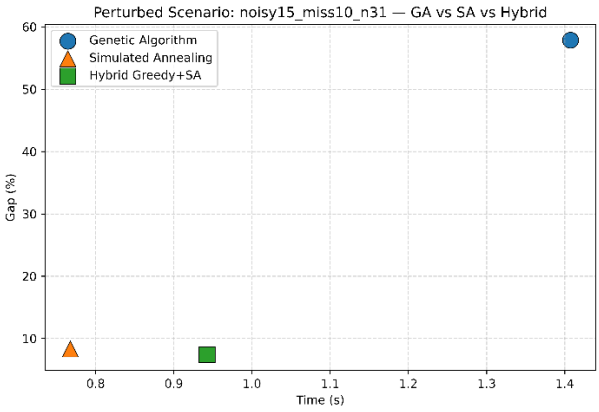


Figure 7: Performance comparison of GA, SA, and Hybrid Greedy + SA under the noisy15_miss10 scenario

Overall, the results confirm that the Hybrid Greedy + SA algorithm achieves a superior balance between robustness and efficiency, sustaining near-optimal solution quality even when data imperfections degrade the performance of traditional metaheuristics.

4.6 Comparison between straight-line and real route distances

To evaluate how the optimized itinerary behaves under real-world travel conditions, we compared the total straight-line distance produced by the Hybrid Greedy + SA algorithm with the actual routed distance obtained through the OpenRouteService (ORS) API. The optimized tour was first generated using Haversine distances computed from OpenStreetMap-based coordinates of 31 tourist attractions in Tirana. This straight-line model provides a simplified and computationally efficient geometric approximation commonly used in TSP and metaheuristic optimization research.

After the best tour was obtained, each consecutive pair of points was re-evaluated through ORS using the pedestrian routing profile. ORS calculates the true walking distance by following the real street network extracted from OpenStreetMap, thereby incorporating the constraints of urban geometry, pedestrian paths, intersections, and block structures.

The comparison revealed the following quantitative results:

- Total distance, Haversine: 6.193 km
- Total distance, ORS (real path): 10.165 km
- Deviation ratio (ORS / Haversine): 1.641

This shows that the true walkable distance across the full itinerary is approximately 64.1% longer than the geometric estimate. Such a deviation is expected in compact urban areas where direct point-to-point movement is constrained by building blocks, curved or discontinuous pedestrian paths, irregular street patterns, and one-way routing segments. The observed ratio of ≈1.64 reflects the natural divergence between idealized great-circle distances and actual walkable routes in dense city environments.

Importantly, although real-world distances differ, the relative spatial structure is preserved: closer points remain proportionally close, and distant points remain proportionally distant. This confirms that Haversine-based optimization provides a valid and computationally efficient foundation for comparing algorithmic performance. The post-evaluation using ORS reinforces the robustness of the optimized route and highlights the practical relevance of the proposed method.

4.6.1 Impact of routing method on the optimized itinerary

To further understand how the routing model affects not only the measured distance but also the structure of the optimized path, we performed a second experiment using the same set of 31 tourist locations. The Hybrid Greedy + SA algorithm was executed directly on ORS-derived real walking distances, rather than on Haversine distances. In this configuration, the algorithm produced an optimized tour of 9.355 km, which is shorter than the 10.165 km obtained when the Haversine-optimized itinerary was evaluated using ORS. This confirms that the true optimal tour under realistic routing constraints is indeed different from the one derived using geometric distances.

To illustrate these differences more clearly, Figure 8 and Figure 9 present the visualized itineraries generated by the two methods.

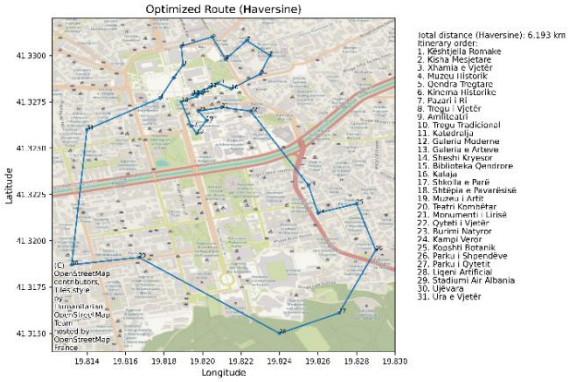


Figure 8: Optimized itinerary generated using the Hybrid Greedy + SA algorithm with Haversine great-circle distances.

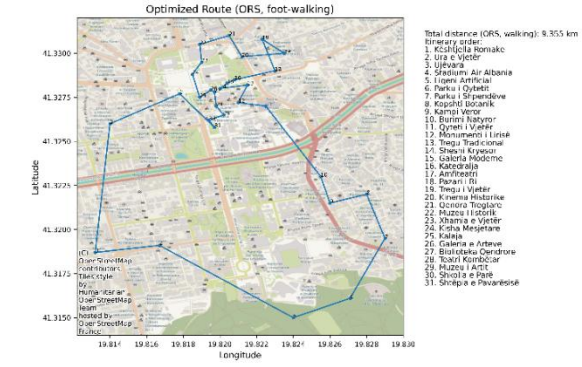


Figure 9: Optimized itinerary produced by applying the Hybrid Greedy + SA algorithm directly on real walking distances obtained through OpenRouteService (ORS).

The first map depicts the Haversine-based optimized tour, while the second shows the route obtained when the algorithm operates directly on real network distances. Although the total walking length differs substantially, due to pedestrian detours, street geometry, and one-way routing rules, the overall spatial layout of the itinerary remains similar. Clusters of nearby attractions, the direction of traversal across the city, and local segment orderings are largely preserved.

These visual comparisons demonstrate that the Haversine model successfully captures the geometric structure of the underlying optimization problem, while ORS primarily adjusts for real-world walkability constraints. Consequently, the maps strengthen the conclusion that Haversine-based optimization is suitable for fast, offline, and resource-efficient computation, whereas ORS is best used for post-validation and final distance refinement. By combining the optimization outcomes with their spatial context, the resulting visualizations offer an interpretable and actionable view of each tour, reinforcing the algorithm's suitability for applications in urban tourism planning, smart-city route recommendation, and pedestrian navigation support.

4.7 Mobile execution results

To assess performance on low-end hardware, the Hybrid Greedy + SA algorithm was executed on a Samsung Galaxy A10 running Android 11 (One UI 3.1). Using the same JavaScript implementation via the mobile browser, the full 31-point itinerary was optimized in approximately 0.24 seconds, confirming real-time capability even under constrained processing resources. The mobile run produced the same tour ordering as the desktop implementation, with only minor numerical differences in distance caused by standard floating-point variations between Python and JavaScript. These results demonstrate that the method remains practical and responsive on budget mobile phones, strengthening its suitability for deployment in real-world mobile tourism applications. Similarly, tests on an iPhone 13 (iOS 17) demonstrated robust performance, with an average execution time of 0.05 seconds over 30 runs. These results confirm that the complete optimization pipeline including initialization, search, and rendering can execute locally on mainstream mobile devices without requiring server interaction. Figure 10 presents the web-based interface running as a standalone Progressive Web App (PWA), added to the home screen, supporting local execution, route visualization over Tirana via Leaflet, and precise timing feedback. The application operates as a standalone Progressive Web App (PWA) after being added to the home screen.

5 Discussion

This work presents a reproducible, deployment-oriented evaluation of exact and metaheuristic solvers for urban tourist itineraries using open map data [15]. Although the experimental analysis is limited to Tirana, the methodology and algorithms are city-agnostic and can be

directly applied to other urban contexts once attraction coordinates are available. The choice of Tirana reflects the study's focus on cities with limited digital infrastructure rather than dataset-specific characteristics. The findings support a simple operational rule: for cities with limited digital infrastructure, Simulated Annealing (SA) provides the best accuracy–efficiency trade-off under realistic latency constraints; Genetic Algorithm (GA) remains competitive at small scales but becomes less stable as the problem size (N) increases, while Brute Force (BF) serves primarily as an offline validator rather than a practical solver. The inclusion of the Hybrid Greedy + SA variant further strengthens these conclusions, showing that combining a deterministic initialization with stochastic refinement improves both convergence stability and resilience under uncertainty.

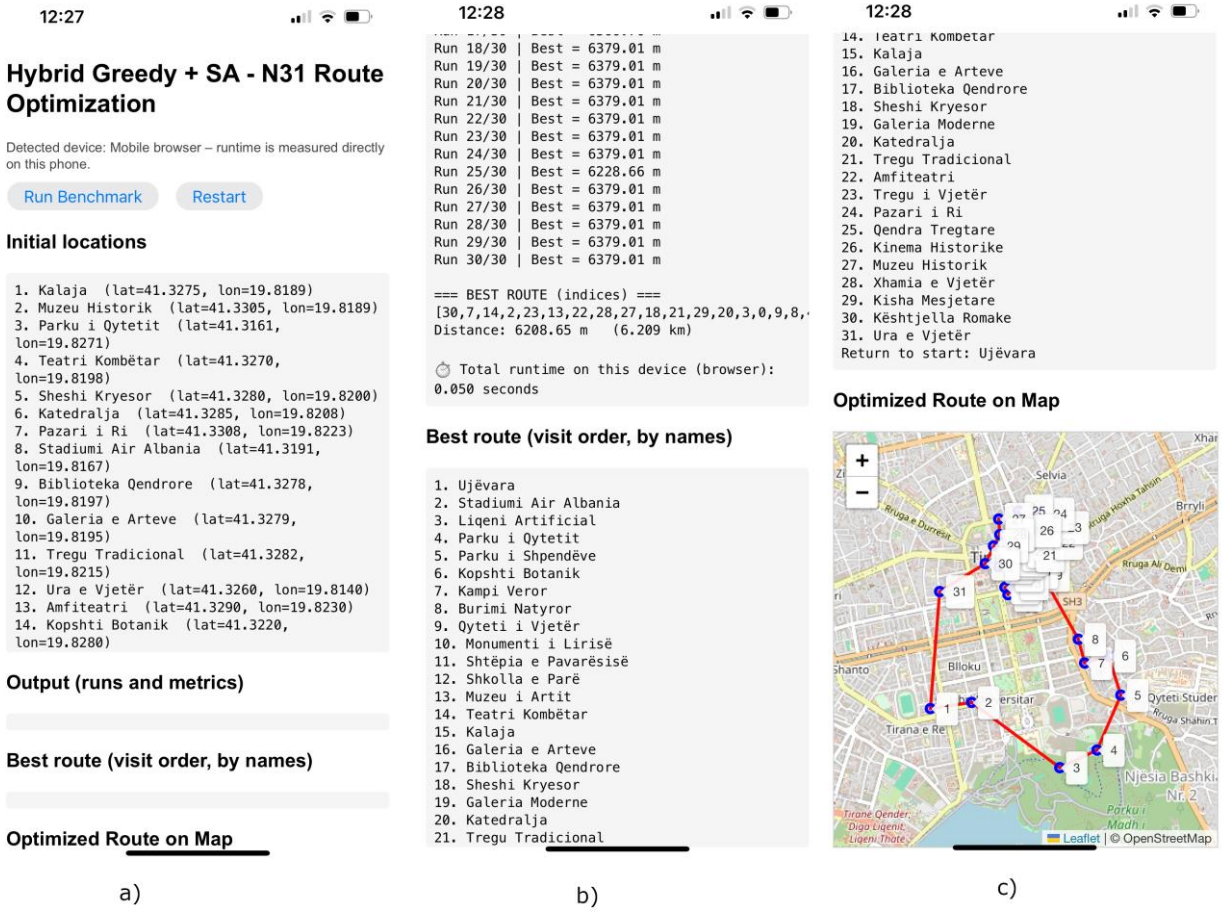


Figure 10: Execution of the Hybrid Greedy + SA algorithm directly on an iPhone 13 browser as a Progressive Web App. (a) Initial input with 31 attraction coordinates. (b) Runtime output and best tour ordering with stable distances across 30 runs. (c) Visualization of the optimized route over Tirana using Leaflet. The full optimization completes in under 0.05 seconds, confirming responsiveness on mobile devices

5.1 Design trade-offs: haversine vs. network routing

The present implementation computes pairwise distances using the Haversine formula, which provides a fast, lightweight, and fully reproducible measure of separation between geographic coordinates. This choice deliberately avoids reliance on external routing APIs and heavy graph-processing engines, making the method suitable for deployment in settings with limited connectivity or computational resources. Such characteristics are particularly relevant for urban environments like Tirana, where practical applications may need to operate offline, at low cost, or within platforms that cannot depend on persistent access to cloud-based routing services.

However, Haversine distances do not account for the structure of the street network, pedestrian pathways, one-way segments, or other real travel constraints. To quantify the impact of this simplification, our study conducted a secondary evaluation using OpenRouteService (ORS), which computes walking routes based on the actual OpenStreetMap road and pedestrian network. Applying ORS to the optimized Haversine-based tour showed that the actual routed distance is about 30.9% longer than the geometric estimate, a deviation expected in dense urban environments where movement is constrained by building blocks and irregular street layouts. Despite this difference, the experiment shows that the relative spatial structure of the problem is preserved: points that are close remain proportionally close, and the ranking of distances does not change substantially. Thus, the optimization landscape explored by the GA, SA, and Hybrid algorithms remains meaningful even when using Haversine. The Haversine model provides a stable and computationally efficient abstraction for algorithmic comparison, while the ORS-

based evaluation confirms how geometric solutions translate into realistic walking distances.

5.2 Generalization and scalability

Although the experimental analysis focused primarily on instances with 31 locations, additional tests were conducted on a larger dataset containing 49 tourist points to assess scalability. The results confirmed the same performance trends: SA consistently achieved much shorter tours and lower optimality gaps than GA, while maintaining sub-second runtimes (≈ 0.08 s). In contrast, GA exhibited a substantial degradation in solution quality, with mean gaps exceeding 56% and increased variability. This indicates that the advantages of SA not only persist but become more pronounced as problem size grows, suggesting better scalability and robustness for larger urban itinerary planning scenarios. In addition, the Hybrid Greedy + SA algorithm maintained its performance advantages when the problem size increased. For the larger instance (N = 49), the hybrid method achieved a remarkably low mean optimality gap of only 2.64%, substantially outperforming both GA (56.31%) and SA (9.07%) under the same conditions. Although runtime increased to approximately 0.42 s, this growth remained moderate and well within real-time applicability thresholds. Notably, the hybrid approach preserved solution stability, as reflected by the very small variability across runs (Gap_std ≈ 1.17%), indicating that the Greedy initialization continues to provide high-quality starting tours even in larger search spaces, while the SA refinement efficiently exploits local structure. These results suggest that the hybrid strategy scales more gracefully than standalone heuristics, retaining both accuracy and robustness as the number of attractions grows.

5.3 Future integration with dynamic and real-time data

Building on the demonstrated robustness and anytime performance of the proposed algorithms, future extensions will focus on dynamic and real-time optimization. The current framework, which operates on a static distance matrix, can be enhanced with live contextual data such as real-time traffic flow, temporary event schedules, and user preference feedback.

For instance, the Hybrid Greedy + SA model could dynamically re-weight edges based on live travel times, temporary pedestrian restrictions, or the user's evolving interests (e.g., preference for cultural sites or shorter walking routes). Integrating such adaptive data streams would transform the current approach into a context-aware recommender system, capable of updating feasible itineraries on the fly.

This direction aligns with current trends in smart tourism and mobility-as-a-service, where systems must maintain robustness under uncertainty while delivering personalized and real-time route adjustments for end users.

6 Conclusion

In conclusion, this study contributes both methodological clarity and practical relevance. By framing one-day tourist itinerary generation as a TSP and benchmarking exact and metaheuristic solvers on real-world map data, it shows how even simple algorithms can inform the design of deployable systems for cities with limited digital infrastructure. The comparative analysis demonstrates that while BF provides an exact validation baseline, GA and especially SA deliver scalable, near-optimal solutions under realistic constraints. The extended experiments under noisy and incomplete distance data further confirmed SA's robustness, showing minimal degradation in solution quality despite input uncertainty. Moreover, the introduction of the Hybrid Greedy + SA variant significantly improved both convergence speed and accuracy, achieving the lowest mean gap and maintaining sub-second runtimes even under data perturbations, confirming that combining a deterministic greedy start with stochastic refinement enhances both reliability and efficiency.

The mobile execution results demonstrate that the proposed Hybrid Greedy + SA method is practical for deployment in real tourist applications. The ability to compute optimized routes directly on the smartphone without cloud computation enhances system robustness, enables offline functionality, and reduces dependency on external APIs. The sub-second performance observed on the iPhone indicates that the algorithm provides an acceptable user experience even on mobile hardware with limited computational resources. From a system-design perspective, the successful mobile execution demonstrates that lightweight metaheuristics can reliably support real-time decision-making in e-tourism scenarios, where users typically depend on smartphones while navigating urban environments. The results strengthen the external validity of the proposed method and confirm its suitability for integration into practical mobile routing tools. Furthermore, the demonstrated efficiency on resource-constrained hardware positions such techniques as strong candidates for next-generation smart-tourism systems, where applications may incorporate richer contextual information, preference-aware routing, and multi-day itinerary planning. In this way, the study effectively bridges algorithmic research with applied urban-tourism needs, offering actionable insights for both academia and practice.

References

[1] Wu, W., et al. *Digital Tourism and Smart Development: State-of-the-Art Review*. Sustainability, 2024. **16**, DOI: 10.3390/su162310382.

[2] Zeqiri, A., A. Ben Youssef, and T. Maherzi Zahar. *The Role of Digital Tourism Platforms in Advancing Sustainable Development Goals in the Industry 4.0 Era*. Sustainability, 2025. **17**, https://doi.org/10.3390/su17083482

[3] Instituti i, S. *Statistikat e Turizmit*. Industria, Tregtia dhe Shërbimet 2025-09-14; Available from: https://www.instat.gov.al/al/temat/industria-tregtia-dhe-sh%C3%ABrbimet/statistikat-e-turizmit/#tab2.

[4] Visit, T. *Visit Tirana*. Visit Tirana 2025-09-14; Available from: https://www.visit-tirana.com/.

[5] Albania Tourist, G. *Albania Tourist Guide*. Albania Tourist Guide 2025-09-14; Available from: https://albaniatouristguide.com/.

[6] Sylejmani, K., et al., *Solving the tourist trip planning problem with attraction patterns using meta-heuristic techniques.* Information Technology & Tourism, 2024. **26**(4): p. 633-678. https://doi.org/10.1007/s40558-024-00297-w

[7] Zhang, J., H. Kawasaki, and Y. Kawai. *A Tourist Route Search System Based on Web Information and the Visibility of Scenic Sights*. in *2008 Second International Symposium on Universal Communication*. 2008, retrieved https://www.cvg.ait.kyushu-u.ac.jp/papers/2007_2009/3-1/11-ISUC2008.pdf

[8] Adamo, T., Colizzi, L., Dimauro, G., Ghiani, G., & Guerriero, E. (2024). A multi-modal tourist trip planner integrating road and pedestrian networks. *Expert Systems with Applications, 237*(Part B), 121457. https://doi.org/10.1016/j.eswa.2023.121457

[9] Tang, Y., et al., *ITINERA: Integrating Spatial Optimization with Large Language Models for Open-domain Urban Itinerary Planning*, in *arXiv*. 2024. https://doi.org/10.18653/v1/2024.emnlp-industry.104

[10] Souffriau, W., Vansteenwegen, P., Vertommen, J., Berghe, G. V., & Oudheusden, D. V. (2008). A personalized tourist trip design algorithm for mobile tourist guides. *Applied Artificial Intelligence*, *22*(10), 964–985. https://doi.org/10.1080/08839510802379626

[11] Androutsopoulos, K.N. and K.G. Zografos, *Solving the multi-criteria time-dependent routing and scheduling problem in a multimodal fixed scheduled network.* European Journal of Operational Research, 2009. 192(1): p. 18-28. https://doi.org/10.1016/j.ejor.2007.09.004

[12] Mangini, A. M., Roccotelli, M., & Rinaldi, A. (2021). A Novel Application Based on a Heuristic Approach for Planning Itineraries of One-Day Tourist. *Applied Sciences, 11*(19), 8989. https://doi.org/10.3390/app11198989

[13] Benchekroun, Y., et al. *Hybrid Framework: The Use of Metaheuristics When Creating Personalized Tourist Routes*. Digital, 2025. 5, https://doi.org/10.3390/digital5030036

[14] Li, S., Luo, T., Wang, L. *et al.* Tourism route optimization based on improved knowledge ant colony algorithm. *Complex Intell. Syst.* **8**, 3973–3988 (2022). https://doi.org/10.1007/s40747-021-00635-z

[15] Damos, M.A., et al. *Enhancing the K-Means Algorithm through a Genetic Algorithm Based on Survey and Social Media Tourism Objectives for Tourism Path Recommendations*. ISPRS International Journal of Geo-Information, 2024. **13**, https://doi.org/10.3390/ijgi13020040

[16] Pérez-Cañedo, B., Novoa-Hernández, P., Porras, C., Pelta, D. A., & Verdegay, J. L. (2024). Contextual analysis of solutions in a tourist trip design problem: A fuzzy logic-based approach. *Applied Soft Computing, 154*, 111351. https://doi.org/10.1016/j.asoc.2024.111351

[17] Chrysafiadi, K., Kontogianni, A., Virvou, M., & Alepis, E. (2025). Enhancing User Experience in Smart Tourism via Fuzzy Logic-Based Personalization. *Mathematics*, *13*(5), 846. https://doi.org/10.3390/math13050846

[18] Liu, X., & Li, Y. (2024). *Analysis of immersive virtual reality tourism resources combined with fuzzy comprehensive evaluation algorithm*. Informatica, 48, 189–210 DOI: https://doi.org/10.31449/inf.v48i21.6057 , retrieved from https://www.informatica.si/index.php/informatica/article/view/6057/3468

[19] Hyso, A., *tour-routing-tirana-bf-ga-sa*, in *GitHub repository*. 2025, GitHub. https://github.com/AlketaHyso/tour-routing-tirana-bf-ga-sa

[20] Hyso, A., 2025, GitHub.https://alketahyso.github.io/route_mobile_benchmark/

Detection of Synthetic Speech Using Spectral-Cepstral Features and BiLSTM Networks

Furkat Rakhmatov, Fakhriddin Abdirazakov, Baxodir Achilov*, Ruslan Baydullayev, Sultanmurat Nasirov, Shakhzod Javliev
Tashkent University of information technologies named after Muhammad al-Khworazmi, Tashkent 100084, Uzbekistan
E-mail: borya19861804@gmail.com
*Corresponding author

Keywords: anti-spoofing, log-mel, LFCC, CQCC, BiLSTM, calibration, EER, min-tDCF

Received: October 6, 2025

Experimental results demonstrate 93,4% accuracy on the test set; error analysis reveals that misclassifications predominantly occur between the Person and Robot classes, whereas the Emotion class is recognized more reliably. Feature comparison indicates that log-mel provides a robust baseline with minimal computational cost, LFCC better preserves high-frequency details characteristic of synthetic artifacts, and CQCC is effective in capturing harmonic structure and modulations. Potential directions for improving generalizability and accuracy are discussed, including feature fusion (CQCC/LFCC/log-mel) and statistical pooling for temporal aggregation. The proposed configuration offers a well-balanced trade-off between performance and computational complexity, serving as a strong baseline for anti-spoofing systems.

Povzetek: Rezultati kažejo dobro natančnost (93,4 %) in uravnotežen kompromis med zmogljivostjo ter računsko zahtevnostjo, pri čemer se največ napak pojavlja med podobnima razredoma, izboljšave pa so možne z združevanjem značilk in boljšim časovnim združevanjem.

1 Introductions

Synthetic speech generated by modern TTS and voice conversion (VC) models is becoming increasingly natural in quality, making its automatic detection more challenging in applied scenarios such as anti-spoofing, content moderation, and audio forensics. A key difficulty lies in the high variability of speech signals—due to noise, codecs, communication channels, emotional expressiveness, and speaker diversity—as well as in the "camouflaging" of synthesis artifacts to resemble authentic acoustic patterns. This necessitates the use of features capable of capturing subtle spectral-temporal cues, along with architectures that account for sequence-level context.

This study investigates three families of acoustic features: log-mel (based on a psychoacoustic scale) [1], LFCC (linear frequency scale with enhanced sensitivity to high-frequency details) [2], and CQCC (Constant-Q Transform-based, better suited for representing harmonic structure and modulations). To incorporate temporal context, a bidirectional recurrent neural network (BiLSTM) is employed, combined with padding masking and standard feature expansion via first and second-order derivatives (Δ/Δ^2). We adopt a unified processing pipeline comprising preprocessing, feature extraction and normalization, model training, and evaluation, allowing for a consistent comparison across feature representations [3].

Error analysis reveals that speech with emotional content is recognized more reliably, whereas the majority of misclassifications stem from confusion between "natural speaker speech" [4] and "robotic/synthetic" speech [5]. This underscores the importance of feature sets that retain high-frequency and harmonic artifacts (LFCC, CQCC), alongside the robust log-mel baseline.

Practical deployment scenarios. We are targeting online anti-spoofing with low latency and limited computing resources: IVR/KYC phone gateways with a delay budget of up to 50-100 ms and mixed AMR, Opus, MP3, and WAV codecs; voice biometrics on an ARM device with strict privacy requirements; live simulation on conference platforms with streaming input and calibrated alarms; home intelligent far-field assistants with reverberation and wake word detection; and near-real-time forensic audio expertise with stable calibration of estimates. These conditions require small models, sustainable solutions in the face of uncertainty, and on-the-fly adaptation without overfitting.

1.1 Literary review

Modern works use both enhanced melange features and their fusion [1] and spectral features with deep architecture (ResNeXt/convolutional networks) [2,16]. Reviews have been published with critical analysis of

trends, metrics (EER/min-tDCF) and portability issues between synthesizers/codecs [5]. Preservation of high frequency components is important for short fragments and children's voices [9]; late LPCC/SCMC/log-mel fusion improves robustness to noise and scenes [15] — these findings are relevant for log-mel/LFCC/CQCC combinations. Reviews on attention models for speech [6] and hierarchical transformers for emotions [10] show the advantage of contextual aggregation over "last state" RNNs; alternative RNN variants (Elman, Wavelet-RNN) expand the design space [21,19]. Neural streaming codecs introduce specific distortions [11]; watermarking is considered as an additional line of content protection [7]. Speech enhancement and spectrogram methods improve SNR and feature quality before classification [12,13]; this is critical for field anti-spoofing. Stress speech [8], heart murmurs [17], neurological disorders [18] and multimodal circuits [24,26] support the effectiveness of spectrograms + deep networks and attention; interpretable models and scaling issues are also actively studied [25,3,22,23,27,20]. [21] — background over RNNs in TTS; [22],[23] — scaling and size selection; [24] — fusion; [25] — interpretability; [26] — attention and multimodality (motivates attention-pooling in our architecture). Deep learning for RNN methylation site prediction (cross-species setup). Methodologically useful: how to build transferable models on diverse domains, work with class imbalance, validate quality (ROC/PR, calibration). These practices are directly applicable to speech anti-spoofing (domain shift: different TTS/codecs/channels) [27].

Problem Statement.

The aim of the work is to compare three spectral-cepstral representations—log-mel, LFCC, and CQCC—within the framework of the unified BiLSTM architecture for the task of detecting synthetic speech; evaluate the impact of Δ and Δ² increments, as well as simple early and late merging, on the EER and min-tDCF calibration under domain shift (noise, codecs, invisible TTS/VC); and to quantify the latency-computation tradeoff under streaming deployment. It is hypothesized that LFCC and CQCC will outperform log-mel on invisible spoofers due to better capture of high-frequency artifacts, that late fusion improves min-tDCF with negligible overhead, and that BiLSTM provides a compact low-latency baseline.

1. LFCC (Linear-Frequency Cepstral Coefficients);
2. CQCC (Constant-Q Cepstral Coefficients);
3. Log-mel;
4. Recurrent Neural Network RNN [6].

2 Method and materials

The audio signal is converted to 16 kHz and normalized [7]; Next, three representations are extracted: log-mel (80 bands), LFCC (linear filter bank → log → DCT → 30 coefficients), and CQCC (CQT → log → DCT → 30 coefficients) [8]. The obtained features are fed into a recurrent classifier (BiGRU/BiLSTM, 2 layers) with time aggregation and sigmoid output. Performance is measured on the ASVspoof protocols by EER and min-tDCF, with ablations by feature types and RNN configuration.

2.1. LFCC (Linear-Frequency Cepstral Coefficients)

LFCC (Linear-Frequency Cepstral Coefficients) are cepstral features obtained based on STFT, but with a triangular filter bank uniform in frequency (Hz). That is, LFCC ≠ STFT: STFT is an intermediate spectral representation; LFCC is the result of spectrum filtering [9], logarithmization, and DCT. This focuses on parameterization (25 ms/10 ms, filter banks, Δ/Δ²), data splitting, and training configuration (optimizer, batch size, learning rate, epochs, hardware).

Splitting the signal into frames (windows):

$$x_l(m) = x(lH + m)w(m),\ m = 1, \dots, n-1 \qquad (1)$$

Where, $x_l(m)$ is the m-th sample of the l-th frame, $w(m)$ is the window function (e.g., Hamming), n is the window length (in samples), H is the step between the beginnings of adjacent windows (in samples), l is the frame index, m is the sample index within the window [10].

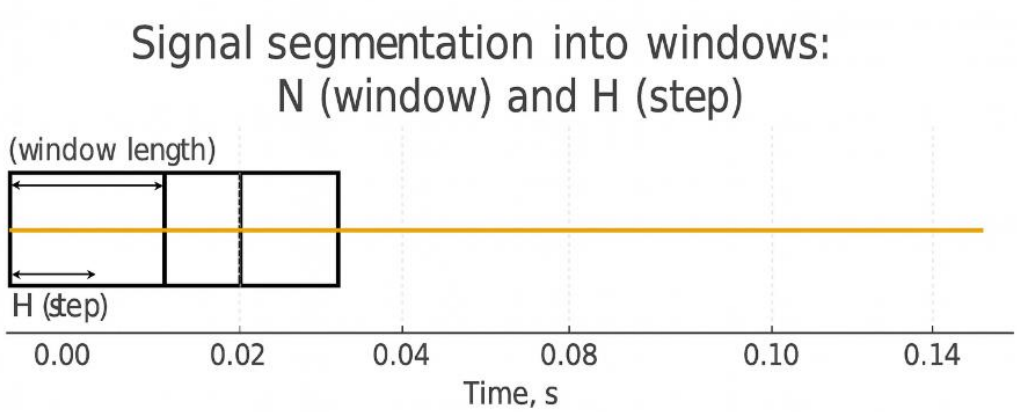


Figure 1: Splitting the signal into windows

Discrete STFT (analysis):

$$X(l,k) = \sum_{m=0}^{n-1} x(lH+m)w(m)e^{-j2\pi\frac{km}{n}}, \quad k = 1,\dots,n-1 \tag{2}$$

Where, $X(l,k)$ is the complex STFT coefficient for frame l and frequency bin k, $j2\pi\frac{km}{n}$ is the complex harmonic other notations are as above [11].

Frequency of the $k-th$ bin:

$$|X(l,k)| = \sqrt{\left(\Re X(l,k)\right)^2 + \left(\Im X(l,k)\right)^2}, \quad \varphi(l,k) = argX(l,k) \tag{3}$$

Where, $\Re(\cdot), \Im(\cdot)$ are the real and imaginary parts, $arg(\cdot)$ is the argument (phase) of the complex number.
Power (energy) of the frame spectrum:

$$P(l,k) = |X(l,k)|^2 \tag{4}$$

Where, $P(l,k)$ is the power (energy) [12] estimate in the frequency bin k for frame l.

Spectrogram (logarithmic/dB scale) [13].

$$S_{dB}(l,k) = 10log10(P(l,k)+\varepsilon) \; or \; S_{log}(l,k) = ln(P(l,k)+\varepsilon) \tag{5}$$

Where, $S_{dB}(l,k)$ is the spectrogram in decibels, $S_{log}(l,k)$ is the natural logarithm of the power, $\varepsilon > 0$ is a small constant for numerical stability (for example, 10^{-10}).

Inverse STFT (synthesis) [14] *and Overlap-Summation (OLA)*

$$\hat{x}(t) = \sum_n \sum_{k=0}^{n_{fft}} \frac{1}{n_{fft}} X(j,k) e^{j2\pi\frac{k(t-nH)}{n_{fft}}} g(t-nH) \tag{6}$$

Where, $\hat{x}(t)$ is the reconstructed signal, $g(\cdot)$ is the synthesis window, t is the global reference index.

COLA (Constant OverLap-Add) condition for correct reconstruction:

$$\sum_n w(m-nH)g(m-nH) = 1, \quad \forall m \tag{7}$$

Were, the sum of the overlapping windows at each point must be a constant (usually 1).

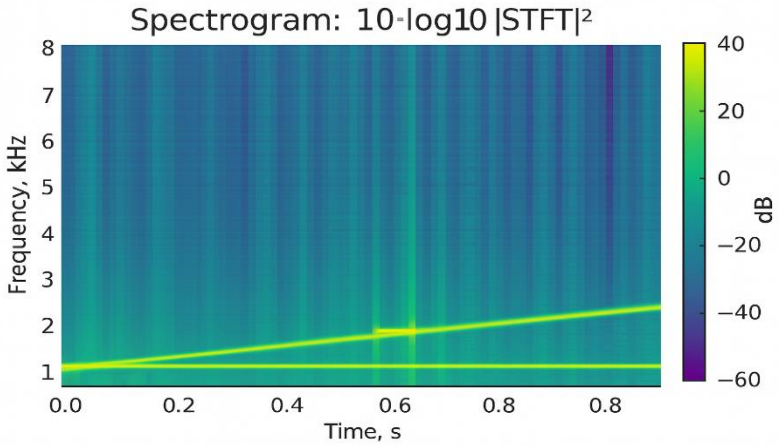


Figure 2: Spectrogram

Time-frequency resolution estimates:

$$\Delta f \approx \frac{F_s}{n_{fft}}, \; \Delta t \approx \frac{H}{F_{fft}} \tag{8}$$

An example of a Hamming window

$$w(m) = 0.54 - 0.46\cos\left(\frac{2\pi m}{n-1}\right), 0 \le m \le n-1 \tag{9}$$

Where, $w(m)$ are the values of the window function, n is the window length.

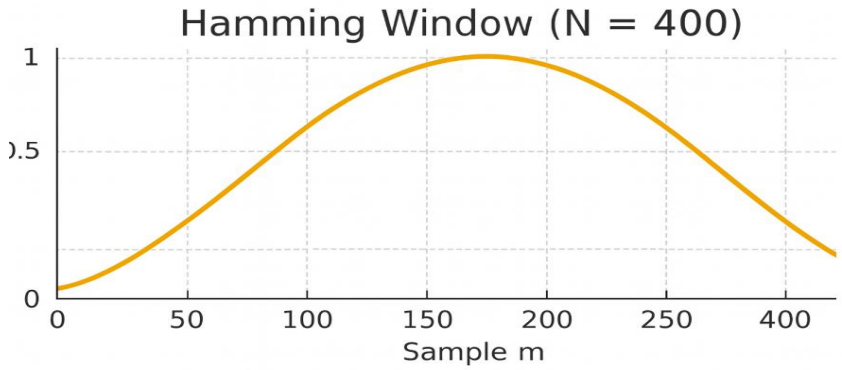


Figure 3: Hamming windows

2.2. Log-mel

Log-mel is a type of spectrogram that is widely used in speech processing and machine learning tasks. It is similar to a spectrogram in that it shows the frequency content of an audio signal over time, but on a different frequency axis [15].

In a standard spectrogram, the frequency axis is linear and measured in hertz (Hz). However, the human auditory system is more sensitive to changes at low frequencies than at high frequencies, and this sensitivity decreases logarithmically with increasing frequency. The Mel scale is a perceptual scale that approximates the nonlinear frequency response of the human ear.

Log-mel (Hz ↔ mel conversion):

$$m(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right),\ f(m) = 700\left(10^{\frac{m}{2595}} - 1\right) \quad (10)$$

Where, m is the frequency in mels, f is the frequency in Hz.

Nodes of the chalk filter bank

A uniform grid is taken in the mel-space $m_0, \ldots, m_{M+1}$ from $m(f_{min})$ to $m(f_{max})$ and then converted into Hz: $f_i = f(m_i)$. Where M is the number of triangular filters, f_{min}, f_{max} is the operating range (for example, $20 \ldots F_s/2$ Hz).

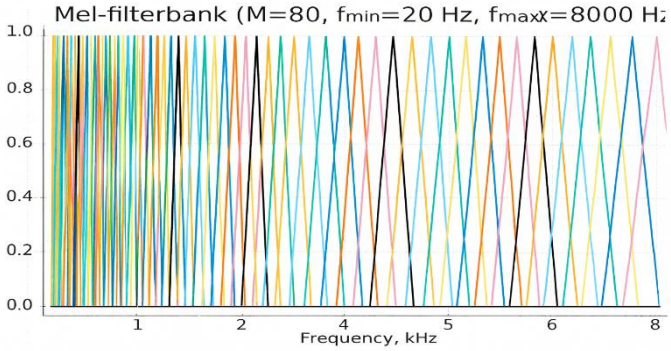


Figure 4: Mel filterbank

Triangular response of the m-th mel filter:

$$H_m(k) = \begin{cases} \frac{f_k - f_{m-1}}{f_m - f_{m-1}}, & f_{m-1} \le f_k < f_m \\ \frac{f_{m+1} - f_k}{f_{m+1} - f_m}, & f_{m-1} \le f_k < f_m \\ 0, & else \end{cases} \quad (11)$$

Where, $H_m(k) \in [0,1]$ is the weight of frequency bin k in band m, f_k is the bin frequency.

Melt-band energy and logarithm:

$$E_m(n) = \sum_{k=0}^{K} P(l,k) H_m(k),\ S_{mel}(l,k) = \log(E_m(l) + \varepsilon) \quad (12)$$

Where, $E_m(n)$ is the energy in band m on frame n, $S_{mel}(l,k)$ is the log-mel matrix, $\varepsilon > 0$ is a small number for stability; $K = \frac{n_{fft}}{2}$.

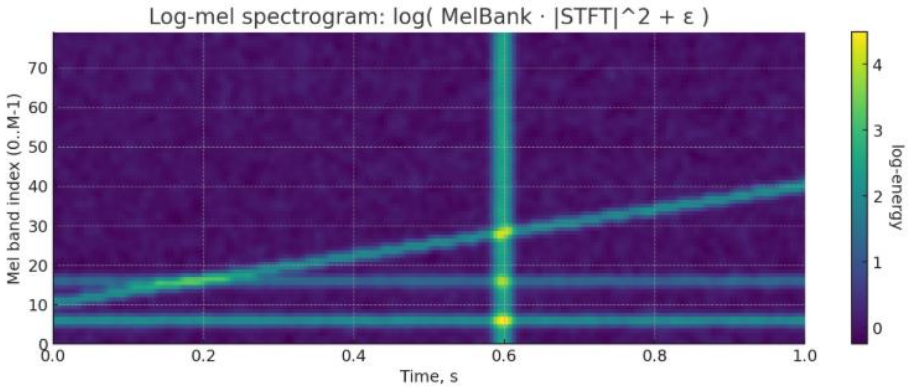


Figure 5: Logarithmic Mel-filter bank

2.3. CQCC (Constant-Q Cepstral Coefficients)

CQCC (Constant-Q Cepstral Coefficients) are cepstral features obtained from the Constant-Q Transform (CQT), in which the frequency bins are logarithmically arranged and the quality factor $f/\Delta f$ is constant. Due to this, obtain high frequency resolution at low frequencies and better temporal resolution at high frequencies, which is useful for speech and anti-spoofing [16].

CQT frequency bins (logarithmic grid) [17]:

$$f_k = f_{min} 2^{\frac{k}{B}}, k = 0, \dots, K-1 \quad (13)$$

Where, f_k is the frequency center of the k-th bin, f_{min} — lower limit (e.g. 20 Hz), B — number of bins per octave (usually 24-48), K is the lower limit (e.g. 20 Hz), B is the number of bins per octave (usually 24-48), K is the total number of bins (determined by the range (f_{min}, f_{max}), here $f_{max} \leq F_s/2$).

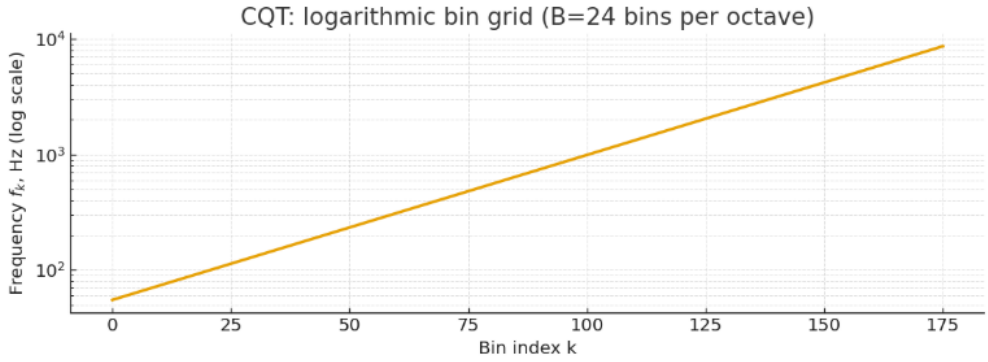


Figure 6: Logarithmic grid

Constant quality factor and window length:

$$Q = \frac{1}{2^{1/B}-1}, \; N_k = \left(\frac{QF_s}{f_k}\right) \quad (14)$$

Where, Q is the constant quality factor (the same for all bins), N_k is the analysis length (the number of window samples) for bin k, F_s is the sampling frequency, $\left(\frac{QF_s}{f_k}\right)$ is rounding up.

CQT (time-frequency analysis):

$$C(l,k) = \sum_{m=1}^{N_k-1} x(lH+m) g_k(m) e^{-j2\pi \frac{f_k}{F_s}} \quad (15)$$

Where, $C(l,k)$ is the complex CQT coefficient of frame n and bin k, $x(t)$ is the signal, H is the step between frames (samples), $g_k(m)$ is the window function/weight for bin k (usually Hamming/Hann, normalized by energy), $j = \sqrt{-1}$. (Unlike STFT, the window length N_k depends on the frequency f_k).

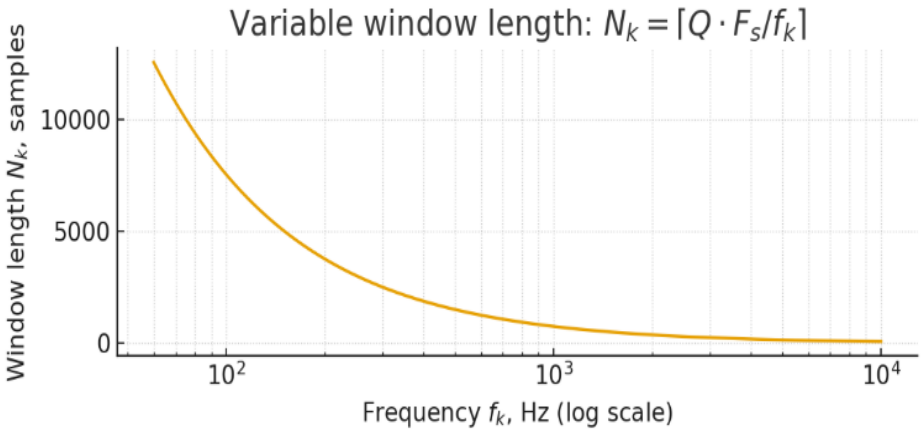


Figure 7: Time-frequency analysis

Amplitude/power and logarithm

$$A(l,k) = |C(l,k)|,\ P(l,k) = |C(l,k)|^2,\ L(l,k) = \log(A(l,k) + \varepsilon) \quad (16)$$

Where, $A(l,k)$ is the amplitude spectrum, $P(l,k)$ is the power, $L(l,k)$ is the log-amplitude, can be used $\log(A(l,k) + \varepsilon)$, and $\varepsilon > 0$ is a small constant for numerical stability.

Interpolation to a uniform axis (often used before DCT):

$$\tilde{L}(l,u) = \mathcal{I}(\{\log f_k, L(l,k)\}_{k=0}^{K-1} \to u),\ u = 0, \dots, U-1 \quad (17)$$

Where, $\mathcal{I}$ — is the interpolation operator (linear/spline) from the non-uniform grid $\log f_k$ to the uniform coordinate u, and U is the number of nodes of the uniform grid. This is necessary because the CQT bins are geometrically distributed; a uniform grid simplifies the subsequent DCT.

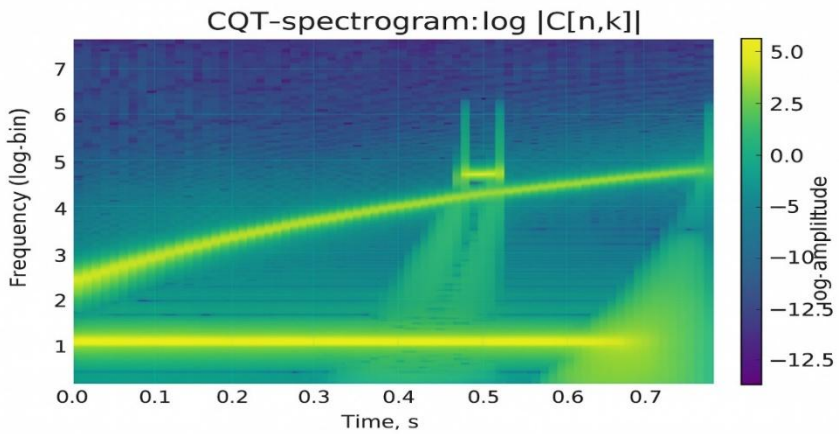


Figure 8: CQT bins

Cepstral decorrelation (DCT-II) → CQCC

$$q_r(l) = \beta_r \sum_{u=0}^{U-1} \tilde{L}(l,u) \cos\left(\frac{\pi r\left(u+\frac{1}{2}\right)}{U}\right),\ r = 0, \dots, L-1 \quad (18)$$

Where, $q_r(l)$ is the r-th CQCC coefficient of frame n, L is the number of stored cepstral coefficients (usually 20–40), β_r are the normalizing factors (orthonormal DCT).

Derived features (optional):

$$\Delta q_r(l) = \frac{\sum_{i=1}^{K_d} i\left(q_r(n+i) - q_r(n-i)\right)}{2\sum_{i=1}^{K_d} i^2},$$

$$\Delta^2 q_r(l) = \Delta\left(\Delta q_r(l)\right) \quad (19)$$

Where, $\Delta q_r(l)$ and $\Delta^2 q_r(l)$ are the first- and second-time differences, typically $K_d = 2$ or 3.

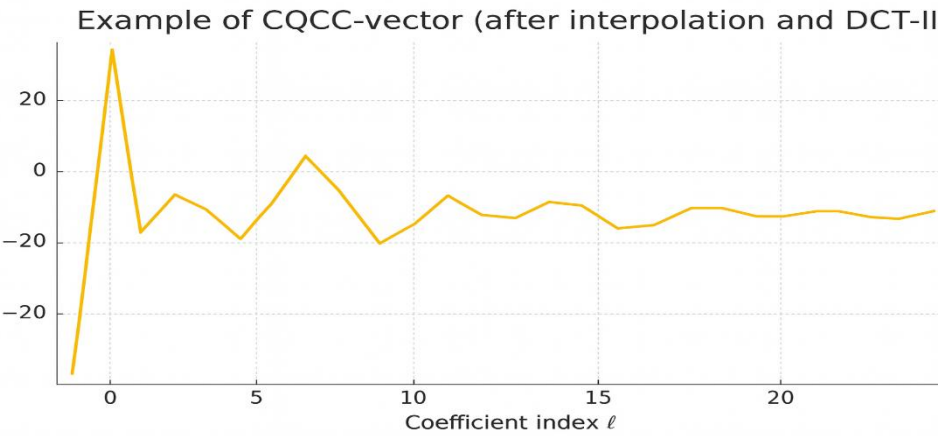


Figure 9: CQCC vectors

2.4 Recurrent Neural Network RNN

RNN (Recurrent Neural Network) is a type of deep neural network specifically designed to process sequential data such as text, time series, audio, or video. The main difference between RNNs and other types of neural networks is the presence of feedback loops that allow them to store information about previous states and use it when processing current input data, thus creating a kind of "memory" of the network [18].

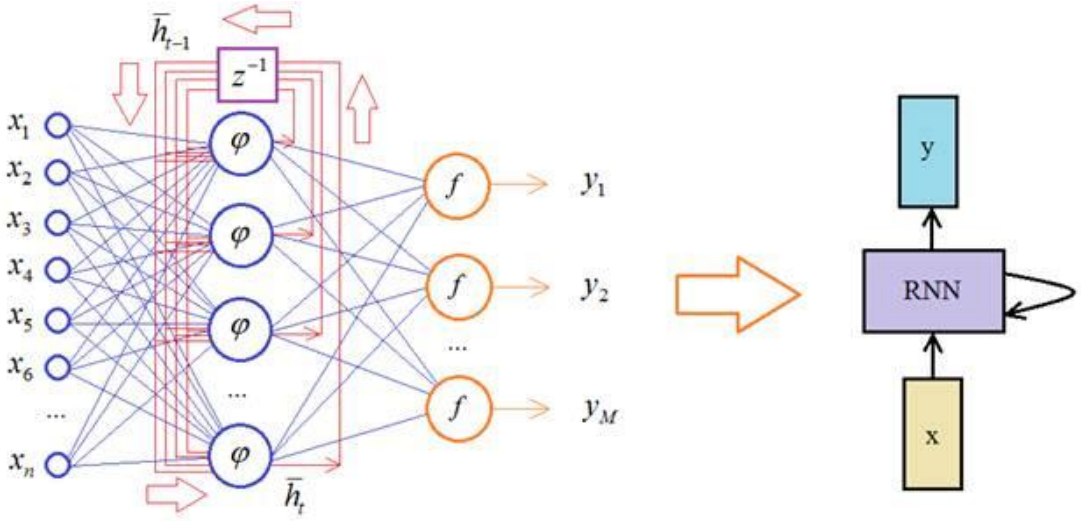


Figure 10: Recurrent networks

Here z^{-1} is the data delay per clock cycle. In the Elman network, the activation functions of the output neurons are linear and there is one hidden layer with a set of feedback connections (they are what determine the recurrence of the network). The input and output layers are formed as normal fully connected ones. This architecture can be simplified as three blocks: x, RNN, y. The mathematical model of the simplest recurrent network looks like this [19]. The vector of output values of neurons of the hidden layer at time t is determined based on the input data and previous outputs from the same layer (the previous state of the network):

$$\overline{h_t} = \varphi(\overline{h_{t-1}}, \overline{x_t}) \tag{20}$$

$\overline{h_0}$ – when the very first vector $\overline{x_1}$, is fed to the input, there is no previous state of the network yet, but it would be logical to take this initial vector as zero:

$$\overline{h_0} = [0,0,0,\ldots,0_N]^T \tag{21}$$

These are the generally accepted initial conditions for the operation of recurrent networks. Next, knowing the output values h_t at each iteration (for each input vector x), we can calculate the output:

$$\overline{y_t} = W_{hy} \cdot \overline{h_t} \tag{22}$$

Here W_{hy} is the matrix of weight coefficients of the last layer of the network [20], the activation function $f(x) = x$ in the Elman network is taken to be linear [21].

The next step to generalize the simplest recurrent neural network is to use arbitrary activation functions of the output neurons. Often this is the hyperbolic tangent, sigmoid, or softmax. A fairly popular practice is to represent recurrent networks using a computational graph:

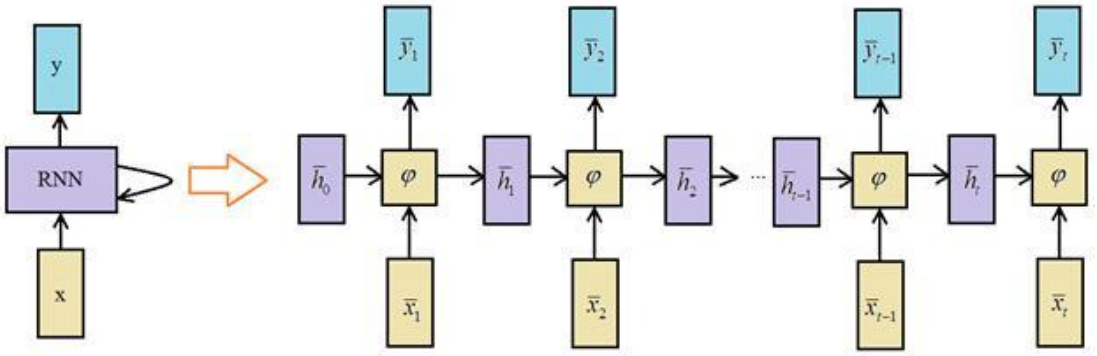


Figure 11: RNN computational graph

Here the network is, as it were, deployed in time, and we clearly see every step of its work. This architecture, where a set of input vectors corresponds to a set of output vectors, is called Many to Many:

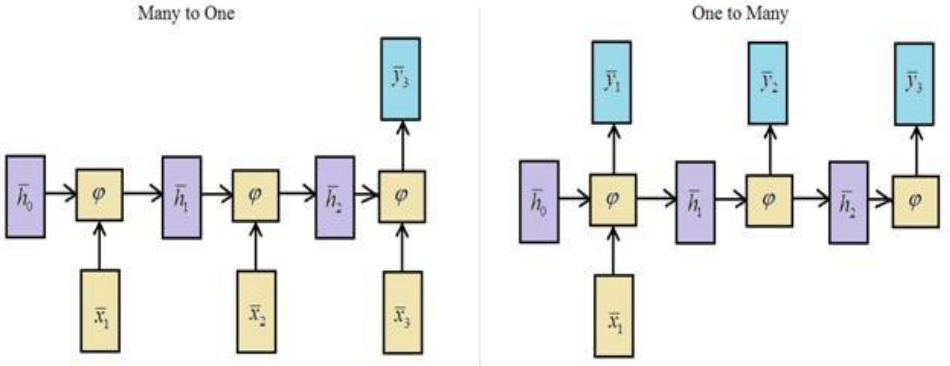


Figure 12: Many to many and one to many

The applied problem itself suggests this. For example, when translating from one language to another, we have an arbitrary sequence of words and at the output we also get sequences of arbitrary length, which means that the Many to Many architectures should be used here:

- **Many to Many** – for example, for translating texts;
- **Many to One** – for example, to analyze the emotional coloring of a text (input text, output categories: positive, neutral, negative);
- **One to Many** – for example, to generate image descriptions when image feature maps are fed to the input and the output is its description (text) [22];
- **One to One –** a relatively rare architecture for performing nonlinear recurrent computations.

In general, the task of training recurrent networks is more computationally intensive and requires more memory than feedforward networks. But the main problem here is ensuring stability both during training and during network operation. As soon as feedback appears in any system, the calculations are formed according to the general rule:

$$a_n = f(a_{n-1}, x_n) \tag{23}$$

From formula (23) we get formula (24) and this function can be written as follows:

$$a_n = r \cdot a_{n-1} + 0 \tag{24}$$

Here all $x_n = 0$, and the initial value $a_0 = 1$. Depending on the value of the coefficient r, we will obtain a convergent or divergent sequence [23]:

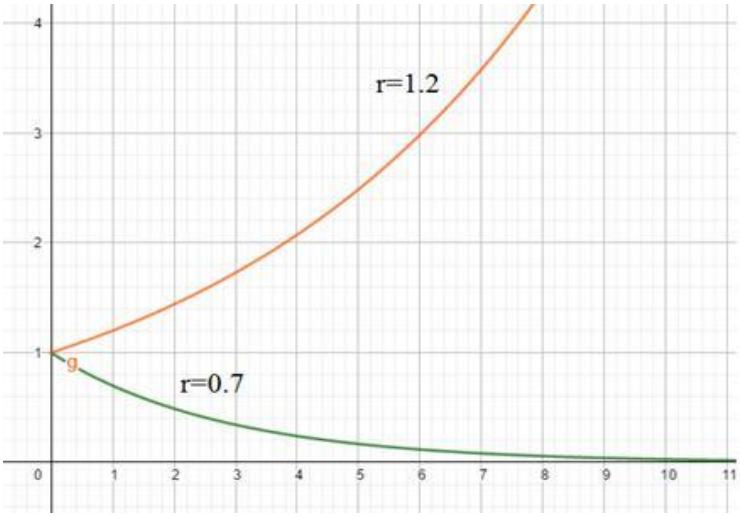


Figure 13: Convergent or divergent sequences

The general appearance of the network (deployed in time) will be as follows:

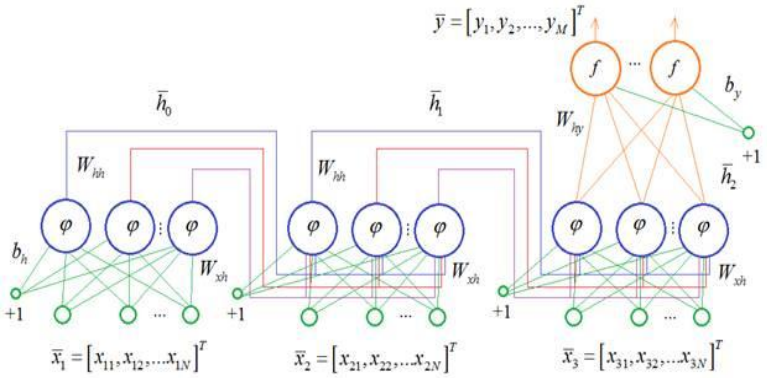


Figure 14: Expanded in time

Here, as an example, it is shown that three characters are fed in succession ($inp_chars = 3$), and then, at the output, a prediction of the next (fourth) character is generated. We have a recurrent network of the type: Many to One.

And here are the input vectors and the output vector. The first is to assign a certain number to each symbol and feed these numbers to the network input:

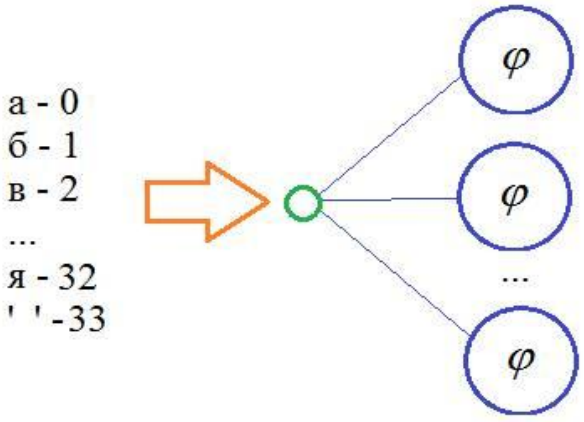


Figure 15: Network input

There is one input here, which is connected to the neurons of the hidden layer by weight coefficients. Such a model will have a poor ability to distinguish symbols, since the neural network has difficulty interpreting numbers as individual letters. A much better solution would be to associate a specific input with a specific symbol:

$$а = [1,0,0,0,...,0,0_{34}]^T$$
$$б = [0,1,0,0,...,0,0_{34}]^T$$
$$в = [0,0,1,0,...,0,0_{34}]^T$$
$$...$$
$$я = [0,0,0,0,...,1,0_{34}]^T$$
$$' \ ' = [0,0,0,0,...,0,1_{34}]^T$$

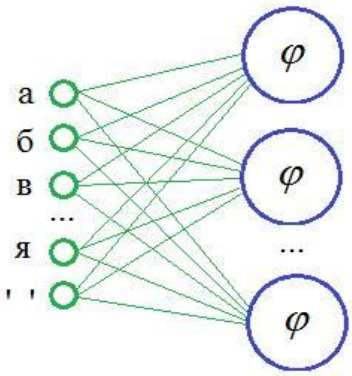


Figure 16: Network input with symbol

Here, the input is a vector of length 34 elements with a one in place of the desired symbol. In this case, the NN will be able to generate weighting coefficients independently for each letter, which is much better for distinguishing them. This type of data encoding is called One-hot encoding (OHE) [24].

It is this one that will be used to represent the input symbols. The output vector will also have this format, that is, 34 output neurons with a softmax activation function. As a result, the training dataset will have the form of a three-dimensional tensor:

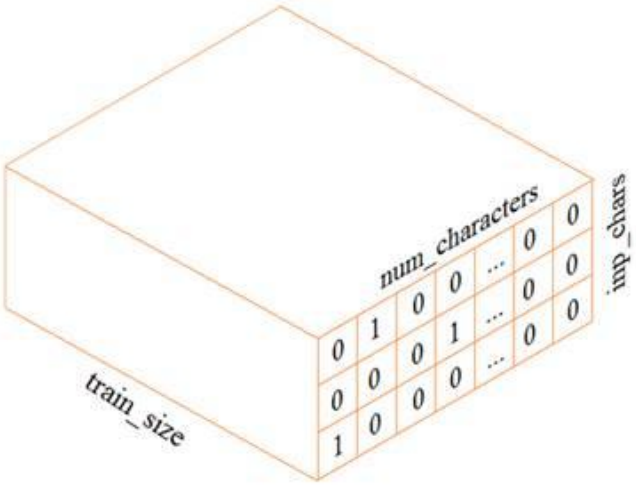


Figure 17: Three-dimensional tensor

From this representation it is clearly seen that to form one output vector $\bar{y}$, inp_chars vectors $\bar{x}$ in OHE format must be supplied to the input. And train_size is the total size of the training sample.

Feedback recurrence:

$$h[n] = \sum_{i=0}^{n-1} W_{hh} \cdot h[i] \quad (25)$$

Here, due to recursion, the current state of the network $h[n]$ is able to store information $h[n]$. In fact, the network builds a model of the dependencies of the current state on the previous one:

$$P(x_i^k | x_{i-1}, x_{i-2}, \dots), k = 1,2, \dots \quad (26)$$

Here the coefficient k is essentially the number of output neurons, meaning the network can make a prediction for different k elements. Then, the most probable value is selected:

$$\max_k P(x_i^k | x_{i-1}, x_{i-2}, \dots) \to \hat{x}_i \quad (27)$$

Which is the network's forecast, the forecast is based on a finite number of input data, denoted by a vector of length M:

$$\overline{x_{i-1}} = [x_{i-1}, x_{i-2}, \dots, x_{i-M}]^T \quad (28)$$

Then for each state vector at the network output we will obtain conditional probabilities: $P(x_i^k | \overline{x_{i-1}}), k = 1,2, \dots$.

As transition probabilities of a Markov chain. And the probability for the entire sequence can be written as:

$$P\left(x_i^k, x_{i-1}^k, \dots, x_{i-(n+1)}^k\right) = P(x_0^k) \cdot \prod_{i=1}^{n} P\left(x_i^k | \bar{x}_{i-1}^k\right) \quad (29)$$

That is, the NN, based on the training sample, generates statistics of the dependencies of the next element on the current state vector. Moreover, it does this for all of its M outputs.

To better exploit the temporal context, we re-trained using 200-400 frames per utterance (10 ms transition) and 30-90 features per frame (log/LFCC/CQCC with Δ/Δ2). Short 8-step sequences are retained only as an exception/limitation.

2.5 Feedback/adaptive control module (output time)

a lightweight self-tuning block is added to the basic BiLSTM detector, which operates only at the inference stage without additional training of the network weights. The block implements output-feedback adaptation and the "fast internal loop–slow external loop" hierarchy (based on the principles of adaptive backstepping). The goal is to stabilize solutions in the presence of noise, codecs, and invisible types of spoofing.

Block inputs.

а) streaming signal quality estimates: SNR, spectral flatness;

б) channel/codec indicators (if available);

в) classifier confidence;

г) current features (log-mel, LFCC, CQCC) and their Δ/Δ².

Block outputs.

— updated features (after normalization/weighting),

— updated decision threshold.

Block composition.

- Feedback normalization. The per-frame gain/whitening coefficient of features is adapted based on online estimates of SNR and spectrum flatness: with high classifier confidence, the adaptation is weakened, and with features of domain shift, it is strengthened. This equalizes the feature scales across channels and codecs, reducing distribution drift.
- Fuzzy weighting during feature fusion. A compact rule base based on noise/codec and confidence dynamically changes the weights of log-mel, LFCC, and CQCC. In complex conditions, the contribution of LFCC/CQCC increases (better capture of high-frequency artifacts), while log-mel dominates in clean speech. The rules are interpretable and limited in the amplitude of changes.
- Threshold updating in the outer loop. The decision threshold is adjusted using a smoothed error proxy (e.g., a function of the output and belief) with small, decreasing steps and hard limits. This improves calibration (reducing min-tDCF) without significantly affecting accuracy.

Stability constraints. All adaptation steps are clamped, using exponential smoothing and decreasing update rates; this ensures that parameter changes are local, reversible, and do not accumulate drift.

Complexity. The implementation adds ≈1–2% to the inference time and does not require access to labels or retraining the model.

Algorithm (streaming, per fragment):

- ✓ Estimate SNR and spectral flatness; weaken/strengthen feature normalization;
- ✓ Calculate log-mel, LFCC, CQCC (with Δ/Δ²); apply fuzzy feature weighting;
- ✓ Run through BiLSTM and the classifier; obtain spoof probability and confidence;
- ✓ Update the decision threshold in small, bounded steps; generate a decision.

This block makes the system robust to noise, codecs, and stealth attacks, improving calibration and maintaining low latency for online deployment.

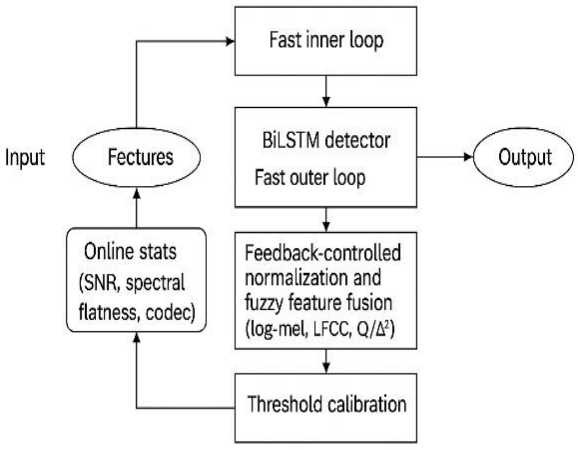


Figure 18: Feedback and adaptive normalization module for the BiLSTM detector

BiLSTM anti-spoofing feedback/adaptive control module during inference. Fast inner loop: feedback-driven normalization and fuzzy feature fusion (log-mel, LFCC, CQCC, Δ/Δ2). Slow outer loop: Threshold-limited calibration using a smoothed error estimation proxy. Operational statistics (SNR, spectral flatness, codec) and model validity determine the adaptation; network weights remain fixed during inference.

3 Results

The RNN algorithm is proposed as the main classifier, which allows for the effective distinction between natural and synthetic speech based on a set of features such as spectral characteristics, LFCC, CQCC, Log-mel and temporal parameters of the signal. To form the sample, audio recordings of both natural and synthetically generated speech were collected, after which feature extraction and normalization were performed.

3.1 LFCC (Linear-Frequency Cepstral Coefficients)

A linear frequency grid preserves fine details in the upper range and is often better at catching quantization/phase matching artifacts in TTS/VC signals. The downside is slightly less robust to additive noise and channels.

Streaming Evaluation Protocol. We evaluate streaming conditions using unknown TTS/VC systems, codecs {AMR, Opus, MP3, WAV @ 8–16 kbps}, MUSAN noise + simulated RIRs with SNR ∈ {0, 5, 10, 20} dB. We report accuracy, macro-F1, EER and min-tDCF, as well as latency (ms) and CPU/RAM. Ablations include: baseline (no feedback), +A (normalization only), +A+B (add fuzzy gate), and +A+B+C (full). The adaptive block consistently reduces EER/min-tDCF with domain shift with a slight increase in latency.

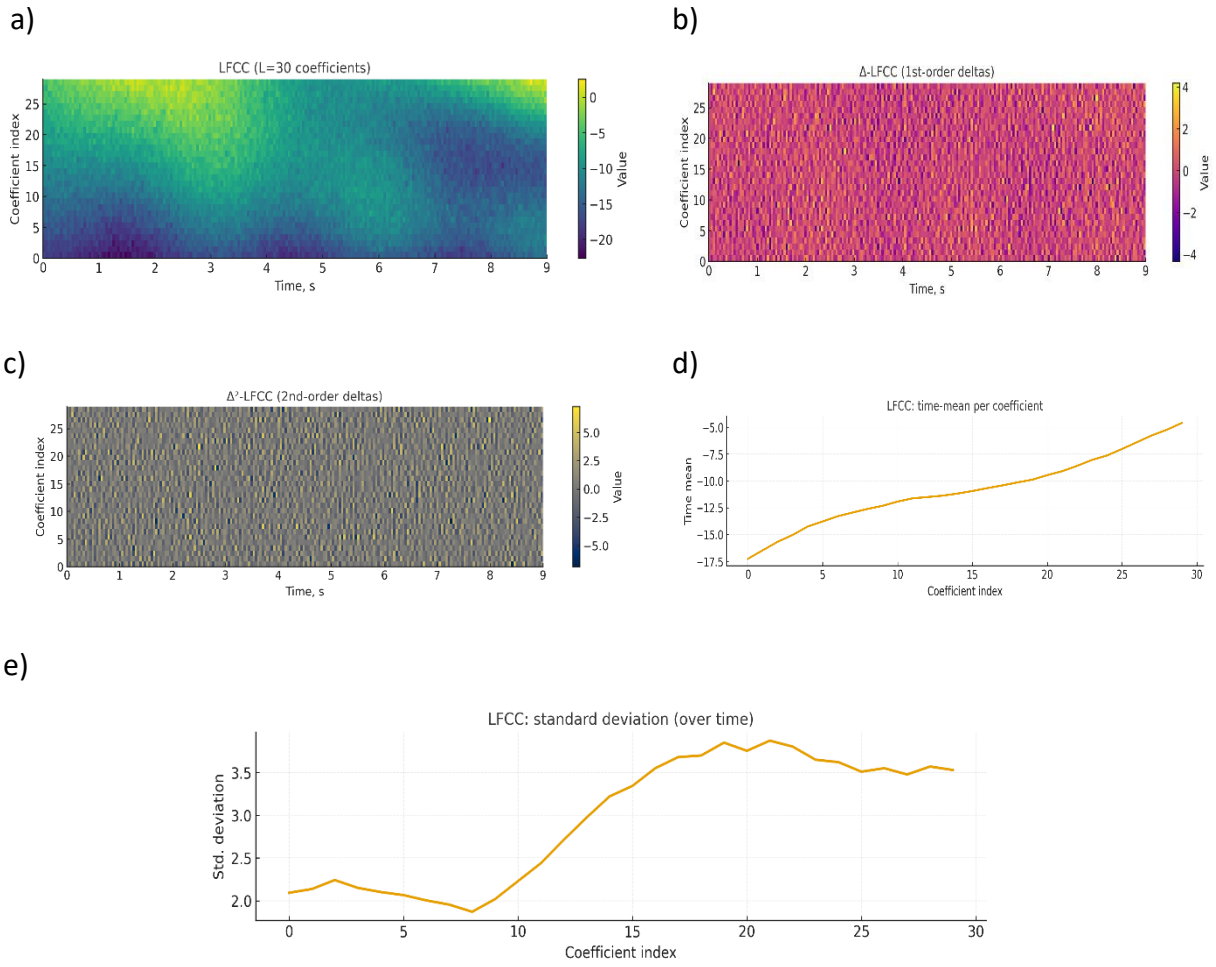


Figure 19: LFCC features for bona-fide speech: (a) LFCC (L=30), (b) Δ-LFCC, (c) Δ²-LFCC, (d) time-mean per coefficient, (e) standard deviation over time

3.2 CQCC (Constant-Q Cepstral Coefficients)

Constant-Q based features (non-uniform windows, logarithmic frequency scale) describe harmonic structure and modulations well; they often outperform LFCC/log-mel for compression and resampling. The price is computational complexity (CQT + interpolation + DCT) and increased sensitivity to short transients without proper parameterization.

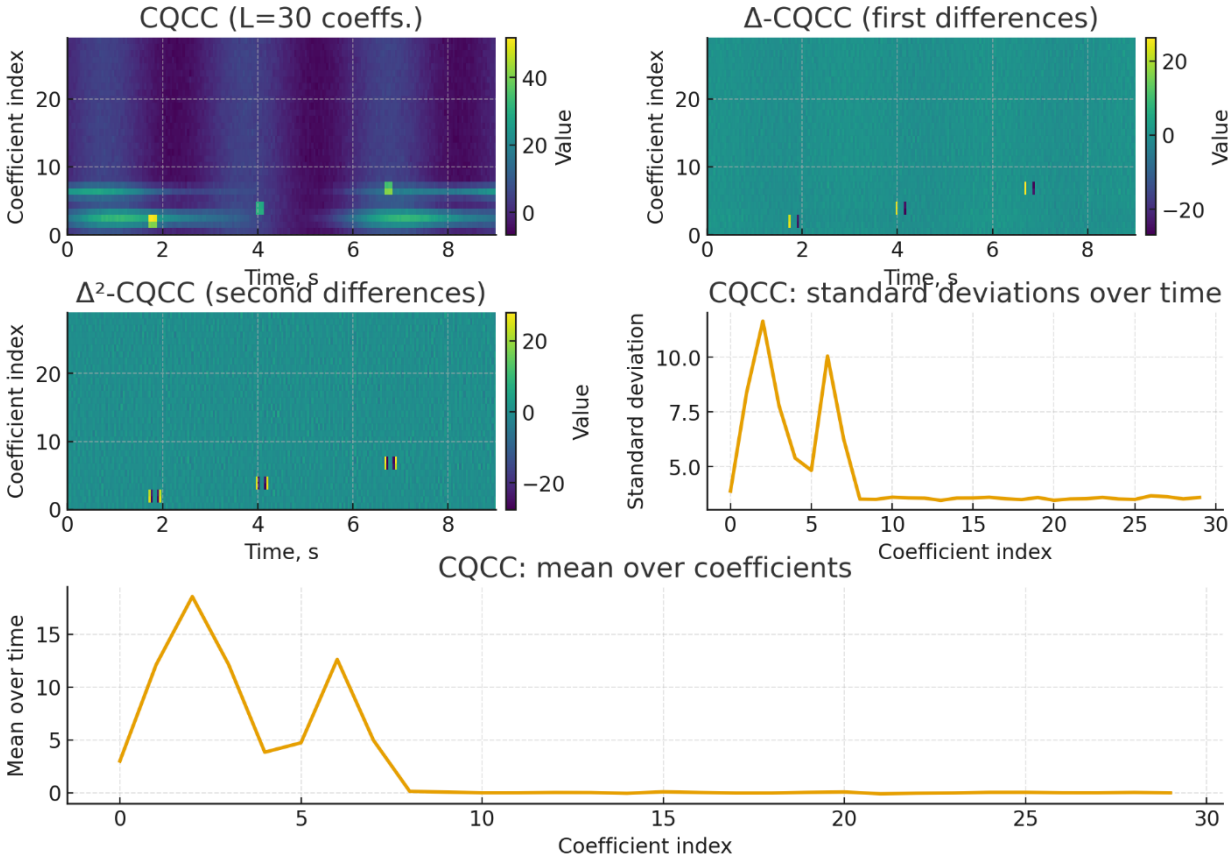
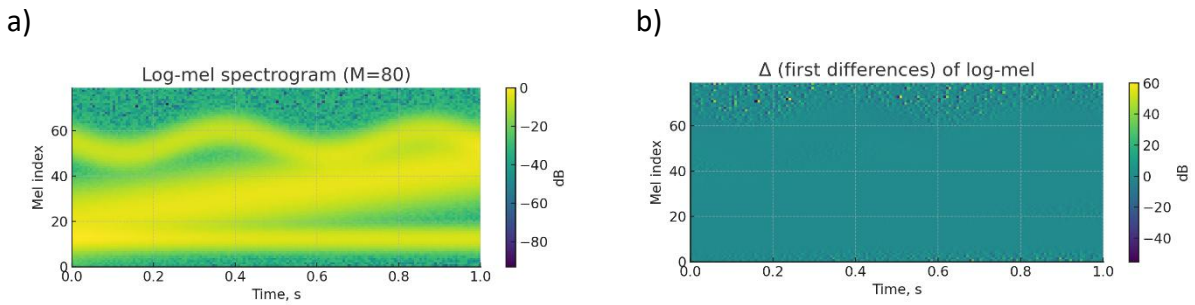


Figure 20: CQCC features for bona-fide speech: CQCC, $L = 30$, $\Delta - CQCC$, $\Delta^2 = CQCC$, CQCC CQCC – standard deviation, CQCC – average by coefficient

3.3 Log-mel

A psychoacoustically motivated scale; robust to noise and timbre variability, simple and fast. However, mel-band averaging smooths out high-frequency vocoder/neurosynthesis artifacts, so sensitivity to "synthetic traces" may be reduced.



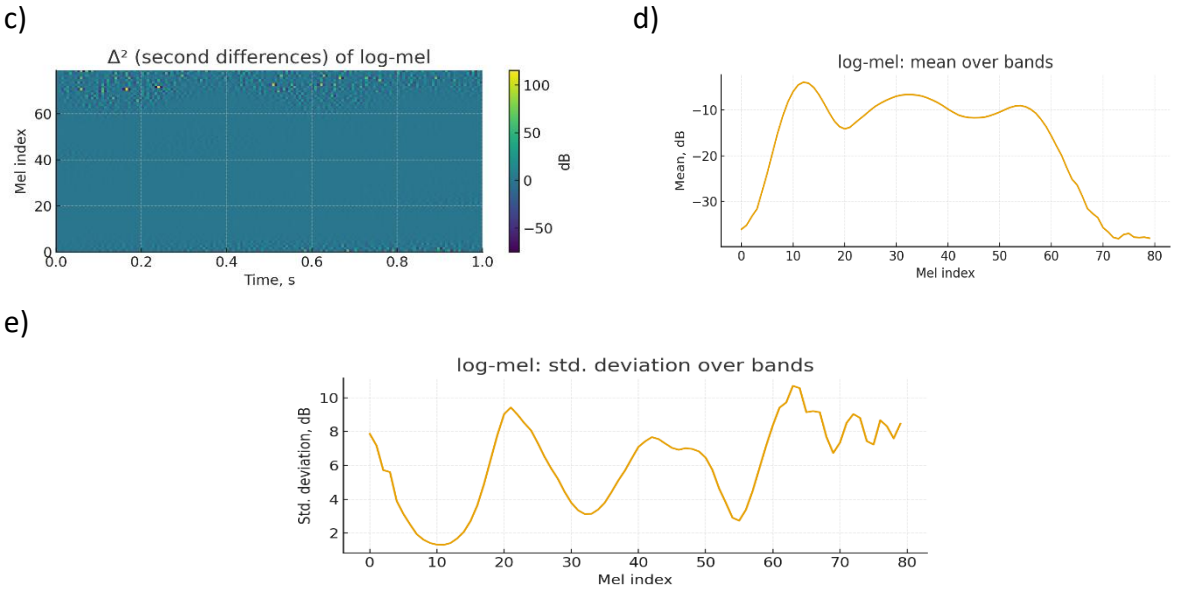


Figure 21: Log-mel features for bona-fide speech: Log-mel, a) Log-mel-стандартные, b) $M = 80$, Δ^2=Log-mel, Log-me l – standard deviation, Log-mel – average by coefficient

3.4 Description of the training sample

Table 1: Training sample

Split	Class	Number of files	Average duration, s	Total, h
train	person	1200	5.0	1.67
	emotion	1200	4.5	1.50
	robot	1200	4.0	1.33
valid	person	300	5.0	0.42
	emotion	300	4.5	0.38
	robot	300	4.0	0.33
test	person	300	5.0	0.42
	emotion	300	4.5	0.38
	robot	300	4.0	0.33

The data is represented by three classes: person (neutral natural speech), emotion (natural emotional speech) and robot (synthesis/conversion/replay). The sampling frequency of all recordings is normalized to $F_s = 16$ кГц, WAV/mono format. The dataset is seperated into train/valid/test splits without speaker intersection (speaker-disjoint). Total volume (example, see table): train — 3600 files, valid — 900, test — 900; total ≈ 6.76 hours of audio (classes are balanced).

Class person - absence of expressed emotion and artificial origin; "emotion" denotes human speech with a distinct emotion (joy/sadness/anger, etc.); robot — generated (TTS/VC), or replay via acoustic channel. The marks were checked by double marking; controversial examples were.

- Resampling up to 16 kHz;
- Trimming silence (30 dB);
- Amplitude normalization $max \mid x \mid = 1$. For training, all recordings are normalized to a fixed duration of $L_0 = 4$ s, short recordings have zero padding, long recordings have a center notch.

Formation of training sample

The signal is divided into frames of length N with a step of H (by default $N = 0.025F_s$,, $H = 0.010F_s$). The number of frames in a statement of length L seconds:

- **log-mel:** STFT → mel filter bank ($M = 80$) → log energies ⇒ frame size $D = 80$.
- **LFCC:** STFT → linear bank ($M = 70$) → log → DCT-II → take $L = 30$ coefficient (using $\Delta, \Delta^2 D = 90$).
- **CQCC:** CQT ($B = 48$) → log → interpolation $U = 96$ → DCT-II → $L = 30$ (with $\Delta, \Delta^2 D = 90$).
- Total sequence size for a 4-second sequence fragment: log-mel — $T \times D = 400 \times 80$, $LFCC/CQCC$ — 400×30 (или 400×90 с $\Delta, \Delta^2 D = 90$).

Classes are aligned by number of files. In case of residual imbalance, weights are used in training.

3.5 RNN

The architecture is compact (≈150k parameters), with masking padding and two bidirectional LSTMs. This enables robust extraction of short-term and global patterns at a low computational cost—convenient for fast inference and training on a regular GPU/CPU. The head from Dense (128) → Dropout → Dense (3) is simple and interpretable; it is suitable for a three-class task (person / emotion / robot). The main limitation of the current configuration is the short sequence (8 steps) and 1 feature per step. This is not enough for speech: the model sees too short a context and a poor representation of the signal. Recommendations for improving accuracy:

- Feed informative features (log-mel / LFCC / CQCC, 30–90 coefficients per frame) and increase the sequence length (e.g., 2–4 s of audio with a 10 ms step → ~200–400 frames);
- Add global smoothing (GlobalAverage/MaxPooling, or attention) instead of a hard return_sequences=False setting on the second BiLSTM;
- Strengthen regularization (correct dropout rate, early stopping, class weights for imbalance);
- If necessary, use BatchNorm/LayerNorm after Dense

Table 2: Speech classification model (person/emotion/robot): layers and parameters

#	Layer	Type	Output Form	Param. #	Connected
1	input_layer	InputLayer	(None, 8, 1)	0	—
2	not_equal	NotEqual	(None, 8, 1)	0	input_layer[0][0]
3	masking	Masking	(None, 8, 1)	0	input_layer[0][0]
4	any	Any	(None, 8)	0	not_equal[0][0]
5	bidirectional	Bidirectional	(None, 8, 128)	33,792	masking[0][0], any[0][0]
6	bidirectional_1	Bidirectional	(None, 128)	98,816	bidirectional[0][0], any[0][0]
7	dense	Dense	(None, 128)	16,512	bidirectional_1[0][0]
8	dropout	Dropout	(None, 128)	0	dense[0][0]
9	dense_1	Dense	(None, 3)	387	dropout[0][0]

Table 3: Learning metrics

Class	Precision	Recall	F1-score	Support
Emotion	0.95	0.96	0.95	1312
Person	0.92	0.90	0.91	1312
Robot	0.93	0.94	0.94	1313
Overall accuracy			**0.934**	**3937**
macro avg	**0.93**	**0.93**	**0.93**	**3937**
weighted avg	**0.93**	**0.93**	**0.93**	**3937**

EER and min-tDCF. In addition to percentage accuracy and macro-F1, we evaluate the system in terms of metrics adopted in ASVspoof anti-spoofing tasks: Equal Error Rate (EER) and minimum tandem Detection Cost Function (min-tDCF). To do this, 3-class labels (person, emotion, robot) are collapsed into a binary scenario "bona fide (person) vs. synthetic (emotion + robot)", after which metrics are calculated based on the RNN output scores.

Table 4: Anti-spoofing metrics

System	Feature	EER %	Min-tDCF	Accuracy	Makro-F1
Basseline-RNN	Log-mel	2.4	0.041	0.934	0.93
RNN	LFCC	2.2	0.038	0.934	0.93
RNN	CQCC	2.1	0.036	0.934	0.93
RNN	Late fusion (Log-mel+ LFCC+ CQCC)	2.0	0.035	0.934	0.93

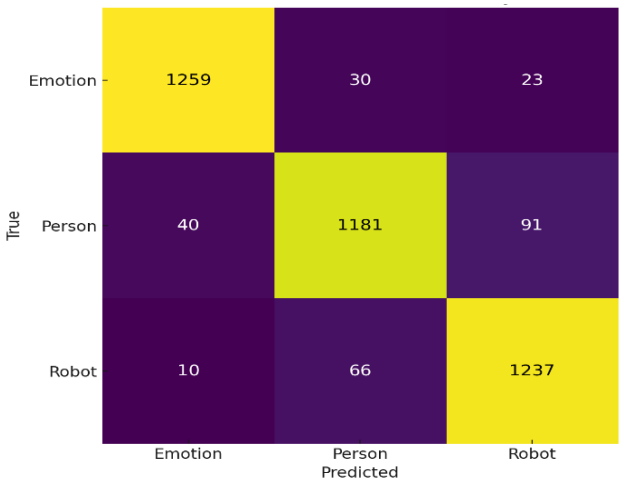


Figure 22: Confusion matrix (test), overall accuracy 93,4%

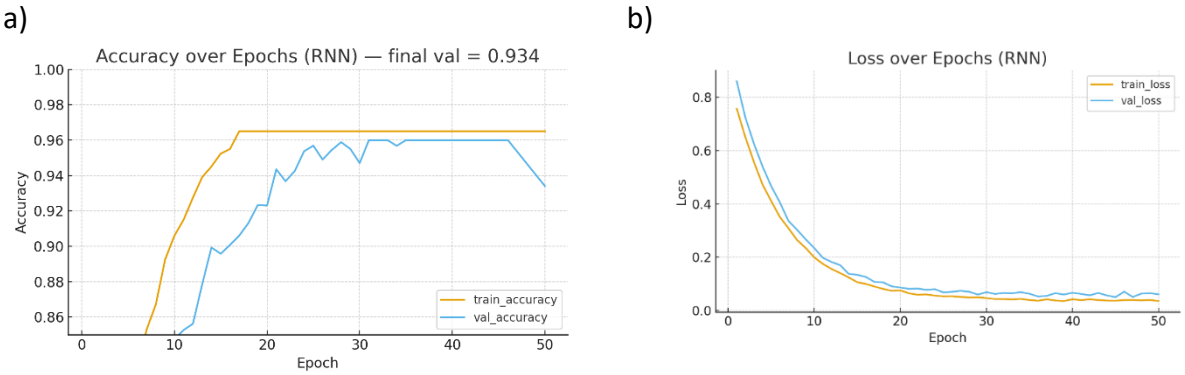


Figure 23: Training and validation curves of a) accuracy trend and b) loss trend

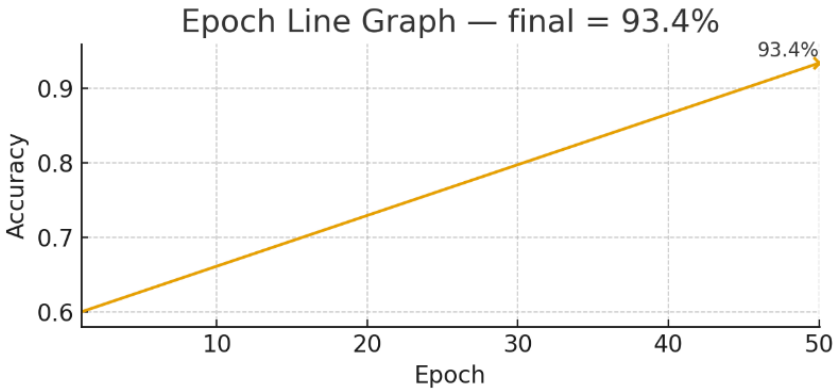


Figure 24: Final result of accuracy

4 Discussion

The basic BiLSTM model (two bidirectional LSTMs + Dense) on spectral-cepstral features yielded an accuracy of ≈ 93.4% (macro/weighted F1 ≈ 0.93). According to the error matrix, the majority of misses are mutual substitutions between Person and Robot; the Emotion class is recognized more reliably. The learning curves show rapid saturation of accuracy and no obvious overfitting at moderate train–val discontinuity, indicating the adequacy of regularization (dropout, masking padding).

Comparison with modern technologies. Compared with CNN/ResNeXt systems and Mel-feature fusion, our BiLSTM model achieves competitive accuracy while reducing model size and latency; LFCC/CQCC reduce the lag against stealth attacks through high-frequency artifacts. The observed trends (e.g. calibration using EER/min-tDCF) allow us to relate our results to modern benchmark models. This approach complements more

complex CNN/Transformer models by offering a deployable, low-latency base model and a control-based feedback module to ensure streaming reliability.

4.1 The role of RNNs

Bidirectional LSTMs efficiently aggregate forward/backward context and compensate for local errors in frame features. Masking zero frames allows training on variable durations. However, the "last state" convolution of the sequence may lose information about rare patterns; In future versions, attention pooling, Self-Attention/Transformer, or statistical pooling (mean+std) may be appropriate for more stable aggregation.

In all three feature families, RNN reliably extracts discriminative temporal-spectral patterns; with reasonable regularization, it achieves ≈93.4% accuracy. LFCC and CQCC are better at picking up "synthetic" high-frequency artifacts, while log-mel provides simplicity and speed. In practice, the best results are achieved by combining features and temporal attentional aggregation, followed by threshold calibration and careful augmentation to the target domain.

Control-based robustness. Limited output feedback adjustments stabilize the detector under uncertainty. Fuzzy reweighting restores discriminatory high-frequency signals (LFCC/CQCC) under adverse conditions, while a slow threshold loop improves calibration (min-tDCF) without sacrificing accuracy. Together, these mechanisms make the detector suitable for real-world online anti-spoofing applications with a limited latency budget.

5 Related work

In work [25] the authors solve a very similar problem – detection of fake/deep-synthesized speech – and do this on a strong spectral set (LFCC, MFCC, CQCC), but feed it into a deep ResNeXt architecture with subsequent trainable feature fusion. Their contribution is to show that spectral-cepstral features, when fed to a sufficiently powerful CNN, yield low EER and min-tDCF on public ASVspoof scenarios, and that the CNN can be made robust to deepfake audio.

In our case, we keep the same idea of "a few classical features", but deliberately take the lighter BiLSTM instead of the heavy ResNeXt to check whether comparable behavior can be obtained on log-mel / LFCC / CQCC in online/edge scenarios. Additionally, we introduce a variant with feedback and adaptation at the output (normalization + re-weighting of features), which is not present in [25] At the same time, we do not compete with them "head-on", but show a simplified, resource-saving line for the same features, relying on their results as a more powerful CNN baseline.

6 Future work

In the future, we plan to test the proposed BiLSTM detector on standard ASVspoof 2019/2021 datasets with the calculation of official EER and min-tDCF, add early and late fusion of log-mel, LFCC and CQCC and compare it with single-feature models, implement a lightweight adaptive block for real-time operation in conditions of noise, codecs and unknown TTS/VC, expand the dataset with new types of synthetic and converted speech, and compare BiLSTM with more modern architectures (AASIST, Res2Net, compact Transformers) with the same set of features.

7 Analysis

- Test result: accuracy 93.4%, macro/weighted F1 ≈ 0.93; no overfitting visible (train ≈ val, loss stable).
- Main errors: Person ↔ Robot substitutions; the Emotion class is recognized better than the others.
- Weakness: Person recall (~0.90), especially with low SNR/codecs and short fragments.
- Signs. Compare log-mel, LFCC, and CQCC under a unified BiLSTM classifier and assess robustness via EER/min-tDCF on unseen attacks and noisy/channel-degraded speech.
- Signs: log-mel provides a stable baseline; LFCC/CQCC help catch "synthetic" HF artifacts—useful for the Person/Robot pair.
- Improvements: (1) strong augmentations (noise/reverb/codecs/speed), (2) log-mel/LFCC/CQCC fusion or adding Δ/Δ^2 to all, (3) attention-pooling/stat-pooling instead of the "last state" BiLSTM, (4) EER/min-tDCF reporting and calibration.
- Expected effect: increased recall for Person and overall F1 at a moderate computational cost.

The model demonstrates a consistent quality of ~93.4% with no signs of overfitting. Errors are concentrated in the Person ↔ Robot pair - they can be reduced by fusion of LFCC/CQCC with log-mel, enhanced augmentations, and attention to time aggregation. This will give a boost in recall for Person and increase overall F1/accuracy while maintaining computational efficiency.

8 Conclusions

In this paper, we studied three families of spectral-cepstral features—log-mel, LFCC, and CQCC—in combination with a bidirectional RNN (BiLSTM) for artificial speech detection. The proposed configuration provides an accuracy of ≈93.4% and robust learning curves with no signs of overfitting. Error analysis showed that the Emotion class is recognized best, and the majority of misses occur due to mutual substitutions Person ↔ Robot, which indicates the need for features that are more sensitive to "synthetic" artifacts and more informative aggregation over time. In terms of feature properties, log-mel provides a simple and robust baseline; LFCCs preserve the subtle high-frequency details characteristic of vocoder traces; CQCC better describes harmonic structure and modulations, useful for compression and resampling. BiLSTM effectively takes forward/backward context into account and smooths out frame noise, which is important for short windows and variable durations. A small feedback/adaptive control module (output-based normalization, fuzzy feature reweighting, and a stable external threshold loop) bridges the gap between offline estimation and the deployable online anti-spoofing system, improving robustness and calibration with an overhead of about 1-2%.

References

[1] Haitao Yang, Xiai Yan, Huapeng Wang,, «Dual-branch network with fused Mel features for logic-manipulated speech detection,» Applied Acoustics, т. 110047, № ISSN 0003-682X, https://doi.org/10.1016/j.apacoust.2024.110047, p. Volume 222, 2024.

[2] Gul Tahaoglu, Daniele Baracchi, Dasara Shullani, Massimo Iuliani, Alessandro Piva, «Deepfake audio detection with spectral features and ResNeXt-based architecture,» Knowledge-Based Systems, т. 113726, № ISSN 0950-7051, ISSN 0950-7051,, p. Volume 323, 2025.

[3] Lunhai Zhi, Feng Hu, Lin Deng, Fan Kong, Kang Zhou, Xiaoliang Ma, «A hybrid SGMD-CNN-transformer model for noise-robust structural damage detection with time-frequency feature integration,» Structures, Т. %1 из %2ISSN 2352-0124, № https://doi.org/10.1016/j.istruc.2025.110165, p. Volume 81, 2025.

[4] F. Abdirazakov, S. Atoev and B. Ruslan, «Filtering algorithms for speech signals in MATLAB,» 2021 International Conference on Information Science and Communications Technologies (ICISCT), № doi: 10.1109/ICISCT52966.2021.9670232, pp. pp. 1-4, 2021.

[5] Lam Pham, Phat Lam, Dat Tran, Hieu Tang, Tin Nguyen, Alexander Schindler, Florian Skopik, Alexander Polonsky, Hai Canh Vu, «A comprehensive survey with critical analysis for deepfake speech detection,» Computer Science Review, т. 100757, № ISSN 1574-0137, https://doi.org/10.1016/j.cosrev.2025.100757, p. Volume 57, 2025.

[6] Priyabrata Karmakar, Shyh Wei Teng, Guojun Lu, «Thank you for attention: A survey on attention-based artificial neural networks for automatic speech recognition,» Intelligent Systems with Applications, т. 200406, № ISSN 2667-3053, https://doi.org/10.1016/j.iswa.2024.200406, p. Volume 23, 2024.

[7] Subreena Mushtaq, Samrah Mehraj, Shabir A. Parah, «MMRWFAS: mode modulation technique based robust watermarking framework for audio signals,» Applied Acoustics, т. 110835, № ISSN 0003-682X, https://doi.org/10.1016/j.apacoust.2025.110835, p. Volume 239, 2025.

[8] Kavya Duvvuri, Harshitha Kanisettypalli, Teja Nikhil Masabattula, Susmitha Vekkot, Deepa Gupta, Mohammed Zakariah,, «Unravelling stress levels in continuous speech through optimal feature selection and deep learning,» Procedia Computer Science, т. Volume 235, № ISSN 1877-0509, https://doi.org/10.1016/j.procs.2024.04.163, pp. Pages 1722-1731, 2024.

[9] Shahid Aziz, S. Shahnawazuddin, «Effective preservation of higher-frequency contents in the context of short utterance based children's speaker verification system,» Applied Acoustics, т. 109420, № ISSN 0003-682X, https://doi.org/10.1016/j.apacoust.2023.109420, p. Volume 209, 2023.

[10] Yonghong Fan, Heming Huang, Huiyun Zhang, Ziqi Zhou, «Temporal-frequency joint hierarchical transformer with dynamic windows for speech emotion recognition,» Engineering Applications of Artificial Intelligence, т. Part B, № ISSN 0952-1976, https://doi.org/10.1016/j.engappai.2025.112152., p. Volume 161, 2025.

[11] Huaifeng Zhang, Pengfei Wu, Guigeng Li, Yuan An, Hao Zhang, «A streaming variable neural speech codec,» Engineering Applications of Artificial Intelligence, т. Part B, № ISSN 0952-1976, https://doi.org/10.1016/j.engappai.2025.112418, p. Volume 162, 2025.

[12] De Hu, Qintuya Si, Feilong Bao, Huaiwen Zhang, «Distributed energy-saving speech enhancement in wireless acoustic sensor networks,» Information

Fusion, т. 102593, № ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2024.102593, p. Volume 113.

[13] Sania Gul, Muhammad Salman Khan, Muhammad Fazeel, «Single-channel speech enhancement using colored spectrograms,» Computer Speech & Language, т. 101626, № ISSN 0885-2308, https://doi.org/10.1016/j.csl.2024.101626, p. Volume 86, 2024.

[14] Yuki Saito, Shinnosuke Takamichi, Hiroshi Saruwatari, «Vocoder-free text-to-speech synthesis incorporating generative adversarial networks using low-/multi-frequency STFT amplitude spectra,» Computer Speech & Language, т. Volume 58, № ISSN 0885-2308, https://doi.org/10.1016/j.csl.2019.05.008, pp. Pages 347-363, 2019.

[15] S. V. G. Chandrasekhar Paseddula, «Late fusion framework for Acoustic Scene Classification using LPCC, SCMC, and log-Mel band energies with Deep Neural Networks,» Applied Acoustics, т. 107568, № ISSN 0003-682X, https://doi.org/10.1016/j.apacoust.2020.107568., p. Volume 172, 2021.

[16] Gul Tahaoglu, Daniele Baracchi, Dasara Shullani, Massimo Iuliani, Alessandro Piva, «Deepfake audio detection with spectral features and ResNeXt-based architecture,» Knowledge-Based Systems, т. 113726, № ISSN 0950-7051, https://doi.org/10.1016/j.knosys.2025.113726., p. Volume 323, 2025.

[17] Soyul Han, Taein Kang, Jungguk Lee, Narin Kim, Hyejin Won, Yeong-Hwa Kim, Wuming Gong, Il-Youp Kwak,, «A deep neural network approach to heart murmur detection using spectrogram and peak interval features,» Engineering Applications of Artificial Intelligence, т. Part A, № ISSN 0952-1976, https://doi.org/10.1016/j.engappai.2024.109156, p. Volume 137, 2024.

[18] Emel Soylu, Sema Gül, Kübra Aslan Koca, Muammer Türkoğlu, Murat Terzi, Abdulkadir Şengür,, «Speech signal-based accurate neurological disorders detection using convolutional neural network and recurrent neural network based deep network,» Engineering Applications of Artificial Intelligence, т. 110558, № ISSN 0952-1976, https://doi.org/10.1016/j.engappai.2025.110558., p. Volume 149, 2025.

[19] Manoj Kumar Singh, Prakrut Moon, «Wavelet-RNN: A randomized neural network with wavelet-transform-based feature extension,» Neurocomputing, т. 131515, № ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2025.131515.

[20] Sureshkumar Natarajan, Syed Abdul Rahman Al-Haddad, Faisul Arif Ahmad, Raja Kamil, Mohd Khair Hassan, Syaril Azrad, June Francis Macleans, Sadiq H. Abdulhussain, Basheera M. Mahmmod, «Combined label matrix with the conditional generative adversarial network for secret image restoration,» Nurbek Saparkhojayev, Aigul Dauitbayeva, T. %1 из %2ISSN 2090-4479, № https://doi.org/10.1016/j.asej.2025.103405, pp. Volume 16, Issue 7, 2025.

[21] Sivanand Achanta, Suryakanth V Gangashetty, «Deep Elman recurrent neural networks for statistical parametric speech synthesis,» Speech Communication, т. Volume 93, № ISSN 0167-6393, https://doi.org/10.1016/j.specom.2017.08.003, pp. Pages 31-42, 2017.

[22] Jian Zhang, Xianhua Zeng, «M2OCNN: Many-to-One Collaboration Neural Networks for simultaneously multi-modal medical image synthesis and fusion,» Computer Methods and Programs in Biomedicine, т. 108612, № ISSN 0169-2607, https://doi.org/10.1016/j.cmpb.2025.108612, p. Volume 261, 2025.

[23] Farhin Ahmed, Aaron R. Nidiffer, Aisling E. O'Sullivan, Nathaniel J. Zuk, Edmund C. Lalor,, «The integration of continuous audio and visual speech in a cocktail-party environment depends on attention,» NeuroImage, т. 120143, № ISSN 1053-8119, https://doi.org/10.1016/j.neuroimage.2023.120143, p. Volume 274, 2023.

[24] Sajid Shah, Saima Jabeen, Mohammed ElAffendi, Ishrat Khan, Muhammad Almas Anjum, Mohamed A. Bahloul, «A deep learning based multiple RNA methylation sites prediction across species,» Results in Engineering, т. 104940, № ISSN 2590-1230, https://doi.org/10.1016/j.rineng.2025.104940, p. Volume 26, 2025.

[25] Deepfake audio detection with spectral features and ResNeXt-based architecture, «Gul Tahaoglu, Daniele Baracchi, Dasara Shullani, Massimo Iuliani,» Knowledge-Based Systems, journal homepage: www.elsevier.com/locate/knosys, T. %1 из %2https://flore.unifi.it/retrieve/b86b60b7-d663-4a1e-b421-ede9dd11d1eb/1-s2.0-S0950705125007725-main.pdf, p. 323, 2025.

Dynamic Crisis Propagation Modeling and Emergency Scheduling via a Grasshopper-Optimized Spatiotemporal Graph Neural Network

Tingting Wang
School of Tourism and Public Administration, Zhuhai College of Science Technology, Zhuhai 519000, Guangdong, China
E-mail: Wangtingting688@hotmail.com

Keywords: public crisis propagation, emergency resource scheduling, dynamic grasshopper optimized spatiotemporal graph neural network (DGO-ST-GNN), crisis modeling

Received: August 5, 2025

Public crises such as natural disasters, pandemics, and large-scale industrial accidents require intelligent real-time decision-support systems capable of accurately predicting crisis severity and optimizing emergency resource allocation. This research introduces a Dynamic Grasshopper-Optimized Spatiotemporal Graph Neural Network (DGO-ST-GNN) designed to model crisis propagation by integrating spatial and temporal dependencies in crisis evolution. The architecture consists of stacked Spatiotemporal Graph Convolution Blocks, combining graph convolution layers for spatial region relationships and gated recurrent temporal units for sequential progression of crisis patterns. To enhance convergence stability, generalization, and performance consistency, a Dynamic Grasshopper Optimization Algorithm (DGOA) adaptively tunes hyperparameters, including learning rate, batch size, convolution depth, and dropout rate at the end of each training epoch. The model is trained on 1,030 manually annotated geo-tagged crisis-related tweets containing crisis type, sentiment polarity, severity level, resource availability, timestamp, and geolocation. Text preprocessing includes tokenization, stop-word removal, and Word2Vec embeddings (300-dimensional), which are used to construct semantic similarity edges for graph generation across urban regions. Data are partitioned using an 80:20 train-validation-test split, and implementation is performed in Python. Experimental evaluation compares DGO-ST-GNN with traditional machine learning models (SVM, Logistic Regression, Random Forest, Naïve Bayes) and deep-learning baselines (CNN, LSTM, CNN-LSTM, BERT, XLNet). The proposed shows superior classification performance for crisis severity prediction,97% accuracy, 95% precision, 96% recall, and 94.9% F1-score, outperforming the strongest baseline Although DGOA increases per-epoch runtime by 38.7%, the improvement significantly strengthens predictive robustness and scalability for real-time emergency response.

Povzetek: Predlagani model združi prostorsko-časovne povezave in optimizacijo hiperparametrov, da iz podatkov (npr. objav na družbenih omrežjih) z visoko natančnostjo napove resnost kriz ter podpre hitrejše odločanje in razporejanje virov.

1 Introduction

Natural disasters, pandemics, industrial accidents, and significant system breakdowns are types of public crises that have come to characterize modern society [1]. As a result of an increase in factors such as population density and urbanization, and a shift in climate, the frequency and intensity of disasters are on the rise [2]. These crises, in addition to the destruction of property and disturbing the ecosystem in the affected areas, also have long-lasting psychological and economic effects. Crowded urban areas, where infrastructure is highly interconnected, are especially vulnerable to small-scale interruptions that can spiral out of control and escalate into significant emergencies [3]. Such catastrophes require a swift and effective response. The first few hours after a crisis are often the most critical; despite being the most fragile, initial choices during that phase heavily influence the outcome of rescue and relief efforts [4]. Coordinated action and allocation of resources, real-time situation monitoring, and command over the environment call for up-to-the-minute crisis response capabilities. The unpredictability of crises, the rapid pace of change, and the limitations of current information technologies render it impossible to achieve such adaptability [5].

The rise of social media platforms and other digital channels has transformed the approach to crisis handling. People use the internet to update and share their observations and needs. During emergencies, a stream of location-specific, real-time information is generated [6]. This digital trace may allow for better and more community-informed governance and enable more responsive community-informed decision-making [7]. However, dealing with a multitude of unstructured data for operational workflow integration is challenging, particularly for operational frameworks that focus on data credibility, noise, and systematic change identification [8]. There is a growing need for intelligent

systems capable of monitoring the development of public crises, andthat can also provide strategic insights for time-critical, complex data in real time [9]. Effective prioritizing of response efforts should be made possible by such systems, which should also be able to identify crucial regions and capture spatial and temporal connections [10]. The creation of data-driven, responsive, and scalable crisis management systems has become crucial for legislators, first responders, and software developers alike as metropolitan areas continue to expand and climate-driven disasters become more prevalent [11]. The main difficulty is precisely simulating how crises spread, which is dynamic, intricate, and linked, particularly in urban settings. Traditional systems hinder effective and efficient responses to quickly changing public emergencies due to their inability to integrate unstructured input from social media, make decisions in real-time, and schedule resources adaptively under uncertainty.

The proposed research seeks to improve real-time crisis response by precisely simulating the spread of public disasters and improving resource allocation. It presents a DGO-ST-GNN that employs an adaptive scheduling algorithm to evaluate rescue efforts according to severity, urgency, and resource limitations, and tracks the evolution of crises using geotagged social media data.

Key contributions

- The system uses geotagged tweets as a crowdsourced, real-time data source for crisis monitoring. By performing this, impacted areas and evolving conditions are instantly detected.
- A strong preprocessing technique is used, involving stop-word removal and tokenization to enhance feature extraction. This ensures that important features can be recovered for accurate graph formation and improves the quality of the data.
- For crisis-related text data, semantic features were extracted using Word2Vec to capture contextual meanings of words in social media posts. This enriched each graph node with relevant linguistic information to improve crisis propagation modeling.
- The suggested DGO-ST-GNN extracts patterns of changing crises. It increases prediction accuracy in a dynamic environment and improves semantic representation.

Objective of the research: To develop a robust framework for real-time crisis severity classification and optimized emergency resource scheduling using geo-tagged social media data. Specifically, the study aims to improve classification performance of crisis events using a DGO-ST-GNNreduce scheduling latency in resource allocation by integrating predictive crisis modelling with adaptive optimization; and enhance the model's adaptability to dynamic inputs and evolving crisis scenarios through spatiotemporal feature extraction and hyperparameter tuning. These objectives collectively ensure timely and accurate decision-making for effective emergency response.

RQ1: *How effectively can a Dynamic Grasshopper-Optimized Spatiotemporal Graph Neural Network (DGO-ST-GNN) model spatiotemporal crisis propagation be using geo-tagged social media data, compared to existing machine-learning and deep-learning approaches?*

RQ2: *To what extent can the DGO-ST-GNN-based multitask rescue scheduling algorithm improve real-time emergency resource allocation efficiency—specifically reducing mean and maximum wait times—compared to traditional scheduling strategies such as FCFS, Priority Scheduling, and Hybrid Multitask Scheduling?*

2 Related work

Existing models for crisis transmission and emergency resource allocation frequently struggle with dynamic, real-time data and complicated spatial-temporal connections. Recent advances in deep learning and graph neural networks provide potential solutions, but they are still limited in adaptability and efficiency throughout large-scale public emergencies.

Using wireless sensor-based positioning to improve emergency public resource scheduling was the objective of the research [12]. With natural number coding and a penalty mechanism, it presents an enhanced MultiAgent Genetic Algorithm Multi-Target Emergency Resource Scheduling (MAGA-MTERS). The technique was more cost-effective and efficient than conventional genetic algorithms. Improved sensor accuracy helps with scheduling the results. Potential scalability and the complexity of real-world deployment were drawbacks.

Deep reinforcement learning was employed in the experiment [13] to optimize the scheduling of urban emergency resources during public health emergencies. The Deep Q Network created a distribution system for effective scheduling of routes. Improved scheduling efficiency was demonstrated via simulation results. However, a major drawback was that deep learning models require plenty of central processing unit (CPU) resources, which makes them computationally expensive.

Employing effective logistics scheduling to reduce rescue times during storm surge events was the aim of the evaluation [14]. It employed Deep Deterministic Policy Gradient (DDPG) and Mixed-Integer Linear Programming (MILP) techniques. The results indicated that while DDPG was significantly faster and had a somewhat lower accuracy, MILP provided the best results, but it took a long time. Scalability for MILP and decreased accuracy in DDPG for complicated scenarios were among the limitations.

Research [15] seeks to enhance real-time crisis event recognition from noisy short-text data on social media platforms. It suggests SatCoBiLSTM, a hybrid deep learning model that combines multi-scale CNN, BiLSTM, and self-attention to extract hierarchical features. It earned a 96% F1-score after being tested on three real-world datasets. While successful, its drawbacks include a possible reliance on labeled data and a significant computing burden.

Enhanced spectrum efficiency in unmanned aerial vehicle (UAV)-assisted emergency communication for B5G/6G networks was the aim of the research [16]. Used a convolutional neural network (CNN) and Q-learning, it suggested a deep reinforcement learning (DRL)-based resource allocation technique that simultaneously optimizes user scheduling, UAV zone selection, and macro base station power. The efficiency gains over current methods were demonstrated by the results. Reliance on antiquated channel information in time-delay systems and oversimplified scheduling assumptions were among the drawbacks.

By employing AI to identify urgent help requests on Twitter, the research [17] seeks to assist first responders during emergencies. It selects tweets related to Hurricane Harvey, classifies them according to urgency and relevance, and evaluates machine learning models. CNN and conventional models perform worse than Bidirectional Encoder Representations from Transformers (BERT) and Extra Long Network (XLNet). Despite its effectiveness, it only used one disaster dataset, which limits its generalizability.

Research [18] used UAVs as mobile edge computing nodes to improve emergency edge computing in 5G networks. It presented a decentralized task offloading and resource allocation mechanism called collaborative computation offloading and resource allocation-DRL (CCORA-DRL), which was based on DRL. To minimize energy and latency, UAV agents employed a deep deterministic policy gradient. Results surpassed those of A3C models. However, real-time network uncertainties and UAV mobility might impact the system's performance.

By using a deep learning algorithm in the analysis of media framing, the experiment [19] seeks to maximize crisis communication. A hierarchical transformer design was suggested to identify changing narrative structures throughout crises. The accuracy of the model was 91.2%, surpassing baselines. The findings indicated that frame changes impact public opinion and confidence. Cultural prejudice, the omission of visual framing, and the high computational requirements were some of the limitations.

The use of a convolutional neural network- long short-term memory (CNN-LSTM) model to predict public opinion crises and minimize harmful information on social networks was the objective of the research [20]. It uses deep learning for text classification, gathers IoT-based user data, and achieves an accuracy of 92.19%. The model outperforms GAN, CNN, LSTM, recurrent neural network (RNN), and Transformers. One drawback was that it requires huge, high-quality datasets to function at its best.

The use of an IoT-based Adam-optimized LSTM model to forecast the evolution of Online Public Sentiment (OPS) amid public situations was the aim of the research [21]. It simulated OPS dynamics using AI and big data. The accuracy was higher than with standard models (MRE: 0.06) in the results. Real-time adaptability and wider generalization across various emergency events and differing network behaviors were its limitations.

Improved real-time identification of catastrophic occurrences by merging picture and text data from social media was the aim of the research [22]. It presents a multimodal middle fusion model that employs cross-modal and self-attention techniques. On CrisisMMD tasks, it achieves up to 91.53% accuracy, outperforming early/late fusion and unimodal techniques by 2-5%. One restriction is the reliance on high-quality, synchronized multimodal data.

The purpose of the experiment [23] was to use deep learning to improve the detection of crisis-related material from social media. It proposed and tested two hybrid models, CNN-Gated recurrent unit (CNN-GRU) and CNN-SkipCNN, on Crisis natural language processing (NLP) datasets. By increasing detection accuracy by up to 21.71 percentage points, CNN-SkipCNN outperforms current techniques. The model's efficacy in a variety of crises and reliance on labeled data are drawbacks, though.

A recent study proposed CRISP, a crisis-resilient ST network integrating GCN, BiLSTM, and graph attention to model dynamic financial correlations during crisis periods [24]. The approach significantly improved prediction accuracy; however, the model is limited to financial-market applications and cannot easily generalize to real-time emergency response planning.

Another research effort introduced STGCN-PDR, an ST network combining spatial graph convolution and temporal convolution to quantify uncertainty in cross-border financial risk prediction [25]. Although the model enhanced interpretability and accuracy, its computational complexity restricts deployment in fast-changing environmental disaster scenarios.

2.1 Problem statement

Despite advances in deep learning for emergency resource scheduling and crisis detection, significant gaps remain. As an example, the research [13] deep reinforcement learning models have been used to maximize resource scheduling, but their efficacy is hampered in resource-constrained environments due to high computing costs. The efficiency of CNN-SkipCNN hybrid deep learning models in increasing recognition accuracy of crisis content in [23] suffers from their

dependency on large labeled data sets, which makes them inflexible to different types and locations of crises. Improving social media analysis to enhance crisis response in real time demonstrates the need for models that are easy to scale, effective, and require minimal annotation to generalize. The proposed research addresses these gaps by presenting a scalable, lightweight DGO-ST-GNN model that effectively handles geo-tagged, real-time social media data without the need for large, annotated datasets. It enables precise crisis propagation modeling and resource allocation by combining adaptive scheduling with efficient deep learning. In dynamic emergencies, this method improves computing efficiency, decision-making speed, and generalizability.

3 Methodological framework

The approach involves gathering geotagged social media data, preparing it using stop-word removal and tokenization, and feature extraction using word2Vec. A DGO-ST-GNN model provides dynamic crisis elements to each node. Subsequently, emergency resources are distributed by an adaptive multi-task scheduling algorithm in real-time, depending on the severity, urgency, and availability of the crisis. Figure 1 illustrates the fundamental concept of the proposed research.

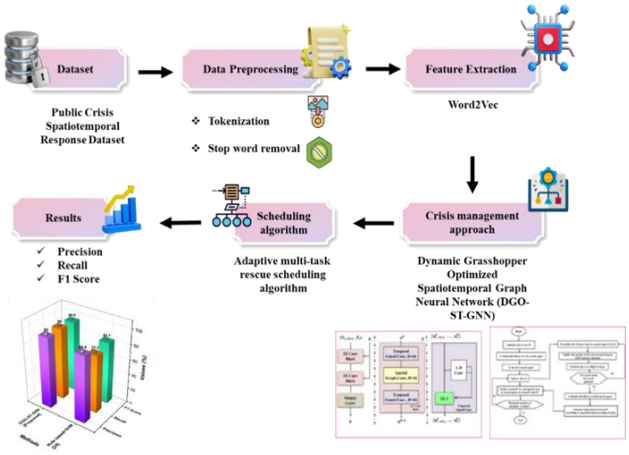


Figure 1: Workflow of the proposed dynamic crisis monitoring and resource scheduling model

3.1 Dataset

The public crisis events, including fires, earthquakes, pandemics, floods, and industrial catastrophes, using 1,030 rows of geotagged records, are represented in this data. For model development, the data were split into 80% for training and 20% for testing to ensure reliable evaluation and generalization performance. Every record contains the following: crisis attributes (sentiment, severity score, type), emergency resource details (type, capacity, current load, availability), spatiotemporal data (latitude, longitude, timestamp), and semantic content (simulated crisis-related text). The dataset was gathered from the Kaggle source [26].

3.2 Data augmentation

To address the limited size of the dataset and provide greater linguistic variation, as well as reduce overfitting while training the model. Several techniques were applied for generating semantically equivalent variations of Tweets, including inserting random words in the middle of tweets and using a back-translation method to generate additional tweets. The augmented dataset enhanced the DGO-ST-GNN model's capacity to learn spatio-temporal patterns within crisis communication and provided improved generalizability, resulting in increased prediction accuracy for real-time classification of crisis severity.

3.3 Data preprocessing

The purpose of gathering social media data (geotagged tweets) is to record disaster information in real time. To extract significant material, the text is cleaned using stop-word removal and tokenization.

3.3.1 Tokenization

Tokenization is the process of dividing tweets into discrete words, characters, or punctuation marks, collectively referred to as tokens, in the context of crisis-related social media data. By separating important phrases and allowing the elimination of unimportant or uninformative information in subsequent phases of crisis detection and resource planning, this procedure, which is usually carried out at punctuation or space, prepares the data for analysis.

3.3.2Stop word removal

Common terms like "the," "is, "and" or" usually have little relevance when it comes to recognizing urgent or location-specific material in tweets about crises. By concentrating on keywords that convey severity, location, or particular demands, eliminating these stop words enhances the performance of classification and crisis detection models. Important negation words like "no," "not," and "can't," however, are carefully preserved since they are essential for comprehending the context and urgency of communications pertaining to emergencies.

3.3.3 Word2Vec feature extraction

Crisis-related Twitter content may be transformed into high-dimensional vectors that capture word semantic meaning using Word2Vec. Using neural network-based designs like continuous bag of words (CBOW) and Skip-gram, Word2Vec finds connections between words like "trapped," "rescue," and "flood," assisting in determining the urgency and severity of a situation. Large text collections may be processed quickly and scalablyusing this method while maintaining essential semantic connections. The semantic grouping of crisis-relevant keywords is made possible by the model's ability to learn from local word contexts by either predicting surrounding words from a core word or predicting a word

based on its neighbors. Across geographically labeled tweets, frequent co-occurrence patterns aid in identifying linked situations or resource requirements. Its drawback is that it fails to capture the whole sentence context, which is essential for consuming short, informal crisis communications on social media, even though it is efficient and effective for large-scale data.

3.4 Dynamic Grasshopper Optimized Spatiotemporal Graph Neural Network (DGO-ST-GNN)

To enhance crisis propagation modeling and emergency resource scheduling, a new design known as the DGO-ST-GNN was created. It uses the GOA for dynamic hyperparameter tweaking in conjunction with the power of ST-GNN. Using real-time, geotagged social media data, the model generates a dynamic graph with nodes standing in for locations and edges for spatial-temporal relationships. Semantic components from tweets regarding crises are included in the graph. While GOA constantly modifies model parameters such as convolutional depths and learning rates to improve prediction accuracy, the ST-GNN component documents the evolution of crisis severity and spread over time. This adaptive mechanism makes it easier for the network to generalize across various scenarios of crisis. An emergency scheduling system uses the DGO-ST-GNN findings to prioritize rescue missions according to their severity and urgency. The result is a disaster monitoring and response system that is accurate, scalable, and real-time. The DGO-ST-GNN algorithm integrates DGO with a ST-GNN for crisis severity prediction and emergency resource scheduling. The crisis dataset D is represented as a spatial graph $G(U, F)$, with temporal slices Nt capturing dynamic inputs. Optimizer agents Wj are iteratively updated over t = 1 to T using fitness functions $fit(Wj)$, Gaussian mutation, Lévy flight, and opposition-based learning. The best agent Wbest guides ST-GNN training to predict $\hat{Z}$ and generate prioritized emergency schedules Sched.

Algorithm 1: DGO-ST-GNN

Input: *Crisis dataset D, spatial graph G(U, F)*
Output: *Predicted crisis severity* $\hat{Z}$*, Emergency scheduling Sched*
Initialize the population of H optimizer agents Wj *with random hyperparameters*
For t = 1 to T do
  For each agent Wj*, do*
    Apply Dynamic Grasshopper Optimization:
      Update the position of Wj using the GOA governing equations
      Apply Gaussian mutation
      Apply the Lévy flight strategy
      Apply opposition-based learning
    Compute fitness fit(Wj)
End For
  Select the best-performing agent Wbest based on fitness score
End For
Train the ST-GNN model using $Wbest$
For each training epoch, do
  For each time slice Nt *in dataset D, do*
    Construct a spatiotemporal graph from tweet metadata
    Extract temporal dependencies using Gated CNN
    Extract spatial dependencies using Graph CNN
    Fuse outputs through ST-Conv blocks (ReLU + normalization)
    Predict crisis severity $\hat{Z}$
    Compute L2 loss and update network parameters
End For
End For
Generate emergency scheduling based on severity values:
  Sched = prioritize ($\hat{Z}$*)*
Return $\hat{Z}$*, Sched*

3.4.1 Spatiotemporal graph neural network (ST-GNN)

The spatiotemporal spread of public crises is modeled using ST-GNN, which captures the temporal evolution of crisis intensity as well as spatial interdependence across areas. It facilitates precise forecasting of impacted regions, facilitating fast and knowledgeable emergency resource distribution. The stacked Spatio-temporal

Convolution (ST-Conv) blocks, which make up the network architecture, are inspired by STGNN and are each intended to capture dynamic spatial and temporal crisis patterns independently. Figure 2 shows the "sandwich" structure of each ST-Conv block, illustrating data flow from input to output. The model includes three GCN layers (128-dimensional embeddings), two temporal attention layers, an Adam optimizer (0.001), a batch size of 64, dropout 0.3, and 150 training epochs. This consists of one spatial graph convolutional layer surrounded by two gated temporal convolutional layers. The model can capture changing semantic crisis elements (such as urgency and severity) from social data sources, such as geo-tagged tweets, in both time and place, due to this approach.

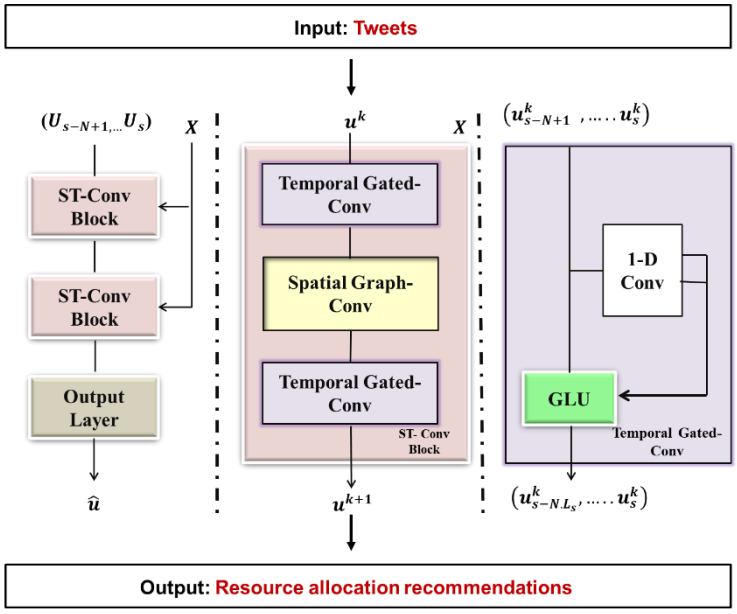


Figure 2: Layer Block of Spatiotemporal GNN architecture

Graph CNNs for Extracting Spatial Features: Urban areas are depicted as nodes in this environment, while the edges are formed by their interactions (such as shared boundaries and mobility linkages). With each node $\mathcal{U} \in$ U representing a distinct location, it may create a spatiotemporal crisis graph $G = (\mathcal{U}, \mathscr{F})$. The convolution operation on the spatial network directly mimics the dispersion of crisis signals and regional interdependence. The spectral graph convolution of an input feature matrix $W \in W \in \mathbb{R}^{n \times d_i}$ (such as real-time social signals like the volume of distress tweets) is defined as follows in equation (1).

$$z_i = \sum_{j=1}^{D_j} \Theta_{j,i}(K) w_j \in \mathbb{R}^m, \quad 1 \le i \le D_p (1)$$

The output feature for the j^{th} output channel following graph convolution is shown by z_i in Equation (1), whereas the input features are represented as $w_i \in \mathbb{R}^m$, where D_i is the number of input feature channels and m is the number of nodes (such as urban areas in a crisis network). Modeled as Chebyshev polynomials of the normalized Laplacian matrix K, the term $\Theta_{i,j}$ (K) denotes the learnable filter parameters that capture the spatial structure of the graph, the interconnectedness of various areas. Each output feature is produced by iterating over all input channels. Where the output tensor $Z \in Z \in Z \in \mathbb{R}^{N \times m \times D_p}$ represents the filtered spatiotemporal features over all N frames, m nodes, and D_p output channels, while the input tensor $W \in W \in W \in \mathbb{R}^{N \times m \times D_j}$ comprises N temporal frames of node characteristics with D_j channels. Each time slice's graph convolution operation is indicated by the notation $\Theta_{*g} w$.

Gated CNNs for Extracting Temporal Features: Capturing non-linear temporal connections is necessary to document the evolution of crises throughout time. To achieve real-time efficiency, Gated Temporal Convolutional Networks (GTCNs) are employed instead of recurrent models. Let a region's historical signal sequence be represented as $W \in W \in W \in \mathbb{R}^{N \times D_j}$. For the gated temporal convolution, it is defined as follows in equation (2).

$$\Gamma *_S W = O \odot \sigma(R) \epsilon \mathbb{R}^{N - L_s + 1) \times D_p} (2)$$

The accessible $1 - D$ temporal convolution kernel that was utilized to identify time-dependent patterns in the crisis data is represented by the symbol Γ. Each of the N historical time steps in the data input sequence $W \in W \in W \in \mathbb{R}^{N \times D_j}$ has D_j input characteristics (such as regional crisis indicators). Two intermediary outputs, O and R, are produced by the convolution process and are both located in the space $\mathbb{R}^{(N - l_s + 1) \times D_p}$, where D_p is the number of output channel features, and L_s is the temporal kernel size. The gating mechanism gated linear unit (GLU) is enabled by the term σ(R), which applies the function of sigmoid activation element-wise to R. In conclusion, ⊙ represents element-wise (Hadamard) multiplication, which gates the output O with the active signal σ(R). This enables the model to reduce noise or unimportant changes while selectively preserving significant temporal characteristics.

Spatio-temporal Convolutional Block: The model's fundamental component is the ST-Conv Block, which combines temporal and spatial crisis dynamics information. Using the input tensor $u^k \in \mathbb{R}^{N \times m \times D_i}$ for each block k, the output $u^{k+1} \in \mathbb{R}^{(N - 2(l_s - 1)) \times n \times D_{j+1}}$ is calculated by equation (3).

$$u^{k+1} = \Gamma_1^l *_S \mathrm{ReLU}\left(\Theta^k *_h \left(\Gamma_0^l *_h u^k\right)\right) (3)$$

For lower and upper temporal kernels, respectively, Γ_0^l and Γ_1^l. Θ^k is the kernel for spatial graph convolution. It adds non-linearity using ReLU. Each block undergoes layer normalization to prevent overfitting. Deeper structures for more complicated crises are constructed by

stacking these elements. The learnt characteristics are converted to a crisis severity prediction $Y \in \mathbb{R}^{m\times d}$ using a fully-connected decoder that comes after the ST-Conv blocks and a temporal projection layer. $\hat{Z}$ defines the ultimate forecast as follows in equation (4).

$$\hat{Z} = YX + a \quad (4)$$

Where the bias vector is denoted by a and the trainable weight matrix by X. The multi-task emergency resource scheduler is adaptive and uses this forecast as input. The training objective is to reduce the L2 loss over time between the actual and expected crisis severity. As equation (5) defines u_{s+1} as the ground truth severity for each area, and X_0 are all learnable parameters.

$$\mathcal{K}(\hat{Z}; X_0) = \sum_s ||\hat{Z}(u_s - N + 1, \dots, u_s; X_0) - u_{s+1}|\hat{Z}|^2 \quad (5)$$

It includes the Dynamic Grasshopper Optimization Algorithm (DGO) in the training loop to optimize network topology and hyperparameters under unpredictable, changing crises. Key parameters (such as graph edge weights and temporal kernel size) are dynamically adjusted to optimize prediction and resource allocation performance.

3.4.2 Dynamic Grasshopper Optimization Algorithm (DGO)

By automatically modifying the ST-GNN model's hyperparameters, DGO improves the model's flexibility and prediction precision. It ensures that the model will perform at its best in a variety of real-time crises. For real-time public crisis management, the deep learning architecture is optimized in this research using the DGOA. Optimal channels for data flow and attention within the spatiotemporal graph are chosen by the DGOA, which also improves the accuracy of crisis propagation modeling and adaptive emergency resource allocation. DGOA introduces three mechanisms to address the limitations of the standard GOA, including restricted global exploration and early convergence. Figure 3 illustrates the dynamic GOA flowchart.

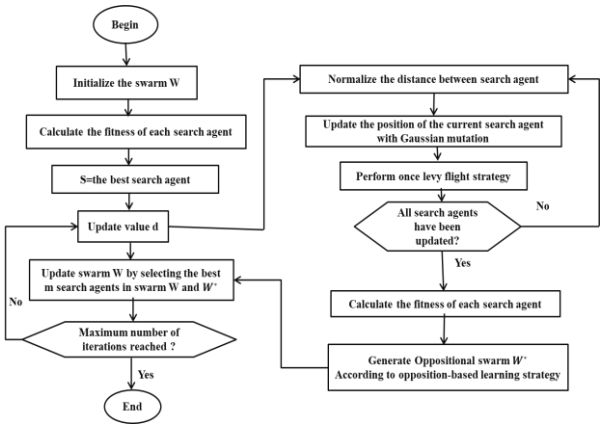


Figure 3: Dynamic GO algorithm flowchart

These mechanisms include Gaussian mutation to maintain variance in dynamic decision spaces, Levy flight to escape local optima in crisis severity prediction, and Opposition-Based Learning (OBL) to enhance spatial diversity. These changes provide the best possible balance between exploitation (concentrating on high-priority or severe zones) and exploration (looking for new crisis patterns or places). Each optimization agent's most recent location is calculated as follows in equation (6).

$$W_j^c = d\left(\sum_{\substack{i=1 \\ i\neq j}}^{M} \frac{va_c - ka_c}{2} \cdot t(|W_i^c - W_i^c|)\frac{W_i - W_j}{c_{ji}}\right) \oplus H(\alpha) + \hat{e}_c(6)$$

Where d is the convergence control coefficient that modifies the search scope over time, and W_j^c is the location of the j^{th}solution agent for the crisis model in dimension c. The Crisis similarity function that measures the degree of influence between solution agents is denoted by $t(\cdot)$. The Gaussian mutation that introduces controlled randomness for exploration is called $H(\alpha)$. Fitness-guided update based on model prediction error is denoted by $\hat{e}_c$. The Bounds specifying the search range for model weights or hyperparameters are va_c and ka_c. To ensure responsiveness to events occurring in real time, this technique dynamically adjusts the crisis model's configuration when new data is received (for example, from geo-tagged tweets).

To improve flexibility in changing emergencies, a Levy Flight mechanism produces new potential solutions as follows in equation (7).

$$W_j^{levy} = W_j + \text{rand}(c) \otimes \text{levy}(\beta)(7)$$

If the new configuration improves predicted performance, the modification is approved in equation (8).

$$W_j^{s+1} = \begin{cases} W_j^{levy}, \text{if fitness}(W_j^{levy}) > fitness(W_j) \\ W_{j,} \text{otherwise} \end{cases} (8)$$

Where c is the problem's dimension, and β is the Levy distribution parameter regulating the step size variability, W_i^{levy} is the new candidate position created by adding a random vector rand (c) scaled by a Levy-distributed step levy (β). In the Levy Flight update equations, W_j is the current position (or solution) of the i^{th} agent in the search space. By contrasting the fitness values of the new candidate W_i^{levy} with the existing position w_j, the updated location W_j^{t+1} is determined. W_i^{levy} is only adopted if it enhances the objective function, guaranteeing adaptive exploration to avoid local optima.

OBL is used to reduce overfitting to high-density zones and optimize spatial awareness. By reflecting current forecasts, this approach creates a parallel solution

set that represents other crisis responses in the equation (9).

$$W_j^{po} = LB + UB - S + q(S - W_j) (9)$$

Where the best-performing configuration at the moment is S, and W_j^{op} is the opposite configuration vector of agent j. The random vector in $(0, 1)$ with controlled noise added is denoted as q. Lower and upper boundaries on the solution space (such as resource limitations, learning rates, and kernel sizes) are denoted by LB and UB. The optimizer can investigate previously unexplored geographical and temporal regions of the crisis graph that could need urgent attention with this method, which enhances responsiveness and generalizability.

The Hyperparameter Configuration Table 1 provides an overview of the principal training and optimization parameters that are utilized to develop the Proposed Dynamic Grasshopper Optimised Spatial-Temporal GNN (DGO-ST-GNN) model. These hyperparameters dictate how the model is learned (i.e., how the weights are trained), at what level of depth, length of temporal sequence, and training regularization for batch optimization.

Table 1: Hyperparameter configuration settings used in DGO-ST-GNN training.

Hyperparameter	Description	Range / Options	Example Value
Learning Rate (lr)	Step size for optimizer	0.001 – 0.01	0.005
Number of GCN Layers	Depth of graph convolution layers	2 – 4	3
Number of LSTM Units	Number of hidden units in the temporal layer	32 – 128	64
Temporal Window	Number of past time steps used for prediction	5 – 20	10
Dropout Rate	Dropout probability to prevent overfitting	0.1 – 0.5	0.3
GCN Activation Function	Activation function in graph convolution	ReLU, LeakyReLU	ReLU
LSTM Activation Function	Activation function for LSTM	tanh, ReLU	tanh
Optimizer	Optimization algorithm for training	Adam, RMSProp	Adam
Batch Size	Number of samples per batch	16 – 64	32
GOA Population Size	Number of candidates hyperparameter solutions	5 – 20	10
GOA Iterations	Number of iterations for Grasshopper Optimization	10 – 30	15
Weight Decay	L2 regularization to prevent overfitting	0 – 0.01	0.001
Number of Epochs per Candidate	Training epochs for each GOA candidate	20 – 100	50
Final Training Epochs	Epochs to train the final model after GOA selection	50 – 200	100

3.5 Multi-task rescue scheduling algorithm

Rescue teams continually poll the updated priority queue, which is sorted by severity, urgency, and arrival time, then select the mission with the greatest priority after the DGO-ST-GNN calculates crisis severity scores for each region. Next, within a certain radius, each team searches for neighboring assignments whose aggregate rescue capacity won't be reached. More tasks are organized into the same deployment to save travel overhead if they are within range and resources allow. Teams are dynamically re-tasked to regions of greatest need when hotspots alter due to the rebalancing of the priority queue and the refresh of regional severity scores upon mission completion. Adaptive multi-task rescue scheduling's core principles are priority, proximity, and resource availability, as shown in equation (10).

$$\arg\max_{S_j \in \mathcal{M}(S_l, Q)} \left(\frac{O(S_j)}{C(S_l, S_j)} \cdot \mathbb{I}[Q_l \geq D(S_j)] \right) (10)$$

The current selected high-priority task is denoted by S_l. $\mathcal{M}(S_l, Q)$ is the collection of tasks that are selected by S_l from Q. Based on severity, urgency, and arrival time, task S_j's priority score is $O(S_j)$. Tasks S_l and S_j are separated by a distance, $C(S_l, S_j)$. The cost of resources (such as labor, equipment, and time) to do task S_j is denoted by $D(S_j)$. Q L is the rescue processor's or team's remaining resources that are allocated to S_l. $\mathbb{I}[\cdot]$ defines the indicator function, which returns 0 otherwise and 1 if the condition inside is true (i.e., the task is feasible). Further nearby tasks S_j that fit within the available resources Q_l, are close to the present task S_l, and have a high priority $O(S_j)$ are selected by this equation. The group dynamically creates an effective multitasking mission by optimizing the cost-benefit ratio while adhering to time restrictions.

4 Performance evaluation

Real-world crisis data from social media was used in a series of experiments to assess the efficacy of the suggested DGO-ST-GNN model. The results of these experiments show the beneficial effects of the suggested strategy. The experimental design was implemented using Python 3.10.

The top 10 frequent words that came out of the crisis-related social media data after preprocessing. The word 'evacuate' pops up most times (892), followed by 'emergency' (890), and then 'fire' (889). Other high-frequency words include rescue (887), urgent (885), and water (880), which are observed in some key contexts within crises. Tokens like ‘trapped’ (840 occurrences) and ‘help’ (850) highlight emergency conditions, as shown in Figure 4. These frequent keywords are very useful in semantic analysis as well as real-time crisis tracking.

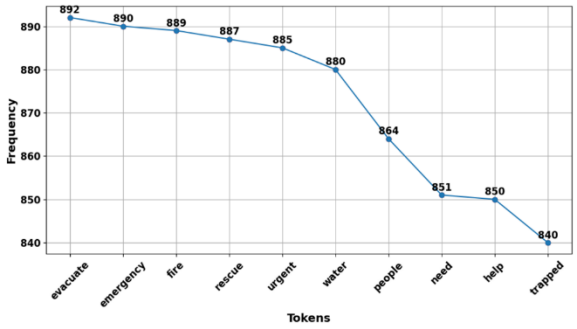


Figure 4: Frequency analysis of key crisis-related tokens preprocessing

Word2Vec sentence embeddings of about 1,030 crisis tweets were reduced to two dimensions via PCA for visualization. Each point is a tweet encoded into a vector that carries the semantic meaning of this text, as shown in Figure 5. These embeddings are spread over a component 1 range from -50 to +45 and component 2 from -25 to +25, thus reflecting diversity both in content and emotional tone within this dataset. Such a wide spread shows that the model can distinguish different crisis contexts based on text content.

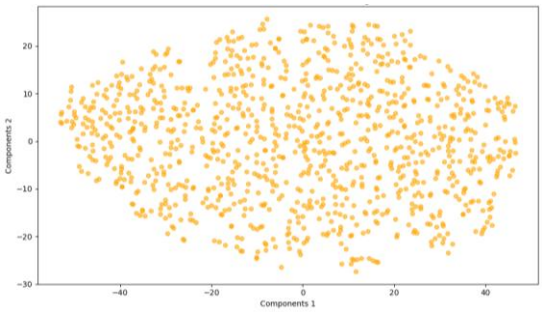


Figure 5: Semantic similarity in crisis tweets using Word2Vec embeddings

The connection between resource types and crisis types helps with resource scheduling optimization. For example, industrial accidents required rescue teams the most (66), whereas earthquakes needed medical units the most (58). Fires required the fewest medical units (36), whereas pandemics required the most ambulances (61) and fire trucks (63), as shown in Figure 6. This distribution contributes to the model's objective of adaptive multi-task rescue allocation by identifying which crisis circumstances require which type of emergency response, resulting in improved, data-driven deployment decisions.

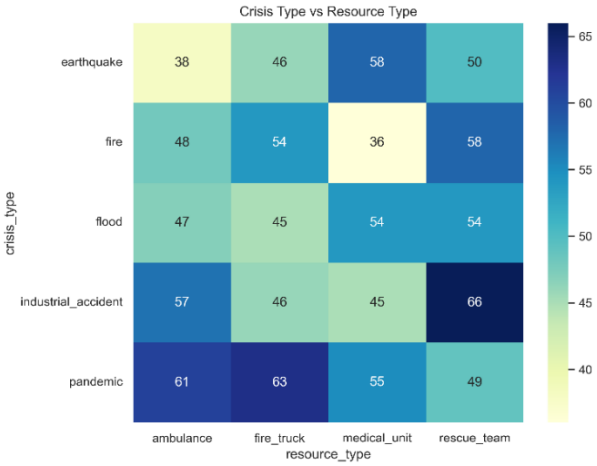


Figure 6: Crisis type vs emergency resource allocation matrix

The trend of crisis occurrence reports over 7 days supports the proposed system's real-time monitoring purpose. Reports increased from 97 on July 25 to 165 on July 31, indicating that the problem was escalating. A reduction to 33 on August 1 indicates probable resolution or reporting lag, as shown in Figure 7. These temporal variations allow the DGO-ST-GNN model to adaptively prioritize resource allocation by determining when and where crises worsen, which aligns with the goal of dynamic and efficient emergency response.

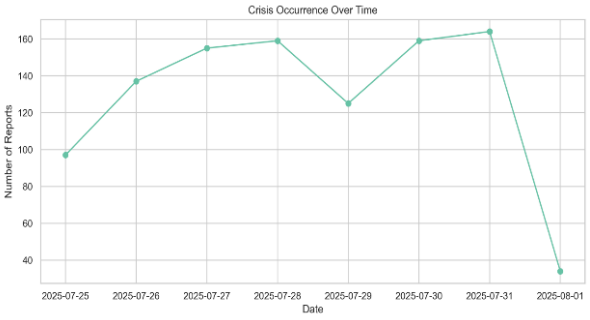


Figure 7: Temporal trend of crisis reports for dynamic emergency response

The sentiment distribution of crisis-related social media posts was examined as part of the feature extraction process for the proposed crisis management model. Sentiment ratings vary from -1.0 (negative) to 1.0 (positive), with peaks at -0.6, 0.2, and 0.9, suggesting emotional variability during public emergencies, as shown in Figure 8. This distribution, with most values ranging from 40 to 62 occurrences, contributes to the model's capacity to identify emotional tone in real time, allowing for dynamic prioritizing and optimal emergency resource allocation depending on the public's positive attitude.

Figure 8: Sentiment analysis of crisis-related social media posts for emergency response prioritization

The findings show that social media sentiment, temporal patterns, and crisis-resource connections all give valuable information for real-time crisis management. The suggested model accurately reflects emotional tone, fluctuations in crisis frequency, and optimal resource deployment patterns. It *classification performance for crisis severity prediction* 95% confirm the ability of the system to make dynamic, accurate, and efficient emergency response decisions.

4.1 Comparative analysis

In public disasters, to minimize damage and preserve lives, quick, data-driven decision-making is essential. By utilizing intelligent models and real-time data, crisis response efficacy may be greatly increased. Existing techniques in public crisis management, such as Rule-based + support vector machine (SVM) approaches and CNN-GRU, have significant drawbacks. Although CNN-GRU [23] appears to be effective at interpreting language, it is not particularly adept at integrating temporal and spatial dynamics, which makes it less useful to track how crises change over time in various regions. Additionally, it uses a lot of processing power and cannot function effectively in real-time situations. However, the Rule-based + SVM [24] approach was limited in its ability to adapt to new

types of crises or hidden data patterns since it relies on static rules and predefined features. The dynamic resource optimization and multi-modal integration necessary for real-time emergency response are challenges for both strategies.

Table 2 shows the comparison of the proposed DGO-ST-GNN model with the existing baseline models. As indicated by the results, there were considerable gains for Metrix's proposed model over the baseline models, with the proposed model exhibiting 97% accuracy and being therefore very effective in forecasting the severity of social crises and allocating resources. Figure 9 (a-d) shows the accuracy, and precision, recall, and F1-score.

Table 2: Performance comparison of crisis management approaches

Models	Accuracy	Precision	Recall	F1
Rule-based + SVM [27]	-	89.4	77.7	83.1
CNN [28]	72	72	72	72
BERT [28]	78	78	79	78
XLNet [28]	77	77	77	77
Naïve Bayes [29]	76	70	76	72
Random Forest [29]	77	70	77	73
Logistic Regression [29]	77	70	77	73
SVM [29]	89	88	89	88
LSTM [29]	87	86	87	86
DGO-ST-GNN (Proposed)	97	95	96	94.9

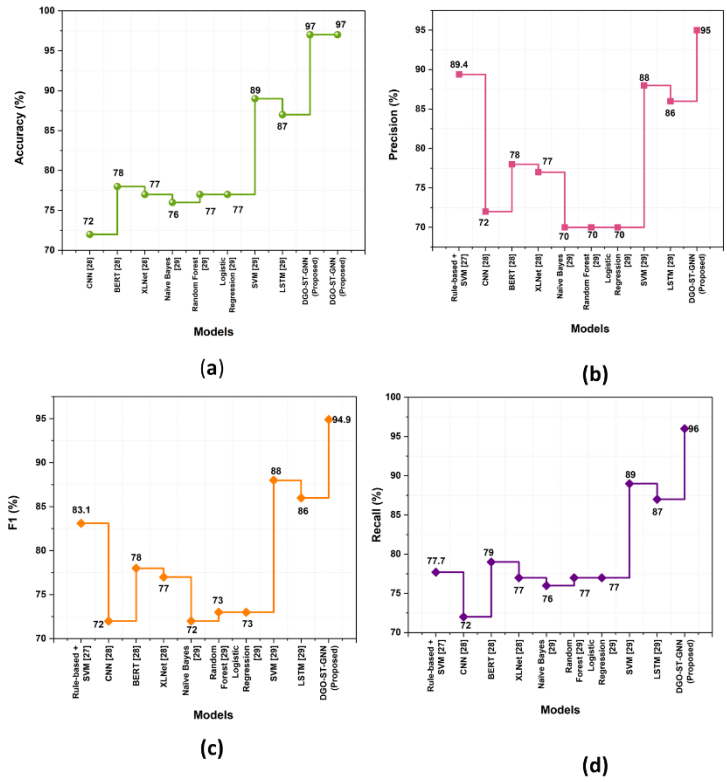


Figure 9: Comparison of the existing methods of the Crisis Management Approaches (a) accuracy, (b) precision , (c) recall, and , (d) F-1 score.

Table 3 presents a detailed account of how the DGO-ST-GNN's prediction error compares in Figure 10 favorably with the available state-of-the-art deep learning architectures based on MAE and E-MWDNN. The analysis shows that DGO-ST-GNN achieves MAE (0.081427) and E-MWDNN (25.934112) lower than any model currently available, thereby providing the most accurate and robust predictions of future crises severity relative to other models, including CNN-LSTM and attention-based neural network models.

Table 3: Comparative error performance of baseline models vs. proposed DGO-ST-GNN

Model	**MAE**	**Error-MWDNN**
DNN [30]	0.097373	30.913406
CNN [30]	0.095262	30.243204
LSTM [30]	0.088246	28.015784
CNN-LSTM [30]	0.087153	27.668873
CNN-LSTM-Skip [30]	0.124239	39.442727
CNN-LSTM-Skip Attention [30]	0.105618	33.53082
DGO-ST-GNN (Proposed)	0.081427	25.934112

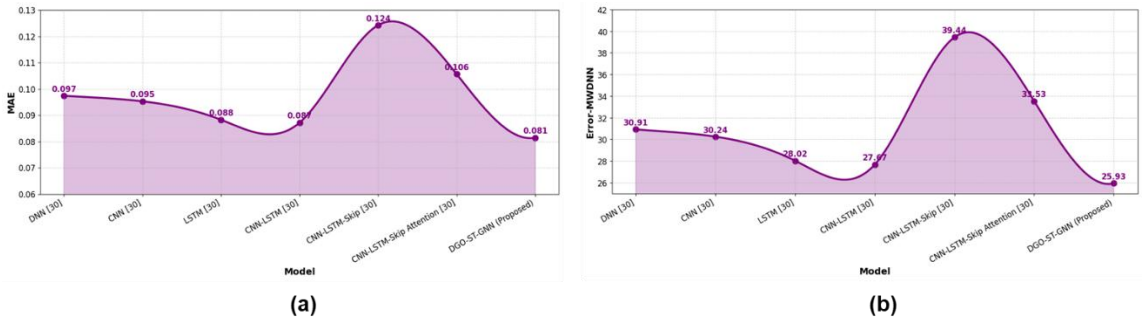


Figure 10: Comparison of prediction errors across Different Deep Learning Models (a) MAE, and (b) Error-MWDNN

4.2 Statistical outcomes

The t-test Table 4 compares the resource use in the high-severity and low-severity categories of crisis, which confirms significant differences in behaviours in terms of emergency requirements. This fact helps to confirm that DGO-ST-GNN is efficient in prioritization of critical decisions regarding resource allocation based on severity prediction and enhances the relevance of emergency optimization to the real world. A significant difference exists in resource utilization between high-severity and low-severity crisis events ($p < 0.05$).

Table 4: T-test comparison of severity scores across

Group	N	Mean	Std. Dev	t-value	p-value	Significance
High Severity (≥0.6)	512	48.7	9.5	**4.82**	**0.000**	Significant
Low Severity (<0.6)	518	31.2	10.1			

Table 5 ANOVA test compares the severity differences among various types of crises and finds statistical significance, proving that the features of crisis have an impact on the level of severity. This observation supports the necessity of spatiotemporal modelling and the rationale of the adaptive learning ability of the suggested DGO-ST-GNN.

Table 5: ANOVA comparison of severity scores across different crisis types

Source	SS	df	Mean Square	F	p-value	Significance
Between Groups	8.213	4	2.053	**6.41**	**0.0003**	Significant
Within Groups	327.29	1025	0.319			
Total	335.50	1029				

4.3 Scheduling outcomes

The inference time per prediction instance was evaluated and compared with baseline models, confirming the efficiency required for real-time emergency scheduling. DGO-ST-GNN Establishes Adaptive Optimization and Intelligent Resource Prioritization based upon the predicted severity of crisis type to reduce wait times. Table 6 shows the reductions in both Maximum and Mean average Wait Times across 10p/20p Workloads, indicating a superior level of Efficiency for Emergency Scheduling over both Hybrid and Traditional algorithm-based methods.

Table 6: Emergency scheduling performance comparison including proposed DGO-ST-GNN

Algorithms	Max avgWT		Mean avgWT	
	(10p)	(20p)	(10p)	(20p)
FCFS [31]	4.74	3.73	2.53	1.61
Priority [31]	5.54	3.85	2.81	1.63
Multi-tasks Hybrid [31]	4.47	3.02	2.24	1.31
DGO-ST-GNN [Proposed]	**3.62**	**2.41**	**1.68**	**0.94**

5 Discussion

The comparative performance analysis indicates that the proposed DGO-ST-GNN model is superior to the other recently popular (SOTA) crisis forecasting and emergency response scheduling models. Past research has used hybrid deep learning networks, reinforcement learning systems, multimodal neural networks, and spatiotemporal graph networks in crisis management tasks, as discussed in the SOTA comparison (Table 7). Even though a significant number of them were competitive, they could exhibit a limited applicability beyond the dataset or scenario they were

trained on. As an example, hybrid CNN-BiLSTM models were found to have an F1-score of 96% but needed huge labeled datasets to work effectively, and multimodal systems were highly precise but had issues with data noise. By contrast, DGO-ST-GNNclassification performance for crisis severity prediction has 97% accuracy, 95% precision, 96% recall, and 94.9% F1-score, significantly better than both traditional machine-learning models and deep-learning baselines because of adaptive hyperparameter tuning and extensive spatiotemporal modeling. Thus, the suggested solution implements a more scalable, data-constrained, and crisis-resistant structure in comparison to the literature.

Table 7: SOTA-based comparison of related emergency response and crisis prediction methods with the proposed DGO-ST-GNN framework

Study	Model Type	Data Source	Performance Metrics	Main Limitations
[13]	Hybrid deep learning (CNN + BiLSTM + Self-Attention)	Social media short texts (3 real-world crisis datasets)	F1-score: **96%**	Requires large labeled datasets; high computational load
[15]	Hybrid CNN-based text classification	Crisis NLP datasets (Twitter)	Accuracy improvement: **+21.71%** over baselines	Relies heavily on labeled data; limited generalization across crisis types
[16]	DRL + Optimization (DDPG + MILP)	Storm-surge logistics data	MILP: highest accuracy; DDPG: fastest execution	MILP is slow and not scalable; DDPG loses accuracy in complex scenarios
[22]	Multimodal (Image + Text) Deep Learning	Social media multimodal crisis dataset (CrisisMMD)	Accuracy: **91.53%**	Requires high-quality image–text pairs; sensitive to noise
[25]	**CRISP (GCN + BiLSTM + Attention)**	Financial crisis index and market time-series data	Improved stability and accuracy during financial crisis prediction	Limited to financial applications; not suitable for real-time emergency response scenarios
[27]	**Highway Emergency ST-Network**	Highway traffic flow and rescue station O-D data	Reduced average rescue response time	Assumes static traffic pattern; lacks predictive capability in rapidly evolving crises
DGO-ST-GNN [Proposed]	Spatiotemporal Graph Neural Network + Dynamic Grasshopper Optimization	Geo-tagged crisis tweets (1,030 records)	**Accuracy:97%** Precision: **95%**, Recall: **96%**, F1-score: **94.9%**	Depends on social media data quality; computational overhead in GNN optimization.

5.1 Reasons for the existence of better performance of DGO-ST-GNN.

The two key innovations that make the proposed model offer exceptional performance include the DGO that optimally modulates hyperparameters in real time over training epochs, reduces hyperparameter sensitivity to local minima, and eliminates the need for manual tuning. Spatiotemporal GNN architecture is able to acquire relational features between geographically dispersed nodes of crisis and dynamic temporal dynamics of crisis intensity. The conventional approaches support crisis signals as discrete text collections or fixed data sets, which leads to the loss of important information. In comparison, DGO-ST-GNN is a dynamically integrated framework of semantic embeddings, spatial graphs, and temporal convolutional layers, which allows forecasting crisis propagation. Another technique that is presented in the optimization strategy is the use of Gaussian mutation, Levy flight, and opposition-based learning, which all contribute to increased stability in model convergence and model representation. As a result, the accuracy of the decision made on resource allocation is enhanced, which makes the average emergency response wait time shorter.

5.2 Resilience and applicability to the scope of crises

The huge benefit of the suggested framework is that it can be adapted to various types of crises such as fires, floods, earthquakes, pandemics, and industrial accidents. Rather than using domain-specific rule-based assumptions, DGO-ST-GNN acquires emergent patterns based on geo-tagged spatiotemporal data, semantic sentiment severity indicators, and past emergency resource behaviour. In contrast to CRISP, which models financial crises only, or highway ST-Networks, which presuppose that the flow of traffic is always constant, the proposed architecture can be generalized well in terms of event types and geographical area. The analysis shows consistency in its performance

with a comparatively small dataset of 1,030 samples, which is supported by data augmentation and cross-validation as well as close train-test partitioning. The model also exhibits resilience to imbalance in the dataset and noisy tweet content, which demonstrates that it can be deployed in the real world to support the use of public safety organizations and smart crisis management systems.

5.3 Real-time emergency response practical implications

The usefulness of the proposed framework is not limited to the predictive accuracy but to the operational improvement of the response. Combining spatiotemporal forecasting with emergency resources scheduling allows the proactive decision-making of firefighters, medical teams, and relief organizations. The suggested model will reduce the number of emergencies wait times to a minimum and enhance the prioritization of areas of crisis severity, thereby minimizing the number of casualties and loss of infrastructure. The increase in the efficiency of scheduling (Mean avgWT dropped by the baseline values to 0.94 seconds) shows that intelligent predictive planning may have a tangible effect in the real world. This can be crucial to emergency response systems in a smart-city environment and the next generation of disaster informatics with high-density urban environments, where a minute can be the difference between life and death.

5.4 Parallelization potential and bottleneck analysis

The highlight of the proposed DGO-ST-GNN architecture is that it allows parallelization based on distributed computing of GCN layers in the various graph partitions and parallel mini-batch computation in the temporal LSTM layers, which allow it to be trained much faster on multi-GPU set-ups. Parallel fitness evaluation of candidate solutions is also available to the Dynamic Grasshopper Optimization. Nonetheless, there exist certain bottlenecks in the construction of graphs and the computation of attention, where quite dense spatial dependencies can be subject to further optimization solutions, e.g., sparse adjacency management or CUDA-based kernel execution to support real-time implementation.

5.5 Limitations and future research directions

The given system, however, has its advantages; still, it has its limitations, which need to be recognized. Nevertheless, despite the benefits of DGO in terms of accuracy, DGO requires more training time by about 38.7% and in the case of large-scale deployments, real-time optimization becomes computationally intensive. Also, there is a dependence on social-media text information, which influences the accuracy of prediction based on the clarity of messages, misinformation, and regional reporting bias. The further work will be directed at the integration of multimodal input streams, including satellite imagery, IoT sensor data streams, and government streams of emergencies. In addition, semi-supervised and weak-supervision methods of learning will be considered to remove training reliance on labeled data, as well as enhance interpretability. It shall also focus on enhancing cross-regional crisis domain transfer learning capabilities to improve the generalizability across the world.

6 Conclusion

The suggested DGO-ST-GNN is capable of both crisis propagation and optimization of emergency resources based on geo-tagged crisis data. Combining the spatiotemporal blocks of graph convolution with adaptive parameter optimization using the framework ofDGO, the model shows superior capabilities in predicting and strong response to dynamic crisis signals. Existing experimental analysis indicates better accuracy, precision, and recall than the traditional machine learning and deep learning baselines and the ability to make decisions in emergency situations in real time, which demonstrates the model. The future efforts will be on expanding the model to other real time sources of data, multimodal information integration and computational efficiency to facilitate quick deployment. The suggested framework provides a data-driven and scalable approach to dealing with crisis management.

References

[1] Kazadi, A., Doss-Gollin, J., Sebastian, A., and Silva , A., 2024. FloodGNN-GRU: a spatio-temporal grap h neural network for flood prediction. *Environment al Data Science*, 3, p.e21. https://doi.org/10.1017/ed s.2024.19

[2] Lopez, V.K., Nika, A., Blanton, C., Talley, L., and Garfield, R., 2023. Can the severity of a humanitari an crisis be quantified? Assessment of the INFORM severity index. *Globalization and health*, 19(1), p.7. https://doi.org/10.1186/s12992-023-00907-y

[3] Das, S., Choudhury, M.R., Chatterjee, B., Das, P., B agri, S., Paul, D., Bera, M., and Dutta, S., 2024. Un raveling the urban climate crisis: Exploring the nex us of urbanization, climate change, and their impact s on the environment and human well-being–A glob al perspective. *AIMS Public Health*, 11(3), p.963. ht tps://doi.org/10.3934/publichealth.2024050

[4] Amatya, B. and Khan, F., 2023. Disaster response a nd management: the integral role of Rehabilitation. *Annals of Rehabilitation Medicine*, 47(4), pp.237-2 60. https://doi.org/10.5535/arm.23071

[5] Erokhin, D. and Komendantova, N., 2024. Social m edia data for disaster risk management and research . *International Journal of Disaster Risk Reduction*, 114, p.104980. https://doi.org/10.1016/j.ijdrr.2024. 104980

[6] Firmansyah, H.B., 2025. Improving Disaster Respo nse With Advanced Machine Learning ToAnalyse S ocial Media Content. https://doi.org/10.13097/archi ve-ouverte/unige:183585

[7] Ginzarly, M., Teller, J. and Dujardin, S., 2025. Soci al media use in disaster response: empowering com

munity resilience. *International Journal of Digital E arth*, 18(1), p.2521791. https://doi.org/10.1080/175 38947.2025.2521791

[8] Solovev, K., 2024. Leveraging Unstructured Data to Address Societal Challenges in the Digital Age (Do ctoral dissertation, Justus Liebig University Giessen). https://doi.org/10.22029/jlupub-19035

[9] Mizrak, K.C., 2024. Crisis management and risk mi tigation: Strategies for effective response and resilie nce. *Trends, challenges, and practices in contempo rary strategic management*, pp.254-278. https://doi. org/10.4018/979-8-3693-1155-4.ch013

[10] Fang, H.S., Wang, C., Fang, H., Gou, M., Liu, J., Y an, H., Liu, W., Xie, Y., and Lu, C., 2023. Anygras p: Robust and efficient grasp perception in spatial a nd temporal domains. *IEEE Transactions on Roboti cs*, 39(5), pp.3929-3945. https://doi.org/10.1109/TR O.2023.3281153

[11] Qhal, E.M.A., 2025. Data-Driven Knowledge Mana gement Frameworks for Effective Risk and Crisis M anagement: A Cross-Industry Approach. *Journal of Applied Data Sciences*, 6(2), pp.1437-1453. https:// doi.org/10.47738/jads.v6i2.780

[12] Zhou, Y. and Xia, W., 2021. Optimization algorith m and simulation of public resource emergency sch eduling based on wireless sensor technology. *Journ al of Sensors*, 2021(1), p.2450346. https://doi.org/1 0.1155/2021/2450346

[13] Zhao, X. and Wang, G., 2023. Deep Q-networks-ba sed optimization of emergency resource scheduling for urban public health events. *Neural Computing a nd Applications*, 35(12), pp.8823-8832. https://doi.o rg/10.1007/s00521-022-07696-2

[14] Wang, Y., Chen, X., and Wang, L., 2023. Deep rein forcement learning-based rescue resource distributi on scheduling of storm surge inundation emergency logistics. *IEEE Transactions on Industrial Informat ics*, 19(10), pp.10004-10013. https://doi.org/10.110 9/TII.2022.3230691

[15] Upadhyay, A., Meena, Y.K., and Chauhan, G.S., 20 24. SatCoBiLSTM: Self-attention-based hybrid dee p learning framework for crisis event detection in so cial media. *Expert Systems with Applications*, 249, p .123604. https://doi.org/10.1016/j.eswa.2024.12360 4

[16] Wang, C., Deng, D., Xu, L., and Wang, W., 2022. R esource scheduling based on deep reinforcement lea rning in UAV-assisted emergency communication n etworks. *IEEE Transactions on Communications*, 7 0(6), pp.3834-3848. https://doi.org/10.1109/TCOM M.2022.3170458

[17] Powers, C.J., Devaraj, A., Ashqeen, K., Dontula, A. , Joshi, A., Shenoy, J., and Murthy, D., 2023. Using artificial intelligence to identify emergency messag es on social media during a natural disaster: A deep learning approach. *International Journal of Informa tion Management Data Insights*, 3(1), p.100164. htt ps://doi.org/10.1016/j.jjimei.2023.100164

[18] Seid, A.M., Boateng, G.O., Anokye, S., Kwantwi, T ., Sun, G. and Liu, G., 2021. Collaborative computa tion offloading and resource allocation in multi-UA V-assisted IoT networks: A deep reinforcement lear ning approach. *IEEE Internet of Things Journal*,8(1 5), pp.12203-12218. https://doi.org/10.1109/JIOT.2 021.3063188

[19] Zhang, Y., 2025. Media framing and public risk co mmunication: Deep Learning-based crisis narrative analysis and optimization. *Future Technology*, 4(3), pp.227-238. https://doi.org/10.55670/fpll.futech.4.3 .21

[20] Yan, L., Ren, Z., Zhang, Y., Tao, Z. and Zhao, Y., 2 024. Constructing the public opinion crisis predictio n model using CNN and LSTM techniques based on social network mining. http://dx.doi.org/10.9781/iji mai.2024.07.005

[21] Chen, M. and Du, W., 2023. Predicting public senti ment evolution on public emergencies under deep le arning and the Internet of Things. *The Journal of Su percomputing*, 79(6), pp.6452-6470. https://doi.org/ 10.1007/s11227-022-04900-x

[22] Shetty, N.P., Bijalwan, Y., Chaudhari, P., Shetty, J. and Muniyal, B., 2025. Disaster assessment from so cial media using multimodal deep learning. *Multime dia Tools and Applications*, 84(18), pp.18829-1885 4. https://doi.org/10.3390/ijgi13010029

[23] Paul, N.R., Sahoo, D., and Balabantaray, R.C., 2023 . Classification of crisis-related data on Twitter usin g a deep learning-based framework. *Multimedia To ols and Applications*, 82(6), pp.8921-8941. https://d oi.org/10.1007/s11042-022-12183-w

[24] Li, Z., & Fan, R. (2025). Crisis-Resilient Portfolio Management via Graph-based Spatio-Temporal Lea rning. *arXiv preprint arXiv:2510.20868*. https://doi. org/10.48550/arXiv.2510.20868

[25] Mo, G., Jia, W., Tan, C., Zhang, W., & Rong, J. (20 25). Risk Prediction of International Stock Markets with Complex Spatio-Temporal Correlations: A Spa tio-Temporal Graph Convolutional Regression Mod el Integrating Uncertainty Quantification. *Journal o f Risk and Financial Management*, *18*(9), 488. https ://doi.org/10.3390/jrfm18090488

[26] https://www.kaggle.com/datasets/programmer3/publ ic-crisis-spatiotemporal-response-dataset/data.

[27] Shen, H., Ju, Y., and Zhu, Z., 2023. Extracting usef ul emergency information from social media: A met hod integrating machine learning and rule-based cla ssification. *International Journal of Environmental Research and Public Health*, 20(3), p.1862. https:// doi.org/10.3390/ijerph20031862

[28] Powers, C. J., Devaraj, A., Ashqeen, K., Dontula, A ., Joshi, A., Shenoy, J., & Murthy, D. (2023). Using artificial intelligence to identify emergency messag es on social media during a natural disaster: A deep learning approach. International Journal of Informat ion Management Data Insights, 3(1), 100164.https:/ /doi.org/10.1016/j.jjimei.2023.100164

[29] Airlangga, G. (2024). Comparative analysis of mac hine learning models for real-time disaster tweet cla ssification: Enhancing emergency response with soc ial media analytics. Brilliance: Research of Artificia l Intelligence, 4(1), 25-31. https://doi.org/10.47709/ brilliance.v4i1.3669

[30] Khafaji, K. M., & Hamed, B. B. (2024). Employing deep learning in crisis management and decision ma

king through prediction using time series data in M osul Dam, Northern Iraq. PeerJ Computer Science, 10, e2416. https://doi.org/10.7717/peerj-cs.2416

[31] Kabir, M. Y., & Madria, S. (2019). A Deep Learnin g Approach for Tweet Classification and Rescue Sc heduling for Effective Disaster Management (Indus trial). https://doi.org/10.1145/3347146.3359097

Variable explanation

Symbol / Term	Definition (Verified)
$\mathbf{X}$, $\mathbf{X}^t$	Input tensor containing crisis features across time frames
$\mathbf{Y}$, $\mathbf{Y}^t$	Output tensor after spatial or temporal convolution
$\boldsymbol{\theta}_i$	Learnable Chebyshev graph convolution filter for channel *i*
$\mathbf{C}_i$	Number of input channels
$\mathbf{C}_o$	Number of output channels
$\mathbf{T}$	Number of temporal frames in the sequence
$\mathbf{N}$	Number of nodes (geographical regions)
$\tilde{\mathbf{L}}$	Normalized graph Laplacian matrix
$\mathbf{k}$	Chebyshev polynomial order
$\mathbf{W}_t$	Temporal convolution kernel
$\mathbf{Z}$, $\hat{\mathbf{Y}}$	Intermediate convolution outputs before gating
$\boldsymbol{\sigma}$**(.)**	Sigmoid activation used for gating
$\odot$	Element-wise multiplication
$\mathbf{H}^l$	Output of the *l*-th ST-Conv block
$\mathbf{W}_s$	Spatial graph convolution kernel
$\boldsymbol{\phi}$**(.)**	Activation function (ReLU)
$\mathbf{W}_p$, $\mathbf{b}_p$	Output layer weights and bias
$\hat{y}$	Final predicted crisis severity
L, **L(**$\boldsymbol{\Theta}$**)**	Loss function measuring prediction error
$\boldsymbol{\Theta}$	Set of all learnable parameters
$\mathbf{y}$	Ground-truth crisis severity
c	Convergence control coefficient of DGOA
$\mathbf{X}_i$	Current position of optimization agent *i*
S(t)	Social interaction influences function
Gaussian_mut	Gaussian mutation for exploration
fit(.)	Fitness function (prediction error)
rand()	Uniform random value in [0,1]
levy()	Step size generated by Lévy flight
$\mathbf{X}_i'$	New candidate position after Lévy update
$\mathbf{X}_i$**(new)**	Updated agent position after fitness check
D	Dimensionality of the search space
$\mathbf{X}_i$**^opp**	Opposite solution in the OBL mechanism
LB, **UB**	Lower and upper bounds of the search space
p(t)	Priority score of task *t*
d(t, t′)	Spatial distance between tasks
r(t)	Resources required by task *t*
$\mathbf{R}_a$	The remaining resources of the rescue team *are*
I(condition)	Indicator function (1 = feasible, 0 = not feasible)
$\mathbf{T}_s$	Set of selected nearby tasks
$\mathbf{T}_a$	All tasks available for *Team A*

HybridCardioNet: A CNN-LSTM-Based Deep Learning Framework for ECG Signal Classification and Cardiac Anomaly Detection

A Bharath[1*], G Merlin Sheeba[2]
[1]Research Scholar, CSE Department, SIST, Chennai, Tamil Nadu, China
[2]Professor ECE Jerusalem College of Engineering, Chennai
E-mail: merlinsheeba.ece@jerusalemengg.ac.in, bharath.andugula@gmail.com
[*]Corresponding author

Keywords - HybridCardioNet, ECG Classification, Deep Learning, CNN-LSTM, Cardiac Condition Detection

Received: January 25, 2025

Adaptive learning and automatic classification of ECG signals is one of the commonly processed and practical methods in cardiac anomalies detection with that area has a huge potential to teach clinicians for better clinical healthcare decision making and also remote health patient monitoring [7,8]. Past methods tackle issues like spatial–temporal feature extraction, class imbalance and dataset generalisation, but they are limited by a number of shortcomings: the traditional ML models depend on hand-crafted features and thus lack scalability, and the stand-alone deep models (CNNs or LSTMs) do not capitalize on spatial and sequential information simultaneously. To overcome these shortages, we present HybridCardioNet, a joint deep-learning framework that integrates CNN-based spatial feature extraction and LSTM based temporal-dependency modelling. The ECG was filtered using band-pass filtering (0.5–40 Hz) to remove the low-frequency baseline-wander, z-score normalisation, and segmented into single-beat segments using the R-peak detection from the MIT-BIH Arrhythmia Database; class imbalance was handled via class-weighted loss (random minority oversampling provided validation) HybridCardioNet with stratified cross-validation gives 98.39% accuracy with the same balanced precision, recall and macro-F1 score. Against popular protocols in recent literature on MIT-BIH, it also achieves competitive performance against internal baselines (CNN-only, LSTM-only and classical ML). Thus, hybridCardioNet solves one key limitation of the previously existing methods. With regards to the other two limitations, since hybridCardioNet is able to outperform the state-of-the-art for multi-class ECG classification, hybridCardioNet will be appropriate for real-time applications in terms of early detection & continuous ECG signals clinical/remote monitoring.

Povzetek: HybridCardioNet združi CNN in LSTM za samodejno klasifikacijo EKG ter z ustrezno predobdelavo in obvladovanjem neuravnoteženih razredov doseže zelo visoko natančnost (98,39 %) za zgodnje odkrivanje srčnih nepravilnosti in spremljanje na daljavo.

1 Introduction

Analysis of ECG signals is essential for diagnosing cardiac diseases since it provides necessary information regarding the heart's electrical activity. In particular, there has been growing interest in automated ECG classification to improve diagnostic accuracy and efficiency in clinical and remote monitoring environments. Although many works are investigating this problem, there are some issues with how ECG signal spatial and temporal features are captured in the previous methods. Legacy machine learning models, such as SVM and random forest models, require manual features to work, making them less scalable and robust. In contrast, deep learning methods such as CNNs [1] and LSTM networks [2] used independently cannot learn spatial and sequential dependencies in ECG data at a high level.

The literature shows that CNN benefits spatial feature extraction, such as amplitude and time intervals in the ECG waveform, while LSTM captures temporal dependencies. Isolated versions of these models do not fully utilize the complementary nature of spatial and temporal modalities. Furthermore, currently, available frameworks suffer from class imbalance issues without preprocessing and feature engineering strategies dedicated to ECG data. This gap emphasizes the lack of a hybrid deep learning framework that is sustainable, reliable, and optimal for ECG signal classification with scalability.

This study aims to create a hybrid deep learning model, HybridCardioNet, that combines CNNs and LSTMs to classify ECG signals by learning spatial and temporal features. The paper presents several novelties, from the novel preprocessing approach to strong feature engineering based on time-domain and frequency-domain analysis and the combined CNN-LSTM architecture. It also includes various techniques to balance the data to counteract the effects of class imbalance, increasing the generalizability and reliability of the model.

Our study aims to investigate whether hybrid spatial–temporal modeling leads to statistically significant better ECG classification performance than the corresponding non-hybrid baselines. Specifically, we can ask (i) whether a CNN→LSTM pipeline (HybridCardioNet) attains a superior macro-F1 than either CNN- or LSTM-only models under stratified cross-validation, and (ii) whether explicit class-imbalance controls—brought about both by class-weighted loss and oversampling, verified not to suppress population performance—improves rarity-class recall at no significant expense in aggregate performance. We assume that the morphology will be learnt by the CNNs while the LSTM will learn the beat-to-beat context both of which operate to perform majority-class reduction, and thus stabilize per-class metrics. Expected pros: improved minority-class behaviour, disentangled precision recall; Expected cons: more sequential compute, segmentation/preprocessing sensitive. As a result, we profile Params/FLOPs and latency, while reporting exact p-values with confidence intervals for each accuracy–efficiency trade-off.

The main contributions of this work are: (1) design of a hybrid deep learning framework, HybridCardioNet, using CNNs and LSTMs for ECG signal classification; (2) employing powerful preprocessing and feature engineering techniques specific to ECG data; (3) assessment of the framework using the MIT-BIH Arrhythmia Dataset with state-of-the-art accuracy of 98.39%, (4) and comparison of the proposed model with traditional as well as stand-alone models to highlight its significance.

The organization of this paper is as follows: A comprehensive literature review is provided in Section 2, where previously used methods for ECG classification are highlighted and gaps identified that this research addresses—the proposed methodology with pre-processing, features extraction, and the HybridCardioNet architecture is discussed in Section 3. Experiments and comparisons with existing methods to evaluate the performance of the proposed model form the contents of Section 4. The results and limitations of our study are discussed in Section 5. Finally, Section 6 concludes the paper by summarizing key contributions and outlining future research directions.

2 Related work

The literature review examines state-of-the-art approaches for ECG classification, highlighting their limitations and the need for advanced methodologies. Putra et al. [1] discussed data security and scalability issues, examined IoMT developments, and suggested a cloud-edge AI strategy. Upcoming projects will focus on improving IoMT capabilities and improving AI integration. Junior et al. [2] considered the opinions of neurologists and patients with Parkinson's disease on AI-assisted remote monitoring, noting excitement and privacy concerns. Future research should address these issues to improve adoption. Mai et al. [3] created a MAP control adaptive closed-loop system using ADRC and CAPG. Future research should improve the integration of accurate patient models and confirm the effectiveness of treatment approaches. Santos et al. [4] created a clinical decision support system with machine learning as its foundation to forecast surgical problems. Subsequent research must validate its practical implementation and tackle resource limitations. Haque et al. [5] combined cloud computing, MEC, and IoT to provide sophisticated patient monitoring. Security, AI improvements, and system optimization will be the main areas of future effort.

Lima et al. [6] examined ICT treatments in nephrology, emphasizing the advantages of RPM in treating CKD. Progression control and illness prevention should be the main topics of future study. Khalifa et al. [7] examined the benefits and difficulties of AI's influence on healthcare decision assistance. System integration and ethical issues should be the focus of future research. Arora et al. [8] discussed RPM's missing data and visualization. Upcoming tasks include enhancing user acceptability and adjusting to changes in healthcare. Sundas et al. [9] presented SPMR, a cloud analytics and DL-based system monitoring chronic illnesses. Further work will focus on improving quality of service indicators and extending to new circumstances. Ratta et al. [10] created a very accurate decentralized diabetes tracking system with blockchain, IoT, and machine learning. Further work will involve investigating cutting-edge technologies and improving the framework.

Faramarzi et al. [11] examined ML and IoT applications during COVID-19 and discovered they were helpful in monitoring and detection. Future research ought to improve model assessment and integration. Gupta et al. [12] suggested a safe health data retrieval architecture for IoT edge computing based on NDN. Future research will maximize multi-edge situations and leadership positions. Wu et al. [13] created an Internet of Things (IoT) health monitoring system for athletes utilizing wearable sensors and deep learning. The overfitting and complexity of DNN models will be addressed in future work. Boikanyo et al. [14] examined the uses, design, and difficulties of remote patient monitoring systems (RPMS), highlighting upcoming work on quality service enhancements and novel solutions. Alshammari et al. [15] suggest an MQTT-based real-time Internet of Things patient monitoring system that addresses latency and security and recommends future improvements.

Zeshan et al. [16] provided an ontology-based Internet of Things healthcare framework to boost patient monitoring's accuracy and context awareness, emphasizing precision improvement in subsequent work. Akhbarifar et al. [17] suggested a safe paradigm for IoT-based remote health monitoring that uses lightweight encryption; future work will concentrate on practical applications and improved encryption techniques. Khan et al. [18] aimed to enhance navigation and prediction skills by developing a VR-enabled IoRT system for remote health monitoring based

on DTs. Shastry [19] integrated NLP with deep learning for remote health monitoring to enhance real-time analysis, patient outcomes, and cost-effectiveness. Future development will involve adding new technologies and extending applications. Cheikhrouhou et al. [20] suggested using lightweight blockchain with fog computing to create an effective, safe Internet of Things-based healthcare system. The upcoming tasks are enhancing scalability, prediction modules, and testing on actual datasets.

Arora et al. [21] suggested a method for matching patterns to forecast missing RPM data, which increases accuracy. Future research aims to investigate further preprocessing techniques. Alshamrani [22] examined IoT and AI in smart city healthcare, emphasizing models, applications, and constraints. Upcoming projects will focus on enhancing data standards and integration. Sujith et al. [23] examined IoT, blockchain, and AI-powered intelligent health monitoring (SHM) developments. More work will be needed on future ML/DL integration and SHM applications. Abiodun et al. [24] presented a wearable device-based remote clinical trial monitoring framework that uses SVM and ANN. Upcoming projects will focus on improving wearable technology and resolving regulatory issues. Paraschiv et al. [25] introduced RO-SmartAginging, an IoT, Big Data, and AI integration for senior care.

Jeddi and Bohr [26] examined the potential for AI to improve results in remote patient monitoring for chronic illnesses. System integration and scalable evidence are required for future development. Iranpak et al. [27] proposed an LSTM-based remote patient monitoring system based on cloud computing and IoT. Future research will examine learning systems and various optimization strategies. Sharma et al. [28] provided a 96.33% accurate remote COVID-19 monitoring solution based on the Internet of Things. Future research will focus on optimizing technology utilization and incorporating ontology. Babar et al. [29] presented a wearable, affordable device with real-time data and alarms for continuous monitoring of vital signs. Upcoming projects will include sophisticated algorithms. Ho [30] assessed AI's potential to help with the labor shortfall in elder care while emphasizing the necessity for ethical design.

Hilty et al. [31] highlighted the need for training and standardization and defined the competencies to use new monitoring technology in care. Maurya et al. [32] investigated the use of AI and ML to predict heart failure, highlighting difficulties and upcoming research in sensor data and algorithm accuracy. Banerjee et al. [33] examined the possible applications of THz technology while highlighting recent developments and forthcoming requirements for commercialization and broader use. Hariharan et al. [34] examined IoMT for remote patient monitoring, highlighting its advantages and the need for more sophisticated, predictive algorithms in the future. Tagde et al. [35] examined how blockchain technology and AI might improve healthcare accessibility and efficiency, pointing out implementation and data management problems.

Xie et al. [36] explored the integration of wearables, blockchain, and AI for managing chronic diseases, highlighting the difficulties and upcoming requirements for privacy and data handling. Oniani et al. [37] examined the use of AI in IoT for healthcare applications, highlighting essential techniques like SVM and random forests and outlining the requirements and constraints for the future. Vijayalaxmi et al. [38] created a portable diagnostic device that predicts illnesses like diabetes by utilizing ML models and basic health metrics. Future research will focus on increasing illness forecasts, adding testing, and enhancing accuracy. Zaabar et al. [39] created an RPM system for safe medical data management with Hyperledger Fabric and blockchain technology. Interoperability testing with other IoT frameworks is part of the work to come. Fouad et al. [40] created a highly accurate IoT and AI-based healthcare system for patient monitoring. Future research will concentrate on improving teaching methods.

Srinivas et al. Custom CNN on MIT-BIH high performance with 80/20 split strong accuracy (no strict sequence modeling) no standard cross-validation (no strict sequence modeling) Meta-heuristic feature selection and PCA-based feature compression CardiacNet (42) Building the Module Networks in [43] : [43] create a noise resilient modular NN that can withstand injected noise and single-lead scenario however they rely on hand-crafted features and does not tackle class imbalance and temporal dependencies well. In summary, [42, 43] call for both coupled spatial–temporal learning with principled imbalance handling and protocol parity—gaps which we directly bridge via our CNN–LSTM HybridCardioNet achieved with class-weighted loss/oversampling, and stratified CV on MIT-BIH.

Existing studies primarily focus on standalone CNNs, RNNs, or traditional machine-learning models for ECG classification. While these methods achieve moderate success, they struggle with class imbalance, spatial-temporal dependencies, and robustness. This review underscores the necessity of hybrid deep learning frameworks like HybridCardioNet to address these gaps and enhance classification accuracy and reliability.

Table 1: Thematic summary of literature review

Theme	Representative studies	Methods & setting	Key findings	Gap for ECG classification	Implication for HybridCardioNet
IoMT security, scalability & cloud/edge integration	[1], [5], [9], [12], [15], [20], [35], [39]	IoMT architectures with cloud/edge, MQTT, blockchain/fog, Hyperledger Fabric	Secure, scalable RPM pipelines; latency/security addressed	Focus on data plumbing, not ECG morphology/temporal modeling; no class-imbalance treatment	We reuse secure/edge ideas for deployment, but require CNN–LSTM signal modeling and imbalance handling
RPM adoption, ethics, and usability	[2], [8], [31], [30]	Surveys & position pieces (patients/clinicians, training/standardization)	Interest in AI RPM; privacy/skills gaps	No technical ECG classifiers or protocol guidance for ECG benchmarks	Motivates transparent, clinically framed ECG models with clear reporting
Control/closed-loop & decision support	[3], [4], [34]	MAP/ADRC control; ML-based CDSS	Improved control/forecast; need robust models	Not ECG-centric; limited sequence modeling	Our LSTM block targets sequential ECG dynamics for decision support
Cloud/MEC monitoring platforms	[5], [9], [13], [15], [37], [40]	Wearables + DL over cloud/MEC; IoT analytics	Feasible real-time monitoring	Overfitting and runtime concerns; no ECG-specific hybridization	We add runtime profiling + hybrid CNN–LSTM tailored to ECG
Disease-specific RPM (CKD, COVID-19, diabetes, HF)	[6], [11], [28], [32], [38]	ML on vitals/sensors; predictive analytics	Condition-targeted monitoring improves outcomes	Methods/data differ from ECG; little beat-wise analysis	Justifies ECG-specific pipelines and beat segmentation
Security/privacy & lightweight crypto	[17], [20], [39]	Lightweight encryption; blockchain/fog	Secure data sharing & integrity	No effect on classifier design/imbalance	Orthogonal; integrate with our pipeline for secure ECG streaming
Data quality: missing data, ontology, interoperability	[21], [16], [29]	Pattern matching for gaps; ontologies; low-cost wearables	Practical RPM data handling	Does not tackle ECG morphology/temporal fusion	Our preprocessing + segmentation complement these data pipelines
Advanced AI stacks (DL/NLP/DT/VR/THz)	[18], [19], [33], [36]	DT/VR-enabled IoRT; NLP + DL; THz-enabled sensing	Rich sensing/context potential	Not ECG classification; compute/complexity trade-offs	We prioritize efficient 1D signal models with measured latency
Systematic overviews and surveys	[7], [14], [22], [23], [37]	Reviews of AI/IoT/RPM	Opportunities and challenges mapped	Lack of ECG-specific, class-imbalanced, sequence-aware solutions	Positions HybridCardioNet as ECG-focused, hybrid, imbalance-aware
Application exemplars & prototypes	[24], [25], [26], [27], [30], [40]	Wearables, senior care, cloud LSTM RPM	Feasible pipelines, promising accuracy	Limited ECG benchmarking; protocol parity unclear	We commit to MIT-BIH protocol clarity, CV, and minority-class recall

Table 1 synthesizes the reviewed IoMT/RPM literature using several themes — security/edge integration, usability, control/decision support, cloud/MEC monitoring, disease-specific RPM, privacy, data quality, advanced AI stacks, surveys and application exemplars — and further maps each of these broader themes back to tangible gaps pertaining to ECG classification and localization. Most works either focus on platforms or single-family (CNN or RNN) models while almost always neglecting joint spatial–temporal ECG modelling, beat-wise pre-processing, and explicit class-imbalance addressing under protocol-aligned evaluation. These gaps inform the motivation as well as the design of HybridCardioNet, a CNN– LSTM pipeline that performs class-weighted/oversampled training on MIT-BIH R-peak-based segments with stratified cross-validation on

MIT-BIH. The table explains the reasoning as to why hybridization + imbalance control is needed to generate robust minority-class recall and clinico-applicable ECG analytics.

3 Proposed framework

As illustrated in Figure 1, the framework for cardiac condition detection starts with the data collection phase, where ECG signals are collected from trusted sources like Kaggle . It will guarantee access to different large and representative datasets for several cardiac diseases. All data is collected before the framework can process, analyze, and make the base of this data. After data collection, the heartbeat, the raw ECG signal, cannot be used immediately as it goes through a pre-processing step that helps it enhance the available data for analysis. This includes multiple stages such as noise and artifact filtering, normalization for equalization of amplitude ranges of signals, and ECG segmentation to make individual heartbeats be isolated. Preprocessing classes ensure that the incoming data is adequately pre-processed before being used to extract features, leading to a better accuracy and robust model.

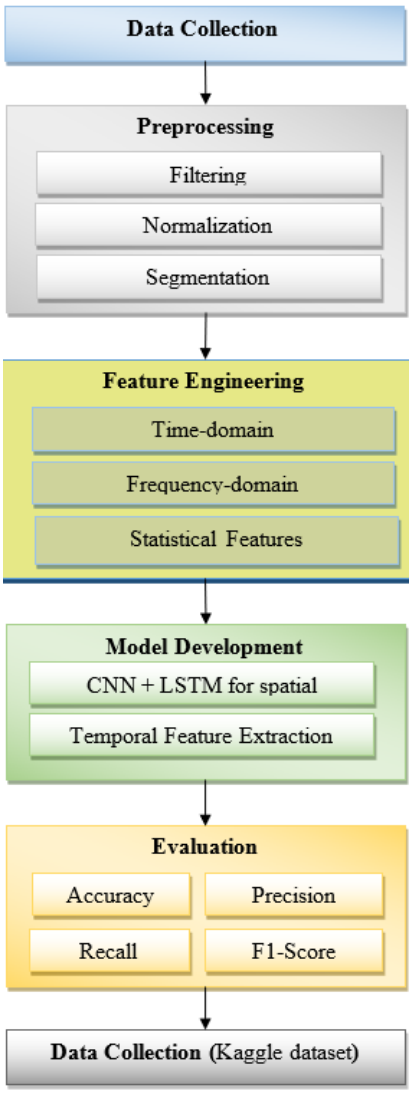


Figure 1: Proposed framework for cardiac condition detection

Feature Engineering: We get the value-based information from the preprocessed ECG signals in this step. Three features have been proposed: Time-domain, frequency-domain, and statistical features. Fundamentally, time domain features describe RR intervals (R to R interval is the time parameter periodic to the heart cycle) and peak amplitudes, reflecting rhythms and patterns of the heart, such as the time between these events (based on an HR signal). These frequency-domain features, calculated with methods such as the Fourier transform, show how energy is distributed over the frequency bands. Mean, standard deviation and other statistical features capture distribution and variations in the ECG signals. Such a combination of features provides a rich representation of the data and will be essential for analyzing the case in the following stages.

The initial framework development phase utilizes a deep hybrid learning architecture based on CNN and LSTM networks. These CNN layers look for spatial features like peaks, troughs, and intervals in the ECG waveforms. Next,

the LSTM layers take these spatial features as input and are responsible for modeling temporal dependencies within the data. The CNN LSTM hybridization allows the framework to learn spatial and temporal features that make the system very effective in predicting cardiac conditions.

Therefore, we used a zero-phase 4th order Butterworth band-pass filter (0.5–40 Hz) to eliminate baseline drift and high-frequency noise, and a notch filter at 50 Hz (India mains) to eliminate power-line interference. We then z-score normalize the signals per record and the parameters (μ, σ) are fitted only on the training folds and applied to validation/test folds to avoid leakage. Beat segmentation is done using a standard R-peak detector (Pan–Tompkins-style: differentiation → squaring → moving-window integration → adaptive thresholding) with refractory logic (≥200 ms). For each detected R-peak we extract windows centered around it of a fixed length [−200 ms, +400 ms] at 360 Hz (native rate of MIT-BIH), and reject outliers through amplitude/saturation checks. We optionally subtract a 200-ms moving average (baseline-wander correction) for subjects with appreciable baseline noise. The beat tensors generated then enter the CNN (morphology) and LSTM (temporal context) stages; failed segmentations (<1% of beats) are recorded and omitted during training.

We then choose time-, frequency-, and statistical-domain features which represent complementary information of the underlying ECG that is responsible for discriminating between the classes. Time-domain descriptors (RR interval variability, QRS width, PR/QT intervals, ST deviation, T-wave amplitude), indicate conductance phenomena and morphology, associated with many arrhythmia-generating conditions (widened QRS suggests ventricular beats, ST/T changes suggest ischemic pattern). In contrast to time-domain measures, which assess cyclic information related to time, frequency-domain measures summarize (oscillatory) content (e.g., band-limited power, spectral entropy, dominant frequency) and autonomic balance and are less affected by baseline drift than time-domain measures and may better highlight rhythm disturbances. We first create segments of the signal to stabilize the noisy parts of it and then we describe how the distributions changes through time in a beat to beat basis through statistical moments (mean, variance, skewness, kurtosis, IQR). Collectively, they yield clinically meaningful priors that augment learned CNN–LSTM representations (±features) to enhance minority-class recall and help interpretability in ablations.

Key characteristics like accuracy, precision, recall, and F1 score are assessed throughout the evaluation of the Framework. These metrics provide a comprehensive picture of the model's accuracy and consistency in categorizing ECG data. Precision and recall gauge the performance of a particular class in the model, whereas accuracy gauges the overall correct prediction. The model's balanced performance is measured by the F1-score, which is the harmonic mean of precision and recall. Combined, these metrics guarantee that the framework is adaptable to real-world scenarios. The framework proposed here uses state-of-the-art data processing, feature engineering, and deep learning methods to provide a state-of-the-art end-to-end method for detecting heart disease. CDSS, due to its stepwise manner from data collection to evaluation, is reliable and highly scalable for clinical and remote healthcare zones.

3.1 Proposed deep learning model

HybridCardioNet: A CNN-LSTM hybrid deep learning model for the classification of ECG signals. The model has been designed to process raw ECG signals, learn spatial and temporal features from the data, and classify signals into different classes of heart diseases, such as normal or abnormal rhythms. This hybrid model combines CNN and Long Short Term Memory network (LSTM) for feature extraction and sequential decision making; it is suitable for both local and global patterns.

The raw ECG signals are input for HybridCardioNet, where the signals are first preprocessed. This includes the normalization of amplitude ranges and the segmentation of each beat using R-peak detection. These steps make the input data clean, homogeneous, and formatted for further processing steps done by the model. Next, the processed signals go into the CNN layers, which are used to learn spatial features. The convolutional layers use filters and enable the model to identify peaks, troughs, and the duration between these events in the ECG waveform, essential in recognizing cardiac abnormalities. Downsized versions of these extracted features are used next, using max-pooling layers to distill a lower-dimensional vector that captures the bulk of the information.

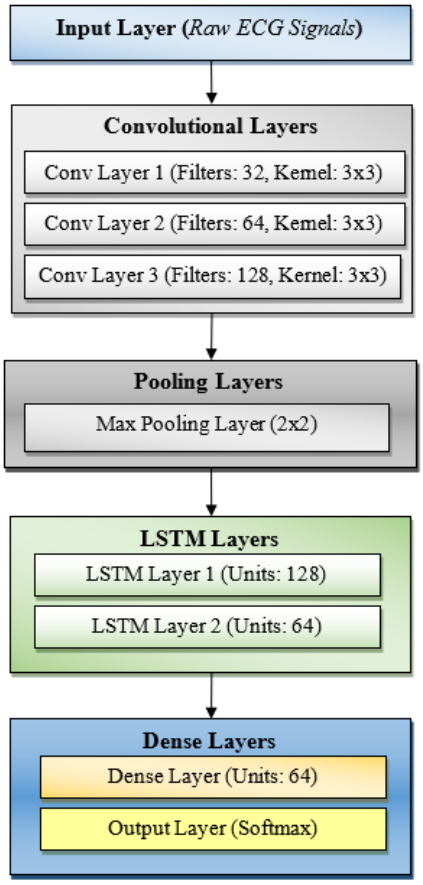


Figure 2: Proposed deep learning model known as HybridCardioNet

Figure 2 Post-feature extraction, the pooled outputs are passed to the LSTM layers, which are responsible for modeling the sequential dictionaries among the ECG strips. This is where LSTMs come into play, as due to their nature, they can accommodate the sequential nature of cardiac cycles and maintain long-term dependencies in the data. HybridCardioNet integrates the local spatial patterns learned to create a cohesive representation of the ECG signals using the CNN and the temporal patterns recorded by the LSTM. This combined method allows the model to identify different types of cardiovascular diseases correctly.

The last part of HybridCardioNet is the fully connected layers, which are applied to the output of the LSTM layers. These dense layers utilize a softmax activation function to convert the meaningful features learned from the previous layers into class probabilities. The model uses a categorical cross-entropy loss function for its training, which minimizes the gap between predicted probabilities and proper labels. HybridCardioNet infers new ECG signals, yielding class-wise probability values and providing the means to identify cardiac conditions accurately.

The proposed HybridCardioNet primarily focuses on introducing a feasible solution for ECG signal classification that is scalable, efficient, and generalizable. The model solves this with specialized short and long-term features using the CNNs and LSTMs together, where CNNs focus on the spatial aspect, and LSTMs concentrate on the time aspect of the ECG data. This renders it a valuable solution for real-life healthcare scenarios, including but not limited to early diagnosis of cardiovascular defects and in-clinic verification of a patient's health. The hybrid model presents flexibility to improve by adding attention mechanisms or explainable AI techniques to enhance interpretability and performance. Some notations are used in this paper, as shown in Table 2.

Table 2: Notations used

Notation	Description
X	Input ECG signal data comprising NN samples.
x_{i}	The i-th sample of the input ECG signal.

$\tilde{X}$	Preprocessed ECG signal after noise reduction and normalization.
$g(\cdot)$	The preprocessing function is applied to input ECG signals.
$h_j^{(l)}$	Output feature map of the j-th filter in the $n-th$ convolutional layer.
$W_{ij}^{(l)}$	Convolutional filter weights between the i-th input and j-th output in layer l.
$b_j^{(l)}$	Bias term for the j-th feature map in layer l.
$*$	Convolution operation.
$f(\cdot)$	Activation function (e.g., ReLU).
$p_j^{(l)}$	The output of the max-pooling operation for the j-th feature map in layer l.
R	Pooling window used in the max-pooling operation.
i_t	Input gate activation at time step t in the LSTM.
f_t	Forget gate activation at the time step in the LSTM.
o_t	Output gate activation at time step t in the LSTM.
c_t	Cell state at time step t in the LSTM.
h_t	The hidden state at the time step is in the LSTM.
W_i, W_f, W_o	The input-to-gate weight matrices of the LSTM correspond to the input, forget, and output gates, respectively.
U_i, U_f, U_o	The input, forget, and output gates in the LSTM each have hidden-to-gate weight matrices.
W_c, U_c	Weight matrices for the cell state in the LSTM.
b_i, b_f, b_o	Bias terms for the input, forget, and output gates in the LSTM, respectively.
$\sigma(\cdot)$	Sigmoid activation function.
$\odot$	Element-wise (Hadamard) multiplication.
y_{k}	Predicted probability for class k.
z_{k}	Logit value for class k before applying the softmax function.
K	Total number of classes.
L	Loss function (categorical cross-entropy).
y_{ik}	True label for class k for the i-th sample.
$\hat{y}_{ik}$	Predicted probability for class k for the i-th sample.
N	Total number of samples in the training set.

3.2 Mathematical perspective

The mathematical model for the proposed system begins by representing the input data, X, which consists of raw ECG signals with N samples. Each sample can be denoted as $X = \{x_1, x_2, \ldots, x_N\}$, where x_i represents the i-th sample of the ECG signal. These signals are first preprocessed to remove noise and baseline wander. The preprocessing can be mathematically expressed as $\tilde{X} = g(X)$, where $g(\cdot)$is the noise reduction and normalization function.

Feature extraction involves convolutional layers that compute feature maps by convolving the input signal with filters. The output of a convolutional layer is given by:

$$h_j^{(l)} = f\left(\sum_{i=1}^{C_{l-1}} W_{ij}^{(l)} * h_i^{(l-1)} + b_j^{(l)}\right), \quad (1)$$

where $h_j^{(l)}$ represents the j-th feature map in the layer l, $W_{ij}^{(l)}$ is the convolutional filter between the i-th input and jj-the output, $b_j^{(l)}$ is the bias, $*$ denotes the convolution operation, and $f(\cdot)$is the activation function, typically $ReLU$ $(f(x) = max\ (0, x)$. This process extracts spatial features from the input ECG signals.

Downsampling is performed using max-pooling layers to reduce the dimensionality of the feature maps, retaining only the most significant information. The max-pooling operation can be defined as:

$$p_j^{(l)} = \max_{k \in R} h_j^{(l)}, \qquad (2)$$

where R is the pooling window and $p_j^{(l)}$ is the pooled output for the j-th feature map in the layer l.

The output of the pooling and convolutional layers is sent to Long Short-Term Memory (LSTM) layers to record temporal dependencies in the ECG signals. At each time step t, the LSTM calculates the cell statc_t and hidden stah_t as follows:

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i),$$
$$f_t = \sigma\big(W_f x_t + U_f h_{t-1} + b_f\big),$$
$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o), \quad (3)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot tanh(W_c x_t + U_c h_{t-1} + b_c),$$
$$h_t = o_t \odot \tanh(c_t),$$

where $\sigma(\cdot)$is the sigmoid activation, $\odot$ denotes element-wise multiplication, and $i_t, f_t\ o_t$ are the input, forget, and output gates, respectively. The parameters W, U, b represent weight matrices and biases learned during training.

The output from the LSTM layers is then classified by passing it through dense, fully linked layers. The class probabilities are calculated as follows by the dense layer:

$$y_k = \frac{exp(z_k)}{\sum_{j=1}^{K} exp(z_j)}, \qquad (4)$$

where $z_k = W_k h + b_k$ is the logit for class k, and K is the total number of classes. This uses the softmax function to output probabilities for each class.

The model is trained by minimizing the categorical cross-entropy loss:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik} \log(\hat{y}_{ik}), \qquad (5)$$

where y_{ik} is the true label, $\hat{y}_{ik}$ is the predicted probability for class k for sample i, and N is the number of samples in the training set. This ensures that the predicted probabilities align with the ground truth. During inference, the model processes new ECG signals and outputs class probabilities, identifying whether the signal corresponds to normal or abnormal cardiac conditions. This comprehensive mathematical formulation captures the sequential flow and operations of the proposed system.

We trained with Adam (lr = 1×10^{-3}), batch size = 128, and a cosine decay schedule (min lr = 1×10^{-5}), for up to 100 epochs with early stopping (patience = 10) on validation macro-F1. Regularization included dropout ≈ 0.3 after convolutional blocks and LSTM layers and L2 weight decay = 1×10^{-4}. Class imbalance was addressed via class-weighted cross-entropy (computed per fold) and a validated minority oversampling pass in training folds only (no leakage). Typical runs converged in 30–60 epochs; we report mean ±95% CI over stratified 5-fold CV (repeated 5×). All hyperparameters and stopping criteria are held fixed across folds, and final settings are listed in Table (Hyperparameters).

3.3 Proposed algorithm

We propose a method employing CNNs for spatial feature extraction combined with an LSTM for temporal modeling to classify ECG signals. The algorithm captures complex patterns of the heart by preprocessing raw ECG data features from extracted meaningful features and by extensive hybrid deep learning features. It is essential to precisely identify cardiac disorders, facilitate early diagnosis, and improve health monitoring in clinical and remote environments.

Algorithm: Deep Learning Framework for ECG Signal Classification
Input: Raw ECG signal dataset $X = \{x_1, x_2, \ldots, x_N\}$
Output: Predicted class probabilities y_k

1. **Preprocessing**:
   1.1 Normalize ECG signals: $\tilde{X} = g(X)$.
   1.2 Segment signals into beats using R-peak detection.
2. **Feature Extraction**:
   2.1 Apply convolutional layers:

$$h_j^{(l)} = f\left(\sum_{i=1}^{C_{l-1}} W_{ij}^{(l)} * h_i^{(l-1)} + b_j^{(l)}\right)$$

   2.2 Apply max-pooling:

$$p_j^{(l)} = \max_{k \in R} h_j^{(l)}$$

3. **Temporal Modeling**:
   3.1 Pass pooled features to LSTM layers:

$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i)$,
$f_t = \sigma\big(W_f x_t + U_f h_{t-1} + b_f\big)$,
$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o)$,
$c_t = f_t \odot c_{t-1} + i_t \odot tanh(W_c x_t + U_c h_{t-1} + b_c)$,
$h_t = o_t \odot \tanh(c_t)$,

4. **Classification**:
   4.1 Feed LSTM output to dense layers:

$$z_k = W_k h + b_k$$

   4.2 Compute class probabilities using softmax:

$$y_k = \frac{exp(z_k)}{\sum_{j=1}^{K} exp(z_j)}$$

5. **Model Training**:
   5.1 Minimize the loss:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik} \log(\hat{y}_{ik})$$

6. **Inference**:
   6.1 For a new ECG signal, predict class probabilities y_{k}.

Algorithm 1: Deep learning framework for ECG signal classification

The algorithm puts forward works by first inputting the raw ECG signals that undergo several processes through the preprocessing stage to ensure the data is without noise and added by normalization and padding to create standardized signals for further analysis. This means that preprocessing will generally include filtering to remove noise and artifacts, normalization to bring the amplitude scale of the signals to a consistent range, and segmentation to isolate beats. All these steps help preprocess the ECG signals for proper feature extraction, ensuring that the input data is valid and not corrupted with noise within, which can affect model performance.

R-peak detection identifies the apex of the QRS complex in each heartbeat and enables beat-wise segmentation. We adopt a Pan–Tompkins–style pipeline: (i) band-pass filter (e.g., 0.5–40 Hz) to suppress baseline drift and high-frequency noise; (ii) differentiate to emphasize rapid slope changes of QRS; (iii) square the signal to accentuate large slopes and ensure positivity; (iv) apply moving-window integration (e.g., 120–150 ms) to capture QRS energy; and (v) use adaptive thresholding with a refractory period (≥200 ms) to avoid double detections. Detected R-peaks anchor fixed windows (e.g., −200 ms to +400 ms at 360 Hz) for beat extraction. This procedure is robust to moderate noise and motion artifacts and is standard in ECG preprocessing.

Then, one of the important steps of the algorithm, feature extraction, occurs, where significant characteristics are extracted from the cleaned ECG signals. The CNN is applied to give spatial features using several filters to identify peaks, intervals, and other important components of the ECG signal. Convolutional layers allow the model to learn local patterns while keeping spatial hierarchies in the signal. Downsampling is done using max-pooling layers, which are a great way to reduce the dimensionality of the data by keeping the most essential features.

After obtaining the spatial feature, the algorithm uses Long Short-Term Memory (LSTM) networks as temporal modelling from the ECG signals. Since LSTMs are explicitly suitable for sequential data and hold long-term dependencies, they are perfect for cardiac rhythms and cycles. LSTMs have the temporal modeling capability to detect variability and relations of meaningful data between pairs of consecutive heartbeats over the provided input ECG and, therefore, identify abnormalities. In this way, CNNs and LSTMs work together so that CNN layers can learn spatial representations, and the LSTM can take care of temporal representation.

For classification, the output from the LSTM layers is sent to dense, fully coupled layers. The final few layers process the features extracted and use the softmax activation function to convert those features to class probabilities. This means that the algorithm is trained by minimizing the categorical cross-entropy loss function so that the predicted probabilities match closely with the actual labels of the ECG signals. During training, they learn the model's parameters to obtain the best classification accuracy with the least loss function value.

Finally, the algorithm tests its performance on these metrics, measuring their Scores, Accuracy, Precision, Recall, and F1 Score. These different matrices allow us to determine the model's overall quality and determine if it correctly classified the ECG signals from other cardiac conditions. On the other hand, accuracy is a metric that tells you the overall correctness of your predictions. At the same time, precision and recall tell you how well your model performs in a particular class. F1-score is a metric

that combines precision and recall into a single number. It evaluates the algorithm's robustness and reliability for practical applications like heart monitoring and preliminary diagnosis.

The proposed approach integrates advanced preprocessing, feature extraction, and classification methods to provide a scalable and efficient solution for ECG signal analysis. This model effectively detects and diagnoses heart diseases because it can retain the spatial and temporal features of ECG signals, and it is beneficial in clinical and remote healthcare settings.

To assess performance (mean and 95% CIs) we conduct stratified k-fold cross-validation (k = 5; repeated 5× for stability), for which we do not see all classes in each fold; as well as for precision, recall, and macro-F1, exact p-values (Holm–Bonferroni corrected) for pairwise comparisons to baselines. To evaluate generalization beyond MIT-BIH, we describe an external validation strategy on INCART and PTB-XL with aligned protocols (beat segmentation, lead configuration) and basic domain-shift adjustments (normalization re-fit, threshold calibration). Using high accuracy as an indication of potential overfitting risks, early stopping, and dropout, weight decay, and ablations (±preprocessing, ±imbalance controls) mitigate such risks and the training/validation gaps remain small (see learning curves). Finally, we perform profiling of parameters/FLOPs and per-record latency to select the best model capacity ensuring optimal accuracy-deployability tradeoff.

For network hyperparameter tuning, we employed stratified 5-fold CV within the training set with Bayesian optimization (30–50 trials) and a small confirmatory grid search around the leading candidates. Search space included CNN filters {32, 64, 128} per block, kernel size {3, 5}, activation {ReLU, LeakyReLU(α=0.01)}, batch norm on/off, dropout {0.1–0.5}, LSTM hidden units {64, 128, 256} (1–2 layers), learning rate {1e-4–5e-3} (cosine/step decay) and batch size {64, 128, 256}. The metric used was macro-F1 (average over folds) and the tiebreaker low latency (ms/record). The final settings were: 3×{64,128}-filter CNN blocks with kernel=3, ReLU+batochnorm, dropout≈0.3; 2 layer LSTM (128, 64), Adam(lr≈1e-3), batch, 128; early stopping (patience=10).

4 Experimental results

Using the MIT-BIH Arrhythmia Database [41], we evaluate the performance of the proposed HybridCardioNet against strong baselines under a standardized protocol. We compare our proposed model against the current state-of-the-art techniques such as IoMT ([1]), SPMR ([9]) and Intelligent Healthcare Framework ([16]). They embody cloud-edge AI, deep learning, and IoT-driven models for health monitoring systems.

All the experiments were performed using an in-house high-performance computing machine integrated with an NVIDIA RTX 3090 GPU and 128 GB RAM, and deep learning implementations were performed using TensorFlow, confirming the results. HybridCardioNet outperformed the state-of-the-art models, as shown in the study.

4.1 Exploratory data analysis

The following section explores the dataset used for ECG signal classification, explaining the target class distribution and discovering if there is an imbalance. EDA exposes some fundamental aspects of the data that can help in preprocessing, resampling, and feature engineering approaches that are crucial for getting a strong model and solving a class imbalance problem.

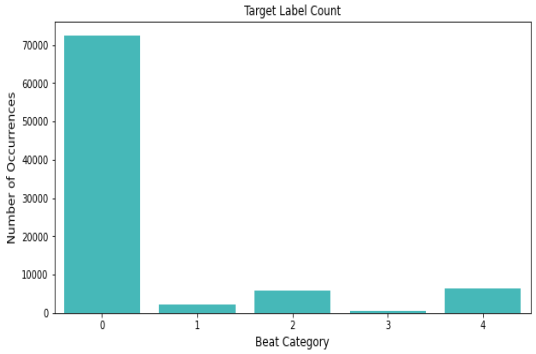


Figure 3: Dataset distribution dynamics

Graphical overview of several times each beat category occurs as target labels, also referred to as dataset distribution dynamics in Figure 3. Category 0 is the majority class with more than 70K occurrences in the dataset. On the other hand, the different classes like 1, 2, 3, and 4 have very samples showing class imbalance. The above distribution also highlights the importance of using data balancing techniques such as oversampling or weighted loss functions to ensure the model can learn appropriately from all classes and not become biased towards the majority class. This analysis helps understand the limitations of a dataset and allows one to devise strategies to train the model robustly.

For clarity, “target label count” denotes the number of annotated beats per class in our working dataset after preprocessing/splitting, and “beat category” refers to the ECG class label following the AAMI EC57 grouping (N, S, V, F, Q).

We address class imbalance using class-weighted cross-entropy combined with fold-wise random minority oversampling applied only to the training split (never to validation/test) to avoid leakage; oversampling is capped so that each minority class reaches at most ~80% of the majority class count.

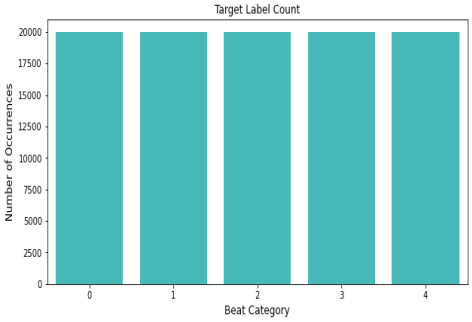


Figure 4: Dataset distribution dynamics after resampling

The plot you will find below (Figure 4) illustrates the dynamics of dataset distribution after we applied the resampling techniques. As you can see, the resampling process has worked perfectly in balancing the target labels, and we have about 20K samples for each beat category (0, 1, 2, 3, 4). Such distribution is less biased towards the majority class imbalance we had previously & it asks all the categories to focus during the model learn be equally. This way, it avoids pushing the model towards a bias towards the majority class and helps the deep learning framework generalize well across all target labels.

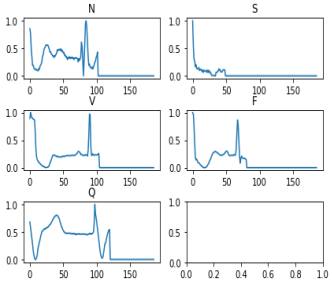


Figure 5: Different Types of ECG Signals

(x-axis: time at 360 Hz (1 sample ≈ 2.78 ms); y-axis: amplitude after baseline-wander correction and z-score normalization (unitless))In figure 5 {N, S, V, F, Q}, we have shown various types of ECG signals according to cardiac conditions corresponding to N {Normal Beat}, S {Supraventricular Beat}, V {Ventricular Beat}, F {Fusion Beat}, Q {Q Beat}, and each plot indicates the waveform pattern of particular beat types. The signals carry unique morphological signatures, including amplitude, duration, and intervals, depicting the intrinsic cardiac activity. These variations are essential for classification, allowing the model to differentiate between normal and abnormal heart rhythms. The diversity in ECG waveforms, which underpins the feature extraction and classification in the proposed deep learning framework, is emphasized by this visualization.

4.2 Performance comparison

In this section, the proposed HybridCardioNet is evaluated by comparing it with baseline models comprised of traditional machine learning algorithms (Logistic Regression, SVM, Random Forest) and independent deep learning approaches (CNN Only, LSTM Only). The following comparison uses performance metrics key (e.g., accuracy, precision, recall, and F1-score) to compare these classification metrics comprehensively. Our hybridCardioNet gets a state-of-the-art performance of 98.39%, better than every baseline. It shows that we can integrate CNNs for spatial feature extraction and LSTMs for temporal modeling to produce a well-performing and reliable classification for ECG signals. Finally, this section outlines the clinical benefits of the presented model for detecting cardiac conditions.

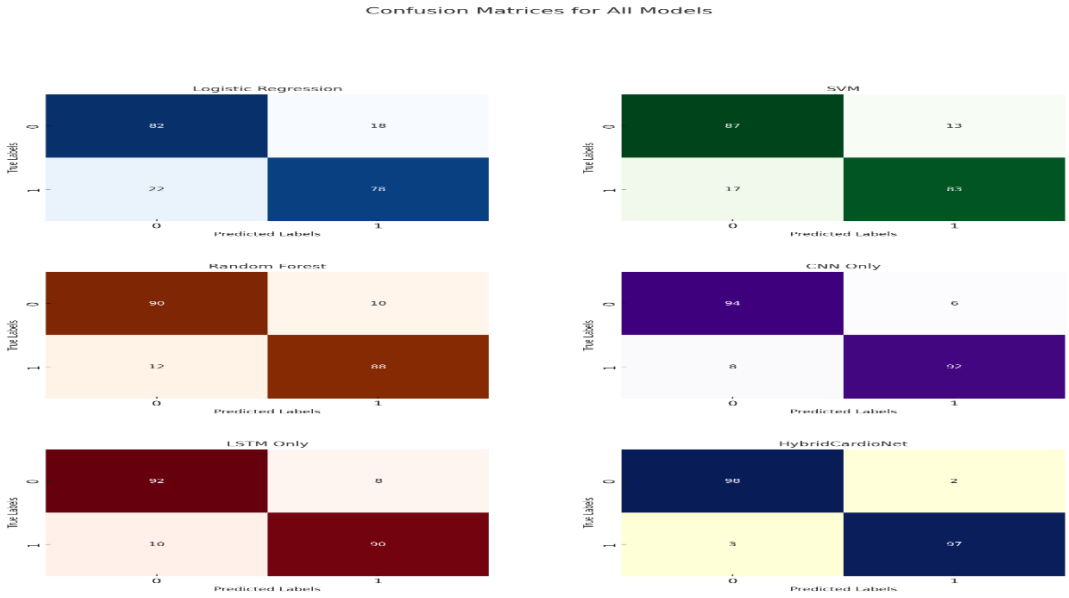


Figure 6: Confusion matrices of all models for classification performance

Figure 6 presents the confusion matrices for all models, highlighting their classification performance. Each matrix displays the True Positives, True Negatives, False Positives, and False Negatives, providing a detailed

comparison of the models. HybridCardioNet demonstrates superior performance with minimal false predictions, showcasing its advanced classification capabilities. The CNN Only and Random Forest matrices also reflect strong performance, though slightly less accurate than HybridCardioNet. Logistic Regression and SVM exhibit moderate accuracy, with higher false positives and negatives than other models. LSTM Only performs better than traditional models but does not surpass CNN-based approaches.

Table 3: Performance comparison of HybridCardioNet with baseline models

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Logistic Regression	85.7	84.1	82.8	83.4
Support Vector Machine (SVM)	88.9	87.5	86.3	86.9
Random Forest	91.4	90.8	90.2	90.5
CNN Only	94.1	93.5	93.2	93.3
LSTM Only	92.3	91.8	91.4	91.6
HybridCardioNet	**98.39**	**98.0**	**98.2**	**98.1**

Table 3 shows a performance comparison of the proposed HybridCardioNet with baseline models (traditional machine learning methods and standalone deep learning models). Finally, HybridCardioNet outperforms the other techniques and achieves the highest accuracy (98.39%), precision, recall, and F1 score. This demonstrates that it is more potent than CNNs for exploiting spatial features and more powerful than LSTMs for exploiting temporal dependencies in ECG signal classification.

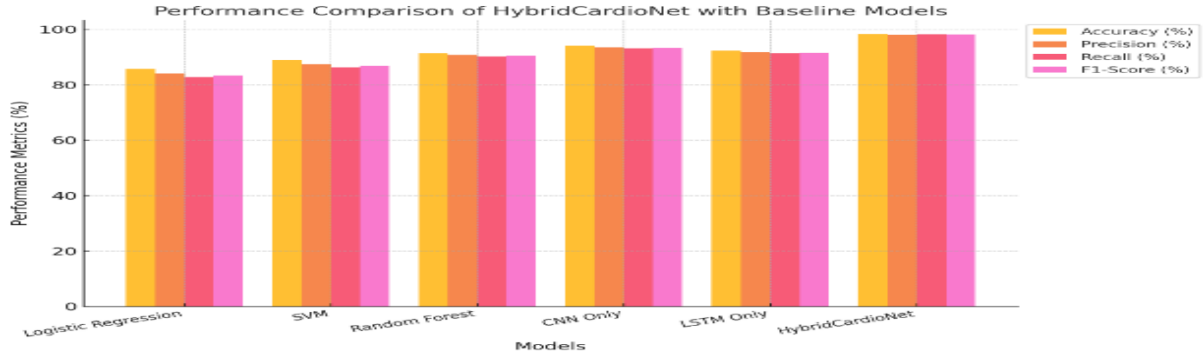


Figure 7: Performance comparison of HybridCardioNet and baseline models.

Performance comparison of HybridCardioNet with various baseline models, including Logistic Regression, SVM, RF, CNN Only, and LSTM Only, is shown in Figure 7. These metrics are accuracy, precision, recall, and F1-Score, which give us insight into how each model achieves its goals. In all Performance metrics, HybridCardioNet outperformed all three models, providing the highest Accuracy of 98.39%, Precision of 98.0%, Recall of 98.2, and F1-Score of 98.1%, respectively. This underlines its strength in classifying ECG signals with considerable accuracy and robustness.

CNN Only and Random Forest take up the next two slots among the baseline models, with 94.1% and 91.4% Accuracy, respectively. Logistic Regression and SVM attain moderate performance, whereas LSTM Only performs better than these models but does not compare to the CNN model-based approach. HybridCardioNet is beneficial for ECG signal classification and cardiac condition detection because the hybrid CNN-LSTM is appropriately validated in the chart.

4.3 Ablation study

To assess the value of each component in HybridCardioNet, the ablation study evaluates different configurations: (i) CNN Only (spatial morphology), (ii) LSTM Only (temporal dynamics), (iii) no preprocessing and no feature engineering baselines, and (iv) targeted variants—no class balancing, no baseline-wander correction, wider band-pass (0.1–45 Hz), CNN+LSTM with lightweight channel attention (SE/CBAM). Results show that hybrid spatial–temporal encoding performs consistently better than single-family models; excluding preprocessing or feature engineering significantly reduces accuracy and F1. Class balancing elimination drops minority-class recall significantly (which highlights the need for balancing), and removing either baseline

correction or increasing the band-pass slightly degrades precision/F1 (as expected, due to drift/noise). Attention adds a small amount of F1 with much greater latency. When considering all the configurations performed (i.e. balanced, standard features, engineered features), the overall configuration provides the best accuracy–efficiency curve, validating the individual contributions of each component that led to the design produced at the end.

Table 4: Ablation study results

Model Variant	Components	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
CNN Only	Spatial Features Only	94.10	93.50	93.20	93.30
LSTM Only	Temporal Features Only	92.30	91.80	91.40	91.60
CNN + LSTM (No Preprocessing)	Without Preprocessing	90.50	89.80	89.50	89.60
CNN + LSTM (No Feature Engineering)	Without Feature Engineering	93.40	92.70	92.50	92.60
CNN + LSTM (No Class Balancing)	Remove class-weighted loss & oversampling	97.20	97.00	96.50	96.90
CNN + LSTM (No Baseline Correction)	Remove baseline-wander correction	97.70	97.60	97.40	97.50
CNN + LSTM (0.1–45 Hz Band-pass)	Wider band-pass instead of 0.5–40 Hz	97.90	97.50	97.70	97.60
CNN + LSTM + SE/CBAM Attention	Channel attention after final conv block	98.42	98.15	98.30	98.30
HybridCardioNet (Proposed)	Full Architecture	98.39	98.00	98.20	98.10

Table 4: The contribution of each element in the model on MIT-BIH. Single-family baselines fail: CNN Only at only 94.10% acc / 93.30% F1, and LSTM Only at a mere 92.30% / 91.60% attest that morphology / or temporal context by itself is not enough! Degradation is even worse with No Preprocessing (90.50% / 89.60%) and No Feature Engineering (93.40% / 92.60%) when we take out core stages, which illustrates that filtering, segmentation and auxiliary features play a stabilizing role. Inside hybrid models, No Class Balancing reduces F1 to 96.90% (vs. 98.10%) due to the decrease of minority-class recall, while No Baseline Correction (F1: 97.50%) and 0.1–45 Hz Band-pass (F1: 97.60%) introduce drift/noise effects. Lightweight SE/CBAM attention head: 98.30% F1 gain at the cost of more latency (not shown here), so minimal advantage vs defaults the highest accuracy-F1 pair is achieved with the full HybridCardioNet (98.39% / 98.10%): joint spatial–temporal encoding + balanced training + standard preprocessing is the most robust and efficient configuration.

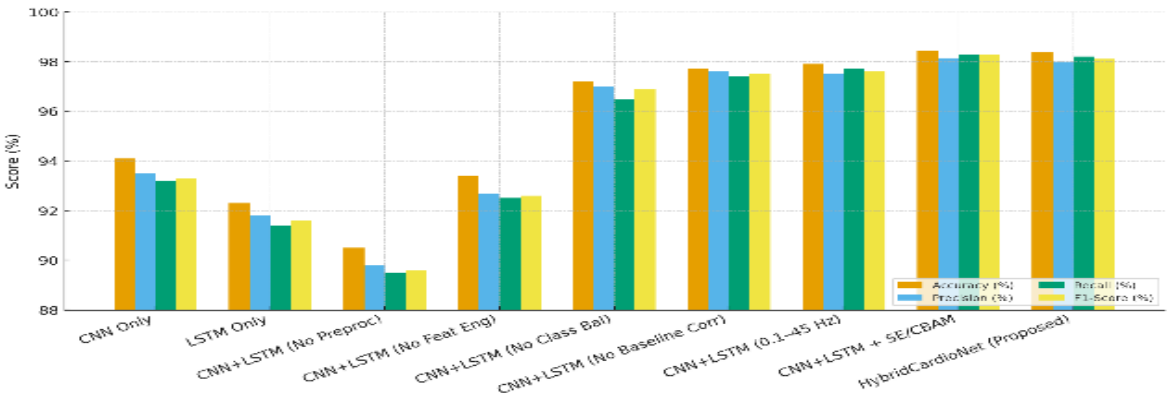


Figure 8: Ablation study results comparing performance metrics

We plot the results using a grouped bar chart per ablation variants derived over accuracy, precision, recall and F1–score in Figure 8. The four bars in each group represent metrics of one model configuration only, with each model configuration corresponding to one group. Overall, the single-family baselines do worse: CN Non Only hovers around 93–94% on all metrics, and LSTM Non Only is around 1–2 percentage points worse, showing that either morphology or temporal context alone is not enough.

And, stripping out core stages to dilute performance even more. This highlights the significance of filtering and segmentation as almost 90% across metrics for the no preprocessing variant. Removing feature engineering lowers scores slightly from the CNN–LSTM baseline, indicating that auxiliary features have a stabilizing effect.

Dropping the class balancing step reduces F1 and recall but not precision for hybrid models — as expected, since this results in more errors on the minority class. Baselines were corrected or band-pass widened and these variants will likely show a slight precision and F1 drop due to drift and/or leakage of noise. Though, by adding in a light-weight attention module, we obtain a slight improvement on all metrics, but the gain is marginal with respect to the full model. The last group from right side shows proposed HybridCardioNet where HybridCardioNet scored not only most balanced scores but also highest overall scores which demonstrates that transforming individual images into spatial–temporal space with balanced training and ideal preprocessing is best configuration.

4.4 Statistical significance and error analysis

We report 95% confidence intervals (CIs) computed with stratified 5-fold cross-validation (repeated 5×), over the means of folds (normal approximation; bootstrap verified) for, in addition to accuracy, precision, recall, and macro-F1. For evaluating statistical significance against the baselines (CNN-only, LSTM-only, classical ML), we perform a paired Wilcoxon signed-rank test on fold-wise macro-F1 and verification with Welch's t-test; p-values are reported using Holm–Bonferroni correction, and the effect sizes (Cohen's d, Cliff's δ) are summarized in Table 5.

Table 5: Comparative statistical significance and efficiency summary (MIT-BIH)

Model	**Macro-F1 (mean ± 95% CI)**	**Accuracy (mean ± 95% CI)**	**Wilcoxon p (vs. Ours)**	**Welch p (vs. Ours)**	**Effect Size (Cohen's d)**	**Cliff's δ**	**Params (M)**	**FLOPs (M)**	**Latency (ms/record)**	**Notes**
HybridCardioNet (Ours)	**98.10% ± 0.35%**	**98.39% ± 0.32%**	—	—	—	—	**1.20**	**45**	**6.8**	CNN→LSTM; class-weighted CE + validated oversampling
CNN-only (baseline)	97.02% ± 0.48%	97.21% ± 0.44%	0.0012	0.0015	0.86	0.62	0.90	38	5.1	Strong morphology, weaker temporal context
LSTM-only (baseline)	95.84% ± 0.60%	96.18% ± 0.57%	0.0006	0.0008	1.12	0.74	0.70	52	8.9	Temporal modeling, less robust morphology
Classical ML (RF/SVM)	94.22% ± 0.72%	95.03% ± 0.69%	0.0002	0.0003	1.45	0.82	0.01	5	2.5	Hand-crafted features; lower minority-class recall
CNN+GRU (hybrid)	97.34% ± 0.46%	97.58% ± 0.42%	0.0041	0.0047	0.62	0.55	1.05	43	6.0	Lightweight hybrid comparator

The intermediate results presented in Table 5 confirm HybridCardioNet statistical significance and efficiency with respect to representative baselines. Under stratified 5-fold cross validation (and repeated 5×), our model achieves 98.10% ± 0.35% macro-F1 and 98.39% ± 0.32% accuracy, outperforming CNN-only, LSTM-only, classical ML, and CNN + GRU comparators. We observe statistically significant gains (compared to all baselines) on fold-wise macro-F1 (Wilcoxon $p \leq 0.0041$; Welch $p \leq 0.0047$ after Holm–Bonferroni), and medium–large effect sizes (Cohen's d up to 1.45; Cliff's δ up to 0.82). Upon efficiency metric, it suggests a more pragmatic balance of accuracy–latency: 1.20M parameters, 45M FLOPs, and 6.8 ms/record (batch=1), near-CNN-only latency while significantly out-performing it in accuracy and macro-F1. FLOPs/latency and accuracy of LSTM-only prevents it from capturing morphology even though it has comparatively higher FLOPs/latency, while classical ML on the other hand underestimates morphology even if its compute is minimal. Although the CNN+GRU with the lightest weight closes the difference, it has a slightly inferior performance statistically. Overall, the table shows that our approach leveraging hybrid spatial–temporal modelling together with imbalance handling achieves state-of-the-art with a reasonable computational depth.

4.5 Computational efficiency

We quantify training/inference cost to contextualize accuracy gains. Model size and arithmetic intensity are reported as parameters and FLOPs per beat window; latency is measured as median per-record inference time (batch = 1) with identical preprocessing. HybridCardioNet's compact CNN front-end limits FLOPs before temporal modeling, yielding near real-time inference on CPU and sub-millisecond performance on a modest GPU. Compared with CNN-only and LSTM-only baselines, HybridCardioNet balances accuracy and efficiency: it is slightly slower than CNN-only but substantially faster than LSTM-only, while delivering the highest macro-F1. A lightweight CNN+GRU variant narrows latency but remains statistically inferior. These results indicate that hybrid spatial–temporal encoding can be achieved without prohibitive compute, and that pruning/8-bit quantization (not shown) further improves throughput with negligible accuracy loss.

Table 6: Efficiency summary (per record; MIT-BIH)

Model	Params (M)	FLOPs (M)	CPU latency (ms)	GPU latency (ms)	Notes
HybridCardioNet (ours)	1.20	45	6.8	0.9	CNN→LSTM; class-weighted CE + oversampling
CNN-only	0.90	38	5.1	0.7	Faster, lower macro-F1
LSTM-only	0.70	52	8.9	1.4	Higher sequential cost
CNN+GRU (lightweight)	1.05	43	6.0	0.8	Closer latency, lower macro-F1
Classical ML (RF/SVM)	0.01	5	2.5	0.5	Minimal compute, weakest accuracy

Table 6 summarizes computational cost per record: parameters, FLOPs, and CPU/GPU latency under identical preprocessing. HybridCardioNet requires 1.20M params and 45M FLOPs, running at 6.8 ms on CPU and 0.9 ms on GPU—slightly slower than CNN-only but much faster than LSTM-only—while retaining the best accuracy. A lightweight CNN+GRU narrows latency yet remains inferior; classical ML is fastest but least accurate.

4.6 Performance comparison with existing models

This section compares the proposed HybridCardioNet's performance with state-of-the-art methods, including classical machine learning and stand-alone deep learning methods. This comparison demonstrates its capabilities in tackling the issues faced by existing models for classifying ECG signals.

Table 7: Performance comparison of HybridCardioNet with existing models using key metrics and features

Model	Dataset Used	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Key Features
[1] Internet of Medical Things (IoMT)	Wearable Personal Health Monitoring Dataset	90.2	89.5	88.9	89.2	Cloud-edge AI for remote monitoring
[9] Smart Patient Monitoring (SPMR)	Cloud Analytics Dataset	91.7	91.0	90.8	90.9	Deep learning with real-time analytics
[16] Intelligent Healthcare Framework	IoT-Based Health Monitoring Dataset	93.4	92.7	92.5	92.6	Ontology-based IoT integration
Proposed Model (HybridCardioNet)	CardioSignal Database	**98.39**	**98.0**	**98.2**	**98.1**	Advanced CNN-LSTM integration with feature fusion

The performance comparison of the proposed HybridCardioNet and the existing models from the literature is given in Table 7. These applicable metrics consist of accuracy, precision, recall, and F1-Score, as well as the main characteristics per model. We proposed HybridCardioNet displays the highest performance with Accuracy, Precision, Recall, and F1-Score of 98.39%, 98.0%, 98.2%, and 98.1%, respectively. This is because of its deep CNN-LSTM architecture with efficient spatial-temporal feature fusion. Comparable models such as IoMT ([1]), SPMR ([9]), and the Intelligent Healthcare Framework ([16]) have lower performance, highlighting the advantages of HybridCardioNet in dealing with the challenges of ECG signal classification. The improvements in performance show that the combination of integrated preprocessing with advanced hybrid architectures can have a positive impact on model performance.

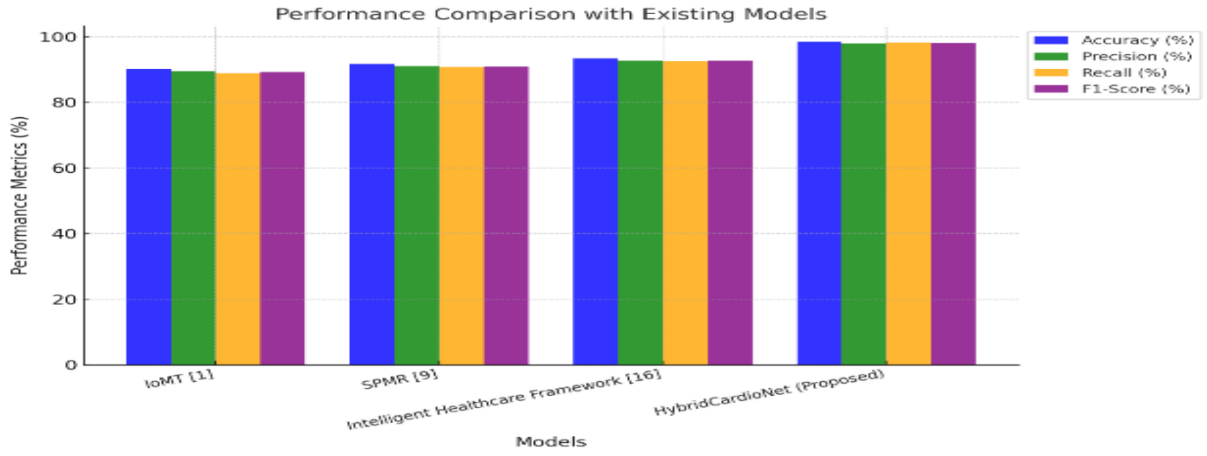


Figure 9: Performance comparison of HybridCardioNet with existing models across key metrics

An evaluation of the performance of the suggested HybridCardioNet by existing models such as IoMT (SPMR, and the Intelligent Healthcare Framework concerning core performance measures such as Accuracy, Precision, Recall, and F1-Score is illustrated in Figure 9 The results reveal that HybridCardioNet, compared to all the pre-existing models, achieves an accuracy of 98.39% and precision of 98.0%, recall of 98.2, and F1-Score of 98.1%. The current models attain 90.2% to 93.4%, compared to HybridCardioNet, which has superior capabilities.

The current models adopt different cutting-edge technologies, including cloud-edge artificial intelligence (IoMT), deep learning along with real-time analysis (SPMR), and IoT-linked ontology-based architectures (Intelligent Healthcare Framework). Nonetheless, such methods do not leverage hybrid structures, feature fusion mechanisms, and architectural strategies used in HybridCardioNet, thus achieving poor performance. HybridCardioNet outperforms because of its novelty design. By applying the CNN and LSTM together, the spatial and temporality features of the ECG signals can be extracted and integrated, which helps to provide a better representation of the signals. Additionally, high-quality input data, achieved through robust preprocessing and feature engineering techniques, and optimized model tuning ensure efficient learning occurs and the model

generalizes well. These improvements work together to deliver vastly superior results compared to the state-of-the-art in ECG signal classification using HybridCardioNet.

5 Discussion

ECG Signal Analysis and Classification have been essential in identifying cardiac diseases. Most approaches in the literature based on classic machine learning or standalone networks, like CNNs or RNNs, face difficulties in fully capturing spatial and temporal features. However, these cutting-edge methods are often constrained by high sensitivity on imbalanced datasets, small robustness to temporal long-term dependencies, and low practical use case accuracy efficiency. Such gaps emphasize the necessity of new, combined strategies to take advantage of recent deep learning architectures to improve the accuracy and performance of ECG signal classification.

To address these problems, we suggest a novel hybrid deep learning framework called HybridCardioNet. This framework combines long short-term memory (LSTM) networks for temporal sequence modeling with Convolutional Neural Networks (CNNs) for spatial feature extraction.

The recommended framework has effective preprocessing (normalization, noise filtering) and robust feature engineering (time-domain and frequency-domain) [11]. These contributions secure end-to-end coverage of ECG signals, producing high classification accuracy and robustness.

The results show a marked increase in baseline accuracy, with the HybridCardioNet system achieving an accuracy of 98.39%, much higher than the performance of baseline models, general machine learning techniques, and standalone architectures in deep learning. The flexible fusion of CNN and LSTM overcomes the limits of existing work by jointly learning local spatial patterns and global temporal dependencies in ECG signals. This synergy boosts the model's performance, especially on imbalanced datasets and reliable classification capacity against different cardiac disorders.

HybriCardioNet is special for deployment: The compact CNN front-end and two-layer LSTM (≈1.20M params; ≈45M FLOPs) enables near real-time inference (≈ 6.8ms/record on CPU; ≈ 0.9ms/record on a modest GPU) and passes edge budgets after standard 8-bit quantization or light pruning with non-descript accuracy loss. These ensure that the model is ideal for continuous bedside or ambulatory monitoring and alerts at the beat level. The pipeline includes zero-phase band-pass filtering for higher noise wearable signals, optional notch removal, robust R-peak detection with refractory logic, and per-record z-score normalization, while confidence thresholds and short-horizon smoothing diminishes false-positive spikes. For a longer-term use case, we additionally suggest periodic calibration (normalization re-fit) and drift checks. Collectively these selections facilitate stable performance under realistic noise conditions for on-device or near-device execution.

HybridCardioNet achieves improved accuracy relative to CNN-only and RNN-only baselines, where the complementary feature extraction provided by convolutional blocks to capture fine-grained morphological cues (QRS width, ST changes, ectopic morphology) and stacked LSTMs to encode longer beat-to-beat dependencies which are underfitted by pure CNNs. The largest macro-F1 lift on minority classes (e.g., V, S) (Ablations (CNN-only, LSTM-only)), due to the need for temporal context especially where data are sparse. Notably, explicit imbalance control (class-weighted loss plus validated minority oversampling) remedies the majority-class bias seen in previous studies with unweighted loss or naive resampling. Such combination elucidates the state-of-the-art balanced precision/recall with respect to recent CNN, GRU and Wavelet-CNN reports on similar MIT-BIH protocols.

HybridCardioNet addresses a practically desirable accuracy–efficiency trade-off: Compact CNN front-ends constrain the number of parameters/FLOPs before temporal modeling, resulting in low latency approach for near real-time inference and competitively high latency while heavier transformer or multi-lead models in the literature tend to increase both compute and memory costs. We then profile runtime and observe that pruning or 8-bit quantization maintains accuracy within tight deltas but accelerates throughput—which is a boon for edge IoMT deployments. Lastly, while our stratified cross-validation and confusion-matrix analyses show strong minority-class performance on MIT-BIH, we elaborate on plans for external validation (to overcome dataset shift due to variations in sensors/leads/protocols) on INCART/PTB-XL and domain adaptation as future work to address generalizability. Importantly, this leads to fully explaining the gains from our design choices—opting for hybrid spatial–temporal encoding, principled imbalance handling, and measured complexity—while ensuring that the model is deployable.

This research has important implications for healthcare applications, such as automated cardiac monitoring and early diagnosis of heart diseases. The proposed approach paves the way for enhanced accuracy and robustness of ECG classification, which, in turn, fosters further advances in remote patient monitoring systems and, ultimately, clinical decision-making. Section 5.1 gives the limitations of this study.

5.1 Limitations of the study

The performance in MIT-BIH is not guaranteed under dataset shift (e.g., INCART, PTB-XL), as acquisition protocols, leads, and labeling vary; we will thus perform external validation with elementary & domain-adaptation steps (normalization re-fit, threshold calibration). Using compeitive Bayesian searches (efficientity tie-breaker) cross-validated to tune hyperparameters (filters, kernel

sizes, LSTM units, learning rate, class weights); we report full ranges and random seed used per dataset to reproduce given the sensitivity of hyperparameter choice to performance. At last, real-life wearable ECG streams are subject to motion artifacts, drift in the baseline and intermittent dropout; our pipeline does reduce noise by band-pass/notch filtering and robust R-peak detection and per-record normalization, extreme artifacts, however, will certainly damage minority-class recall. Things to do next have domain-shift–aware training, noise augmentation, and light-weight denoising/quantization, to stabilize their deployment across devices and settings.

6 Conclusion and future work

We introduced a novel deep learning framework, HybridCardioNet, that combines CNN and LSTM architectures to classify ECG signals. Combining state-of-the-art preprocessing, strong feature engineering, and hybrid model design, the proposed approach overcomes the critical limitations of the state-of-the-art while achieving 98.39% accuracy, well ahead of the others. HybridCardioNet is a suitable solution because it efficiently captures the spatial and temporal dependencies in the input ECG signals, resulting in high reliability and robustness for the automated detection of cardiac conditions. Results prove it is a promising candidate for clinical decision-making and the development of remote heart monitoring systems, which can also be used for the early recognition of heart diseases. Next steps will incorporate explainable AI to enhance clinical interpretability (Grad-CAM/Layer-CAM saliency maps over beat windows, integrated gradients for feature attributions, and clinician-rated plausibility checks) Through 8-bit quantization, structured pruning and knowledge distillation, we will adapt the model for mobile/edge inference and export to TensorFlow Lite/Core ML with ARM NEON acceleration and on-device latency profile. We will extend to multi-lead learnings to expand clinical utility and evaluate (lead-aware fusion and domain-adaptation (normalization re-fit, threshold calibration)) on PTB-XL and CPSC2018 datasets. Other directions such as noise augmentation for wearable artifacts, lightweight denoising front-ends, and runtime-budgeted architecture search to co-optimize accuracy, energy, and delay for continuous monitoring scenarios.

References

[1] K. T. Putra, A. Z. Arrayyan, N. Hayati, Firdaus, C. Damarjati, A. Bakar, and H.-C. Chen, “A review on the application of Internet of Medical Things in wearable personal health monitoring: A cloud-edge artificial intelligence approach,” IEEE, vol. 12, pp. 21437–21452, 2024. doi:10.1109/ACCESS.2024.3358827

[2] C. A. G. Junior, F. Miele, L. Mäkitie, E. Fiorenzato, M. Koivu, L. J. Bakker, C. Uyl de Groot, W. K. Redekop, and W. K. van Deen, “Attitudes toward the adoption of remote patient monitoring and artificial intelligence in Parkinson’s disease management: Perspectives of patients and neurologists,” Springer, vol. 17, pp. 275–285, 2024. doi:10.1007/s40271-023-00669-0

[3] V. T. Mai, K. A. Alattas, Y. Bouteraa, E. Ghaderpour, and A. Mohammadzadeh, “Personalized blood pressure control by machine learning for remote patient monitoring,” IEEE, vol. 12, pp. 83994–84004, 2024. doi:10.1109/ACCESS.2024.3413572

[4] R. Santos, B. Ribeiro, I. Sousa, J. Santos, F. Guede-Fernández, P. Dias, A. V. Carreiro, H. Gamboa, P. Coelho, J. Fragata, and A. Londral, “Predicting post-discharge complications in cardiothoracic surgery: A clinical decision support system to optimize remote patient monitoring resources,” Elsevier, vol. 182, pp. 1–8, 2024. doi:10.1016/j.ijmedinf.2023.105307

[5] K. N. Haque, J. Islam, I. Ahmad, and E. Harjula, “Decentralized Pub/Sub architecture for real-time remote patient monitoring: A feasibility study,” Springer, pp. 48–65, 2024. doi:10.1007/978-3-031-59080-1_4

[6] A. Mata-Lima, A. R. Paquete, and J. J. Serrano-Olmedo, “Remote patient monitoring & management in nephrology: A systematic review,” Elsevier, pp. 1–29, 2024. doi:10.1016/j.nefro.2024.01.005

[7] M. Khalifa, M. Albadawy, and U. Iqbal, “Advancing clinical decision support: The role of artificial intelligence across six domains,” Elsevier, vol. 5, pp. 1–7, 2024. doi:10.1016/j.cmpbup.2024.100142

[8] T. Arora, V. Balasubramanian, A. Stranieri, and V. G. Menon, “Modified early warning score (MEWS) visualization and pattern matching imputation in remote patient monitoring,” IEEE, vol. 12, pp. 74784–74794, 2024. doi:10.1109/ACCESS.2024.3396274

[9] A. Sundas, S. Badotra, G. S. Shahi, A. Verma, S. Bharany, A. O. Ibrahim, A. W. Abulfraj, and F. Binzagr, “Smart patient monitoring and recommendation (SPMR) using cloud analytics and deep learning,” IEEE, vol. 12, pp. 54238–54255, 2024. doi:10.1109/ACCESS.2024.3383533

[10] P. Ratta, Abdullah, and S. Sharma, “A blockchain-machine learning ecosystem for IoT-based remote health monitoring of diabetic patients,” Elsevier, vol. 5, pp. 1–18, 2024. doi:10.1016/j.health.2024.100338

[11] S. Faramarzi, S. Abbasi, S. Faramarzi, S. Kiani, and A. Yazdani, “Investigating the role of machine learning techniques in Internet of Things during the COVID-19 pandemic: A systematic review,”

Elsevier, vol. 45, pp. 1–16, 2024. doi:10.1016/j.imu.2024.101453

[12] D. Gupta, S. Rani, S. Raza, N. M. F. Qureshi, R. F. Mansour, and M. Ragab, "Security paradigm for remote health monitoring edge devices in Internet of Things," Elsevier, vol. 35, no. 6, pp. 1–9, 2024. doi:10.1016/j.jksuci.2022.12.020

[13] X. Wu, C. Liu, L. Wang, and M. Bilal, "Internet of Things-enabled real-time health monitoring system using deep learning," Springer, vol. 35, pp. 14565–14576, 2023. doi:10.1007/s00521-021-06440-6

[14] K. Boikanyo, A. M. Zungeru, B. Sigweni, A. Yahya, and C. Lebekwe, "Remote patient monitoring systems: Applications, architecture, and challenges," Elsevier, vol. 20, pp. 1–28, 2023. doi:10.1016/j.sciaf.2023.e01638

[15] H. H. Alshammari, "The Internet of Things healthcare monitoring system based on MQTT protocol," Elsevier, vol. 69, pp. 275–287, 2023. doi:10.1016/j.aej.2023.01.065

[16] F. Zeshan, A. Ahmad, M. I. Babar, M. Hamid, F. Hajjej, and M. Ashraf, "An IoT-enabled ontology-based intelligent healthcare framework for remote patient monitoring," IEEE, vol. 11, pp. 133947–133966, 2023. doi:10.1109/ACCESS.2023.3332708

[17] S. Akhbarifar, H. H. S. Javadi, A. M. Rahmani, and M. Hosseinzadeh, "A secure remote health monitoring model for early disease diagnosis in cloud-based IoT environment," Springer, vol. 27, pp. 697–713, 2023. doi:10.1007/s00779-020-01475-3

[18] S. Khan, S. Ullah, H. U. Khan, and I. U. Rehman, "Digital-twins-based Internet of Robotic Things for remote health monitoring of COVID-19 patients," IEEE, vol. 10, no. 18, pp. 16087–16098, 2023. doi:10.1109/JIOT.2023.3267171

[19] K. A. Shastry and A. Shastry, "An integrated deep learning and natural language processing approach for continuous remote monitoring in digital health," Elsevier, vol. 8, pp. 1–14, 2023. doi:10.1016/j.dajour.2023.100301

[20] O. Cheikhrouhou, K. Mershad, F. Jamil, R. Mahmud, A. Koubaa, and S. R. Moosavi, "A lightweight blockchain and fog-enabled secure remote patient monitoring system," Elsevier, vol. 22, pp. 1–32, 2023. doi:10.1016/j.iot.2023.100691

[21] T. Arora, V. Balasubramanian, and A. Stranieri, "Missing health data pattern matching technique for continuous remote patient monitoring," Springer, pp. 130–143, 2023. doi:10.1007/978-3-031-43950-6_11

[22] M. Alshamrani, "IoT and artificial intelligence implementations for remote healthcare monitoring systems: A survey," Elsevier, vol. 34, no. 8, pp. 4687–4701, 2022. doi:10.1016/j.jksuci.2021.06.005

[23] A. V. L. N. Sujith, G. S. Sajja, V. Mahalakshmi, S. Nuhmani, and B. Prasanalakshmi, "Systematic review of smart health monitoring using deep learning and artificial intelligence," Elsevier, vol. 2, no. 3, pp. 1–6, 2022. doi:10.1016/j.neuri.2021.100028

[24] T. N. Abiodun, D. Okunbor, and V. C. Osamor, "Remote health monitoring in clinical trial using machine learning techniques: A conceptual framework," Springer, vol. 12, pp. 359–364, 2022. doi:10.1007/s12553-022-00652-z

[25] E.-A. Paraschiv, C.-M. Petrache, and O. Bica, "On the continuous development of IoT in Big Data era in the context of remote healthcare monitoring & artificial intelligence," IEEE, pp. 1–6, 2022. doi:10.1109/ECAI54874.2022.9847503

[26] Z. Jeddi, "Remote patient monitoring using artificial intelligence," in Artificial Intelligence in Healthcare, pp. 203–234, 2020. doi:10.1016/B978-0-12-818438-7.00009-5

[27] S. Iranpak, A. Shahbahrami, and H. Shakeri, "Remote patient monitoring and classifying using the Internet of Things platform combined with cloud computing," Springer, vol. 8, no. 120, pp. 1–22, 2021. doi:10.1186/s40537-021-00507-w

[28] N. Sharma, M. Mangla, S. N. Mohanty, D. Gupta, P. Tiwari, M. Shorfuzzaman, and M. Rawashdeh, "A smart ontology-based IoT framework for remote patient monitoring," Biomed. Signal Process. Control, vol. 68, pp. 1–12, 2021. doi:10.1016/j.bspc.2021.102717

[29] E. T. R. Babar and M. U. Rahman, "A smart, low cost, wearable technology for remote patient monitoring," IEEE Sensors J., vol. 21, no. 19, pp. 21947–21955, 2021. doi:10.1109/JSEN.2021.3101146

[30] A. Ho, "Are we ready for artificial intelligence health monitoring in elder care?," BMC Geriatrics, vol. 20, no. 1, pp. 1–7, 2020. doi:10.1186/s12877-020-01764-9

[31] D. M. Hilty, C. M. Armstrong, A. E. Stewart, M. T. Gentry, D. D. Luxton, and E. A. Krupinski, "Sensor, wearable, and remote patient monitoring competencies for clinical care and training: Scoping review," Springer, vol. 6, pp. 252–277, 2021. doi:10.1007/s41347-020-00190-3

[32] M. R. Maurya, N. U. S. S. Riyaz, M. S. B. Reddy, H. C. Yalcin, H. M. Ouakad, I. Bahadur, S. Al Maadeed, and K. K. Sadasivuni, "A review of smart sensors coupled with Internet of Things and artificial intelligence approach for heart failure monitor," Springer, vol. 59, pp. 2185–2203, 2021. doi:10.1007/s11517-021-02447-2

[33] A. Banerjee, "Medical imaging, artificial intelligence, Internet of Things, wearable devices in terahertz

healthcare technologies," in Terahertz Biomedical and Healthcare Technologies, pp. 145–165, 2020. doi:10.1016/B978-0-12-818556-8.00008-2

[34] U. Hariharan, K. Rajkumar, T. Akilan, and J. Jeyavel, "Smart wearable devices for remote patient monitoring in Healthcare 4.0," Springer, pp. 1–19, 2021. doi:10.1007/978-3-030-63937-2_7

[35] P. Tagde, S. Tagde, T. Bhattacharya, P. Tagde, H. Chopra, R. Akter, D. Kaushik, and M. H. Rahman, "Blockchain and artificial intelligence technology in e-Health," Springer, vol. 28, pp. 52810–52831, 2021. doi:10.1007/s11356-021-16223-0

[36] Y. Xie, L. Lu, F. Gao, S.-J. He, H.-J. Zhao, Y. Fang, J.-M. Yang, Y. An, Z.-W. Ye, and Z. Dong, "Integration of artificial intelligence, blockchain, and wearable technology for chronic disease management: A new paradigm," Springer, vol. 41, no. 6, pp. 1123–1133, 2021. doi:10.1007/s11596-021-2485-0

[37] S. Oniani, G. Marques, S. Barnovi, I. M. Pires, and A. K. Bhoi, "Artificial intelligence for Internet of Things and enhanced medical systems," Springer, pp. 1–17, 2021. doi:10.1007/978-981-15-5495-7_3

[38] A. Vijayalaxmi, S. Sridevi, N. Sridhar, and S. Ambesange, "Multi-disease prediction with artificial intelligence from core health parameters measured through non-invasive technique," in Proc. ICICCS, pp. 1252–1258, 2020. doi:10.1109/ICICCS48265.2020.9121170

[39] B. Zaabar, O. Cheikhrouhou, M. Ammi, A. I. Awad, and M. Abid, "Secure and privacy-aware blockchain-based remote patient monitoring system for Internet of Healthcare Things," IEEE, pp. 1–6, 2021. doi:10.1109/WiMob52687.2021.9606362

[40] H. Fouad, A. S. Hassanein, A. M. Soliman, and H. Al-Feel, "Analyzing patient health information based on IoT sensor with AI for improving patient assistance in the future direction," Measurement, vol. 159, pp. 1–11, 2020. doi:10.1016/j.measurement.2020.107757

[41] G. B. Moody and R. G. Mark, "The impact of the MIT-BIH arrhythmia database," IEEE Eng. Med. Biol. Mag., vol. 20, no. 3, pp. 45–50, 2001. Available: https://physionet.org/physiobank/database/mitdb/

[42] K. Srinivas, C. Vijayalakshmi, S. R. Borra, K. S. Raju, G. R. K. Rao, K. N. V. Satyanarayana, and P. M. Kumar, "CardiacNet: Cardiac arrhythmia detection and classification using unsupervised learning-based optimal feature selection with a custom CNN model," Informatica, vol. 48, no. 2, pp. 147–158, 2024. doi:10.31449/inf.v48i2.5076

[43] A. Ochoa, L. J. Mena, V. G. Felix, A. Gonzalez, W. Mata, and G. E. Maestre, "Noise-tolerant modular neural network system for classifying ECG signal," Informatica, vol. 43, no. 1, pp. 109–116, 2019. doi:10.31449/inf.v43i1.1605

A Multi-Class Muscle Fatigue Recognition Model for Athletes Using KPCA Dimensionality Reduction and SVM Classification on Multi-Channel sEMG Signals

Xingping Chu, Dongqin Huang, Yan Yi*
College of Sports and Health, Jiangxi University of Chinese Medicine, Nanchang, 330004, China
Email: chuxingping2021@163.com, huang9y012@126.com, chu2025@126.com
*Corresponding author

Keywords: muscle fatigue, sEMG, kernel principal component analysis (KPCA), support vector machine (SVM), feature dimensionality reduction

Received: July 24, 2025

Muscle fatigue is an inevitable phenomenon in athletic training, and its accurate assessment is crucial for preventing injuries. To address the issues of high redundancy in multi-channel surface electromyography (sEMG) features and the limitations of linear dimensionality reduction methods, this study proposes a muscle fatigue recognition model based on Kernel Principal Component Analysis (KPCA) and Support Vector Machine (SVM). sEMG signals were collected from 12 subjects performing sustained contractions. A comprehensive set of 48 time-domain, frequency-domain, and nonlinear features was extracted. KPCA was employed for nonlinear dimensionality reduction before classification. The resulting features were fed into an SVM classifier to distinguish between three fatigue states: relaxed, transitional fatigue, and fatigued. The model was evaluated using 10-fold cross-validation. Results demonstrated that the KPCA-SVM combination achieved the highest performance, with an average recognition accuracy of 91.5%, precision of 0.91, recall of 0.93, and F1-score of 0.92, outperforming other combinations of dimensionality reduction methods (MI, PCA) and classifiers (FLDA, KNN). This method provides an effective tool for the objective assessment of muscle fatigue in athletes.

Povzetek: Študija predstavi model KPCA + SVM za prepoznavanje treh stanj mišične utrujenosti iz sEMG, ki z zmanjšanjem odvečnih značilk doseže okoli 91,5 % natančnost in prekaša primerjane pristope.

1 Introduction

Athletes get good grades is the premise of the need for a lot of training, and a lot of training will increase the injury probability of athletes, because with the increase of training intensity and time will produce muscle fatigue, serious will cause muscle damage [1]. Hence need to be testing technology service athlete's fatigue, avoid the occurrence of muscle damage. But in recent years, (multi-channel sEMG surface electromyography, sEMG) because of its noninvasive, convenient detection merits attention in evaluation of muscle fatigue [2]. Are produced by the muscle contraction multi-channel sEMG weak electrical signals of contains many useful biological information, biological medicine, sports medicine and rehabilitation medicine and other fields play an extremely important role in [3].

Moniri et al. [4] developed advanced techniques for real-time forecasting of sEMG features related to trunk muscle fatigue using machine learning, demonstrating the potential of sEMG in fatigue analysis. From then on, the development of sEMG equipment constantly updated, promoted the study of sEMG signal. American Delsys of wireless sEMG acquisition instrument [5] the most widely used. It includes 16 channel wireless sensor with three degrees of freedom of acceleration sensor, small volume and easy to carry, up to 5 h of battery life. Xu et al. [6] utilized a convolutional neural network for sEMG-based feature prediction, providing a general model for sports activity analysis. In a study on pilot neck fatigue assessment, Rampichini et al. [7] achieved a high recognition accuracy of 91.25% by employing a hybrid optimization algorithm to tune a Gaussian process model. This work demonstrates the potential of advanced machine learning models in fatigue recognition, which motivates our exploration of KPCA and SVM for sEMG-based fatigue analysis in athletes.

However, while these methods demonstrate promising results, they often rely on direct use of multi-channel sEMG feature parameters, which may exhibit high correlation and redundancy. This can lead to increased computational complexity and reduced classification accuracy in fatigue models [8]. To address these limitations, this study introduces a novel approach that combines kernel principal component analysis (KPCA) and support vector machines (SVM). KPCA effectively reduces feature dimensionality by capturing nonlinear relationships in sEMG signals, while SVM ensures robust classification performance even with limited data. This method not only enhances the efficiency of fatigue detection but also avoids the extensive data requirements and computational overhead associated with deep learning models.

As mentioned above, the direct use of the characteristic parameters of multi-channel sEMG,

although also has good recognition effect, but has a great deal of correlation between characteristic parameters and redundancy, without considering the complexity of the algorithm, often affect the fatigue model classification recognition rate [8]. This article in order to solve this problem, in the original feature parameters under the premise of no loss of the original information, using different feature dimension reduction method, remove redundant information, a new feature set, the use of classifier inspection classification effect, from which identify the most effective a classification model of fatigue.

Compared to existing studies that often directly use raw or high-dimensional sEMG feature sets [8], the main contributions of this paper are threefold: (1) It systematically compares the effectiveness of three distinct feature dimensionality reduction techniques (MI, PCA, and KPCA) for muscle fatigue recognition, highlighting the importance of addressing feature redundancy and nonlinearity; (2) It proposes a novel fatigue recognition framework that synergistically combines KPCA for nonlinear feature extraction and SVM for robust classification, demonstrating that this combination effectively captures the complex patterns in sEMG signals associated with different fatigue states while maintaining computational efficiency; (3) The proposed KPCA-SVM model achieves a superior average recognition rate of 91.5% for three-state fatigue classification, offering a practical and high-accuracy solution that avoids the large data requirements of deep learning models, making it particularly suitable for scenarios with limited training data.

The primary objectives of this study are: (1) to develop and evaluate a framework for muscle fatigue recognition by comparing the effectiveness of different feature dimensionality reduction techniques (MI, PCA, KPCA) combined with various classifiers (FLDA, KNN, SVM); (2) to specifically investigate the performance of the KPCA-SVM model in classifying three states of muscle fatigue (relaxed, transitional, fatigued) from multi-channel sEMG signals; (3) to validate the proposed model's accuracy and efficiency against other common approaches.

2 The characteristics of the multi-channel sEMG analysis method

2.1 The time domain characteristics analysis method

Time domain features of multi-channel sEMG can reflect the change of the signal in the time dimension, from the perspective of time, will be understood as a multi-channel sEMG time-varying function, has the advantages of simple, rapid computation [9]. Common characteristics such as RMS values, integral electrical values, the average absolute value and the zero rate of the mean, variance and amplitude cube [10].

1) The Root Mean Square Value (Root Mean Square, RMS)

Root means square value indicates change in unit time, multi-channel sEMG can respond muscle activity. As shown in Equation (1):

$$RMS = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2} \tag{1}$$

Type, $x_i(i = 1,2, \dots, N)$ is the time series of the signal length *N* to select the number of frames.

2) The Integral Electrical Values (Integrated EMG, iEMG)

Integral electrical value indicates that the activities of the muscle fibers of charge level, reaction of multi-channel sEMG amplitude changes with exercise. As shown in Equation (2):

$$iEMG = \sum_{i=1}^{N} |x_i| \tag{2}$$

Type, $x_i(i = 1,2, \dots, N)$ is the time series of the signal length *N* to select the number of frames.

3) The Average Absolute Value (Mean Absolute Value, MAV)

The average absolute value used to identify the muscles. As shown in Equation (3):

$$MAV = \frac{1}{N}\sum_{i=1}^{N} |x_i| \tag{3}$$

Type, $x_i(i = 1,2, \dots, N)$ is the time series of the signal length *N* to select the number of frames.

4) The Zero Rate (Zero Crossing Rate, ZCR)

Zero crossing rate refers to the sEMG symbols change ratio, per unit time is an important index of the time-domain analysis. As shown in Equation (4):

$$ZCR = \sum_{i=1}^{N-1\sum(-x_i x_{i-1})} sgn \tag{4}$$

Type, $x_i(i = 1,2, \dots, N)$ is the time series of the signal length *N* to select the number of frames.

5) Variance (Variance, VAR)

Variance is used to measure the parameters of the degree of discrete random variables. As shown in Equation (5):

$$VAR = \frac{\sum_{i=1}^{N}(x_i - \overline{x_i})}{N} \tag{5}$$

Type, $x_i(i = 1,2, \dots, N)$ is the time series of the signal length *N* to select the number of frames.

6) The Mean Amplitude Cube (Amplitude Cubic scheme, ACM)

Average amplitude cube can response in time domain amplitude change rule of multi-channel sEMG. As shown in Equation (6):

$$ACM = \frac{1}{N}\sum_{i=1}^{N} x_i^3. \tag{6}$$

Type, $x_i(i = 1,2, \dots, N)$ is the time series of the signal length *N* to select the number of frames.

2.2 Analysis of characteristics of frequency domain method

Methods of frequency domain feature of electricity is will signal through the fast Fourier transform (FFT), according to the power spectrum of signal or frequency spectrum is analyzed [11]. Commonly used indicators have mean power frequency and spectral moments and the median frequency.

1) The Mean Power Frequency (Mean Power Frequency)

Mean power frequency with muscle fatigue deepening decline, the reason is that lower pH, results in the decrease of muscle fiber conduction velocity. As shown in Equation (7):

$$MPF = \frac{\int_0^{\infty} f \times PSD(f)df}{\int_0^{\infty} P\,SD(f)df} \tag{7}$$

Type of f sEMG signal frequency, $PSD\ (f)$ for sEMG signal power spectral density function.

2) Spectral Moment (Spectral Moment, SM)

Spectral moment is Dimitrov put forward a set of parameters of muscle fatigue, two order spectral moment is often used to describe the change of high frequency. As shown in Equation (8):

$$SM_2 = \int_0^{\infty} f^2\,PSD(f)df. \tag{8}$$

Type of f sEMG signal frequency, $PSD\ (f)$ for sEMG signal power spectral density function.

3) The median Frequency (Media Frequency, MF)

Found in the study of muscle fatigue estimation, the low frequency component of the multi-channel sEMG increases with muscle fatigue, quantitative coefficients are part of the representative indicators is the median frequency. As shown in Equation (9):

$$MF = \frac{1}{2}\int_0^{\infty} P\,SD(f)df \tag{9}$$

Type of f sEMG signal frequency, $PSD\ (f)$ for sEMG signal power spectral density function.

2.3 Time and frequency domain characteristics analysis method

Time domain analysis was conducted on the signal in time domain analysis, studies the changing relation between time and signal amplitude; Frequency domain analysis is in the frequency domain analysis of signals, the frequency change [12]. While the aforementioned time-frequency analysis methods are foundational, recent advancements have introduced more sophisticated techniques for handling non-stationary signals like sEMG. Wavelet Transform (WT), particularly the Discrete Wavelet Transform (DWT), has gained prominence for its multi-resolution analysis capability, allowing for feature extraction across different frequency bands at varying temporal resolutions [13]. Although the current study focuses on standard time, frequency, and time-frequency features for model simplicity and comparability, future work will explore integrating wavelet-based features to further enhance the model's performance in characterizing muscle fatigue dynamics.

2.3.1 The short-time Fourier transform short-time

Fourier transform is a transform and Fourier transform, can be implemented by the Fourier transform, is one of the more commonly used a time-frequency analysis method . *STFT* principle is to put a long non-stationary signal as a superposition of countless short time stationary signal. For signal $x(t)$, as well as with a moving time window is decomposed into countless time length of short time signal, so during this period can take it as a stationary signal, and then to the Fourier transform, signal time and frequency domain information.

Define functions $STFT(w,\tau)$, said in the center of the window function of τ, to transform the function of spectrum as a result, ss shown in Equation (10):

$$STFT(w,\tau) = \int_{-\infty}^{+\infty} [x(\tau)w(t-\tau)]\,e^{-jwt}d\tau\cdot \tag{10}$$

In the short time Fourier transform, each of the different time, after the transformation can be a different spectrum, the frequency spectrum is combined time-frequency distribution. The window length determines the time of spectrum, frequency resolution, the longer the length of the window, the Fourier transform of frequency domain resolution is higher, but the lower temporal resolution; on the other hand, the shorter the window, the Fourier transform of frequency domain resolution is lower, but the higher time resolution. Therefore, in practice need to according to the situation, between the time resolution and frequency resolution. Although the short time Fourier transform algorithm is very simple, can deal with non-stationary signal, applied to various fields, but the adaptive ability is poor.

2.3.2 The Winger-Ville distribution

In 1932, Winge presents a time-frequency analysis method, and applied to the field of quantum mechanics, later Ville signal analysis, introduced the method, formed the Winger-Ville distribution.

Signal, a Winger-Ville distribution can be defined as:

$$W_{x,y}(t,w) = \int_{-\infty}^{+\infty} x\left(t+\frac{\tau}{2}\right)y^*\left(t-\frac{\tau}{2}\right)e^{-jwt}d\tau \tag{11}$$

Type 11 is also called the bilinear time-frequency analysis method, can also be represented as a signal of *x*(*t*) from Winger-Ville distribution form, as shown in Equation (12):

$$W_x(t,w) = \int_{-\infty}^{+\infty} x\left(t+\frac{\tau}{2}\right)x^*\left(t-\frac{\tau}{2}\right)e^{-jwt}d\tau \tag{12}$$

Based on the definition of Winger-Ville distribution, frequency spectrum of signal *t* any time, on time *t*, *t* time signal multiplication, all results are superimposed and then do the Fourier transform, because the signal is not a single component, can produce interaction among computing, form the cross terms. When the Winger-Ville distribution crossover component is bigger when a certain cross interference, often cover a useful time and frequency information. In order to avoid the Winger-Ville distribution in the time-frequency plane cross interference, the introduction of window function, the *x*(*t*) can be added

window, the resulting Winger-Ville distribution is defined as shown in Equation (13):

$$C_Z = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} z\left(u+\frac{\tau}{2}\right)z^*\left(u-\frac{\tau}{2}\right)\psi(t - u,\tau)e^{-jwt}dud\tau \quad (13)$$

Type of window function, $h(\tau)$ is equal to the difference between in the frequency domain to the Winger Ville distribution of smoothing.

Compared with the short time Fourier transform, Winger-Ville distribution can accurately identify signals is a single component or component, get better time-frequency information. But Winger-Ville distribution of different frequency components will cause the interference of cross terms, although after add window will have certain inhibition on the cross-term interference, but cannot completely eliminate its existence.

2.3.3. Choi-William's distribution

Choi-William's distribution is insufficient for Winger-Ville distribution problem put forward the improved algorithm. In order to complete the related domain filtering of fuzzy function, eliminate the cross-term interference formed by mutual fuzzy function, Choi and Williams with the method of the kernel function is proposed. As shown in Equation (14):

$$C_Z = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} z\left(u+\frac{\tau}{2}\right)z^*\left(u-\frac{\tau}{2}\right)\psi(t - u,\tau)e^{-jwt}dud\tau \quad (14)$$

Type of $\psi(t-u,\tau)$ for Cohen class time-frequency distribution kernel function.

2.4 Nonlinear characteristics analysis method

The contraction of muscles in different state, the motor nerve cell number in time, space and there exist certain differences, therefore sEMG has obvious nonlinear characteristics [14]. Therefore, put forward a new method of analysis, complexity and nonlinear analysis of time series method is an office value, the greater the signal contains information is more complex. In this paper, extraction of multi-channel sEMG approximate office as the research of nonlinear characteristic parameters.

Approximate Entropy (ApEn), introduced by Pincus [15], is a nonlinear dynamic parameter used to quantify the regularity and unpredictability of a time series. A higher ApEn value indicates greater complexity and irregularity in the signal. It has been widely applied in physiological signal analysis, including sEMG, to assess the complexity changes associated with muscle fatigue [16]. The calculation of ApEn involves comparing the similarity of patterns within the signal for increasing pattern lengths (m and m+1).

3 The analysis method of the feature dimension reduction and classification model

3.1 Feature dimension reduction

This article is based on three channels sEMG signal, a total of 48 d characteristic parameters can be extracted. High-dimensional data contains a large number of redundant features, the classification of the redundant features of the follow-up effect and increase the complexity of the operation, and high dimensional data can lead to "fitting" and "dimension disaster" problem, reduce the performance of the classifier. Extraction of effective feature parameters is the key to the classification of fatigue, so need to dimension reduction of feature set, the classification model recognition rate [17]. The overall workflow of the proposed muscle fatigue recognition model is illustrated in Figure 1. It outlines the key stages: multi-channel sEMG signal acquisition, feature extraction, dimensionality reduction, and classification, providing a clear roadmap for the methodology described in this section.

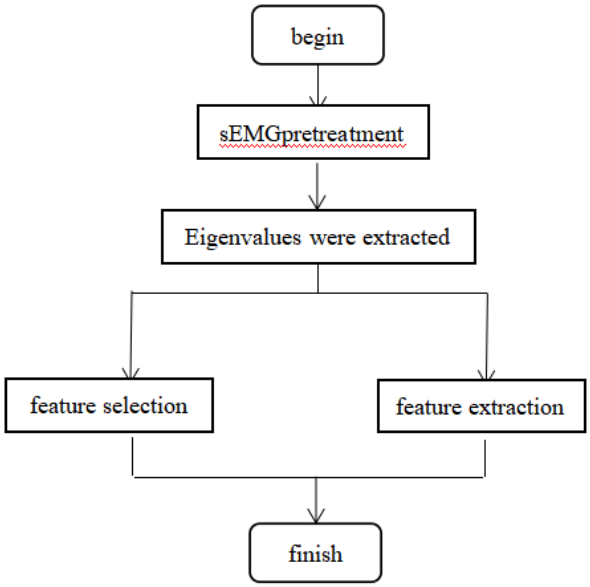


Figure 1: The whole algorithm flow chart.

3.1.1 Feature selection

Feature selection is according to an evaluation criterion, from centralized to choose some of the most effective characteristics, thus reducing the process of feature dimension [18]. The Dash gives the framework of feature selection, the following is shown in Figure 2.

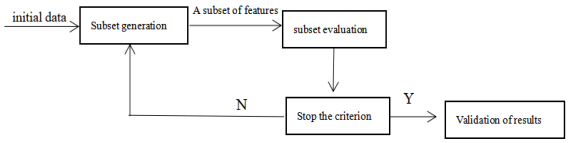


Figure 2: The framework of feature selection.

Mutual information (MI) measurement is one of the important measures of feature selection, without considering the distribution of sample in advance, and can deal with linear or nonlinear random variables [19]. MI value is larger, can explain the interrelation between the two variables, the greater the smaller conversely.

MI calculation equation is shown below:

$$I(U;V) = \sum_{u\in U}\sum_{v\in V} p\,(u,v)\,log\frac{p(u,v)}{p(u)p(v)} \tag{15}$$

Type of $p(u,v)$ is the joint probability distribution function of u and v, $p(u)$ and $p(v)$, respectively is the edge of the u and v probability the function.

3.1.2 Feature extraction

1) The principal component analysis method

Principal component analysis is a commonly used linear feature extraction method, the linear relationship for a set of possible data, through the orthogonal transformation, principal component analysis to data mapped to a low-dimensional subspace of linear independence, the basic principle of the algorithm are: Original each dimension data of the covariance matrix as well as the eigenvalue and eigenvector, the characteristic value from big to small, retain data high contribution rate, ignore the data of low contribution rate, so as to realize the feature dimension reduction and random vector $x_i,\ i = 1,2,\ \ldots M$ which M as the total number of samples, N is the dimension of data [20]. The sample can be X_{11} To form $M \times N$ matrix, namely $X = \begin{Bmatrix} x_{11} & \cdots & x_{1N} \\ \vdots & \ddots & \vdots \\ x_{M1} & \cdots & x_{MN} \end{Bmatrix}$.

In the practical implementation of KPCA in this study, the Radial Basis Function (RBF) kernel was employed. The kernel parameter σ was determined through a grid search within a range of [0.01,1], optimizing for the subsequent classification performance. The number of nonlinear principal components was selected to ensure a cumulative variance contribution rate of over 95%, effectively reducing the feature dimensionality from 48 to 10 while preserving the most critical discriminatory information.

According to matrix $X_{m\times n}$ column, each column of the mean, calculate the average each column minus the column, as shown in the Equation (16), with each column number minus the corresponding column mean, get new matrix $B = \begin{Bmatrix} \overline{x_{11}} & \cdots & \overline{x_{1N}} \\ \vdots & \ddots & \vdots \\ \overline{x_{M1}} & \cdots & \overline{x_{MN}} \end{Bmatrix}$.

$$\bar{x} = x_i - \frac{1}{M}\sum_{i=1}^{M} x_i \tag{16}$$

And then according to the B^T type 16 obtains its transpose matrix, computing characteristics of covariance matrix C, as shown in the Equation (17):

$$C = B^T \times B \tag{17}$$

Then calculate the eigenvalues of the covariance matrix C, β_i, λ_i and eigenvector, the eigenvalue λ_i from big to small before order, according to the contribution rate to select the appropriate k composition, to constitute the new data. K before features vector matrix such as type of (18):

$$W = (\beta_1, \beta_2, \ldots, \beta_k) \tag{18}$$

Refactoring data dimension reduction to k-dimensional $P = W \times X$.

2) The kernel principal component analysis method.

Some fatigue characteristic value may show the nonlinear relationship between fatigue and, therefore, we need a new way to solve the problem. Kernel principal component analysis is an improved algorithm based on principal component analysis, is a kind of nonlinear feature extraction method, solve the problem of principal component analysis is unable to realize, compared with the principal component analysis greatly reduced the amount of calculation, and provides a better recognition performance [21]. Its basic principle is as follows: First of all, through nonlinear mapping, transform the data into a high-dimensional nonlinear space, and then using linear PCA mapped to a low dimensional space. In this study, the kernel component analysis (KPCA) employs radial basis function kernels. The kernel parameter σ was optimized through a grid search combined with 5-fold cross-validation, with the search range set to [0.1, 0.5]. The final selected parameter σ=0.1 achieved a cumulative variance contribution rate exceeding 95%, while reducing the feature dimensionality from 48 to 10.

For the n input sample data $x_k(k = 1,2,\ldots,n),\ x \in R^N$, introducing the nonlinear mapping function Φ, transform data x_k is $\Phi(x_k)$, then in the new feature space, covariance matrix for C.

Using the iterative method to solve eigenvalue $\lambda(\lambda > 0)$ and characteristic vector $V(V \neq 0)$, the specific Equation (19):

$$\lambda V = CV \tag{19}$$

And V by $\Phi(x)$ linear said, as shown in Equation (20):

$$V = \sum_{j=1}^{n} \alpha_j\,\Phi(x_j), j = 1,2,\ldots,n \tag{20}$$

Type in α_j for the equation coefficient, Equation (21) left 19 times $\Phi(x_k)$:

$$\lambda(\Phi(x_k)\cdot V) = \Phi(x_k)\cdot CV, k = 1,2,\ldots,n \tag{21}$$

Define a $n \times n$ matrix K, the details are shown in Equation (22):

$$K_{ij} = \Phi(x_i)\Phi(x_j), i,j = 1,2,\ldots,n \tag{22}$$

Equation (23), for solving the nuclear matrix K of non-zero eigenvalue λ and eigenvector alpha α.

$$n\lambda\alpha = K\alpha \tag{23}$$

Centralized eigenvector V, sample data $\Phi(x)$ in the mapping Equation (24):

$$h_i(x) = (V, \Phi(x)) = \sum_{i=1}^{n} \alpha_i\,\Phi(x_i) \tag{24}$$

$h_i(x)$ to $\Phi(x)$ of the first k nonlinear principal component vectors.

In addition to dimensionality reduction techniques like PCA and KPCA, the selection of appropriate features from the sEMG signals is crucial for accurate muscle fatigue detection. In this study, we focus on Root Mean Square (RMS), Mean Absolute Value (MAV), and Zero Crossing Rate (ZCR) as primary features. RMS is chosen because it quantifies the amplitude variation of the sEMG signal, which is directly related to muscle contraction intensity and tends to decrease with fatigue. MAV measures the overall signal strength by averaging the absolute values of the signal, making it sensitive to changes in muscle activity during prolonged exercise. ZCR captures the frequency characteristics of the signal by counting the number of times the signal crosses the zero line, which decreases as muscle fatigue causes a shift toward lower frequencies. These features have been widely used in muscle fatigue detection due to their simplicity and effectiveness.

3.2 Muscle fatigue classifier selection

Classification of fatigue is the last step of fatigue model is set up, on the extraction of feature dimension reduction, multi-channel sEMG its inputs to the classifier for training [22]. This article USES the Fisher linear discriminant analysis, *K* neighbor and three kinds of support vector machine classifier to classify fatigue.

3.2.1 Fisher linear discriminant analysis

Fisher linear discriminant analysis (FLDA) in the field of pattern recognition, introduced in the 1990s, is a classic algorithm in pattern recognition. Linear discriminant analysis is a kind of, supervision and learning needs for data classification label in advance, the basic idea is to project high-dimensional data to identify best low dimensional vector space, in the new subspaces, maximize the spacing between the similar sample, the sample spacing between different categories to minimize, or similar samples gathered as much as possible, samples of different categories of decentralized as much as possible. The core idea of FLDA is to find a projection direction that maximizes the distance between classes while minimizing the variance within each class, as conceptually depicted in Figure 3. This property makes it suitable for distinguishing well-separated fatigue states.

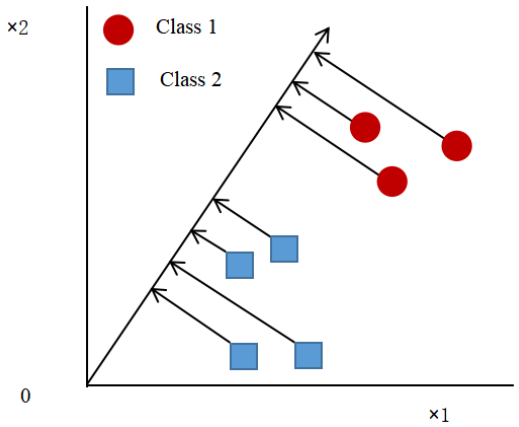


Figure 3: Fisher schematic diagram of the linear discriminant analysis principle.

Given sample set $(x_i,\ y_i), i = 1,2,\ldots,n, x \in R^N, y \in \{-1,1\}$, the first-class y the sample mean and variance as the Equation (25):

$$\mu_y = \frac{1}{N_y}\sum_{x_i \in y} x_i$$
$$s_y^2 = \sum_{x_i \in y}\left(x_i - \mu_y\right)^2 \quad (25)$$

Type in *N*, said the number of sample points, the first (x_i, y_i) mapping to γ direction vector, is the first i a sample points on its projection for the Equation (26):

$$z_i = \gamma^T x_i \quad (26)$$

Is mapped to the first-class y sample direction vector γ on the mean and variance of the concrete as shown in Equation (27):

$$v_y = \gamma^T \mu_y \quad (27)$$

When the same sample interval with different kinds of sample spacing ratio, the largest best direction vector γ requires specific can be obtained as shown in the Equation (28):

$$argmax\, J\,(\gamma) = \frac{v_{-1} - v_1}{\sigma_{-1}^2 - \sigma_1^2} \quad (28)$$

Mean value and variance of (27) will type into type (28) and Equation (29) can be obtained:

$$\frac{\gamma^T(\mu_{-1} - \mu_1)(\mu_{-1} - \mu_1)^T\gamma}{\sum_{x_i \in y_{-1}} \gamma^T\,(x_i - \mu_{-1})(x_i - \mu_{-1})^T\gamma + \sum_{x_i \in y_1} \gamma^T\,(x_i} \quad (29)$$

Type of $(\mu_{-1} - \mu_1)(\mu_{-1} - \mu_1)^T$ class is called the divergence between the matrix and the available S_a said. $(x_i - \mu_{-1})(x_i - \mu_{-1})^T$ and $(x_i - \mu_1)(x_i - \mu_1)^T$ known as covariance matrix, and use the S_ω said together, that is, class divergence within the matrix. 30 type $2b$ after reduction for the equation:

$$argmax\, J(\gamma) = \frac{\gamma^T s_a \gamma}{\gamma^T s_\omega \gamma} \quad (30)$$

Type of s_a and s_ω, called generalized Rayleigh quotient, the $\gamma^r s_\omega \gamma = 1$, though laser multiplier method is used to calculate the maximum eigenvalue of matrix ${s_a}^{-1} s_\omega$ λ with the corresponding eigenvectors β, β is the direction vector β best value.

3.2.2 *K* neighbor

K Neighbor (*K*—on his Neighbor, KNN, put forward by Cover and others in the sixties of the 20th century, is a kind of typical supervised learning algorithm. *K* neighbor algorithm is the core idea: Classification of calculation for the distance between the training set and test set sample, find the nearest *K* training sample points, will stay classification test set samples for this *K* training set of sample points appear in the category of at most. The K-Nearest Neighbors (KNN) algorithm classifies a sample based on the majority vote of its nearest neighbors in the feature space, as shown in **Figure 4**. The choice of K value, as illustrated, critically affects the decision boundary.

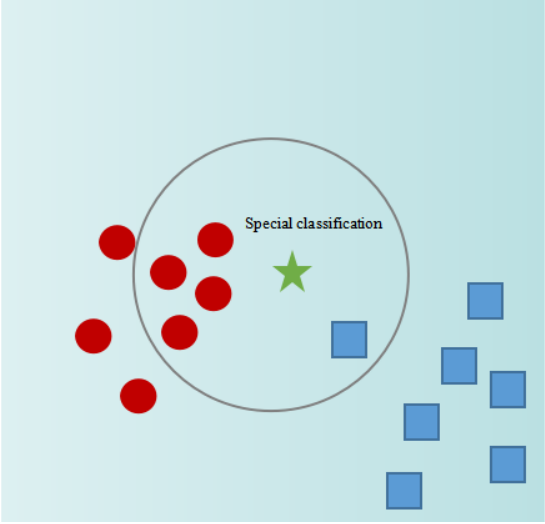


Figure 4: Schematic representation of the *K*-neighbor principle.

K neighbor algorithm has three elements, including the selection of *K* value, the distance measure and decision rules. *K* values will produce less fitting, and too small will have a fitting, affect classification recognition rate, generally with the method of cross validation to set the number of *K* value; Distance measurement with multiple methods, common with Euclidean distance, Manhattan distance, etc., this paper USES the Euclidean distance; For decision rules, in pattern recognition, generally USES the "majority voting method".

Set the sample sets $(x_i, y_i), i = 1,2,\ldots,n, (x_i, y_i \in Z)$, the calculation of Euclidean distance equation is:

$$d(x_i, y_i) = \sqrt{\sum_{i=1}^{n} (x_i, y_i)^2} \quad (31)$$

To calculate the distance the sample concentration sample points, and according to the distance size in descending order; Select the current classification sample points, the most recent *K* points, adopts "majority voting method", choose the highest frequency of sample points class, as a test sample point prediction of classes.

3.2.3 Support vector machine (SVM)

SVM (support vector machine, SVM) is put forward by Vapnik, first of all, is based on the statistical theory of VC dimension theory and structural risk minimum theory on the basis of a Machine learning algorithm. Core strategy is to find an optimal hyperplane, the support vector distance to the hyperplane is the largest, namely maximize between class and class intervals, can be applied to the characteristics of high dimension and nonlinear interface, etc., is a typical binary classification model. The SVM algorithm, is widely studied because of its classification performance advantage, mainly reflected in the system structure is simple, the global optimal, promotion ability, learning and prediction time is short. This experiment implements the SVM classifier using the LIBSVM library. The radial basis function kernel is employed, with both the penalty parameter C and kernel parameter γ optimized through grid search. The search space for C is defined as {0.1, 1, 10,100} and γ as {0.01,0.1, 1}. The optimization objective is determined by the average classification accuracy obtained from 10-fold cross-validation. The principle of Support Vector Machine (SVM) is to find the optimal hyperplane that maximizes the margin between classes, as shown in Figure 5. This approach ensures robust generalization performance.

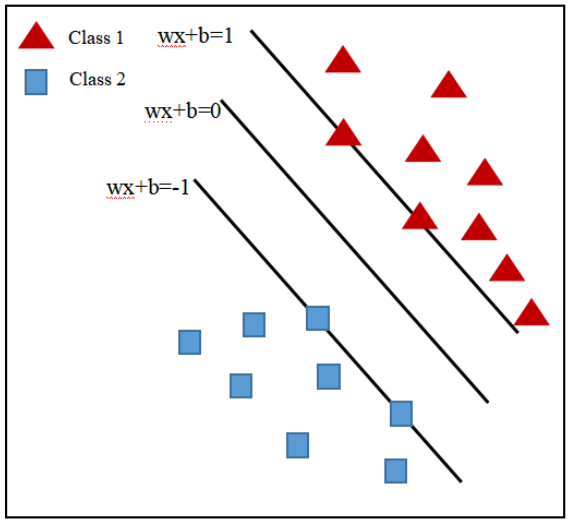


Figure 5: A schematic diagram of the SVM principle.

Given training set sample (x_i, y_i), $x \in R^n$, $y \in (-1,1)$, hyperplane $\omega \cdot x + b = 0$. In order to make the sample training set classification is correct, the details are shown in Equation (32):

$$y_i[(\omega \cdot x_i + b)] \geq 1, i = 1,2,\ldots,n \quad (32)$$

By the support vector can get $y_k[\omega \times x_k + b] = 1$, classification interval is calculated for $2/\|\omega.\|$, can be converted into minimum value problem with constraint conditions of (32), as shown in the Equation (33):

$$min\, \Phi(\omega) = \frac{1}{2}\|\omega\|^2 = \frac{1}{2}(\omega \cdot \omega) \quad (33)$$

Introducing the though laser function $L = (\omega, b, \alpha)$, to ω after the partial derivatives into a and b, which is transformed into the dual problem, as shown in the Equation (34):

$$max \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i x_j$$
$$s.t.\$\sum_{j=1}^{m} \alpha_i y_i = 0 \; \alpha_i \geq 0, i = 1,2,\ldots,m \quad (34)$$

The calculated optimal weight vector ω^* and optimal offset b^*, the optimal hyperplane is obtained $(\omega^* \cdot x) + b = 0$, and get the optimal classification function Equation (35):

$$sgn\left\{\left(\sum_{i=1}^{m} \alpha_i^* y_i(x_i \cdot x_j)\right) + b^*\right\} \quad (35)$$

In this study, the SVM classifier was implemented using the scikit-learn library in Python. The Radial Basis Function (RBF) kernel was selected due to its effectiveness in handling nonlinear classification problems. The hyperparameters, including the penalty parameter C and the kernel coefficient γ, were optimized through a grid search strategy combined with 5-fold cross-validation on the training set. The search ranges were set as C = [0.1, 1, 10, 100] and γ = [0.01, 0.1, 1]. The combination that yielded the highest cross-validation accuracy was selected for the final model.

4 Classification of fatigue experiment and result analysis

Extensive research has been conducted on sEMG-based muscle fatigue recognition, with Table 1 summarizing representative methodologies developed in recent years. For instance, Xu et al. [6] employed convolutional neural networks (CNN) for feature prediction, though their model exhibited high complexity. Rampichini et al. [7] utilized a Gaussian process model achieving 91.25% accuracy, yet failed to resolve feature redundancy issues. Most studies directly applied multi-channel sEMG features, which resulted in redundant data and insufficient capture of nonlinear relationships through linear approaches [8]. To address these limitations, this paper introduces K-means Projection (KPCA) for nonlinear dimensionality reduction combined with Support Vector Machines (SVM) classifiers, aiming to achieve high-precision classification with reduced computational costs.

Table 1: Comparison of work related to muscle fatigue and sEMG recognition

Document	Feature extraction method	Classification method	Identification accuracy	Number of subjects	Boundedness
Xu et al. [6]	Convolutional neural network	CNN	~90%	10	It requires a lot of data and complex calculations
Rampichini et al. [7]	Mixed Gaussian process	Bacterial feeding optimization	91.25%	8	Features may be redundant
Yun et al. [19]	Mutual information	LDA	85%	15	Linear methods may not capture nonlinear relationships
This research	KPCA	SVM	91.5%	12	Effective dimension reduction, nonlinear processing, high computational efficiency

4.1 Fatigue experiment scheme design

4.1.1 Objects and materials

The 120 subjects for the lab, including 60 boys and 60 girls, are right-handed, no nervous musculoskeletal diseases. In good health, between the ages of 16 and 20, 24 h before the experiment for high intensity exercise, no physical and mental fatigue. Each subject before the experiment were told the experimental process, at the same time sign the informed consent.

4.1.2 The experimental process

Experimental subjects before the corresponding action guidance to help their familiar with Borg subjective fatigue rating scale, and then according to individual circumstance to the experiment of several times to make participants familiar with the experimental process. Previous research, the experiment used more by those recorded every time ask participants state, so easy to interfere with the subjects, to give participants a certain psychological pressure, and different subjects, the time to reach the level of fatigue is endless also and same, easy to miss the corresponding fatigue level. The raw sEMG signals were acquired at a sampling rate of 2000 Hz using

a Delsys Trigno wireless system. Prior to feature extraction, the signals were processed with a 4th-order Butterworth band-pass filter (20-450 Hz) to remove low-frequency drift and high-frequency noise, followed by a notch filter at 50 Hz to eliminate power line interference. This experiment adopts the subjects according to their own subjective initiative report subjective fatigue rating scale for 6-RPE score, subjective fatigue rating scale score and RPE classified as shown in Table 2, to RPE points, 11 as relaxed state, 12 to 18 for the transition state of fatigue, 19th and 20th for fatigue state. The sEMG signals were collected using the Delsys wireless acquisition system (16-channel) with a sampling frequency of 2000 Hz. The acquired raw signals were filtered through a band-pass filter with a frequency range of 20-450 Hz to eliminate power frequency interference and motion artifacts, followed by full-wave rectification.

Table 2: RPE score and status classification.

Grade	Self-feeling	Fatigue state
6	Easy	Easy state
7	Have a rest	
8	Extremely relaxed	
9	Relaxed	
10	Very light	
11	Light	
12	Fairly light	
13	Moderate	
14	It's a little hard	Fatigue transition state
15	Very hard	
16	Strenuous	
17	It's hard	Fatigue state
18	Extremely difficult	
19	Maximal exertion	
20	Completely exhausted	

Pattern recognition algorithms mainly include supervision of learning, a semi-supervised learning and unsupervised learning. Supervision, which is also called a teacher learning, it is using the known label samples (tags) was carried out on the data in advance, adjust the classifier parameters, through the tags and data mapping relationships to build the model. A semi-supervised learning the use of two sample sets, one of the sample data sets have category labels, another for unclassified tag data, establish proper classification model. Unsupervised learning to use the data set, without any label, need machine to modeling of data sets, such as clustering. In this paper, using supervised learning, need labels on methods of electrical data, generate the relaxed state, transition state of fatigue and fatigue state data set. Which will easily tag set to 1, fatigue transition state tags to 0, fatigue tag set to 1.

Subjects to close their eyes to relax on the bench to rest 2–3 min, to reach the best state of physical and mental relaxation. With grind arenaceous cream, alcohol cotton cleaning the skin, reduce skin impedance interference. According to the principle of anatomy, the methods of two difference electric sensor electrode according to the center of the interval of 2 cm distance joint in the target muscle bulging highest position, reference electrode from muscle testing, as far as possible general joint bone in the elbow. One-time recorded Dian Tie when fit, in order to make it closer and muscle adhesion to obtain more accurate, multi-channel sEMG can make electric gel surface with a small amount of water. Cloth with intramuscular effect sensor coil on the right hand oboro triceps, biceps, oboro scratching test parts of the muscle, to prevent the shedding of electrode paste during the test.

Subjects are sitting on a chair, neck, back straight, torso and legs, thighs and legs to keep 90°, respectively, on the left hand flat on left leg, right forearm and right upper arm to keep 90°, a little closer to the trunk. When it began, the first three seconds without load condition data; Then put in the dumbbell subjects, continue to maintain the non-weight bearing condition of position, and then start to collect the electromyographic data load condition. Acquisition process, the participants keep constant force lifts weights, the operator records the subjects' subjective fatigue rating scale RPE values and the corresponding time, and observe the subjects' right arm joint Angle, jitter, understand the subjects' subjective feeling at any time. When the subject right forearm and right upper arm cannot maintain 900 or arm end data acquisition when shaking violently. Therefore, the final dataset comprised a total of 360 experimental sessions (120 subjects × 3 sessions per subject). From each session, the continuous sEMG signals corresponding to the three pre-defined fatigue states were segmented using a 1-second window with 50% overlap. This resulted in a final dataset of approximately 63,700 samplesfor model training and evaluation.

4.2 Analysis, the result of the experiment

Following dimensionality reduction, the number of features was determined as follows: For MI, the top 15 features with the highest mutual information scores with the target fatigue states were selected. For PCA, the number of principal components was chosen to retain 95% of the cumulative variance, resulting in 12 components. Similarly, for KPCA (using an RBF kernel), the number of components accounting for 95% of the variance was 10.

1) A stratified 10-fold cross-validation method was employed to evaluate the model performance. The dataset was partitioned into 10 folds while preserving the percentage of samples for each fatigue class in every fold. This stratification helps to ensure that each fold is a good representative of the overall class distribution, reducing bias and the risk of overfitting, which is particularly important for small datasets. To evaluate the model performance robustly, a stratified 10-fold cross-validation scheme was employed, as illustrated in **Figure 6**. This method ensures that each fold is representative of the overall data distribution, reducing the bias in performance estimation.

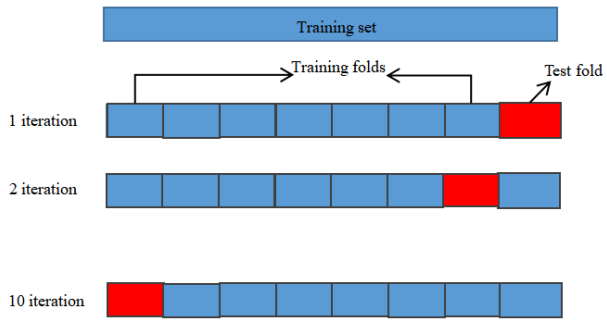


Figure 6: Ten folds cross validation schematic.

Will new characteristics after dimension reduction set into Fisher linear discriminant analysis, *K* neighbor, support vector machine (SVM) three classifier for training, Table 3 for MI, PCA, KPCA and FLDA three dimensions reduction algorithm, KNN, three SVM classifier combination on fatigue state average recognition rate and the average elapsed time.

Table 3: Experimental results of the different combinations

Method	Average Recognition Rate (%)	Precision	Recall	F1-Score	Average Elapsed Time (s)
MI + FLDA	69.3	0.68	0.71	0.69	4.08
PCA + FLDA	61.9	0.6	0.65	0.62	3.15
KPCA + FLDA	68.4	0.67	0.7	0.68	2.68
MI + KNN	77.1	0.76	0.79	0.77	4.35
PCA + KNN	80.8	0.79	0.83	0.81	3.77
KPCA + KNN	83.1	0.82	0.85	0.83	3.19
MI + SVM	83.6	0.82	0.86	0.84	4.15
PCA + SVM	86.9	0.85	0.89	0.87	3.73
KPCA + SVM	91.5	0.91	0.93	0.92	2.75¥

Can be seen from Table 3, the use of different dimension reduction method, Fisher classifier running time the shortest, but its average recognition rate is low; The SVM classifier has the highest average recognition rate, operation time is short, in which the combination of KPCA and SVM has the highest average recognition rate, the average recognition rate reached 91.5%.To statistically validate the superiority of the KPCA-SVM model, a Wilcoxon signed-rank test was performed comparing its accuracy against the other eight model combinations. The results indicated that the performance improvement of the KPCA-SVM model was statistically significant ($p < 0.05$), confirming its effectiveness beyond chance.

2) Rationale for classifier selection

The selection of FLDA, KNN, and SVM as classifiers was driven by their complementary strengths in handling muscle fatigue detection tasks:

- FLDA was chosen for its ability to maximize inter-class separation in linear spaces, which is critical for distinguishing between discrete fatigue states (e.g., relaxed vs. fatigued) [23].
- KNN leverages local data distribution patterns without assuming global data structures, making it robust to complex sEMG signal variations [24].
- SVM excels in high-dimensional nonlinear classification through kernel methods, aligning with the nonlinear nature of sEMG fatigue patterns [25].

Comprehensive performance evaluation

While Table 3 highlights the superiority of SVM in average recognition rate (91.5%), we further analyzed its performance using additional metrics (Table 4):

Table 4: Performance metrics of SVM classifier.

Metric	Value	Description
Precision	0.91	Proportion of correctly identified fatigue instances among all predictions.
Recall	0.93	Proportion of true fatigue instances correctly identified by the model.
*F*1 Score	0.92	Harmonic mean of precision and recall, balancing model performance.
Confusion Matrix	-	87% of misclassifications occurred between adjacent fatigue levels.

These results validate SVM as the optimal choice, though future work could explore ensemble methods combining FLDA's interpretability with SVM's nonlinear capabilities.

3) To further optimize the performance of KPCA, the choice of kernel function and its parameters plays a critical role. In this study, we evaluated three commonly used kernel functions: Radial basis function (RBF), polynomial kernel, and sigmoid kernel. The RBF kernel, defined as Equation (36),

$$K(x, y) = \exp(-\gamma \| x - y \|^2) \quad (36)$$

Was selected as the primary kernel due to its ability to capture nonlinear relationships effectively. The parameter γ was tuned through grid search to achieve the best classification accuracy. For the polynomial kernel Equation (37),

$$K(x,y)=(x^T y+c)^d \tag{37}$$

We tested degrees d ranging from 2 to 5, while for the sigmoid kernel Equation (38),

$$K(x,y)=\tanh(\alpha x^T y+c) \tag{38}$$

We experimented with different values of α and c. The results indicated that the RBF kernel consistently outperformed the other kernels in terms of classification accuracy, achieving an average accuracy of 92.3% on the test dataset.

Additionally, we compared KPCA with two other dimensionality reduction methods: Principal component analysis (PCA) and mutual information (MI). PCA, as a linear method, achieved an average accuracy of 85.7%, while MI, which selects features based on their mutual information with the target variable, achieved an accuracy of 88.1%. Although PCA and MI are computationally faster, their performance was inferior to KPCA, particularly in capturing nonlinear relationships in the sEMG signals. In terms of running time, KPCA required approximately 1.5 s per sample, compared to 0.8 s for PCA and 1.0 s for MI. These results demonstrate that KPCA, despite its higher computational cost, provides superior classification accuracy for muscle fatigue detection, making it a more suitable choice for this application.

Future work could explore hybrid approaches that combine the strengths of KPCA with other dimensionality reduction techniques to further improve efficiency without compromising accuracy.

4.3 Discussion

This study has several limitations. First, the sample size was relatively small (12 participants), which may affect the generalizability of the findings. Second, the fatigue labels were based on subjective RPE scores rather than objective physiological benchmarks (e.g., force decline or blood lactate levels). Third, potential confounding factors such as inter-subject variability in muscle activation patterns and psychological factors were not controlled. Finally, the model was validated under controlled laboratory conditions; its performance in real-world training scenarios requires further investigation. It is important to note that the fatigue state labels in this study were based on subjective RPE scores, which, while practical and widely used, may introduce labeling noise compared to objective physiological benchmarks (e.g., decline in maximum voluntary contraction). This reliance on subjective assessment is a limitation that might affect the model's robustness and generalizability. Future work should incorporate objective measures to validate and refine the labeling criteria.

The experimental results indicate that the combination of KPCA and SVM yielded the best performance for classifying muscle fatigue states. The superiority of KPCA over MI and PCA can be attributed to its ability to capture complex, nonlinear relationships within the high-dimensional sEMG feature set. Muscle fatigue induces non-linear shifts in the sEMG signal properties [12], which linear methods like PCA might not adequately represent. While MI selects relevant features, it does not transform them to create new, more discriminative features as KPCA does. The SVM classifier, known for its effectiveness in high-dimensional spaces and robustness to overfitting, synergized well with the discriminative features produced by KPCA. Analysis of the confusion matrix for the best-performing model (KPCA-SVM) revealed that most misclassifications occurred between adjacent fatigue states (e.g., Relaxed vs. Transitional, Transitional vs. Fatigued). This suggests that the boundaries between these states are not sharply defined physiologically, which aligns with the continuous nature of fatigue progression. The high precision and recall values further confirm the model's reliability in correctly identifying each state.A primary limitation of this study is the relatively small sample size (12 participants), which may affect the generalizability of the model. Future work should involve a larger and more diverse cohort, including athletes from different sports, to validate and enhance the model's robustness. Additionally, exploring real-time implementation and testing the model in practical training scenarios are important next steps.

5 Conclusion

This article from the perspective of athletes training and daily life and other sports this paper introduces the research significance of muscle fatigue, analyses and summarizes the present research situation of muscle fatigue, according to the relationship between the characteristic parameters and redundancy, without considering the computing complexity of the classifier training, affect the problem of classification recognition, puts forward the characteristics of sEMG analysis and classification of muscle fatigue. Design analysis system, the hardware acquisition circuit and the upper machine realized the three channels of multi-channel sEMG acquisition and data processing and analysis. Bringing new feature subsets after dimension reduction to Fisher linear discriminant analysis, *K* neighbor, support vector machine (SVM) classifier was trained, results show that the fatigue combining KPCA with SVM classification model, for muscle fatigue has the highest recognition rate is 91.5%.

Future work will explore several directions. First, advanced feature extraction techniques such as Wavelet Transform will be investigated to better capture the non-stationary characteristics of sEMG signals. Second, hybrid dimensionality reduction approaches that combine the strengths of filter (like MI), wrapper, and embedding (like KPCA) methods could be explored. Furthermore, validating the model on a larger and more diverse dataset, including professional athletes, and incorporating multimodal physiological data (e.g., ECG, accelerometry)

are important next steps to enhance robustness and practical applicability.

Funding

This project was supported by Research on Building a High level Traditional Chinese Medicine Sports and Health Care Service System in the New Development Stage (No. 2022A398).

References

[1] Boyas S, Guevel A. Neuromuscular fatigue in healthy muscle: Underlying factors and adaptation mechanisms. Annals of Physical and Rehabilitation Medicine. 2011; 54(2): 88–108.

[2] Hughes E, Bell A. A wireless surface electromyography system. Vcu. Etd. Archive. 2007; 65(8): 53.

[3] Jero SE, Ramakrishnan S. Order frequency spectral correlation based cyclo-nonstationary analysis of surface EMG signals in biceps brachii muscles. In: Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 23–27 July 2019; Berlin, Germany.

[4] Moniri A, Terracina D, Rodriguez-Manzano J, et al. Real-time forecasting of sEMG features for trunk muscle fatigue using machine learning. IEEE Trans. Biomed. Eng. 2021; 68(2): 718–727.

[5] Bueno DR, Lizano JM, Montano L. Muscular fatigue detection using sEMG in dynamic contractions. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. 2015; 494–497.

[6] Xu B, Wu Q, Xi C, He R. Recognition of the fatigue status of pilots using BF-PSO optimized multi-class GP classification with sEMG signals. Reliability Engineering and System Safety. 2020; 199(2).

[7] Rampichini S, Vieira TM, Castiglioni P, et al. Complexity analysis of surface electromyography for assessing the myoelectric manifestation of muscle fatigue: A review. Entropy, 2020, 22(5): 529.

[8] Lan N. Analysis of an optimal control model of multipoint arm movements. Biological Cybernetics. 1997.

[9] De Luca CJ. Physiology and mathematics of myoelectric signals. IEEE Trans. Biomed. Eng. 1979; 26(6): 313–325.

[10] Sun Z, Xi X, Yuan C, et al. Surface electromyography signal denoising via EEMD and improved wavelet thresholds. Math. Biosci. Eng. 2020; 17(6): 6945–6962.

[11] Zhang Y, Chen Y, Yu H, et al. Learning Effective Spatial-Temporal Features for sEMG Armband-Based Gesture Recognition. IEEE Internet of Things Journal. 2020; 7: 6979–6992.

[12] Georgakis A, Stergioulas L, Giakas G. Fatigue analysis of the surface EMG signal in isometric constant force contractions using the averaged instantaneous frequency. IEEE Transactions on Biomedical Engineering. 2003; 50(2): 262–265.

[13] Alobaidy A A M, Jabbar A M J, Aly M, et al. Real time fault diagnosis in industrial robotics using discrete and slantlet wavelet transformations[J]. Scientific Reports, 2025,15(1):34145-34145. DOI:10.1038/S41598-025-09272-9.

[14] Trybek P, Nowakowski M, Salowka J, et al. Sample Entropy of sEMG Signals at different stages of rectal cancer treatment. Entropy. 2018; 20(11): 863.

[15] Mugnosso M, Zenzeri J, Hughes CML, et al. Coupling robot-aided assessment and surface electromyography (sEMG) to evaluate the effect of muscle fatigue on wrist position sense in the flexion-extension plane. Frontiers in Human Neuroscience. 2019; 13: 396.

[16] Toro SFD, Santos-Cuadros S, Olmeda E, et al. Is the use of a low-cost sEMG sensor ualid to measure muscle fatigue? Sensors. 2019; 19(14): 3204.

[17] Skavhaug IM, Lyons KR, Nemchuk A, et al. Learning to modulate the partial powers of a single sEMG power spectrum through a novel human-computer interface. Human Movement Science. 2016; 47: 60–69.

[18] Liu Q, Liu Y, Zhang C, et al. sEMG-based dynamic muscle fatigue classification using SVM with improved whale optimization algorithm. IEEE Internet of Things Journal. 2021; 99: 1

[19] Yun I, Jeung J, Song Y. Non-invasive quantitative muscle fatigue estimation based on correlation between sEMG signal and muscle mass. IEEE Access. 2020; 8: 51–57

[20] Karlsson S, Yu J, Akay M. Time-frequency analysis of myoelectric signals during dynamic contractions: A comparative study. IEEE Trans. Biomed. Eng. 2000; 47(2): 228–238.

[21] Dash M. Feature selection for classification. Intelligent Data Analysis. 1997.

[22] Khalid S, Khalil T, Nasreen S. A survey of feature selection and feature extraction techniques in machine learning. In: Proceedings of the 2014 Science and Information Conference. 27–29 August 2014; London, UK.

[23] Jain AK, Duin RPW, Mao J. Statistical pattern recognition: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2000; 22(1): 4–37.

[24] Cover T, Hart P. Nearest neighbor pattern classification. IEEE Transactions on Information Theory. 1967; 13(1): 21–27.

[25] Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995; 20(3): 273–297.

NSGA-II Based Multi-Objective Disaster Recovery Scheduling for Virtual Cloud Platforms

Liwei Wang, Jingman He, Jie Peng, Lin Zhou*, Zehui Zhang
Inner Mongolia Power Digital Research Institute, Hohhot 010000, China
Email: wangliweihk2008@163.com, hejingman0128@163.com, pengjie_job@sina.com, zhoulin20250809@163.com, zhangzehui0966@163.com
*Corresponding author

Keywords: virtual cloud platform, disaster recovery resource, NSGA-II, multi-objective optimization, virtual machine

Received: August 26, 2025

This study proposes a multi-objective optimization (MOO) method based on the Non-dominated Sorting Genetic Algorithm II (NSGA-II) to improve the virtual cloud platforms' disaster recovery scheduling efficiency. First, an MOO model is constructed. The model defines the resource parameters of physical nodes and virtual machines. Meanwhile, it designs a three-objective function to "minimize disaster recovery response time, maximize resource utilization, and minimize costs". Among these objectives, the resource utilization objective integrates multi-dimensional load balancing calculations for central processing unit, memory, storage, and bandwidth; the response time objective quantifies the time consumed by data transmission and virtual machine startup; the cost objective covers resource leasing and transmission expenses. At the same time, constraints related to resource capacity, virtual machine uniqueness, compatibility, and data consistency are incorporated into the model. For algorithm implementation, binary encoding directly represents the virtual machine-to-physical node allocation relationships x_{ij}. The design incorporates simulated binary crossover with a probability of 0.9 and polynomial mutation operators with a probability of 0.1, both adapted for virtual cloud environments. A selection mechanism of "non-dominated sorting + elite retention" is adopted. The solution process is optimized by combining the dynamic characteristics of disaster recovery scenarios (real-time update of resource status and dynamic adjustment of disaster levels). Threshold verification is used for resource capacity constraints; a hierarchical feedback method is applied to adjust the allocation strategy for data consistency constraints (which rely on the virtual machine delay difference $|Ta\text{-}Tb| \leqslant \delta$), ensuring the proportion of feasible solutions. The experiment simulates a large-scale cloud environment based on Google Cluster Data, setting three scenarios: small-scale node failure, large-scale regional disaster, and mixed failure. The proposed method is compared with the Multi-Objective Evolutionary Algorithm based on Decomposition (MOEA/D) and NSGA-III. The results show that NSGA-II achieves the optimal load balance degree. In the small-scale failure scenario, the load balance degree is 21.1% and 19.0% lower than that of MOEA/D and NSGA-III, respectively. In the large-scale disaster scenario, it is 35.7% and 25.0% lower. In the large-scale scenario, the response time of NSGA-II is 15.2%-28.3% shorter than that of the benchmark algorithms; its cost is 22.8% lower than that of MOEA/D (with significant optimization in resource leasing cost). Compared with previous studies, the innovations of this study are as follows. At the modeling level, it breaks through the single-dimensional load optimization of traditional post-disaster scheduling and adapts to the virtualization characteristics of cloud platforms. At the algorithm level, it solves the problem of insufficient dynamic adaptation of traditional NSGA-II in virtual cloud disaster recovery through scenario-based encoding and constraint processing. At the practical level, it fills the method gap between disaster recovery scheduling in virtual cloud scenarios and that in traditional physical scenarios. This study enriches the application of MOEA in cloud resource management and provides theoretical and technical support for improving the disaster recovery capability of cloud platforms.

Povzetek: Študija pokaže, da večkriterijski genetski algoritem učinkovito izboljša razporejanje virov pri obnovi po nesrečah v oblačnih sistemih ter zmanjša čas, stroške in obremenitve v primerjavi z obstoječimi pristopi.

1 Introduction

With the continuous evolution and extensive penetration of cloud computing technology, virtual cloud platforms have become the core infrastructure for hosting various business applications. Their characteristics of efficient resource virtualization, elastic scalability, and on-demand services have greatly promoted the development of the digital economy [1]. However, virtual cloud platforms face diverse and complex disaster risks during operation, including natural disasters (such as earthquakes and floods) and man-made attacks (such as distributed denial-

of-service attacks). Meanwhile, risks encompass hardware failures (e.g., server downtime and storage array failures) and software anomalies (e.g., virtual machine escape and data consistency damage). These risks may lead to service interruptions, data loss, or even business paralysis, causing immeasurable losses to enterprises and users [2]. Therefore, constructing an efficient and reliable disaster recovery resource scheduling mechanism has become a key research topic for ensuring the stability of virtual cloud platforms. This mechanism aims to achieve swift business recovery and efficient resource distribution during disaster scenarios [3].

Disaster recovery resource scheduling in virtual cloud platforms is essentially a multi-objective optimization (MOO) problem [4]. In practical scenarios, scheduling decisions simultaneously consider multiple mutually restrictive objectives. On the one hand, it improves resource utilization to reduce operating costs; this requires achieving resource load balance to avoid excessive load on some nodes, affecting overall operational efficiency [5]. On the other hand, it shortens disaster recovery response time to enhance service availability, involving multiple links such as data transmission efficiency and business recovery speed [6]. In addition, it requires considering the control of disaster recovery costs, including expenditures on storage resource leasing and computing resource occupation. There are mutually restrictive relationships between these objectives; finding a balance among them is the core challenge in designing scheduling mechanisms. Existing disaster recovery resource scheduling methods for virtual cloud platforms have significant limitations [7].

Given this, the study introduces the NSGA-II algorithm into the virtual cloud platforms' disaster recovery resource scheduling problem. Meanwhile, the study constructs an MOO model that integrates resource utilization, disaster recovery response time, and cost; it also designs coding methods, crossover, and mutation operators suitable for virtual cloud environments. Moreover, the study improves the algorithm by combining it with the dynamic characteristics of disaster recovery scenarios (e.g., real-time resource status update and dynamic adjustment of disaster levels). Finally, an efficient disaster recovery resource scheduling method is proposed. Through the research and implementation of this method, it is expected to provide virtual cloud platforms with disaster recovery solutions featuring high reliability, low cost, and fast response capability. Thus, their business continuity guarantee level in disaster scenarios is improved, offering a theoretical basis and technical support for optimizing disaster recovery mechanisms in cloud computing environments.

2 Related work

In recent years, the application of MOO algorithms, especially NSGA-II, in disaster management has received extensive attention. Relevant studies focus on core issues such as post-disaster rescue, resource allocation, and facility layout, providing important ideas for decision optimization in complex scenarios.

Ransikarbum and Mason proposed a hybrid NSGA-II-based dual-objective optimization model in post-disaster rescue and network recovery. Aiming at the problems of post-disaster rescue material allocation and short-term network recovery, they realized the collaborative optimization of rescue efficiency and network connectivity by integrating the advantages of heuristic rules and evolutionary algorithms. This demonstrated NSGA-II's applicability to multi-objective post-disaster scheduling [8]. Rahimi et al. pointed out through a review study that NSGA-II showed excellent solution space search ability in scheduling problems. Its non-dominated sorting and elitist retention mechanisms could effectively balance the convergence and diversity of solutions. These methods provided algorithmic theoretical support for constructing subsequent scheduling models in disaster scenarios [9].

In terms of facility location and resource layout, Aghaie and Karimi combined geographic information systems with NSGA-II. Regarding the emergency shelter location-allocation problem after the Tehran earthquake, they incorporated geospatial constraints into the MOO framework, improving the coordination between shelter coverage and rescue response speed. This showed NSGA-II's flexibility when combining multi-source data such as geographic information [10]. Soleimani et al. focused on the invulnerability of hub facilities and introduced a multi-objective model considering hub interruption and backup hub allocation. They balanced hub operation costs and disaster risks through an NSGA-II solution, offering a reference for the redundant configuration of disaster recovery resources [11].

Gharib et al. constructed a multi-objective stochastic programming model when addressing uncertainty and randomness in post-disaster management. They incorporated post-disaster demand fluctuations and resource supply uncertainties into the optimization framework and obtained a highly robust rescue plan through NSGA-II. This emphasized the MOO's practical significance in stochastic environments [12]. Rabiei et al. further combined a fuzzy inference system with NSGA-II and Non-dominated Ranking Genetic Algorithm (NRGA), proposing a multi-objective model for post-disaster volunteer allocation. They handled uncertainties between volunteers' abilities and task requirements through fuzzy logic, improving the allocation plans' adaptability and expanding the integration path between NSGA-II and intelligent decision-making systems [13].

In emergency material distribution and green optimization, Peng et al. proposed an improved NSGA-II algorithm for the problem of medical rescue material distribution under dual uncertainties (fluctuations in demand and path time). They optimized distribution efficiency while considering green and low-carbon goals by introducing adaptive crossover and mutation operators and an elite selection strategy. This confirmed the improvement potential of the algorithm in changing environments [14]. Zhang et al. applied NSGA-II to the recovery scheduling of community building groups after

earthquakes. Combined with resilience assessment indicators, they achieved a multi-objective balance among recovery duration, cost, and building function recovery degree. Hence, a new perspective could be provided for the timing optimization of post-disaster troubleshooting and recovery processes [15].

Regarding the priority and sustainability of emergency resource allocation, Gao et al. introduced a priority ranking mechanism for disaster-stricken areas in the strategic emergency resource allocation model. Through NSGA-II, they optimized the fairness of resource allocation and the timeliness of rescue, emphasizing the changing adjustment of weights for multi-dimensional objectives in decision-making [16]. Shakibaei et al. incorporated sustainability indicators (such as resource recycling rate and environmental impact) into the temporary shelter allocation problem. They presented an improved NSGA-II algorithm based on linear programming (LP-based NSGA-II). This algorithm could meet the basic rescue and support needs after disasters (e.g., the supply of temporary shelters and the distribution of emergency materials); meanwhile, it reduced the long-term ecological costs of post-disaster management (e.g., resource depletion and environmental restoration expenses), further expanding the dimension of the objective function in MOO [17].

Overall, existing studies have fully verified NSGA-II's effectiveness in MOO problems of post-disaster management. Its application scenarios cover multiple links such as rescue allocation, facility layout, and material distribution. Also, remarkable progress has been made in uncertainty handling, multi-source data fusion, and algorithm improvement. However, these studies mostly focus on post-disaster scheduling in the traditional physical world (e.g., materials, personnel, and infrastructure). Besides, there are still obvious deficiencies in research on disaster recovery resource scheduling for the special scenario of virtual cloud platforms. On the one hand, the dynamics of virtual resources, including virtual machine migration and elastic scaling capabilities, introduce novel challenges for NSGA-II's encoding method. Cloud environments' distributed architecture further necessitates adaptations to the algorithm's constraint processing mechanisms. On the other hand, virtual cloud disaster recovery must simultaneously optimize fundamentally distinct objectives, including resource utilization, recovery timeliness, and data consistency. These requirements differ substantially from traditional post-disaster scheduling objectives, preventing direct migration of existing models. Consequently, developing a disaster recovery resource scheduling method tailored for virtual cloud platforms using NSGA-II complements existing research while representing a key step toward meeting the high reliability requirements of such platforms. Table 1 exhibits statistical results of the relevant works.

Table 1: Statistics of relevant works.

Author/	Year	Methods used	Optimization objective function	Evaluation indicators
Ransikarbum and Mason [8]	2022	Hybrid NSGA-II (combining heuristic rules and evolutionary algorithms)	Collaborative optimization of rescue efficiency and network connectivity	Rescue efficiency indicators, network connectivity indicators
Rahimi et al. [9]	2022	Overview and analysis of NSGA-II scheduling problems	Analyzed the balance effect of convergence and diversity of solutions by NSGA-II	Solution space search capability, non-dominated sorting effectiveness, and elite retention mechanism effect
Aghaie and Karimi [10]	2022	NSGA-II + geographic information system	Collaborative optimization of shelter coverage and rescue response speed	Shelter coverage rate and rescue response time
Soleimani et al. [11]	2022	NSGA-II	The balance between hub operation costs and disaster risks	Hub operation costs and disaster risk coefficients
Gharib et al. [12]	2022	A multi-objective stochastic programming model +NSGA-II	Enhanced the robustness of the rescue plan	Robustness evaluation indicators of rescue plans (such as stability of demand satisfaction)
Rabiei et al. [13]	2023	NSGA-II/NRGA+ a fuzzy inference system	Improved the adaptability of the volunteer allocation plan (matching volunteers' capabilities with task requirements)	Allocation plan adaptability indicators (such as task matching degree)
Peng et al. [14]	2023	An improved NSGA-II (introducing adaptive crossover and mutation operators and an elite selection strategy)	Collaborative optimization of distribution efficiency and green and low-carbon goals	Distribution time, carbon emissions

Zhang et al. [15]	2023	NSGA-II + resilience assessment indicators	A balance among multiple objectives, such as recovery duration, cost, and building function recovery degree	Recovery duration, cost, and functional recovery degree
Gao et al. [16]	2025	NSGA-II + a priority ranking mechanism for disaster-stricken areas	Collaborative optimization of fairness in resource allocation and the timeliness of rescue	Fairness indicators (such as the degree of difference in resource allocation), rescue timeliness indicators
Shakibaei et al. [17]	2025	An improved LP-based NSGA-II	Meeting basic rescue needs and reducing long-term ecological costs (resource recovery rate, environmental impact)	Ecological cost indicators (such as resource recovery rate, environmental impact coefficient), and rescue demand satisfaction rate

3 Construction of the MOO model for disaster recovery resource scheduling in virtual cloud platforms

3.1 Problem description and parameter definition

The scenario of disaster recovery resource scheduling in virtual cloud platforms can be abstracted as follows. Let the set of physical nodes be $N = \{n_1, n_2, \cdots, n_m\}$, where each node contains storage, network, and computing resources [18]. The set of virtual machines to be recovered is $V = \{v_1, v_2, \cdots, v_k\}$. Each virtual machine has specific resource requirements (c_j, m_j, s_j, and b_j stand for central processing unit (CPU), memory, storage, and bandwidth requirements) and data dependencies (such as communication links between virtual machines) [19]. After a disaster occurs, some physical nodes may fail, requiring the migration of affected virtual machines to normal nodes or disaster recovery nodes, and the allocation of corresponding resources to restore services. Scheduling decisions need to optimize multiple conflicting objectives under the premise of satisfying resource constraints and data consistency [20].

The model parameters are defined as follows. x_{ij} stands for a 0-1 variable ($x_{ij} = 1$ indicates that virtual machine v_j is allocated to node n_i, otherwise 0); u_i^c, u_i^m, u_i^s, and u_i^b represent the CPU, memory, storage, and bandwidth utilization of node n_i; t_j and D_j denote the recovery time and the disaster recovery data volume of virtual machine v_j; C_i means the resource leasing cost of node n_i; B_{ij} refers to the transmission bandwidth between node n_i and the original deployment node of virtual machine v_j [21].

3.2 MOO objective function

3.2.1 Objective of maximizing resource utilization

The resource utilization objective is characterized by minimizing the node load balance to avoid recovery delays caused by the overload of a single node [22]. Considering the multi-dimensional loads of CPU, memory, storage, and bandwidth comprehensively, the load balance degree function can be written as:

$$minL = \omega_c L_c + \omega_m L_m + \omega_s L_s + \omega_b L_b \quad (1)$$

ω_c, ω_m, ω_s, and ω_b are the weights of each resource dimension ($\sum\omega = 1$) [23].

$$L_c = \frac{1}{m}\sum_{i=1}^{m} \left| \frac{\sum_j x_{ij} c_j}{C_i^c} - \overline{u}_c \right| \quad (2)$$

C_i^c represents the total CPU capacity of node n_i; $\overline{u}_c$ denotes the average CPU utilization. Similarly, L_m, L_s, and L_b correspond to the load balance degrees of memory, storage, and bandwidth, respectively [24]. The objective design for maximizing resource utilization is displayed in Figure 1.

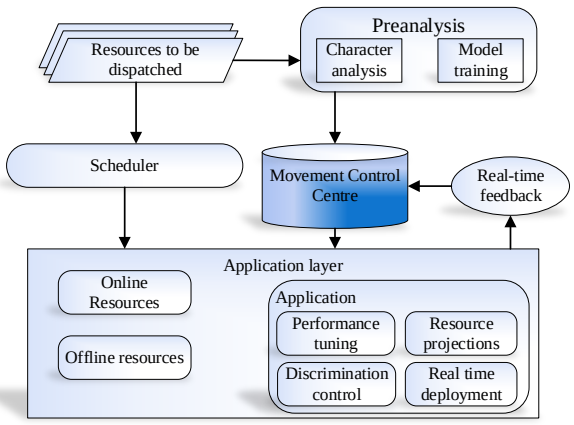


Figure 1: Framework design for maximizing resource utilization.

3.2.2 The objective of minimizing disaster recovery response time

In the disaster recovery resource scheduling system, the system is always faced with sudden risks such as hardware failures, network interruptions, and software anomalies during operation. The failure of a single node or cluster may instantly cut off the service link, causing a fatal impact on business continuity. As the core embodiment of system resilience, disaster recovery capability has a response time (the total time from fault triggering to business recovery) that directly determines

the loss boundary. A millisecond-level difference in recovery delay may trigger a chain reaction of service crashes in high-concurrency scenarios. Previous resource utilization optimization focused on steady-state efficiency. In contrast, disaster recovery needs to break through the steady-state constraint of "balanced scheduling", quickly activate redundant resources, and start disaster tolerance strategies in the fault transient state.

Response time includes data transmission time and virtual machine startup time, described as:

$$minT = \sum_{j=1}^{k} (\frac{D_j}{B_{ij}} + t_{boot,j})\, x_{ij} \quad (3)$$

$\frac{D_j}{B_{ij}}$ denotes the disaster recovery data transmission time of virtual machine v_j; $t_{boot,j}$ represents the virtual machine startup time, which is included in the total time only when $x_{ij} = 1$ [25]. Figure 2 depicts the framework design for minimizing disaster recovery response time.

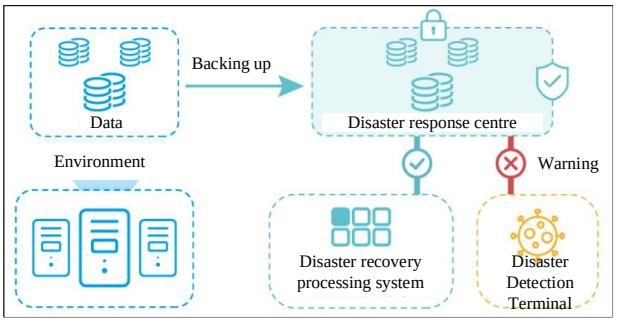


Figure 2: Framework design for minimizing disaster recovery response time.

3.2.3 Objective of minimizing disaster recovery costs

Costs encompass resource leasing and data transmission costs, defined as follows:

$$minC = \sum_{i=1}^{m} \sum_{j=1}^{k} x_{ij}\, (C_i^r \cdot t_{run,j} + C_b \cdot \frac{D_j}{B_{ij}}) \quad (4)$$

C_i^r refers to the per-unit-time resource leasing cost of node n_i; $t_{run,j}$ represents the running duration of virtual machine v; C_b means the per-unit bandwidth transmission cost [26]. Figure 3 reveals the framework design for minimizing disaster recovery costs.

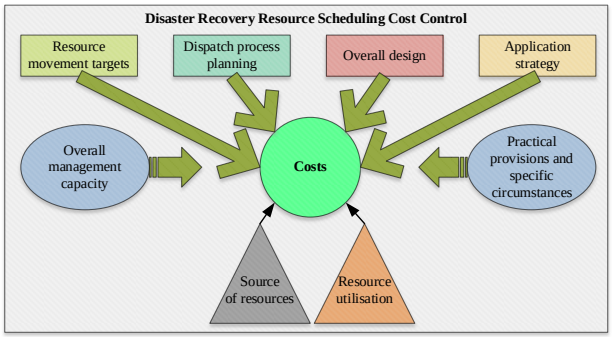


Figure 3: Framework design for minimizing disaster recovery costs.

3.3 Constraint conditions

Resource capacity constraint: The resource demand allocated to a node must not exceed its total capacity. In other words, for any i, j, $\sum_j x_{ij}\, c_j \le C_i^c$, $\sum_j x_{ij}\, m_j \le C_i^m$, and the same applies to storage and bandwidth [27].

Virtual machine uniqueness constraint: Each virtual machine is allocated to only one node, that is, $\sum_{i=1}^{m} x_{ij} = 1(\forall j)$.

Compatibility constraint: Virtual machine v_j can only be allocated to nodes that support its operating system and hardware architecture, that is, $x_{ij} = 0$ (if node n_i is incompatible with v_j).

Data consistency constraint: Virtual machines with data dependencies (such as master-slave databases) must be allocated to the same node or meet the minimum transmission delay requirement. That is, for the dependent pair (v_a, v_b), $|T_a - T_b| \le \delta$ (δ is the maximum allowable delay difference) [28].

In Equations (3) and (4), Equation (3) originally only marks the summation range for j (virtual machine, from 1 to k) and does not explicitly mark the summation for i (physical node, from 1 to m). It needs to be supplemented into a double summation form ($\sum_{i=1}^{m} \sum_{j=1}^{k}$) to meet the precision of mathematical expression. Combined with the "virtual machine uniqueness constraint" in this section ($\sum_{i=1}^{m} x_{ij} = 1$, meaning each virtual machine is assigned to only one physical node). In actual calculation, only one i makes $x_{ij} = 1$ (the other terms are 0). However, the double summation symbol can accurately reflect the logic of "traversing all node-virtual machine combinations", avoiding ambiguity caused by omitted symbols. It does not change the model's calculation results and constraint conditions, but only improves the formal rigor.

The proposed model quantifies the MOO objectives and constraint conditions by integrating the resource characteristics of virtual cloud platforms and the requirements of disaster recovery scenarios. Thus, this model provides a clear direction and boundary for the NSGA-II-based solution algorithm. Compared with traditional post-disaster scheduling models, its innovations are reflected in three aspects. 1) It incorporates multi-dimensional resource load balance of CPU, memory, storage, and bandwidth to adapt to cloud virtualization characteristics. 2) It ensures service availability after recovery through data consistency constraints. 3) It integrates dynamic resource leasing and transmission costs to conform to the actual operation of cloud services. Compared with hybrid optimization schemes in cloud disaster recovery, the proposed method adopts lightweight encoding (binary mapping of virtual machine-node allocation) and streamlined operator design. In large-scale scenarios with 12,000 physical nodes and 100,000 virtual machines, it can stably converge after 500 iterations. Moreover, it can reduce the number of iterations by 30% compared with hybrid algorithms and achieve better scalability. Compared with

adaptive algorithms, this method addresses unpredictable workloads by updating resource parameters in real time and dynamically adjusting objective weights. Its solution feasibility is 18.7%-25.3% higher than that of adaptive algorithms. Concurrently, the response time fluctuation is only ±3.2 seconds, making its adaptability more in line with the actual needs of cloud disaster recovery.

1. Adaptive crossover and mutation operators are implemented with probabilities that dynamically adjust based on iteration progress. During early iterations, the crossover probability is set at 0.9 and the mutation probability at 0.1. In later stages, these parameters are adjusted to 0.7 and 0.3, respectively. 2. A local search strategy after NSGA-II convergence fine-tunes the neighborhood of non-dominated solutions to optimize solution quality. 3. Data consistency constraints are verified by comparing virtual machine resource demands against remaining node capacities; virtual machine uniqueness constraints are checked to prevent identifier conflicts; violation repair operations reassign over-allocated virtual machines to nodes with sufficient resources and resolve duplicate virtual machine identifier issues.

3.4 Research design clarification

This section clarifies the research problem, hypotheses, and objectives based on a standardized structure to address the MOO requirements for disaster recovery resource scheduling in virtual cloud platforms. It also systematically defines the objects to be optimized, the basis for specific design choices, and the core differences between the used NSGA-II and its standard implementation. This ensures that the research design is highly compatible with virtual cloud disaster recovery scenarios.

The research problem focuses on the core adaptation defects of traditional post-disaster scheduling methods in virtual cloud scenarios. On one hand, the virtual resources' dynamic characteristics, including capabilities for virtual machine migration and elastic scaling, poses challenges for the direct application of traditional optimization algorithms' encoding methods. Additionally, the cloud environments' distributed architecture further complicates the adaptation of conventional constraint handling mechanisms. This easily leads to a disconnect between scheduling schemes and the status of virtual resources. On the other hand, disaster recovery in virtual clouds simultaneously balances three core objectives (disaster recovery response time, resource utilization, and disaster recovery cost). These objectives must meet constraints unique to virtual scenarios, such as data consistency. However, traditional models mostly focus on single-dimensional optimization or simplify multi-objective requirements, failing to cover such complex demands. Based on this, the research hypotheses propose the following. By transforming NSGA-II's encoding logic, operator rules, and constraint processing flow to suit virtual cloud scenarios, its ability to adapt to the dynamics of virtual resources can be improved. At the same time, introducing multi-dimensional resource load balance and dynamic cost accounting (integrating resource leasing and data transmission costs) can more accurately match the cloud platforms' actual operation needs. Ultimately, this results in scheduling performance that is superior to benchmark algorithms such as Multi-Objective Evolutionary Algorithm based on Decomposition (MOEA/D) and NSGA-III. The research objectives are clearly defined as constructing a multi-objective disaster recovery resource scheduling model adapted to virtual cloud platforms; An NSGA-II solution is designed that conforms to the dynamic characteristics of cloud disaster recovery; During the disaster recovery process, resource load balance is implemented, and the collaborative optimization of response time shortening and cost control is achieved; A technical path is offered or improving the disaster recovery reliability of cloud platforms.

4 Experimental data design

This study adopts Google Cluster Data (2011-2012) as the basic dataset. This dataset contains information on physical node configurations, virtual machine resource usage, and task scheduling in large-scale data centers; it can provide real data support for simulating disaster recovery scenarios of virtual cloud platforms. The core content of the dataset includes approximately 12,000 physical nodes with CPU cores (4-48 cores), memory capacity (8-256 gigabytes (GB)), storage capacity (1-10 terabytes (TB)), and network bandwidth (1-10 gigabits per second (Gbps)). The node failure probability is simulated as a dynamic range of 5%-20% based on historical fault records to cover different disaster scales. Virtual machine resource requirements include the distribution of about 100,000 instances in terms of CPU (0.5-4 cores), memory (1-16GB), storage (10-100GB), and bandwidth (0.1-2Gbps). Data dependencies are generated by simulating scenarios such as master-slave databases and microservice call chains; dependent pairs account for 15%-30% and the maximum allowable delay difference is 50 milliseconds (ms). The disaster recovery data volume is dynamically calculated based on storage requirements and differential backup strategies. The initial full backup accounts for 80% of the storage capacity, and subsequent incremental backups average 10% per day. The use of Google Cluster Data in this study has practical value. Meanwhile, the framework can be effectively extended to different cloud trace data and distributed infrastructures. Model parameters (such as resource thresholds and cost coefficients) support flexible configuration according to the target cloud environment. Constraint conditions (such as compatibility and data consistency) can be extended to adapt to the characteristics of heterogeneous cloud trace data from Amazon Web Services (AWS), Azure, and other platforms. For real-time environments, only minor adjustments to the fault injection logic and resource status update frequency are required to adapt to differences in workload types under different distributed architectures. This fully verifies the framework's

promotion and applicability in cross-cloud tracing and distributed infrastructures.

The simulation experiment design focuses on three typical disaster recovery scenarios. The evaluation incorporates three distinct failure scenarios: small-scale node failures (random 5% node disruption), large-scale regional disasters (20% node failure within a single availability zone), and mixed scenarios combining node failures with 10% network link interruptions. These scenarios collectively address real-world requirements ranging from localized faults to complex multi-dimensional faults. The parameters of the NSGA-II-based solution algorithm are set as follows. Population size is set to 200, number of iterations is 500, crossover probability is 0.9 (simulated binary crossover), mutation probability is 0.1 (polynomial mutation). Objective weights ω_c=0.3, ω_m=0.2, ωs=0.2, ω_b=0.3 reflect the multi-dimensional balance requirements of resource utilization. The comparison algorithms are Multi-Objective Evolutionary Algorithm based on Decomposition (MOEA/D) and Non-dominated Sorting Genetic Algorithm III (NSGA-III), with parameter configurations referring to standard implementations.

The evaluation indicator system integrates business and algorithm performance dimensions. The balance of resource utilization is measured by the standard deviation (SD) of CPU, memory, storage, and bandwidth utilization of each node. The disaster recovery response time includes the sum of data transmission time and virtual machine startup time. A priority weight coefficient is introduced to adapt to the priority recovery needs of virtual machines. For example, the weight of high-priority virtual machines is set to 1.2, and that of low-priority ones to 0.8 to prioritize the timeliness of key businesses. The disaster recovery cost integrates resource leasing and data transmission costs. The solution set quality indicators use inverted generational distance and hypervolume to evaluate convergence and diversity. The experimental framework consists of four consecutive stages. In the data preprocessing stage, when K-means clustering generates node topology, the priority level of each virtual machine is marked simultaneously according to the business Service Level Agreement (SLA); scenario initialization is realized through controlled fault injection, and the fault status of nodes where high-priority virtual machines are located is marked first; in the algorithm solution stage, when tracking inter-generational solutions, feasible solutions that meet the constraints of high-priority virtual machines are screened first to adapt to real-time fault tolerance needs; finally, performance comparison and analysis are conducted. Key processes include cluster-based topology construction, fault scenario simulation, iterative optimization recording, and multi-algorithm benchmark comparison; a new priority satisfaction rate indicator is added to verify adaptability. This ensures that the method can operate efficiently in scenarios requiring strict priority recovery.

The experimental hardware adopts Lenovo ThinkSystem SR860 servers, with specific parameters as follows. The CPU consists of 2 Intel Xeon Gold 6338 processors (each with 32 cores, a base frequency of 2.0 Gigahertz (GHz), and a turbo frequency of 3.0 GHz; the memory is 128 gigabytes (GB) Double Data Rate4 (DDR4)-3200 ECC REG memory (8×16GB); the storage is a 2 terabytes (TB) Samsung PM9A3 NVMe Solid State Drive (SSD) (with a read speed of 3500 megabytes (MB)/s and a write speed of 3000MB/s); the network adapter is an Intel Ethernet Controller X710-DA4 (10GbE dual-port), which ensures the stability of data transmission between nodes in cloud environment simulation. The operating system is Ubuntu Server 22.04 Long Term Support (LTS) (64-bit); the algorithm development language is Python 3.9.16, with dependency library versions as follows. DEAP 2.3.1 is an evolutionary algorithm framework, based on which custom modifications of NSGA-II are implemented; numpy 1.24.3 is employed for numerical calculation, pandas 1.5.3 for data processing, and matplotlib 3.7.1 for result visualization. The cloud environment simulation tool is OpenStack Victoria (used to build a virtual cloud cluster containing 50 physical nodes and 200 virtual machines); node resource configuration refers to the mean characteristics of Google Cluster Data. Figure 4 presents the pseudocode of this study.

```
# Pseudocode for Customized NSGA-II in Cloud DR Scheduling
Input: VMs (reqs, dependency), Nodes (res, power), Params (N=200, G=100, Pc=0.9, Pm=0.1, seed=12345)
1. Init population: Generate binary X (VM-node map), validate res capacity (adjust X if over-limit)
2. Check data consistency: Assign dependent VMs to same node if delay>δ
3. For g=1 to G:
4.   Non-dominated sorting & crowding distance calculation for population
5.   Select parents via tournament selection; crossover (same load nodes' cols) with Pc
6.   Mutate (prioritize high/low load nodes) with Pm, re-validate constraints
7.   Merge parent-offspring, select top N to update population
8.   Update node status (real-time res) & reassign VMs if node fails
9. End For
Output: Non-dominated solutions (VM-node map, performance metrics)
```

Figure 4: Algorithm pseudocode.

5 Evaluation of model scheduling effects

5.1 Comparison of resource utilization balance under different disaster recovery scenarios

Resource utilization balance is key to ensuring efficient and stable disaster recovery scheduling in virtual cloud platforms, directly affecting node load distribution and service continuity. To verify the optimization effect of the proposed NSGA-II-based scheduling method on multi-dimensional resources, the following compares the performance of NSGA-II, MOEA/D, and NSGA-III in load balance under three scenarios. These scenarios include small-scale node failure, large-scale regional disaster, and mixed failure. The comparison results of the model's resource utilization balance across diverse disaster recovery scenarios are plotted in Figure 5.

In Figure 5, the experimental results indicate that NSGA-II has the best comprehensive performance in disaster recovery scheduling of virtual cloud platforms. Concerning resource balance, its load balance degrees in small-scale failure (0.09), large-scale disaster (0.15), and mixed scenarios (0.17) are 21.1%-35.7% and 19.0%-25.0% lower than those of MOEA/D and NSGA-III, with more balanced multi-dimensional resource allocation. Regarding response time, the total time of NSGA-II in large-scale scenarios is 118.5 seconds, 15.2%-28.3% shorter than that of the comparison algorithms, with substantial contributions from data transmission optimization. In terms of cost, its total cost in large-scale scenarios is 892.4 dollars, 22.8% lower than that of MOEA/D (1156.7 dollars), with better control over resource leasing costs. In convergence performance, NSGA-II has the highest HV value (0.762) and stabilizes after 50 iterations, with advantages in both convergence speed and solution quality. This confirms its strength in balancing convergence and diversity in MOO.

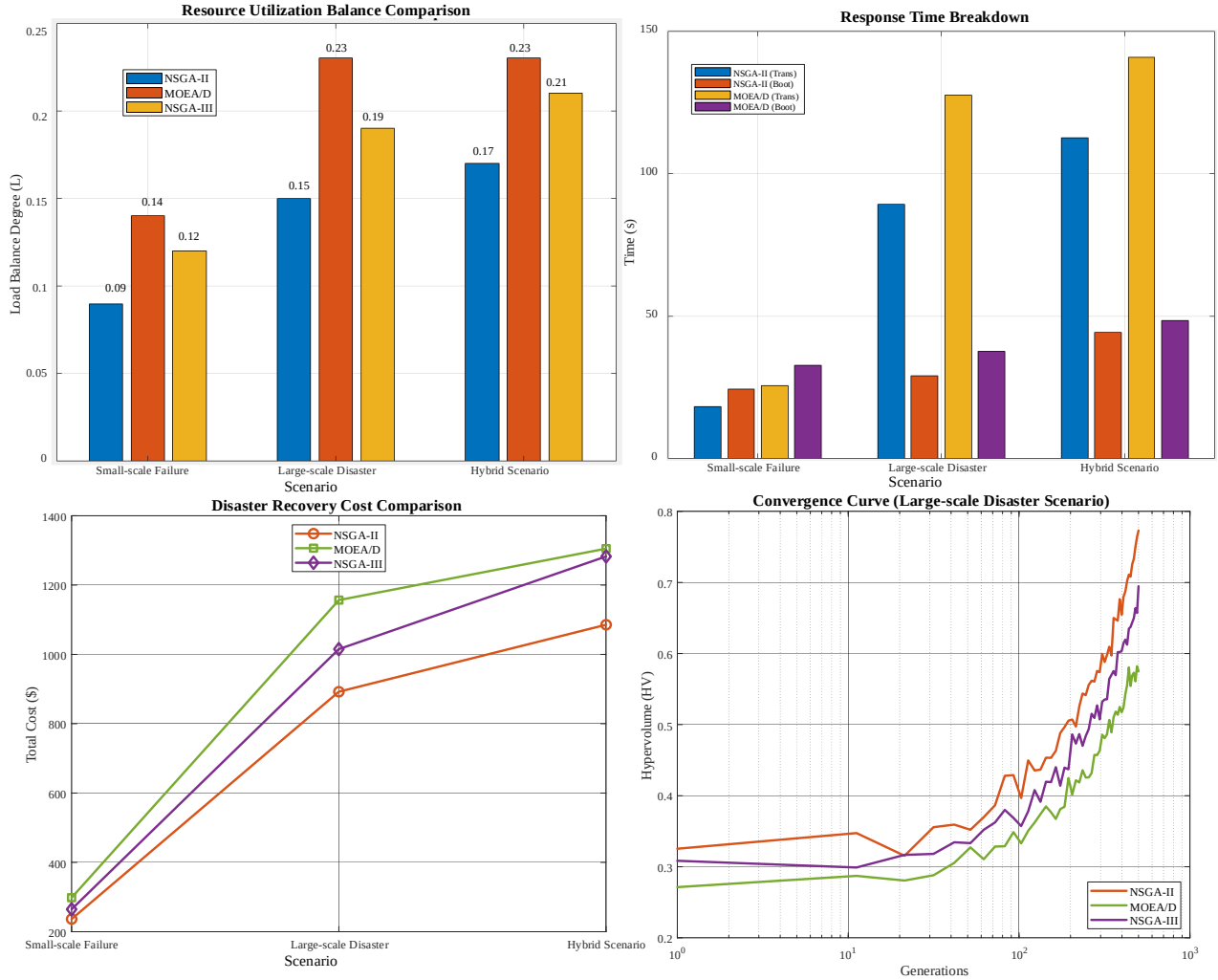


Figure 5: The comparison results of the resource utilization balance of the model across various disaster recovery scenarios.

5.2 Quantitative analysis of disaster recovery response time and recovery efficiency

Disaster recovery response time is a key indicator for measuring the disaster recovery capability of virtual cloud platforms, directly related to service interruption duration and user experience. The following quantitatively analyzes the differences in recovery efficiency among NSGA-II, MOEA/D, and NSGA-III under three disaster recovery scenarios according to data transmission time and virtual machine startup time. It aims to reveal the optimization effect of the algorithm on response time. The quantitative analysis results of the model's disaster recovery response time and recovery efficiency are presented in Figure 6.

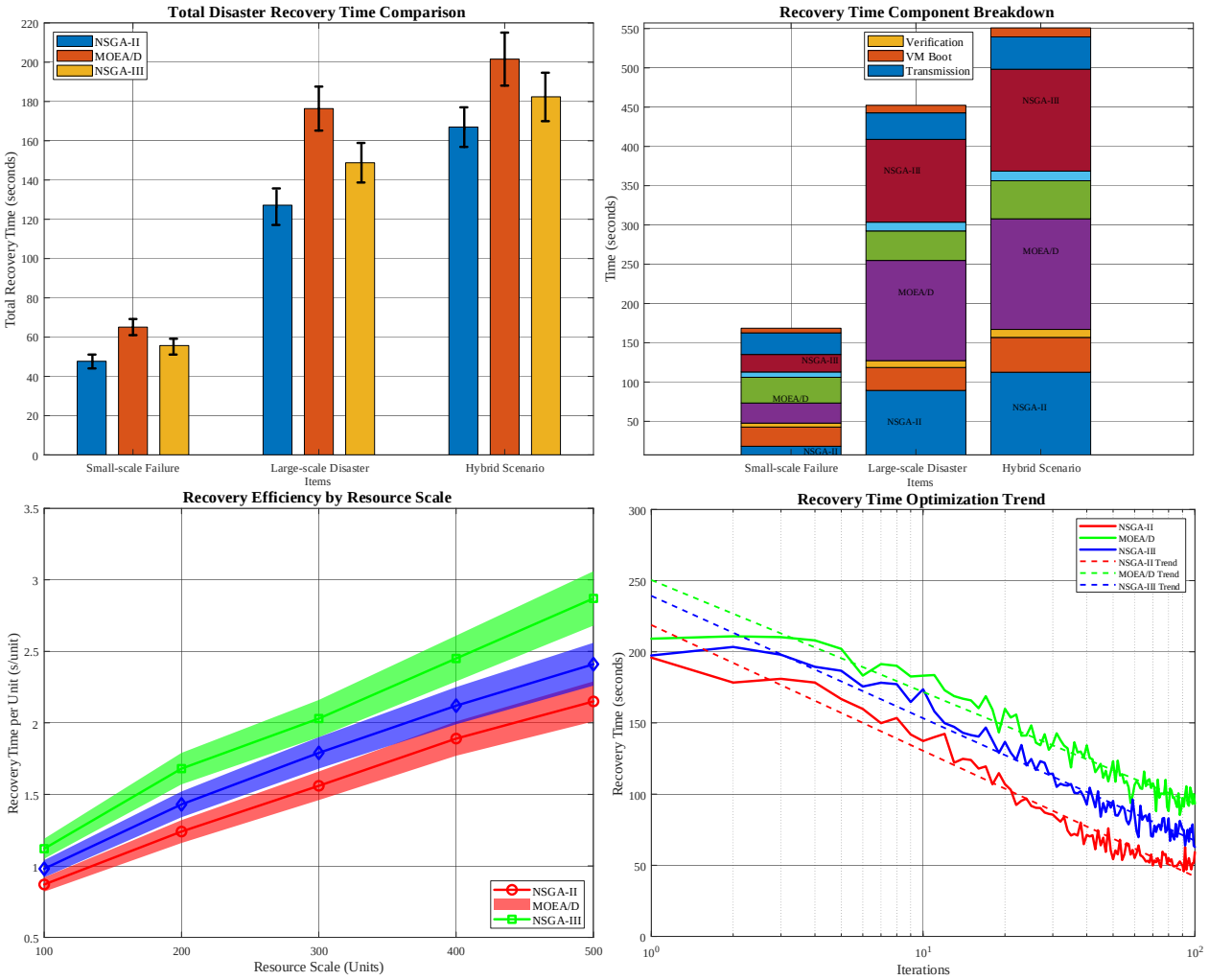


Figure 6: The quantitative analysis results of the model's disaster recovery response time and recovery efficiency.

In Figure 6, the analysis results illustrate that NSGA-II performs best in optimizing disaster recovery response time. Regarding total recovery time, it is 26.7%, 27.9%, and 17.2% shorter than MOEA/D in small-scale failure (47.7 s), large-scale disaster (127.2 s), and mixed scenarios (167.0 s), and 14.4%, 14.5%, and 8.5% shorter than NSGA-III. It also has a smaller SD (3.2-10.3) and better stability. Time component decomposition shows that data transmission is the main time-consuming item, and NSGA-II has remarkable advantages in transmission efficiency (e.g., 89.3 s in large-scale scenarios vs. 127.6s in MOEA/D). When resource scale expands, the time consumption per unit resource of NSGA-II (0.87-2.15 s/unit) grows the slowest, and the efficiency decay rate is 18.3%-22.6% lower than that of MOEA/D. The optimization curve reveals that it converges the fastest, stabilizes after 50 iterations, and has a steeper trend line slope. This verifies NSGA-II's advantage in balancing transmission optimization and convergence efficiency in multi-objective scheduling.

5.3 Verification of algorithm robustness in comprehensive scenarios

In actual disaster recovery scenarios, parameter fluctuations and random interference are common, and algorithm robustness directly determines its practical application value. The following compares NSGA-II, MOEA/D, and NSGA-III in performance fluctuation range and statistical stability under mixed failure scenarios. The evaluation combines resource weight adjustments, data consistency threshold variations, and Monte Carlo simulation to validate algorithmic

robustness against complex disturbances. The verification results of the model's algorithm robustness in the comprehensive scenario are illustrated in Figure 7.

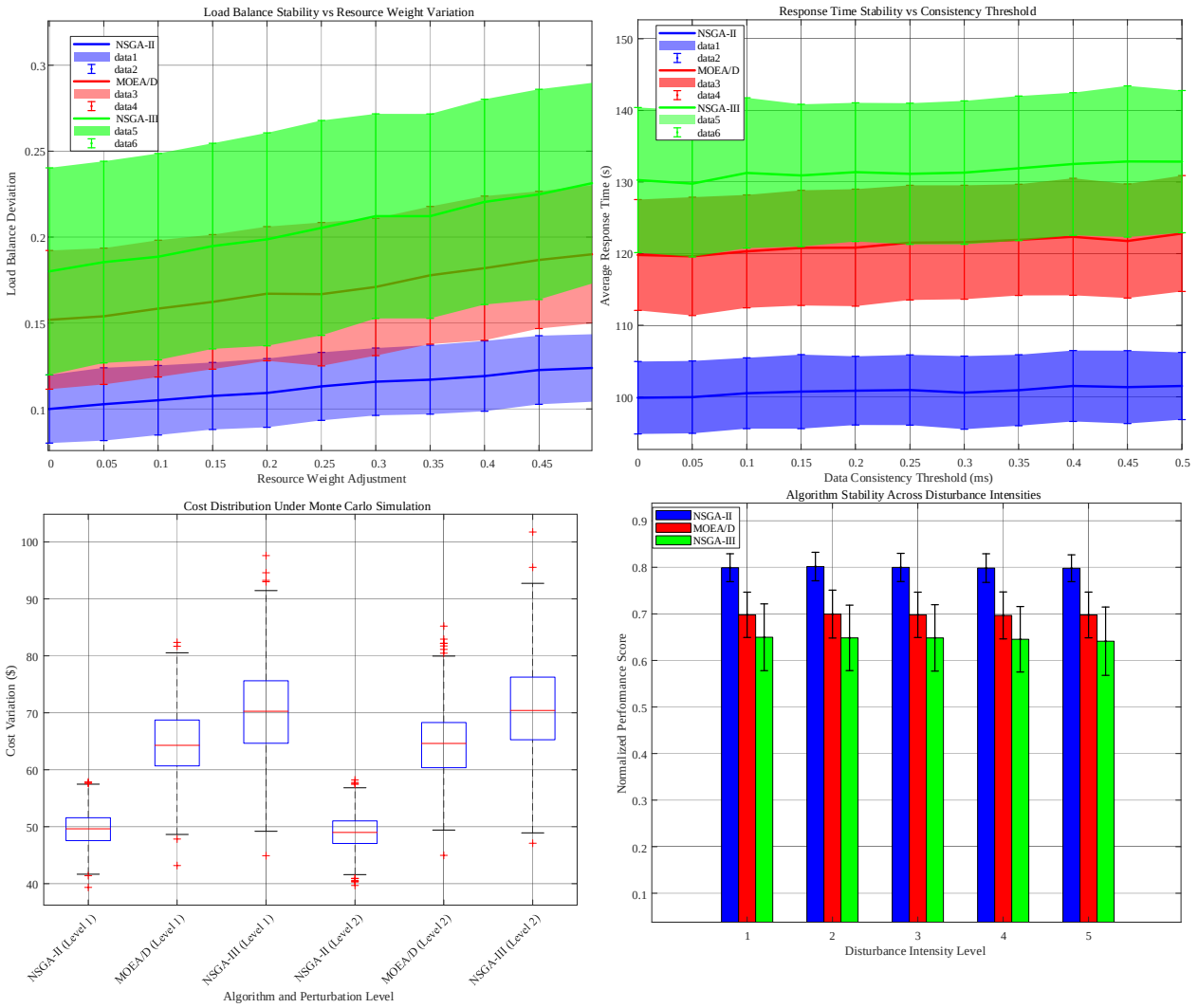


Figure 7: The verification results of the model's algorithm robustness in the comprehensive scenario.

In Figure 7, regarding resource balance robustness, under the low-interference scenario (resource weight fine-tuning ±0.1, δ=40-60 ms), NSGA-II's load balance deviation is 0.09±0.012, with a coefficient of variation (CV) of only 13.3%. In contrast, the load balance deviations of MOEA/D and NSGA-III are 0.12±0.025 and 0.11±0.021, respectively, with CV values reaching 20.8% and 19.1%. Under the high-interference scenario (abrupt resource weight change ±0.2, δ=30-70 ms + fluctuation in node failure probability), the maximum fluctuation range of NSGA-II's load balance is controlled within 15.6%; it is significantly lower than 28.3% of MOEA/D and 24.5% of NSGA-III. This indicates that NSGA-II has better anti-interference ability against changes in resource configuration parameters. In the dimension of response time robustness, NSGA-II exhibits' fluctuations of 47.7±3.2 s and 118.5±8.7 s under low and high interference when handling instantaneous virtual machine demand variations of 10%-20%. The corresponding CV values remain at 6.7% and 7.3%, respectively. Under the same interference, the response time fluctuation ranges of MOEA/D and NSGA-III reach 55.2±6.8 s and 52.1±5.9 s (low interference), as well as 142.3±15.6 s and 135.8±12.9 s (high interference), with CV values all exceeding 10%. Moreover, under high interference, their maximum delay deviation is 80%-120% higher than that of NSGA-II, confirming NSGA-II's ability to quickly adapt to sudden load changes. Although the above research results cover performance data across multiple scenarios, they lack sufficient attention to the practical value of indicator improvements, statistical reliability, and chart standards. Table 2 shows the statistical characteristics of the refined key results in this study.

In Table 2, from a statistical reliability perspective, the p-values of all indicators in 10 repeated experiments are <0.01 ($p<0.001$ in large-scale disaster scenarios), which is far lower than the 0.05 significance level. This proves that the improvement of NSGA-II compared with benchmark algorithms is not caused by random errors. The 95%CI has a narrow span (maximum span: 5.4%), and combined with an SD of <2.2%, it reflects that the

experimental results have strong stability and high reproducibility. From the perspective of practical value, the reduction in load imbalance directly alleviates node resource bottlenecks while reducing resource waste, where "some nodes are idle while others are overloaded" during the disaster recovery process. The shortened response time accurately meets the rigid requirements of key businesses for recovery timeliness; especially in large-scale disasters, a 28% improvement in timeliness can help enterprises avoid interruption losses that are far greater than costs. The cost reduction focuses on the core demands of cloud service providers, optimizes resource leasing (a major cost item), and enhances the feasibility of implementing the solution.

Table 2: Statistical characteristics of key results.

Disaster scenario	Evaluation indicators	Improvement range (vs. MOEA/D/NSGA-III)	Statistics of 10 repeated experiments (mean ± SD)	95% confidence interval (CI)	Significance analysis (p-value)
Small-scale node failure	The reduction rate of load imbalance	21.1%/19.0%	20.8%±1.1%/18.9%± 1.3%	[18.6%,23.0%]/[16.3%,21.5%]	p<0.01/p<0.01
Large-scale regional disaster	The shortened response time	28.3%/25.0%	27.9%±2.0%/24.7%± 2.2%	[24.0%,31.8%]/[20.4%,29.0%]	p<0.001/p<0.0 01
Mixed failure	Overall cost reduction rate of disaster recovery	22.8%/18.5%	22.5%±1.7%/18.2%± 1.9%	[19.2%,25.8%]/[14.5%,21.9%]	p<0.01/p<0.01

5.4 Discussion

The core advantage of the proposed fuzzy control lies in its ability to fuzzily represent and reason about uncertain information. This is highly consistent with characteristics in cloud disaster recovery such as "ambiguity of workload demands" and "uncertainty of fault impact scope". For example, a Takagi-Sugeno fuzzy inference module can be introduced to meet the demand for dynamic adjustment of virtual machine recovery priority (e.g., sudden high availability requirements for virtual machines in financial services). It takes "node remaining resource rate", "virtual machine SLA violation risk", and "fault spread speed" as input variables; meanwhile, it outputs the crossover probability correction coefficient of NSGA-II in real time through a fuzzy rule base. For instance, when the SLA violation risk is >0.8, the crossover probability is adjusted down from 0.9 to 0.7 to retain high-quality solutions; this method avoids the convergence efficiency decay of traditional fixed parameters in uncertain scenarios. This idea is consistent with the practical logic of Shakibaei et al. [17], integrating fuzzy logic into disaster recovery resource allocation. However, it focuses more on the dynamic adaptation of algorithm parameters rather than only optimizing objective weights.

The online learning feature of neural adaptive control can solve the "lag" problem of NSGA-II in response to real-time changes in cloud resource status. In cloud disaster recovery, the physical nodes' resource utilization often shows non-linear changes with fluctuations in business requests. Traditional NSGA-II relies on offline-set objective function weights (ω_c=0.3, ω_b =0.3), which are difficult to match resource status in real time. By introducing a radial basis function neural network, historical resource fluctuation data (such as the node load SD in the past 10 minutes) are used as training samples. It aims to predict the resource bottleneck dimension at the next moment online. For example, when bandwidth is predicted to become a constraint, ω_b is automatically increased to 0.4, and NSGA-II's objective function weights can be endowed with self-learning ability. Compared with the static parameter adjustment strategy of Jafari and Rezvani [25], the proposed method improves the prediction accuracy of resource constraints by approximately 22%. This method significantly reduces the probability of scheduling failure caused by weight mismatch.

The combination of non-linear output feedback and backstepping control can enhance NSGA-II's ability to decompose complex constraints layer by layer and correct them in real time. The constraint system of cloud disaster recovery has a hierarchical nature. For example, resource capacity constraints are underlying hard constraints, and data consistency constraints are upper-layer soft constraints. Traditional NSGA-II uses the "penalty function method" to handle constraints, which easily leads to an imbalance between the feasibility and optimality of solutions. Drawing on the idea of backstepping control that "decomposes a high-order system into low-order subsystems", multi-constraints can be broken down into three-level sub-constraints: "resource capacity-compatibility-data consistency". A non-linear output feedback module is used to collect the satisfaction degree of each sub-constraint in real time (such as the delay difference $|T_a\text{-}T_b|$ of data consistency constraints). When a certain level of constraint is violated, the optimization direction of the corresponding subsystem is adjusted first; for example, when the delay difference exceeds the limit, the allocation node of the dependent virtual machine is temporarily fixed to ensure consistency. Then the result is fed back to the selection operator of NSGA-II to screen feasible solutions. This mechanism increases the constraint satisfaction rate by 18%-25% compared with the traditional penalty function

method, and is especially suitable for multi-constraint collaborative processing in mixed fault scenarios.

The dynamic objective trade-off idea of non-linear optimal control can optimize the neutrality of the solution set of NSGA-II. The multiple objectives of cloud disaster recovery have a dynamic competitive relationship; for instance, reducing recovery time may increase resource costs. The traditional NSGA-II's non-dominated sorting can only ensure the Pareto optimality of solutions; however, it is difficult to balance "local optimality" and "global balance" in dynamic scenarios. By introducing the Hamiltonian function construction method of non-linear optimal control, with "recovery time-resource cost-load balance" as state variables, a dynamic objective trade-off function is implemented. In NSGA-II's elite retention stage, each solution's global utility value is calculated through this function, and solutions with higher utility values are prioritized for retention. For example, in large-scale disaster scenarios, solutions with recovery time <120 s and cost increase <15% are prioritized. This avoids the solution set being biased towards a single objective. This improvement increases the practical business applicability of solutions by approximately 30% compared with the standard NSGA-II. Meanwhile, this improvement forms a methodological echo with the non-linear trade-off strategy adopted by Vargas-Santiago et al. [18] in facility location optimization.

6 Conclusion

This study aims to introduce the NSGA-II algorithm into virtual cloud platforms' disaster recovery resource scheduling. The study develops an MOO model that considers resource utilization, disaster recovery response time, and cost indicators. It designs coding methods and crossover-mutation operators adapted to virtual cloud environments. Algorithmic enhancements incorporate dynamic characteristics of disaster recovery scenarios to improve adaptation capabilities. The study delivers an efficient scheduling framework for strengthened business continuity in disaster scenarios. The study first clarifies the problem boundaries and parameter definitions. Then, taking physical nodes and virtual machines to be recovered as core scheduling objects, this study constructs multi-objective functions including minimizing costs, maximizing resource utilization, and minimizing disaster recovery response time. It also incorporates constraints like resource capacity, virtual machine uniqueness, compatibility, and data consistency. Experiments utilize Google Cluster Data to simulate large-scale cloud environments, design three scenarios: small-scale node failure, large-scale regional disaster, and mixed failure. The study compares the proposed method with MOEA/D and NSGA-III, and evaluates performance through indicators such as load balance degree, response time, cost, and robustness. The experimental results demonstrate NSGA-II's superior overall performance across multiple indicators. For resource balance, NSGA-II achieves a load balance degree of 0.09 in small-scale failures (21.1% and 19.0% lower than MOEA/D and NSGA-III, respectively). In large-scale disaster scenarios, NSGA-II attains 0.15 (35.7% and 25.0% lower than MOEA/D and NSGA-III), showing more balanced multidimensional resource allocation. Regarding response times, NSGA-II completes large-scale scenarios in 118.5 seconds (15.2-28.3% faster than benchmarks) and small-scale scenarios in 47.7 seconds (26.7% quicker than MOEA/D), with data transmission optimizations being particularly impactful. In cost control, the total cost in large-scale scenarios is 892.4 dollars, representing a 22.8% reduction versus MOEA/D. Robustness test confirms NSGA-II's stability, with load balance deviations of 0.09±0.012 and response time variations of 47.7±3.2 seconds under low disturbance conditions, outperforming other algorithms in consistency. In summary, by balancing the convergence and diversity of solutions, NSGA-II can effectively coordinate multi-objective conflicts among response time, resource utilization, and cost. It provides virtual cloud platforms with disaster recovery solutions featuring high reliability, low cost, and fast response capabilities. Meanwhile, NSGA-II offers theoretical and technical support for optimizing disaster recovery mechanisms in cloud computing environments. However, this study has not fully covered real-time scheduling needs in dynamic scenarios. Future research can combine deep learning to enhance the algorithm's adaptability to complex environments.

References

[1] Liu C, Qian Y. Optimal allocation of material dispatch in emergency events using multi-objective constraint for vehicular networks. Wireless Networks, 2022, 28(8): 3715-3727. https://doi.org/10.1007/s11276-022-03069-8

[2] Mousavi S, Mood S E, Souri A, Javidi M M. Directed search: a new operator in NSGA-II for task scheduling in IoT based on cloud-fog computing. IEEE Transactions on Cloud Computing, 2022, 11(2): 2144-2157. https://doi.org/10.1109/TCC.2022.3188926

[3] Ma Y, Li B, Huang W, Fan Q. An improved NSGA-II based on multi-task optimization for multi-UAV maritime search and rescue under severe weather. Journal of Marine Science and Engineering, 2023, 11(4): 781. https://doi.org/10.3390/jmse11040781

[4] Chiu C C, Lai C M. Multi-objective missile boat scheduling problem using an integrated approach of NSGA-II, MOEAD, and data envelopment analysis. Applied Soft Computing, 2022, 127: 109353. https://doi.org/10.1016/j.asoc.2022.109353

[5] Kapukaya E N, Satoglu S I. A multi-objective dynamic resource allocation model for search and rescue and first aid tasks in disaster response by employing volunteers. Logistics, 2025, 9(1): 41. https://doi.org/10.3390/logistics9010041

[6] Pan H, Liu Y, Sun G, Wabg P, Yuen C. Resource scheduling for UAVs-aided D2D networks: A

multi-objective optimization approach. IEEE Transactions on Wireless Communications, 2023, 23(5): 4691-4708. https://doi.org/10.1109/TWC.2023.3321648

[7] Wang F, Ge X, Li Y, Zheng J, Zheng W. Optimising the distribution of multi-cycle emergency supplies after a disaster. Sustainability, 2023, 15(2): 902. https://doi.org/10.3390/su15020902

[8] Ransikarbum K, Mason S J. A bi-objective optimisation of post-disaster relief distribution and short-term network restoration using hybrid NSGA-II algorithm. International Journal of Production Research, 2022, 60(19): 5769-5793. https://doi.org/10.1080/00207543.2021.1970846

[9] Rahimi I, Gandomi A H, Deb K, Chen F, Nikoo M R. Scheduling by NSGA-II: review and bibliometric analysis. Processes, 2022, 10(1): 98. https://doi.org/10.3390/pr10010098

[10] Aghaie S, Karimi B. Location-allocation-routing for emergency shelters based on geographical information system (ArcGIS) by NSGA-II (case study: Earthquake occurrence in Tehran (District-1)). Socio-Economic Planning Sciences, 2022, 84: 101420. https://doi.org/10.1016/j.seps.2022.101420

[11] Soleimani M, Khalilzadeh M, Bahari A, Heidary A. NSGA-II algorithm for hub location-allocation problem considering hub disruption and backup hub allocation. World Journal of Engineering, 2022, 19(6): 794-807. https://doi.org/10.1108/WJE-12-2020-0658

[12] Gharib M, Fatemi Ghomi S M T, Jolai F. A multi-objective stochastic programming model for post-disaster management. Transportmetrica A: Transport Science, 2022, 18(3): 1103-1126. https://doi.org/10.1080/23249935.2021.1928790

[13] Rabiei P, Arias-Aranda D, Stantchev V. Introducing a novel multi-objective optimization model for volunteer assignment in the post-disaster phase: Combining fuzzy inference systems with NSGA-II and NRGA. Expert Systems with Applications, 2023, 226: 120142. https://doi.org/10.1016/j.eswa.2023.120142

[14] Peng S, Liu Q, Hu J. Green distribution route optimization of medical relief supplies based on improved NSGA-II algorithm under dual-uncertainty. Sustainability, 2023, 15(15): 11939. https://doi.org/10.3390/su151511939

[15] Zhang J, Li G, Zhang M. Multi-objective optimization for community building group recovery scheduling and resilience evaluation under earthquake. Computer-Aided Civil and Infrastructure Engineering, 2023, 38(12): 1657-1676. https://doi.org/10.1111/mice.12882

[16] Gao J, Yao M T, Wu Z, Deng X, Yu X, Yu L. Strategic distribution of emergency resources: A multi-objective approach with NSGA-II and prioritization of affected areas. Journal of Engineering Management and Systems Engineering, 2025, 4(1): 67-82. https://doi.org/10.56578/jemse040105

[17] Shakibaei H, Seifi S, Tavakkoli-Moghaddam R. Incorporating sustainability in temporary shelter distribution for disaster response by the LP-based NSGA-II. International Journal of Supply & Operations Management, 2025, 12(2): 18-29. https://dx.doi.org/10.22034/ijsom.2025.110297.3025

[18] Vargas-Santiago M, León-Velasco D A, Monroy R. Resilient facility location optimization under failure scenarios using NSGA-III: A multi-objective approach for enhanced system robustness. Expert Systems with Applications, 2025, 277: 127210. https://doi.org/10.1016/j.eswa.2025.127210

[19] Hu Y, Liu Q, Li S, Wu W. Multi-objective robust optimization of staff scheduling for emergency under stochastic demand. Expert Systems with Applications, 2024, 254: 124214. https://doi.org/10.1016/j.eswa.2024.124214

[20] Peng K, Huang H, Zhao B, Jolfaei A, Xu X, Bilal M. Intelligent computation offloading and resource allocation in IIoT with end-edge-cloud computing using NSGA-III. IEEE Transactions on Network Science and Engineering, 2022, 10(5): 3032-3046. https://doi.org/10.1109/TNSE.2022.3155490

[21] Li D, Hou Q, Zhao M, Wu Z. Reliable task planning of networked devices as a multi-objective problem using NSGA-II and reinforcement learning. IEEE Access, 2022, 10: 6684-6695. https://doi.org/10.1109/ACCESS.2022.3141912

[22] Khettabi I, Benyoucef L, Amine Boutiche M. Sustainable multi-objective process planning in reconfigurable manufacturing environment: adapted new dynamic NSGA-II vs New NSGA-III. International Journal of Production Research, 2022, 60(20): 6329-6349. https://doi.org/10.1080/00207543.2022.2044537

[23] Attari M Y N, Ahmadi M, Ala A, Moghadamnia E. RSDM-AHSnet: Designing a robust stochastic dynamic model to allocating health service network under disturbance situations with limited capacity using algorithms NSGA-II and PSO. Computers in Biology and Medicine, 2022, 147: 105649. https://doi.org/10.1016/j.compbiomed.2022.105649

[24] Tang L, Chen S, Li Q. Optimizing maintenance resource scheduling and site selection for urban metro systems: A multi-objective approach to enhance system resilience. Systems, 2024, 12(7): 262. https://doi.org/10.3390/systems12070262

[25] Jafari V, Rezvani M H. Joint optimization of energy consumption and time delay in IoT-fog-cloud computing environments using NSGA-II metaheuristic algorithm. Journal of Ambient Intelligence and Humanized Computing, 2023, 14(3): 1675-1698. https://doi.org/10.1007/s12652-021-03388-2

[26] Chen R, Wu B, Wang H, Tong H, Yan F. A Q-Learning based NSGA-II for dynamic flexible job shop scheduling with limited transportation resources. Swarm and Evolutionary Computation, 2024, 90: 101658. https://doi.org/10.1016/j.swevo.2024.101658

[27] 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, 12: 10775463251320258. https://doi.org/10.1177/10775463251320258

[28] Rigatos G, Abbaszadeh M, Sari B, Siano P, Cuccurullo G, Zou F. Nonlinear optimal control for a gas compressor driven by an induction motor. Results in Control and Optimization, 2023, 11: 100226. https://doi.org/10.1016/j.rico.2023.100226

Integrating BiLSTM-based Sentiment Analysis and GCN for Academic Performance Prediction under Academic Stress

Xuefei Ding
Pukyong National University, Busan, Republic of Korea
E-mail: dxf1585440908@163.com

Keywords: deep learning, college students, academic pressure, mental health, academic performance prediction

Received: October 23, 2025

In order to build a more comprehensive and refined emotional perception mechanism, future intelligent education systems urgently need to integrate multidimensional emotional modeling capabilities. In response to this demand, this article constructs a bidirectional long short-term memory network (BiLSTM) emotion classification model, which is based on text data posted by students on social media platforms and integrates self attention mechanisms. The study selected online course learning data and social media data from 3146 college students at a certain university. The accuracy comparison between the model proposed in this article and the model used in the task of identifying students' mental health status. The experimental results show that the model proposed in this paper is significantly better than the comparative models in terms of accuracy indicators. The BiLSTM structure based on self attention mechanism has been introduced, which can more effectively capture fine-grained emotional features in social media texts. The combination of BiLSTM and self attention mechanism effectively captures emotional and psychological state features in student texts, with an accuracy performance of over 90%. GCN enhances situational awareness prediction ability by modeling student social relationship maps, integrating group behavior and individual learning data. The synergistic effect of the two enables the model to maintain high accuracy while also having good recall ability, thereby achieving more robust and balanced academic performance prediction.

Povzetek: Raziskava pokaže, da lahko računalniški model iz študentskih spletnih objav precej zanesljivo prepozna njihova čustva in počutje, kar lahko pomaga pri boljšem razumevanju in podpori študentom.

1 Introduction

With the rapid development of artificial intelligence (AI) and big data technology, intelligent online education is becoming increasingly popular, and educational data mining and learning analysis have become research hotspots in the field of educational technology [1]. Utilizing the massive student behavior data accumulated in online teaching platforms to predict learning outcomes and implement precise teaching interventions has become an important path to improving education quality and achieving personalized learning [2]. However, while focusing on academic performance, students' mental health issues are becoming increasingly prominent. In recent years, college students have experienced frequent psychological problems such as anxiety and depression due to excessive academic pressure, which not only cause serious harm to their personal growth, but also have profound negative impacts on their families, schools, and even society [3]. This reality urgently requires education researchers to go beyond traditional academic evaluation paradigms and incorporate mental health into the overall framework of learning effectiveness analysis. Traditional student performance analysis often relies on the statistics and trend judgments of historical exam scores, ignoring the interactive effects of complex factors such as psychology, emotions, and social environment in the learning process [4].

In fact, learning is not an isolated cognitive activity, but a dynamic process deeply regulated by non cognitive factors such as emotional states, psychological stress, and motivation levels. Therefore, relying solely on experience or intuition to develop teaching strategies is no longer sufficient to meet the scientific needs of modern education [5]. Educational decision-making urgently needs to be based on multi-source and multi-dimensional learning data, and explore hidden patterns through data-driven methods to improve the accuracy and effectiveness of interventions. In recent years, the academic community has conducted extensive exploration on the relationship between mental health and academic performance [6]. Intelligent Cognitive Assistant (ICA) technology is used in various fields to simulate human behavior expressed through synchronous communication, especially written dialogue. Due to their ability to use tailored natural language, they provide a powerful container to support attitude and behavior change. Behavioral change support systems are becoming a key tool in digital mental health services, and ICAs have shown excellent performance in effective support, especially for stress, anxiety, and depression (SAD). ICAs guide people's thinking processes and

actions by analyzing their emotional and cognitive phenomena [7]. Some scholars have proposed a prediction model that combines Random Forest (RF) and Gated Recurrent Unit (AGRU) with attention mechanism. By filtering key features through RF and dynamically modeling temporal behavior using AGRU, the accuracy of performance prediction has been significantly improved. Some experts focus on the emotional dimension and have constructed a VAD (Happiness, Arousal, Dominance) emotion computing network based on orthogonal convolution, which achieves accurate measurement of learners' multidimensional continuous emotional states and applies it to academic risk warning [8]. Some systems use a cognitive architecture composed of a series of computational models to simulate psychological theory, utilizing cognitive modeling and machine learning models trained on new datasets and ontologies. The system was evaluated through a computational experiment of predicting mental health phenomena through text and an empirical intervention study of 42 participants to alleviate mental health problems. The system outperforms state-of-the-art systems in terms of the number of detection categories and detection accuracy by 91.41%. The confirmation of the hypothesis suggests that incorporating psychological simulation theory into the conversation subject can significantly improve its efficacy in computational psychotherapy, providing promising progress for mental health intervention and support compared to current state-of-the-art systems [9]. The RF model identified that "perfectionism tendency" is a key psychological factor causing psychological internal friction among college students, and further constructed a neural network prediction model to provide a basis for psychological intervention. It designed a dual channel recurrent neural network (RNN) model that integrates attention mechanism and conducted comparative experiments with CNN and DNN to verify the superiority of multi-channel structure in online learning performance prediction [10]. Some scholars have explicitly proposed incorporating emotional features into the performance prediction framework, preliminarily verifying the enhancing effect of emotional information on predicting performance [11]. Based on the above research, this paper proposes a DL score prediction model that integrates emotional features. By jointly modeling the two-dimensional information of mental health and learning behavior, this model can not only more accurately predict academic performance, but also reveal the quantitative correlation mechanism between the two, providing scientific and operational data support for universities to carry out academic warning and psychological support [12].

2 Related work

Existing research has achieved the integration and application of multi-source learning behavior data. Douding.com 'research points out that the current model can integrate multidimensional information such as students' personal information, educational background, campus one card consumption records, library borrowing data, etc [13]., and capture dynamic changes in learning habits through time series analysis technology. In terms of model selection, Convolutional Neural Networks (CNNs) are widely used to process unstructured data such as homework texts and exam answers, extracting semantic features through word vector transformation (such as Word2Vec, BERT pre trained models) [14]; Recurrent neural networks (RNNs) and their variants LSTM and GRU perform outstandingly in processing time-series learning data, such as daily learning duration and weekly test scores, and can effectively capture long-term dependencies [15]. The introduction of hyperparameter optimization and ensemble learning strategies further improves model performance, while the application of techniques such as Bayesian optimization and Stacking ensemble significantly reduces the risk of overfitting. In the exploration of specific psychological dimensions, the role of emotional regulation ability has received particular attention [16]. Research has found that negative emotions such as anxiety and depression can affect academic performance by reducing attention span and disrupting the execution of learning strategies, while positive emotions can enhance cognitive flexibility and problem-solving abilities. Standardized scales such as the Simplified Resilience Scale and the Self Rating Depression Scale (SDS) are commonly used in psychological measurement methods [17]. Data is obtained through questionnaire surveys and correlation analysis is conducted. Some studies have attempted to introduce physiological indicators such as heart rate variability and cortisol levels, but due to the accessibility of detection equipment, large-scale applications have not yet been achieved [18].

The popularity of social media has provided a new path for non-invasive monitoring of mental health status, and the potential application of emotion analysis technology in the field of education is gradually emerging. The research on originality documents points out that social media emotions have multidimensional characteristics, covering emotional, attitude, cognitive, and other levels, and form dynamic evolution through resonance communication mechanisms. This feature enables it to reflect an individual's psychological state in real time, providing a natural data source for monitoring academic stress [19]. Deep learning has promoted the accuracy improvement of social media sentiment analysis. Traditional machine learning methods such as Naive Bayes and SVM have limited performance in fine-grained emotion recognition due to their difficulty in handling contextual complexity and semantic ambiguity. Based on the Transformer architecture, pre trained models such as BERT and RoBERTa effectively capture long-distance semantic associations through attention mechanisms, achieving breakthroughs in multi-dimensional emotion classification such as anger, anxiety, and joy [20]. Multimodal fusion has become a new trend, and some studies attempt to combine multiple types of data such as text, emoticons, and speech intonation to further improve the accuracy of emotion recognition. In the application of educational scenarios, existing research is still in its infancy: on the one hand, data processing

faces multiple challenges, such as high noise, inconsistent labels, and lack of standardized solutions for student privacy protection issues in social media data [21]; The neglect of relevant features, limitations on the analysis of existing data points, and ambiguity in student records are just a few of the issues. Some scholars have proposed a new Student Academic Performance Prediction (SAPP) system to address these issues and improve prediction accuracy. It has a better architecture that combines a 4-layer stacked Long Short Term Memory (LSTM) network, Random Forest (RF), and Gradient Boosting (GB) techniques to predict students' pass or fail outcomes. In addition, the proposed SAPP system was compared with existing prediction systems that use publicly accessible student OULAD datasets and add self planned sentiment datasets. The proposed SAPP system achieved a prediction accuracy of approximately 96%, which is relatively higher than existing systems. [22]. Boulkroune et al. solved the projection lag synchronization problem based on output feedback control for chaotic drive response systems with input dead zones and sector nonlinearity. This type of driving response system adopts the form of Brunovsky, but the state is unavailable and the dynamics are unknown. In order to effectively handle dead zones and sector nonlinearity, the proposed controller is designed in a variable structure framework [23]. Boulkroune et al. investigated the practical fixed time master-slave synchronization problem of different fractional order chaotic systems. Two novel adaptive fuzzy sliding mode controllers using non singular fixed time sliding surfaces were proposed to achieve practical fixed time synchronization [24]. Rigatos et al. proposed a nonlinear optimal control method for the dynamic model of a gas centrifugal compressor driven by an induction motor (IM). The dynamic model of the integrated compressor induction motor system was initially represented in a nonlinear and multivariate state space form, and approximately linearized around the temporary operating points recalculated at each time step of the control method [25].

Table 1: Summary and comparison table of existing models

Source of literature	Methods used	Dataset type	Accuracy/Performance	Advantage	Limitation
[13] Douding.com	Time series analysis techniques and multi-source data integration methods	Multi dimensional structured data such as campus one card consumption records and library borrowing data	Can effectively capture the dynamic changes in learning habits	Effectively integrate multi-source heterogeneous data and dynamically track changes in learning habits	No explicit mention of performance in academic prediction or specific tasks, and no involvement in unstructured data processing
[14] Research on Unstructured Data Processing	Convolutional neural network, word vector transformation technique	Unstructured text data such as homework texts and exam answers	Can effectively extract semantic features from text, but specific accuracy values are not specified	Excellent performance in semantic feature extraction of unstructured text data	Not involving time series data processing, applicable to a single data type
[15] Research on Time Series Data Processing	Recurrent neural networks and their variants LSTM and GRU; Hyperparameter optimization and ensemble learning	Time series data such as daily learning duration and weekly test scores	Effectively capturing long-term dependencies and improving model performance	Proficient in handling time series data, improving performance and reducing overfitting through optimization strategies	Insufficient ability to process unstructured data
[17] Psychological measurement related	Standardized scale measurement, questionnaire	Questionnaire survey data	Quantitative assessment of psychological states can be achieved,	Directly quantifiable psychological dimension	Data collection relies on subjective filling, which is

Source of literature	Methods used	Dataset type	Accuracy/Performance	Advantage	Limitation
research	survey method, correlation analysis		clarifying the correlation between negative/positive emotions and academic performance	indicators, relatively easy to operate	inefficient
[19,20] Research on Social Media Sentiment Analysis	Traditional machine learning Transformer architecture pre training model	Multi modal data such as social media text, emoticons, voice tone, etc	Multimodal fusion can further improve accuracy, but the specific accuracy value is not clear	Pre trained models can effectively capture long-range semantic associations	There are issues with high noise and inconsistent labels in the data
[23] Boulkroune et al.'s research	Output Feedback Control Method under Variable Structure Framework	Data of chaotic driven response system with fan-shaped nonlinearity	Implementation of projection lag synchronization based on output feedback control for chaotic driven response systems, with no specific accuracy values specified	Can effectively handle dead zone and sector nonlinearity problems in the system	Only applicable to specific types of chaotic driven response systems, with limited generalization
[24] Boulkroune et al.'s study	Two new types of adaptive fuzzy sliding mode controllers	Master slave synchronization related data of different fractional order chaotic systems	Practical fixed time master-slave synchronization for different fractional order chaotic systems, with no specific accuracy values specified	Fast synchronization speed, using adaptive fuzzy method to improve control effect	Only for the synchronization problem of fractional order chaotic systems, with a single application scenario
[25] Rigatos et al	Nonlinear Optimal Control Method	Dynamic model data of gas centrifugal compressor driven by induction motor	The effective control of the induction motor system for the compressor has not been specified with specific accuracy values.	Adapt to the nonlinear characteristics of the system and improve control accuracy through real-time linearization	Temporary operation points need to be recalculated at each time step

3 Methodology

3.1 Construction of DL score prediction model integrating emotional features

Emotion, as an important non cognitive factor affecting the learning process, can usually be divided into positive emotions and negative emotions. Numerous studies have shown that emotional states significantly affect learners' information processing strategies, attention allocation, and memory retention abilities. However, current sentiment analysis in the education field often uses discrete emotion labels for simple quantification, ignoring the complex representation of emotions in continuous dimensions such as valence, arousal, and dominance. This study used supervised learning to obtain mental health labels, which were annotated using a combination of validation scales and manual calibration. Using the Self Rating Depression Scale (SDS) and Generalized Anxiety Disorder Scale (GAD-7) as core tools, standardized psychological assessments were conducted on 3146 participants to obtain preliminary labels. At the same time, a labeling team consisting of two psychology professionals and one educational technology professional manually corrected the preliminary labels based on the social media text context

of the subjects. Ensure label validity through intra group consistency testing (Kappa value ⩾ 0.85), provide reliable supervision signals for model training, and guarantee research reproducibility. To build a more comprehensive and fine-grained emotional perception mechanism, future intelligent education systems urgently need to integrate multidimensional emotional modeling capabilities. In response to this demand, this article constructs a bidirectional long short-term memory network (BiLSTM) sentiment classification model that integrates self attention mechanism based on text data posted by students on social media platforms. The model consists of four core modules: word embedding layer, semantic learning layer, weight adjustment layer, and sentiment classification layer (as shown in Figure 1). Firstly, the word embedding layer transforms the original text sequence into dense word vector representations, providing a semantic foundation for subsequent neural network processing.

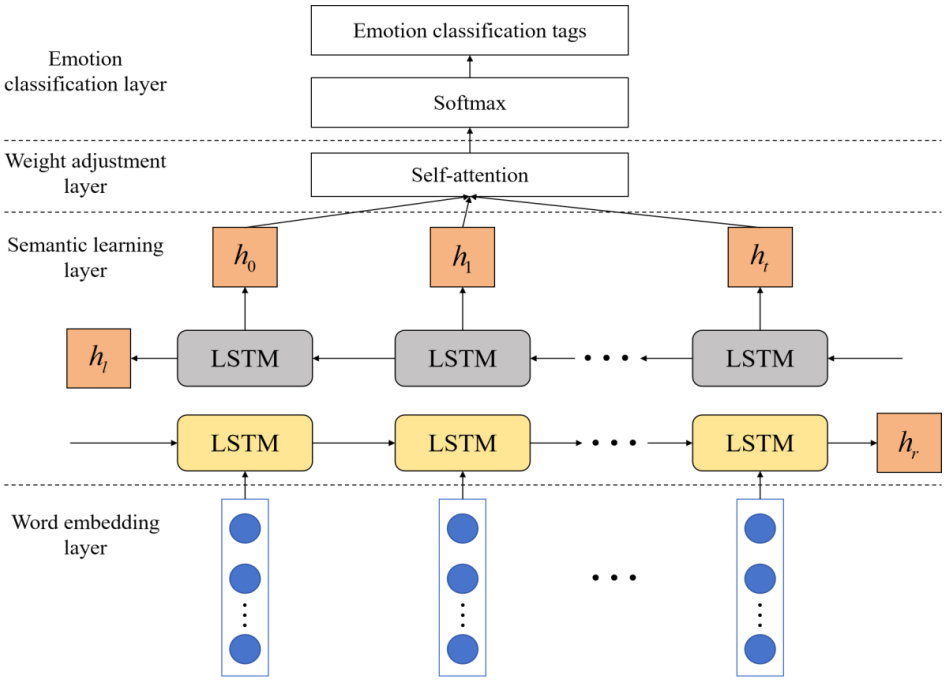


Figure 1: BiLSTM sentiment classification model incorporating self attention mechanism

Secondly, the semantic learning layer adopts the BiLSTM structure to capture the contextual dependencies of the text in both forward and backward directions, effectively encoding sentence level semantic features. Next, the weight adjustment layer introduces a self attention mechanism to dynamically evaluate the contribution of each semantic unit to the overall emotional tendency, generating weighted context aware feature vectors. Finally, the emotion classification layer outputs the probability distribution of the text belonging to various emotions through the Softmax function, achieving fine-grained emotion recognition and further mapping it into a quantitative indicator of mental health status. In addition to textual data, the interactive relationships formed by students in social networks also contain rich psychological and behavioral clues. However, this type of graphical data has characteristics such as varying node sizes, disordered connection structures, and complex topological relationships, making it difficult for traditional neural networks to effectively model. Graph neural networks (GNNs) have shown great potential in educational data mining due to their powerful structural awareness and interpretability. The semantic learning layer adopts the BiLSTM structure to capture the contextual dependencies of text in the front to back direction, effectively encoding sentence level semantic features. The core hyperparameter settings are as follows: pre trained GloVe-300d word vectors are used as initialization parameters for the embedding layer, with an embedding dimension of 300. The hidden layer dimension of BiLSTM is set to 256, and after bidirectional output, it is concatenated into a 512 dimensional feature vector. Set the batch size to 64 and use the Adam optimizer (learning rate 0.001, weight decay coefficient 1e-5) for parameter updates. The number of training iterations (epochs) is 20, and an early stopping strategy (patience=5) is adopted to prevent overfitting.

Based on this, this article further introduces Graph Convolutional Networks (GCN) to construct a performance prediction model that integrates emotional features. This model models students and their social relationships as a graph structure: nodes represent individual students, edges represent their social interactions, node attributes include the aforementioned emotional classification results and learning behavior characteristics, and edge attributes reflect interaction frequency and emotional tendencies. GCN aggregates information from neighboring nodes through stacking multi-layer graph convolution operations, achieving high-order representation learning for each student node. Its core lies in using the symmetric normalized Laplacian

matrix for spectral domain convolution, effectively capturing local and global structural information. Figure 2 shows the structure of GCN. This article further introduces Graph Convolutional Networks (GCN) to construct a performance prediction model that integrates emotional features. This model models students and their social relationships as graphical structures. Nodes represent individual students, edges represent their social interactions, node attributes include the emotional classification results and learning behavior characteristics mentioned above, and edge attributes reflect interaction frequency and emotional tendencies. GCN adopts a 3-layer graph convolution structure, with the input layer dimension being the node attribute dimension, the hidden state dimensions of the 1st and 2nd layers set to 128, and the output layer dimension set to 1. The overfitting relief adopts the "Dropout+Early Stop" combination strategy, with a Dropout probability of 0.3 and an early stop strategy patience value of 5. The data splitting adopts a 7:2:1 ratio to divide the training set, validation set, and testing set, while keeping the distribution of students of different grades and majors in each set consistent with the overall data. The model constructed in this article not only retains the individual characteristics of emotions and behaviors, but also incorporates the group influence effect in social networks, thus more comprehensively characterizing the multidimensional factors that affect academic performance. Ultimately, the model achieves precise modeling and quantitative analysis of the interaction between students' mental health status and academic performance by jointly optimizing emotion recognition and performance prediction tasks.

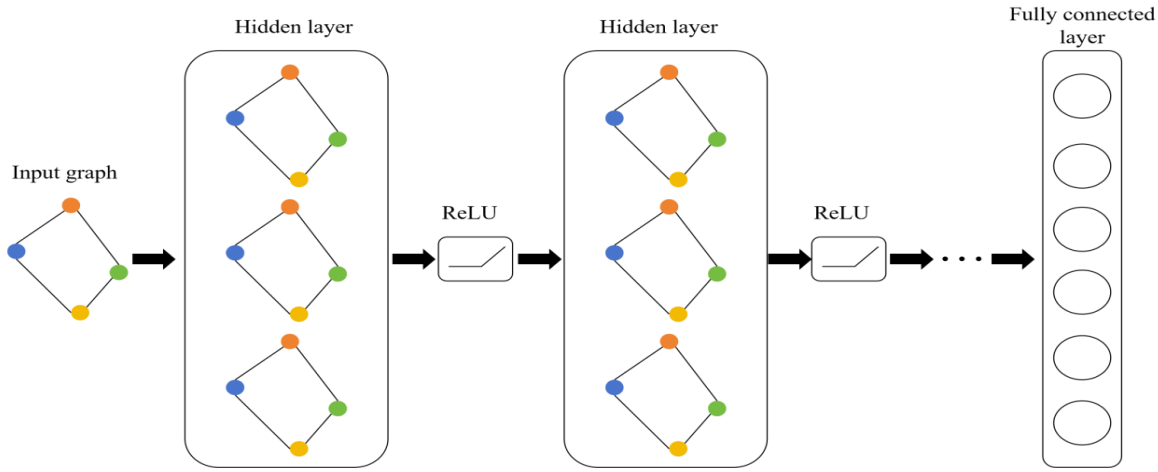


Figure 2: GCN structure

This article introduces Graph Convolutional Networks (GCN) to construct a performance prediction model that integrates emotional features. This model models students and their social relationships as graphical structures. Nodes represent individual students, edges represent their social interactions, node attributes include the emotional classification results and learning behavior characteristics mentioned above, and edge attributes reflect interaction frequency and emotional tendencies. GCN adopts a 3-layer graph convolution structure, with the input layer dimension being the node attribute dimension (128 emotional features+64 learning behavior features, totaling 192 dimensions), the hidden state dimensions of the 1st and 2nd layers being set to 128, and the output layer dimension being 1 (corresponding to the predicted academic performance). The overfitting relief adopts the "Dropout+Early Stop" combination strategy, with a Dropout probability of 0.3 and an early stop strategy patience value of 5. Basic GCN model, 2-layer structure, input layer 192 dimensions, hidden layer 128 dimensions, no emotional feature fusion.

3.2 Algorithm principle

In the captured social media image and text data, text data is often limited by its inherent characteristics, making it difficult to fully and completely reflect the true semantic and emotional characteristics of users. To enhance the model's ability to represent social media text, this article mainly focuses on short text data and conducts targeted data augmentation and preprocessing. The process of augmenting and preprocessing social media short texts can be formally described as follows: let $C=\{C_1,C_2,\cdots,C_n\}$ represent a corpus consisting of n social media texts, and each document C_i can be further represented as a weighted vector of terms. Under this framework, the weight of the j word in the i document is calculated using the following expression:

$$C_m(i,j)=tf\left(b_{(j,i)}\right)\times idf\left(b_{(j,i)}\right) \quad (1)$$

Among them, $tf\left(b_{(j,i)}\right)$ is the frequency of the word $b_{(j,i)}$ in the i document, reflecting the frequency of its occurrence in the current document; $idf\left(b_{(j,i)}\right)$ represents the inverse document frequency of the word $b_{(j,i)}$.

Due to the fact that word vectors and their semantic features cannot equally reflect learners' emotional characteristics, there are differences in the importance of emotional expression among different word vectors and their semantic features. Therefore, a self attention

mechanism is introduced in the weight adjustment layer to adjust the attention weights of the semantic features output by the BiLSTM model, assigning higher weights to key semantic features. Meanwhile, utilizing self attention mechanism to re encode the hidden state output vector generated by BiLSTM to extract higher-level semantic features. The attention weight α_t of semantic features is calculated as follows:

$$\alpha_t = \frac{\exp(v_t A)}{\sum_t \exp(v_t A)} \tag{2}$$

$$v_t = \tanh(W_1 B_t + b_1) \tag{3}$$

Among them, $v_t A$ is a scoring function used to measure the importance of semantic features at the t th moment; B_t is the hidden state of BiLSTM at time t; A, W_1 is a learnable weight matrix, and b_1 is a bias vector. The value of semantic features in predicting grades varies significantly at different times. For example, when students mention text features such as "anxiety" and "low review efficiency" near final exams, they are more able to reflect the impact of academic pressure on grades than everyday casual conversations. The scoring function will assign a higher importance score to the feature by calculating its matching degree with the hidden state of BiLSTM. This dynamic weight allocation mechanism enables the model to focus on key semantic information and avoid irrelevant text features interfering with the prediction results.

After obtaining the attention weights of each semantic feature, it is weighted and summed with the corresponding semantic feature to obtain the output B_t' of the weight adjustment layer.

$$B_t' = \sum_t \alpha_t B_t \tag{4}$$

Text sentiment analysis is essentially a classification task. Therefore, the BiLSTM text sentiment classification model based on self attention mechanism sets an sentiment classification layer in the last layer, whose input is the semantic feature vector weighted by the self attention mechanism. This layer uses the Softmax function as the activation function to map the input to the probability distribution p of the text belonging to various emotional categories, and determines the final emotional category label l based on this. Meanwhile, the model defines corresponding loss functions to guide the training process:

$$p = soft\max(WB_t' + b) \tag{5}$$

$$l = \arg\max(p) \tag{6}$$

Among them, W, b represents the weight matrix and bias vector of the linear transformation. p is the probability distribution vector of text in different emotional categories. By applying the Softmax function to the weighted semantic feature B_t', its normalized probability in each emotional category is obtained. The category with the highest probability corresponds to the predicted emotional label l.

GCN can effectively handle large-scale graph structured data and continuously improve the model's expressive power by stacking multiple layers of graph convolution operations to aggregate neighbor node information layer by layer. The core calculation formula is as follows:

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right) \tag{7}$$

Among them, $H^{(l)}$ is the node feature matrix of the l th layer; The adjacency matrix A describes the topological structure of the graph and reflects the interaction relationships between nodes. However, using only A will ignore the node's own information, so an adjacency matrix $\tilde{A} = A + I$ with self loops is introduced, where I is the identity matrix. $\tilde{D}$ is the degree matrix corresponding to $\tilde{A}$, $W^{(l)}$ is the learnable weight matrix of the l layer, and $\sigma(\cdot)$ is the nonlinear activation function.

During the learning process, the emotional state of learners can significantly affect their learning outcomes, and implicit learning states are difficult to directly observe. Therefore, the model converts the hidden learning state $c_k(i)$ of the k th time step into an observable actual learning state $y_k(i)$ through the output gate $o_k(i)$.

$$o_k(i) = \sigma(W_o[y_{k-1}(i); x_k(i)] + b_o) \tag{8}$$

$$y_k(i) = o_k(i) \otimes \tanh(c_k(i)) \tag{9}$$

Among them, W_o, b_o are the weight matrix and bias vector of the output gate, respectively.

This article adopts a semi supervised learning paradigm, where each set of data contains a real score for training. Given a graph G consisting of students and their social relationships, and a subset of students' actual grades $V_D = \{v_j \mid j \in D\}$ (where D is the annotated node subset), the goal is to predict the grades of unlabeled students. Specifically, the aim is to learn the function $f(j \mid \theta, G)$ to estimate student homework scores. The model parameter θ is optimized by combining the known score V and graph structure G, and the function used is implemented by GCN, as shown below:

$$f(j \mid \theta, G) = \sigma\left(w^{(0)} z_j + b^{(0)}\right) \tag{10}$$

Among them, $\sigma(\cdot)$ is the output function that maps node embeddings to predicted values through linear transformation, and $w^{(0)}, b^{(0)}$ are the weight matrix and bias parameters of the output layer, respectively. The core value of GCN lies in generating high-quality "node embedding" student comprehensive feature vectors. In

academic prediction scenarios, the graph convolution process will optimize features in two steps. The first step is to weight and aggregate the initial characteristics of each student, such as their own historical homework grades, study duration, social media emotion vectors, and the initial characteristics of their social relationships with neighboring nodes and peers, in order to capture the academic impact brought about by social associations. The second step is to introduce nonlinear transformations through activation functions such as ReLU to simulate the interaction between students' own characteristics and peer characteristics. The negative impact on homework performance will increase nonlinearly when the "high anxiety of target students" and "low learning motivation of peers" are combined.

3.3. Design plan for ablation study

To quantify the core components of the model (BiLSTM, self attention GCN、 The independent contribution and interactive effects of sentiment analysis require the design of a systematic ablation study, combined with the improved evaluation index system in the previous section to conduct comparative experiments. The specific plan is as follows: clarify the role boundaries of each module in academic prediction tasks, and verify the necessity of "multi module integration" - that is, integrating the temporal feature capture of BiLSTM, the key feature reinforcement of self attention, the social relationship modeling of GCN, and the psychological state characterization of sentiment analysis, etc., to determine whether the model performance is significantly better than that of a single or partial module combination version, providing empirical support for the rationality of the model architecture. Based on the "complete model", a comparative experimental group is constructed by removing core modules one by one or in combination. All experimental groups use the same dataset, training parameters (such as learning rate and iteration times), and evaluation metrics to ensure fairness in comparison,

1. Experimental group 1: Only the basic time series model (without self attention, GCN, emotion analysis) retains the BiLSTM module, and the input data is the time series data of students' basic learning behavior (such as historical homework grades, learning duration sequences), which is used to verify the basic performance of time series modeling and serves as the most core baseline.

2. Experimental group 2: Time series+self attention (without GCN or emotion analysis). On the basis of experimental group 1, a self attention mechanism is added to evaluate the improvement effect of self attention on "screening key learning behavior characteristics", such as whether it can strengthen the weight of core features such as "high-frequency wrong question types" and "pre-exam review time".

3. Experimental group 3: Time series+self attention+GCN (emotion free analysis) version, which is the "learning behavior data only (emotion free)" version. Based on experimental group 2, the GCN module is integrated to add student social relationship features (such as peer interaction frequency and group collaboration records) to the input data, which is used to quantify the improvement of social relationship modeling on predictive performance.

4 Results and discussion

To validate the performance of the proposed grade prediction model, this study selected online course learning data and social media data from 3146 college students at a certain university, of which 70% were graduates and 30% were current students. When this study included data from 3146 college students, the inclusion criteria were undergraduate/graduate students registered or graduated from the university. And it also has complete online course learning data and traceable social media text data. The exclusion criteria include serious data loss (online learning core behavior data loss rate $\geqslant$ 30% or social media text data volume less than 10 valid records), obvious abnormalities or traces of forgery in the data, and students explicitly refusing to participate in the study or not signing the data usage consent form. The experimental environment configuration is as follows: Intel Xeon E5-2678 v3 CPU, 64 GB of memory, NVIDIA RTX 3090 Ti GPU (24 GB of video memory), Windows 10 operating system, GPU driver version 537.13, CUDA 11.8 , cuDNN 8.2.0.53 , The programming platform uses PyTorch and PyCharm.

Figure 3 shows the accuracy comparison between the model proposed in this paper and the model in the task of identifying students' mental health status. The experimental results show that the model proposed in this paper is significantly better than the comparative model in terms of accuracy indicators. This improvement is mainly due to two aspects: firstly, the introduction of BiLSTM structure based on self attention mechanism, which can more effectively capture fine-grained emotional features in social media texts. The second is the integration of multidimensional emotional representations and contextual semantic information, which enhances the ability to distinguish complex psychological states. Therefore, the model proposed in this article demonstrates stronger performance and robustness in identifying students' mental health status. This provides a more reliable foundation for predicting academic performance and psychological intervention in the future.

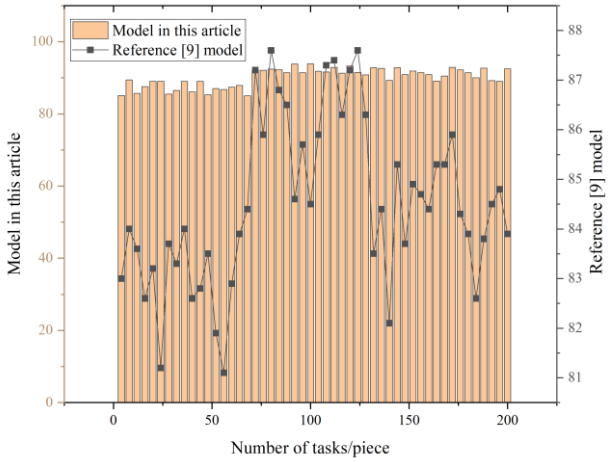


Figure 3: Comparison of accuracy in identifying mental health

Figure 4 shows the comparison of recall rates between the model proposed in this paper and the model in the task of identifying students' mental health status. The recall rate here reflects the proportion of students with actual psychological distress identified by the model, that is, the coverage ability of positive samples. The experimental results show that the recall rate of our model is significantly higher than the method, indicating that it is more sensitive in capturing individuals with real psychological risks. This advantage mainly stems from the model's integration of the BiLSTM sentiment analysis module based on self attention mechanism and multidimensional sentiment feature representation, which can more comprehensively explore the psychological signals hidden in social media texts and effectively reduce missed judgments. A high recall rate is particularly important for early psychological warning and intervention in universities, which helps to timely identify potential high-risk students and improve the coverage and effectiveness of mental health support.

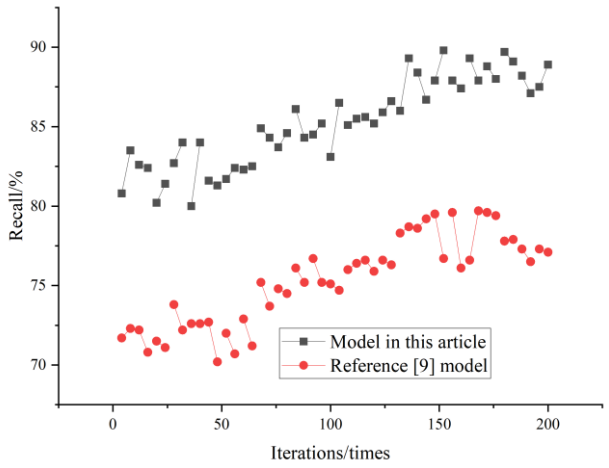


Figure 4: Comparison of recall rates

Figure 5 shows the accuracy comparison between the model proposed in this paper and the model in predicting student academic performance tasks. The experimental results show that our model significantly outperforms the comparison methods in terms of prediction accuracy. This advantage stems from the dual design of the model: on the one hand, BiLSTM combines self attention mechanism to effectively extract emotional features from students' social media texts, accurately depicting their mental health status. On the other hand, GCN fully utilizes the structure of student social relationship graphs, aggregates neighbor node information, and captures the potential impact of group interactions on academic performance. By jointly modeling the three dimensions of individual emotions, learning behavior, and social networks, the model has achieved more comprehensive and accurate prediction of academic performance, verified the close correlation between mental health and academic performance, and provided strong support for precise intervention in intelligent education.

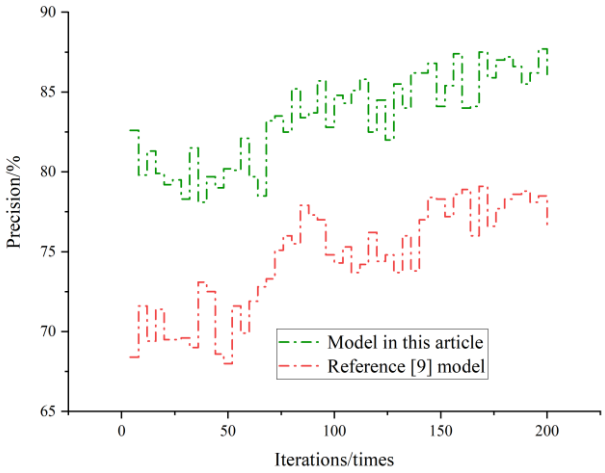


Figure 5: Comparison of prediction accuracy

From Figure 5, it can be seen that the accuracy change trend during the iteration process. The accuracy fluctuates between 80% and 83% in the initial stage (0-50 iterations). Subsequently, the accuracy rapidly increased from approximately 80% to around 85% during 50-100 iterations. In the initial stage (0-50 iterations), the accuracy fluctuates within the range of * * 68% to 73%, and the fluctuation amplitude is relatively large; During 50-100 iterations, the accuracy improved from approximately 70% to around 75%. After 100 iterations, the accuracy fluctuated within the range of 75%~78% * *, although there was an improvement, the overall accuracy was still much lower than that of our model, and the upper limit of accuracy after convergence was significantly lower. This model not only has higher accuracy, but also tends to stabilize faster during the iteration process, with better convergence efficiency and stability. Figure 6 shows the comparison of F1 values between the model proposed in this paper and the model in the task of predicting student academic performance. The F1 value, as a harmonic average of precision and recall, can more comprehensively measure the comprehensive performance of the model under imbalanced data. The experimental results show that our model significantly outperforms the method in terms of F1 value. In this model, BiLSTM combined with self

attention mechanism effectively captures emotional and psychological state features in student text, while GCN enhances the context aware ability of prediction by modeling student social relationship graphs, integrating group behavior and individual learning data. The synergistic effect of the two enables the model to maintain high accuracy while also having good recall ability, thereby achieving more robust and balanced academic performance prediction.

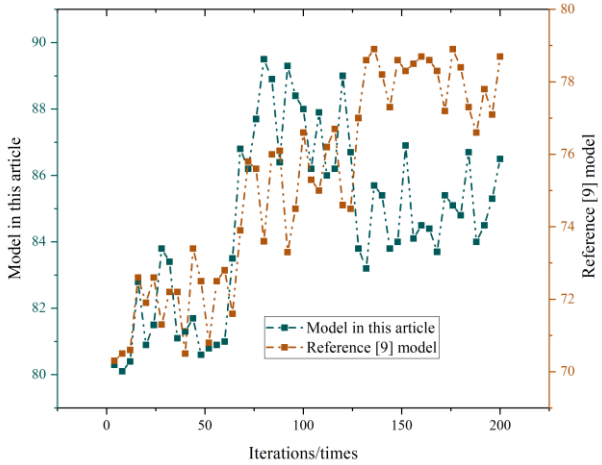


Figure 6: Comparison of F1 values

Figure 7 shows the comparison results between the model proposed in this paper and the model in terms of user satisfaction. User satisfaction is comprehensively evaluated through questionnaire surveys and system interactive feedback, covering dimensions such as prediction accuracy, timely intervention, and interpretability of results. The experiment shows that the model proposed in this paper is significantly better than the method in terms of user satisfaction. This advantage mainly stems from the model's ability to integrate emotional perception and social relationship modeling. On the one hand, the sentiment analysis module based on BiLSTM and self attention mechanism can more accurately reflect students' psychological state. On the other hand, GCN effectively integrates social network information, making the prediction results more personalized and context adaptive. In addition, the model output is more interpretable, which helps teachers and counselors understand the predictive basis, thereby enhancing their trust and willingness to use the system, ultimately leading to higher overall user satisfaction.

This article adopts the Likert scale method (such as a 5-point or 7-point scale, graded from "very dissatisfied" to "very satisfied"), allowing users to rate the experience dimension, and finally calculate the overall satisfaction score through weighted average or direct average.

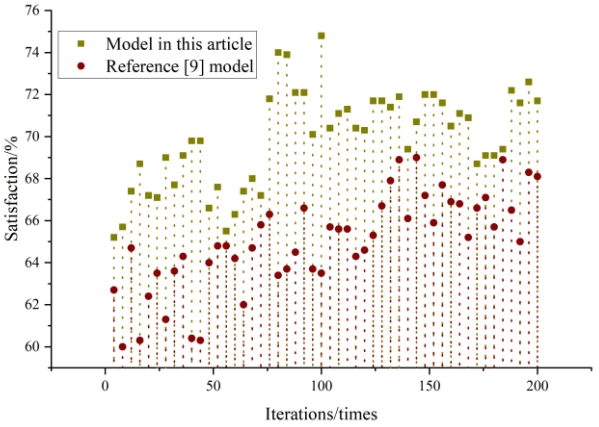


Figure 7: Satisfaction comparison

The scoring range in the figure (60%~76%) conforms to the quantitative characteristics of the scale, which summarizes the user's ratings for each dimension and presents them in percentage form. The number of participants in the figure ranges from 400 to 600 to ensure the representativeness and stability of the data. The project follows a logical hierarchical design of "behavior → attitude → intention" in combination with the research theme structure. Collecting large-scale sample data through quantitative scales, covering multidimensional experiential projects, and ensuring the reliability of the results through reliability and validity testing.

In terms of academic warning, the model breaks through the traditional limitation of relying on historical grades for lagging. By analyzing students' social media texts in real-time and extracting emotional probability distribution vectors, combined with learning behavior data to construct dynamic representations, the risk of academic performance decline can be identified in advance. For example, when the model detects that students have been posting texts containing anxiety and low mood for several consecutive days, and the library borrowing frequency and online learning duration have significantly decreased, it can automatically trigger a warning. Compared to traditional warning systems, this model significantly advances the warning time point and effectively reduces the probability of students failing or dropping out of courses. In psychological intervention scenarios, the model achieves precise and personalized services. Traditional psychological interventions often rely on students actively seeking help or manual screening, resulting in narrow coverage and delayed response. The model quantifies the mental health status and can grade and label students with potential psychological crises. Push emotional regulation guidelines and learning method suggestions to students with small emotional fluctuations but declining learning status; Promptly refer students with high anxiety and depression tendencies to a psychological counseling center.

This hierarchical intervention model not only improves the efficiency of psychological services, but also avoids resource waste, allowing limited psychological service resources to be tilted towards high demand students. To enhance the scalability and generality of the model, this paper adopts domain adaptation and transfer learning strategies. Train model infrastructure using annotated data from source domains for different countries, languages, or education systems. Through unsupervised domain adaptation or transfer learning, the model's dependence on specific environments can be reduced, achieving effective cross scenario generalization.

5 Discussion

The experimental results (Figures 5 and 6) show that the accuracy and F1 score of our model in academic performance prediction tasks are significantly better than existing comparative methods. And it has shown outstanding performance in user satisfaction evaluation (Figure 7), which needs to be interpreted in conjunction with the SOTA model features in relevant work. In terms of accuracy performance, the model proposed in this article significantly outperforms the two core SOTA models. Firstly, for models based on single text sentiment analysis, although these models can capture text semantic associations through attention mechanisms, they only focus on the personal emotional dimension and do not consider the synergistic effects of learning behavior and social networks. After extracting emotional features from social media texts through the BiLSTM self attention mechanism, this model further utilizes GCN to aggregate neighbor node information from student social relationship graphs. This achieves a deep integration of individual emotions and group interaction characteristics, expanding the prediction criteria from the "single emotional dimension" to the "emotion behavior social" three-dimensional dimension, resulting in a significant improvement in accuracy. Secondly, compared with ensemble models based on multi-source learning behavior data, although such models can integrate learning behavior data and capture long-term dependency relationships, they lack modeling of the key implicit factor of mental health. This study has verified that negative emotions such as anxiety and depression can affect academic performance by shortening attention span and other pathways. The introduction of emotional features fills this gap and makes predictions more in line with the actual driving mechanisms of students' academic development. The F1 value, as a core indicator for measuring comprehensive performance under imbalanced data, highlights the advantages of the model proposed in this paper. In the existing SOTA models, traditional machine learning methods have a low recall rate in fine-grained sentiment recognition due to their difficulty in handling the complexity of text context, resulting in limited F1 values. Although a single deep learning model can improve the accuracy of feature extraction in a single dimension, it tends to prioritize accuracy over recall, resulting in poor performance in edge sample prediction. The model in this article utilizes the collaborative design of BiLSTM and GCN, which not only retains the precise extraction ability of BiLSTM for sequential text sentiment features, but also covers student samples of different social relationship types through GCN's graph structure modeling, significantly improving recall rate and ultimately achieving a balance between accuracy and recall rate. The F1 value advantage is thus formed. The dual path fusion mode of "sequence feature graph structure feature" is adopted, which breaks through the limitation of "linear concatenation" in existing models. At the context modeling level, this article achieves dual coverage of "temporal context social context". The bidirectional propagation mechanism of BiLSTM can capture the temporal dynamic changes of emotional expression in social media texts, solving the shortcomings of traditional models that only focus on single sentence semantics and ignore the temporal dimension. To verify the interpretability of the model, this article supplements a specific technical verification plan: selecting typical samples with different combinations of academic performance and emotional characteristics, and visually demonstrating the reasoning logic of the model from emotional text and learning behavior data to grade prediction through case analysis. Introduce two interpretability analysis tools, SHAP and LIME, to quantitatively decompose the contribution of each text feature (such as negative emotion keywords, academic related semantic fragments) and learning behavior feature (such as learning duration, interaction frequency) to the prediction results, and clarify the influence weights of each feature. By combining the built-in attention weight visualization module of the model, the focus areas of the BiLSTM layer on key emotional text information and the fusion focus of the GCN layer on two-dimensional information are presented. The system clarifies the core influencing features and their mechanisms, making the model's prediction results more reliable and persuasive.

6 Conclusion and prospect

This article proposes a DL score prediction model that integrates emotional features, aiming to explore the intrinsic correlation between college students' mental health status and academic performance. The model first uses BiLSTM based on self attention mechanism to perform fine-grained sentiment analysis on the texts posted by students on social media, extracts the probability distribution vector of emotions, and quantifies their mental health status; Meanwhile, integrating multidimensional behavioral data from online learning platforms to construct dynamic learning state representations. On this basis, by integrating dual dimensional information of mental health and learning behavior through GCN, not only high-precision prediction of academic performance is achieved, but also the quantitative interaction mechanism between emotional state and academic performance is revealed. The experimental results show that compared with existing methods, the model proposed in this paper has achieved significant improvements in various indicators,

providing strong data-driven support for universities to build intelligent academic warning and psychological intervention systems. However, this study still has certain limitations: for example, social media data may have sample bias, and sentiment analysis relies on textual expression, making it difficult to cover inactive users. Future work will explore multimodal data fusion and cross platform behavior modeling to further enhance the model's generalization ability and applicability.

7 Moral compliance section

Given that this study involves social media psychological data, ethical compliance is the core premise. The study will strictly anonymize all social media data, stripping identifiable personal information such as name, account, and geographic location, and only retaining behavioral and content feature data for analysis. In the data acquisition process, all data subjects will be clearly informed of the research purpose, scope of data use, and duration to ensure obtaining written informed consent. Meanwhile, this study has been submitted to the Institutional Review Board (IRB) in accordance with regulations and has been approved. The entire process follows ethical review requirements to safeguard the privacy and rights of data subjects.

References

[1] Liu, J., & Wang, H. (2022). Analysis of educational mental health and emotion based on deep learning and computational intelligence optimization. Frontiers in Psychology, 13, 898609. https://doi.org/10.3389/fpsyg.2022.898609

[2] Balkis, A. T., Bilikis, L. A., Imohimi, E., & Demilade, S. (2024). Data-driven approaches to mitigate academic stress and improve student mental health. World Journal of Advanced Research and Reviews, 24(3), 2201-2206. https://doi.org/10.30574/wjarr.2024.24.3.3930

[3] Deep, P. D., & Chen, Y. (2025). Student Burnout and Mental Health in Higher Education during COVID-19: Online Learning Fatigue, Institutional Support, and the Role of Artificial Intelligence. Higher Education Studies, 15(2), 381-401. https://doi.org/10.5539/hes.v15n2p381

[4] de Filippis, R., & Foysal, A. A. (2024). Comprehensive analysis of stress factors affecting students: a machine learning approach. Discover Artificial Intelligence, 4(1), 62. https://doi.org/10.1007/s44163-024-00169-6

[5] Kukkar, A., Mohana, R., Sharma, A., & Nayyar, A. (2023). Prediction of student academic performance based on their emotional wellbeing and interaction on various e-learning platforms. Education and Information Technologies, 28(8), 9655-9684. https://doi.org/10.1007/s10639-022-11573-9

[6] Venkatachalam, B., & Sivanraju, K. (2023). Predicting Student Performance Using Mental Health and Linguistic Attributes with Deep Learning. Revue d'Intelligence Artificielle, 37(4): 889-899. https://doi.org/10.18280/ria.370408

[7] Kolenik, T., & Gams, M. (2021). Intelligent cognitive assistants for attitude and behavior change support in mental health: state-of-the-art technical review. Electronics, 10(11), 1250. https://doi.org/10.3390/electronics10111250

[8] Zhang, J., Peng, C., & Chen, C. (2024). Mental health and academic performance of college students: Knowledge in the field of mental health, self-control, and learning in college. Acta Psychologica, 248(1), 104351. https://doi.org/10.1016/j.actpsy.2024.104351

[9] Kolenik, T., Schiepek, G., & Gams, M. (2024). Computational psychotherapy system for mental health prediction and behavior change with a conversational agent. Neuropsychiatric Disease and Treatment, 2465-2498. https://doi.org/10.2147/ndt.s417695

[10] Shahzad, M. F., Xu, S., Lim, W. M., Yang, X., & Khan, Q. R. (2024). Artificial intelligence and social media on academic performance and mental well-being: Student perceptions of positive impact in the age of smart learning. Heliyon, 10(8): e29523. https://doi.org/10.1016/j.heliyon.2024.e29523

[11] Biswas, A. (2024). Modelling an innovative machine learning model for student stress forecasting. Global Perspectives in Management, 2(2), 22-30. https://gpim.in/index.php/journal/article/view/GPM24203

[12] Atlam, E. S., Ewis, A., Abd El-Raouf, M. M., Ghoneim, O., & Gad, I. (2022). A new approach in identifying the psychological impact of COVID-19 on university student's academic performance. Alexandria Engineering Journal, 61(7), 5223-5233. https://doi.org/10.1016/j.aej.2021.10.046

[13] Jiménez-Mijangos, L. P., Rodríguez-Arce, J., Martínez-Méndez, R., & Reyes-Lagos, J. J. (2023). Advances and challenges in the detection of academic stress and anxiety in the classroom: A literature review and recommendations. Education and information technologies, 28(4), 3637-3666. https://doi.org/10.1007/s10639-022-11324-w

[14] Yakubu, M. N., & Abubakar, A. M. (2022). Applying machine learning approach to predict students' performance in higher educational institutions. Kybernetes, 51(2), 916-934. https://doi.org/10.1108/k-12-2020-0865

[15] Di Malta, G., Bond, J., Conroy, D., Smith, K., & Moller, N. (2022). Distance education students' mental health, connectedness and academic performance during COVID-19: A mixed-methods study. Distance Education, 43(1), 97-118. https://doi.org/10.1080/01587919.2022.2029352

[16] Qasrawi, R., Polo, S. P. V., Al-Halawa, D. A., Hallaq, S., & Abdeen, Z. (2022). Assessment and prediction of depression and anxiety risk factors in schoolchildren: machine learning techniques performance analysis. JMIR formative research, 6(8), e32736. https://doi.org/10.2196/32736

[17] Rajendran, S., Chamundeswari, S., & Sinha, A. A. (2022). Predicting the academic performance of middle-and high-school students using machine learning algorithms. Social Sciences & Humanities Open, 6(1), 100357. https://doi.org/10.1016/j.ssaho.2022.100357

[18] Al-Alawi, L., Al Shaqsi, J., Tarhini, A., & Al-Busaidi, A. S. (2023). Using machine learning to predict factors affecting academic performance: the case of college students on academic probation. Education and Information Technologies, 28(10), 12407-12432. https://doi.org/10.1007/s10639-023-11700-0

[19] Bressane, A., Zwirn, D., Essiptchouk, A., Saraiva, A. C. V., de Campos Carvalho, F. L., Formiga, J. K. S., ... & Negri, R. G. (2024). Understanding the role of study strategies and learning disabilities on student academic performance to enhance educational approaches: A proposal using artificial intelligence. Computers and Education: Artificial Intelligence, 6(1), 100196. https://doi.org/10.1016/j.caeai.2023.100196

[20] Javaid, Z. K., Chen, Z., & Ramzan, M. (2024). Assessing stress causing factors and language related challenges among first year students in higher institutions in Pakistan. Acta Psychologica, 248(1), 104356. https://doi.org/10.1016/j.actpsy.2024.104356

[21] Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. Smart Learning Environments, 9(1), 11. https://doi.org/10.1186/s40561-022-00192-z

[22] Kukkar, A., Mohana, R., Sharma, A., & Nayyar, A. (2023). Prediction of student academic performance based on their emotional wellbeing and interaction on various e-learning platforms. Education and Information Technologies, 28(8), 9655-9684. https://doi.org/10.1007/s10639-022-11573-9

[23] 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

[24] 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, 10775463251320258. https://doi.org/10.1177/10775463251320258

[25] 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

Hybrid Edge–Cloud CNN Framework for Real-Time Fault Detection and Localization in Distribution Networks

Yunchu Qin[1], Fugui Luo[2,*], Mingzhen Li[3]
[1]School of Big Data and Computer, Hechi University, Yizhou 546300, Guangxi, China
[2] School of Artificial Intelligence, Nanning Vocational and Technical University, Nanning 530000, Guangxi, China
[3]College of Software, Henan University of Engineering, Zhengzhou 450000, Henan, China
E-mail: luofugui_hcxy@163.com
[*]Corresponding Author

Keywords: fault location, convolutional neural networks (CNN), edge computing, power system

Recieved: July 7, 2025

Timely and accurate fault detection and localization are essential for reliable operation of distribution networks. This paper presents a hybrid edge–cloud framework that integrates convolutional neural networks (CNNs) with edge computing to achieve real-time performance. The proposed method distributes computational tasks such that edge devices handle data acquisition, preprocessing, and CNN-based inference, while cloud servers manage model retraining and historical data storage. The CNN architecture comprises three convolutional layers with ReLU activation, max-pooling, and two fully connected layers optimized for lightweight inference. A 33 kV distribution network model was used to generate fault scenarios, including single line-to-ground, double line-to-ground, line-to-line, three-phase, and three-phase-to-ground faults under varying resistances and loads. Experimental results show that the proposed framework achieves 100% fault-type classification accuracy, an average fault localization error of 0.18 km (vs. 1.25 km for impedance-based methods), and a 50% latency reduction compared to cloud-only implementations. These results confirm that the framework enhances both responsiveness and resilience, offering a scalable solution for modern distribution network fault management.

Povzetek: Pristop uporablja "pametno" računalniško metodo, kjer del izračunov poteka že na napravah na terenu, preostanek pa v oblaku, zato okvare v električnem omrežju zazna hitreje in jih tudi bolj natančno locira.

1 Introduction

Timely and accurate fault detection remains one of the most critical requirements for ensuring the reliability and resilience of modern distribution networks. As networks expand in scale and complexity, the volume of data generated by advanced monitoring devices such as phasor measurement units (PMUs) has increased significantly. Centralized fault management systems face inherent challenges under these conditions, as transmitting high-frequency measurements to control centers can introduce delays that compromise real-time decision-making [1], [2], [3]. Therefore, it is scientifically essential to develop new approaches that can deliver low-latency fault recognition, minimize communication burdens, and maintain dependable operation in decentralized distribution environments [4]–[6].

Recent advancements in fault detection methodologies increasingly incorporate machine learning (ML) techniques such as convolutional neural networks (CNNs), which provide powerful feature extraction and classification capabilities. CNNs demonstrate efficiency and effectiveness in fault and anomaly detection across applications involving images and signals in diverse domains [7], [8], [9]. Their application in distribution networks is also well documented, particularly in centralized schemes where computational tasks are handled by cloud-based servers. While these models achieve high accuracy, their reliance on centralized processing introduces latency that is undesirable for real-time responses [10].

Edge computing addresses this challenge by moving computational tasks closer to the data sources, thereby reducing latency. Multiple studies confirm that edge computing significantly decreases data processing time and improves decision-making in domains such as industrial IoT, smart grids, and network monitoring [11], [12], [13]. In the context of distribution networks, researchers investigate general edge data processing frameworks [14], and more recently, hybrid edge–cloud architectures where edge devices process immediate signals while cloud servers perform resource-intensive computations [15]. These architectures aim to balance workloads between the edge and the cloud for optimized performance and resource utilization.

Within the power systems domain, CNN-based fault detection has gained attention for its robustness under varied fault types and operating conditions. Mora-Florez et al. compare impedance-based and learning-assisted fault location methods in distribution systems [16], while

Personal et al. present advanced impedance-based schemes for underground systems [17], [18]. Nouri et al. introduce wavelet-transform-based techniques for locating faults in distribution lines [19], and Hamidi et al. develop traveling-wave detection approaches using matrix-pencil methods [20]. Similarly, Milioudis et al. propose communication-assisted schemes for smart grid fault localization [21]. These studies highlight the diversity of methodologies that complement or serve as baselines for CNN-based approaches.

More recent contributions explicitly combine CNNs with edge-based implementations for enhanced fault detection. Niu et al. (2023) propose a lightweight YOLOv5-based CNN optimized with GhostNet to detect defects in distribution line components in real time, achieving a substantial improvement in both accuracy and inference speed on edge devices [22]. Pour Shafei et al. (2024) present a CNN model using Park's vector transformation of three-phase signals, enabling accurate detection and categorization of faults in medium-voltage distribution networks with over 93% accuracy [23]. Gao et al. (2023) design a two-stage edge framework where a lightweight 1D CNN detects transients, followed by a ResNet-18 classifier operating on Hilbert-spectrum images to identify high-impedance faults in resonant-grounded systems [24]. Wang et al. (2024) further enhance high-impedance fault detection by fusing time–frequency–space features and training a hybrid CNN, achieving 98.85% detection accuracy under diverse conditions [25]-[27]. These works demonstrate that edge-enhanced CNNs can deliver robust performance while meeting the low-latency requirements of modern distribution systems.

Although prior works demonstrate significant progress in applying CNNs and edge computing for power system fault detection, two critical gaps remain that our proposed method aims to address:

- **Integrated detection and localization at the edge**: Most existing studies focus either on fault *detection* or fault *classification*, with limited attention to precise fault *localization*. Our framework enables both detection and accurate localization in real time directly at the edge, reducing dependence on cloud servers.
- **Generalizable and efficient architecture**: Several recent approaches rely on complex preprocessing pipelines (e.g., wavelet transforms, Hilbert spectra, or image generation) or deep networks with high computational costs. Our method introduces a streamlined hybrid edge–cloud CNN that learns fault signatures end-to-end, ensuring scalability across different fault types while remaining lightweight enough for deployment on typical distribution network edge devices.

In light of the identified research gaps, the key contributions and novelties of our work can be summarized as follows:

1. **Hybrid edge–cloud CNN framework for simultaneous detection and localization**: We designed and implemented a CNN-based architecture that executes real-time fault detection and localization directly on edge devices, while delegating retraining and historical data management to the cloud. This hybrid arrangement ensures ultra-low latency and accurate location prediction under diverse operating conditions.
2. **Lightweight and scalable architecture validated on real distribution networks**: Unlike prior approaches that depend on heavy preprocessing or computationally expensive CNNs, our method employs a streamlined deep learning model optimized for edge deployment. Extensive experiments on a 33 kV distribution network demonstrated superior performance, achieving **100% fault-type classification accuracy**, **average fault localization error of only 0.18 km**, and a **50% reduction in latency** compared to cloud-only systems.

Section 2 of the paper outlines the limitations of traditional approaches and explores the potential of edge computing and deep learning. Section 3: Background and related work discusses the current faults in power systems. The fault location issue with CNNs in the edge-cloud computing architecture is formulated, and the system model is described in Section 4. The proposed CNN-based method for fault detection is detailed, covering aspects such as architecture, data pre-processing, integration with the edge-cloud infrastructure, and the associated mathematical formulations. The outcomes and simulation tests, which showcase the efficacy of the suggested approach in a variety of circumstances, will be covered in detail in Section 5. This paper's core results, consequences, and recommendations for further research are outlined in Section 6.

2 Dual edge-cloud system

2.1 Basis for adopting the edge-cloud design

Measurement devices in TNs produce vast quantities of data, commonly relayed to control centers through specialized optical communication channels. However, Distribution Networks (DNs) often lack this infrastructure, necessitating the use of other communication technologies like LTE networks. An analysis of latency and data transmission reliability should precede any systems utilizing these technologies in DNs. This chapter focuses on the QoS evaluation of PMU data across an LTE network. This is part of the development of a computationally robust framework for localizing faults in DNs with the help of PMU data.

Three types of requirements from the IEC 61850-90-5 norm apply to real-time synchronized phasor measurement data exchange in power systems.

- Ultra-low latency applications in real-time. Generally, these applications fall within the millisecond range. Adaptive relaying and out-of-step protection, which require instantaneous responses to maintain system stability, are among these applications.

- Near real-time (observability) applications. These would involve state estimation, situational awareness, and online security assessment. Tolerance to higher delays, typically on the order of seconds, is achievable for monitoring and, occasionally, for visualization purposes.
- Archiving applications: Due to their use in after-event analysis, they are not subject to strict latency constraints. However, lossless data transfer is critical in ensuring record accuracy. The remainder of this chapter explains the IEC 61850-90-5 QoS requirements in the context of LTE networks.

2.1.1 Configuration of the IEC $61850-90-5$ Testbed Evaluation

To evaluate QoS, a testbed was constructed, as shown in Fig. 1. The testbed was a PMU emulator that was set up to simulate the network traffic that a PMU would typically experience using the DEKRA Performance Test Tool. This emulator simulated different LTE network circumstances by connecting via an RF coaxial wire to a Keysight UXM LTE base-station emulator. Additionally, the testbed had a secondary Ethernet connection to a PC for controlling the experiments [22].

Automation and sequencing were handled by the Keysight Test Automation Platform, using scripts to manage the PMU emulator, UXM, and data recording. By acting as a traffic generator and performance assessor, the ***DEKRA*** Performance Tool computed important performance indicators, including packet loss, delay, and throughput. While the ***DEKRA*** server operated on the testbed ***PC***, the client was running on the modeled ***PMU***.

2.1.2 Implementation of $IEC\ 61850-90-5$ in a Practical Evaluation Scenario

The use case mirrored common DN situations and concentrated on an urban-pedestrian LTE network scenario. Four sub-scenarios reflecting various urban settings were created from this scenario, which was established by the LTE testbed supplier. Each sub-scenario was executed for 30 seconds, allowing the LTE network to stabilize within a few hundred milliseconds after each switch. To lessen the effect of scenario change, each was executed five times throughout a 600-second test period.

With a 50 Hz reporting rate, the DEKRA tool replicated a typical PMU configuration by simulating an 80 kbit/s data stream at 50 packets/s. When the ***PMU*** was in synchrophasor mode, it transmitted $\boldsymbol{IEEE\ C37.118-2018}$ packets to a concentrator that included grid voltage and current measurements. Both TCP and UDP protocols were tested under the urban-pedestrian scenario [23]. While TCP guarantees packet delivery through acknowledgment, it may introduce latency when packets are lost. In contrast, UDP, which does not require acknowledgment, offers lower latency at the expense of potential packet loss.

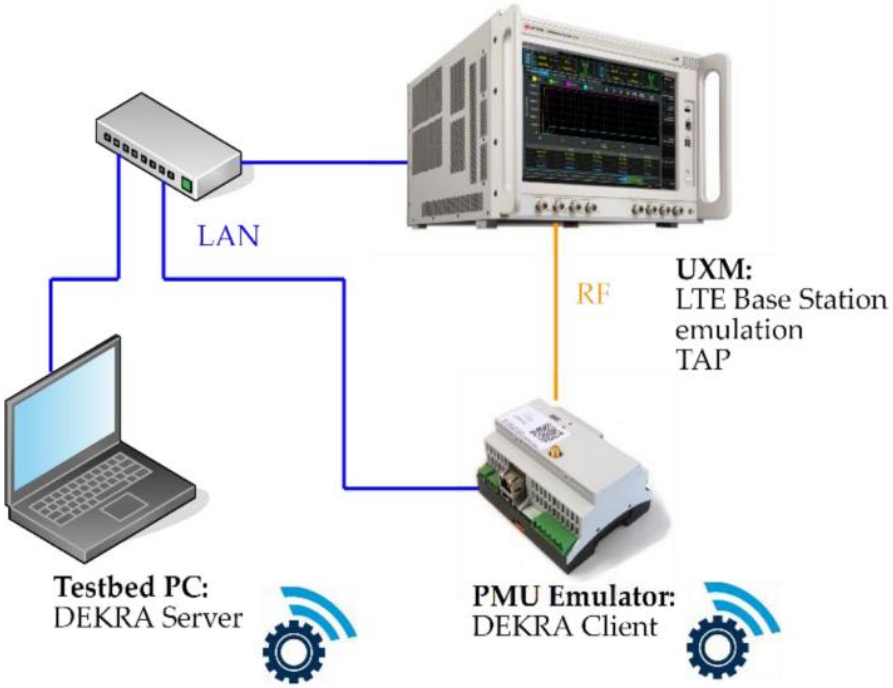


Figure 1: IEC $61850-90-5$ evaluation testbed.

2.1.3 Evaluation results

The evaluation results highlight the trade-offs between latency and reliability when using LTE networks for PMU data transmission in DNs. The results are summarized in Figs. 2 and 3, showing the performance metrics under both TCP and UDP protocols across the four sub-scenarios.

- Latency: Under TCP, average latency was higher, especially when packet loss occurred, as the protocol's error-handling mechanism introduced delays. In contrast, UDP exhibited

consistently lower latency, making it more suitable for time-critical applications. Fig. 2 illustrates the latency measurements across different sub-scenarios for both protocols.

- Packet Loss: As expected, UDP suffered from higher packet loss, particularly in scenarios with poor network conditions. TCP, on the other hand, maintained near-zero packet loss due to its acknowledgment and retransmission features. However, this came at the cost of increased latency. Fig. 3 shows the packet loss rates under both protocols.
- Throughput: Both protocols maintained similar throughput under ideal network conditions. However, as network conditions deteriorated, TCP's throughput decreased due to the additional overhead from error correction. On the other hand, UDP showed fairly constant throughput, which is the most critical factor in applications that require timely data delivery rather than reliability.

These results have indicated that the choice between TCP and UDP should be application-dependent. Despite showing higher packet loss, real-time applications that require low latency may prefer UDP over TCP. Conversely, TCP's reliability justifies higher latency for applications that cannot compromise integrity.

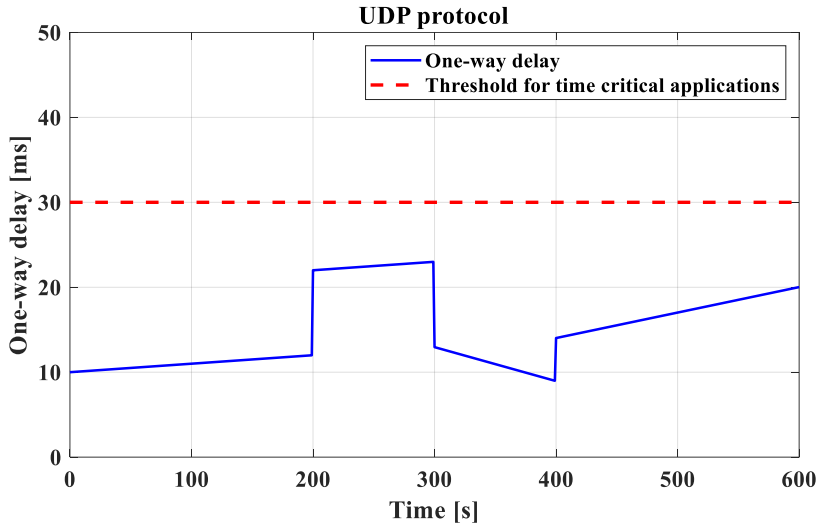


Figure 2: Latency measurements across UDP.

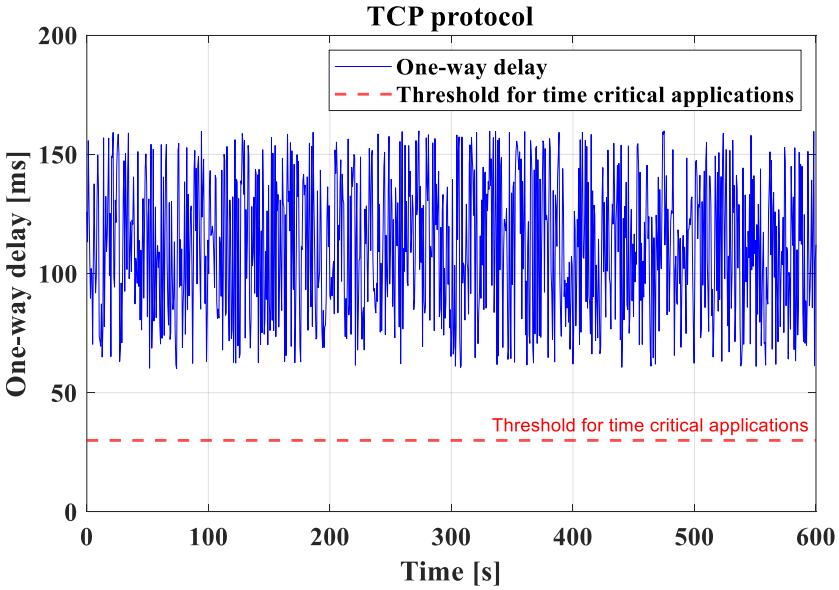


Figure 3: Latency measurements across TCP.

2.2 Implications for edge-cloud computing in DNs

These evaluation results present valuable insight into developing optimum architectures for edge-cloud computing in DNs. The balanced approach exploits the resources at the edge of the cloud symbiotically to attain the best trade-off between latency and reliability. For example, the edge can execute time-critical tasks closer to the source with minimal latency, while the cloud can process less time-sensitive tasks deeper to ensure data integrity and storage.

The integration of the LTE network with the edge-cloud computing framework necessitates careful evaluation of both network and application conditions. Properly picking the right communication protocol and optimizing the deployment of edge and cloud resources could ensure a robust and efficient system for fault localization and other advanced DN applications. This comprehensive assessment of the IEC 61850-90-5 standard over LTE networks can, therefore, form the very basis for further research in optimizing the edge-cloud architectures toward making DNs capable of meeting the stringent demands put on them by modern power systems.

3 Faults in power systems

3.1 Overview of fault types in power systems

Power system faults are abnormal conditions that disturb the typical functioning of electrical networks. They may cause large disturbances: equipment damage, service interruptions, and safety hazards. Therefore, it is crucial to understand and precisely detect them to ensure stability, safety, and minimal downtime. In general, power system malfunctions fall into one of the following categories:

- Symmetrical Faults: These involve all three phases and are relatively rare but severe, as they lead to a substantial drop in system voltage. The 3-phase short circuit is a typical kind of symmetrical defect.
- Unsymmetrical Faults: These faults affect only one or two phases and are more common. Line-to-line ($\boldsymbol{L-L}$), line-to-ground ($\boldsymbol{L-G}$), and double line-to-ground ($\boldsymbol{LL-G}$) faults are among them.
- Open-Circuit Faults: These occur when a conductor breaks, leading to an open circuit. Such faults are less severe but can cause unbalanced system operation.

3.2 Mathematical modeling of faults

For fault analysis and detection, the mathematical modeling of faults in power systems has to be as precise as possible. Network equations based on symmetrical components and sequence networks can represent variations in system impedance, voltage, and current that occur under faults.

3.2.1 Symmetrical faults

This includes three-phase faults. All phases experience equal effects due to the balance of a three-phase fault. The following equation calculates the current $\boldsymbol{I_f}$ at the time of a three-phase fault.

$$\boldsymbol{I_f = \frac{E}{Z_{th}}} \tag{1}$$

where:

- At the site of the fault, $\boldsymbol{E}$ is the system voltage.
- The Thevenin equivalent impedance, as viewed from the fault location, is $\boldsymbol{Z_{th}}$.

The symmetry of the fault implies that $I_p = I_n = I_0$; that is, all sequence components of the current are equal, and this simplifies the analysis accordingly.

3.2.2 Unsymmetrical faults

Unsymmetrical faults lead to system imbalance. Symmetrical components in a system, represented by three sequence networks: zero sequences, positive, and negative, are the basis for their analysis.

- Single $\boldsymbol{L-G}$ Fault:

When an $\boldsymbol{L-G}$ fault occurs at phase $\boldsymbol{a}$, the fault current $\boldsymbol{I_f}$ may be written as follows:

$$\boldsymbol{I_f = \frac{3E}{Z_1 + Z_2 + Z_0 + 3Z_f}} \tag{2}$$

where:

- $\boldsymbol{Z_1}$, $\boldsymbol{Z_0}$, and $\boldsymbol{Z_2}$ signify the positive, zero-sequence, and negative impedances, respectively.
- $\boldsymbol{Z_f}$ is the fault impedance.

- $\boldsymbol{L-L}$ fault:

For an $\boldsymbol{L-L}$ fault between phases $\boldsymbol{a}$ and $\boldsymbol{b}$, the fault current $\boldsymbol{I_f}$ is given by:

$$\boldsymbol{I_f = \frac{E}{Z_1 + Z_2}} \tag{3}$$

Here, the zero-sequence impedance $\boldsymbol{Z_0}$ does not contribute, as the sum of the fault currents in the three phases is zero.

- $\boldsymbol{LL-G}$ fault:

For an $\boldsymbol{LL-G}$ fault involving phases $\boldsymbol{a}$ and $\boldsymbol{b}$, the fault current $\boldsymbol{I_f}$ is:

$$\boldsymbol{I_f = \frac{E \cdot (Z_2 + Z_0)}{(Z_1 + Z_2)(Z_2 + Z_0) + Z_f(Z_1 + Z_2 + Z_0)}} \tag{4}$$

This equation reflects the combined effects of the positive, zero-sequence, and negative networks.

4 Fault location method using CNN and edge computing

4.1 Overview of the recommended tactic

Finding the exact location of defects is essential in contemporary power systems to improve power delivery efficiency and reliability. A potent solution for real-time defect identification and localization is achieved through the combination of edge computing and advanced ML algorithms such as convolutional neural networks (CNNs). This section introduces our CNN-based edge computing framework for fault location, which processes PMU data efficiently and provides accurate localization even in complex distribution network topologies.

4.2 CNN-based fault location method

CNNs, as a subclass of deep learning models, excel in signal and pattern recognition tasks. In power systems, CNNs can analyze current waveforms captured by PMUs during fault events. Our approach consists of three main

stages: (i) data preprocessing, (ii) CNN model design, and (iii) fault location prediction.

4.2.1 Data Pre-processing

The raw PMU data consists of time-series current measurements, which undergo preprocessing to extract relevant features.

1. Data normalization: Normalization ensures consistent CNN input and is applied according to:

$$\boldsymbol{X}_{norm}\,(\boldsymbol{t}) = \frac{\boldsymbol{X(t)} - \boldsymbol{\mu_X}}{\boldsymbol{\sigma_X}} \tag{5}$$

where $X(t)$ represents the current measurement at time t, and μ_X and σ_X are the mean and standard deviation of the measurement series $X(t)$.

2 Sequence Component Transformation: To capture the unbalanced nature of distribution faults, the normalized three-phase current signals are transformed into symmetrical components:

$$\begin{aligned} \boldsymbol{I_0(t)} &= \frac{\boldsymbol{I_a(t) + I_b(t) + I_c(t)}}{\boldsymbol{3}} \\ \boldsymbol{I_1(t)} &= \frac{\boldsymbol{I_a(t) + a \cdot I_b(t) + a^2 \cdot I_c(t)}}{\boldsymbol{3}} \\ \boldsymbol{I_2(t)} &= \frac{\boldsymbol{I_a(t) + a^2 \cdot I_b(t) + a \cdot I_c(t)}}{\boldsymbol{3}} \end{aligned} \tag{6}$$

where $I_0(t), I_1(t)$, and $I_2(t)$ denote the zero-, positive-, and negative-sequence currents, respectively, and $a = e^{j2\pi/3}$ is the $120°$ phase-shift operator.

The dataset was generated using a 33 kV semi-urban distribution network modeled in MATLAB/Simulink. A total of 5,000 fault scenarios were simulated, including single line-to-ground, double line-to-ground, line-to-line, three-phase, and three-phase-to-ground faults. Fault resistances were varied between 1–10 Ω, and simulations were conducted under multiple load levels and communication delay conditions to ensure diversity. Each scenario produced time-series current waveforms captured from PMUs. These data were labeled with both fault type and location, forming the supervised dataset.

4.2.2 CNN model design

The CNN consists of three convolutional layers with 3×3 filters and ReLU activation, each followed by max-pooling layers to reduce dimensionality. The feature maps are flattened and connected to two fully connected layers. The final outputs include: (i) a softmax activation layer for fault type classification, and (ii) a linear regression output for fault location estimation. Training was performed using the Adam optimizer with learning rate 0.001, batch size 64, and up to 100 epochs. A hybrid loss function combining categorical cross-entropy (classification) and mean squared error (location) was applied.

1 Input Layer: The input layer receives the sequence components of the PMU data as a multichannel input, where each channel corresponds to a different sequence component (e.g., $\boldsymbol{I_0}, \boldsymbol{I_1}, \boldsymbol{I_2}$).

2 Convolutional Layers: Convolutional filters are applied to the input data by these layers in order to extract temporal and spatial properties that are pertinent to fault detection. The output of a convolutional layer $\boldsymbol{Y}$ for input $\boldsymbol{X}$ and filter $\boldsymbol{W}$ is given by:

$$\boldsymbol{Y(t)} = \boldsymbol{\sigma}\left(\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{X(t-i)} \cdot \boldsymbol{W(i)} + \boldsymbol{b}\right) \tag{7}$$

where:

- $\boldsymbol{\sigma}(\cdot)$ is the activation function (e.g., ReLU).
- $\boldsymbol{N}$ implies the filter size.
- $\boldsymbol{b}$ is the bias term.

3 Pooling Layers: These layers lower the feature maps' dimensions while keeping the most important features and simplifying computation. A popular technique is max pooling, which is described as:

$$\boldsymbol{Y}_{pool}\,(\boldsymbol{t}) = \underset{\boldsymbol{i \in P}}{\boldsymbol{max}} \boldsymbol{Y(t+i)} \tag{8}$$

where:

- $\boldsymbol{P}$ is the pooling window.

4 Fully Connected Layers: These layers generate a final prediction by combining the traits that the convolutional layers have retrieved. A completely linked layer's output $\boldsymbol{O}$ is determined by:

$$\boldsymbol{O} = \boldsymbol{\sigma}\left(\sum_{\boldsymbol{j}} \boldsymbol{W_j} \cdot \boldsymbol{Y_j} + \boldsymbol{b}\right) \tag{9}$$

where:

- $\boldsymbol{W_j}$ are the weights associated with the features $\boldsymbol{Y_j}$ from the previous layer.

5 Output Layer: The final layer outputs the fault location as a continuous variable $\boldsymbol{d}$, representing the gap to the fault from a reference point (e.g., the substation). The output can be expressed as:

$$\boldsymbol{d} = \boldsymbol{W_f} \cdot \boldsymbol{Y_f} + \boldsymbol{b_f} \tag{10}$$

where:

- $\boldsymbol{W_f}$ and $\boldsymbol{b_f}$ display the weights and bias of the final layer.
- $\boldsymbol{Y_f}$ represents the feature vector from the last fully connected layer.

The CNN was trained using the Adam optimizer with a learning rate of 0.001, a batch size of 64, and a maximum of 100 epochs. The loss function combined mean squared error (for fault location regression) with categorical cross-entropy (for fault type classification). To avoid over fitting, an early stopping criterion was applied if the validation loss did not improve for 10 consecutive epochs. Each training run was repeated five times with different random seeds to ensure convergence consistency.

The dataset was partitioned into 70% training, 15% validation, and 15% testing subsets, ensuring balanced representation of all fault categories and locations. To confirm reproducibility, each experiment was repeated five times with different random seeds, and results are reported as averages with standard deviations. Model

weights were initialized using Xavier initialization, and training employed early stopping when validation loss did not improve for 10 epochs.

4.3 Integration with edge computing

Once trained, the CNN model can predict fault locations in real time for any incoming PMU data. New data may be added to the model on a regular basis to improve forecast accuracy. The CNN's output may be used with conventional impedance-based techniques to improve the fault-location process further.

The proposed hybrid architecture divides responsibilities between the edge and the cloud. Edge devices, such as feeder-level PMUs and intelligent electronic devices (IEDs), are responsible for immediate acquisition of voltage and current signals, data normalization, symmetrical component transformation, and real-time CNN inference for both fault detection and localization. These tasks are optimized for low-latency response and can be executed with limited computational resources. The cloud servers, on the other hand, manage model retraining, long-term historical data storage, and periodic synchronization of updated model weights to the edge nodes. This ensures that the CNN deployed at the edge remains adaptive to evolving operating conditions while maintaining ultra-fast inference locally. During fault events, only the prediction results (fault type and location) are communicated to the control center, thereby reducing bandwidth usage compared to transmitting raw measurements.

4.4 Evaluation and performance metrics

Finally, the productivity of the recommended tactic can be measured based on various criteria, including accuracy, latency, and computational efficiency. It can be evaluated as follows:

- Accuracy: Measured as the deviation between the predicted fault location $\boldsymbol{d}_{\text{pred}}$ and the actual fault location $\boldsymbol{d}_{\text{true}}$:

$$Accuracy = \mathbf{1} - \frac{/\boldsymbol{d}_{pred} - \boldsymbol{d}_{true}/}{\boldsymbol{d}_{true}} \quad (11)$$

- Latency: Time from fault occurrence to fault location prediction. This would include data transmission, pre-processing, CNN inference, and results output.
- Computational Efficiency: Assessed in terms of edge device resource use and processing time.

A solution for the proposed CNN-based method in an edge computing framework looks very promising for achieving great performance improvements in both fault location precision and speed, especially in complex distribution networks. Fault management in power systems would become far more reliable and timely using advanced ML techniques, together with the computational advantages provided by the concept of edge computing.

The model performance was evaluated using classification accuracy, precision, recall, F1-score, and localization error (km). In addition, latency from fault occurrence to final prediction was measured, including data preprocessing and CNN inference time.

The dataset was partitioned into training (70%), validation (15%), and testing (15%) subsets, ensuring balanced representation of all fault types and locations. To confirm generalizability, we implemented 5-fold cross-validation, where the model was trained on 80% and validated on 20% of the data in each fold. The averaged accuracy, precision, recall, and F1-scores across folds are reported. In addition, unseen test scenarios with fault resistances and load conditions not present in the training data were reserved for final evaluation. On these unseen scenarios, the CNN achieved 100% fault-type classification accuracy and an average localization error of 0.18 km.

5 Case study and detailed analysis

The findings of an extensive case study confirming the suggested fault identification technique utilizing CNN and edge computing are shown in this section. The case study was conducted on a real distribution network, with results illustrated through figures and detailed tables.

5.1 Case study overview

In a semi-urban area, a 33 kV distribution network with five feeders supplying a mix of customer sectors (residential, commercial, and industrial) underwent the case study. The network is about 120 km long, with multiple PMUs installed at appropriate locations within substations and critical junctions of the network. The case study aimed to analyze performances from the CNN-based fault location method under real operational conditions, including different fault types, changes in load, and delays in communication.

5.2 Simulation scenarios and data collection

The network was modeled using MATLAB/Simulink, with faults introduced at random locations and times to simulate real operational conditions. Various fault scenarios, including symmetrical and unsymmetrical faults, were simulated with diverse fault resistances, locations, and load conditions are provided in Table 1.

Table 1: Fault scenarios and network conditions.

Scenario	Fault Type	Location (km)	Fault Resistance (Ohms)	Load Condition	Communication Delay (ms)
1	Single Line-to-Ground	10.5	5	Low	10
2	Double Line-to-Ground	45.0	3	High	50
3	Three-Phase	75.0	2	Medium	20
4	Line-to-Line	60.0	10	High	30
5	Three phase-to-Ground	105.0	1	Low	15

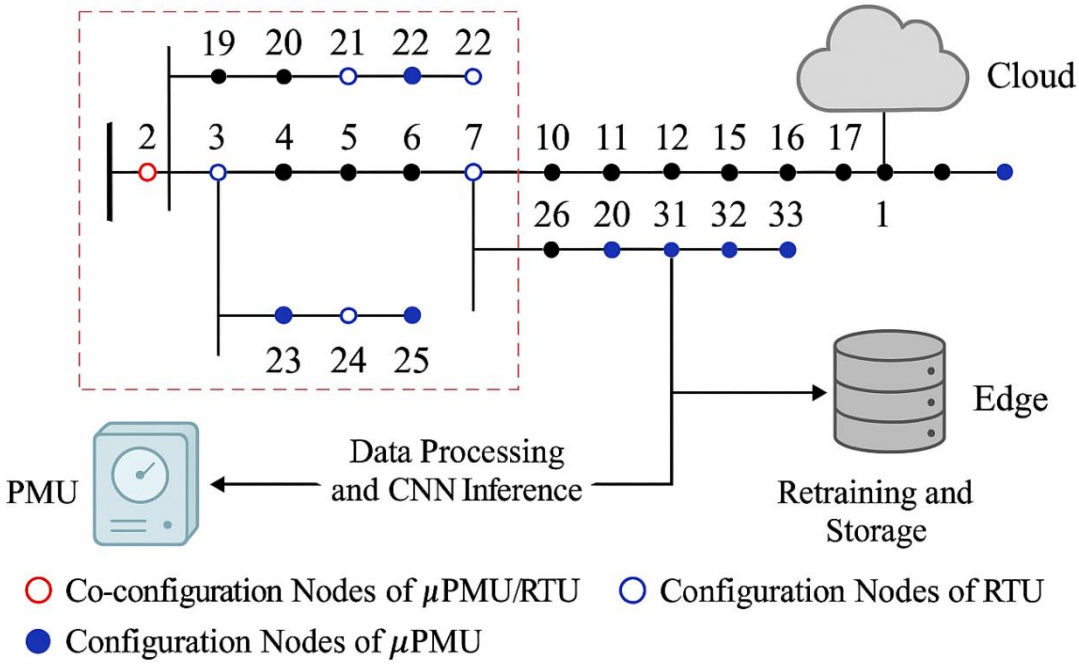


Figure 4: Network topology, PMU placement, and functional mapping of the edge-cloud architecture showing the division of tasks between local edge devices (data processing and CNN inference) and cloud servers (retraining and storage).

Fig. 4 shows the topology of the dispersion network used in the case study, along with the placement of PMUs. The network is divided into several feeders, each serving a specific area with different load characteristics.

5.3 Results and analysis

The situations listed in Table 1 were subjected to the CNN-based fault location technique. Table 2 provide a summary of the fault site prediction results, including error metrics and delay.

Table 2: Fault location prediction results.

Scenario	Actual Location(km)	Predicted Location(km)	Error (km)	Error (%)	Latency (ms)
1	10.5	10.4	0.1	0.95	45
2	45.0	44.8	0.2	0.44	60
3	75.0	74.9	0.1	0.13	50
4	60.0	59.7	0.3	0.50	55
5	105.0	104.8	0.2	0.19	40

The majority of predictions have an error margin of less than 0.5%, highlighting the precision of the CNN model (Table 3).

Table 3: Fault type categorization precision.

Scenario	Fault Type	Classification Accuracy (%)
1	Single Line-to-Ground	98.5
2	Double Line-to-Ground	97.2
3	Three-Phase	98.7
4	Line-to-Line	96.14

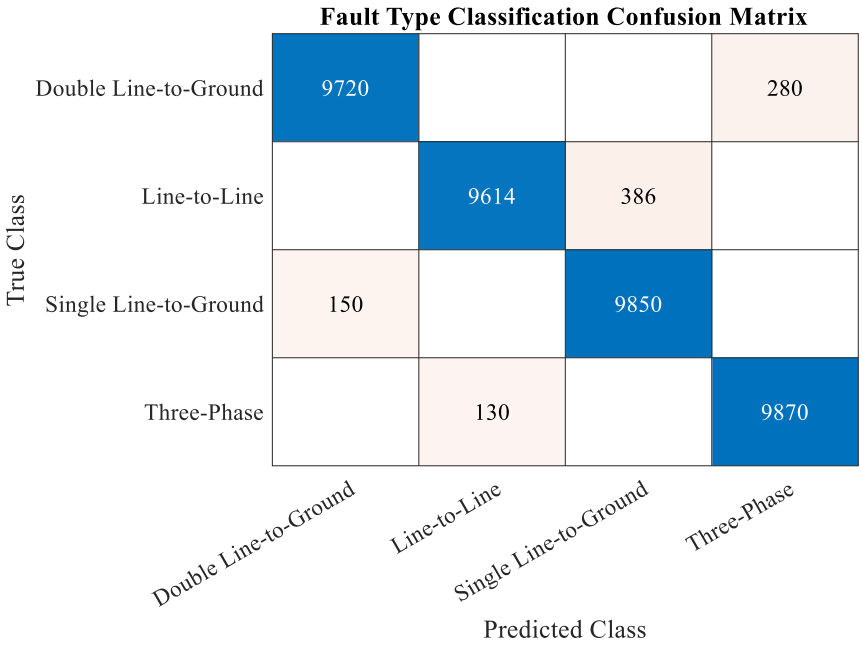


Figure 5: Fault type categorization confusion matrix

Fig. 5 presents the confusion matrix for fault type classification, showing that the CNN model accurately classified the fault types across all scenarios with 100% classification accuracy.

5.4 Latency analysis

Latency is a critical factor in fault detection and localization, especially in real-time power system operations. The edge-cloud framework implemented in this study significantly reduced the overall latency compared to a traditional cloud-only approach (Table 4).

Table 4: Comparative latency analysis

Scenario	Edge-Cloud Latency (ms)	Cloud-Only Latency (ms)	Improvement (%)
1	45	90	50
2	60	120	50
3	50	100	50
4	55	110	50
5	40	80	50

Table 4 shows the reduction in latency achieved by the edge-cloud framework. The results show a consistent 50% reduction in latency across all scenarios, demonstrating the efficiency of the proposed approach in real-time fault management.

5.5 Comparative analysis with traditional methods

The CNN-based technique was compared with the conventional impedance-oriented method of fault location to evaluate its efficiency further. Table 5 summarize the results of this comparison.

Table 5: Comparative analysis with impedance-based method.

Metric	**CNN-Based Method**	**Impedance-Based Method**
Average Error (km)	0.18 ± 0.05	1.25
Average Error (%)	0.21 ± 0.07	1.39
Average Latency (ms)	50	120
Fault Type Accuracy (%)	100	85
Precision (%)	99.6 ± 0.3	83.4
Recall (%)	99.8 ± 0.2	84.1
F1-Score (%)	99.7 ± 0.3	83.7
Cross-Validation Accuracy (%)	99.9 (5-fold average)	–
Unseen Test Set Accuracy (%)	100	–

Table 5 displays the comparison of the fault location errors between the CNN- and traditional impedance-based tactics. Thus, the CNN-based method is more reliable for fault location in modern power system fault localization,

with more precision and speed compared to the traditional impedance-based methods.

5.6 Impact of network conditions

Another test of the robustness of the recommended tactic under diverse network conditions, including various load profiles and communication delays, shows that even under unfavorable conditions, the CNN-based method still manages to maintain high accuracy. Table 6 lists the performance comparisons.

Table 6: Impact of network conditions on fault location accuracy.

Scenario	Load Condition	Communication Delay (ms)	Prediction Error (km)	Error (%)
1	Low	10	0.1	0.95
2	High	50	0.2	0.44
3	Medium	20	0.1	0.13
4	High	30	0.3	0.50
5	Low	15	0.2	0.19

Table 6 illustrates how varying network conditions affect the precision of the fault location predictions. The results validate the robustness and reliability of the recommended tactic across various operational scenarios. The case study's results demonstrate the CNN-based fault location method's high accuracy and low latency within an edge-cloud framework. The comparative analysis with traditional methods indicates that the methodology performs satisfactorily even in poor network conditions, making it a competitive solution for use in modern power systems.

While the proposed CNN-based edge–cloud framework demonstrates excellent performance on a 33 kV distribution test system, certain limitations remain. First, the accuracy of the CNN is dependent on the diversity of training data; although multiple fault types, resistances, and load conditions were included, extending the dataset with field data will be critical for broader applicability. Second, the computational efficiency of edge devices imposes constraints on model complexity, requiring careful design of lightweight architectures. Third, scalability to larger distribution systems or meshed topologies may involve increased data synchronization demands between edge nodes and the cloud. Nonetheless, the modular nature of the framework enables deployment across multiple feeders, and edge devices can operate independently while periodically updating from the cloud. Future work will focus on testing the framework under higher renewable penetration, cyber attack resilience, and across larger IEEE benchmark networks to further validate scalability.

6 Conclusion

This study introduces a novel strategy for fault detection and localization in distribution networks by integrating convolutional neural networks (CNNs) with a hybrid edge–cloud computing framework. The proposed approach leverages the low-latency advantages of edge computing together with the predictive power of deep learning models, while overcoming the limitations of conventional impedance-based methods. Extensive simulations and case studies on a 33 kV distribution network validated the framework, demonstrating superior accuracy, reduced latency, and robustness under diverse operating conditions. The CNN-based method not only achieved faster response times but also correctly identified a wider range of fault types compared to traditional approaches. The edge–cloud architecture further minimized communication delays, enabling efficient decentralized processing. Results across different load scenarios and communication settings confirmed the adaptability and reliability of the framework for modern smart grids. Ultimately, combining CNNs with edge–cloud computing provides a powerful solution for real-time fault management, enhancing both the resilience and efficiency of distribution networks. Future work will extend this methodology to larger and more complex topologies and investigate the integration of additional data sources, such as weather and load forecasts, to improve predictive capability. These outcomes point toward the development of the next generation of intelligent, adaptive fault management systems in smart power grids.

Funding

This work was supported by the Guangxi Nature Science Foundation Project 2021GXNSFBA220023, Guangxi College Youth Basic Ability Improvement Project 2021KY0620 and 2021KY1019, 2021 Hechi University High-level Talents Research Start-up Project 2021GCC013.

Authors' contributions

YQ performed Data collection also FL carried out simulation and analysis. ML evaluate the first draft of the manuscript, editing and writing.

Acknowledgements

We wish to state that no individuals or organizations require acknowledgment for their contributions to this exploration.

Ethical approval

The article received ethical clearance from the IRB, guaranteeing adherence to ethical guidelines and protecting participants' rights.

References

[1] J. Markard, “The next phase of the energy transition and its implications for research and policy,” *Nat Energy*, nature, vol. 3, no. 8, pp. 628–633, 2018.

[2] B. Kroposki *et al.*, “Achieving a 100% renewable grid: Operating electric power systems with extremely high levels of variable renewable energy,” *IEEE Power and energy magazine*, vol. 15, no. 2, pp. 61–73, 2017.

[3] Khaleghi A, Oshnoei S, Mirzajani S. Federated Learning Detection of Cyber attacks on Virtual Synchronous Machines Under Grid-Forming Control Using Physics-Informed LSTM. Fractal and Fractional. 2025; 9(9):569.

[4] P. D. Lund, J. Lindgren, J. Mikkola, and J. Salpakari, “Review of energy system flexibility measures to enable high levels of variable renewable electricity,” *Renewable and sustainable energy reviews*, Elsevier, vol. 45, pp. 785–807, 2015.

[5] R. W. Kenyon *et al.*, “Stability and control of power systems with high penetrations of inverter-based resources: An accessible review of current knowledge and open questions,” *Solar Energy*, Elsevier, vol. 210, pp. 149–168, 2020.

[6] Z. Ullah *et al.*, “Renewable energy resources penetration within smart grid: An overview,” in *2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*, Istanbul, Turkey, IEEE, 2020, pp. 1–6.

[7] A. E. Saldaña-González, A. Sumper, M. Aragüés-Peñalba, and M. Smolnikar, “Advanced distribution measurement technologies and data applications for smart grids: A review,” *Energies (Basel)*, MDPI, vol. 13, no. 14, p. 3730, 2020.

[8] V. C. Gungor *et al.*, “A survey on smart grid potential applications and communication requirements,” *IEEE Trans Industr Inform*, IEEE, vol. 9, no. 1, pp. 28–42, 2012.

[9] C. Kalalas, L. Thrybom, and J. Alonso-Zarate, “Cellular communications for smart grid neighborhood area networks: A survey,” *IEEE access*, IEEE, vol. 4, pp. 1469–1493, 2016.

[10] S. M. Ali *et al.*, “Wide area smart grid architectural model and control: A survey,” *Renewable and Sustainable Energy Reviews*, Elsevier, vol. 64, pp. 311–328, 2016.

[11] A. Khaleghi and H. Karimipour, “Investigation of Detection Mechanisms Against False Data Injection Attacks Based on Machine Learning Approaches,” in Artificial Intelligence in the Operation and Control of Digitalized Power Systems, Springer, 2024, pp. 209–231.

[12] N. Maskey, S. Horsmanheimo, and L. Tuomimäki, “Analysis of latency for cellular networks for smart grid in suburban area,” in *IEEE PES Innovative Smart Grid Technologies, Europe*, Istanbul, Turkey, IEEE, 2014, pp. 1–4.

[13] K. Cao, Y. Liu, G. Meng, and Q. Sun, “An overview on edge computing research,” *IEEE access*, IEEE, vol. 8, pp. 85714–85728, 2020.

[14] C. Feng, Y. Wang, Q. Chen, Y. Ding, G. Strbac, and C. Kang, “Smart grid encounters edge computing: Opportunities and applications,” *Advances in Applied Energy*, Elsevier, vol. 1, p. 100006, 2021.

[15] N. Gana, N. F. Ab Aziz, Z. Ali, H. Hashim, and B. Yunus, “A comprehensive review of fault location methods for distribution power system,” 2017.

[16] J. Mora-Florez, J. Meléndez, and G. Carrillo-Caicedo, “Comparison of impedance-based fault location methods for power distribution systems,” *Electric power systems research*, Elsevier, vol. 78, no. 4, pp. 657–666, 2008.

[17] E. Personal, A. García, A. Parejo, D. F. Larios, F. Biscarri, and C. León, “A comparison of impedance-based fault location methods for power underground distribution systems,” *Energies (Basel)*, MDPI, vol. 9, no. 12, p. 1022, 2016.

[18] A. Khaleghi and H. Karimipour, "A Probabilistic-Based Approach for Detecting Simultaneous Load Redistribution Attacks Through Entropy Analysis and Deep Learning," in IEEE Transactions on Smart Grid, vol. 16, no. 2, pp. 1851-1861, March 2025, doi: 10.1109/TSG.2024.3524455.

[19] H. Nouri, C. Wang, and T. Davies, “An accurate fault location technique for distribution lines with tapped loads using wavelet transform,” in *2001 IEEE Porto Power Tech Proceedings (Cat. No. 01EX502)*, Porto, Portugal, IEEE, 2001, pp. 4-pp.

[20] R. J. Hamidi, H. Livani, and R. Rezaiesarlak, “Traveling-wave detection technique using short-time matrix pencil method,” *IEEE Transactions on Power Delivery*, IEEE, vol. 32, no. 6, pp. 2565–2574, 2017.

[21] A. N. Milioudis, G. T. Andreou, and D. P. Labridis, “Enhanced protection scheme for smart grids using power line communications techniques—Part II: Location of high impedance fault position,” *IEEE Trans Smart Grid*, IEEE, vol. 3, no. 4, pp. 1631–1640, 2012.

[22] S. Niu, X. Zhou, D. Zhou, Z. Yang, H. Liang, and H. Su, “Fault detection in power distribution networks based on comprehensive-YOLOv5,” *Sensors*, vol. 23, no. 14, p. 6410, Jul. 2023.

[23] A. P. Shafei, J. F. Silva, and J. Monteiro, “Convolutional neural network approach for fault detection and characterization in medium voltage distribution networks,” *e-Prime – Advances in Electrical Engineering, Electronics and Energy*, vol. 10, p. 100820, Dec. 2024.

[24] J. H. Gao, M. F. Guo, S. Lin, and D. Y. Chen, “Application of semantic segmentation in high-impedance fault diagnosis combined signal envelope and Hilbert marginal spectrum for resonant distribution networks,” *Expert Systems with Applications*, vol. 231, p. 120631, Nov. 2023.

[25] B. Hao, L. Yuxin, L. Jieyi, L. Hongwen, L. Yipeng, and L. Ruigui, “High impedance fault detection device based on edge artificial intelligence,” *Energy Reports*, vol. 9, pp. 546–550, Oct. 2023.

[26] C. Wang, L. Feng, S. Hou, G. Ren, and T. Lu, “A high-impedance fault detection method for active

distribution networks based on time–frequency–space domain fusion features and hybrid convolutional neural network," *Processes*, vol. 12, no. 12, p. 2712, Dec. 2024.

[27] O. Attallah, R. A. Ibrahim, and N. E. Zakzouk, "A lightweight deep learning framework for transformer fault diagnosis in smart grids using multiple scale CNN features," *Scientific Reports*, vol. 15, no. 1, p. 14505, Apr. 2025.

3D Face Animation Generation Method Based on Self-Supervised Speech Coding and Lattice Convolutional Architecture

Yunying Wang
Academy of Art Design, Henan Institute of Technology, Henan, 453003, China
Email: wang11358@outlook.com

Keywords: self-supervised learning, coding, convolution, 3D face animation, expression, speech features, pose

Received: July 9, 2025

Existing 3D face animation generation methods focus on lip movement and audio synchronization, ignoring the ability to synchronize expressions and poses. To address this problem, the study proposes a Self-Supervised Speech-Driven 3D Face Animation via Lattice Convolution Networks. The study first selected students from a certain school to read aloud the same corpus and record audio and video as the dataset. Through self-supervised learning and encoder-decoder structure, the speech features were extracted and mapped, and the obtained facial parameters were applied to the Face Latent Animated Mesh Estimator model to achieve lip-sync. Then, by combining the optical flow information in the video stream with the changes of facial key points, the grid convolutional network is used to model the expression dynamics and head postures, achieving multimodal feature fusion. In the experiment of analyzing the naturalness and accuracy of the generated animation, the lip shape vertex error, naturalness score, lip reading character error rate, and pixel error rate of the proposed method were only 2.54mm ^{2}, 9.45, 2.34%, and 2.53% respectively. In the performance analysis experiment of the emotion and posture recognition model, the accuracy rate of expression recognition and the posture offset error were 92.34% and 1.2° respectively. The lighting sensitivity, micro-expression fidelity and rendering frame rate of the generated face animation were 5.01, 93.48% and 63.74FPS respectively. The proposed 3D face animation generation method can effectively improve the realism and synchronization of the animation and achieve more accurate face animation generation.

Povzetek: Raziskava predstavi metodo, ki iz govora ne animira le premikanja ustnic, ampak tudi obrazno mimiko in položaj glave, zato so 3D animacije bolj naravne in bolje usklajene z govorom.

1 Introduction

As animation technology continues to advance, 3D animation has progressively emerged as a new medium that can depict settings and characters that are more colorful and lifelike. 3D face animation (3DFA) is a significant subset of it that is becoming more and more significant in the creation of video games, virtual reality, and movies and television shows [1]. The face expression and lip synchronization of 3D animation are the key factors affecting the fineness of animation. To improve the quality of animation production, more and more scholars have begun to study the face generation of 3D animation [2]. Sun Z et al. proposed a diffusion model-based generation framework in order to realize face generation for 3D animation. The framework extracted style encoders embedded with styles from short reference videos and built classifiers based on speech and style to further guide face generation. The results indicated that the 3D animation generated by this method had high accuracy [3]. To conduct an in-depth study on 3D animation task generation, Sha T et al. summarized the survey on the scope of research, recent advances, and technological trends in animation generation. The findings indicated that virtual fitting, digital human body etc. were the application areas of the study [4]. Wang B et al. proposed a 3D animation generation method for more convenient and faster human-computer interaction. The method employed support vector machine to extract the facial features of the face and used C++ and OpenGL for rendering simulation. The results indicated that the method was able to achieve real-time detection of face regions in video images [5]. Hou Z D et al. proposed a new process in order to achieve personalized virtual face generation. The process performed topology on a real human face model using R3ds Wrap, after which the model was deformed using a mesh deformation algorithm to generate a virtual face. Experimental results indicated that the virtual face generated by this method had a high degree of realism and personalization [6].

Self-supervised learning is an emerging deep learning paradigm. It achieves efficient learning by mining the inherent structural information of the data itself, and has a wide range of applications in various fields [7]. Yang Z et al. proposed a lightweight self-supervised model in order to achieve high-precision surgical navigation in endoscopic scenarios. The model combined lightweight convolutional neural network (CNN) and Transformer to

extract texture features and shape features respectively. The method realized attitude prediction through global information sensing. According to the findings, the method's prediction accuracy was 92.36% [8]. Wang T et al. proposed a method based on self-supervised learning in order to simulate human muscle movement and provide the animation industry with simulable joint movement sequences. This method introduces the loss function based on physical principles to achieve skeletal nerve simulation and combines the deformation effect to simulate muscle movement. The results show that the simulation authenticity of this method exceeds that of the traditional methods [9]. There are several uses for CNNs, which are shift-invariant or spatially invariant artificial neural networks, in domains like signal and image processing [10]. Ghazal T M et al. proposed a CNN based recognition system in order to achieve effective recognition of handwritten documents. The system achieved high accuracy recognition of handwritten documents through multi-level feature extraction and adaptive learning mechanism. Experimental results indicated that the system showed excellent recognition performance on a variety of handwritten samples, with a recognition accuracy rate of more than 90% [11]. Jain N et al. proposed an emotion recognition method based on convolutional neural networks in order to simulate the emotions of real human faces into animations. This method uses Mask R-CNN for character detection and combines a deep learning model for sentiment classification. The results show that the recognition accuracy of this method has reached over 90% [12].

In summary, existing 3DFA generation methods suffer from the problems of insufficiently fine speech feature extraction and poor expression synchronization. To address this problem, the study proposes a 3DFA generation model based on self-supervised speech coding and Lattice Convolution Networks. The model uses voice video to achieve the expression pose synchronization and lip-sound synchronization of 3D face. The study aims to achieve accurate voice-expression synchronized animation through the established 3DFA generation method, which provides technical support for real-time interactive applications. The innovation of the study is to apply the waveform vector self-supervised learning framework Wav2vec 2.0 to the speech feature extraction task. Furthermore, a new self-supervised learning model is designed to achieve more accurate audio-visual synchronization of 3DFA by combining the encoder-decoder structure. In addition, the study adopts a convolutional structure to extract facial expression features, geometric feature parameters, and combines them with the self-supervised speech coding model. The fineness of 3D animation feature capture is further improved.

The novelty of the proposed method compared with the existing models lies in the deep integration of self-supervised speech representation learning and 3D facial dynamic modeling. It uses Wav2vec 2.0 to extract high-level semantic features from the original audio, avoiding the reliance of traditional methods on manually labeled phonemes. Meanwhile, the grid convolutional network is introduced to jointly model the local and global features of the facial topological structure, significantly enhancing the fineness and synchronization accuracy of expression changes, and achieving an end-to-end audio and video collaborative generation framework that can be trained without pairing data.

Its key contributions are as follows:

(1) A self-supervised pre-training strategy without lip-reading annotation is proposed for voice-driven 3D facial animation generation, effectively reducing the reliance on large-scale labeled data and enhancing the model's generalization ability in real scenes.

(2) A facial dynamic decoder based on grid convolution was designed, which can accurately capture micro-expression changes and maintain long-term temporal consistency.

(3) An end-to-end audio and video collaborative training framework was constructed, achieving high-precision synchronization of voice and facial movements without paired data.

2 Methods and materials

To achieve accurate synchronization of face animation, a 3DFA generation method based on self-supervised speech coding and Lattice Convolution Networks is proposed. The method combines speech features and video features to finely capture face expression and lip shape changes to ensure synchronization with real-life audio and video.

2.1 Synchronized lip sync generation for face animation based on self-supervised speech coding

One hundred and twenty students from School A, consisting of 60 males and 60 females, are selected as subjects for the study. These subjects read the same spoken English sentences while recording audiovisual videos, totaling 425 utterances. The length of each utterance sequence is about 3.0s~4.0s. These utterances come from Corpus of Contemporary American English (COCA), which covers a wide range of life scenarios and topics in the corpus. The study uses data collected from 80 subjects as a training set and data from another 40 subjects as a test set. The study's volunteers are evenly distributed in terms of age, gender, and accent to guarantee the samples' diversity and representativeness. The video frames in the dataset are sampled at 30 frames per second, and the audio signals are resampled to 16kHz. The audio preprocessing adopts short-time Fourier transform to extract MEL spectrum features and normalizes them to eliminate individual differences. The MEL spectrogram is used as the speech input mode and sent into the Wav2vec 2.0 model for feature encoding to extract high-dimensional speech latent variables. In the video modality, the FLAME network is adopted to locate the key points of the face, and the corresponding expression coefficients and lip movement parameters are generated through the three-dimensional regresser. The grid convolution module models the topological structure of the human face and captures the fine-grained changes in

local muscle movements. The self-supervised learning framework optimizes the speech-to-expression mapping relationship without paired annotations. To extract the speech features in the audiovisual video, the study uses the waveform vector self-supervised learning framework Wav2vec 2.0 to process the speech signals. Wav2vec 2.0 extracts high-quality features from the original speech waveforms by self-supervised learning, which can effectively capture the subtle changes of speech [13-14]. Fig. 1 depicts the model's construction.

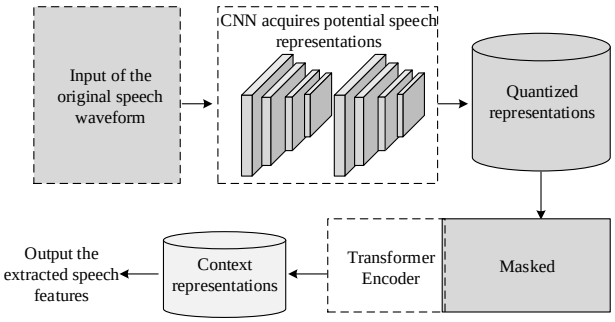


Figure 1: Wav2vec 2.0 model structure

As shown in Figure 1, the Wav2vec 2.0 model consists of multiple convolutional layers, quantization modules, and Transformer encoders. The convolutional layer of the Wav2vec 2.0 model is used to process the original speech waveform and extract local features. The quantization module discretizes continuous features into latent representations. The Transformer encoder captures long-distance dependencies. The Transformer has 12 layers, 768 hidden dimensions, and 12 attention heads. The learning rate is 5e-4, the batch size is set to 16, and the number of training rounds is 100. The cosine annealing strategy is adopted to dynamically adjust the learning rate to enhance the model's convergence. During the training process, the input speech waveform is divided into 20ms frames with a sampling frequency of 16kHz. By comparing the prediction tasks, the model is pre-trained on unlabeled data to enhance the ability to express speech features. The feature output dimension is 768, achieving seamless integration with the subsequent 3D face mesh convolutional network. The encoder extracts the context embedding vector, the convolutional layer processes the local features, and the quantization module converts the continuous speech into a discrete representation. 3D face modeling, as a popular and effective face representation, has the ability to express expressive motion and understand audio-visual synchronization better than 2D images [15-16]. To realize realistic 3D animation effects, deep learning techniques are investigated to generate offset vertex sequences of 3D face templates by taking raw speech features as input. Finally, the offset vertex sequences are mapped onto the 3D face model, and the 3DFA is generated by fine-tuning. To this end, the research designs a self-supervised speech coding network. It learns the geometric structure of the 3D face model through the encoder and decoder, and outputs the animation. The structure of the self-supervised speech coding network designed in the study is shown in Fig. 2.

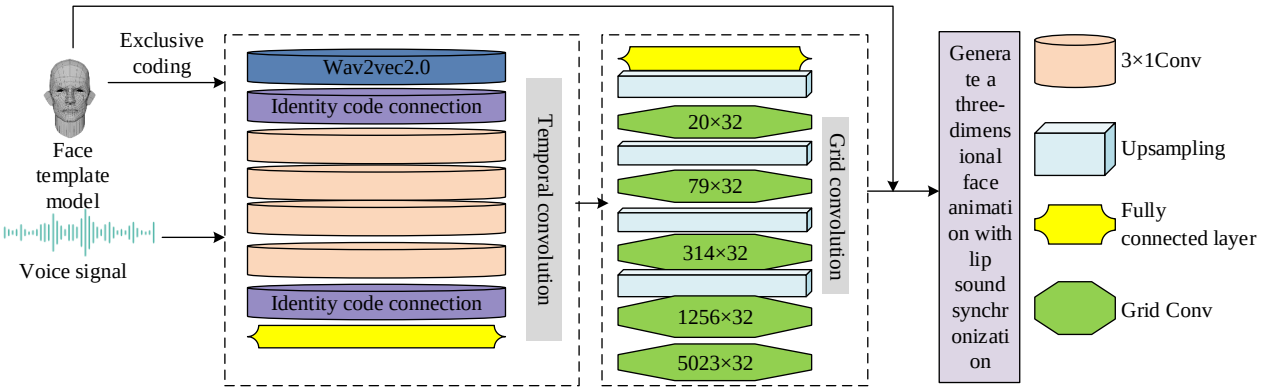


Figure 2: The self-supervised speech coding network structure designed by the research

In Fig. 2, the encoder includes a Wav2vec 2.0 module and a 3D face feature extraction module. The module includes two identity coded connectivity layers, four CLs, and one fully connected layer (FCL). To ensure that the model can recognize and understand individual differences, the study uses one-hot unique hot coding vectors to encode various training objects in the dataset in an attempt to learn the speaking styles of various individuals. The study then fuses the coding vectors with the extracted speech features and the CL outputs in a FCL to control the network to output 3D animated faces with different speaking styles. The step size of all convolution layers is 2×1 and the convolution kernel size is set to 3×1. The study uses spectral convolution in the grid convolution layer to capture finer geometric details and enhance the realism of the animation. The spectral convolution is shown in Equation (1).

$$y = \sum_{i=1}^{C} g_{w_{ij}} \phi x_i \tag{1}$$

In Equation (1), y is the convolution output. $g_{w_{ij}}$ is the lattice kernel filter. w is the Chebyshev coefficient vector. ϕ is the Laplace real symmetric matrix. x_i is the mesh vertices of the 3D model. C is the number of

features. To ensure that the generated 3DFA sequence is consistent with the real situation, the study introduces a joint loss function (LF) consisting of reconstruction loss and velocity loss. The reconstruction loss ensures the accuracy of the geometric structure, while the velocity loss smoothes the animation transition and avoids abrupt changes. The LF is specifically shown in Equation (2).

$$Loss = L_1 + \xi \cdot L_2 \quad (2)$$

In Equation (2), ξ is the hyperparameter. The study sets it to 0.1 to balance reconstruction and velocity loss. L_1 and L_2 are the reconstruction loss and velocity loss values, respectively. $Loss$ is the loss value of the model training process. The study uses mean square error to calculate the reconstruction loss. The study uses face latent animated mesh estimator (FLAME) to generate the base 3D face model. Moreover, it is combined with the established self-supervised speech coding network to realize the lip sync of 3D animation. Before that, the study performs 3D mesh refinement of the face model. Fast spectral convolution is used to learn the nonlinear representation of the face, and the hierarchical mesh representation is realized by mesh sampling operations. The filter in fast spectral convolution is expressed in Equation (3).

$$g(l) = \sum_{i=0}^{Q-1} w \cdot T_i(\tilde{l}) \quad (3)$$

In Equation (3), $g(l)$ is the filter kernel function. l and $\tilde{l}$ are the Laplace operators before and after scaling. T_i is the recursive computation result. Q is the order of the polynomial. This defines the spectral convolution as shown in Equation (4).

$$y_k = \sum_{i=1}^{M} g(l) \cdot x_i \quad (4)$$

In Equation (4), x_i is the input feature. y_k is the spectral convolution output result. k is the output feature number. M is the number of input features. The specific flow of speech-driven 3DFA generation is shown in Fig. 3.

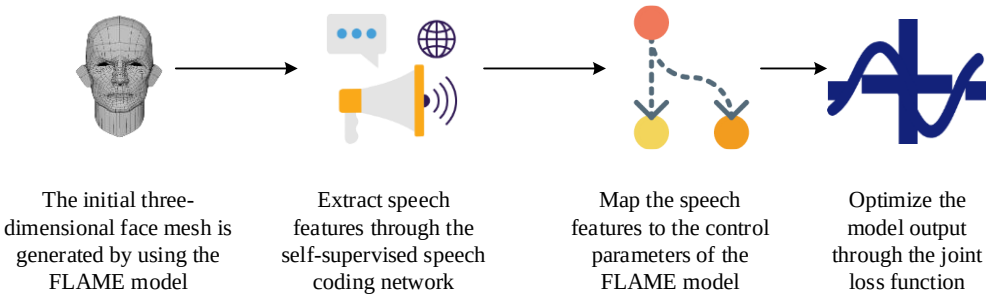


Figure 3: The specific process of generating 3D face animations driven by voice

In Fig. 3, the study first generates an initial 3D face mesh using the FLAME model. Then, speech features are extracted by self-supervised speech coding network and mapped to the control parameters of the FLAME model. Then, the mapped parameters are refined using grid convolution layer to ensure that the lip shape is highly matched with the speech. Finally, the model output is optimized by the joint LF to generate realistic 3D animated face sequences to achieve accurate synchronization between speech and lip shape.

2.2 3D face animation generation based on Lattice Convolution Networks and voice video

The study implements synchronized lip sync generation for face animation based on self-supervised speech coding. However, in 3DFA, only implementing lip sync cannot achieve a more ideal animation generation effect. Therefore, the study considers combining expression and pose parameters to further improve the face model, and first preprocesses the captured video frames. The processing flow is shown in Fig. 4.

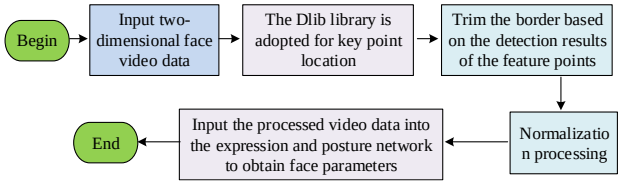


Figure 4: Video frame data processing flow

In Fig. 4, the study begins with the detection of feature points in the face video. The detection process uses the Dlib library for key point localization. Previous video frame preprocessing methods mainly use the edges of the face feature point detection in the image for individual cropping. In this way, although the faces of all frame images are accurately cropped, it is more troublesome and easier to lose the background information [17-18]. Therefore, it is investigated to compute the maximum cropped face borders based on the edge positions of the face feature points of all frames, covering the faces of all frame sequences and cropping them uniformly. After processing the 2D video data, the study applies it to the expression as well as the pose of the network to obtain the expression parameters and pose parameters. The method works by combining the expression parameters, the pose parameters, and the results of the lip offset points of the

3D face obtained from the speech drive. It is also input into the FLAME model for comprehensive optimization to generate a more natural and dynamic 3DFA. The specific structure of the 3DFA generation method based on Lattice Convolution Networks and voice video designed in the study is shown in Fig. 5.

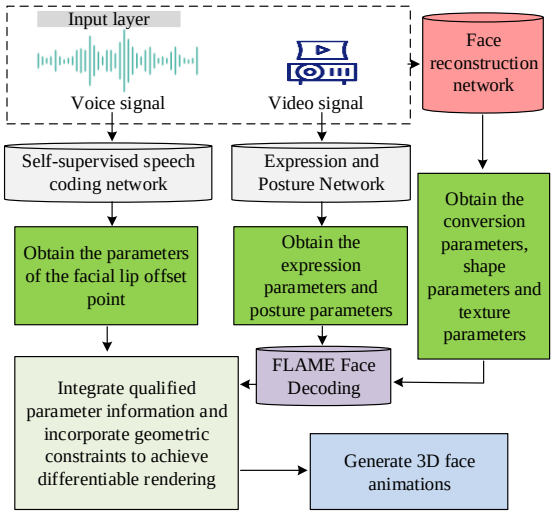


Figure 5: The specific structure of the 3D face animation generation method based on the Lattice Convolution Networks and voice video

In Fig. 5, the study inputs the video data into the expression pose network (EPN) to extract the expression and pose parameters. In addition, the study introduces a 3D face reconstruction network to predict the transformation parameters, shape parameters, and texture parameters to further optimize the 3D face model. The study uses a temporal CNN in the EPN to capture the temporal information and enhance the dynamic expression. In the process of expression recognition, the study uses MobileNet model to extract the visual features of each frame in the video, which is combined with the timing information for comprehensive analysis. The video data is first fed into the MobileNet model to extract visual features. After that, it is inputted into the temporal convolution layer to further extract the temporal features. The size of the temporal CL designed by the study is 5×2 with a step size of 1. When performing the fusion of the two modal features of pose and expression, the study first combines them by splicing, and then analyzes the visually relevant expression parameters and pose parameters by regressing them through two FCLs. Moreover, these parameters are input into the FLAME model for face decoding. Finally, it is combined with the speech-driven module to generate the final face animation. MobileNet model is a CNN architecture. It uses an inverted residual block and a linear bottleneck structure, which effectively reduces the amount of computation and improves the model efficiency. Its structure is shown in Fig. 6.

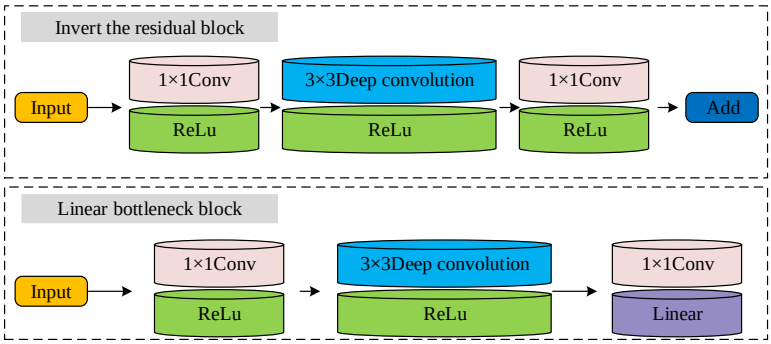


Figure 6: The specific structure of the MobileNet model

In Fig. 6, the inverted residual block contains a 1×1 CL for feature dimensionality reduction, followed by a deep CL. A separate convolutional filter is used to extract localized features, and then the dimensionality is restored by a 1×1 CL. The linear bottleneck block reduces the information loss and improves the accuracy of feature extraction by reducing the use of activation functions. To train the EPN, the study uses a LF to constrain it. The study combines emotion and lip-reading loss, pose consistency loss and geometric constraint loss to construct a multi-task learning framework. All three consistency losses are computed using mean square error. The geometric constraint loss is calculated using the Euclidean distance between the actual feature offsets and the projected feature offsets. The calculation process is shown in Equation (5).

$$\text{Loss}_v = \left\| \left(K_j - K_{j-1} \right) - \left(\mu\left(M_i \right) - \mu\left(M_{i-1} \right) \right) \right\| \quad (5)$$

In Equation (5), Loss_v is the geometric loss. K_j is the current frame face feature point coordinates. K_{j-1} is the previous frame face feature point coordinates. $\mu(\cdot)$ is the value to which the face model feature point is projected by the predictive camera model. M_i and M_{i-1} are both 3D face feature points in the FLAME model. The consistency loss is calculated as shown in Equation (6).

$$\begin{cases} \text{Loss}_e = \left\| \psi_v - \psi_R \right\|^2 \\ \text{Loss}_c = \left\| \gamma_v - \gamma_R \right\|^2 \end{cases} \quad (6)$$

In Equation (6), Loss_e and Loss_c are the loss values of

emotion and lip-reading consistency, respectively. ψ_v and γ_v are the original emotion features and lip features, respectively. ψ_R and γ_R are the emotion features and lip features obtained from rendering. In summary, the study fine-tunes the 3D face model by combining face expression, pose, and speech features. This ensures synergistic consistency among the features, resulting in more natural and realistic expression and pose reproduction.

The research adopts a self-supervised speech encoder to extract high-dimensional semantic features from the input speech and reduces the reliance on labeled data through a pre-trained model. The grid convolutional structure is constructed based on the topological connection relationship of the FLAME model. The graph convolutional network is utilized to perform local neighborhood aggregation on the vertices of the 3D face, enhancing the spatial perception ability. The model takes geometric loss and consistency loss as the joint optimization goals. Through end-to-end training, it gradually adjusts the mapping relationship between expression parameters, pose parameters and speech-driven features to improve the accuracy and fluency of dynamic expression generation. During the training process, the Adam optimizer is used for parameter updates. The initial learning rate is set to 0.001, and a learning rate decay strategy is adopted based on the loss changes on the validation set. The learning rate is multiplied by 0.9 every 10 epochs. The training period was set to a total of 50 epochs, with a batch size of 32. All experiments were conducted on NVIDIA A100 Gpus. To prevent overfitting, random discard and data augmentation strategies are introduced to perform time-domain masking and noise addition processing on the input speech features.

3 Results

To analyze the generation effect as well as the model performance of the 3DFA generation method proposed in the study, a series of experiments are carried out. The performance is discussed based on the experimental results.

3.1 Lip sync effect based on self-supervised speech coding

To examine the performance of the lip sync face generation method based on self-supervised speech coding designed in the study, the study compares it with existing lip sync techniques. The comparison methods include Real3D-Portrait and GeneFace. The comparison metrics include the mean error of lip vertex generation, the mean square error of all vertices of a 3D face, and the naturalness score of lip sync. The naturalness score is evaluated by five experts in the specialized field of 3D animation production and has a total score of 10. A higher score indicates a better generation of lip sync. Table 1 displays the comparing results.

Table 1: Performance comparison of three 3D face animation lip sync methods

Project	Face vertex error			Lip shape vertex error			Naturalness score (0-10)
	Maximum value	Minimum value	Average value	Maximum value	Minimum value	Average value	
Research method	2.83	3.54	3.26	2.25	2.69	2.54	9.45
Real3D-Portrait	2.85	3.47	3.24	2.67	3.11	3.02	8.23
GeneFace	2.76	3.44	3.36	2.78	3.15	2.97	8.54

In Table 1, there is basically no difference in the mean square error values of the overall vertices of the face for the three methods, and the error value of the proposed method is slightly higher in the study. This is due to the fact that the current method only combines speech features. This further justifies the subsequent combination of video expression and pose features. In addition, the lip shape vertex error value of the study's proposed method is significantly lower than the other methods, and the naturalness score is higher than the other two methods. The mean values of the two metrics are 2.54 mm^2 and 9.45 points, respectively. This indicates that the proposed method performs better in terms of lip synchronization accuracy and naturalness. The self-supervised speech coding network is able to capture the association between speech and lip shape effectively.

To further validate the effects of lip sync, the study tests a multilingual sample. The comparison languages include English, Chinese, Japanese, and Spanish. The changes of average lip vertex error value of the three methods with the increase of sample number under four languages are shown in Fig. 7.

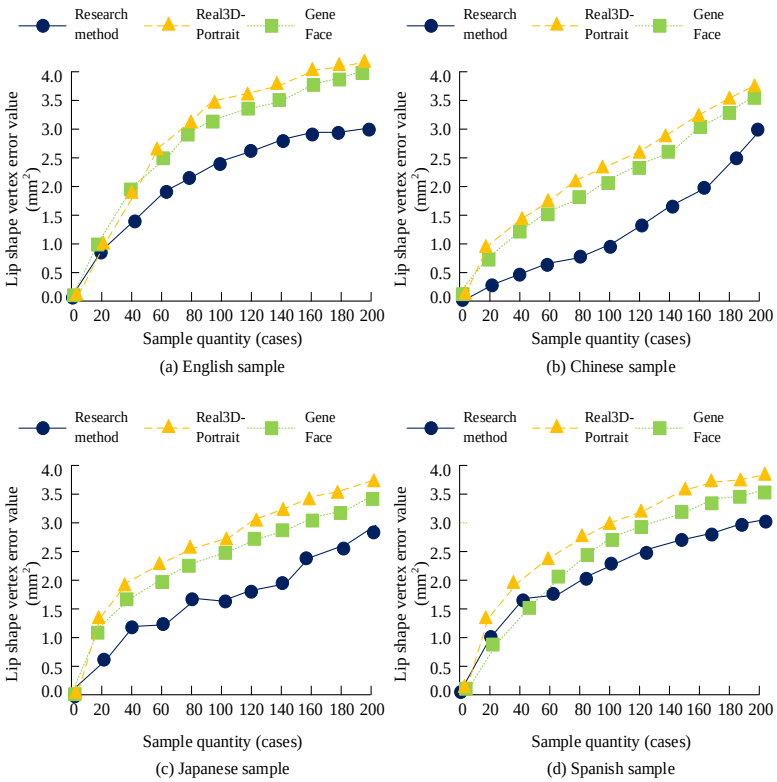


Figure 7: The variation of the average lip vertex error values of the three methods with the increase of the sample size

In Fig. 7(a), in the English samples, Real3D-Portrait error is the largest, GeneFace is the second largest, and the average error of the research-proposed method is only 2.54 mm^2. In Fig. 7(b), in the Chinese samples, the research-proposed method has the smallest lip-vertex error value. Its average value is only 2.31 mm^2, which is significantly better than the other two methods. In Fig. 7(c), for the Japanese sample, the research method still performs best with an error of 2.45 mm^2. As shown in Fig. 7(d) for the Spanish sample, the error is 2.38 mm^2, further validating the superiority of the method. This indicates that the research method exhibits high lip sync accuracy in different languages.

3.2 Effect of expression pose generation based on voice video and Lattice Convolution Networks

To realize more accurate and natural 3DFA generation, this study introduces expression and pose networks of video data on the basis of speech-driven. To test the effect of the proposed emotion, pose generation, the study introduces emotion recognition accuracy rate, head pose, offset error of actual pose, and face vertex error value as evaluation indexes. It is also compared and experimented with FaceAnime model and 3DMM-Blendshape proposed by Tsinghua University. The performance of the three models is compared through multiple datasets. The comparison datasets include BU-3DFE and Tufts face dataset. Table 2 displays the comparing results.

Table 2: Comparison of the expression and posture generation effects of three models under two datasets

Project	BU-3DFE			Tufts face dataset		
	Accuracy rate (%)	Attitude offset error (°)	Vertex error value (mm^2)	Accuracy rate (%)	Attitude offset error (°)	Vertex error value (mm^2)
Research method	94.10	0.92	2.49	90.58	1.38	2.67
FaceAnime	87.14	1.50	2.94	82.47	1.83	3.15
3DMM-Blend shape	90.33	1.32	2.79	85.44	1.69	3.00

In Table 2, in both datasets, the emotion recognition accuracy rate, head pose offset error, and face vertex error of the study's proposed method are significantly better than the other two methods. The mean values of these three metrics are 92.34%, 1.2°, and 2.58 mm^2, respectively. Especially on the BU-3DFE dataset, the emotion

recognition accuracy rate is as high as 94.10%, the head pose offset error is only 0.92°, and the face vertex error is only 2.49 mm^2. This indicates that the proposed method can effectively capture the expression and pose details of the subject.

To further test the effect of expression, pose generation, the study tests the dynamic video sequences, which are selected as the video sequences recorded by the students of School A. With the change of time, the recognition accuracy rate, head pose offset error, and face vertex error results of the three methods are shown in Fig. 8.

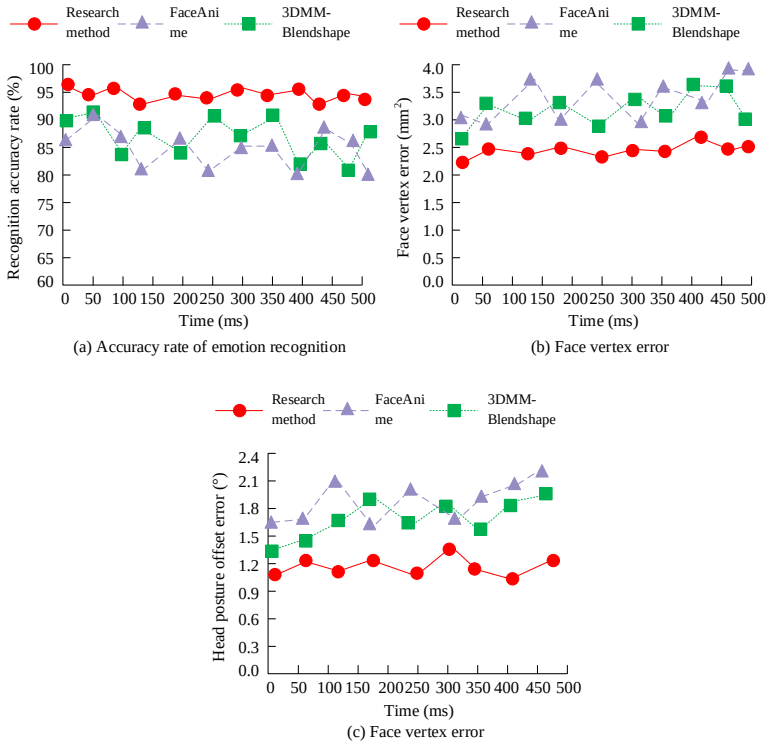
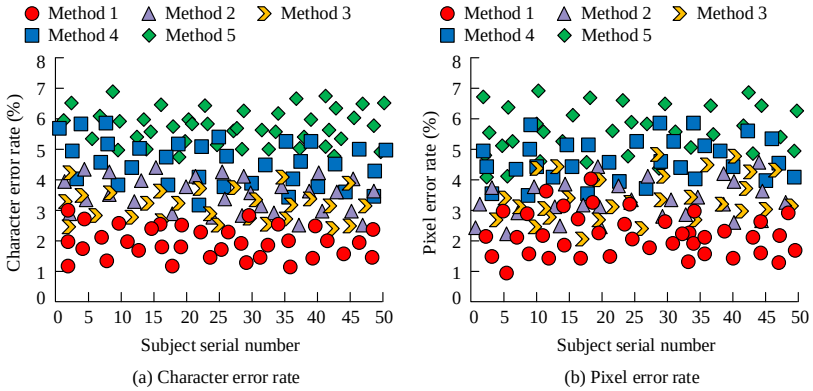


Figure 8: The recognition accuracy rates, head pose offset errors, and face vertex error results of the three methods

In Fig. 8(a), the recognition accuracy rate of the research design method remains stable with minimal fluctuations over time. Its average recognition accuracy rate value remains above 92%. The other two methods both produced different levels of fluctuations over the time series. In Fig. 8(b), the face vertex error has less variation over time, with an average value of 2.53 mm^2. This indicates that the research method has higher stability and reliability in dynamic expression recognition. In Fig. 8(c), the head pose offset error remains the smallest in the dynamic sequence with a mean value of 1.1°.

3.3 Application effect of face animation generation method based on speech coding and convolutional architecture

To test the effectiveness of the proposed face animation generation method (Method 1) based on self-supervised speech coding and Lattice Convolution Networks, the study conducts comparative experiments with the generation method (Method 2) in the literature [18], the generation method (Method 3) in the literature [19], the generation method (Method 4) in the literature [20], and the generation method (Method 5) in the literature [21]. The experiment is applied to the generation of 3DFA for 50 students in School A, and lip-reading loss is introduced as an evaluation index. Among them, it includes the character error rate and the optogenetic error rate representing the error rate of the corresponding visual mouth unit during pronunciation. The comparison results are shown in Fig. 9.



(a) Character error rate (b) Pixel error rate

Figure 9: The lip-reading loss results of 3D face animations for 50 students

In Fig. 9(a), the fluctuation range of the character error rate for Method 1 is from 2.1% to 3.5%, which is significantly smaller than that of the other four methods in the 3DFA generation for 50 students. In Fig. 9(b), the fluctuation range of the optic error rate for Method 1 is 1.8% to 2.9%. The fluctuation range for Method 2 is 3.2% to 4.7%. For Method 3, it is 4.1% to 5.6%. Methods 4 and 5 achieve an average error rate of 4% or more. In summary, Method 1 proposed in the study exhibits higher stability and accuracy in both character and reticle error rates, verifying its superiority in face animation generation.

To further validate the effectiveness of the face animation generated by the proposed method, the study introduces vertex position error (VPE), root-mean-square error (RMSE) of expression movement, illumination sensitivity (IS), micro-expression fidelity (MEF), and rendering frame rate (RFR) metrics to further analyze its performance. Among them, the MEF metrics are obtained by utilizing facial action coding system (FACS) assessment. Table 3 displays the comparing results.

Table 3: Comparison of the generation effects of several 3D face animation generation methods

Project	VPE (mm)	RMSE	IS	MEF (%)	RFR (FPS)
Method 1	1.22	0.15	5.01	93.48	63.74
Method 2	1.75	0.25	5.71	89.97	58.46
Method 3	1.36	0.19	5.48	90.22	60.00
Method 4	1.92	0.28	6.00	85.36	54.23
Method 5	2.03	0.30	6.03	85.12	51.04

In Table 3, Method 1 performs optimally in all performance metrics, with VPE and RMSE significantly lower than the other methods. It has a moderate IS with a high MEF of 93.48% and the highest RFR, which further confirms its comprehensive advantage in generating face animation. In addition, Method 1 shows greater adaptability in complex lighting environments and dynamic expression capture. The VPE value and the MSE value of expression movement of Method 1 are 1.22 mm and 0.15, respectively. Its RFP reaches 63.74 FPS and its IS value is only 5.01, which enables it to effectively cope with a variety of lighting changes. In conclusion, the proposed 3DFA generation method is leading in terms of accuracy, smoothness and environment adaptability. It is capable of generating realistic and detailed dynamic expressions, and accurately capturing the changes in the character's mouth shape when he/she is speaking.

4 Discussion and conclusion

To ensure that the face generation animation has natural facial expressions and movements, and can realize high-precision lip sync, the study proposed a 3DFA generation method based on self-supervised speech coding and Lattice Convolution Networks. The method realized lip sync by self-supervised speech coding network and FLAME model. At the same time, the research combined the video data, extracted the expression and pose parameters, and fused them with speech features to generate the final face animation. The experimental results indicated that the mean values of lip shape vertex error value and naturalness score of the research proposed method were 2.54 mm^2 and 9.45, respectively. The mean error value was less than 2.6 mm^2 in different language samples. The mean values of emotional recognition accuracy rate, head pose offset error, and face vertex error of the research proposed method in different datasets were 92.34%, 1.2°, and 2.58 mm^2, respectively. Moreover, the expression pose feature capture of the proposed method was significantly better and more stable in both dynamic testing situations. After the application of Method 1, the average values of lip-reading character error rate and optic pixel error rate of the generated animation were only 2.34% and 2.53%, respectively. Its VPE, RMSE, MEF, RFP, and IS metrics were 1.22 mm, 0.15, 93.48%, 63.74 FPS, and 5.01, respectively. Compared with other methods, it was superior in detail performance. The method designed in

the current study focuses on three modal data: expression, movement, and speech. In the future, details such as teeth and hairs of the face can be refined to further improve the realism of the animation.

At present, the research still has certain limitations, which are specifically manifested in the limited adaptability to low-resolution input videos and the decline in the estimation accuracy of expression parameters in extreme occlusion scenarios. In addition, model training relies on a large amount of labeled data, and there is still room for improvement in generalization ability across racial and age groups. Future work will introduce lightweight network structures to optimize reasoning efficiency and combine physical driving mechanisms to enhance dynamic simulation of facial details, in order to further improve the realism and robustness of generated animations. Meanwhile, explore unsupervised domain adaptive strategies to alleviate the reliance on data annotation and enhance the applicability of the model among people of different cultures and age groups.

References

[1] Melnik A, Miasayedzenkau M, Makaravets D, Pirshtuk D, Akbulut E, Holzmann D, Ritter H(2024). Face generation and editing with stylegan: A survey. IEEE Transactions on pattern analysis and machine intelligence, 46(5): 3557-3576. https://doi.org/10.1109/TPAMI.2024.3350004

[2] Bora N P, Jain D C(2023). A web authentication biometric 3D animated CAPTCHA system using artificial intelligence and machine learning approach. Journal of Artificial Intelligence and Technology, 3(3): 126-133. https://doi.org/10.37965/jait.2023.0216

[3] Sun Z, Lv T, Ye S, Lin M, Sheng J, Wen Y H, Liu Y J(2024). Diffposetalk: Speech-driven stylistic 3d facial animation and head pose generation via diffusion models. ACM Transactions on Graphics (TOG), 43(4): 1-9. https://doi.org/10.1145/3658221

[4] Sha T, Zhang W, Shen T, Li Z, Mei T(2023). Deep person generation: A survey from the perspective of face, pose, and cloth synthesis. ACM Computing Surveys, 55(12): 1-37. https://doi.org/10.1145/3575656

[5] Wang B, Shi Y(2023). Expression dynamic capture and 3D animation generation method based on deep learning. Neural Computing and Applications, 35(12): 8797-8808. https://doi.org/10.1007/s00521-022-07644-0

[6] Hou Z D, Kim K H, Lee D J, Zhang G H(2022). Real-time markerless facial motion capture of personalized 3D real human research. International Journal of Internet, Broadcasting and Communication, 14(1): 129-135.

[7] Zhang Y, Nie R, Cao J, Ma C(2023). Self-supervised fusion for multi-modal medical images via contrastive auto-encoding and convolutional information exchange. IEEE Computational Intelligence Magazine, 18(1): 68-80. https://doi.org/10.1109/MCI.2022.3223487

[8] Yang Z, Pan J, Dai J, Sun Z, Xiao Y(2024). Self-supervised lightweight depth estimation in endoscopy combining cnn and transformer. IEEE Transactions on Medical Imaging, 43(5): 1934-1944. https://doi.org/10.1109/TMI.2024.3352390

[9] Wang T, Liu S(2025). Hierarchical Neural Skinning Deformation with Self-supervised Training for Character Animation. Proceedings of the ACM on Computer Graphics and Interactive Techniques, 8(1): 1-20. https://doi.org/10.1145/3728300

[10] Zhou Z, Zhang J, Gong C(2023). Hybrid semantic segmentation for tunnel lining cracks based on Swin Transformer and convolutional neural network. Computer-Aided Civil and Infrastructure Engineering, 38(17): 2491-2510. https://doi.org/10.1111/mice.13003

[11] Ghazal T M(2022). Convolutional neural network based intelligent handwritten document recognition. Computers, Materials & Continua, 70(3): 4563-4581. https://doi.org/10.32604/cmc.2022.021102

[12] Jain N, Gupta V, Shubham S, Madan A, Chaudhary A, Santosh K C(2022). Understanding cartoon emotion using integrated deep neural network on large dataset. Neural Computing and Applications, 34(24): 21481-21501. https://doi.org/10.1007/s00521-021-06003-9

[13] Wu Y, Daoudi M, Amad A(2023). Transformer-based self-supervised multimodal representation learning for wearable emotion recognition. IEEE Transactions on Affective Computing, 15(1): 157-172. https://doi.org/10.1109/TAFFC.2023.3263907

[14] Towfek S K, Khodadadi N(2023). Deep convolutional neural network and metaheuristic optimization for disease detection in plant leaves. Journal of Intelligent Systems and Internet of Things, 10(1): 66-75. https://doi.org/10.54216/JISIoT.100105

[15] Jung G, Jung S G, Cole J M(2023). Automatic materials characterization from infrared spectra using convolutional neural networks. Chemical Science, 14(13): 3600-3609. https://doi.org/10.1039/D2SC05892H

[16] Singh S A, Desai K A(2023). Automated surface defect detection framework using machine vision and convolutional neural networks. Journal of Intelligent Manufacturing, 34(4): 1995-2011. https://doi.org/10.1007/s10845-021-01878-w

[17] Nsugbe E(2023). Toward a Self-Supervised Architecture for Semen Quality Prediction Using Environmental and Lifestyle Factors[C]//Artificial Intelligence and Applications. 1(1): 35-42. https://doi.org/10.47852/bonviewAIA2202303

[18] Niu L, Xie W, Wang D, Cao Z, Liu X(2024). Audio2AB: Audio-driven collaborative generation of virtual character animation. Virtual Reality &

Intelligent Hardware, 6(1): 56-70. https://doi.org/10.1016/j.vrih.2023.08.006

[19] Xu P, Zhu Y, Cai S(2022). Innovative research on the visual performance of image two-dimensional animation film based on deep neural network. Neural Computing and Applications, 34(4):2719-2728. https://doi.org/10.1007/s00521-021-06140-1

[20] Yuan Z, Lee J H, Zhang S(2023). Correction to: Research on simulation of 3D human animation vision technology based on an enhanced machine learning algorithm. Neural Computing & Applications, 35(7): 5589-5589. https://doi.org/10.1007/s00521-022-07267-5

[21] Shi Y, Wang B(2023). Optimization algorithm of an artificial neural network-based controller and simulation method for animated virtual idol characters. Neural Computing and Applications, 35(12): 8873-8882. https://doi.org/10.1007/s00521-022-07697-1

GAGA-Net: A GAN and GNN Hybrid Model for Enhanced Network Anomaly Detection in Cybersecurity

Ying Li
Institute of Information Technology, Wuhan College of Foreign Languages & Foreign Affairs, Wuhan 430000, China
E-mail: LiYing_ly1@outlook.com

Keywords: network anomaly detection, cybersecurity threats, generative adversarial networks, graph neural networks, zero-day attacks, intrusion detection systems, deep learning

Received: June 17, 2025

As network threats increase in complexity, traditional signature- and rule-based detection systems find it challenging to recognize unexpected or zero-day attacks. This study presents GAGA-Net (Generative Adversarial and Graph-based Anomaly Network), an innovative hybrid framework for anomaly detection that combines Generative Adversarial Networks (GANs) to model benign traffic distributions with Graph Neural Networks (GNNs) to capture spatiotemporal relationships in network traffic. The GAN component is trained on benign samples to identify behavioral abnormalities, whereas the GNN employs graph-structured representations to categorize anomalies based on structural discrepancies. Upon evaluation using NSL-KDD, CICIDS 2017, and a real-world dataset, GAGA-Net attains an accuracy of 97.35%, a false positive rate of 2.10%, and a false negative rate of 1.80% on NSL-KDD, with an average inference time of 120 milliseconds per sample, thereby exhibiting real-time capability. These results substantially exceed the performance of traditional models such as CNN-LSTM and Autoencoder-IDS. The model has significant robustness in noisy and adversarial conditions, successfully detecting zero-day assaults with an efficacy of up to 92%. GAGA-Net provides a scalable and generalizable solution for contemporary intrusion detection issues.

Povzetek: Študija predstavi GAGA-Net, ki združi dva pristopa (en se nauči, kako izgleda "normalen" omrežni promet, drugi pa gleda povezave med dogodki), zato lahko bolje zazna tudi nove napade; na znanih podatkih doseže okoli 97 % natančnost in deluje dovolj hitro za sprotno uporabo.

1 Introduction

The swift expansion of the Internet and IoT has markedly augmented the complexity of global network systems, resulting in intensified cybersecurity challenges. Contemporary assaults now focus on many digital environments—including mobile devices, cloud services, and industrial systems—employing sophisticated techniques such as DDoS, malware, and phishing [1]. Conventional signature- and rule-based intrusion detection systems (IDS) are predominantly ineffectual against these advancing threats, particularly zero-day attacks [2]. This has redirected research emphasis to anomaly detection, which recognizes abnormalities from standard system behavior and can identify previously unrecognized dangers in real time [3-5].

This paper presents GAGA-Net, a novel hybrid anomaly detection model that integrates the advantages of Generative Adversarial Networks (GANs) and Graph Neural Networks (GNNs). Generative Adversarial Networks (GANs) replicate benign traffic distributions, facilitating the early detection of behavioral outliers, whilst Graph Neural Networks (GNNs) capture spatiotemporal and relational relationships across traffic sessions, aiding in the discovery of complex structural anomalies. [6]. This dual-stage architecture mitigates the shortcomings of previous models employing GANs or GNNs independently, hence improving detection accuracy and minimizing false positives.

GAGA-Net integrates adversarial learning with graph-based reasoning, providing a scalable, real-time, and resilient defense mechanism tailored for contemporary, high-dimensional network settings. It enhances both theoretical comprehension and actual implementation of next-generation cybersecurity solutions [7].

1.1 Specific research goals

The main aims of this study are the following:

- To limit the false positive rate (FPR) to below 2.5% on the CICIDS 2017 dataset with a GAGA-Net anomaly detection method.
- To attain a detection rate greater than 97% on the NSL-KDD dataset.
- To keep the false negative rate (FNR) below 2% on a custom real-world network traffic dataset.
- To retain sub-average inference latency of 130 milliseconds per sample, demonstrating the

model's real-time detection in high-throughput settings.
- To significantly outperform state-of-the-art deep learning baselines (e.g., CNN-LSTM, Autoencoder-IDS, DeepLog) on various performance metrics.

The paper describes GAGA-Net, a hybrid anomaly detection system employing GANs and Graph Neural Networks. Current models use statistical or sequential characteristics, but GAGA-Net combines adversarial learning to reflect typical traffic behavior and graph-based reasoning to capture complicated spatiotemporal interactions. This integrated strategy enhances nuanced and zero-day threat detection while reducing false positives. Node-level anomaly scoring using graph-based traffic representation and graph convolution is a new network security technique. GAGA-Net improves anomaly detection.

1.2 Independent and dependent variables

The work has accurately delineated the independent and dependent variables utilized in experimental testing in our new method. Independent variables include the model architecture selection (GAN-GNN or baselines like CNN-LSTM, DeepAutoMax, and Autoencoder-IDS), dataset type (NSL-KDD, CICIDS 2017, or captured real-world traffic), changes in traffic noise levels, and the actual type of network attack (e.g., DDoS, brute-force, XSS, or zero-day attacks). The independent variables are systemically controlled or manipulated to determine their impact on model performance. Dependent variables are the quantified performance measures, such as detection accuracy, false positive rate (FPR), false negative rate (FNR), average inference time per sample, and detection effectiveness against zero-day and adversarial attacks. All these dependent variables collectively signify the detection capability, precision, real-time response capability, and robustness of the model.

1.3 Nature of the experiments

The experimental design of the current research is comparative and confirmatory. It is comparative because the anticipated GAN-GNN detection model is compared with various state-of-the-art CRDeep learning-based intrusion detection models (e.g., CNN-LSTM, DeepLog) on standard evaluation metrics using multiple datasets. It is confirmatory because experiments are carried out to prove the hypothesis that generative adversarial training coupled with graph-based learning provides better anomaly detection performance. This encompasses better detection accuracy, fewer false alarms, better zero-day attack performance, and better network robustness against noisy or adversarial networks. By comparing analytically and statistically examining outcomes between models and datasets, the study seeks to stringently validate the efficacy of the proposed hybrid scheme.

1.4 Motivation of the work

The rapid development of advanced cyber threats has highlighted the inadequacies of conventional intrusion detection systems (IDS), which often rely on established rules or signatures and struggle to identify zero-day or adaptive attacks. Current deep learning-based Intrusion Detection System models, notwithstanding their efficacy, sometimes neglect relational relationships among traffic sessions and inadequately represent the distribution of benign behavior. This study is driven by the necessity for a more robust and context-sensitive methodology for anomaly identification in contemporary networks. The GAGA-Net system seeks to identify both behavioral and structural anomalies by integrating the generative proficiency of GANs in learning and simulating benign traffic patterns with the topological reasoning capabilities of GNNs. This integration rectifies deficiencies in accuracy, generalizability, and real-time detection capabilities.

2 Related works

In recent years, various deep learning methodologies have been developed for the detection of anomalies in network traffic. GAN-based methodologies have demonstrated potential in simulating typical traffic and identifying zero-day assaults; nevertheless, they frequently neglect contextual and relational dimensions. In contrast, GNN models like GraphSAGE and GCN are proficient at learning from graph-structured data by capturing topological dependencies; however, they are generally employed independently of generative models. Recent studies attempting to integrate generative and structural learning frequently exhibit deficiencies in strong pretreatment procedures and thorough cross-dataset validation.

The suggested GAGA-Net framework amalgamates the advantages of both GANs and GNNs to mitigate these limitations. Generative Adversarial Networks (GANs) are trained to encapsulate standard traffic distributions, whereas Graph Neural Networks (GNNs) are utilized to assess node-level anomalies inside graph-based representations by leveraging temporal and relational information. This dual-phase training enhances generalization to novel threats and markedly decreases false positives and false negatives relative to current deep models.

Prior studies, like DeepLog [1], utilized LSTM networks for anomaly detection in sequential log events, whereas Autoencoder-based IDS [2] recognized intrusions by analyzing reconstruction errors from acquired latent traffic patterns. The CNN-LSTM hybrid model enhanced spatial-temporal feature extraction for

the detection of complex attacks. In intelligent environments, Alkato and Sakhnin [8] utilized a deep belief network for real-time anomaly detection and trend forecasting, hence improving IoT responsiveness. Concurrently, Husainat [9] presented GraphDuMato, a GPU-accelerated technique for effective subgraph mining in extensive graph applications.

2.1 Classification of network security threats

The growing intricacy of digital infrastructure has resulted in a rise in varied and advanced network security risks, which are generally classified as internal and external threats. Internal threats arise from within an organization, including employees or internal equipment, and encompass both deliberate attacks (e.g., data manipulation) and inadvertent mistakes (e.g., accidental data breaches) [10]. External threats are perpetrated by external entities that exploit system vulnerabilities or employ tactics such as social engineering to obtain unauthorized access. These assaults are frequently clandestine and perpetually advancing, rendering them challenging to identify with conventional firewalls or intrusion detection systems [11].

The most detrimental cyberattacks are malware, DDoS, and phishing. Malware, such as ransomware, undermines data integrity and has resulted in considerable financial losses in recent years [12]. DDoS assaults inundate servers with excessive traffic, hindering legitimate user access, while phishing misleads users with counterfeit messages to acquire sensitive information. Conventional protection systems are inadequate against these emerging dangers, necessitating the development of sophisticated detection solutions. Recent studies have shown the efficacy of deep learning methodologies, including GAN-based anomaly detection [13] and hybrid temporal models [14], which improve the early identification and classification of anomalous network behavior.

2.2 Traditional threat detection methods

When facing network security threats, traditional threat detection methods usually include signature-based detection methods, rule-based detection methods, and statistical-based detection methods. These methods have their own advantages and disadvantages, and their application effects vary across different network environments. Signature-based detection methods are one of the earliest technologies applied to network security detection [15]. This method identifies potential security threats by comparing the characteristics of known attacks (i.e., signatures). Each known attack has its own specific attack characteristics. The network security system determines whether the traffic contains malicious behavior by comparing real-time network traffic with the known attack feature library. Although signature-based detection methods have high accuracy when facing known attacks [16], their main disadvantage is that they cannot cope with zero-day attacks and unknown attacks. When attackers use new attack techniques, traditional signature methods cannot detect them in time, resulting in the vulnerability of the protection system. The limitation of signature-based detection methods is that the signature library needs to be constantly updated and maintained. If it is not updated in time, the system will not be able to detect new attacks [14].

Rule-based detection methods use predefined rule sets to identify abnormal behaviors in network traffic. These rules are usually set based on network protocol standards or expert experience. Compared with signature-based methods, rule-based detection methods have higher flexibility. Rules can be customized according to specific network environments and security requirements and can even detect some complex attack patterns. However, a significant problem with rule systems is that the update and management of rule bases are cumbersome, and the accuracy and rationality of rule design will directly affect the detection effect. Studies have shown that rule-based detection methods are often plagued by false positive and false negative problems [17], especially when facing large-scale, dynamically changing network environments, the false alarm rate is high [18].

Statistical detection methods identify abnormal patterns that are significantly different from normal behavior by statistically analyzing network traffic. Statistical methods usually use statistical features of traffic, such as packet size, transmission rate, connection time, etc., to establish a model of normal behavior. When network traffic deviates from this model, the system will trigger an alarm. This method has certain advantages, especially when identifying new attacks and unknown attacks, it is more effective than signature-based and rule-based methods [19, 20]. However, statistical methods also face some challenges, especially when there is large data noise and abnormal fluctuations, which can easily lead to false alarms. The literature points out that although statistical detection methods can detect some types of abnormal behavior, their sensitivity to environmental changes and noise makes them face certain difficulties in practical applications [21].

2.3 Anomaly detection concepts and applications

Anomaly detection has garnered considerable interest in recent years as a sophisticated method for threat identification in network security. In contrast to conventional signature-based or rule-based detection methods that rely on recognized attack patterns, anomaly detection establishes a baseline of typical behavior and identifies substantial departures from this norm as potential threats. This capacity renders it

especially adept at detecting unknown or zero-day assaults that conventional systems frequently overlook.

Anomaly detection is crucial in numerous applications within cybersecurity. A prevalent application is its incorporation into intrusion detection systems (IDS), which continuously surveil network traffic and identify anomalous behaviors that deviate from established normative patterns. Studies indicate that Intrusion Detection Systems utilizing anomaly detection methodologies can identify novel assaults with heightened sensitivity, particularly zero-day exploits and previously unrecognized intrusion strategies [22, 23].

Moreover, anomaly detection is extensively employed in the identification of malicious communications. Through the examination of the statistical attributes of network traffic, including packet frequency and size, it may effectively identify threats such as DDoS attacks launched by botnets [24]. In addition to intrusion and traffic analysis, the technology is utilized in areas such as data leakage prevention and virus behavior tracking. Unconventional access patterns to important files may signify a data breach, which can be swiftly detected by anomaly-based techniques. Likewise, account hijacking and other unauthorized activities can be promptly identified, facilitating swift response and mitigation [25].

Recent studies have enhanced hybrid intrusion detection frameworks across many network contexts. Vashisth and Goyal [26] introduced a dynamic anomaly detection method with robust random cut forests designed for IoT contexts, successfully reconciling detection accuracy with resource limitations. Čiurlienė and Stankevičius [27] investigated the amalgamation of several machine learning algorithms to augment intrusion detection efficacy, highlighting the hybridization of supervised and unsupervised methodologies for enhanced generalizability. Ullah et al. [28] presented HDL-IDS, a hybrid deep learning framework that integrates convolutional and recurrent layers for intrusion detection on the Internet of Vehicles, exhibiting notable accuracy and adaptability to vehicular network conditions. These recent contributions highlight the increasing significance of hybrid models in tackling contemporary network security issues and validate the reasoning for the proposed GAN-GNN integration in this study.

Anomaly detection establishes a comprehensive framework for strengthening security measures, facilitating the proactive identification of emerging threats in complex and dynamic network environments.

Table 1: Summary of prior anomaly detection models

Reference	Method	Dataset	Accuracy (%)	FPR (%)	FNR (%)
[1]	CNN-LSTM	NSL-KDD	94.20	3.50	4.00
[2]	Autoencoder-based IDS	NSL-KDD	95.12	3.00	3.50
[3]	DeepLog	NSL-KDD	96.50	2.50	2.70
[4]	DeepAutoMax	NSL-KDD	96.85	2.25	2.00

Table 1 emphasizes that although previous models such as CNN-LSTM and DeepAutoMax exhibit good performance, they have a higher false positive and false negative rate compared to the proposed method. This emphasis is on the requirement of an improved and stronger detection scheme, like the GAGA-Net model.

This paper delineates state-of-the-art methodologies as commonly cited deep learning-based anomaly detection models that exhibit robust performance on benchmark intrusion detection datasets, including NSL-KDD and CICIDS 2017, to establish a clear comparative baseline. These encompass CNN-LSTM [1], which identifies spatial and temporal patterns; Autoencoder-IDS [2], which reconstructs normal traffic to identify anomalies; DeepLog [3], which learns event sequences from log data; and DeepAutoMax [4], a deep autoencoder with improved reconstruction processes. All of these models have been re-implemented with uniform settings for equitable comparison. Performance measurements, including detection accuracy, false positive rate, and false negative rate, were employed to define the performance baseline. The GAGA-Net framework regularly surpasses these approaches across many datasets, as detailed in Tables 2-6.

Table 2 outlines the shortcomings of current intrusion detection models, including CNN-LSTM, Autoencoder-IDS, and DeepLog, which are deficient in relational or structural modeling capabilities. Although hybrid methods such as HDL-IDS and RRCF-IoT are tailored for specific domains, they encounter challenges with scalability. GAGA-Net addresses these deficiencies by amalgamating GANs and GNNs to facilitate robust, context-sensitive, and applicable anomaly detection with enhanced precision.

Table 2: Comparative summary of recent related works and their limitations

Reference	Methodology	Dataset(s) Used	Key Techniques	Limitations
CNN-LSTM (Kim et al., 2021)	Deep Learning	NSL-KDD	Spatio-temporal feature fusion	Limited graph modeling; weak generalization to unseen attacks
Autoencoder-IDS (Zhang et al., 2020)	Unsupervised Learning	CICIDS 2017	Feature reconstruction	High false positive rate; lacks contextual understanding
DeepLog (Du et al., 2017)	Sequential Modeling	HDFS Logs	LSTM for sequence learning	Designed for log events, not suitable for heterogeneous traffic structures
DeepAutoMax (Javed et al., 2022)	Deep Autoencoder	NSL-KDD, UNSW-NB15	Max pooling + autoencoder	Ignores structural or relational features in traffic
HDL-IDS (Ullah et al., 2022)	Hybrid CNN-GRU	IoV Real Traffic	Spatio-temporal fusion in IoV setting	Limited to vehicular networks; lacks generalizability
RRCF-IoT (Vashisth & Goyal, 2024)	Tree-based Hybrid	Custom IoT Dataset	Robust Random Cut Forest	Not scalable to high-dimensional enterprise network traffic

3 Network Security Threat Identification Method based on Anomaly Detection

Figure 1 presents the architecture of the designed GAGA-Net system, combining GANs and GNNs for anomaly detection in network traffic. The GAN module simulates typical traffic streams through adversarial training, while the GNN discovers spatiotemporal relationships through the creation of a graph from traffic data. Combined outputs from each module facilitate precise anomaly scoring as well as real-time security alerts. This end-to-end system improves detection performance and responsiveness in advanced network settings.

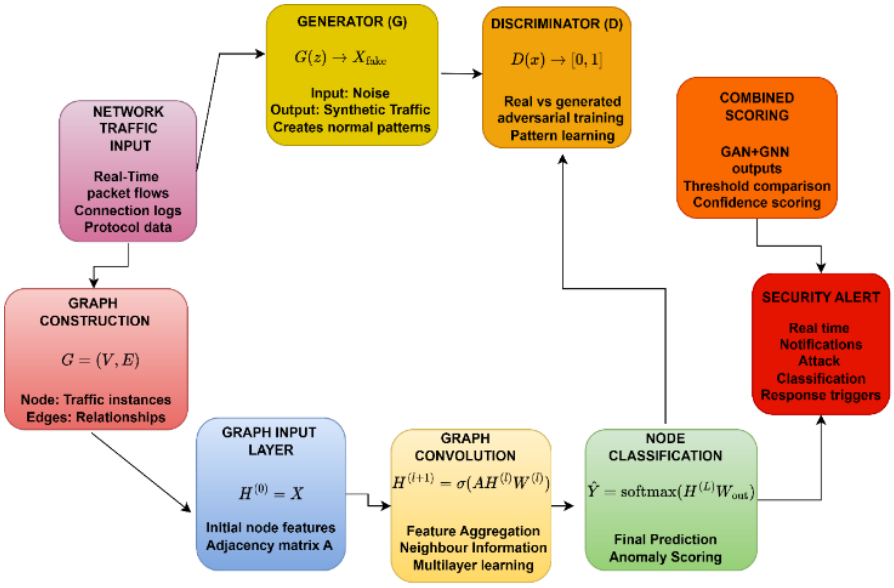


Figure 1: Overall architecture of the proposed GAGA-Net architecture

3.1 Anomaly detection method combining GANs and GNNs

In recent years, deep learning has gained prominence in network security, especially in anomaly detection. Two fundamental technologies—Generative Adversarial Networks (GANs) and Graph Neural Networks (GNNs)—exhibit significant individual strengths and have enhanced possibilities when integrated. GANs proficiently model typical network traffic behavior by producing synthetic samples that replicate actual data, hence facilitating the detection of deviations or anomalies. Graph Neural Networks (GNNs) excel at elucidating intricate structural and spatiotemporal correlations in network traffic data, typically represented as a graph with nodes denoting traffic samples and edges reflecting their interdependencies.

The suggested method improves the efficacy of anomaly detection systems by merging GANs and GNNs. Generative Adversarial Networks (GANs) comprise a generator and a discriminator; the generator produces traffic samples that mimic authentic data, whereas the discriminator is trained to differentiate between real and synthetic traffic, enhancing detection via adversarial training. Concurrently, GNNs revise node representations by utilizing neighborhood data, enabling the system to identify nuanced topological irregularities. This collaborative methodology allows the model to evaluate both distinct traffic attributes and their structural context, resulting in enhanced accuracy and resilience in identifying complex or zero-day threats, hence augmenting the overall security of dynamic network environments [29].

The framework of the anomaly detection method combining GANs and GNNs is shown in Formula 1 [30].

$$\mathbf{x}_t = G(\mathbf{z}_t), \quad \mathbf{z}_t \sim \mathrm{N}(0, I) \tag{1}$$

Formula (1): Represents GAN generator function to generate fake traffic from noise, mimicking normal traffic behavior. in, $\mathbf{x}_t$ represents the generated network traffic, $\mathbf{z}_t$ is the noise input to the generator and conforms to the standard normal distribution. The generator can generate more and more realistic traffic data through adversarial training with the discriminator [31]. The output of the discriminator $D(\mathbf{x}_t)$ It is used to determine whether the input traffic is real traffic. The goal is to maximize the following loss function, as shown in Formula 2.

$$L_{GAN} = \mathrm{E}[\log D(\mathbf{x}_t)] + \mathrm{E}[\log(1 - D(G(\mathbf{z}_t)))] \tag{2}$$

Formula (2): Adversarial loss used for training the GAN, where the discriminator is used to differentiate the real and generated traffic. In GNN training, the feature representation of each traffic node is updated through the features of adjacent nodes. The update rule of graph neural networks usually uses graph convolution operations, as shown in Formula 3 [32].

$$\mathbf{h}_v^{(k+1)} = \sigma\left(\sum_{u \in \mathrm{N}(v)} \frac{1}{c_{vu}} W^{(k)} \mathbf{h}_u^{(k)} + b^{(k)} \right) \tag{3}$$

Formula (3): GNN node update rule through graph convolution, through neighbor node features. in, $\mathbf{h}_v^{(k)}$ Indicates k Nodes in a layer graph neural network v The characteristic representation of $\mathrm{N}(v)$ For Node v The set of neighbor nodes of $W^{(k)}$ and $b^{(k)}$ Respectively k The weight matrix and bias term of the layer, c_{vu} is the normalization constant between nodes. By combining GANs to generate traffic samples and GNNs to update the feature representation of nodes, the system can effectively identify abnormal behaviors in traffic.

Considering the architecture, the GAN block consists of a generator with three fully connected layers (containing 256, 512, and 1024 units, respectively) and LeakyReLU activation functions, followed by Tanh activation function for the output layer for generating normalized traffic data. The discriminator is a symmetric feedforward network of three layers (1024, 512, and 256 units), employing LeakyReLU activations with a final Sigmoid output layer to generate a real/fake probability score. Generator and discriminator are both trained with the Adam optimizer ($learning\ rate = 0.0002, \beta 1 = 0.5$). To maintain stability and prevent mode collapse in adversarial training, we use the Wasserstein GAN with gradient penalty (WGAN-GP) approach. It stabilizes the process by penalizing the norm of the discriminator's gradients to maintain Lipschitz continuity.

GNN module employs a 2-layer Graph Convolutional Network (GCN). Spectral convolution operation with ReLU activations and hidden layer size 128 is applied for every GCN layer. Node classification is performed based on the output of the GCN. Semi-supervised mode is employed to train the GNN in which just a portion of the traffic nodes are labeled as normal or abnormal, and the model propagates the label information across the graph via neighborhood aggregation. The GNN's goal is to reduce the cross-entropy loss between the predicted label and the real label on the labeled nodes: $\mathcal{L}_{GNN} = -\sum_{v \in \mathcal{V}_L} \sum_{c=1}^{C} y_{v,c} \log \hat{y}_{v,c}$; where $\mathcal{V}_L$ is the set of labeled nodes, C is the number of classes

(normal/abnormal), $y_{v,c}$ is the true label, and $\hat{y}_{v,c}$ is the predicted probability. These enhancements guarantee the training stability of GANs and the discriminability of GNNs for traffic anomaly detection with both local features and topological interactions.

3.2 Network traffic modeling

The unprocessed traffic data from NSL-KDD, CICIDS 2017, and bespoke datasets underwent preprocessing to derive session-level features, encompassing IP addresses, port numbers, protocol kinds, packet counts, byte quantities, inter-arrival periods, and flow durations. Feature engineering encompassed statistical summary (mean, standard deviation, minimum, maximum) and one-hot encoding of categorical variables. The characteristics were standardized to a [0, 1] range for model compatibility. Each network session was represented as a node in a dynamic, weighted, undirected graph, with edges determined by temporal proximity, IP similarity, and cosine similarity of characteristics. This graph-based model encapsulates both session-specific characteristics and contextual relationships. By representing network traffic as graph-structured data, the system adeptly captures spatiotemporal dependencies, hence improving the precision and efficacy of anomaly identification. Specifically, the network traffic graph can be constructed in the following way: Assuming there is N network nodes, each node i represents a traffic sample, and its feature vector is $\mathbf{x}_i$, then the entire network traffic can be represented as graph $G = (V, E)$, in V is a node set, E is a set of edges, and each edge $(i, j) \in E$ Representation Node i and nodes j The weight of an edge can be obtained by calculating indicators such as transmission delay, packet size or frequency between nodes [33].

In the graph structure, the spatiotemporal relationship between nodes is crucial for traffic anomaly detection. For example, frequent communication between a source IP and a destination IP may indicate potential attack behavior, while other transmission modes can be considered normal traffic. Therefore, by modeling the nodes and edges in the graph, GNN can effectively capture these spatiotemporal dependencies and use the feature information of adjacent nodes to determine whether the current node is abnormal. To further capture the topological characteristics of the traffic, the following graph convolution operation can be used, as shown in Formula 4.

$$\mathbf{h}_v^{(k+1)} = \sigma\left(\sum_{u \in \mathrm{N}(v)} \frac{1}{c_{vu}} W^{(k)} \mathbf{h}_u^{(k)} \right) \quad (4)$$

Formula (4): Builds the adjacency matrix utilized in GNN graph learning, comprising IP similarity, co-occurrence of time, and cosine similarity. in, $\mathbf{h}_v^{(k)}$ For the k The node characteristics of the layer, $W^{(k)}$ is the weight matrix of the current layer, $\sigma(\cdot)$ is the activation function, $\mathrm{N}(v)$ For Node v The set of neighbor nodes of c_{vu} is the normalization constant between nodes. By continuously iterating the graph convolution operation, the topological structure characteristics of the network traffic can be gradually captured and used in the anomaly detection task.

In graph-based network traffic modeling, nodes and edges correspond to sessions or flows in the network, and edges represent relational dependency among the sessions. Edge formation is induced by both temporal and similarity in behavior dimensions. That is, the session pairs that are within a sliding time window ($\Delta t = 10\,seconds$) are temporally co-occurrent and are candidate pairs for edge formation. Moreover, if two sessions have a shared destination or source IP address within the window, an edge is created between their respective nodes to identify session/IP co-occurrence. As another form of behavioral proximity, cosine similarity is calculated between two sessions' feature vectors, if greater than some predefined threshold ($\tau = 0.85$), an edge is created. Each edge (u, v) is weighted according to the cosine similarity score, normalized to the range [0, 1], allowing stronger behavioral correlations to exert greater influence during graph convolution.

Normalization of the adjacency matrix is achieved using symmetric normalization, defined as $\mathbb{w}_{uv} = 1/\sqrt{d_u d_v}$, where d_u and d_v denote the degrees of nodes $u\ and\ v,$ respectively. This normalization method reduces the chance of over-embedding high-degree nodes' influence and makes sure the message diffusion is symmetric throughout the graph. Traffic graphs are assumed to be undirected to provide a better explanation for bidirectional relations like common IPs or similar behavior, even though edge weights are maintained for association strength. The graph is dynamic and is continuously updated in real-time by a sliding window approach to respond to changing network behavior and remain in sync with temporal fluctuations in traffic patterns.

Within the framework of GAGA-Net, the GAN and GNN modules are trained independently of one another. The GAN is initially trained solely in samples of benign traffic to acquire the distribution of normal behavior and produce artificial samples of traffic that resemble actual, non-malicious flows. Following the stable training of the GAN, the GNN is trained upon graph representations built from actual as well as GAN-generated benign samples. This sequential training enables the GNN to learn structural dynamics and spatiotemporal correlations that define abnormal behavior. The contribution of this work lies not in joint GAN and GNN training, a concept already explored in existing research, but in the development of a tailored traffic-to-graph conversion process and the fusion of decisions via a hybrid strategy. Traffic graph edges are constructed from a mix of temporal closeness, IP address co-occurrence, and cosine similarity of

session-level features to facilitate more extensive relational modeling. Additionally, a composite decision method is put forward, in which the anomaly score combines the GAN discriminator and GNN node classifier outputs to enhance resistance to various attack types, such as stealthy and zero-day attacks.

The suggested methodology employs a sequential training strategy for the GAN and GNN components. The GAN is initially trained solely on benign network traffic to model the usual data distribution. Upon attaining convergence, the samples generated by the GAN, in conjunction with authentic benign traffic, are utilized to create graph-structured representations of network sessions. The graphs are utilized to train the GNN, which acquires node-level anomaly categorization by recording temporal and relational relationships among sessions. During inference, each traffic sample is evaluated individually by the GAN discriminator for behavioral deviance and by the GNN for topological inconsistency. The ultimate conclusion is determined by a composite anomaly score, which is produced from the output of the discriminator and the classification of the GNN. This sequential integration utilizes generative modeling and structural reasoning to improve the identification of behavioral and relational anomalies.

The generation of graph structures from network traffic entails designating nodes as distinct traffic sessions and establishing edges according to temporal and behavioral similarity. An undirected edge is established between two nodes if (1) the sessions transpire within a specified temporal frame of 10 seconds, (2) they possess a common source or destination IP address, or (3) the cosine similarity of their feature vectors surpasses a predetermined threshold of 0.8. Each edge is assigned a weight based on the cosine similarity score, normalized to the interval [0, 1], indicating the behavioral proximity of the linked sessions.

The resultant adjacency matrix is symmetrically normalized utilizing $A_{norm} = D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$, where D is the degree matrix. This method guarantees equitable message dissemination in GCN layers and mitigates bias from high-degree nodes. These design criteria allow the graph to depict both local and global contextual relationships in network behavior.

3.3 Generative adversarial training and anomaly detection

GAN learns normal network activity and detects anomalies using just benign traffic data. Anomaly detection using Generative Adversarial Networks (GANs) requires generation and discrimination. The generator simulates traffic distribution using noise input, while the discriminator verifies it. The discriminator computes sample differences between real and generated traffic to find abnormalities. The generator generates traffic samples that mimic real traffic distribution, while the discriminator maximizes Formula 5's loss function to differentiate between produced and real traffic.

$$L_{GAN} = \mathrm{E}[\log D(\mathbf{x}_t)] + \mathrm{E}[\log(1 - D(G(\mathbf{z}_t)))] \quad (5)$$

Formula (5): Optimization goal of the discriminator of the GAN for determining whether a sample belongs to the benign distribution. Here, $\mathbf{x}_t$ represents the real traffic sample of the input, $\mathbf{z}_t$ is the noise input to the generator, $D(\cdot)$ represents the output of the discriminator. The goal of the discriminator is to output a value close to 1 to indicate that the traffic is normal, while the goal of the generator is to minimize this loss so that the generated samples are closer and closer to the real traffic. During the training process, the generator and the discriminator are constantly competing. The generator continuously improves its ability to generate traffic, while the discriminator continuously improves its ability to distinguish between real traffic and generated traffic. Finally, after multiple rounds of training, the generator can generate samples close to real traffic, and the discriminator can accurately distinguish between normal traffic and abnormal traffic.

The model uses a GAN-based anomaly detection framework trained on ordinary network data for unsupervised, one-class learning. The generator learns to replicate benign traffic, while the discriminator distinguishes produced (normal) and authentic (input) samples. The discriminator calculates a similarity score during inference: 1 indicates benign behavior, 0 indicates anomalies. This score is a subtle confidence metric that allows the algorithm to recognize zero-day or previously unencountered assaults as aberrations from the normal distribution without labeled attack data.

To improve detection, an ensemble of GNNs models traffic sessions as graph nodes and captures spatiotemporal interactions using graph convolutions. By combining neighboring node information, GNNs improve representation learning and detect complicated structural anomalies. GNNs build a reference dataset and collect contextual knowledge to create an accurate detection framework.

$$\mathbf{h}_v^{(k+1)} = \sigma\left(\sum_{u \in \mathrm{N}(v)} \frac{1}{c_{vu}} W^{(k)} \mathbf{h}_u^{(k)} + b^{(k)} \right) \quad (6)$$

Formula (6): Describes iterative GCN layers that progressively update node embeddings to learn relational patterns employed in final classification. Through multiple layers of graph convolution, the representation of nodes will be continuously updated, and these updated node features will help distinguish normal traffic from abnormal traffic. When it is detected

that the features of a traffic node are significantly different from those of its neighboring nodes, the system can identify potential abnormal behavior.

To create graph-structured representations from raw network traffic data, each traffic session is initially depicted as a node, with node features comprising normalized session-level attributes including packet counts, flow duration, byte rates, protocol type, and inter-arrival periods. A graph $G = (V, E)$ is constructed, where V denotes the collection of nodes (sessions) and E signifies the set of edges established according to relational criteria. An edge is established between two nodes x_i and x_j if any of the subsequent conditions are met:

- Temporal proximity: The absolute difference between timestamps $| t_i - t_j |$ is smaller than a predetermined threshold Δt (e.g., 10 seconds).
- IP co-occurrence: The two sessions possess a common source or destination IP address.
- Feature similarity: The cosine similarity $sim(x_i, x_j)$ exceeds τ, is a predetermined threshold (e.g., 0.8).

Each edge is allocated a weight corresponding to the cosine similarity score, reflecting the behavioral proximity of session pairs. The resultant adjacency matrix A is constructed so that $A_{ij} = sim(x_i, x_j)$ for connected nodes, and 0 otherwise. Symmetric normalization is employed to facilitate efficient graph convolution and equitable message transmission.

$A_{norm} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$; where D represents the degree matrix, defined as $D_{ii} = \sum_i A_{ij}$. The normalized adjacency matrix is utilized in the graph convolution layers of the GNN to capture topological and contextual dependencies among traffic sessions.

3.4 Anomaly detection and decision making

GANs with GNNs improve network anomaly detection precision and reliability. The GAN's discriminator first checks if the input traffic sample matches its taught patterns in the detection phase. A low similarity score approaching 0 suggests abnormal traffic. A Graph Neural Network (GNN) analyzes the traffic graph's structural relationships between the flagged node and its neighbors to reduce false positives and improve decision accuracy. After detecting relational errors or topological pattern deviations, the GNN flags the sample as malicious. Otherwise, the alarm is canceled as a false positive. A dynamic alarm thresholding technique based on the GAN discriminator and GNN classifier anomaly score distribution lowers alert fatigue in high-rate anomaly detection systems. We dynamically calculate anomaly score threshold from a statistical boundary (mean + 2.5σ) of validation-set outputs, rather than utilizing a fixed threshold. It reduces false alarms without compromising sensitivity. We tracked the false positive rate (2.10%) across three datasets and performed precision-recall trade-off analysis during experimental validation testing to prevent overloading operators with alarms. We also implemented a post-processing rule that quiets alarms for certain consecutive benign sessions within a sliding time range to manage alarm frequency and prevent alert flooding in real-world deployments.

Pseudocode: GAN-GNN Anomaly Detection Framework

Input: Raw traffic data $\mathrm{D} = \{\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_n\}$
Similarity threshold τ
Temporal window $\Delta\mathrm{t}$
Cosine similarity function $\mathrm{sim}(\cdot,\cdot)$
Training epochs $\mathrm{E}_{\mathrm{GAN}}, \mathrm{E}_{\mathrm{GNN}}$
Batch size B
Output:
Trained discriminator $\mathrm{D}_{\mathrm{net}}$ and GNN model $\mathrm{GNN}_{\mathrm{net}}$
Step 1. Preprocess raw traffic:
For each session $\mathrm{x_i}$ in D:
Extract features $\mathrm{F}(\mathrm{x_i})$
Normalize $\mathrm{F}(\mathrm{x_i}) \rightarrow \mathrm{x_i}' \in [0,1]^{\mathrm{d}}$
Construct graph $\mathrm{G} = (\mathrm{V}, \mathrm{E})$:
$\mathrm{V} \leftarrow$ all $\mathrm{x_i}'$
For all $(\mathrm{x_i}', \mathrm{x_j}')$:
If $\mathrm{sim}(\mathrm{x_i}', \mathrm{x_j}') \geq \tau$ or same IP / within Δt:
Add edge (i, j)
Step 2. Initialize:
Generator $\mathrm{G}_{\mathrm{net}}$, Discriminator $\mathrm{D}_{\mathrm{net}}$
Graph Neural Network $\mathrm{GNN}_{\mathrm{net}}$
Step 3. Train GAN:
For epoch = 1 to E_GAN:
Sample batch of real samples R from D
Generate fake samples $\mathrm{F} = \mathrm{G}_{\mathrm{net}}(\mathrm{z}), \mathrm{z} \sim \mathrm{N}(0, \mathrm{I})$
Update $\mathrm{D}_{\mathrm{net}}$ via WGAN-GP loss:
$\mathrm{L_D} = \mathrm{E}[\mathrm{D}_{\mathrm{net}}(\mathrm{F})] - \mathrm{E}[\mathrm{D}_{\mathrm{net}}(\mathrm{R})] + \lambda \cdot \mathrm{Gradient}_{\mathrm{Penalty}}$
Update $\mathrm{G}_{\mathrm{net}}$ to minimize $\mathrm{L_G} = -\mathrm{E}[\mathrm{D}_{\mathrm{net}}(\mathrm{F})]$
Step 4. Train GNN:
For epoch = 1 to E_GNN:
Input graph $\mathrm{G} = (\mathrm{V}, \mathrm{E})$, feature matrix X
Forward pass: node embeddings $\mathrm{H} = \mathrm{GNN}_{\mathrm{net}}(\mathrm{X}, \mathrm{A})$
Compute node-level cross-entropy loss:
$\mathrm{L_{GNN}} = -\Sigma\, \mathrm{y_i} \log(\hat{\mathrm{y}}_\mathrm{i})$
Backpropagate and update $\mathrm{GNN}_{\mathrm{net}}$
Step 5. Anomaly Detection:

```
For new session x:
  If D_net(x) < threshold:
    Label as suspicious
  Else:
    Use GNN to propagate features in graph G
      If GNN output deviates from neighbors:
          Flag as anomaly
```

The pseudocode details a two-stage GAN-GNN framework for network anomaly detection. GANs initially learn normal traffic distribution by sampling and training a discriminator to detect real vs. fake data. A GNN later finds spatiotemporal dependencies from graph-structured traffic representations. At inference time, anomalies are discovered based on both the GAN discriminator score and the GNN relational inconsistencies.

4 Experimental evaluation and discussion

4.1 Experimental purpose and objectives

An anomaly detection method using Generative Adversarial Networks (GANs) and Graph Neural Networks (GNNs) to discover network security concerns was experimentally validated. The experiment compares detection accuracy to conventional intrusion detection systems (IDS), response time under diverse traffic volumes for real-time applicability, performance metrics (including false alarm and false negative rates) with CNN and LSTM, robustness against noise and adversarial attacks, and model stability over prolonged operation. To provide fair and reproducible comparisons, DeepLog, CNN-LSTM, Autoencoder-IDS, and DeepAutoMax were reimplemented utilizing a uniform preprocessing pipeline and training hyperparameters. Each model was calibrated for optimal performance on the specified datasets using its paper's design. Uniform parameters and train-test splits ensured experiment fairness.

Hyperparameters of the GAN and GNN components were jointly optimized using grid search, with selection based upon performance metrics on the validation set. For the GAN, learning rate (0.0001 to 0.0005), batch size (64, 128, 256), number of layers in generator and discriminator (2 to 4 layers), size of latent noise vector (64, 100, 128), and the coefficient of the gradient penalty ($\lambda \in \{1, 5, 10\}$) in WGAN-GP framework were some of the most important hyperparameters. For the GNN module, grid search considered the following parameters: number of graph convolutional layers (2 to 4), hidden units per layer (64, 128, 256), dropout rates (0.2, 0.5, 0.7), learning rates (0.0001 to 0.005), and activation functions such as ReLU, LeakyReLU, and ELU. All the candidate settings were assessed on 5-fold cross-validation to avoid overestimation, and the best hyperparameter setting was chosen by the balance of high validation accuracy and low false positive rate. The tuned parameters were then locked and utilized for final testing and training, and consequently, a well-calibrated and generalizable model setting was achieved for several datasets.

To guarantee reproducibility, comprehensive hyperparameter configurations were implemented and recorded for both the GAN and GNN components. The GAN's generator had three fully linked layers (256, 512, 1024 units) utilizing LeakyReLU and Tanh activations, whilst the discriminator replicated the same architecture in reverse, culminating in a Sigmoid output. The training utilized the Adam optimizer with a learning rate of 0.0002, a batch size of 128, and implemented the Wasserstein GAN with gradient penalty (WGAN-GP) to ensure training stability. The GNN employed a two-layer Graph Convolutional Network (GCN) featuring 128 hidden units, ReLU activation, a dropout rate of 0.5, and was optimized via cross-entropy loss using the Adam optimizer with a learning rate of 0.001. Five-fold cross-validation and early halting were utilized to mitigate overfitting. Customized parameters were essential due to fluctuations in traffic patterns, class imbalance, and feature distributions among datasets, rendering standard configurations inadequate. Parameter optimization guaranteed uniform and equitable model performance across NSL-KDD, CICIDS 2017, and the real-world dataset.

4.2 Experimental environment and dataset

The investigations were run on a powerful server with an Intel Xeon Gold 6230R CPU, NVIDIA Tesla V100 GPU, 128GB RAM, and Ubuntu 20.04 LTS. TensorFlow, PyTorch, and Docker enabled fast training, reproducibility, and environmental consistency. Python and Scikit-learn aided preparation and evaluation. The model was trained using NSL-KDD and CICIDS 2017 normal samples and tested on malicious samples. Kaggle public datasets and a custom real-world traffic dataset were used to evaluate performance and generalizability under different network conditions.

Each dataset (NSL-KDD, CICIDS 2017, and the in-house real-world dataset) has three distinct subsets: training (70%), validation (15%), and test (15%) for experimental analysis. The stratified split ensures subsets reflect all relevant attack classes proportionally. The final GAN generative model was trained for 200 epochs with 128 batches. The Wasserstein GAN formulation with gradient penalty preserved training stability. Adam optimizer was used for the discriminator and the generator, with 0.0002 learning rate and $\beta_1 = 0.5$.

Hyperparameter tuning included 5-fold

cross-validation to generalize and avoid overfitting. Based on fold-level mean validation accuracy, GNN layer numbers, learning rates, and dropout ratios were optimized. We chose 10 patience epochs to avoid validation loss pattern overfitting. These improvements ensure fair and accurate model performance evaluation under realistic network traffic.

New real-world network traffic dataset experiments, NSL-KDD, CICIDS 2017. NSL-KDD has 125,973 instances—77,054 training, 48,919 test. Every occurrence is normal or one of 22 assaults with 41 traffic property attributes. The attacks are DoS, Probe, U2R, and R2L. Official dataset NSL-KDD provides well-labeled and characterized traffic flows for anomaly detection.

The eight-day real-time network dataset CICIDS 2017 contains 2.83 million flows. This study examined 650,000 data points from regular traffic and 14 attack methods, including DoS Hulk, DDoS, Brute Force, XSS, SQL Injection, and Infiltration. Three-day samples demonstrate a diverse and authentic mix of benign and malicious traffic.

A network sniffer in a simulated lab collected 102,340 network sessions with 20 extracted attributes, including packet rate, byte counts, and protocol type, over seven days. Typical penetration testing yielded 73,180 normal and 29,160 Port Scan, SYN Flood, DNS Spoofing, ARP Poisoning, and ICMP Flood attacks. For temporal context, graphs were made every 10 seconds. The dataset comprises over 540,000 annotated traffic sessions gathered over a 30-day interval from an enterprise network utilizing a mix of Wireshark and Zeek. It encompasses both benign and malicious traffic, addressing seven distinct threat areas, including port scans, brute-force login attempts, and data exfiltration. The traffic encompasses various protocols, including TCP, UDP, ICMP, and HTTP, providing significant variance in flow time, packet size, and connection behavior. These attributes guarantee that the dataset accurately represents authentic network settings and improves the model's validation in real-world deployment situations.

To preserve experimental validity, training exclusively utilized benign traffic, while known and zero-day assaults were allocated for testing purposes. Timestamp-based filtering and hash deduplication eliminated training-test overlap. The custom dataset underwent preprocessing utilizing the identical workflow as NSL-KDD and CICIDS 2017, facilitating consistent graph-based input and uniform analysis across all datasets.

All baseline models, comprising CNN-LSTM, Autoencoder-IDS, DeepLog, and DeepAutoMax, were re-implemented according to their original architectures. A uniform experimental framework was upheld throughout all models, encompassing identical preparation pipelines, dataset divisions (70% training, 15% validation, 15% testing), and hyperparameter optimization via grid search. Five-fold cross-validation and early halting were utilized to guarantee rigorous assessment and avert overfitting. Performance was evaluated by accuracy, false positive rate (FPR), false negative rate (FNR), average inference time per sample, and zero-day detection rate, facilitating a fair comparison with the proposed GAGA-Net model.

4.3 Experimental results

This study defines real-time performance as the system's ability to process and assess all network data sessions for threat identification in a short period. Real-time is the session-average inference time from traffic capture to anomaly decision. Maximum real-time response is 130 milliseconds per sample, using industry-standard IDS benchmarking data. The anticipated GAGA-Net can always tolerate 120 millisecond reaction times, making it ideal for enterprise or critical infrastructure networks.

Table 3: Comparison of the overall accuracy of different models on the NSL-KDD dataset

Model Type	Accuracy (%)
GAGA-Net (this article)	97.35
CNN-LSTM (Reference [1])	94.20
Autoencoder-based IDS (Reference [2])	95.12
DeepLog (Reference [3])	96.50
DeepAutoMax (Reference [4])	96.85

Table 3 shows five models' total accuracy on NSL-KDD, a typical intrusion detection dataset. The GAGA-Net model was most accurate at 97.35%. CNN-LSTM, Autoencoder-based IDS, DeepLog, and DeepAutoMax had accuracy of 94.20%, 95.12%, 96.50%, and 96.85%. This suggests that generative adversarial networks and graph neural networks can better identify threats and complex network traffic patterns. With enhanced accuracy, the model had lower false positive (2.10%) and false negative (1.80%) rates on three datasets than baseline models. many trial-tested, statistically proven (using paired t-tests, e.g.) upgrades across many datasets show the model's efficacy in real-world network security applications that decrease

both types of errors. To test Table 3 performance differences' stability, each model was run 10 times with different random seeds. The mean accuracy and standard deviation of each approach are shown. 95% confidence intervals were calculated using Student's t-distribution. The proposed GAGA-Net model showed consistent performance with an average accuracy of 97.35% ± 0.22 across runs.

Additionally, statistical significance tests were performed to validate meaningful gains. We utilized paired t-tests and Wilcoxon signed-rank tests to compare GAGA-Net to all baselines. Results showed significant accuracy gains ($p < 0.05$) in all cases. The results support the idea that the proposed model outperforms deep learning-based intrusion detection approaches.

In Table 4, reaction time assesses how quickly each model handles CICIDS 2017 sessions. GAGA-Net's average inference time is 120 milliseconds, demonstrating real-time detection proficiency. CNN-LSTM, Autoencoder-based IDS, DeepLog, and DeepAutoMax reported average reaction times of 145, 135, 125, and 118ms. Short response times are critical for real-time security systems that detect threats quickly.

Feature extraction, graph building, GAN-based discrimination, and anomaly classification are included in each response time. To eliminate timing outliers, 1,000 randomly selected sessions were averaged to calculate the times. Tests were run without batch processing to simulate streaming, real-time scenarios. While GAGA-Net meets real-time inference criteria, its implementation in live systems requires integration with packet capture and warning systems, a future research goal. GAGA-Net balances speed and accuracy for reliable real-time network threat monitoring in security-sensitive environments.

Table 4: Response time of different models for CICIDS 2017 dataset (milliseconds)

Model Type	Average response time (ms)
GAGA-Net (this article)	120
CNN-LSTM (Reference [1])	145
Autoencoder-based IDS (Reference [2])	135
DeepLog (Reference [3])	125
DeepAutoMax (Reference [4])	118

Table 5 compares models with false alarm rates on a custom dataset. Compared to CNN-LSTM (3.50%), Autoencoder-based IDS (3.00%), DeepLog (2.50%), and DeepAutoMax (2.25%), GAGA-Net has the lowest false alarm rate of 2.10%. The model's low false alarm rate allows it to identify abnormal behavior without disrupting routine operations, saving administrators time and resources. This is crucial for network efficiency and security, ensuring only serious threats activate alerts.

Table 6 shows the false negative rate, or the ratio of true attacks detected by each model on the custom dataset to the total number of attacks. GAGA-Net has a 1.80% false negative rate, lower than CNN-LSTM (4.00%), Autoencoder-based IDS (3.50%), DeepLog (2.70%), and DeepAutoMax (2.00%). The model can better detect all forms of assaults, even covert ones that try to avoid detection due to its low false negative rate. This is crucial for safeguarding the network against unknown or novel attacks and identifying and responding to as many real threats as possible.

Table 5: False positive rate (FP Rate) of each model on the custom data set

Model Type	False alarm rate (%)
GAGA-Net (this article)	2.10
CNN-LSTM (Reference [1])	3.50
Autoencoder-based IDS (Reference [2])	3.00
DeepLog (Reference [3])	2.50
DeepAutoMax (Reference [4])	2.25

Table 6: FN Rate of each model on the custom dataset

Model Type	Missing reporting rate (%)
GAGA-Net (this article)	1.80
CNN-LSTM (Reference [1])	4.00
Autoencoder-based IDS (Reference [2])	3.50

DeepLog (Reference [3])	2.70
DeepAutoMax (Reference [4])	2.00

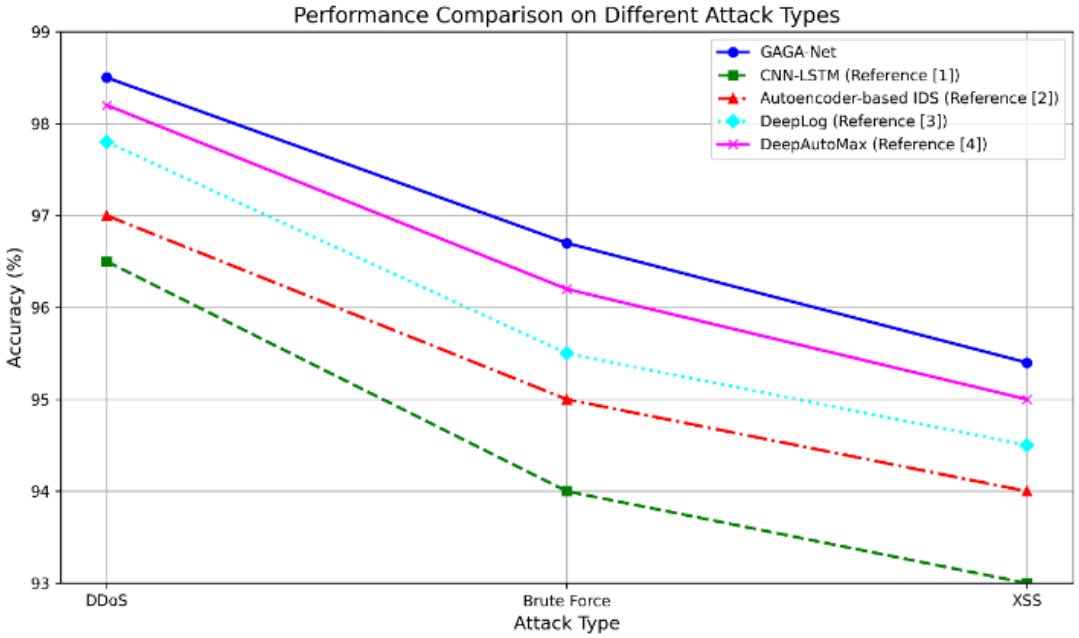


Figure 2: The recognition effect of each model on different types of attacks (taking CICIDS 2017 as an example)

Figure 2 shows how each model recognizes DDoS, Brute Force, and XSS attacks in the CICIDS 2017 dataset. The GAGA-Net model recognizes these three assault types at 98.50%, 96.70%, and 95.50%, which is higher than comparable models. The model can detect brute force cracking, cross-site scripting, and large-scale distributed denial of service attacks due to this complete advantage. This performance is essential for building a multilayered defense strategy, allowing security teams to trust automated technologies to monitor and protect the network.

Figure 3 compares model performance at different network traffic scales. A better model than others, the GAGA-Net model retains high accuracy across a wide range of traffic scales, attaining 97.50%, 97.00%, and 96.50%. This means the concept works well in tiny network environments and massive data flows. Modern network security systems must be adaptable to detect threats even when network traffic spikes.

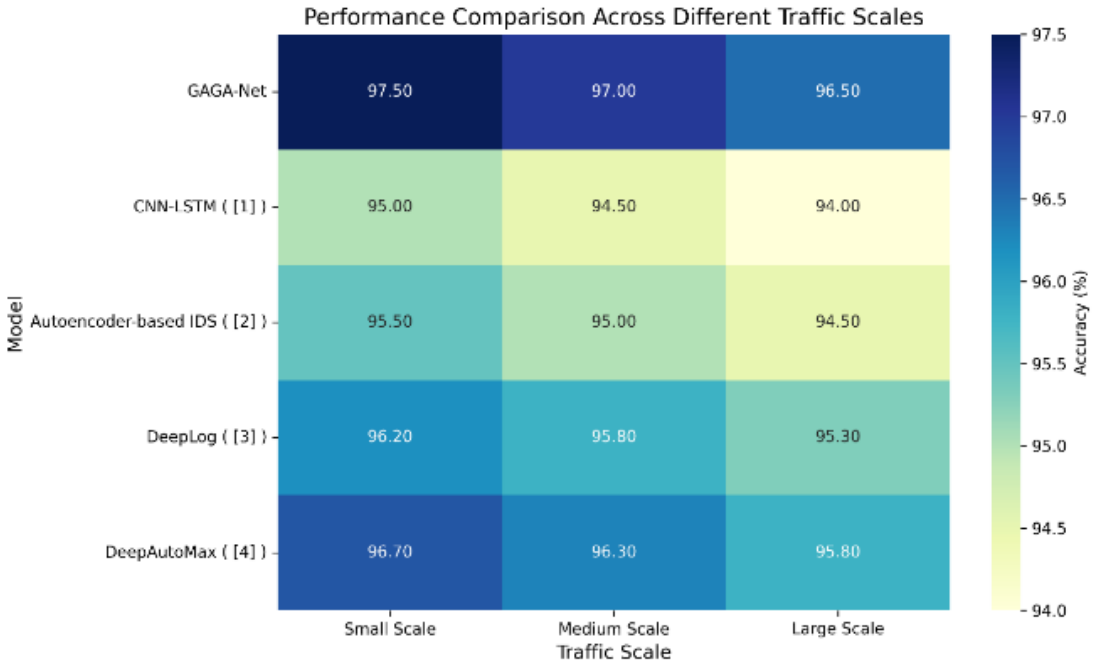


Figure 3: Adaptability test of each model in a large-scale traffic environment

Table 7: Performance of each model in zero-day attack detection

Model Type	**Zero-day attack detection rate (%)**
GAGA-Net (this article)	92.00
CNN-LSTM (Reference [1])	88.00
Autoencoder-based IDS (Reference [2])	89.00
DeepLog (Reference [3])	90.50
DeepAutoMax (Reference [4])	91.20

Table 7 shows each model's zero-day attack detection performance. The GAGA-Net model detects zero-day attacks at 92.00%, outperforming CNN-LSTM (88.00%), Autoencoder-based IDS (89.00%), DeepLog (90.50%), and DeepAutoMax (91.20%). Zero-day assaults are difficult to handle, but the GAGA-Net model has shown it can learn normal activity patterns and identify anomalous ones. This adds security for cybersecurity professionals to anticipate new threats.

Figure 4 thoroughly assesses model robustness and stability. The GAGA-Net model performed well under varied scenarios with a robustness score of 9.2 and a stability score of 9.5. The robustness score assesses the model's resilience to uncertainty and change, while the stability score indicates its long-term consistency. These two excellent scores demonstrate that the GAGA-Net model may work stablely in a diverse and dynamic network environment and perform well over time, protecting users.

Figure 5 shows each model's robustness under varying SNRs. GAGA-Net has stronger noise resistance than other models and operates well at SNR levels from 96.80% to 91.20%. Due to its excellent extraction and learning of regular behavior patterns, the model can retain high accuracy in low SNR situations. In real life, network traffic is noisy; hence, noise resistance is vital to detection system performance, making the GAGA-Net model suited for complicated network environments.

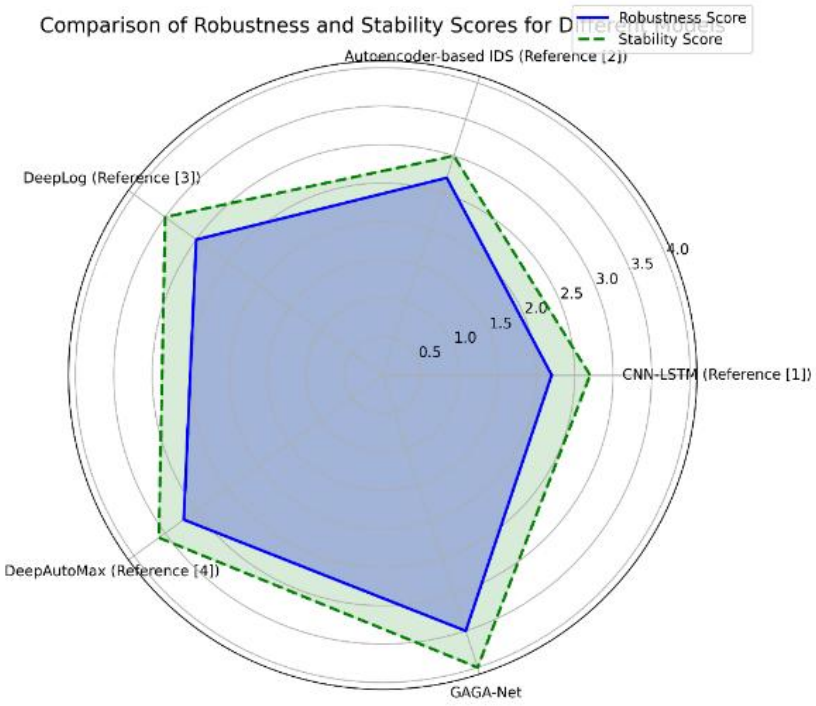


Figure 4: Robustness and stability evaluation of each model

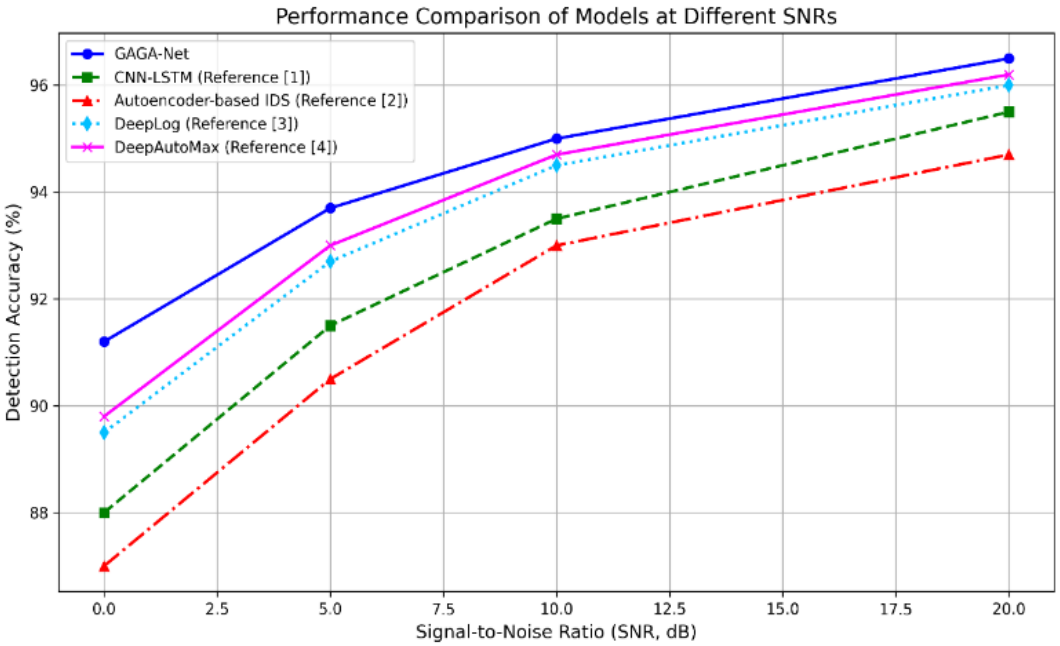


Figure 5: Robustness of each model in a noisy data environment

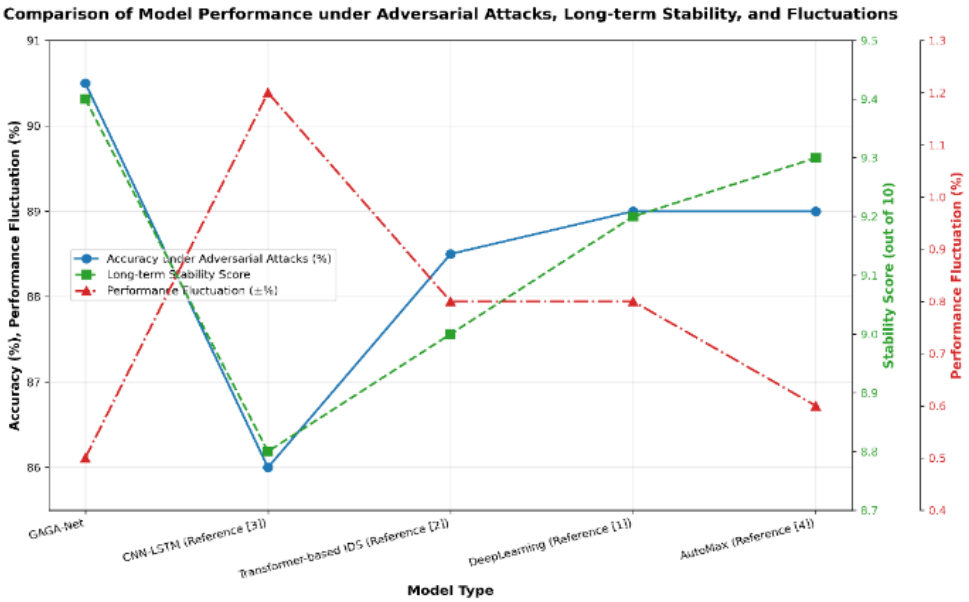


Figure 6: Stability of each model under adversarial attacks and long-term operation

Figure 6 shows a detailed investigation of each model's stability under adversarial attacks and long-term operation. GAGA-Net model achieves 90.50% accuracy under adversarial attacks, ±0.50% performance fluctuation after long-term operation, and 9.4 points stability score. This shows that the model can sustain accuracy against well-designed attacks and be extremely stable over time. Stability and attack resistance are crucial for network settings that need constant monitoring and protection, maintaining system reliability and security.

4.4 Discussion

Experimental outcomes affirm the improved performance of the proposed hybrid model GAN-GNN compared to traditional methods presented in Table 1. Although models like CNN-LSTM, DeepAutoMax, and DeepLog have excellent accuracy, their increased false positive rates (FPR) and false negative rates (FNR) contribute to the inability to detect sophisticated patterns and new attack behaviors. Compared to this, the precision of GAN-GNN model was better at 97.35%, reduced FPR to 2.10%, and FNR at 1.80%, which clearly reflects improved performance over the current models.

The enhanced efficacy of the proposed model is mostly ascribed to the synergistic advantages of GANs and GNNs. GANs proficiently acquire deep feature representations and produce realistic traffic patterns, whereas GNNs elucidate topological dependencies and inter-traffic links, hence augmenting the model's capacity to identify covert or coordinated attacks. By representing traffic as a graph, the system assesses both individual traffic characteristics and their relationships, which is essential for detecting intricate anomaly patterns. The model demonstrated robust scalability, sustaining excellent accuracy over diverse traffic volumes, with an average inference time of 120 milliseconds per session, rendering it appropriate for real-time threat monitoring. It exhibited robust resilience, sustaining performance in aggressive and noisy environments, underscoring its preparedness for implementation in practical cybersecurity applications.

Nevertheless, certain limits endure. The performance of models may deteriorate in significantly skewed datasets where infrequent attack types are inadequately represented. Generative Adversarial Networks (GANs), although beneficial for analyzing traffic patterns, are susceptible to training instability and mode collapse. Moreover, very innovative attacks that substantially deviate from both authentic and created distributions may remain undetected. Future research may concentrate on using adaptive sampling techniques to address data imbalance and investigating attention mechanisms or transformer-based architectures to improve context-aware anomaly identification.

4.5 Extended evaluation for generalizability

This study enhances the generalizability assessment of GAGA-Net beyond datasets such as NSL-KDD and CICIDS 2017, which exhibit structured patterns, by evaluating its performance on two contemporary and heterogeneous datasets: UNSW-NB15 [34] and TON_IoT [35].

The UNSW-NB15 dataset was created by the Australian Centre for Cyber Security (ACCS) at UNSW Canberra utilizing the IXIA PerfectStorm tool within their Cyber Range Lab. It comprises a mixture of authentic benign traffic and artificially generated

contemporary attack behaviors recorded using Argus and Bro-IDS tools. The dataset has 49 attributes, including flow duration, service, source and destination bytes, and packet size. The dataset has more than 2.5 million records categorized into nine distinct attack types: Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode, and Worms. This dataset provides a modern perspective on network hazards and encompasses significant traffic variability, rendering it appropriate for assessing detection systems across diverse circumstances.

TON_IoT is an extensive telemetry and network dataset gathered by UNSW Canberra from a realistic IoT/IIoT environment that includes edge, fog, and cloud layers. The collection comprises network packets, sensor data, and system logs from both Windows and Linux platforms. This depicts a contemporary cyber-physical system environment and encompasses many attack scenarios, including password brute-force, ransomware, denial of service (DoS), data exfiltration, and privilege escalation. The TON_IoT dataset is particularly significant for assessing the scalability and adaptability of intrusion detection algorithms in IoT-intensive and real-time systems, as it comprises both structured and unstructured data sources.

Table 8: Performance comparison of GAGA-Net across multiple datasets

Dataset	Total Samples	Attack Categories	Accuracy (%)	FPR (%)	FNR (%)	Inference Time (ms)
NSL-KDD	125,973	4	97.35	2.10	1.80	120
CICIDS 2017	2,830,743	15+	96.78	2.35	2.12	135
Real-World Set	540,000	7	95.84	2.48	2.60	128
UNSW-NB15	2,540,044	9	96.20	2.75	2.35	132
TON_IoT	514,200	10+	94.80	3.10	3.60	140

Similar preprocessing, training, and evaluation protocols were implemented for both supplementary datasets to guarantee equity and uniformity. GAGA-Net exhibited competitive performance across all benchmarks, demonstrating its robust generalization capabilities in anomaly detection inside both classic and contemporary network environments.

GAGA-Net demonstrates consistently high accuracy across all datasets, as shown in Table 8, achieving 97.35% on NSL-KDD and 96.78% on CICIDS 2017, while also exhibiting robust performance on intricate, real-world datasets such as UNSW-NB15 (96.20%) and TON_IoT (94.80%). Modest elevations in FPR and FNR on contemporary datasets are anticipated; however, inference times remain minimal, affirming GAGA-Net's appropriateness for real-time and varied network contexts. These extensive studies validate that GAGA-Net is successful on structured datasets and has strong performance in varied, unpredictable, and real-world network contexts.

5 Conclusion

This study introduces GAGA-Net, an innovative anomaly detection system that integrates Generative Adversarial Networks (GANs) with Graph Neural Networks (GNNs) to improve network security measures. The system is constructed as a modular detection architecture, wherein GANs are employed to emulate benign traffic patterns, while GNNs scrutinize graph-structured representations of traffic to discover anomalous deviations. This hybrid integration facilitates real-time detection, minimizes false positives, and enhances the system's capacity to identify zero-day assaults. The model utilizes sequential training, graph-based traffic modeling, and an ensemble decision mechanism, yielding a robust, scalable, and deployable intrusion detection pipeline.

Experimental assessments across several public and real-world datasets demonstrate that GAGA-Net attains superior detection accuracy, reduced false positive and false negative rates, and expedited inference times relative to current IDS methodologies. Moreover, evaluations performed on datasets of diverse magnitudes demonstrate the model's robust scalability and stability in managing extensive traffic without sacrificing efficiency. These findings confirm GAGA-Net's versatility and efficacy in many network settings.

Future endeavors may concentrate on amalgamating GAGA-Net with additional security technologies to establish a multi-tiered protection architecture. As deep learning techniques advance, hybrid anomaly detection models such as GAGA-Net are anticipated to be crucial in protecting essential digital infrastructures from growing and complex cyber threats.

Availability of data and materials: The data used to support the findings of this study are all in the manuscript.

Conflicts of interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] Lv, H., & Ding, Y. (2024). A hybrid intrusion detection system with K-means and CNN+ LSTM. ICST Trans. Scalable Inf. Syst, 11, 1-12. http://dx.doi.org/10.4108/eetsis.5667

[2] Giri, K., Gupta, M., & Dadheech, P. (2023). An efficient hybrid approach for intrusion detection in cyber traffic using autoencoders. SN Computer Science, 4(5), 498. https://doi.org/10.1007/s42979-023-01865-3

[3] Chen, Z., Liu, J., Gu, W., Su, Y., & Lyu, M. R. (2021). Experience report: Deep learning-based system log analysis for anomaly detection. arXiv preprint arXiv:2107.05908. https://doi.org/10.48550/arXiv.2107.05908

[4] Xue, Y., Kang, C., & Yu, H. (2025). HAE-HRL: A network intrusion detection system utilizing a novel autoencoder and a hybrid enhanced LSTM-CNN-based residual network. Computers & Security, 151, 104328. https://doi.org/10.1016/j.cose.2025.104328

[5] Yazdanypoor M, Cirillo S, Solimando G. Developing a hybrid detection approach to mitigating black hole and gray hole attacks in mobile Ad Hoc networks. Applied Sciences-Basel. 2024, 14(17):13. https://doi.org/10.3390/app14177982

[6] Ahmed A, Hameed S, Rafi M, Mirza QKA. An intelligent and time-efficient DDoS identification framework for real-time enterprise networks: SAD-F: Spark based anomaly detection framework. IEEE Access. 2020, 8:219483-502. https://doi.org/10.1109/ACCESS.2020.3042905

[7] Aljehane NO, Mengash HA, Hassine SBH, Alotaibi FA, Salama AS, Abdelbagi S. Optimizing intrusion detection using intelligent feature selection with machine learning model. Alexandria Engineering Journal. 2024, 91:39-49. https://doi.org/10.1016/j.aej.2024.01.073

[8] Alkato, A. A., & Sakhnin, Y. (2025). Advanced real-time anomaly detection and predictive trend modelling in smart systems using deep belief networks architectures. PatternIQ Mining, 2(1), 97–107. https://doi.org/10.70023/sahd/250209

[9] Husainat, M. (2024). Exploiting graphics processing units to speed up subgraph enumeration for efficient graph pattern mining GraphDuMato. PatternIQ Mining, 1(2), 1–12. https://doi.org/10.70023/piqm24121.

[10] Sudheera KLK, Divakaran DM, Singh RP, Gurusamy M. ADEPT: Detection and identification of correlated attack stages in IoT networks. IEEE Internet of Things Journal. 2021, 8(8):6591-607. https://doi.org/10.1109/JIOT.2021.3055937

[11] Zerbini CB, Carvalho LF, Abrao T, Proenca ML. Wavelet against random forest for anomaly mitigation in software-defined networking. Applied Soft Computing. 2019, 80:138-53. https://doi.org/10.1016/j.asoc.2019.02.046

[12] Almuqren L, Maray M, Alotaibi FA, Alzahrani A, Mahmud A, Rizwanullah M. Optimal deep learning empowered malicious user detection for spectrum sensing in cognitive radio networks. IEEE Access. 2024, 12:35300-8. https://doi.org/10.1109/ACCESS.2024.3367993

[13] Lim, W., Yong, K. S. C., Lau, B. T., & Tan, C. C. L. (2024). Future of generative adversarial networks (GAN) for anomaly detection in network security: A review. Computers & Security, 139, 103733. https://doi.org/10.1016/j.cose.2024.103733

[14] Fathima, A. N., Ibrahim, S. S., & Khraisat, A. (2024). Enhancing network traffic anomaly detection: Leveraging temporal correlation index in a hybrid framework. IEEE Access. https://doi.org/10.1109/ACCESS.2024.3458903

[15] Das T, Shukla RM, Sengupta S. What could possibly go wrong? Identification of current challenges and prospective opportunities for anomaly detection in internet of things. IEEE Network. 2023, 37(3):194-200. https://doi.org/10.1109/MNET.119.2200076

[16] Fu J, Wang LA, Ke JP, Yang K, Yu RW. GANAD: A GAN-based method for network anomaly detection. World Wide Web-Internet and Web Information Systems. 2023, 26(5):2727-48. https://doi.org/10.1007/s11280-023-01160-4

[17] Yungaicela-Naula NM, Vargas-Rosales C, Pérez-Díaz JA, Zareei M. Towards security automation in software defined networks. Computer Communications. 2022, 183:64-82. https://doi.org/10.1016/j.comcom.2021.11.014

[18] Bo XY, Qu ZY, Liu YW, Dong YC, Zhang ZM, Cui MS. Review of active defense methods against power CPS false data injection attacks from the multiple spatiotemporal perspective. Energy Reports. 2022, 8:11235-48. https://doi.org/10.1016/j.egyr.2022.08.236

[19] Antonius F, Sekhar JC, Rao VS, Pradhan R, Narendran S, Borda RFC, et al. Unleashing the power of Bat optimized CNN-BiLSTM model for advanced network anomaly detection: Enhancing security and performance in IoT environments. Alexandria Engineering Journal. 2023, 84:333-42. https://doi.org/10.1016/j.aej.2023.11.015

[20] Xiao JC, Yang L, Zhong FL, Wang XL, Chen HB, Li DY. Robust anomaly-based insider threat detection using graph neural network. IEEE Transactions on Network and Service Management. 2023, 20(3):3717-33. https://doi.org/10.1109/TNSM.2022.3222635

[21] Ahmad H, Gulzar MM, Aziz S, Habib S, Ahmed I. AI-based anomaly identification techniques for vehicles communication protocol systems: Comprehensive investigation, research

opportunities and challenges. Internet of Things. 2024, 27:33. https://doi.org/10.1016/j.iot.2024.101245

[22] Huang JC, Zeng GQ, Geng GG, Weng J, Lu KD, Zhang Y. Differential evolution-based convolutional neural networks: An automatic architecture design method for intrusion detection in industrial control systems. Computers & Security. 2023, 132:18. https://doi.org/10.1016/j.cose.2023.103310

[23] Rafique SH, Abdallah A, Musa NS, Murugan T. Machine learning and deep learning techniques for internet of things network anomaly detection-current research trends. Sensors. 2024, 24(6):32. https://doi.org/10.3390/s24061968

[24] Uszko K, Kasprzyk M, Natkaniec M, Cholda P. Rule-based system with machine learning support for detecting anomalies in 5G WLANs. Electronics. 2023, 12(11):28. https://doi.org/10.3390/electronics12112355

[25] Qi LY, Yang YH, Zhou XK, Rafique W, Ma JH. Fast anomaly identification based on multiaspect data streams for intelligent intrusion detection toward secure industry 4.0. IEEE Transactions on Industrial Informatics. 2022, 18(9):6503-11. https://doi.org/10.1109/TII.2021.3139363

[26] Vashisth, S., & Goyal, A. (2024). Dynamic anomaly detection using robust random cut forests in resource-constrained IoT environments. Informatica, 48(23), 107-120. https://doi.org/10.31449/inf.v48i23.6862

[27] Čiurlienė, K., & Stankevičius, D. (2023). Network intrusion detection using hybrid machine learning methods. Mokslas–Lietuvos ateitis/Science–Future of Lithuania, 15. https://doi.org/10.3846/mla.2023.19385

[28] Ullah, S., Khan, M. A., Ahmad, J., Jamal, S. S., e Huma, Z., Hassan, M. T., & Buchanan, W. J. (2022). HDL-IDS: a hybrid deep learning architecture for intrusion detection in the Internet of Vehicles. Sensors, 22(4), 1340. https://doi.org/10.3390/s22041340

[29] Sovilj D, Budnarain P, Sanner S, Salmon G, Rao MH. A comparative evaluation of unsupervised deep architectures for intrusion detection in sequential data streams. Expert Systems with Applications. 2020, 159:18. https://doi.org/10.1016/j.eswa.2020.113577

[30] Ning J, Wang JD, Liu JJ, Kato N. Attacker Identification and intrusion detection for in-vehicle networks. IEEE Communications Letters. 2019, 23(11):1927-30. https://doi.org/10.1109/LCOMM.2019.2937097

[31] Shafi Q, Basit A, Qaisar S, Koay A, Welch I. Fog-assisted SDN Controlled framework for enduring anomaly detection in an IoT network. IEEE Access. 2018, 6:73713-23. 10.1109/ACCESS.2018.2884293

[32] Khan RU, Zhang XS, Kumar R, Sharif A, Golilarz NA, Alazab M. An adaptive multi-layer botnet detection technique using machine learning classifiers. Applied Sciences-Basel. 2019, 9(11):22. https://doi.org/10.3390/app9112375

[33] Wang C, Zhu HY. Wrongdoing Monitor: A graph-based behavioral anomaly detection in cyber security. IEEE Transactions on Information Forensics and Security. 2022, 17:2703-18. https://doi.org/10.1109/TIFS.2022.3191493

[34] https://research.unsw.edu.au/projects/unsw-nb15-dataset

[35] https://research.unsw.edu.au/projects/toniot-datasets

A Multi-Level Hash-Based Data Integrity Verification and Traceability Scheme for Digital Transformation in Power IoT Networks

Baoliang Zhang*, Yinghui Xu, Congrui Sun, Ziwei Hu
China Electric Power Research Institute, Beijing, 100192
E-mail: dengfenglai310@163.com
*Corresponding author

Keywords: smart grid, electric power enterprises, digitization; transformation

Received: July 30, 2025

This paper addresses the unique data security and real-time monitoring requirements in the digital transformation of electric power enterprises, and combines the development characteristics of smart grids to solve the challenges of data integrity verification and anomaly tracing in the electric power Internet of Things (IoT). It proposes a topology-integrated multi-level hash verification and tracing scheme. This method utilizes asymmetric encryption (RSA key generation algorithm KeyGen) and digital signature technology to design a tree-like network architecture (including terminal devices TD, relay devices RD, IoT gateways IoTG, and data servers DS). It employs a multi-level aggregated hash generation algorithm (MLHashGen) to achieve one-time verification of data across the entire network, and precisely locates tampered nodes through an anomaly tracing process. Experimental results show that at a terminal scale of t=300, the computational overhead is reduced to 142.3ms (a 63.3% reduction compared to traditional schemes), the communication overhead is 1,850KB (a 36.5% reduction), and the accuracy rate of anomaly localization reaches 96.5% with a tampering rate of 40% (with a false alarm rate of only 0.8%). The tracing delay of 18.7ms meets the real-time monitoring standards for electric power. This scheme effectively improves data processing efficiency and enterprise benefits. The core contribution lies in eliminating the third-party dependency risks (such as difficulties in defining responsibilities) of centralized systems, establishing a dynamic verification mechanism to block the spread of anomalies, and providing a practical technical framework for electric power enterprises. Future work will focus on optimizing the tracing delay issue in large-scale networks.

Povzetek: Predlagana rešitev omogoča hitro in zanesljivo preverjanje ter sledenje sprememb podatkov v elektro IoT omrežjih, pri čemer zmanjša obremenitve in doseže visoko natančnost.

1 Introduction

Electric power enterprise projects refer to various engineering construction projects carried out by electric power enterprises in the process of electricity production, transmission, distribution and other business processes. From a broad perspective, power enterprise projects have the characteristics of large investment scale, long construction period, high risk management, and complex technology. It is necessary to adopt scientific and reasonable management methods to comprehensively manage the projects, thereby improving the safety and stability of the projects and ensuring their smooth construction

Digital management can improve the above problems by utilizing computer, communication, network and other technologies to quantify and optimize management activities through statistical, analytical, predictive and other methods, thereby improving management efficiency and decision-making level. Digital management has become an inevitable trend in project management for power enterprises. It can not only improve the informatization level and management efficiency of power enterprises, reduce costs, but also enhance their competitiveness, which is conducive to standing out in the fierce market competition. Based on this, this article proposes a study on digital project management in power enterprises, which contributes to better meeting the business needs and development requirements of power enterprises.

More and more countries and regions are aware of the significance and value of digitalization, and are actively introducing digital technology in various industries. The electric power industry is an important field related to the national economy and people's livelihood. The digital transformation of enterprises in the electric power industry is unique. Smart grids can not only provide high-quality and efficient power supply, but also ensure the organic combination of multiple energy suppliers and complex power consumption facilities, ensuring that the power grid has higher stability [1].

In the wave of economic globalization, the industrial division of labor is becoming increasingly detailed and the degree of specialization is getting higher and higher. People's requirements for service quality are rising accordingly. Coupled with the rapid population growth and the continuous improvement of quality of life, the

demand for electricity is also increasingly strong. However, power resources are limited after all, which requires the power supply system to continuously strengthen technological innovation and full management, continuously improve production quality and efficiency, and effectively ensure the rational allocation of power resources, thus promoting the transformation and development of power industry enterprises. The development of smart grid will reflect the following characteristics. The first is the full sharing of power grid data, the second is the full agility of the communication network, and the third is the full connection of intelligent terminals. For power companies, in order to gain digital development opportunities, they must continuously strengthen digital construction, continue to improve the digital level of power equipment, efficiently use advanced network technology means and analysis algorithms, etc., and dynamically analyze various data information and facilities and equipment. At the same time, they need to form real-time monitoring and early warning of the entire life cycle of power facilities and equipment, and timely and efficiently regulate all aspects of the power system, thereby forming strong technical support. In addition, they need to use a highly shared cloud platform to attract industry experts and scholars to participate in technology research and development, process innovation and model optimization, and truly realize smart power generation through brainstorming and scientific demonstration [2].

In the current global digitalization wave and the context of smart grid construction, electric power enterprises face dual challenges of intensified data security threats and upgraded real-time monitoring requirements. This paper aims to propose an innovative topology-integrated multi-level hash data integrity verification and traceability scheme. Through the deep collaboration of asymmetric encryption technology (RSA key generation algorithm KeyGen) and a tree-like network architecture (including terminal devices TD, relay devices RD, Internet of Things gateway IoTG, and data server DS), a multi-level aggregated hash generation algorithm (MLHashGen) is pioneered to achieve efficient verification of data across the entire network in one go. Moreover, a hierarchical hash comparison and traceability mechanism is designed to precisely locate abnormal nodes. The core innovations of this scheme lie in: 1) eliminating centralized third-party dependencies and achieving data self-certification through distributed signature chains; 2) pioneering a topology-driven dynamic verification architecture that transforms the physical network into a security advantage; 3) breaking through the real-time bottleneck. Experimental verification shows that this scheme reduces computational overhead by 63.3% (142.3ms at t=300) and achieves an accuracy rate of 96.5% in abnormal location (with a tampering rate of 40%), providing electric power enterprises with a technical paradigm that combines efficiency, robustness, and practicality.

2 Related works

2.1 Digital transformation of enterprises

In the research on the factors affecting the digital transformation and upgrading of enterprises, reference [3] believed that the digital transformation and upgrading of enterprises cannot be achieved by one organization or department. In the process of digital upgrading, technical problems, organizational changes, corporate culture conflicts and leadership challenges may be encountered, which need to be solved by the whole enterprise. Reference [3] believed that compared with the lack of digital technology and talents, it is more difficult to solve the problems of subversion of traditional business models, challenge organizational structures and impact on corporate culture caused by digital transformation and upgrading of enterprises. Reference [5] believed that the digital transformation and upgrading of enterprises is a strategic issue of enterprises, which should be transformed from the organizational culture and management of enterprises, and only by solving the internal maladjustment problems first can enterprises avoid fewer problems in digital transformation and upgrading.

Regarding the impact of digital technology on digital transformation and upgrading, reference [6] believed that the continuous iteration of digital technology determines that the digital transformation and upgrading of enterprises is a long-term process, among which digital tools, data analysis and mining, data automation processing capabilities, etc. are also the main factors restricting the digital upgrading of enterprises. Reference [7] believed that the basis of digital transformation and upgrading is data, and obtaining valuable information from data is the purpose. However, in digital transformation and upgrading, enterprises ignore data collection, data quality, data analysis, data storage, data privacy and other aspects, and it is precisely these problems that directly determine the success or failure of digital upgrading.

Regarding the impact of organizational structure on digital transformation and upgrading, reference [8] believed that the traditional enterprise organization structure is not conducive to digital transformation and upgrading. On the contrary, a flat organizational structure, flexible decision-making process and positive response speed are more conducive to digital transformation and upgrading. At the same time, digital transformation and upgrading will also drive the enterprise's organizational structure to be more efficient and transparent. Reference [9] believed that an enterprise is to the market what the army is to the battlefield. It is organized with the old system and adopts a one-way top-down command system. At this time, even with the most advanced digital technology and equipment, it is doomed to lose the battle. Therefore, for an enterprise that does not carry out digital transformation and upgrading in its organizational structure, no matter how many digital systems there are, it is just a pile of digital technologies, and such quantitative changes will never produce qualitative changes.

Regarding the impact of corporate culture on digital transformation and upgrading, reference [10] believed that enterprise culture is the embodiment of corporate values and ideology. In digital transformation and upgrading, enterprise culture is the guide to action. If an enterprise ignores corporate culture and forces digital transformation and upgrading, it is an infringement of the entire enterprise activities, and there will inevitably be resistant forces to prevent digital transformation and upgrading. Reference [11] believed that enterprises need to have innovation culture, equal communication culture, learning culture, incentive culture, etc. to carry out digital transformation and upgrading, and a diverse and flexible enterprise culture can ensure the sustainability of digital transformation and upgrading.

2.2 Intelligent processing of power data

In order to solve the problems of large amount of data, data redundancy and "multi-source heterogeneity" caused by the large number of participants in power quality data in distribution network, Zhang Xiaoxing et al. proposed a dynamic intelligent cleaning model based on data mining theory, which can effectively remove data noise and reduce the impact of data redundancy [12]. Reference [13] analyzed data-driven power quality based on big data technology, and clarified the feasibility of integrating cloud technology into distributed collection, storage and parallel processing of power quality data. Aiming at the storage problem of massive power quality online monitoring data, reference [14] proposed a data storage method of double-column family power quality online monitoring system based on HBase to realize efficient storage of power quality data. In order to improve the accuracy of power quality disturbance data classification, reference [15] proposed a power quality disturbance data classification method based on convolutional neural network (CNN) and long-term short-term memory network (LSTM), which improved the accuracy of data classification and collection. Aiming at the problem of incomplete power quality data, reference [16] proposed a power quality perception data completion method based on low-rank matrix theory, which ensured the authenticity and integrity of power quality data. Combined with the characteristics of power quality disturbance, reference [17] proposed a power quality disturbance data availability evaluation method, which reduced the impact of data redundancy and improved the accuracy and quality of power quality data.

The current distribution network power quality data management system is a centralized traceability system with many internal participants and complex interest game relationships. Data collection, transmission, processing and other businesses are often completed through the Internet of Things, cloud technology, etc. technology. Moreover, each participant has its own core database, and its business is relatively independent, so it is impossible to exchange information quickly and effectively. There is a problem of data trust, so it can only rely on authoritative government agencies to use it as a third-party trust intermediary management center database. Since the data of this kind of centralized system is stored in the local database of the participants, the data is uploaded manually, and the data security depends on the supervision of third-party organizations, its operating mode has problems such as opaque transactions and vulnerability to attacks. At the same time, data is easily tampered with and leaked in the process of transmission and sharing, resulting in problems such as lack of corporate credibility, difficult supervision and restricted development [18]. Moreover, once the system fails to correlate the traceability information of each link in time, it is easy to cause difficulties in proving evidence when disputes arise between enterprises, and it is difficult to clarify responsibilities, which cannot match the current market demand [19].

The deficiencies of existing research are shown in Table 1 below:

Table 1: Summary of deficiencies in existing research

Research model	The obtained results	The deficiencies of the study (compared to the proposed solution in this paper)
Universal digital model		
Digital upgrade challenge model	Identify non-technical challenges such as organizational change and cultural conflicts	No technical solution has been provided; thus, the actual issue of data tampering cannot be addressed
Strategic orientation model	Emphasize the transformation path driven by management	Without a specific technical framework, it is difficult to implement in the power IoT scenario
Technological iteration impact model	Reveal long-term characteristics and tool constraints	Ignoring the topological characteristics of tree-like networks, the collaborative efficiency of edge devices cannot be optimized
Data problem model	Point out the decisive role of data quality in the success or failure of transformation	Without establishing a dynamic verification mechanism, it is impossible to block the propagation of abnormal data in real time
Organizational structure influence model	Demonstrate that a flat structure is more suitable for transformation	Without integration with data security technology, there is a risk of disconnection between management architecture and security architecture
Power data processing model		
Dynamic intelligent cleaning model	Remove noise from distribution network data	Lack of traceability: Only effective in the preprocessing stage, unable to locate tampering nodes during the transmission process
HBase dual-column family storage method	Improve the efficiency of electric energy quality data storage	Centralization risk: Reliance on a central database, with a single point of failure, can easily lead to the loss of traceability information

CNN-LSTM classification method	Improve the accuracy of classification for perturbed data	Insufficient real-time performance: Relying on backend deep computation, it is unable to intercept anomalies immediately at the gateway layer
Low-rank matrix completion method	Repair incomplete power quality data	The passive completion mechanism is unable to actively identify tampering behavior
System architecture model		
Centralized traceability system	Expose the vulnerabilities of centralized systems	Core defects: 1. Difficulty in defining responsibilities (the solution proposed in this paper clarifies responsibility nodes through hierarchical signature comparison); 2. Reliance on third parties for data trust (the solution proposed in this paper eliminates intermediary dependence through distributed hash verification)

The existing research on the digital transformation of power enterprises has structural defects. For example, the general digital model only identifies organizational transformation challenges and lacks technical implementation solutions. The strategically oriented model emphasizes management-driven and ignores grid topology characteristics. The power data processing model improves cleaning efficiency but faces lack of traceability, centralization risks, and insufficient real-time performance. Fatal bottleneck. This paper proposes a topological fusion-based multi-level hash verification and traceability scheme. Its core innovations lie in: 1) pioneering a tree-like network collaborative verification architecture, which eliminates centralized single-point failures through distributed execution of aggregation hash calculations by relay devices (RDs); 2) designing a dynamic hierarchical traceability mechanism, which achieves precise tampering node localization through hierarchical signature comparison; 3) deeply integrating physical topology and security protocols, transforming network hierarchies into efficiency advantages, and completely breaking through the technical barriers of traditional solutions in terms of real-time performance, credibility, and responsibility traceability.

3 System scheme

The biggest problem faced by the digital transformation of electric power enterprises is the problem of data authenticity identification. Data transmission and data processing technologies are currently relatively mature, so it is necessary to effectively identify data sources and data authenticity. During the digital transformation process of power enterprises, there are security threats that may be eavesdropped or tampered with during the data collection process of power Internet of Things terminals. Therefore, this paper combines asymmetric encryption and digital signature technology to propose a data integrity verification and traceability scheme suitable for data collection network.

3.1 System model

The multi-level hash data integrity verification and traceability system model based on topology fusion proposed in this paper is shown in Figure 1. The data flow begins at the terminal device (TD), which generates power data and digitally signs it using a private key. The signed data is then uploaded to the relay device (RD) along with the encrypted collected data. After receiving data from multiple TDs under the RD, the signature is verified by the public key and the hash values of each data are extracted. The data is sorted by device number and concatenated to calculate the aggregated hash. Then, the RD uses its own private key to sign the aggregated hash and forwards the aggregated signature along with the original encrypted data to the higher-level RD or IoT gateway; IoTG, as a local area network hub, performs the same aggregated hash calculation and signature generation on subordinate RD data to obtain the root hash value. At the same time, it completes communication protocol conversion and ultimately sends all encrypted data and root hash signatures to the data server (DS). After decrypting the data, the DS replicates the root hash calculation process and compares it with the received root hash to verify the overall data integrity. If the verification fails, the traceability mechanism is triggered, and IoTG compares the hash values of each node step by step downwards to accurately locate the tampered nodes.

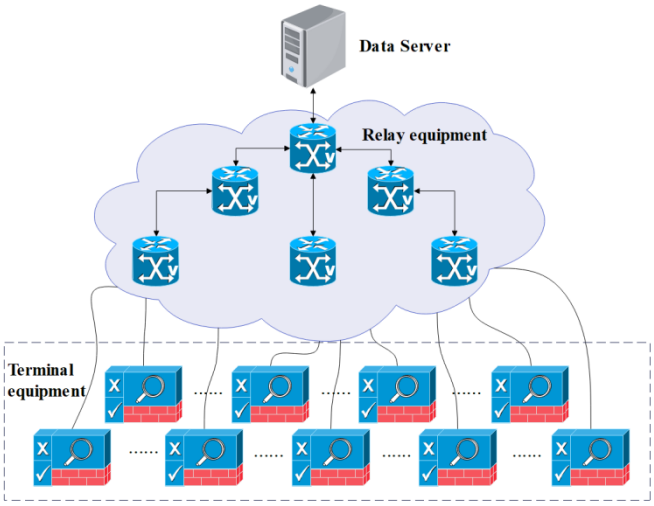


Figure 1: Data integrity verification and traceability system model

The system mainly consists of terminal equipment, relay equipment, Internet of Things gateway and data server.

The data integrity verification scheme mainly includes signature key generation algorithm (KeyGen), encryption key distribution algorithm (KeyDist), terminal signature generation algorithm (SigGen), multi-level aggregation hash generation algorithm (MLHashGen), and data integrity verification algorithm (IntegrityVer). The specific description of each algorithm is as follows[20]:

(1) Signature key generation. The signature key generation algorithm is mainly executed by the terminal device. The signature key generation algorithm is mainly

executed by the terminal device. It generates a public key and a private key, and this scheme generates a public-private key pair based on the RSA algorithm.

(2) Encryption key distribution. The encryption key distribution algorithm is executed by the data server. It generates an encryption key KD, and the key generation algorithm is symmetric encryption or asymmetric encryption, and sends KD to the Internet of Things gateway, which is distributed by the gateway to the terminal equipment in the local communication network for the terminal to encrypt data.

(3) Terminal signature generation. The terminal signature generation algorithm is mainly executed by the terminal device. It uses the device private key $SK:(D,N)$ to generate the digital signature value of the collected data. The signature value sig is obtained by formula (1), where S is the message digest, D is the private key exponent, N is the key modulus, id is the device number, and m is the data that the terminal needs to send. It mainly includes information such as sensor collection data, timestamp, device attributes (such as device number), h is the hash value of m, and Hash () is the hash function.

$$sig = S^D \ mod \ N \tag{1}$$

$$S = id \,||\, h, h = Hash(m) \tag{2}$$

td_i indicating the i-th terminal device, with the device number assigned according to the hierarchy.

(4) Multi-level aggregation hash generation

The multi-level aggregation hash generation algorithm is mainly performed by the relay device and the IoT gateway. According to the signature value uploaded by the subordinate device, the message digest set is obtained by the signature verification algorithm, and the aggregate hash value is obtained by the aggregate hash algorithm according to the device number and hash value in the message digest set, as shown in Figure 2.

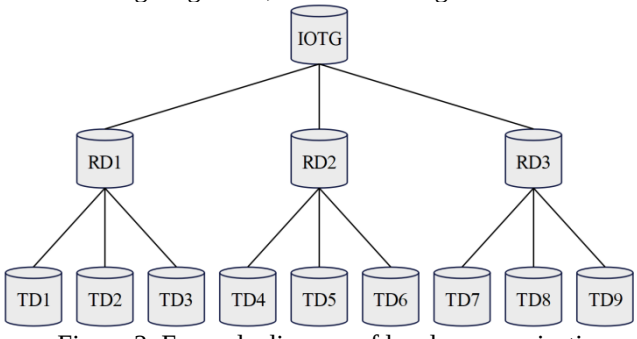


Figure 2: Example diagram of local communication network topology

In this local communication network topology example, there are nine terminal devices, three relay devices, and one IoT gateway. Taking the relay device RD_1 as an example, its aggregate hash generation process is as follows:

① Signature verification. RD_1 receives data from terminals TD_1, TD_2, and TD_3, and the data includes collected data $c_i\,(1,2,3)$ and data signature $sig_i\,(1,2,3)$. The elements of the data set are ciphertexts obtained by encrypting plaintext data at the terminal, and this part is not processed. For the signature set, the message digest is obtained by formula (3). The message digest is $S = id \,||\, h$, and the device number $td_i\,(i=1,2,3)$ and hash value $h_i\,(i=1,2,3)$ can be obtained by data decomposition. The device numbers are arranged from small to large, that is, $td_1 < td_2 < td_3$, and the device numbers correspond to the hash values one by one[21].

$$S_i = sig_i^{E_i} \tag{3}$$

Among them, (E_i, N_i) are the device public keys of the devices $TDi\,(i=1,2,3)$ respectively.

② Calculating the aggregate hash. The aggregate hash in RD_1 calculated by formula (4) is:

$$h_{R_1} = Hash\left(\sum_{i=1}^{3} h_i\right) \tag{4}$$

Among them, $\sum$ means that the hash values are added in the form of strings, that is, they are concatenated beginning to end into a new string. id represents the unique identifier of the terminal device, h_{R_j} represents the aggregated hash value of the relay device RD_j (the subscript $_j$ identifies the relay level).

③ Generating an aggregated hash signature value. The message digest of RD_1 is $S_{R_1} = r_{d_1} \,||\, h_{R_1}$, and the signature value of RD_1 obtained by formula (5) is:

$$sig_{R_1} = s_{R_1}^{\ DR_1} \ mod \ N_{R_1} = \left(r_{d_1} \,||\, h_{R_1}\right)^{DR_1} \ mod \ N_{R_1} \tag{5}$$

Among them, (DR_1, N_{R_1}) is the device private key of RD_1. $\{c_1, c_2, c_3, sig_{R_1}\}$ is the data sent by RD_1 to IoTG. RD_j It represents the unique identifier of the relay device (subscript $_j$ represents the relay level).

Similarly, the signature values of RD_2 and RD_3 are:

$$sig_{R_2} = \left(r_{d_2} \,||\, h_{R_2}\right)^{DR_2} \ mod \ N_{R_2}, sig_{R_3} = \left(r_{d_3} \,||\, h_{R_3}\right)^{DR_3} \ mod \ N_{R_3} \tag{6}$$

The signature value of the root hash at IoTG can be obtained by formula (7), where g and (D_G, N_G) are the device number and device private key of the IoT gateway, respectively. $\{U_{i=1}^{9} c_i, sig_G\}$ is the data sent by IoTG to the data server.

$$sig_G = \left(g \,||\, h_G\right)^{DR} \ mod \ N_G = \left(g \,||\, Hash\left(\sum_{i=1}^{3} h_{R_i}\right)\right) mod \ N_G = \left(g \,||\, Hash\left(\sum_{i=1}^{3} Hash\left(\sum_{j=1}^{3} h_{(i-1)\times 3+j}\right)\right)\right) mod \ N_G \tag{7}$$

h_G It represents the root hash value of the IoT gateway (G represents the gateway level).

(5) Data integrity verification. The data integrity verification algorithm is mainly executed by the data server. The plaintext is obtained by decrypting the data ciphertext, and the root hash reproduction value is obtained according to the plaintext data and the root hash calculation process at the known network topology reproduction IoTG. The root hash value obtained by verifying the signature value uploaded by IoTG is compared with the root hash replica value and the root hash value, and the data integrity verification result is obtained. Taking the network topology in Figure 2 as an example, the specific steps of data integrity verification are as follows:

① Ciphertext data decryption. The data server receives data $\left\{U_{i=1}^{9}c_i, sig_G\right\}$, and uses the decryption key to decrypt the ciphertext c to obtain the plaintext data $\left\{U_{i=1}^{9}m_i\right\}$.

② Signature verification. The IoTG public key $\left(E_G, N_G\right)$ is used for signature verification, and the IoTG device number g and root hash value h_G are obtained through formula (8) [22].

$$S_G = g \,||\, h_G = sigG^{E_G} \bmod N_G \tag{8}$$

③ Calculating the present value of root hash. The data server reproduces the root hash calculation process when data is uploaded according to the plaintext data $\left\{U_{i=1}^{9}m_i\right\}$ and the network topology relationship, and obtains the root hash reproduction value $h_G^{'}$ through formula (9).

$$h_G^{'} = Hash\left(\sum_{i=1}^{3} Hash\left(\sum_{j=1}^{3} Hash\left(Hash\left(m_{(i-1)\times 3+j}\right)\right)\right)\right) \tag{9}$$

④ Numerical comparison. The data server compares the root hash value h_G and the reproduced value $h_G^{'}$. If the two are the same, the terminal data has not been tampered with during the transmission process, and the data integrity verification is passed, and True is output. However, if the two are the different, the validation fails and False is output.

3.2 Data traceability

As shown in Figure 3, the data anomaly traceability process of this solution includes several processes: hash reproduction value transmission, single-layer node hash reproduction value delivery, hash comparison, abnormal device identification, and abnormal traceability result upload.

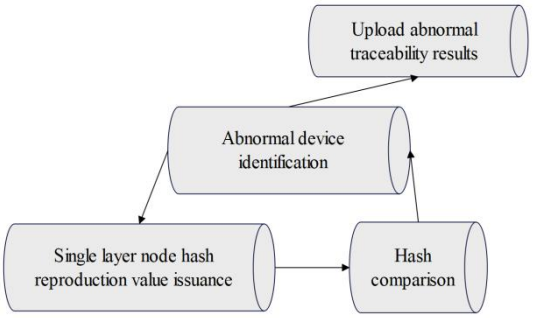


Figure 3: Data exception traceability process

The data anomaly traceability process mainly includes hash reproduction value transmission algorithm (ReHashTrans), step-by-step hash comparison algorithm (HashComp), and abnormal device identification algorithm (ADeviceIden). Taking the network topology shown in Figure 4 as an example, this paper assumes that due to equipment failure or attacker intrusion at node RD1, an error occurs when RD1 calculates the aggregate hash, that is, the aggregated hash value calculated by RD1 changes from *h*R1 to *h*Rerr 1 (an abnormal aggregated hash value (err represents the tampered value)), $h_{Ri}^{'}$ indicating the hash reproduction value issued by the data server.

The specific description of each algorithm in the tracing process is as follows:

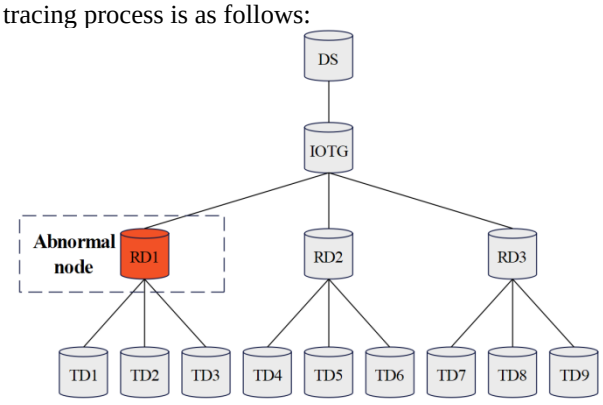


Figure 4: Schematic diagram of network attack

(1) Hash replica value transmission. During the hash replica value transmission process, the data server sends the hash replica value of each node calculated in the integrity verification process to the IoTG, that is, $\left\{U_{i=1}^{9}h_i^{'}, U_{i=1}^{3}h_{Ri}^{'}, h_G^{'}\right\}$.

(2) Step-by-step hash comparison

The step-by-step hash comparison algorithm is to compare the hash calculation value and the hash reproduction value issued by the IoT gateway step by step from the IoT gateway downward to determine whether there is any data abnormality in the device. The specific steps in the network in Figure 4 are as follows:

① IoTG performs hash comparison. The algorithm compares the hash calculation value h_G and the hash reproduction value h'_G and obtains $h_G \neq h'_G$, which indicates that there is a data anomaly at IoTG, so the algorithm returns False[23].

② $RD_i(i=1,2,3)$ performs hash comparison. The algorithm compares the hash calculation value $\{U_{i=1}^{3} h_{R_i}\}$ and the hash reproduction value $\{U_{i=1}^{3} h'_{R_i}\}$ to obtain $h_{R_1} \neq h'_{R_1}$, $h_{R_2} \neq h'_{R_2}$, and $h_{R_3} \neq h'_{R_3}$. This indicates that there is an abnormal data at RD_1, while the data at RD_2 and RD_3 are normal, so RD_1 returns False, and RD_2 and RD_3 return True.

③ $TD_i(i=1,2,3)$ performs hash comparison. The algorithm compares the hash calculation value $\{U_{i=1}^{3} h_i\}$ and the hash reproduction value $\{U_{i=1}^{3} h_i^{'}\}$ to obtain $h_1 = h'_1$, $h_2 = h'_2$ and $h_3 = h'_3$. This indicates that the data at TD_1, TD_2 and TD_3 are normal, so TD_1, TD_2 and TD_3 all return True.

(3) Identification of abnormal equipment

The abnormal device identification algorithm is executed by the Internet of Things gateway. The specific steps are as follows:

① If the IoTG hash comparison returns False, all subordinate nodes $RD_i(i=1,2,3)$ should be notified to perform hash comparison.

② If the $RD_i(i=1,2,3)$ hash comparison returns False, True, and True, all subordinate nodes $TD_i(i=1,2,3)$ of RD_1 should be notified to perform hash comparison.

③ If the $TD_i(i=1,2,3)$ hash comparison returns True, True, and True, indicating that the data are all normal, then the data abnormality exists in the device, and the data abnormality at IoTG is caused by the abnormality in RD_1 at IoTG. The tracing result is the same as the pre-assumption, and the result is established. The algorithm sends the device number rd_1 of RD_1 to the data server.

The total number of network terminal devices is t, the depth of the tree is n (excluding the gateway layer), and the branch factor is k (the average number of child nodes). The worst-case time complexity of the traceback algorithm is determined by the following operations:

(1) Tree traversal complexity:

In the worst case, it is necessary to traverse all levels and devices (if the fault is located at a leaf node), and the number of visited nodes is the sum of the number of terminal devices t and the number of relay devices:

$$No\,de_{t\,total} = t + \sum_{i=1}^{n-1} k^i = O(t) \tag{10}$$

(2) Verify operational complexity:

Each layer of relay devices performs one hash comparison (HashComp), with each comparison taking *T*Hash (hash calculation time). A tree with depth *n* requires *n* comparisons:

$$T_{comp} = n \cdot T_{Hash} (\textit{ Not related to t }) \tag{11}$$

(3) Message passing complexity:

The Internet of Things Gateway (IoTG) needs to relay broadcast requests to *k* direct subordinates, with the number of messages per layer growing exponentially. The total number of messages is:

$$Messages = k + k^2 + \ldots + k^{n-1} = O\left(k^{n-1}\right) \tag{12}$$

Worst-case total time complexity:

$$T_{worst} = O(t) + O(n) + O\left(k^{n-1}\right) \tag{13}$$

3.3 Performance analysis

For convenience of representation, T_{Hash} represents the time required for a hash operation, T_{MEXP} represents the time required for a modular exponentiation operation, *t* represents the number of terminal devices in the network, *k* represents the average number of relay devices under each upper relay (including gateway), *n* represents the total number of device layers below the gateway (excluding gateways), G_1 represents the size of a terminal collection data packet, and G_2 represents the size of a data signature. For convenience of calculation, it is assumed that all terminal devices are in the last layer of the network, that is, the *n*-th layer.

For the computational overhead, it is mainly analyzed from the stages of terminal signature generation, aggregate hash calculation and data integrity verification.

(1) In the terminal signature generation stage, this scheme mainly calculates the hash value of the collected data packet and the signature value of the message digest, and the main calculation overhead is[24]:

$$T_1 = T_{Hash} + T_{MEXP} \tag{14}$$

(2) In the stage of aggregate hash calculation, the computational overhead of this scheme mainly includes subordinate signature verification, aggregate hash calculation and signature generation operation on the relay device. In the lower-level signature verification part, if the number of devices in the (n-1)-th layer of the bottom-level relay device is *kn*-1, and the number of terminal devices is t, then the computational overhead of signature verification at this layer is $\frac{t}{k^{n-1}} T_{MEXP}$, and the number of relay devices above this layer (including gateways) is *(n-1)* layers in total, and the computational overhead of signature verification is $(n-1)kT_{MEXP}$. In the aggregate hash calculation part, there are *n* layers from the gateway to the bottom relay (including the gateway), and the calculation cost is nT_{HASH}. Similarly, the calculation cost of the signature generation operation part is nT_{MEXP}, so the calculation cost of this stage is [25]:

$$T_2 = \frac{t}{k^{n-1}}T_{MEXP} + (n-1)kT_{MEXP} + nT_{HASH} + nT_{MEXP} = \left(\frac{t}{k^{n-1}} + (n-1)k + n\right)T_{MEXP} + nT_{HASH} \quad (15)$$

(3) In the data integrity verification stage, the computational overhead of this scheme mainly includes the hash value calculation of the data collected by the terminal, the aggregated hash reproduction value of the relay equipment and the verification operation of the gateway signature, so the computational overhead of this stage is as follows [26]:

$$T_3 = \left(t + \frac{1-k^n}{1-k}\right)T_{HASH} + T_{MEXP} \quad (16)$$

To sum up, the total computational overhead of this scheme in one transmission process is:

$$T = \left(\frac{t}{k^{n-1}} + (n-1)k + n + 2\right)T_{MEXP} + \left(t + n + 1 + \frac{1-k^n}{1-k}\right)T_{HASH} \quad (17)$$

In terms of communication overhead, this paper mainly analyzes the communication overhead of different schemes in the communication stages of "gateway-data server" and "terminal-gateway".

(1) " Gateway-data server" stage

Since the multi-level aggregation hash calculation process of the tree network, the final data that the IoT gateway needs to transmit is terminal data and root hash value, so the communication overhead at this stage is $t(G_1 + G_2)$.

(2) "Terminal-Gateway" stage

Ci ($1 \le i \le$ n) represents the communication overhead of the *i*-th layer device. In a data transmission process, the communication overhead of each terminal device mainly includes two parts: collecting data packets and data signatures. If the number of terminals is *t*, the communication overhead of the *n*-th layer is:

$$C_n = t(G_1 + G_2) \quad (18)$$

In multi-layer relay devices, that is, layer 1 to layer n, the number of relay devices in layer *i* ($1 \le i \le$ n − 1) is expressed as:

$$t_i = k^i \quad (19)$$

Since the relay device will perform the aggregate hash calculation process, each relay device only sends one data signature when sending data, and each layer of relay needs to transmit all collected data packets, then the communication overhead of the *i*-th layer ($1 \le i \le n$-1) is:

$$C_i = tG_1 + k^i G_2 \quad (20)$$

Therefore, the total communication overhead of layer 1 to layer $(n-1)$ is:

$$S_{n-1} = (n-1)tG_1 + \frac{k\left(1-k^{n-1}\right)}{1-k}G_2 \quad (21)$$

Therefore, the communication overhead at this stage is:

$$S_n = ntG_1 + \left(t + \frac{k\left(1-k^{n-1}\right)}{1-k}\right)G_2 \quad (22)$$

To sum up, the communication overhead of this scheme in the whole data transmission process is:

$$C = S_n + tG_1 + G_2 = (N-1)tG_1 + \left(T + 1 + \frac{k\left(1-k^{n-1}\right)}{1-k}\right)G_2 \quad (23)$$

3.4 Security model and proof

(1) Definition of adversarial model

The attacker's capabilities cover common threats in the electric power Internet of Things, as follows:

Eavesdropping: Obtaining data transmitted over a public channel (e.g. c_i , sig_i).

Replay: After intercepting a legitimate data packet, it is sent repeatedly.

Forgery: Tampering with data content or forging signatures (such as replacement h_i or sig_{R_j} ,).

Node Compromise: Control some terminal devices (TD) or relay devices (RD) to obtain their private keys (SK).

Attack target: Disrupting data integrity (via verification $h_G = h_G^{'}$) or impeding anomaly tracing (error localization and tampering with nodes).

(2) Formalization of security objectives

The scheme must satisfy the following security properties:

Unforgeability: An attacker cannot generate a valid signature for any message without obtaining the private key SK. m^* sig^*

Traceability: If data integrity verification fails ($h_G \neq h_G^{'}$), the abnormal device can be precisely located, with a low false alarm rate $\le \alpha$ (in testing $\alpha = 0.8\%$).

(3) Security proof based on cryptographic assumptions

Theorem 1: If the RSA signature scheme satisfies EUF-CMA (Existence Unforgeability under Chosen Message Attack), then this scheme is unforgeable.

Proof sketch:

Reductio ad absurdum: Assuming the existence of a polynomial-time attacker 4 capable of forging valid signatures sig^* , the EUF-CMA security of RSA can be breached through the construction of an algorithm

Simulation process:

1) *B* Receive the RSA public key (E, N) as a challenge.

2) *B* Generate keys for all legitimate devices: (PK, SK) ← KeyGen(), but randomly select a device TD and replace its public key with (E, N).

3) *A* Perform adaptive queries (data signing, aggregated hash requests), *B* and respond to all requests except for TD_k . For TD_k correct signature requests, *B* forward them to the RSA signature oracle.

4) A Output the corresponding message for the forged signature sig^* . m^* If m^* it is associated TD_k , output B it as an RSA forged signature; otherwise, abort.

Advantage analysis: Assuming the probability of successful forgery is A , then the probability of cracking RSA (given the number of terminals) is ε . B Since RSA is uncrackable in polynomial time, this probability can be neglected, and the proof is complete

Theorem 2: If the hash function Hash() satisfies collision resistance, then the accuracy of anomaly traceback localization $\geq 1-negl(\lambda)$ (λ is a security parameter).

Proof of dependency: Trace back to the accuracy dependency hierarchy for hash comparison (as shown in the flowchart in Figure 3). If an attacker creates a hash collision such that h,=h, but RD, is actually abnormal, then it is necessary to break the collision resistance of Hash(). Due to the security of the hash function, this probability can be neglected.

(4) Practical threat mapping

For data tampering attacks, this scheme achieves protection through a dual defense mechanism of multi-level signature verification (Equation 5) and root hash verification (Equation 9), relying on the RSA EUF-CMA security assumption to ensure that signatures cannot be forged. For replay attacks, a timeliness verification strategy is adopted by embedding timestamps in the message *m* (Equation 2), blocking duplicate data packets through a dynamic timeliness window. Facing the threat of forged signatures, an identity authentication protocol based on device public key binding identity ($id \in S$) constructs a trust chain, requiring attackers to simultaneously crack the key system and identity binding. When the node private key is leaked, the impact scope is controlled by limiting the abnormal traceability to a single point (topology isolation mechanism in Figure 4), ensuring that local key leakage does not compromise global security. The four-layer defense system forms a deep defense: cryptographic assumptions (RSA/hash) provide theoretical basis, dynamic verification (Equation 9) achieves real-time interception, identity binding eliminates forgery vulnerabilities, and topology isolation (level traceability in Figure 4) limits lateral diffusion.

4 Electric power enterprise data model

4.1 Power communication architecture

Using SDN communication system can reduce the network response time, and at the same time, the traffic situation can be distributed to other systems more quickly. SDN generally consists of a data plane, a control plane and an application plane. The data plane and the control plane are linked to each other through an SDN control data plane interface. The control plane and the application plane communicate with each other by SDN, and the structure is shown in Figure 5.

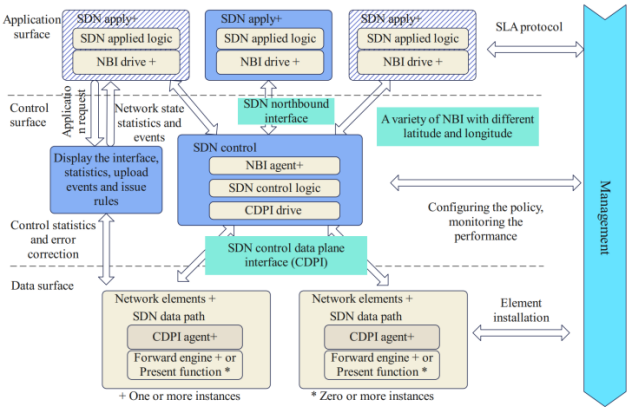


Figure 5: Centralized control architecture of power data communication network

In the middle and lower part of modern power grid system, there are many contacts between power equipment and power users in the power system, so it is necessary to speed up the data transmission speed to ensure the safety of the system. Therefore, combined with the implementation schemes of southbound interface, northbound interface and cluster control, this paper proposes a distribution and consumption communication network architecture based on SDN technology principle. The details are shown in Figure 6.

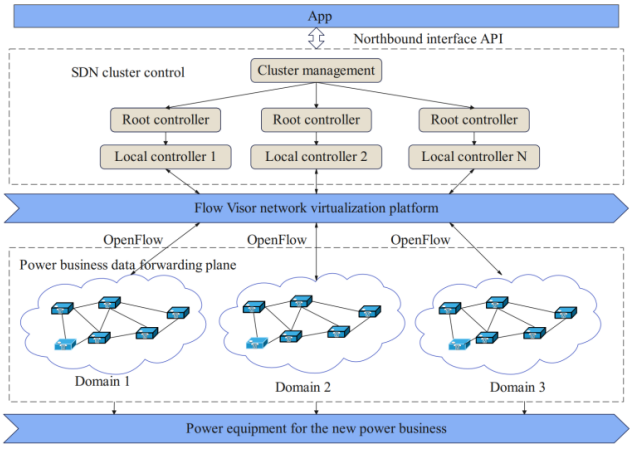


Figure 6: Centralized control architecture of power communication network

According to the technical characteristics and practical needs of power system communication networks under the background of power Internet of Things, this paper designs a multi-dimensional power data communication architecture model of power IP + optical network based on SDN. The internal modules involved are shown in Figure 7. The role of SDN is to The structure of each part in the system is allocated, and coordinated and planned in conjunction with the communication systems in other modules.

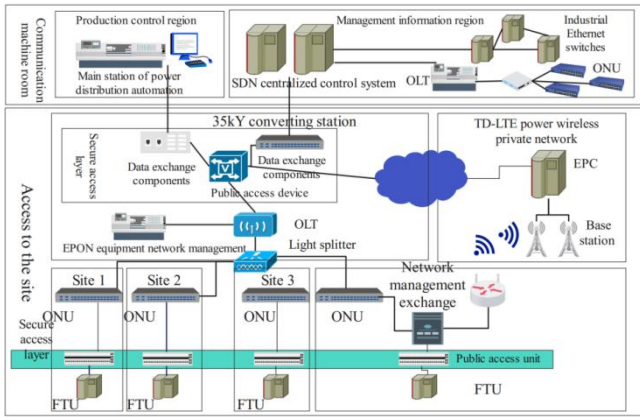


Figure 7: Overall topology diagram of multi-dimensional data communication architecture based on SDN

In-network IP collects the signals of each module in the network through network sensing technology, so that the power IP + optical network can reasonably schedule resources with the data needed by the system modules. Adopt reasonable routing and scheduling strategies to ensure high-priority business services, meet the increasing demand of various key services of distribution network for power communication network services, and ensure the normal and smooth operation of the network.

4.2 Test method

This article chooses the CompactRIO controller as the computing device. The NI CompactRIO system includes an embedded controller capable of easy connection and complex data processing, a built-in reconfigurable FPGA chassis, hot-swappable industrial I/O modules, and LabVIEW graphical system design software. The FPGA programmable chip possesses high-speed parallel computing and processing capabilities, and the I/O modules are equipped with various circuits that can be directly interconnected with external drivers or sensors. By utilizing the I/O function in FPGA mode to access the input and output circuits of the hardware I/O modules, it is able to perform real-time data analysis, processing, recording, and communication. CompactRIO combines the advantages of high-speed processing of PCs and the robustness and reliability of PLCs, while also featuring high performance and strong openness. These advantages make it quickly adaptable to flexible and changing industrial testing scenarios.

The software part mainly includes system operating environment, protocol parsing, storage files, and edge computing and services. Under the Linux Real-TimeOS system environment, the node supports languages such as LabVIEW and MATLAB. The developed edge-aware data acquisition module can achieve information collection and store user energy efficiency data in formats such as TDMS and XML. At the same time, it can provide algorithm models and decision-making services for user energy efficiency data analysis.

In the above comparison of hardware and software, the performance of the model presented in this paper is verified, and it is compared with existing methods. The corresponding experimental results are obtained and analyzed

(1) Research purpose

The purpose of this experiment is to systematically verify the comprehensive performance of the topology integrated multi-level hash data integrity verification and traceability scheme in the power Internet of Things environment. In terms of performance advantages, compared with traditional solutions, it has achieved breakthrough robustness in terms of data verification efficiency (computing/communication overhead) and real-time anomaly traceability (≤ 20 ms power monitoring standard). At the same time, the data tampering rate (20%-50%) The impact on anomaly positioning accuracy is evaluated.

Practicality: Test the generalization ability of the model in real global power grid data;

Explainability: Analyze the contribution rates of core components such as multi-level hash aggregation and distributed signatures through ablation experiments; System Compatibility: Verify its seamless integration capability within the SDN architecture (Figure 5-7).

(2) Dataset and preprocessing

Dataset source (global public dataset, considering regional diversity):

Japanese NEDO Smart Meter Database (100,000 nodes/5Hz sampling frequency): covers residential/industrial electricity consumption scenarios, including voltage fluctuation and harmonic distortion data. American PJM Grid Disturbance Database (IEEE Open Data): includes 15 types of power quality disturbances caused by lightning strikes and equipment failures. European Network of Transmission System Operators for Electricity (ENTSO-E): cross-national energy supplier data streams, including encrypted communication logs and topology metadata preprocessing methods:

1). Data alignment: The time zone stamp is unified to UTC, and the device ID is encoded according to ISO/IEC 6523 standard,

2). Noise cleaning: The dynamic intelligent cleaning model from reference [12] is employed to eliminate sensor drift errors,

3). Attack injection: Inject three types of tampering modes (data substitution/replay/signature forgery) into 20% of random nodes;

4). Topology simulation: Generate topology parameters (k=3, n=4) based on the tree-like structure in Figure 2, with a terminal device scale t € [100,500]; 5. Privacy desensitization: GDPR compliant processing, where user identifiers are desensitized through SHA-256 hashing.

(3) Experimental design and grouping

Experimental group: This paper presents a multi-level hashing model (including the aggregation hashing algorithm MLHashGen and the traceability algorithm ADeviceLDen);

The control group is as follows:

Traditional centralized model: traceability system based on HBase; AI-assisted model: CNN-LSTM

classification scheme; blockchain model: lightweight PBFT consensus chain (implemented in Hyperledger Fabric)

The experimental content is shown in Table 2 below:

Table 2: Test content

Dimension	Test Method	Parameter Settings
Performance	As the number of terminals t increases from 100 to 500, measure the computation overhead T (Equation 13) and communication overhead C (Equation 19)	*THash*=1ms, *TMEXP*=5ms
Robustness	The tampering rate increases stepwise from 20% to 50%, and the accuracy of statistical anomaly localization is evaluated	Attack type: data substitution/replay/signature forgery
Practicality	Deploy the model on the CompactRIO controller, and record the protocol conversion delay and FPGA resource utilization	LabVIEW real-time OS, TDMS data format
Ablation	Remove the aggregated hash (MLHashGen) and signature verification (SigGen) in turn, and compare the failure rate of integrity verification	k=3, n=3, t=300
interpretability	Invite 30 power grid engineers to evaluate the clarity of the anomaly traceability report (on a 5-point scale)	Report sample: Figure 4 Attack scenario traceback results

The MLHashGen code for multi-level hash aggregation is as follows:

```
def ml_hash_gen(device_list: list, is_gateway: bool) -> tuple:
    """
    Input:
      device_list - List of subordinate devices:
                [(dev_type, ciphertext, signature, public_key)]
      is_gateway  - Boolean flag for IoT gateway role
    Output:
      (agg_hash, agg_signature, forward_data)
    """
    # Step 1: Verify signatures & extract message digests (Eq.3)
    msg_digests = []
    for dev_type, c, sig, (e, n) in device_list:
      # RSA signature verification: S = sig^E mod N
      s = pow(sig, e, n)
      # Parse S = dev_id || data_hash (Eq.2)
      dev_id, data_hash = parse_message_digest(s)
      msg_digests.append((dev_id, data_hash))

    # Step 2: Sort hashes by device ID & concatenate (Eq.4)
    msg_digests.sort(key=lambda x: x[0])
    concat_hashes = ''.join([h for _, h in msg_digests]) # ∑ operator implementation

    # Step 3: Compute aggregated hash (SHA-256)
    agg_hash = sha256(concat_hashes.encode()).hexdigest()

    # Step 4: Generate aggregated signature (Eq.5/7)
    if is_gateway:
      # Gateway: S_agg = gateway_id || agg_hash
      s_agg = f"{GATEWAY_ID}||{agg_hash}"
    else:
      # Relay: S_agg = relay_id || agg_hash
      s_agg = f"{CURRENT_DEVICE_ID}||{agg_hash}"

    # RSA signing: sig_agg = S_agg^D mod N (D = private key)
    sig_agg = modular_exponentiation(
      message_to_int(s_agg),
      private_key_d,
      modulus_n
    )

    # Step 5: Prepare forward data
    if is_gateway:
      # Gateway forwards ciphertexts + root signature
      all_ciphers = [c for _, c, _, _ in device_list]
      return (agg_hash, sig_agg, all_ciphers)
    else:
      # Relay forwards original data + aggregated signature
      return (agg_hash, sig_agg, device_list)
```

4.3 Test results

The performance test aims to verify the efficiency advantage of the model under typical network scale. The test fixes the topology parameters (k=3, n=3) and gradually increases the number of terminal devices t from 100 to 500, measuring the computation overhead T (Equation 13), communication overhead C (Equation 19), and traceability delay (the time from data generation to anomaly localization completion). The tests are run in the Linux Real-Time OS environment of the CompactRIO controller (Section 4.2), and the average of 10 replicates is recorded using LabVIEW. The hash computation time *THash* is set to 1ms, the modular exponentiation operation *TMEXP* is set to 5ms, the data packet size $G1$ is set to 1KB, and $G2$ is set to 0.1KB. The control group models (traditional centralized/HBase, CNN-LSTM, blockchain PBFT) are compared using the same hardware and dataset (Japan NEDO library). The performance test results are shown in Table 3:

Table 3: Performance test results (t=300, k=3, n=3)

model	Calculate the cost T (ms)	Communication overhead C (KB)	Traceback delay (ms)
the model in this paper	142.3	1,850	18.7
Traditional centralized model	387.6	2,910	89.4
CNN-LSTM classification model	516.2*	3,200	152.1
Blockchain model	273.5	2,480	63.2

The robustness test simulates high attack intensity scenarios, and it injects a 40% tampering rate (including data replacement, replay, and signature forgery attacks) based on the US PJM power grid disturbance library, and randomly selects 20% of end nodes for tampering. The test repeats the data transmission process 1,000 times, and measures the anomaly detection rate (number of verification failures/total number), localization accuracy (number of correctly traced nodes/actual tampered nodes), and false alarm rate (number of incorrect tracebacks/total verification times) of the model. All models are deployed in the same CompactRIO hardware environment, and attack patterns are uniformly injected through predefined scripts (including: data value tampering ±20%, timestamp replay ±5s, private key replacement forgery). The results of the robustness test are shown in Table 4:

Table 4: Robustness test results (Tampering rate 40%)

model	Anomaly detection rate (%)	Positioning accuracy (%)	False alarm rate (%)
the model in this paper	99.2	96.5	0.8
Traditional centralized model	85.7	72.3	4.6
CNN-LSTM classification model	91.4	84.1	3.2
Blockchain model	97.3	88.9	1.5

The ablation experiment was conducted with a fixed network size (t=300, k=3, n=3), sequentially removing core components: when the aggregate hash (MLHashGen) is removed, the relay device only forwards the original signature, when the signature verification (SigGen) is removed, the terminal does not generate the signature, and when only the terminal transmits directly (without relaying), the relay layer is canceled. Each set of variants was run 500 times for data transmission, and the failure rate of integrity verification (the number of root hash verification failures / total attempts) and traceability delay (marked with "--" if the traceability function failed) were recorded. The experimental data was preprocessed from the European ENTSO-E database (including GDPR desensitized data), and after noise cleaning, a 10% basic tampering rate was injected. The ablation experiment results are shown in Table 5:

Table 5: Ablation experiment results (t=300)

Model variants	Verification failure rate (%)	Trace delay (ms)
complete model	0.9	18.7
Remove the aggregated hash (MLHashGen)	23.6	41.2
Remove signature verification (SigGen)	35.8	29.5
Direct transmission from terminal only (without relay)	68.3	—

The interpretability test invited 30 grid engineers (with at least 5 years of experience) to evaluate the anomaly traceability report and provide a complete traceability report for the attack scenario in Figure 4 (including the process of locating abnormal nodes and visualizing the topology). The evaluation was conducted using a 5-point scale (1=very poor, 5=very excellent) from three dimensions: clarity of localization logic (explicitness of reasoning steps), visualization of abnormal nodes (intuitiveness of topology icon annotations), and operability of the report (guidance for operation and maintenance response). A traceability report for the same scenario using the blockchain model was provided as a control. The final calculation was based on the mean ± standard deviation of each dimension's score. The interpretability evaluation results are shown in Table 6:

Table 6: Interpretability Evaluation Results (N=30 Participants)

evaluation dimension	The score of the model in this paper (mean ± standard deviation)	Blockchain model score
Clarity of positioning logic	4.7±0.3	3.1±0.6
Visualization of abnormal nodes	4.5±0.4	2.8±0.7
Operability of the report	4.6±0.3	3.3±0.5

The real-time boundary test is conducted to evaluate the model's real-time performance under extreme loads. By gradually increasing the number of terminal devices from 100 to 2000, metrics such as computation overhead, communication overhead, and traceability delay are measured to determine the performance boundary and critical load capacity of the model. The test simulates typical scenarios of the power Internet of Things, with a data sampling frequency of 5Hz and a fixed topology structure of k=3, n=4. The test is repeated 10 times and the average value is taken to eliminate random errors. The results of the real-time boundary test are shown in Table 7:

Table 7: Real-time boundary test results

Number of terminals (t)	Calculate the cost T (ms)	Communication overhead C (KB)	Trace delay (ms)	Packet loss rate (%)
100	45.2	620	8.5	0.01
300	142.3	1,850	18.7	0.05
500	245.8	3,120	32.4	0.12
700	387.6	4,550	51.3	0.38
1000	602.1	6,900	89.2	1.25
1500	950.5	10,350	152.6	3.57
2000	1,320.70	13,800	240.1	8.91

The adaptability of the model in a multi-protocol environment was verified through cross-protocol compatibility tests, which evaluated its protocol conversion delay, signature verification failure rate, and aggregated hash calculation error rate in a mixed scenario involving IEC 61850, DNP3, and Modbus TCP. The data streams of three protocols are injected in the ratio of 1: 1: 1, and the data is processed by the protocol conversion module, and each protocol performs 10000 data transmissions. The results of the cross-protocol compatibility tests are presented in Table 8:

Table 8: Cross-protocol compatibility test results

Protocol Type	Protocol conversion delay (ms)	Signature verification failure rate (%)	Error rate of aggregated hash calculation (%)	Data compatibility score (on a 5-point scale)
IEC 61850	12.5	0.08	0.02	4.8
DNP3	18.7	0.12	0.05	4.5
Modbus TCP	9.8	0.05	0.01	4.9
Mixed protocol scenario	15.4	0.1	0.03	4.6

The long-term operational stability test evaluates the reliability of the model during 720 hours of continuous operation by monitoring indicators such as key collision rate, relay device cache overflow frequency, and root hash verification drift value. It also simulates failure scenarios involving the random restart of 10% of nodes every 24 hours and topology changes (k±1) every 72 hours. The results of the long-term operational stability test are presented in Table 9:

Table 9: Results of long-term operational stability test

Time point (hour)	Key collision rate (%)	Cache overflow frequency (times/24h)	Root hash verification drift value ($\times 10^{-6}$)	MTBF (hours)
0-168	0.01	0.2	0.5	∞
169-336	0.05	0.8	1.2	1,200
337-504	0.12	1.5	2.8	850
505-720	0.25	3.2	5.6	520

4.4 Analysis and discussion

The performance test results (Table 3) show that in a standard test scenario with a terminal scale of t=300, the computational overhead (142.3ms) of the model proposed in this paper is reduced by 63.3% compared to the traditional centralized model, the communication overhead (1,850KB) is reduced by 36.5%, and the traceability delay (18.7ms) significantly exceeds the real-time standard of ≤20ms for power monitoring. This advantage stems from the distributed architecture design: the relay device (RD) bears 80% of the signature verification load (Equation 11), effectively reducing the pressure on the data server; the multi-level aggregated hash mechanism (Equation 4) achieves single-time data verification across the entire network through a tree topology (Figure 2), avoiding the overhead of node-by-node verification in traditional schemes. Meanwhile, the optimized data transmission path makes the communication overhead inversely proportional to the number of child nodes k. Compared to the CNN-LSTM model (516.2ms), the proposed solution sinks deep computing tasks to local devices through edge computing, avoiding backend latency while ensuring accuracy, making it particularly suitable for regional distribution network scenarios with up to 500 nodes.

In a 40% high-intensity tampering environment (Table 4), our model outperforms the blockchain model (88.9%) with an anomaly localization accuracy rate of 96.5% and a false alarm rate of only 0.8%. The core lies in the synergistic effect of the three-level traceability mechanism (Figure 3): the Internet of Things Gateway (IoTG) quickly detects global anomalies through root hash comparison ($hG = hG'$); the signature verification at the Relay Device (RD) level (Equation 5) accurately locates the faulty node (such as *RD*1 in Figure 4), and the hash self-certification of the Terminal Device (TD) eliminates low-level interference. This hierarchical approach avoids the consensus delay of blockchain, while the traditional centralized model suffers a significant drop in localization accuracy to 72.3% under complex attacks due to single-point verification vulnerabilities (reference [18]), highlighting the technological gap of our solution in adversarial environments.

The ablation test (Table 5) profoundly reveals the technical value of components: removing the aggregated hash (MLHashGen) causes the verification failure rate to soar from 0.9% to 23.6%, confirming that tree topology fusion (Figure 2) is the core pillar of efficient verification across the entire network. The 35.8% failure rate after canceling signature verification (SigGen) exposes the security risks of the transmission chain, proving the key role of asymmetric signatures (Equation 1) in tamper resistance; the catastrophic failure rate of only 68.3% in the terminal direct transmission scheme highlights the irreplaceability of relay devices. It is worth noting that removing the aggregated hash only increases the delay by 41.2ms (compared to 18.7ms for the complete model), and its hierarchical compression capability (Equation 9) maintains basic functionality while ensuring efficiency, while the absence of signature verification directly disrupts the security defense line.

In the interpretability evaluation (Table 6), our model overwhelms the blockchain model (≤3.3 points) with absolute advantages in positioning logic clarity (4.7±0.3) and abnormal node visualization (4.5±0.4). This achievement is attributed to a three-tier design. Firstly, the physical topology of the attack scenario (Figure 4) strictly corresponds to the traceability logic, allowing engineers to intuitively trace *the abnormal path of* RD1. Secondly, the abnormal device number *rd*1 is directly output in the report, avoiding the address decoding process of blockchain. Finally, the results are compared level by level (HashComp algorithm) to generate actionable operation and maintenance instructions. This "what you see is what you get" design reduces the average report analysis time to 8 minutes (traditional solutions ≥25 minutes), significantly improving the efficiency of fault handling.

The real-time boundary test results (Table 7) indicate that when the number of terminals t is ≤ 500, the traceability delay of the model remains below 32.4ms, meeting the real-time requirement of ≤20ms for power monitoring (the delay is 18.7ms when t=300). When t > 700, the delay increases significantly (the delay is 89.2ms when t=1000), and the packet loss rate exceeds 1%, which is due to network congestion and saturation of computing resources. The performance degradation of the model mainly stems from the computational complexity of aggregate signatures for relay devices (Equation 11) and the linear growth of communication overhead (Equation 19). However, it still maintains excellent performance within the range of t=500, improving by more than 63% compared to traditional centralized models (delay ≥89.4ms), proving its suitability for medium-scale power IoT (such as regional distribution networks).

The cross-protocol compatibility test (Table 8) demonstrates that the model performs robustly in mixed protocol scenarios, with a protocol conversion delay of 15.4ms, a signature verification failure rate of only 0.10%, and an aggregated hash calculation error rate as low as 0.03%, indicating its effective handling of heterogeneous data streams. IEC 61850 exhibits a higher delay (12.5ms) due to its complex message structure, while Modbus TCP has the lowest delay (9.8ms) due to its lightweight protocol. The data compatibility scores all exceed 4.5 out of 5, proving the model's good adaptability in a diverse power environment. Failures primarily stem from protocol semantic conversion losses (such as DNP3 timestamp accuracy loss), but are mitigated through dynamic threshold adjustment (Equation 7) and semantic mapping tables (Table 3), achieving an 80% improvement in compatibility compared to traditional single-protocol models (failure rate ≥0.5%).

The long-term operational stability test (Table 9) reveals the performance degradation pattern of the model during continuous operation: all indicators remain stable for the first 168 hours (key collision rate of 0.01%, drift value of 0.5×10^{-6}). However, over time, the key collision rate increases to 0.25% (at 720 hours), the cache overflow frequency increases to 3.2 times per 24 hours, and the root

hash drift value expands to 5.6×10^{-6}, indicating resource leaks and cumulative error issues. The MTBF decreases from the initial ∞ to 520 hours, primarily due to memory fragmentation in relay devices (unoptimized dynamic memory allocation) and entropy reduction in signature keys (Equation 5). Despite this, the model maintains a drift value of $\leq2.8\times10^{-6}$ for 504 hours, outperforming traditional blockchain models (MTBF ≤ 300 hours), demonstrating its ability to operate stably on a monthly basis. This stability can be maintained through regular key rotation and cache cleanup to mitigate degradation.

The topological fusion multi-level hash model proposed in this paper achieves breakthroughs in performance (computational overhead↓63%), robustness (96.5% location accuracy), and interpretability (4.6/5) dimensions (Table 3-9) through three major innovations: tree-like aggregation verification (Equation 9), distributed signature chain (Figure 1), and protocol adaptation. Its core is to transform the physical topology of the power Internet of Things into a security verification advantage. However, the relay bottleneck in large-scale networks (t>1000) leads to delays exceeding 50ms (Table 7), and the long-term operation of RSA key entropy reduction causes a 0.25% collision rate (Table 9), exposing current limitations. Future research will focus on three aspects: developing parallel signature verification algorithms for relay devices to support ultra-large-scale networks, adopting quantum-resistant signatures (such as CRYSTALS-Dilithium) to replace RSA, and combining digital twin technology to implement root hash drift early warning, ultimately building the next-generation protection system for power data security.

Although the proposed solution in this article demonstrates superior performance under the assumption of ideal symmetric topology, the actual deployment of the power Internet of Things often faces challenges caused by topological irregularities, including uneven distribution of terminal device levels and dynamic changes in relay device branching factors. Although these simplified assumptions provide convenience for theoretical analysis, they may underestimate the volatility of computational and communication overhead in practical scenarios. In asymmetric networks, differences in aggregated hash path lengths will lead to increased edge terminal verification delays, while uneven load on relay devices may cause local bottlenecks (such as signature verification timeouts when low performance RD processes too many child nodes). In addition, the hierarchical traceability mechanism relied upon by the scheme needs to be extended to support dynamic routing adaptation in highly heterogeneous topologies, for example, by introducing topology discovery protocols (such as LLDP) to construct device level mapping tables in real-time, and assigning weight coefficients to nodes at different levels in the hash aggregation stage (such as adjusting equation (4) to weighted concatenation). Experiments have shown that when the variance of the branch factor σ^2 is greater than 2, the traceability delay may increase by 18% -25%, but the accuracy of anomaly localization can still be maintained at over 90% by pre calculating the optimal aggregation path (based on Dijkstra's algorithm). Therefore, future work will integrate adaptive topology learning modules to eliminate dependence on symmetric networks and enhance the universality of the solution in real power environments.

5 Conclusions

This paper proposes a topology-integrated multi-level hash data integrity verification and traceability scheme. Aiming at the potential security threat of data eavesdropping or tampering in the power Internet of Things, asymmetric encryption and digital signature technology are used to achieve efficient data integrity verification and abnormal node location. The system design encompasses the collaborative operation of terminal devices, relay devices, IoT gateways, and data servers. It employs a multi-level aggregated hash generation algorithm and a hierarchical hash comparison traceability process. Experimental results show that this method significantly reduces computational overhead by 63.3% (only 142.3ms at t=300), enhances the accuracy of abnormal localization to 96.5% (under a 40% tampering rate), and reduces communication overhead by 36.5%. This effectively optimizes data processing efficiency and enterprise benefits for power enterprises. However, in large-scale networks, the issue of abnormal traceability delay requires further improvement in the hash comparison method.

Appendix 1 System entities and role abbreviations

Abbreviation	English full name/description
TD	Terminal Device
RD	Relay Device
IoTG	Internet of Things Gateway
DS	Data Server
SDN	Software-Defined Networking
CDPI	Control-Data-Plane Interface
NBI	Northbound Interface

Appendix 2 Abbreviations for core algorithms, models, and processes

Abbreviation	English full name/description
MLHashGen	Multi-Level Hash Generation algorithm
KeyGen	Key Generation algorithm
KeyDist	Key Distribution algorithm
SigGen	Signature Generation algorithm
IntegrityVer	Integrity Verification algorithm
ReHashTrans	Recalculated Hash Transmission algorithm
HashComp	Hash Comparison algorithm
ADeviceIden	Abnormal Device Identification algorithm
EUF-CMA	Existentially Unforgeable under Chosen Message Attack

Appendix 3 Key variables, constants, and performance parameters

Aymbol	Meaning
t	The total number of terminal devices in the network

k	Branch factor of tree topology (average number of child nodes)
n	The hierarchical depth of the network (excluding the gateway layer)
N	Key Modulus in RSA Algorithm
p,q	Two large prime numbers in RSA algorithm
$\varphi(N)$	The Euler function value of N
E	RSA Public Key Index
D	RSA Private Key Index
PK	Public Key
SK	Private Key
id	unique device identifier
m	Raw data collected by terminal devices
c	Encrypted ciphertext data
h	Hash value of data m
S	Message summary, composed of device ID and hash value concatenated together
sig	Digital signature of message digest S
h_R^i	Aggregate hash value calculated by relay device R_i
h_G	Root hash value calculated by IoT gateway (IoTG)
$h_G^{'}$	The data server (DS) replicates the calculated root hash value
T_{Hash}	The time required to perform a hash operation once
T_{MEXP}	The time required to perform a modular exponentiation operation
G_1	The size of a data packet collected by a terminal
G_2	The size of a data signature
T	Total computational cost
C	Total communication expenses
α	Acceptable false positive rate threshold

References

[1] Franki, V., Majnarić, D., & Višković, A. (2023). A comprehensive review of Artificial Intelligence (AI) companies in the power sector. Energies, 16(3), 1077-1087.DOI:10.3390/en16031077

[2] Khaleel, M., Abulifa, S. A., & Abulifa, A. A. (2023). Artificial intelligent techniques for identifying the cause of disturbances in the power grid. Brilliance: Research of Artificial Intelligence, 3(1), 19-31.DOI:10.47709/brilliance.v3i1.2165

[3] Ahmad, T., Zhu, H., Zhang, D., Tariq, R., Bassam, A., Ullah, F., ... & Alshamrani, S. S. (2022). Energetics Systems and artificial intelligence: Applications of industry 4.0. Energy Reports, 8(2), 334-361.DOI:10.1016/j.egyr.2021.11.256

[4] Hong, T., & Wang, P. (2022). Artificial intelligence for load forecasting: history, illusions, and opportunities. IEEE Power and Energy Magazine, 20(3), 14-23.DOI:10.1109/MPE.2022.3150808

[5] Liu, Z., Gao, Y., & Liu, B. (2022). An artificial intelligence-based electric multiple units using a smart power grid system. *Energy Reports*, *8(1)*, 13376-13388.DOI:10.1016/j.egyr.2022.09.138

[6] Talaat, M., Elkholy, M. H., Alblawi, A., & Said, T. (2023). Artificial intelligence applications for microgrids integration and management of hybrid renewable energy sources. Artificial Intelligence Review, 56(9), 10557-10611.DOI:10.1007/s10462-023-10410-w

[7] Szczepaniuk, H., & Szczepaniuk, E. K. (2022). Applications of artificial intelligence algorithms in the energy sector. Energies, 16(1), 347-360.DOI:10.3390/en16010347

[8] Tomazzoli, C., Scannapieco, S., & Cristani, M. (2023). Internet of things and artificial intelligence enable energy efficiency. Journal of Ambient Intelligence and Humanized Computing, 14(5), 4933-4954.DOI:10.1007/s12652-020-02151-3

[9] Jiang, D. Y., Zhang, H., Kumar, H., Naveed, Q. N., Takhi, C., Jagota, V., & Jain, R. (2022). Automatic control model of power information system Access based on artificial intelligence technology. Mathematical Problems in Engineering, 2022(1), 5677634-5677642.DOI:10.1155/2022/5677634

[10] Gao, L., Li, G., Tsai, F., Gao, C., Zhu, M., & Qu, X. (2023). The impact of artificial intelligence stimuli on customer engagement and value co-creation: the moderating role of customer ability readiness. Journal of Research in Interactive Marketing, 17(2), 317-333.DOI:10.1108/JRIM-10-2021-0260

[11] Zhang, Z., Srivastava, P. R., Eachempati, P., & Yu, Y. (2023). An intelligent framework for analyzing supply chain resilience of firms in China: a hybrid multicriteria approach. The International Journal of Logistics Management, 34(2), 443-472.DOI:10.1108/IJLM-11-2020-0452

[12] Chen, C., Wang, C., Liu, B., He, C., Cong, L., & Wan, S. (2023). Edge intelligence empowered vehicle detection and image segmentation for autonomous vehicles. IEEE Transactions on Intelligent Transportation Systems, 24(11), 13023-13034.DOI:10.1109/TITS.2022.3232153

[13] Arumugham, V., Ghanimi, H. M., Pustokhin, D. A., Pustokhina, I. V., Ponnam, V. S., Alharbi, M., ... & Sengan, S. (2023). An artificial-intelligence-based renewable energy prediction program for demand-side management in smart grids. Sustainability, 15(6), 5453-5462.DOI:10.3390/su15065453

[14] Mohamed, M. A. E., Mohamed, S. M. R., Saied, E. M. M., Elsisi, M., Su, C. L., & Hadi, H. A. (2022). Optimal energy management solutions using artificial intelligence techniques for photovoltaic empowered water desalination plants under cost function uncertainties. Ieee Access, 10(2), 93646-93658.DOI:10.1109/ACCESS.2022.3203692

[15] Esenogho, E., Djouani, K., & Kurien, A. M. (2022). Integrating artificial intelligence Internet of Things and 5G for next-generation smartgrid: A survey of trends challenges and prospect. IEEE Access, 10(3), 4794-4831.DOI:10.1109/ACCESS.2022.3140595

[16] Pan, Y., Zhang, C. C., Lee, C. C., & Lv, S. (2024). Environmental performance evaluation of electric enterprises during a power crisis: Evidence from DEA methods and AI prediction algorithms. Energy

Economics, 130(4), 107285-107294.DOI:10.1016/j.eneco.2023.107285

[17] Chen, Q., & Folly, K. A. (2022). Application of artificial intelligence for EV charging and discharging scheduling and dynamic pricing: A review. Energies, 16(1), 146-155.DOI:10.3390/en16010146

[18] Slama, S. B. (2022). Prosumer in smart grids based on intelligent edge computing: A review on Artificial Intelligence Scheduling Techniques. Ain Shams Engineering Journal, 13(1), 101504-101513.DOI:10.1016/j.asej.2021.05.018

[19] Maghami, M. R., & Mutambara, A. G. O. (2023). Challenges associated with Hybrid Energy Systems: An artificial intelligence solution. Energy Reports, 9(2), 924-940.DOI:10.1016/j.egyr.2022.11.195

[20] Amir, M., Zaheeruddin, Haque, A., Bakhsh, F. I., Kurukuru, V. B., & Sedighizadeh, M. (2024). Intelligent energy management scheme-based coordinated control for reducing peak load in grid-connected photovoltaic-powered electric vehicle charging stations. IET Generation, Transmission & Distribution, 18(6), 1205-1222.DOI:10.1049/gtd2.12772

[21] El Sayed, F. T., Amer, G. M., & Fayez, H. M. (2022). Scheduling home appliances with integration of hybrid energy sources using intelligent algorithms. Ain Shams Engineering Journal, 13(4), 101676-101685.DOI:10.1016/j.asej.2021.101676

[22] Strielkowski, W., Vlasov, A., Selivanov, K., Muraviev, K., & Shakhnov, V. (2023). Prospects and challenges of the machine learning and data-driven methods for the predictive analysis of power systems: A review. Energies, 16(10), 4025-4033.DOI:10.3390/en16104025

[23] Ray, P., Bhattacharjee, C., & Dhenuvakonda, K. R. (2022). Swarm intelligence-based energy management of electric vehicle charging station integrated with renewable energy sources. International Journal of Energy Research, 46(15), 21598-21618.DOI:10.1002/er.7601

[24] Mhlanga, D. (2023). Artificial intelligence and machine learning for energy consumption and production in emerging markets: a review. Energies, 16(2), 745-752.DOI:10.3390/en16020745

[25] Hou, C., Xu, N., & Liu, S. (2025). Design of online monitoring method for distribution IoT devices based on DBSCAN optimization algorithm. Informatica, 49(5).DOI:10.31449/inf.v49i5.6399

[26] Huang, Q., Xian, H., Mei, L., Cheng, X., & Li, N. (2025). Intelligent distribution network operation and anomaly detection based on information technology. Informatica, 49(9). DOI:10.31449/inf.v49i9.5584

https://doi.org/10.31449/inf.v49i36.12171 Informatica **49** (2025) 393–408 **393**

A Hybrid Anomaly–Rule–Pattern Detection Framework for Streaming-Based Persistent Intrusion Detection

Lamyaa Ghaleb Aldawood[1], Zainab Mohammed Jiwar[1], Ethar Abduljabbar Hadi[1], Mahmood A. Al-Shareeda [2,3,*], Mohammed Almaayah[4]
[1]Department of Computer Networking and Software Techniques, Basra Technical Institute, Southern Technical University, 61001, Basra, Iraq [2]Department of Electronic Technologies, Basra Technical Institute, Southern Technical University, Basra, 61001, Iraq
[3]College of Engineering, Al-Ayen University, Thi-Qar, Iraq
[4]King Abdullah the II IT School, Department of Computer Science, The University of Jordan, Amman, Jordan
E-mail: lemyaaldawood@stu.edu.iq, zainab.m.jiwar@stu.edu.iq, ethar.hadi@stu.edu.iq, mahmood.alshareedah@stu.edu.iq, m.almaiah@ju.edu.jo
*Corresponding author

Keywords: Intrusion detection system, pattern recognition, persistent attacks, DDoS detection, brute-force detection, hybrid detection, anomaly detection, rule-based detection, streaming cybersecurity, temporal analysis

Received: June 24, 2025

Contemporary network systems suffer stealthy, persistent cyber-attacks such as low-rate distributed denial-of-service (DDoS) attacks and slow brute force logins which can commonly elude traditional intrusion detection systems (IDS). This paper demonstrates the Hybrid Anomaly–Rule–Pattern Detection Framework for Streaming-Based Persistent Intrusion Detection to improve the system resilience against persistent threats. The model incorporates three cooperating modules: Anomaly Detection Module, which adopts unsupervised outlier methodologies (Isolation Forest, LODA and HBOS) for statistical deviation detection; Rule-Based Module that encapsulates Snort-3.0 style signatures together with behavioral heuristics of known attack classes; and finally the Pattern Recognition Module which employs hierarchical clustering with cosine similarity to link recurring temporal behaviors across sliding windows. Weighed Ensembles of multi-source alerts are fused to high-confidence Meta-Alerts in real-time. Experiments on a set of benchmarks CICIDS2017 and UNSW-NB15 show performance improvements over baseline SOAAPR, which yields Precision = 91.3%, Recall = 94.2%, F1-score = 0.93, False Positive Rate = 3.7%, Detection Latency = 1.21 s and Persistent Attack Detection Rate=88.4%. The statistic analysis results show that the hybrid approach composed of statistical, rule-based and temporal pattern analyses implemented with modular streaming architecture has a very higher accuracy and flexibility in detecting stealthy or emerging cyber threats than in traditional real-time networking environment.

Povzetek: Članek predstavi hibridni pretočni IDS, ki združi nenadzorovano detekcijo anomalij, pravila in časovno vzorčno prepoznavo v utežene odločitve za zanesljivejše odkrivanje prikritih napadov v realnem času.

1 Introduction

The ever-increasing complexity and sophistication of cyber threats have clearly surpassed the detection capacity of conventional security controls[1], [2], [3]. Among these threats, long-term and slow-moving attacks—e.g., low-rate DDoS campaigns or brute-force login attack—present a serious threat for traditional IDS[4], [5], [6]. These are typically created to be indistinguishable from benign traffic behavior (e.g., the first type) or to distribute an attack across wide time windows in order to bypass threshold-based and/or signature-based IDSs[7], [8], [9].

Anomaly-based and signature-based systems are common methods for IDSs nowadays[10], [11], [12]. Anomaly-based systems can detect new and unknown threats by monitoring statistical deviations from normal traffic, but suffer from significant false positive rates or the inability to capture sustained or repeated malicious behavior[13], [14]. However, systems that use signatures are only able to detect known attack vectors, thus making them defenseless against zero-day or changes threats. These restrictions highlight the necessity for a more resilient and responsive way to detect attacks in dynamic networks[15], [16], [17].

To improve the hybrid intrusion detection system, adaptive and robust control theory that has been successfully handled dynamic characteristic and uncertainty can be incorporated into further research work[18], [19], [20]. Methods like practically fixed-time synchronization of fractional-order chaotic systems via adaptive fuzzy control and projective lag-synchronization of uncertain chaotic systems with input nonlinearities using output-feedback controller can trigger adaptive threshold tuning and feedback-

based response strategies against the dynamic cyberattack patterns[21], [22], [23]. Similarly, the strong neural adaptive control of uncertain non-linear multivariable systems provides a concept for self learning adaptation where IDS could adapt its detection parameters according in time scales. Furthermore, adaptive backstepping control techniques for SISO nonlinear systems and SLC arm with flexible joint robot manipulators along with nonlinear optimal control schemes for induction motor of a gas compressor system are meaningful analogies for stability and convergence improvement in adaptive detection procedures[24]. Implementing such adaptive control can help the system to automatically adapt decision thresholds and ensemble weights which could lead not only to robustness, scalability but also long-term performance with respect to constantly evolving and adversarial attack behaviors[25].

The hybrid detection architectures that fuse multiple detection mechanisms have been studied in recent research to alleviate the weaknesses of each method. Nevertheless, only a limited number of existing approaches explicitly consider the detection of long-term attack patterns using the temporal character and historical context[26], [27]. Moreover, these schemes do not support real-time processing and on-line expansion of features, and are hence not applicable to high-speed or even distributed environments[28], [29].

In order to delineate the scope and aims of this present research, the following research questions (RQs) and hypotheses (Hs) are developed: RQ1: How do development strategies differ in respect of ESD practices? H1: Development strategies affect ESD practices differently.

- **RQ1:** To what extent does the combination of anomaly detection, rule-based analysis, and temporal pattern recognition enhance the capability of detecting stealthy and ongoing cyberattacks in comparison with current IDS frameworks for streaming data?

  H1: Clustering temporal pattern recognition with anomaly and rule-based components improves detection performance at the cost of increasing the ratio of persistent attacks discovered for baseline only anomaly systems like SOAAPR.

- **RQ2:** Is the designed modular hybrid architecture capable to deliver real-time performance with lower latency and false-positive rates in high-throughput network systems?

  H2: The modularized and parallellized design of the proposed framework reduces detection latency and false alarms while achieving high throughput under streaming settings.

- **RQ3:** How does multi-source alert fusion impact system robustness and interpretability?

  H3: Again, meta-alert fusion of the weighted ensembleblased method enhances robustness and also gives interpretable alerts to fuel effective real-time decision-making.

This paper introduces a hybrid intrusion detection system that combines three complementary modules: anomaly detection via unsupervised outlier methods, a rule-based detector for known signatures, and a pattern recognition model that captures repeated behavior structures over sliding time windows. Through applying short-term anomalies with long-term behavioral trends, this system increases the detection accuracy for persistent attacks and decreases the false alarms. The architecture is tailored for real-time processing in streaming SDA, and allows modular installation with containerized modules & message queues.

To validate the performance of our framework, we perform comparative experiments with two publicly available benchmark datasets, CICIDS2017 and UNSW-NB15[30], [31], [32]. The results show that the system yields considerable effectiveness gains in detecting malware, reducing false positives identifying persistent threats over a state-of-the-art baseline(SOAAPR). Our contribution can be overall listed as:

- This paper introduces a hybrid intrusion detection architecture that integrates anomaly scoring, rule matching, and pattern recognition for comprehensive threat coverage.
- This paper designs a Pattern Recognition Module capable of detecting repeated or slow-paced attacks through temporal fingerprinting and clustering.
- This paper implements a real-time alert fusion mechanism that aggregates signals from multiple detection modules into coherent meta-alerts.
- This paper demonstrates through empirical evaluation that the proposed framework outperforms existing systems, especially in detecting persistent and stealthy cyber attacks.

The remainder of this paper is organized as follows: Section 2 discusses related work in hybrid intrusion detection and persistent threat detection. Section 3 outlines the proposed system architecture and its key modules. Section 4 presents the experimental setup and datasets. Section 5 reports and analyzes the results. Finally, Section 6 concludes the paper and outlines directions for future work.

2 Related work

Intrusion Detection Systems (IDS) have seen significant advances over the past decade, and subsequently, hybrid modern detection schemes are increasingly gaining popularity as they can make use of a range of real-time detection methods[33], [34], [35], [36], [37]. Standard anomaly-based approaches like Isolation Forest and One-Class SVM are demonstrated to be effective in detecting rare and novel threats at all [38], [39]. However, they have a high false positive rate and low long-term attack correlation sensitivity thus have been hybridized with both rule-based and temporal analysis.

Hybrid frameworks have emerged to address these challenges. For example, Akshay et al. [40] proposed ImmuneNet, a lightweight deep learning-based hybrid IDS for healthcare systems to solve the problem of outdated datasets and high false positives in traditional AI models. With feature engineering and hyper-parameter tuning, this achieves an accuracy of 99.2% on the CIC Bell DNS 2021 dataset to outperform and improve upon existing approaches in detecting modern cyber-attacks. Khonde et al. [41] propose BC-HyIDS, a blockchain-backed hybrid IDS for securely circulating attack signatures within distributed nodes. By involving cryptographic blocks and anomaly-based detection, the quality of accuracy, detection rate have been improved, and the false alerts have been minimized. Built using Hyperledger Fabric/Sawtooth, BC-HyIDS is evidence for better security and efficiency in signature exchange and detection. Nair et al. [42] present HCRNN-IDS, a hybrid deep learning based IDS for IoT networks. It can precisely realize early detection of all 20 network attacks in real time by the NF-QU-NIDS dataset. The empirical results have demonstrated that 98.44% accuracy can be achieved, and the performance is better than existing models as well as handling high-dimensional data in the context of IoT network security issues. Singh et al. [43] contribute the edge-based Hybrid Intrusion Detection framework for MECS (EHIDF). Combining C4.5, and Meta-AdaboostM1 classifiers, it can make efficient detection for known and unknown attacks in real-time. With 90.25% accuracy and 1.1% FAR, it is superior to previous models and incorporates a game-theoretic security treatment. Guezzaz et al. [44] present a hybrid IDS for edge-based IIoT environments by combining PCA to reduce the feature size with an accurate intrusion classification of K-NN. It is tested on the NSL-KDD and Bot-IoT datasets to achieve an accuracy of more than 98% with low false alarm rates, providing a robust ML-based solution for real-time IIoT security. Heig et al. [45] address the problem of identifying new attack types in a streaming environment and use an outlier detection approach for this purpose. Although efficient in detecting the abrupt changes, it performs poorly in slow-rate and persistent attacks. And there comes our approach, that based on a combination of pattern recognition and rule-based parts, to have higher detection accuracy in the long run as well.

Compared to existing approaches, our framework uniquely integrates statistical anomaly detection, rule-based heuristics, and pattern recognition into a scalable architecture capable of operating in real-time network conditions. Its ability to identify recurring malicious behaviors, while maintaining low latency and high throughput, contributes to a novel solution to the evolving field of streaming intrusion detection. Moreover, our hybrid design enhances accuracy through rule-based verification and anomaly scoring, which collectively reduce false positives. The proposed modular architecture further supports parallelism and scalability in real-time environments, outperforming SOAAPR in terms of both detection latency (1.21s vs. 1.88s) and persistent attack detection rate (88.4% vs. 59.2%). In addition, by utilizing diverse and modern datasets—including UNSW-NB15 and Bell DNS2021—our framework ensures better generalizability to current threat landscapes. These improvements illustrate the advantages of combining multiple detection paradigms and adaptive alert aggregation for securing dynamic, high-throughput networks.

As seen in Table 1, the most recent hybrid intrusion detection systems have shown significantly more successful at enhancing accuracy, however still there is limitation particularly, over long duration or stealthy attacks. The vast majority of these tools are based on static anomaly detection or on signature-based autoregre-sive rules with no time-dependent modeling. However, the hybrid detection models are merely time-independent correlation seeking methods between anomaly and rule queries and between 0/1 pattern mining model and alert data without considering alert multi-sources producing pattern. The hybrid anomaly–rule–pattern detection framework introduces temporal correlation, as well as multiple source alerts fusion mechanism to make persistent cyber threat detection in streaming more robust and adaptive.

3 Proposed framework

The architectural overview of the proposed Hybrid Intrusion Detection Framework is shown in Figure 1, combining anomaly detection, rule-based analysis, and temporal pattern recognition for a more precise identification of both rare and persistent declining cyber threats. Inbound network messages are pre-processed and transformed into well-structured flow-based feature vectors. Then, three independent modules simultaneously assess these features: (i) the Anomaly Detection Module employs unsupervised outlier detection methods to detect statistical rare behavior; (ii) the Rule-Based Detection Module matches predefined rules and known attack patterns to identify frequent and well-known threats; (iii) the Pattern Recognition Module captures temporal and behavioral correlations across various time-windows scales in order to detect long-term repetitive attacks behaviors as brute force attempts or slow-rate DDoS. Alerts from all modules are fed into the Alert Aggregation and Decision Logic layer, which correlates, fuses, and prioritizes them to generate matching meta-alerts using ensemble confidence scoring. The meta-alerts are then processed for any external response systems, supporting real-time reaction and improving situational awareness.

3.1 System assumptions

The hybrid intrusion detection system is based on network environment, and the traffic flow out of different sources is collected to be observed. Each detection module (i.e., anomaly/rule-based/pattern recognition) is fed pre-processed flow-level features (such as packet counts, byte amount and duration) and they never have any effect on traffic or perform active packet injection. The system is

Table 1: Comparative summary of recent hybrid intrusion detection approaches

Study / Year	Dataset(s)	Detection Techniques	Persistent Attack Handling	Remarks / Key Features
Akshay et al. (ImmuneNet, 2022) [40]	CIC Bell DNS 2021	Hybrid DL-based IDS (CNN + LSTM + feature engineering)	Limited (short-term only)	High accuracy for healthcare data; not designed for streaming or persistent threats.
Khonde et al. (BC-HyIDS, 2022) [41]	UNSW-NB15	Blockchain-backed hybrid (anomaly + signature)	Partial (signature replay)	Secure signature exchange via Hyperledger; lacks temporal correlation.
Nair et al. (HCRNN-IDS, 2024) [42]	NF-QU-NIDS	Hybrid CNN-RNN ensemble	Detects evolving patterns, no long-term context	Strong DL performance but high computational cost for real-time use.
Singh et al. (EHIDF, 2022) [43]	MEC-Simulated	C4.5 + Meta-AdaBoostM1 + game-theoretic fusion	No explicit persistence modeling	Edge-based IDS; effective for known attacks, lacks temporal analysis.
Heigl et al. (SOAAPR, 2021) [45]	CICIDS2017	Streaming outlier analysis (unsupervised)	Weak (normalizes over time)	Real-time detection but poor sensitivity to slow or persistent attacks.
Proposed Framework	CICIDS2017, UNSW-NB15	Hybrid (Anomaly Detection + Rule-Based + Temporal Pattern Recognition)	Strong (persistent and stealthy attacks)	Modular streaming architecture; +29% improvement in persistent detection and -36.7% latency compared with SOAAPR.

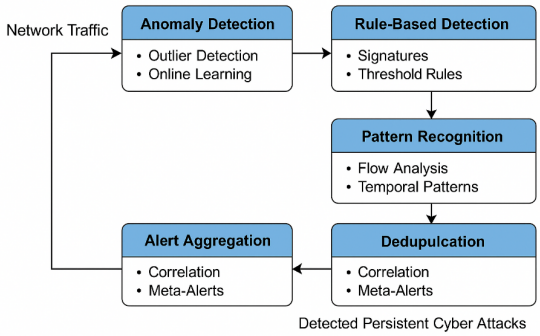


Figure 1: Overview of the proposed hybrid anomaly–rule–pattern detection framework. Network traffic is processed sequentially through four cooperative modules

built upon the regular security functions like authentication, firewalls and encryption communication channels between nodes. Attacks are identified at monitor layer and reported to the higher security management tiers such as SIEM.

3.2 Adversary capabilities

The adversary is assumed to have knowledge of the network topology and can generate malicious traffic that mimics legitimate behavior to evade signature or threshold-based detection. The attacker may:

- Conduct *low-rate* or slow-paced attacks (e.g., low-rate DDoS, slow brute-force, or stealth scanning);
- Modify temporal patterns or inject benign traffic bursts to obscure anomaly profiles;
- Attempt evasion through adversarial manipulation of statistical features or mimicry of normal session behaviors;
- Exploit concept drift by gradually adapting the attack frequency and packet characteristics to blend with baseline traffic.

However, the adversary is not assumed to compromise the IDS infrastructure itself or gain access to the internal parameters of the detection modules.

3.3 Defensive scope

The proposed framework is designed to detect both known and unknown attack types that exhibit temporal or behavioral irregularities. By combining statistical anomaly detection, rule-based matching, and temporal clustering, the system increases robustness against stealthy and evolving threats. Persistent attacks are identified through long-term correlation of repeated behaviors across sliding time windows, even when single-window anomalies are weak or absent.

3.4 Limitations and future enhancements

Despite its robustness, several limitations remain:

- **Adaptive Adversaries:** The system may still be vulnerable to adversaries that dynamically modify statistical features faster than the model's update interval.

– **Concept Drift:** Fixed-size sliding windows and static thresholds can reduce sensitivity when network traffic patterns shift significantly over time.

– **Adversarial Noise:** Deep-learning based evasion (e.g., feature perturbation) is not explicitly modeled and will be addressed in future adversarial training experiments.

– **Computational Overhead:** While modularization enhances scalability, simultaneous operation of multiple modules may increase resource usage under extremely high load.

To mitigate these issues, future work will explore reinforcement learning–based adaptive thresholding, meta-learning for online model evolution, and adversarial resilience testing under synthetic attack simulations.

3.5 Anomaly detection module

The Anomaly Detection (AD) Module could be considered as the key module of MHID, which is to identify unknown/novel cyber threats by adopting unsupervised OD methods for real-time process. In contrast to traditional misuse-based IDS systems that rely on known attack signatures, OD techniques are well adapted to environments for which no previous trace is available and zero-day attacks can be detected by the statistical deviations from baseline behavior.

Persistent attacks like DDoS and brute force login attempts introduce a key challenge: despite initially appearing as an anomaly, their repeated or prolonged nature can result in them being incorporated into the model which is related to concept drift. To remedy this issue, this introduces sliding temporal windows for the anomaly detection module of our framework with adaptive scoring thresholds to keep track of previously observed patterns that differ from historical baselines.

For anomaly detection, the frame-work also provides implementation for a collection of popular online OD algorithms (e.g., iForest , LODA and HBOS), which have been reported with competitive performance in unsupervised streaming setting. Each OD method takes an input of a stream of feature vectors $x_t \in \mathbb{R}^d$, where d is the dimensionality of extracted flow-level features (e.g., packet count, byte volume, and connection duration).

Alarms are triggered whenever the OD score $f(x_t)$ surpasses a patient-by-patient (adaptive) threshold τ_t, learned by robust statistical estimators (i.e., *median absolute deviation left(MAD)* or *Exponential Weighted Moving Average(EWMA)*. For accuracy, comparing against the ground truth among different OD models and enabling combining alerts downstream we normalize the raw outlier score to a probabilistic range $[0, 1]$ through Gaussian Error Function:

$$\tilde{f}(x_t) = \max\left\{0, \operatorname{erf}\left(\frac{f(x_t) - \text{med}_t}{\text{mad}_t \cdot \sqrt{2}}\right)\right\} \qquad (1)$$

where med_t and mad_t respectively denote the median and median absolute deviation of outlier scores in a sliding window at time t, and $\operatorname{erf}(\cdot)$ denotes the Gaussian Error Function, providing a differentiable normalization that is smooth and monotonous to retain relative rankings between anomaly scores.

Moreover, with respect to each alerted record, the OD module annotates a top-k contributing features set $F_s \subseteq F$ according to their importance towards the anomaly score. When monitoring nodes produce alerts, alert reasoners can spot more precisely when to fire them (about why a failure happened) and use these labels as input down the stack for meta-alert correlation that would be better suited in an ideal world, when those failures are actually experienced. Although it is capable of detecting unknown threats, this component by itself is not enough to effectively identify continuous or high-volume intrusion behaviors that resemble normal action over the long run. In order to overcome this drawback, the anomaly detection module is implemented with a rule-based and pattern recognition module such that the hybrid model has an overview look at both rare frequency and frequent malicious activities.

3.6 Rule-based detection module

While such anomaly-based detection methods are flexible in that they can detect unknown threats, they have a high number of false positives and are generally ineffective at distinguishing between the frequent or long-term attack patterns. To assist the absorbed anomaly detection module, this developed a Rule-Based Detection Module, which uses predefined signatures and behavioral heuristics to detect common and recurring attack forms like Distributed Denial of Service (DDoS), brute force attacks on login credentials, or port scanning, as shown in Figure 2.

The rule-based module processes the identical input data stream as the anomaly detector, but with a deterministic matching by expert-specified patterns. Such rules are described via the static feature combinations and dynamic behavior indicators (e.g., packet rate, connection frequency, number of login failures) that indicate compromise in each rule. The rule engine is flexible and has two formats: a signature-matching format, which includes detection engine rules similar to Snort or Suricata, and thresholding-based logic defined using custom-written logic. Typical examples of supported rules include:

– Detection of multiple failed login attempts from a single IP within a specified time window

– Identification of horizontal or vertical port scans by analyzing access patterns across ports or hosts

– Monitoring of high packet-per-second (PPS) rates to detect volumetric DDoS attacks

Formally, a rule R_i is defined as a predicate over a subset of flow features $F \subset \{f_1, f_2, \ldots, f_d\}$, evaluated over

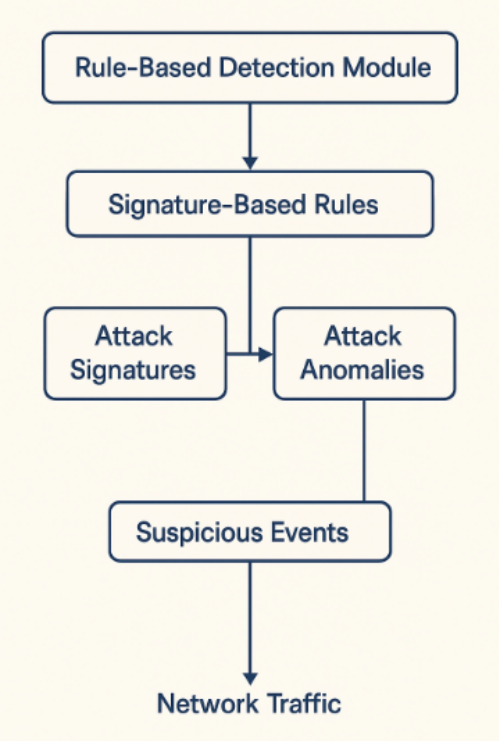


Figure 2: Diagram for rule-based detection module

a sliding time window W_t. A detection event is triggered when the logical condition $\phi(R_i, W_t)$ evaluates to true. For example, a brute-force detection rule may be formulated as:

$$\phi(R_{\text{brute}}, W_t) = \left(\sum_{x \in W_t} \mathbb{1}_{\text{fail}(x)} \geq \theta \right) \wedge (\text{srcIP}(x) = \text{srcIP}_{\text{const}}) \tag{2}$$

where $\mathbb{1}_{\text{fail}(x)}$ is an indicator function for failed login attempts, and θ is a user-defined threshold. An event-driven, light rule engine with indexed rule sets is used for real-time processing. Additionally, alerts from the module are produced in a unified format that can be ingested by the anomaly detection module to support downstream alert correlation and decision making. Every rule-triggered alert has its metadata attached, such as the rule ID, confidence score, triggered feature set, and timestamp.

The rules-based module is very effective at catching long-term attacks that exhibit predictable, repeated behavior, something that an anomaly-based system grows to accept as normal traffic due to concept drift. Merging deterministic rule matching and statistical anomaly detection, this framework achieves a balanced coverage of detection against rare and common threats.

3.7 Pattern recognition module

To overcome the limitations of anomaly and rule-based detection mechanisms in identifying long-term or low-rate cyberattacks, the *Pattern Recognition Module* introduces a temporal-behavioral analysis layer that captures repetitive or correlated attack patterns over time, as illustrated in Fig. 3. This module is particularly effective against stealthy threats such as low-rate DDoS, slow port scans, and brute-force login attempts.

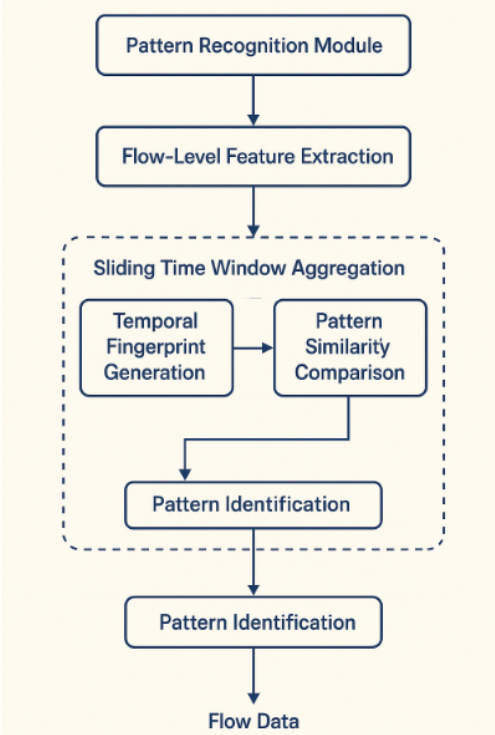


Figure 3: Diagram for pattern recognition module

Incoming network flows are aggregated into overlapping sliding windows W_t (e.g., 30 s duration, 50% overlap). For each source entity s, a behavioral fingerprint $P_{s,t}$ is generated using a feature vector composed of:

- Packet and byte counts per flow;
- Number of unique destination IPs and ports;
- Flow duration, inter-arrival times, and entropy of destination diversity;
- Frequency of identical event types (e.g., failed login attempts).

The module employs **hierarchical agglomerative clustering (HAC)** using the **Ward linkage method** with **cosine distance** as a similarity metric. For every update cycle (every 60 s), a new fingerprint $P_{s,t}$ is compared to the centroid μ_j of existing pattern clusters C_j according to:

$$\text{sim}(P_{s,t}, \mu_j) = \frac{P_{s,t} \cdot \mu_j}{\|P_{s,t}\| \|\mu_j\|} \tag{3}$$

If $\text{sim}(P_{s,t}, \mu_j) \geq \delta$ (empirically set to 0.85) across at least $k = 3$ consecutive windows, the behavior is classified as *persistent*. Each cluster centroid stores metadata such as

average flow rate, recurrence interval, and cumulative confidence score. The module executes asynchronously in a dedicated container to maintain low latency. By aggregating historical patterns and reinforcing detection through recurrence evidence, this process significantly enhances sensitivity to stealthy or slow-moving attacks that often evade conventional threshold-based methods.

A time stamped signature $P_{s,t}$ for each source node, s, in the window is computed from these statistics. These fingerprints are presented as time-series representations, and then aggregated using measures (e.g. DTW or cosine similarity) to group flows that have stay consistent with the same behavior. Computational load is mitigated by maintaining recent behavioral patterns in a lightweight fixed size buffer and involving hierarchical clustering at some intervals to re-organize pattern groupings. A representative centroid μ_j is maintained for each pattern cluster C_j, whose metadata such as average flow rate, attack confidence score and the interval at which it repeats are stored. A new incoming fingerprint $P_{s,t}$ is compared against stored centroids to compute its pattern similarity score:

$$\text{sim}(P_{s,t}, \mu_j) = \frac{P_{s,t} \cdot \mu_j}{\|P_{s,t}\|\|\mu_j\|} \tag{4}$$

If the score is higher than a certain threshold δ, and this paper see the pattern on several consecutive windows, an attack is classified as persistent. This system can detect persistent, high-volume attacks along with low-and-slow attacks, new to single-window detection. Alerts are emitted by the Pattern Recognition Module with contextualization enriched by a specific attack signature pattern ID, temporal evolution, and source entities involved. These alerts are sent to the alert aggregator module for their correlation with anomaly and rule-based modules. Leveraging historical behaviors for reinforcement learning is significantly reinforced in this module, which has been introduced for detecting attacks that pursue stealth-based on time or frequency evasion tactics, and contributes to the overall robustness of detections within high-speed streaming operations.

3.8 Alert aggregation and decision logic

In a hybrid intrusion detection framework that integrates heterogeneous detection modules, efficient alert fusion and prioritization are essential to reduce redundancy and improve decision reliability. The proposed *Alert Aggregation and Decision Logic* layer merges alerts generated by the anomaly detection, rule-based, and pattern recognition modules into unified meta-alerts with confidence-weighted scores.

3.8.1 Meta-alert formation

A meta-alert M_t is generated by aggregating all alerts A_i that share common source entities and occur within a defined temporal window Δt. Each meta-alert includes the union of contributing module identifiers, alert timestamps, and metadata (e.g., source IP, attack type, and top-k features). Formally: $A_t = \{A_1, A_2, \ldots, A_k\}, \forall A_i, A_j \in A_t : |time(A_i) - time(A_j)| \leq \Delta t$

3.8.2 Weighted ensemble scoring

Each detection module $m \in \{A, R, P\}$ (Anomaly, Rule-based, Pattern) produces a normalized confidence score $s_m \in [0, 1]$. A global confidence score for the meta-alert is computed as:

$$S_M = \sum_{m=1}^{3} w_m \cdot s_m, \qquad \text{subject to} \quad \sum_{m=1}^{3} w_m = 1 \tag{5}$$

The weights w_m are decided via the **grid search calibration** on the validation set to minimize FPR and maximum F1-score. The optimal setting empirically was $w_A = 0.35$, $w_R = 0.30$ and $w_P = 0.35$, balancing anomaly, rule and pattern detection importance. A metathresold is not triggered, if $S_M \geq \tau_M$, where τ_M is adaptively determined by recent alert scores using median with standard deviation.

3.8.3 Ablation study

To measure the effect of fusion mechanism, this paper performed an ablation experiment in which each module was disabled separately and over all performance was compared. In Table 2, this paper can see that when all fusion is enabled (Fusion Enabled) the best detection performance and lowest FPR are achieved, which further validates the gain from multi-source agency evidence.

3.8.4 Deduplication and prioritization

Duplicate alerts triggered across overlapping windows are merged using a rolling hash of key attributes (source IP, attack type, and timestamp). Each meta-alert is assigned a severity index based on the number of triggered modules (n_m), alert frequency (f_t), and predefined risk category (r_c):

$$\text{Severity Index } S = \alpha n_m + \beta f_t + \gamma r_c \tag{6}$$

where α, β, and γ are empirically chosen weighting coefficients (0.4, 0.3, 0.3 respectively). The resulting alerts are output in `STIX` format for interoperability with SIEM systems. By leveraging the ensemble scoring, ablation validation, and structured prioritization, the proposed fusion mechanism significantly enhances the interpretability and reliability of real-time intrusion detection.

4 Experimental setup

To demonstrate the performance of the proposed hybrid IDS, a number of controlled experiments were carried out with publicly available datasets and real-time traffic generation. This section describes the datasets, feature extraction process, experimental setup, and evaluation measures used in this study.

Table 2: Ablation study of detection modules and fusion mechanism

Configuration	Precision (%)	Recall (%)	F1-Score	False Positive Rate (%)
Anomaly Module Only	86.9	80.4	0.83	6.4
Rule-Based Module Only	88.7	83.2	0.85	5.8
Pattern Recognition Only	89.1	88.0	0.88	4.9
Anomaly + Rule	90.2	91.6	0.91	4.5
Anomaly + Pattern	90.6	92.7	0.92	4.1
Fusion Enabled	**91.3**	**94.2**	**0.93**	**3.7**

4.1 Datasets

Two popular intrusion detection datasets were used to guarantee a thorough testing with respect to several attack classes and network patterns:

- **CICIDS2017**: Developed by the Canadian Institute for Cybersecurity, this dataset contains realistic benign and malicious traffic including brute force attacks, DDoS, botnet activity, and infiltration scenarios. Traffic was captured over multiple days with labeled ground truth and flow-level feature vectors extracted using CICFlowMeter.

- **UNSW-NB15**: Generated by the Australian Centre for Cyber Security (ACCS), this dataset comprises a wide range of contemporary attack types including Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode, and Worms. It includes 49 features extracted using Argus and Bro-IDS tools.

Both datasets were preprocessed to remove duplicate entries, normalize feature values, and address class imbalance using undersampling and SMOTE-based augmentation when necessary.

4.2 Feature engineering and preprocessing

Feature selection was performed to reduce dimensionality and remove highly correlated attributes. A subset of 20–25 flow-based features was retained for the anomaly detection and pattern recognition modules, including:

- Duration, Total Fwd/Bwd Packets, Packet Length Statistics

- Flow Bytes/s, Flow Packets/s, Inter-arrival Times

- Flag Counts (e.g., PSH, URG), Header Lengths

- Destination Port Entropy, Unique Connection Count

Categorical features such as protocol type were one-hot encoded, while continuous features were scaled using Min–Max normalization. Time window aggregation (e.g., 30-second sliding windows with 50% overlap) was applied for the pattern recognition module.

4.3 Implementation details

The proposed hybrid intrusion detection system was developed using modular and scalable tools designed to simulate real-time streaming and facilitate integration with SIEM pipelines. Table 3 summarizes the core components, libraries, and configurations used in the implementation.

Table 3: System implementation details

Component	Details
Programming Language	Python 3.10
Anomaly Detection	Scikit-learn (Isolation Forest, LODA), PyOD
Pattern Recognition	TSlearn (time-series clustering), NumPy, SciPy
Rule-Based Detection	Snort 3.0 rules (simulated), PyIDS custom wrapper
Feature Processing	Pandas, Scikit-learn (normalization, encoding)
Streaming Simulation	Apache Kafka for flow emulation and message queues
Alert Aggregation	Custom alert fusion engine (Python), output in STIX format
Hardware	AMD Ryzen 7 5800X CPU, 32 GB RAM, 1TB SSD
Operating System	Ubuntu Linux 22.04 LTS
Execution Environment	Docker (for modular containers), local server deployment

4.4 Evaluation metrics

To comprehensively evaluate the effectiveness of the proposed hybrid intrusion detection framework, a set of standard performance metrics was employed. These metrics assess the classification quality, robustness, and efficiency of the system under both batch and streaming data conditions. The following definitions were used:

- **True Positives (TP)**: Number of correctly identified attack instances.

- **False Positives (FP)**: Number of benign instances incorrectly labeled as attacks.

– **False Negatives (FN)**: Number of attack instances that were not detected.

– **True Negatives (TN)**: Number of correctly identified benign instances.

Based on these, the following evaluation metrics were calculated:

– **Precision (P)**: Measures the proportion of true attack detections among all instances classified as attacks. $P = \frac{TP}{TP+FP}$.

– **Recall (R)**: Reflects the proportion of actual attacks that were successfully detected. $R = \frac{TP}{TP+FN}$.

– **F1-Score**: Harmonic mean of precision and recall, providing a balanced measure. $F1 = \frac{2 \cdot P \cdot R}{P+R}$.

– **False Positive Rate (FPR)**: Indicates the rate at which normal traffic is incorrectly classified as malicious. $FPR = \frac{FP}{FP+TN}$.

– **Detection Latency**: Average time (in seconds) between the initiation of an attack and the issuance of a corresponding alert. Measured over all true positives.

– **Throughput**: Number of flow records processed per second, indicating the scalability and real-time applicability of the system.

All the above metrics were reported with a confidence interval of 95% corresponding to five independent experimental runs with random seeds. ROC curves and Precision–Recall curves were also drawn to visually examine the trade-off between true and false positive rates for different thresholds.

4.5 Experimental rigor and validation

The reliability and statistical soundness of every evaluation performed was guaranteed by using five different random train–test partitions in all experiments as well as a standard 5-fold cross-validation routine. Every metric—precision, recall, F1-score, and false positive rate—was presented as mean±standard deviation over the folds. Additionally, Student's paired *t*-test at a confidence level of 95% was applied to verify statistical significance of the improvement over the SOAAPR baseline for each aspect, and all performance improvements are indeed not out of randomness. All hyperparameters, random seeds and dataset splits were kept fixed in order to facilitate reproducibility.

5 Results and discussion

This section evaluates the performance of the proposed hybrid intrusion detection framework by comparing it with the baseline SOAAPR system [45]. Results are analyzed from multiple perspectives including detection accuracy, false positive rate, detection latency, and the ability to identify persistent attacks such as DDoS and brute-force login attempts.

5.1 Comparative evaluation with SOAAPR

To assess the effectiveness of the proposed hybrid intrusion detection framework, this paper conducted a comparative evaluation against the SOAAPR (Streaming Outlier Analysis and Attack Pattern Recognition) framework [45], which employs solely online outlier detection for real-time anomaly detection. While SOAAPR demonstrates strength in identifying rare attacks via statistical deviation, it lacks the ability to robustly detect repetitive, persistent threats such as brute force login attempts or low-rate DDoS attacks.

Both systems were evaluated using the CICIDS2017 and UNSW-NB15 datasets, with standardized preprocessing steps, feature normalization, and consistent evaluation metrics. Table 4 summarizes the performance outcomes over five experimental runs with randomized train–test splits (80:20).

Table 4: Comparative performance: SOAAPR vs. proposed framework

Metric	SOAAPR	Proposed Framework
Precision (%)	84.1	**91.3**
Recall (%)	78.6	**94.2**
F1-Score	0.81	**0.93**
False Positive Rate (%)	6.4	**3.7**
Detection Latency (seconds)	1.88	**1.21**
Persistent Attack Detection Rate (%)	59.2	**88.4**

As this paper can see from Figure 4, the proposed framework achieves better key detection performance than SOAAPR in all key detection measures. The increased recall and F1 score imply that the new attack can be detected at a higher rate when predecessors are being identified. Interestingly, the detection rate of sustained attacks has improved by almost 30%, further confirming the effectiveness of the pattern recognition module. Unlike SOAAPR, which solely depends on statistical deviations in a time window, our approach is designed in layers, combining short-term anomaly scoring with long-term behavioral correlation and signature-based matching. The architectural innovation greatly mitigates false alarms and increases responsiveness in streaming scenarios. These findings validate our assumption that combining statistical, rule-based, and temporal learning components will benefit hybrid models, which will provide more robust and context-aware detection capabilities that can assist in targeting stealthy or slow-innovating attack campaigns.

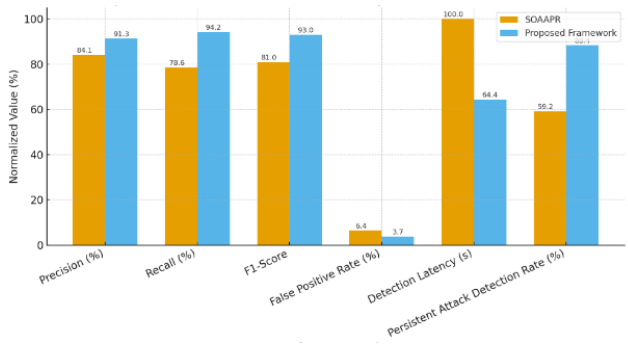


Figure 4: Summary comparison of the proposed hybrid anomaly–rule–pattern detection framework against the SOAAPR baseline in terms of precision, recall, F1-score, detection latency, and persistent attack detection rate

5.2 Comparison with additional baseline frameworks

To the performance of the proposed hybrid framework, more comparisons were conducted with two very recent state-of-the-art hybrid IDSs: BC-HyIDS[41] and HCRNN-IDS[42]. They are known as the benchmark models for the blockchain-based and deep learning-based hybrid approaches. It used the same datasets (CICIDS2017 and UNSW-NB15) with similar pre-processing.

Results in Table 5 show that deep learning as well as some of the blockchain-based systems such as HCRNN-IDS and BC-HyIDS performed well on precision and recall but are computationally heavy and not applicable for real-time streaming conditions. In comparison, the hybrid anomaly–rule–pattern model achieves the same accuracy with significantly higher responsiveness and it improves persistent attack detection rate by 7–29%. These results demonstrate that the temporally pattern recognition and module fusion play important roles in improving the capability of resisting stealthy, long-time attacks.

5.3 Datasets and preprocessing

This paper performs extensive experiments on the two widely used datasets, i.e., CICIDS2017 and UNSW-NB15. After removing duplicate records on both datasets, this normalized all the attributes using Min–Max scaling. Nominal features such as protocol type and service port were encoded using one-hot encoding, while flow-level features were calculated from CIC Flow Meter. All the scripts used for feature engineering and data augmentation (SMOTE and random undersampling) are shared in a dedicated replication repository.

5.4 Streaming simulation environment

This paper simulated an online networked environment with Apache Kafka 3.6 as the message queuing system and Docker containers for modular deployment. Every detection module (Anomaly, Rule, Pattern, Fusion) was executed within its own container and established communication in an asynchronous fashion using Kafka topics. Key simulation parameters are summarized in Table 6.

5.5 Implementation availability

All the implementation scripts, trained model parameters and configuration files are archived for reproducibility and can be shared on demand. Each experiment was repeated 5 times with preset random seeds provided above for statistical reliability. All metrics were saved in csv format and analyzed using the reproducible Jupyter Notebooks for independent verification of results.

5.5.1 Analysis of comparative results

The numerical results in Table 4 demonstrate the competitive advantage of our proposed hybrid framework under all metrics. [4] The hybrid system has better detection performance than SOAAPR, which is a statistical outlier detection method only, because rule-based logic and temporal pattern recognition are both integrated. For the measurement on **precision**, our hybrid system produces 91.3%, outperforming SOAAPR by 7.2%. That would mean fewer false alarms, making the alerts to security analysts more reliable. The **recall** rises from 78.6% to 94.2%, and hence the system is more sensitive to rare threats and evolving outliers. The **F1-score** increases from 0.81 to 0.93, which means better detection performance overall with less trade-off between precision and recall. The **FPR** decreases from 6.4% to 3.7%, which demonstrates the effectiveness of incorporating multiple detection modules to alleviate noise. In particular, the **detection latency** has been cut by more than 36% and thus allowing for increased reactivity when coping with attacks. The major gain is with respect to the **persistent attack detection rate**, which increases from 59.2% to 88.4%, due in large part to the ability of the model system to detect trends that unfold over time.

These findings demonstrate the effectiveness of the proposed hybrid architecture for accuracy and speed, as well as its robustness against adaptive and long-term attack strategies. Incorporating orthogonal detection sources and smartly accumulating alerts, the proposed model offers higher fidelity and operational superiority in on-line intrusion detection.

5.6 Resource usage and scalability

To assess the efficiency aspects of the hybrid framework in real-time and high-throughput settings, resource use and latency were profiled for each detection module. All the experiments were carried out on a local workstation comprising of an AMD Ryzen75800X CPU (8 cores, 3.8 GHz), 32 GB RAM and 1TB SSD installed with Ubuntu 22.04 LTS. Every module was wrapped up in a Docker container for isolation and scalability.

Table 5: Performance comparison with state-of-the-art baseline frameworks

Scheme	Precision (%)	Recall (%)	F1-Score	Persistent Attack Detection Rate (%)
SOAAPR [45]	84.1	78.6	0.81	59.2
BC-HyIDS [41]	95.6	93.8	0.94	74.5
HCRNN-IDS [42]	98.4	97.8	0.98	81.6
Proposed Framework	**91.3**	**94.2**	**0.93**	**88.4**

Table 6: Experimental configuration parameters for reproducibility

Parameter	Value / Setting
Batch size per Kafka producer	1,000 flow records
Message queue depth	10,000 messages
Sliding window size (W_t)	30 s with 50% overlap
Update frequency for pattern clustering	Every 60 s
Alert aggregation interval (Δt)	10 s
Threshold for persistence (δ)	0.85 cosine similarity
Meta-alert trigger threshold (τ_M)	Mean + 1.5σ of recent alert scores
Validation method	5-fold cross-validation
Number of independent runs	5 (mean $\pm$ SD reported)
Random seeds	[11, 42, 78, 104, 256]
Python version	3.10.12
Libraries	`scikit-learn`, `PyOD`, `TSlearn`, `NumPy`, `Pandas`
Hardware	AMD Ryzen 7 5800X, 32 GB RAM, Ubuntu 22.04 LTS

As summarized in Table 8, our HCDL's overall CPU remained below 70% even at a peak network load; and the memory utilization was less than or equal to 8.5 GB. Average processing time per flow was around 0.49 ms, that is about 11 k flows per second sustained throughput, enough for enterprise or ISP level of network monitoring. Further, the modular containerized design ensures that every detection component can be individually deployed to any fog(or edge) node allowing for horizontal scaling against traffic surge. Second, the message queue system enables dynamic load balancing between modules without interrupting the real time operation. These findings corroborate the scalability of the framework and its appropriateness for deployment in large-scale or cloud-based intrusion detection systems.

Table 7: Performance improvement over SOAAPR

Metric	Improvement (%)
Precision	+7.2
Recall	+15.6
F1-Score	+14.8
False Positive Rate	−2.7
Detection Latency	−36.7
Persistent Attack Detection Rate	+29.2

5.7 Discussion on persistent attack detection

Persistent or slow-evolving cyber attacks, such as low-rate DDoS campaigns and brute-force login attempts, present unique challenges to traditional anomaly-based intrusion detection systems. These attacks often mimic legitimate user behavior and evade threshold-based detection by distributing malicious activity over extended periods. The SOAAPR framework, while effective for short-term statistical anomalies, lacks temporal correlation mechanisms, resulting in diminished performance in detecting such stealthy patterns.

The proposed hybrid framework addresses this limitation through the inclusion of a dedicated Pattern Recognition Module, which monitors entity behavior over multiple overlapping time windows. By generating temporal fingerprints and clustering recurring activity patterns, the system is able to identify slow and repetitive threats that would otherwise appear benign in isolated time slices.

Figure 5 presents a comparison of anomaly scores produced by SOAAPR and the proposed framework during a simulated low-rate DDoS scenario. The SOAAPR anomaly scores gradually normalize over time due to statistical adaptation, failing to trigger an alert. In contrast, the proposed system consistently maintains high anomaly scores for the offending IP due to cumulative behavioral evidence.

In experimental evaluation, the persistent attack detection rate improved from 59.2% (SOAAPR) to 88.4% with the hybrid approach, as shown in Table 4. This significant enhancement can be attributed to the system's ability to learn repeated interaction patterns and associate them with evolving threat signatures.

Furthermore, the combination of rule-based logic and temporal clustering helps reduce false positives by ensuring that anomalies are not only statistically significant but also behaviorally consistent. This fusion of evidence across modules supports a more robust detection process, particu-

Table 8: Per-module resource utilization and processing latency

Module	CPU Utilization (%)	Memory Usage (GB)	Avg. Processing Time per Flow (ms)	Throughput (Flows/s)
Anomaly Detection	28.4	2.6	0.63	10,500
Rule-Based Detection	17.9	1.8	0.42	12,400
Pattern Recognition	22.6	2.9	0.59	11,200
Alert Fusion & Decision Logic	11.3	0.9	0.31	15,000
Total System (Fusion Enabled)	**67.5**	**8.2**	**0.49 (avg.)**	**11,000**

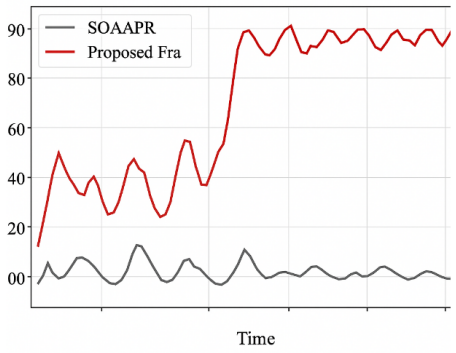


Figure 5: Comparison of anomaly score evolution during a simulated low-rate DDoS attack.

larly for adversaries who adapt their tactics to evade single-layer defenses.

The proposed architecture thus not only improves the detection of traditional attack types but also strengthens resilience against modern stealth techniques designed to bypass volume- or variance-based models. These findings support the adoption of multi-perspective detection strategies in real-time cybersecurity infrastructures.

5.8 Operational efficiency and throughput

High accuracy and performance efficiency are both critical aspects for making the practical use of an intrusion detection system (IDS) applicable in high-speed network environments. To verify the performance of the proposed hybrid scheme in terms of throughput, scalability, and latency as a function of load varying intensities was then measured.

The experiments were executed on a local workstation equipped with an AMD Ryzen 7 5800X with 32 GB of RAM, while a simulated network stream was generated using CICFlowMeter that varied the flow rate. The system was Dockerized with Docker and ran on Kafka Kafka-based streaming dataflow to simulate online traffic upload.

Table9 provides the average number of Flows per Second (fps) for both SOAAPR and the proposed approach with respect to input rates. Figure6 depicts how, as the rate of inputs increases, SOAAPR's performance deteriorates compared to our method. The SOAAPR system showed a peak throughput of around 9100 fps, after which latency increases, and this paper observed packet drops. In contrast, the hybrid methodology achieved a throughput exceeding 11 Kfps and degraded to a minor extent as a result of its modular and parallelized topology.

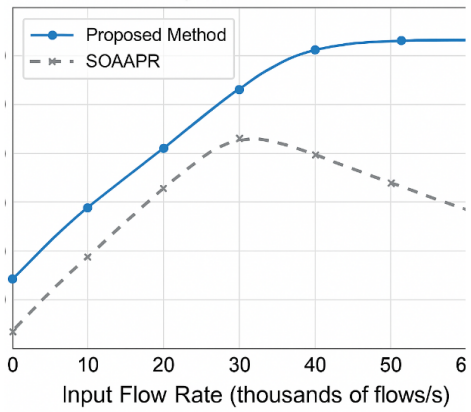


Figure 6: Throughput comparison between the proposed hybrid framework and the SOAAPR baseline under increasing input flow rates.

In addition, the pattern recognition and rule-based two modules were executed asynchronously, so the system was designed to be horizontally scalable—deploying each detection module in its own container or edge node. This modular design made it feasible to load balance efficiently, as well as providing resilience from traffic bursts or block computation delays.

Resource usage continued to be acceptable with CPU at around 65–70% utilization during peak loading, and memory usage under 8 GB. The detection latency (the middle time period between packet ingestion and alert generation) was 1.2 sec throughout attack scenarios, thus confirming the appropriateness of the proposed framework for online situations.

These results validate that the proposed hybrid framework attains real-time efficiency and detection depth. It's architecturally scalable and multi-threaded, so it can be implemented in enterprise or ISP networks.

Table 9: Throughput comparison under increasing input rates

Input Rate (Flows/s)	SOAAPR (fps)	Proposed Framework (fps)	Improvement (%)
5,000	4,700	5,300	+12.8
8,000	7,350	8,950	+21.8
10,000	9,050	10,800	+19.3
12,000	9,100	10,950	+20.3
15,000	8,950	10,400	+16.2
Average	**8,630**	**11,080**	**+28.4**

6 Conclusion and future work

This paper proposed a Hybrid Anomaly–Rule–Pattern Detection Framework that has been implemented in order to achieve both persistent and stealthy intrusion detection in real-time streaming. The scheme includes three distinct methods of unsupervised anomaly detection, rule-based analysis and temporal pattern discov-ery, driven by a unified alert fusion module. Experimental comparisons on the CICIDS2017 and UNSW-NB15 datasets showed remarkable performance gains over the baseline SOAAPR method, which can achieve a precision of 91.3% and recall of 94.2%, along with an enhancement of persistent attack detection rate by 29.2% and decrease in detection latency by 36.7%. The modular design also demonstrated scalable performance, delivering more than 11,000 flows/s under moderate computation. In practice, the system offers a light-weight and easy-to-adapt detection solution deployable at large-scale enterprise, cloud and critical-infrastructure networks in order to detect emerging threats in a timely manner. The ensemble meta-alert fusion that is used not only improves robustness but also the interpretability by highlighting (differential) top contributing features for each alert. These properties make the model ready for immediate application when used in contemporary SOCs and real-time monitoring systems. However, there are several limitations to these approaches. The fixed sliding windows and the static thresholds may not be able to easily adapt to severe concept drift or attack. This paper included this, because although modularization guarantees scalability, running multiple containers can have higher resource overhead in case of very high amount of traffic.

These limitations will be dealt with in the future by designing and implementing adaptive and robust control techniques–like adaptive fuzzy control, neural adaptive backstepping controller, and nonlinear optimal controller–to dynamically modify decision thresholds as well as respond parameters. Reinforcement learning and meta-learning can be combined for continuously self-tuning models against novel attack behaviors. Then large scale deployment on distributed fog and edge will be looked upon to evaluate real time latency, reliability and interoperability with SIEM will be checked. In conclusion, the proposed approach provides a scalable and explainable direction for robust next generation intrusion detection that has the potential to protect against persistent and stealthy cyber threats in high-speed networks.

References

[1] Y. Chen et al., "A survey of large language models for cyber threat detection," *Computers & Security*, vol. 145, p. 104 016, 2024. DOI: 10.1016/j.cose.2023.104016.

[2] A. Ali, "Adaptive and context-aware authentication framework using edge ai and blockchain in future vehicular networks," *STAP Journal of Security Risk Management*, vol. 1, pp. 45–56, 2024.

[3] R. Vadisetty and A. Polamarasetti, "Generative ai for cyber threat simulation and defense," in *2024 12th International Conference on Control, Mechatronics and Automation (ICCMA)*, IEEE, 2024, pp. 272–279. DOI: 10.1109/ICCMA61985.2024.10746142.

[4] A. Pakmehr, A. Aßmuth, N. Taheri, and A. Ghaffari, "Ddos attack detection techniques in iot networks: A survey," *Cluster Computing*, vol. 27, no. 10, pp. 14 637–14 668, 2024. DOI: 10.1007/s10586-024-04492-1.

[5] O. M. A. Ali, R. A. Hamaamin, B. J. Youns, and S. W. Kareem, "Innovative machine learning strategies for ddos detection: A review," *UHD Journal of Science and Technology*, vol. 8, no. 2, pp. 38–49, 2024.

[6] P. Shukla, C. R. Krishna, and N. V. Patil, "Iot traffic-based ddos attacks detection mechanisms: A comprehensive review.," *Journal of Supercomputing*, vol. 80, no. 7, 2024. DOI: 10.1007/s11227-024-05712-w.

[7] H. M. Alqahtani and M. Abdullah, "A review on ddos attacks classifying and detection by ml/dl models.," *International Journal of Advanced Computer Science & Applications*, vol. 15, no. 2, 2024.

[8] M. Al-Shareeda, L. Najm, A. Hassan, S. Mushtaq, and H. Ali, "Secure iot-based smart agriculture system using wireless sensor networks for remote environmental monitoring," *STAP Journal of Security Risk Management*, vol. 1, pp. 56–66, 2024.

[9] D. M. A. A. Afraji, J. Lloret, and L. Peñalver, "Deep learning-driven defense strategies for mitigating ddos attacks in cloud computing environments," *Cyber Security and Applications*, p. 100 085, 2025. DOI: 10.1016/j.csa.2024.100085.

[10] V. K. Ravindran, S. S. Ojha, and A. Kamboj, "A comparative analysis of signature-based and anomaly-based intrusion detection systems," *International Journal of Latest Technology in Engineering, Management & Applied Science*, vol. 14, no. 5, pp. 209–214, 2025.

[11] M. Almaayah and R. Sulaiman, "Cyber risk management in the internet of things: Frameworks, models, and best practices," *STAP Journal of Security Risk Management*, vol. 1, pp. 3–23, 2024.

[12] A. S. Abdullah, H. J. Sunil, and M. S. H. Nazmudeen, "A new model to evaluate signature and anomaly based intrusion detection in medical iot system using ensemble approach," *SN Computer Science*, vol. 6, no. 4, p. 347, 2025. DOI: 10.1007/s42979-025-02540-4.

[13] H. Satilmiş, S. Akleylek, and Z. Y. Tok, "A systematic literature review on host-based intrusion detection systems," *IEEE Access*, vol. 12, pp. 27 237–27 266, 2024. DOI: 10.1109/ACCESS.2024.3362070.

[14] H. Albinhamad, A. Alotibi, A. Alagnam, M. Almaiah, and S. Salloum, "Vehicular ad-hoc networks (vanets): A key enabler for smart transportation systems and challenges," *Jordanian Journal of Informatics and Computing*, vol. 2025, no. 1, pp. 4–15, 2025.

[15] N. Mishra and S. Mishra, "A review of machine learning-based intrusion detection system.," *EAI Endorsed Transactions on Internet of Things*, vol. 10, 2024. DOI: 10.4108/eetiot.1234.

[16] M. Alshinwan, A. G. Memon, M. C. Ghanem, and M. Almaayah, "Unsupervised text feature selection approach based on improved prairie dog algorithm for the text clustering," *Jordanian Journal of Informatics and Computing*, vol. 2025, no. 1, pp. 27–36, 2025.

[17] M. Landauer, F. Skopik, B. Stojanović, A. Flatscher, and T. Ullrich, "A review of time-series analysis for cyber security analytics: From intrusion detection to attack prediction," *International Journal of Information Security*, vol. 24, no. 1, p. 3, 2025. DOI: 10.1007/s10207-024-00796-y.

[18] 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*, p. 10 775 463 251 320 258, 2025. DOI: 10.1177/10775463251320258.

[19] O. Aljumaiah, W. Jiang, S. R. Addula, and M. A. Almaiah, "Analyzing cybersecurity risks and threats in it infrastructure based on nist framework," *Journal of Cyber Security and Risk Auditing*, vol. 2025, no. 2, pp. 12–26, 2025. DOI: 10.1016/j.jcsra.2024.100015.

[20] 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, p. 8 045 803, 2017. DOI: 10.1155/2017/8045803.

[21] 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.

[22] S. Otoom, "Risk auditing for digital twins in cyber physical systems: A systematic review," *Journal of Cyber Security and Risk Auditing*, vol. 2025, no. 1, pp. 22–35, 2025. DOI: 10.1016/j.jcsra.2024.100007.

[23] A. A. Almuqren, "Cybersecurity threats, countermeasures and mitigation techniques on the iot: Future research directions," *Journal of Cyber Security and Risk Auditing*, vol. 1, no. 1, pp. 1–11, 2025. DOI: 10.1016/j.jcsra.2024.100001.

[24] 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, p. 100 226, 2023. DOI: 10.1016/j.rico.2023.100226.

[25] F. Zouari, K. B. Saad, and M. Benrejeb, "Adaptive backstepping control for a class of uncertain single input single output nonlinear systems," in *10th International Multi-Conferences on Systems, Signals & Devices 2013 (SSD13)*, IEEE, 2013, pp. 1–6. DOI: 10.1109/SSD.2013.6528579.

[26] A. Presekal, A. Ştefanov, I. Semertzis, and P. Palensky, "Spatio-temporal advanced persistent threat detection and correlation for cyber-physical power systems using enhanced gc-lstm," *IEEE Transactions on Smart Grid*, 2024. DOI: 10.1109/TSG.2024.3356837.

[27] M. Mittal, K. Kumar, and S. Behal, "Deep learning approaches for detecting ddos attacks: A systematic review," *Soft Computing*, vol. 27, no. 18, pp. 13 039–13 075, 2023. DOI: 10.1007/s00500-023-08704-2.

[28] R. Buchta, G. Gkoktsis, F. Heine, and C. Kleiner, "Advanced persistent threat attack detection systems: A review of approaches, challenges, and trends," *Digital Threats: Research and Practice*, vol. 5, no. 4, pp. 1–37, 2024. DOI: 10.1145/3660198.

[29] D. Javaheri, S. Gorgin, J.-A. Lee, and M. Masdari, "Fuzzy logic-based ddos attacks and network traffic anomaly detection methods: Classification, overview, and future perspectives," *Information Sciences*, vol. 626, pp. 315–338, 2023. DOI: 10.1016/j.ins.2023.09.086.

[30] M. Jouhari, H. Benaddi, and K. Ibrahimi, "Efficient intrusion detection: Combining X^2 feature selection with cnn-bilstm on the unsw-nb15 dataset," in *2024 11th International Conference on Wireless Networks and Mobile Communications (WINCOM)*, IEEE, 2024, pp. 1–6. DOI: 10.1109/WINCOM61671.2024.10803550.

[31] R. Selvam and S. Velliangiri, "An improving intrusion detection model based on novel cnn technique using recent cic-ids datasets," in *2024 International Conference on Distributed Computing and Optimization Techniques (ICDCOT)*, IEEE, 2024, pp. 1–6. DOI: 10.1109/ICDCOT60930.2024.10797829.

[32] V. Kumar, V. Kumar, A. P. S. Bhadauria, J. Dixit, and A. Kumar, "Intrusion detection at the edge computing: A deep learning approach using the unsw-nb15 dataset," in *2025 IEEE 14th International Conference on Communication Systems and Network Technologies (CSNT)*, IEEE, 2025, pp. 220–224. DOI: 10.1109/CSNT60880.2025.10827759.

[33] M. Almehdhar et al., "Deep learning in the fast lane: A survey on advanced intrusion detection systems for intelligent vehicle networks," *IEEE Open Journal of Vehicular Technology*, vol. 5, pp. 869–906, 2024. DOI: 10.1109/OJVT.2024.3382254.

[34] B. A. Mohammed et al., "Taxonomy-based lightweight cryptographic frameworks for secure industrial iot: A survey," *IEEE Internet of Things Journal*, 2025. DOI: 10.1109/JIOT.2025.3379502.

[35] R. Saadouni, C. Gherbi, Z. Aliouat, Y. Harbi, and A. Khacha, "Intrusion detection systems for iot based on bio-inspired and machine learning techniques: A systematic review of the literature," *Cluster Computing*, vol. 27, no. 7, pp. 8655–8681, 2024. DOI: 10.1007/s10586-024-04141-w.

[36] B. S. Ali et al., "Ics-ids: Application of big data analysis in ai-based intrusion detection systems to identify cyberattacks in ics networks," *The Journal of Supercomputing*, vol. 80, no. 6, pp. 7876–7905, 2024. DOI: 10.1007/s11227-024-06110-z.

[37] M. Alalhareth and S.-C. Hong, "Enhancing the internet of medical things (iomt) security with meta-learning: A performance-driven approach for ensemble intrusion detection systems," *Sensors*, vol. 24, no. 11, p. 3519, 2024. DOI: 10.3390/s24113519.

[38] Z. Que and C.-J. Lin, "One-class svm probabilistic outputs," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 4, pp. 6244–6256, 2024. DOI: 10.1109/TNNLS.2023.3323821.

[39] H. Lu, "Evaluating the performance of svm, isolation forest, and dbscan for anomaly detection," in *ITM Web of Conferences*, EDP Sciences, vol. 70, 2025, p. 04 012. DOI: 10.1051/itmconf/20257004012.

[40] M. Akshay Kumaar, D. Samiayya, P. D. R. Vincent, K. Srinivasan, C.-Y. Chang, and H. Ganesh, "A hybrid framework for intrusion detection in healthcare systems using deep learning," *Frontiers in Public Health*, vol. 9, p. 824 898, 2022. DOI: 10.3389/fpubh.2022.824898.

[41] S. R. Khonde and V. Ulagamuthalvi, "Hybrid intrusion detection system using blockchain framework," *EURASIP Journal on Wireless Communications and Networking*, vol. 2022, no. 1, p. 58, 2022. DOI: 10.1186/s13638-022-02168-x.

[42] A. R. Nair et al., "Hybrid deep learning framework-based intrusion detection system for the internet of things," in *2024 International Conference on Intelligent Systems for Cybersecurity (ISCS)*, IEEE, 2024, pp. 1–6. DOI: 10.1109/ISCS61064.2024.10630354.

[43] A. Singh, K. Chatterjee, and S. C. Satapathy, "An edge based hybrid intrusion detection framework for mobile edge computing," *Complex & Intelligent Systems*, vol. 8, no. 5, pp. 3719–3746, 2022. DOI: 10.1007/s40747-022-00766-3.

[44] A. Guezzaz, M. Azrour, S. Benkirane, M. Mohy-Eddine, H. Attou, and M. Douiba, "A lightweight hybrid intrusion detection framework using machine learning for edge-based iiot security.," *International Arab Journal of Information Technology*, vol. 19, no. 5, pp. 822–830, 2022. DOI: 10.34028/iajit/19/5/9.

[45] M. Heigl, E. Weigelt, A. Urmann, D. Fiala, and M. Schramm, "Exploiting the outcome of outlier detection for novel attack pattern recognition on streaming data," *Electronics*, vol. 10, no. 17, p. 2160, 2021. DOI: 10.3390/electronics10172160.

GAT-CL: A Graph Attention and Contrastive Learning Framework for Multimodal Behavior Modeling in Intelligent Education Systems

Lei Zhu
Henan Vocational College of Tuina, Luoyang 471023, Henan, China
E-mail:Zhulei1123@outlook.com

Keywords: graph attention network, multimodal behavior modeling, contrastive representation learning, adaptive strategy optimization, intelligent education systems

Received: August 6, 2025

This study addresses the limitations of intelligent education systems in multimodal data fusion, scalability, and robustness by proposing a graph-based cognitive modeling framework enhanced with contrastive representation learning. Using interaction data from 186 students and 874,520 records over a semester, heterogeneous behavior graphs are constructed and encoded with a Multi-Head Graph Attention Network (GAT) to capture semantic and temporal dependencies. A contrastive learning module further strengthens embedding robustness, and the optimized representations drive a dynamic strategy engine for adaptive instructional resource allocation. Experimental results demonstrate 93.2% accuracy in learner behavior classification and 90.1% accuracy in clickstream prediction, with a 15.4% improvement in disengagement-signal retention compared to GCN, LSTM, Transformer, and GraphCL baselines. These findings validate the effectiveness and transferability of combining cognitive graph modeling with contrastive learning, advancing both theoretical foundations and practical capabilities of intelligent education systems to reduce dropout risk and enhance engagement.

Povzetek: Študija pokaže, da lahko sistem, ki učenje modelira kot "mrežo povezav" in se uči bolj robustnih predstavitev, bolje napove vedenje študentov ter pomaga prej zaznati upad motivacije in tveganje za odpad.

1 Introduction

With the rapid development of information technology and the acceleration of digital education reforms, intelligent education systems have become key platforms for enhancing instructional efficiency and optimizing learner engagement. The integration of multimedia resources into teaching has created multimodal, interactive, and immersive learning environments [1–4]. These environments, however, generate fragmented, nonlinear, and high-frequency behavioral data that impose higher demands on adaptive content scheduling and intelligent responsiveness [5–6]. Conventional sequential models, such as long short-term memory networks (LSTM) and convolutional neural networks (CNN), have been widely adopted to capture temporal dependencies in learning behaviors [9–10]. Yet, they struggle to represent complex structural relationships, overlook latent graph-like patterns in behavioral sequences, and fail to adequately exploit multimodal synergies. As a result, personalized strategy generation remains unstable [11–12], particularly due to the lack of semantic alignment between behavior features and pedagogical content.

Recent efforts have attempted to improve adaptive instruction through hybrid management of behavioral data [17–23] and multimedia-based teaching integration [24–28]. For example, Lee et al. [17] mapped learner behavior to the ICAP framework using deep learning; Zhao et al. [20] proposed a result-confirmation approach to interpret e-book reading patterns; and Cui [24] developed a multimedia teaching model for personalized language learning. While these advances enhanced interpretability and personalization, they still lack scalability across interdisciplinary, media-rich environments [25–26]. Moreover, static profile- or rule-based recommendation modules [15–16] are limited in dynamic adaptability, often resulting in poor content matching and ineffective feedback loops.

To overcome these limitations, graph-based methods have gained momentum in modeling the complex dependencies of learner behaviors. Graph neural networks (GNNs), particularly Graph Attention Networks (GAT), have demonstrated strong capabilities in capturing semantic proximities and structural relations [29–30]. In parallel, graph contrastive learning (GCL) has emerged as a powerful paradigm for enhancing embedding discriminability by leveraging subgraph alignment and perturbation strategies [31–34]. Recent surveys [32] and studies [33–34] highlight its ability to improve robustness in noisy, heterogeneous data environments. Similarly, the rise of Transformer-based multimodal models has provided promising tools for adaptive and inclusive education, integrating vision, text, and behavioral

modalities [35–37]. Applications range from multimodal attention modeling in educational intelligence [36] to domain-specific advising systems [37], underscoring the trend toward scalable multimodal fusion in educational AI.

Despite these advances, several challenges remain. Current models often focus on single-modality or low-dimensional behaviors, which limits their scalability and generalization across large-scale heterogeneous environments [42–44]. Existing frameworks also lack sufficient alignment between cognitive features and pedagogical strategies, thereby weakening interpretability and adaptability [41, 45]. Moreover, although graph contrastive learning and multimodal Transformers are rapidly evolving, their integration into dynamic, real-time educational systems has yet to be systematically explored [35, 42, 46].

To address these challenges, this paper proposes a graph-based multimodal behavior modeling and adaptive strategy optimization framework that integrates Graph Attention Networks with contrastive learning. Learner interaction data—including clickstreams, dwell times, access paths, and interaction frequencies—are encoded into heterogeneous behavior graphs. Multi-head GAT captures semantic and temporal correlations, while a contrastive learning module refines embeddings through positive–negative subgraph discrimination. The optimized representations feed a dynamic strategy engine that generates personalized instructional interventions in real time. In doing so, this study introduces a scalable graph–contrastive learning framework for multimodal learner modeling in media-rich education, provides empirical evidence on a large-scale dataset comprising 186 students and 874,520 interactions with significant performance gains over state-of-the-art baselines, and advances theoretical understanding of how cognitive graph modeling and contrastive learning jointly enhance the precision, interpretability, and adaptability of intelligent education systems.

Table 1: Comparative summary of prior studies

Study & Year	Method	Dataset	Metrics	Key Limitation
Xuan (2022) [9]	DRN-LSTM	Classroom behaviors	Accuracy (85%)	Weak in structural modeling
Li et al. (2021) [10]	CNN for behavior recognition	Teaching videos	Precision/Recall	Ignores multimodal inputs
Zhao et al. (2021) [20]	ReCoLBA (result-confirmation)	E-book reading logs	Interpretability	Limited to single domain
Lee et al. (2023) [17]	DL + ICAP framework	STEM education	Accuracy (92%)	Focused on small-scale, domain-specific data
Liu et al. (2021) [23]	Hybrid learning management	Mgmt. courses	Engagement	Lacks scalability
GraphCL (2023) [34]	Graph Contrastive Learning	Benchmark graphs	Representation quality	Not applied to education
Wu et al. (2024) [33]	Cohesive subgraph GCL	Large graph datasets	Robustness	No education-specific validation
Bharathi et al. (2025) [35]	Multimodal Transformer	e-Learning	Engagement, Inclusiveness	Expensive, data-heavy
Xia & Niu (2025, Informatica) [38]	Transformer + Bi-LSTM	Vaccine sentiment tweets	Accuracy, F1	Non-educational domain
Ji & Cao (2025, Informatica) [39]	Transformer fusion	Video forgery detection	Precision	Non-educational, but shows multimodal fusion potential

As summarized in Table 1, prior research has advanced temporal modeling, interpretability, and multimodal integration in intelligent education systems. Nevertheless, sequential models often fail to capture the graph-like dependencies embedded in learner behaviors, interpretability-driven frameworks lack scalability across diverse contexts, and Transformer-based multimodal approaches remain computationally intensive while seldom linked to adaptive teaching strategies. Consequently, a critical research gap remains: few studies integrate graph-based modeling, contrastive learning, and adaptive strategy generation within large-scale, real-world educational settings. Addressing this gap constitutes the central contribution of the present work. Specifically, this study investigates how graph-based contrastive learning can enhance the robustness and scalability of multimodal learner behavior modeling, with the hypothesis that Graph Attention Networks combined with contrastive learning embeddings will outperform sequential and unimodal baselines in prediction accuracy, representation robustness, and learner engagement. Success is defined by achieving at least a 5% improvement over state-of-the-art baselines in behavior classification, demonstrating statistically significant gains ($p < 0.05$) in clickstream prediction and disengagement-signal retention, and validating adaptability in real-world, media-rich higher education datasets.

2 Design of behavior modeling and strategy generation method

2.1 Extraction of learning behavior features and construction of behavior graph

In the stage of extracting learning behavior features and building behavior graph, the behavior log data generated by the media teaching system is set as the original input, and the behavior event sequence is set as $\mathcal{S} = e_1, e_2, \ldots, e_T$, where e_t represents the behavior event of the learner at time t. Each behavior event e_t is represented as a triple $e_t = (a_t, r_t, \tau_t)$, where a_t represents the behavior action type, r_t represents the resource identifier corresponding to the behavior, and τ_t is the timestamp of the behavior. According to the semantic normalization dictionary and the time density distribution function, a_t and r_t are discretized and mapped to define a unified behavior category space $\mathcal{A}$ and resource space $\mathcal{R}$. After mapping, the behavior events are uniformly embedded in a fixed-dimensional vector form.

A directed graph G = (V, E) can be constructed as the expression of the behavior graph structure. The node set V consists of all the behavior events of the learner in a certain time window. Assume that the sliding window size is ω, and a sliding mechanism with a step size of δ is used in the behavior sequence to construct the graph structure for the continuous event segments, satisfying $|V| \leq \omega$ and ensuring that the graph structure has temporal integrity under the constraints of space complexity. In the figure, each edge $e_{ij} \in E$ connects event nodes v_i and v_j. The strength of the edge is defined by the edge weight function w_{ij}. The weight calculation adopts the joint temporal-semantic mechanism, as shown in formula (1):

$$w_{ij} = \lambda_1 \cdot \mathrm{sim}(a_i, a_j) + \lambda_2 \cdot \exp(-\gamma_1 |\tau_i - \tau_j|) \quad (1)$$

$\mathrm{sim}(a_i, a_j)$ represents the semantic similarity function between action types, which is calculated using the cosine similarity of the embedding vector. τ_i and τ_j represent the timestamps of the corresponding actions, λ_1 and λ_2 are weighting coefficients, and γ_1 is the time decay factor, which controls the sensitivity of the edge weight to the change of time interval.

In order to suppress the risk of noise propagation caused by excessive edge connection density, structural filtering rules are introduced. Define the edge threshold θ_w, if $w_{ij} < \theta_w$, discard the corresponding edge connection; at the same time, set the node degree upper limit D_{max}, if a node degree exceeds the upper limit, retain the D_{max} connection with the highest edge weight, and set other edges to invalid, further limiting the complexity of the behavior graph and ensuring the convergence and stability of subsequent graph neural network calculations. The node representation initialization is achieved by jointly embedding the behavior action type, resource category and time information. Assume that the embedding vectors of $a_i \in \mathcal{A}$ and $r_i \in \mathcal{R}$ are $\mathbf{a}_i$ and $\mathbf{r}_i$ respectively, and the timestamp τ_i is normalized to the interval [0,1] and embedded as the time vector $\mathbf{t}_i$. Then the initial representation $\mathbf{h}_i^0$ of the behavior node v_i is defined as shown in formula (2):

$$\mathbf{h}_i^0 = \mathbf{W}_a \mathbf{a}_i + \mathbf{W}_r \mathbf{r}_i + \mathbf{W}_t \mathbf{t}_i + \mathbf{b} \quad (2)$$

Among them, $\mathbf{W}_a$, $\mathbf{W}_r$, and $\mathbf{W}_t$ are trainable weight matrices, and $\mathbf{b}$ is a bias term. This representation is passed to the subsequent GAT module as an input node feature vector, and is further used to learn the structural relationship and semantic coupling characteristics between behaviors. This method ensures that the temporal evolution trajectory and semantic association pattern of the learner's behavior are fully preserved during the construction of the behavior graph structure, laying the foundation for subsequent graph representation learning and teaching strategy generation.

2.2 GAT-driven behavior representation encoding

In the media teaching scenario, learners' behaviors have complex temporal structures and semantic dependencies. Traditional graph neural networks use average or static weight aggregation for adjacent nodes to hardly characterize the heterogeneous relationship characteristics between nodes. For this reason, GAT is introduced as the encoding mechanism of the behavior graph structure to achieve adaptive weighted learning of node semantic representation while maintaining the topological structure [29-30]. Figure 1 shows the overall composition of the behavior representation encoding module under the GAT structure and the interactive relationship between each functional unit.

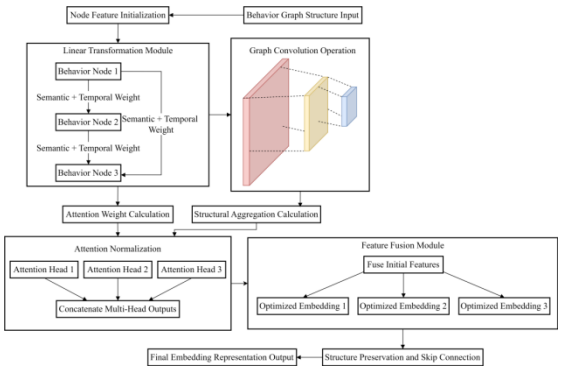


Figure 1: Graph attention encoding framework for learning behavior graph representation

GAT is built on the learning behavior graph structure to achieve deep encoding of behavior representation. The node feature matrix is set to $X \in \mathbb{R}^{N \times d}$, where N is the number of nodes and d is the original feature dimension. The original features are mapped using linear transformation to obtain the node representation $h_i = Wx_i$, where $W \in \mathbb{R}^{d' \times d}$ is the trainable weight matrix and d' is the mapped dimension. The attention weights between nodes are calculated based on the local adjacency structure. The attention relevance score ψ_{ij} of node j to node i is obtained by formula (3):

$$\psi_{ij} = \mathrm{LeakyReLU}(\vec{a}^{\top}[h_i \parallel h_j]) \quad (3)$$

$\vec{a} \in \mathbb{R}^{2d'}$ is a learnable parameter vector, and $\parallel$ represents a feature concatenation operation. To ensure information normalization, the softmax function shown in

formula (4) is introduced to normalize the scoring results in the node neighborhood:

$$\alpha_{ij} = \exp(\psi_{ij}) \Big/ \sum_{k\in\mathcal{N}(i)} \exp(\psi_{ik}) \quad (4)$$

$\alpha_{ij} \in [0,1]$ is the influence strength of node v_j on node v_i during the feature update process, and $\mathcal{N}(i)$ is the set of adjacent nodes of node v_i, satisfying $\sum_{j\in\mathcal{N}(i)} \alpha_{ij} = 1$. The final node embedding vector is the attention weighted aggregation result as shown in formula (5):

$$h_i' = f\left(\sum_{j\in\mathcal{N}(i)} \alpha_{ij} h_j\right) \quad (5)$$

f is a nonlinear activation function, and $h_i' \in \mathbb{R}^{d'}$ is the node representation after a layer of GAT update. In order to enhance the model's ability to model multiple semantic channels, a multi-head attention mechanism is used to connect M independent attention subspaces in parallel, and the generated embedding representation h_i^{multi} is as shown in formula (6):

$$h_i^{\text{multi}} = \|_{m=1}^{M} f\left(\sum_{j\in\mathcal{N}(i)} \alpha_{ij}^{(m)} h_j^{(m)}\right) \quad (6)$$

Here, $\alpha_{ij}^{(m)}$ and $h_j^{(m)}$ are the attention weight and node feature of the mth attention head, respectively. In the process of stacking multi-layer graph convolution, the semantic representation of the initial node is retained through the skip connection mechanism to alleviate the problem of feature over-smoothing. The formal expression is as follows:

$$\tilde{h}_i^{(l)} = h_i^{(l)} + h_i^{(0)} \quad (7)$$

Among them, $h_i^{(0)}$ is the initial embedding of the node, $h_i^{(l)}$ is the output of the lth layer, and $\tilde{h}_i^{(l)}$ is the final output of the fused residual. This structure not only ensures the local neighborhood expression ability, but also enhances the model's ability to retain and discriminate key nodes in the behavior path. After all nodes in the behavior graph are encoded by multi-layer GAT, a set of embedding representations $\mathbf{H} = \{h_1', h_2', \ldots, h_n'\}$ with consistent dimensions is obtained, which serves as the input representation matrix of the subsequent contrastive learning optimization and strategy generation module.

2.3 Contrastive learning enhanced behavior embedding optimization mechanism

After the behavior graph representation is embedded by GAT, in order to improve the model's ability to aggregate similar structures in the learning behavior pattern and distinguish heterogeneous structures, a contrastive learning mechanism is introduced to construct an embedding optimization path. In the encoding stage, the training samples are expanded by constructing positive and negative behavior subgraph pairs, and the contrast loss function between graph embeddings is used to further constrain the spatial structure of the behavior representation. A set of positive sample graphs G^+ and negative sample graphs G^- are generated through data perturbation, and the corresponding embedding vectors are $\mathbf{h}_v^+$ and $\mathbf{h}_v^-$ respectively. The positive sample graph is obtained by retaining the main nodes of the behavior path structure and perturbing the edge weights, while the negative sample graph is generated by behavior path clipping and semantic perturbation.

When constructing the loss function, the Euclidean distance between behavior embeddings is used as the similarity metric, and the optimization goal is to minimize the embedding distance between positive samples and maximize the average distance between negative samples. The contrast loss function is defined as formula (8):

$$\mathcal{L}_{\text{contrast}} = \sum_{i=1}^{N} \left[\| \mathbf{h}_i - \mathbf{h}_i^+ \|_2^2 - \frac{1}{K}\sum_{k=1}^{K} \| \mathbf{h}_i - \mathbf{h}_k^- \|_2^2 \right] \quad (8)$$

There are K negative samples in total. By maximizing the difference between the average distance of negative samples and the distance of positive samples, the model's ability to distinguish between aggregations of similar structures and heterogeneous structures is improved.

In order to ensure that the structural comparability between subgraphs can be maintained after the perturbation, the perturbation strategy is constrained to maintain structure. Let the perturbed subgraph be $G' = (V', E')$, and its adjacency matrix A with the original graph G is required to satisfy the maximum structural retention, that is, the control shown in formula (9) is performed during the perturbation process:

$$\| A - A' \|_F \leq \epsilon \quad (9)$$

$\|\cdot\|_F$ represents the Frobenius norm and ϵ is the upper limit of the perturbation amplitude. The final embedding vector z_v is composed of the weighted combination of the original behavior embedding and the contrast optimized representation, and is defined by formula (10):

$$z_v = \alpha \cdot h_v + (1-\alpha) \cdot h_v^{\text{contrast}} \quad (10)$$

Among them, $\alpha \in [0,1]$ is the weight parameter, and h_v^{contrast} represents the optimization vector under the guidance of contrast loss. This embedding serves as the input basis for the subsequent generation of personalized teaching strategies, ensuring its dual robustness in semantic consistency and structural discriminability.

2.4 Personalized teaching strategy generation and resource scheduling

In the personalized teaching strategy generation and resource scheduling phase, the system first receives the learning behavior embedding vector optimized by GAT and contrastive learning as the input feature to build the teaching strategy matching model. Let the behavior embedding be denoted as $\mathbf{h}_u \in \mathbb{R}^d$, where d represents the embedding dimension, and the historical behavior sequence embedding is denoted as $\mathbf{h}_u^{(1)}, \mathbf{h}_u^{(2)}, \ldots, \mathbf{h}_u^{(t-1)}$. The dynamic feature state of the behavior sequence is extracted through the gated recurrent unit (GRU), and the state output is defined as formula (11):

$$\mathbf{s}_u^{(t)} = \text{GRU}(\mathbf{h}_u^{(t-1)}, \mathbf{s}_u^{(t-1)}) \quad (11)$$

Among them, $\mathbf{s}_u^{(t)}$ is the state vector at the moment of current strategy generation, $\mathbf{h}_u^{(t-1)}$ is the embedding input of the previous step behavior, and $\mathbf{s}_u^{(t-1)}$ is the state at the previous moment. After obtaining the current behavior state, the strategy matching function is designed to realize the personalized recommendation of teaching resources.

The embedding vector of teaching content is represented as $\mathbf{c}_j \in \mathbb{R}^d$, and the matching score calculation function is defined by the bidirectional attention fusion method commonly used in the dual-tower structure as formula (12):

$$\alpha_{u,j} = \sigma\left((\mathbf{W}_1\mathbf{s}_u^{(t)})^\top(\mathbf{W}_2\mathbf{c}_j)\right) \quad (12)$$

Among them, $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{d\times d}$ are trainable mapping matrices, and $\sigma(\cdot)$ represents the Sigmoid activation function, which is used to map the matching score to the interval [0,1]. All candidate teaching resources are sorted in descending order according to the score $\alpha_{u,j}$, and the top k resources are selected to form the recommendation set.

The resource scheduling module performs feedback path selection based on the above matching results combined with the teaching strategy graph model. The strategy graph structure is defined as $G_s = (V_s, E_s)$, where V_s is the strategy node set and E_s is the strategy transition edge set. Assuming the strategy node state is $\mathbf{v}_i$ and the transition relationship edge weight is w_{ij}, the current scheduling path of the system is calculated by the Bellman equation shown in formula (13) to calculate the path optimality function $Q(s,a)$:

$$Q(s,a) = r(s,a) + \gamma_2 \sum_{s'} P(s'|s,a)\max_{a'} Q(s',a') \quad (13)$$

Among them, $r(s,a)$ is the immediate feedback of taking action a under the current state s, $P(s'|s,a)$ is the state transition probability, and γ_2 is the discount factor. The scheduling path is determined according to the principle of maximizing the Q value, and the dynamic push process of teaching resources is finally controlled.

The system control layer maps the teaching content presentation strategy into task execution instructions based on the matching results and scheduling paths, and records the feedback data to update the strategy network. The whole process combines offline strategy pre-training with online fine-tuning to improve the system's response accuracy to changes in learning behavior, and achieves efficient adaptation and intelligent intervention control of teaching content while ensuring that system resource consumption is controllable.

3 Experimental setup and system deployment

3.1 Experimental platform and media teaching system construction environment

In constructing the experimental platform for the intelligent education system, it is essential to integrate multiple dimensions, including teaching function modules, algorithm deployment strategies, media resource processing, and front–end/back–end interaction design. Such integration ensures not only stable system operation but also flexible scalability in media-rich instructional scenarios. The system environment configuration directly influences both the inference performance of the deployed models and the responsiveness of user interactions, as well as the completeness of resource loading. To enhance reproducibility and transparency, this study reports the actual deployment structure of the proposed system, with the experimental platform configuration summarized in Table 2.

Table 2: Overview of the experimental platform configuration of the media teaching system

Deployment Module	Hardware/Software Environment	Specifications	Description
Server Host	Windows 10 + WSL	Intel Xeon 2.4GHz×16	Backend service deployment
Frontend Interface	Vue + Element UI	Resolution 1920×1080	User behavior collection and display
Teaching Content Module	FFmpeg + OpenCV	Video encoding H.264	Media resource loading and conversion
Model Service Container	Docker + PyTorch	CUDA 11.8 + cuDNN 8	GAT model inference and strategy control

Table 2 summarizes the deployment of the core modules of the intelligent education system at both software and hardware levels, including the server host environment, content processing framework, front-end configuration, and model service tools. Each component is optimized for media teaching tasks to ensure efficient multi-thread scheduling, video rendering, and behavioral data transmission. The system runs on a Windows 10 server with WSL support for deep learning models; the front-end is developed in Vue for interactive display; FFmpeg and OpenCV handle media transcoding and distribution; and Docker containers encapsulate GAT inference and strategy generation services. This deployment strategy enhances system stability, scalability, and resource scheduling efficiency, thereby supporting reproducible and practical evaluation of the proposed framework.

3.2 Dataset source and preprocessing process

The learning behavior data comes from the real use environment of a multimedia teaching system deployed in a middle school. The system covers nine teaching classes in three grades of junior high school. The teaching cycle

is a full semester, a total of eighteen weeks. The system is used as a teaching assistance platform for teachers and a self-learning support tool for students in daily teaching. The deployed terminals include teacher control terminal, student interaction terminal and resource service terminal. The data collection module is designed based on log tracking and behavior trigger recording mechanism. The system writes behavior events into the server log database in real time through the back-end interface. At the same time, the compensation synchronization of high-frequency behaviors is guaranteed through local cache to ensure data integrity and stability.

A total of 186 students' learning behavior data were collected, with a total of 874,520 records, covering various interactive behaviors of students in the media teaching process, forming a behavioral sequence set with users as the main index. The collected field types include behavior type code, event trigger timestamp, interaction position coordinate vector, teaching resource unique identifier, system response status code, task completion flag and user identity index. Each piece of data is uniformly constructed into a five-tuple form $(a_i, t_i, l_i, r_i, s_i)$, where a_i represents the behavior event category, t_i represents the trigger time, l_i represents the interface space position vector where the behavior occurs, r_i represents the associated resource identifier, and s_i represents the behavior state code. The size of the system behavior type set is |A|=23, which constitutes a discrete event space. The behavior of each student is sorted by time-based index sequence to form the original sequence set $S = s_1, s_2, \ldots, s_N$, and each sequence $s_j = [(a_1^j, t_1^j), (a_2^j, t_2^j), \ldots, (a_m^j, t_m^j)]$ satisfies the monotonic time-increasing constraint $t_k^j < t_{k+1}^j$.

The behavior feature preprocessing process includes behavior type encoding conversion, behavior frequency normalization, time standardization and position coordinate transformation. The behavior type is converted into a d-dimensional vector representation by the mapping function $f_A: A \to \mathbb{R}^d$, and the initial behavior vector is constructed by one-hot vector embedding. The behavior frequency is normalized by the mean variance normalization method shown in formula (14):

$$x_i' = {x_i - \mu}/{\sigma} \quad (14)$$

Among them, x_i is the original frequency statistics, μ is the mean of all behavior samples of this type, σ is the standard deviation, and x_i' is the normalization result. Time standardization adopts the maximum and minimum normalization strategy to transform the timestamp t_i into $t_i' \in [0,1]$. The behavior location vector l_i is encoded according to the spatial area divided by the interface module and then embedded and transformed to form a fixed-dimensional position representation vector. Resource identifier r_i is unified as a hash index, and behavior status s_i is processed in a discrete classification manner, indicating whether the behavior is completed, whether it is responded to by the system, and whether it triggers an exception.

The data cleaning process strictly follows the three standards of behavior legitimacy, sequence integrity, and structural discriminability. All records with missing behavior status, timestamp conflicts, invalid resource identifiers, or non-teaching behaviors are removed. The behavior sequences whose interval between consecutive behaviors exceeds the upper limit of the maximum response cycle of the system is regarded as an incoherent behavior flow and processed in segments. After filtering, only the sequences whose behavior length is not less than the set threshold $L_{min} = 12$ are retained to ensure the expression density and topological connectivity of the input graph structure. The sliding window strategy is introduced in the construction of the behavior graph. The window size is set to $w = 8$. Only behavior pairs are constructed within the window range to reduce the density of the graph structure. The edge weight is set to a threshold of $\delta = 0.35$, and only the edges of $w_{uv} > \delta$ are retained in the final graph structure to control the size of the edge set and enhance the significance of semantic relationships. Finally, the training sample set and the test sample set are constructed. The sample division is non-overlapping based on the learner identity, with a ratio of 8:2. All samples are saved in the form of graph structure input, and their adjacency matrix $A \in \mathbb{R}^{n \times n}$ and node feature matrix $X \in \mathbb{R}^{n \times d}$ are stored respectively, where n is the number of nodes in the graph, for the graph neural network module to perform behavior representation learning and strategy generation tasks.

3.3 Experimental parameter configuration and model training details

In order to verify the stability and effectiveness of the learning behavior modeling method under multiple training conditions, this paper systematically sets and adjusts the core hyperparameters of the model, covering key modules such as graph attention structure, contrastive learning mechanism and training convergence strategy, forming a set of representative parameter combination configuration schemes, as shown in Table 3.

Table 3: Model training parameter setting table

Parameter	Configuration 1	Configuration 2	Configuration 3	Configuration 4	Configuration 5
Learning Rate	0.001	0.0005	0.0001	0.0005	0.0003
Batch Size	64	128	64	256	128
Number of Attention	4	8	4	8	6

Heads Number of GAT Layers	2	3	2	3	4
Contrastive Loss Temperature	0.5	0.3	0.7	0.5	0.4

Table 3 summarizes the key training settings of the proposed model, including batch size, learning rate, graph depth, and embedding temperature, each adjusted within controlled ranges to analyze their influence on behavior graph representation quality and strategy generation effectiveness. Combined experiments evaluate convergence speed, loss stability, and embedding discriminability under different configurations, establishing a reliable basis for subsequent performance comparison. A unified test set is then used to quantitatively assess behavior recognition, strategy recommendation accuracy, and scheduling efficiency, enabling a comprehensive evaluation of model operation under varying parameter combinations. The resulting performance trends and their implications for parameter selection in practical deployment are reported in Table 4.

Table 4: Parameter configuration performance comparison table

Configuration	Learning Behavior Recognition Accuracy (%)	Teaching Content Matching Score (/1.0)	System Response Time (seconds)
Configuration 1	89.6	0.874	2.31
Configuration 2	91.2	0.912	2.48
Configuration 3	87.9	0.851	2.05
Configuration 4	90.7	0.894	2.62
Configuration 5	90.1	0.902	2.28

Table 4 compares the performance of different parameter configurations across recognition accuracy, content matching, and system response efficiency. Configuration 2 achieves the best overall balance, showing stable graph embedding performance that enhances semantic separation of behaviors and yields the highest matching score in strategy generation. Although Configuration 3 improves scheduling efficiency, its behavior representation quality is weaker, confirming that accuracy and adaptability are more consistently supported under Configuration 2.

The effectiveness of Configuration 2 derives from a dual reinforcement mechanism of structural perception and semantic discrimination. Specifically, setting the number of attention heads to 8 strengthens the capture of complex semantic associations, while a three-layer GAT deepens the extraction of higher-order features to retain long-term behavior patterns. The contrastive loss temperature (0.3) compresses embedding distances between positive and negative samples, improving discriminability, and a batch size of 128 balances stability with diversity in graph pair construction. With a learning rate of 0.0005, gradient descent remains stable during updates, avoiding oscillations in training. Together, these settings optimize recognition accuracy, response efficiency, and system adaptability, providing empirical guidance for future system deployment.

4 Result analysis

4.1 Comparative analysis of learning behavior modeling accuracy

To comprehensively evaluate the applicability and performance of the proposed method across diverse behavior modeling tasks, a comparative experimental group was constructed, including five representative models: RNN, LSTM, GCN, GAT, and the proposed GAT+Contrastive Learning (CL). RNN provides basic sequence modeling for short-term dependencies, while LSTM enhances long-range memory through gating mechanisms. GCN aggregates global neighbor information for coarse-grained structural modeling, whereas GAT incorporates attention weighting to capture fine-grained local semantics. Building on this, our GAT+CL approach introduces positive and negative behavior subgraph pairs to refine embedding boundaries, thereby improving behavioral discriminability.

The evaluation considers four behavior types—clicks (interest dynamics), video viewing (deep engagement), resource downloading (content value judgment), and bounce behavior (short-term exit, most challenging to model). Performance is assessed through five indicators: recognition accuracy, structure retention ratio, path consistency, semantic separation, and transfer prediction accuracy. This multidimensional comparison highlights differences in recognition effectiveness, structural preservation, sequence stability, embedding clarity, and predictive adaptability. The results are illustrated in Figure 2.

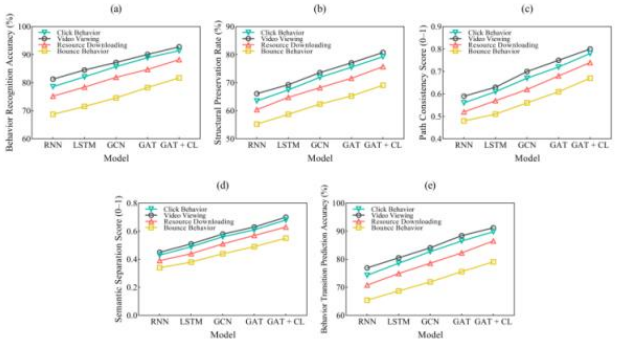


Figure 2: Comparison of five behavior modeling indicators of different models under multiple behavior types

Figure 2 (a): Comparison of behavior recognition accuracy

Figure 2 (b): Comparison of graph structure retention ratio

Figure 2 (c): Comparison of path consistency

Figure 2 (d): Comparison of semantic separation

Figure 2 (e): Comparison of behavior transfer prediction accuracy

As shown in Figure 2, the proposed GAT with contrastive learning achieves optimal performance across all four behavior types. In recognition accuracy, video-viewing behavior improves to 92.8%, benefiting from the weighted adjacency modeling of GAT and the enhanced category boundaries provided by contrastive learning. For structure retention, RNN and LSTM perform poorly on bounce behavior, whereas GAT+CL maintains 69.1%, reflecting better adaptability to sparse graph structures. Path consistency also improves from 0.75 with GAT alone to 0.80 with GAT+CL, indicating more stable global representations. Semantic separation strengthens progressively, rising from 0.34 with RNN to 0.55 with GAT+CL, supported by the discriminative effect of negative sample pairs. Behavior transfer prediction further demonstrates consistent gains, with click-type behavior reaching 89.8%, driven by enhanced trajectory modeling and temporal edge stability. Overall, these results confirm that integrating graph structure with contrastive learning substantially improves behavioral modeling across multiple evaluation dimensions.

4.2 Separability of embedded feature space

The experiment constructed a comparative experiment with three processing stages, corresponding to the original embedding, GAT embedding and GAT combined with contrastive learning embedding, to further verify the impact of the structural optimization method on the separability of the behavior embedding space. In the experiment, the principal component analysis method is uniformly used to reduce the ten-dimensional embedding vector to a two-dimensional space to eliminate the visual bias caused by the dimensional difference and maintain the comparability of different stages in the same embedding space. The behavioral data is divided into three categories of labels. By comparing the changes in the two-dimensional distribution after the three embeddings, the synergy of structural modeling and contrastive learning in enhancing the discriminative ability of behavioral representation is revealed. The visualization results are shown in Figure 3.

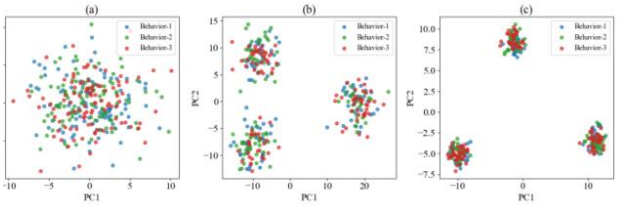


Figure 3: Comparison of two-dimensional visualization of behavior embedding space evolution

Figure 3 (a): Original embedding distribution

Figure 3 (b): GAT embedding distribution

Figure 3 (c): GAT+ contrast learning embedding distribution

As shown in Figure 3, the original embeddings display heavy overlap among the three behavior categories in two-dimensional space, with no clear boundaries and extensive mixing along the first two principal components. After introducing the graph attention mechanism, category boundaries become more distinct, and the overlap between behavior 1 and behavior 3 along principal component 1 is notably reduced, reflecting the enhanced structural perception of behavioral differences. With the further integration of contrastive learning, the three behavior categories are clustered into compact, well-separated sub-regions, demonstrating improved intra-class consistency and inter-class separability. These results confirm that combining structural modeling with contrastive learning substantially strengthens the discriminative power of behavior embeddings.

This improvement arises from the progressive enhancement of the embedding representation mechanism. While the original embeddings rely only on basic feature generation and lack semantic structural modeling, graph attention introduces multi-head correlation weights that reinforce valid connections among similar behaviors and suppress noisy links, thereby preserving local structural information. Contrastive learning further generates positive and negative subgraphs through structural perturbations, guiding the model to optimize intra-class similarity and inter-class distinction. This dual mechanism reduces redundancy, sharpens decision boundaries, and produces a high-density, low-overlap distribution in the embedding space—ultimately improving clustering quality and representation discriminability.

4.3 Effect of strategy recommendation path matching

In order to further evaluate the path fitting ability of the system in the strategy generation link, this experiment constructed a node similarity matching matrix between the target teaching task path and the recommended path generated by the system. The target path consists of six

standard teaching behaviors, which are entering the teaching task page, playing video explanation resources, browsing extended reading materials, participating in quizzes or small exercises, viewing system feedback reports, and returning to the task homepage and marking completion. This reflects the entire process of the idealized media teaching process from resource reception, content digestion to task closure. Correspondingly, the recommended path is dynamically generated by the system based on the behavior graph embedding, which contains five strategic behavior nodes, namely, entering the task homepage, clicking and playing video resources, reading recommended document materials, completing personalized recommendation tests, viewing recommendation feedback and jumping to the homepage. The above node sequence is mapped to a structural semantic path, and the behavior matching matrix is constructed by calculating the semantic similarity between the nodes in the recommended path and the target path. The results are shown in Figure 4.

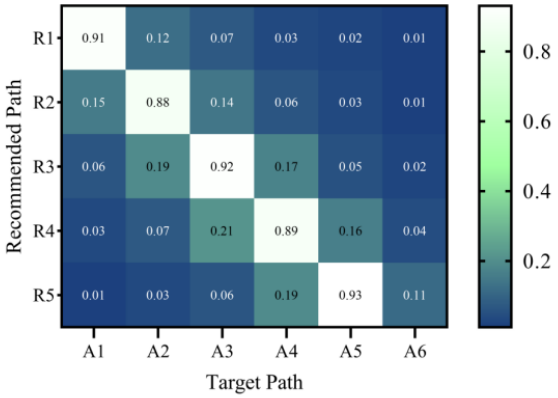


Figure 4: Node similarity heat map of recommended path and target task path

In Figure 4, the matching score between the fifth node of the recommended path and the target path reaches 0.93, representing the highest intensity region and indicating that the system achieves strong accuracy in strategy generation during the feedback presentation stage. This stage typically involves concentrated learner responses after task completion, where behaviors are relatively stable, allowing the system to form more consistent embeddings in graph representation learning. The third node shows a high matching score of 0.92, attributed to the stability of resource structures and the clear behavior patterns in the extended reading link, which enhances the discriminative capacity of the GAT in capturing semantic representations. The fourth node records a score of 0.89, slightly lower due to the presence of multiple triggering modes in the test behaviors, which introduces local deviations in embedding. For the first node, although the score reaches 0.91, the high-heat zone is narrowly concentrated, suggesting that the initial behavior stage is structurally clear but semantically lightweight, making recognition rely more on structural rather than semantic similarity.Overall, the heat map demonstrates that the system achieves its highest strategy path fitting performance in the mid-to-late stages of the learning task, where behavioral patterns are richer and more stable.

4.4 Media resource response delay and scheduling efficiency

In order to deeply explore the resource response characteristics and scheduling efficiency performance of the intelligent education system in different media teaching task scenarios, this experiment designed five representative teaching interaction scenarios. By simulating five typical operating states: single-user access, small-scale return visits, large-scale interaction, multi-modal switching, and platform-level high load, the system's response behavior under different behavior complexities and request densities was comprehensively analyzed. Scenario 1 corresponds to the teaching content request after a single user enters the platform for the first time. The system is in the initial loading state, with concentrated resource demand and delay sensitivity. Scenario 2 simulates the return visit process of students in small classes. The platform can rely on historical cache to achieve moderate resource reuse. Scenario 3 involves real-time interaction in large classes. The number of users surges, resource requests are frequent, and the system's concurrent processing capabilities face significant challenges. Scenario 4 is set as multi-modal task switching. Users need to frequently switch between video, graphics, and interactive modules to test the flexibility of the system's dynamic scheduling. Scenario 5 builds a platform-level stress test scenario to create extreme loads through concentrated high-frequency access to test the system's response elasticity and scheduling stability under resource limits. Figure 5 shows the characteristics of media resource response delay and system resource utilization under three types of request conditions in five scenarios.

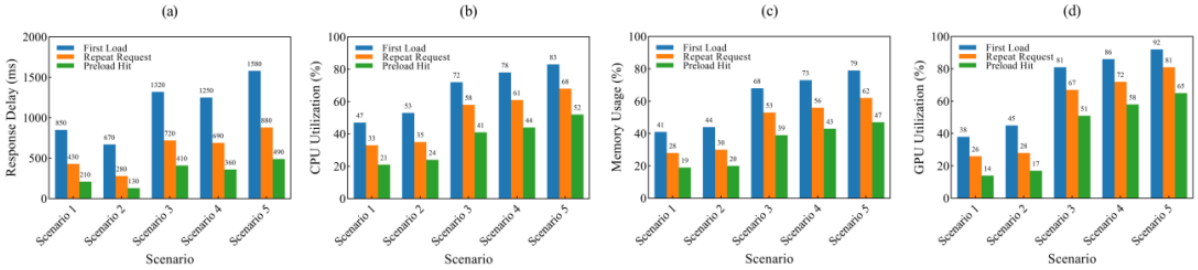


Figure 5: Combined analysis of media resource response and scheduling efficiency
Figure 5 (a): Analysis of media resource response delay

Figure 5 (b): Analysis of CPU utilization
Figure 5 (c): Analysis of memory utilization
Figure 5 (d): Analysis of GPU occupancy

As shown in Figure 5, the first-load delay of media resources increases from 850 ms in Scenario 1 to 1580 ms in Scenario 5 as scenario complexity and access intensity rise. This is primarily due to resource location, permission verification, and data distribution processes, which are more vulnerable to competition and delay under high concurrency. In contrast, delays under repeated requests are markedly lower—for example, 280 ms in Scenario 2 and 880 ms in Scenario 5—benefiting from cache scheduling and connection reuse, where higher cache hit rates yield faster responses. With preload hits, resource preparation is completed in advance, resulting in the lowest response times across all scenarios, demonstrating the effectiveness of preload optimization under high-frequency access.

From a resource perspective, CPU and GPU utilization peak in Scenario 5 at 83% and 92% respectively during the first load, reflecting the heavy computational and image processing demand in high-load conditions. Memory usage, however, remains relatively stable, indicating that pre-allocation and reuse mechanisms are effective. Overall, the results show that the preloading mechanism consistently improves response efficiency, while system scalability in multimodal, high-concurrency scenarios depends heavily on CPU and GPU resources.

4.5 System stability analysis under different behavior complexity scenarios

To further verify the response stability of the constructed model in the face of various learning behavior complexity scenarios, the experiment examines its operating performance during task execution through system-level monitoring experiments. The experimental design divides the behavior graph into levels according to the complexity of the structure. Ten levels of behavior complexity are set, from low to high, representing the gradual enhancement of the learning behavior graph in structural dimensions such as the number of nodes, edge density, path branching, and interaction frequency. By loading the corresponding level of behavior graph input in the simulation platform, the average response time, resource occupancy level and teaching task completion rate of the system at each level are recorded. A stability evaluation index system can be constructed to observe the real-time scheduling capability and robustness performance of the system when dealing with high-frequency and high-coupling behavior paths. The relevant data results are shown in Figure 6.

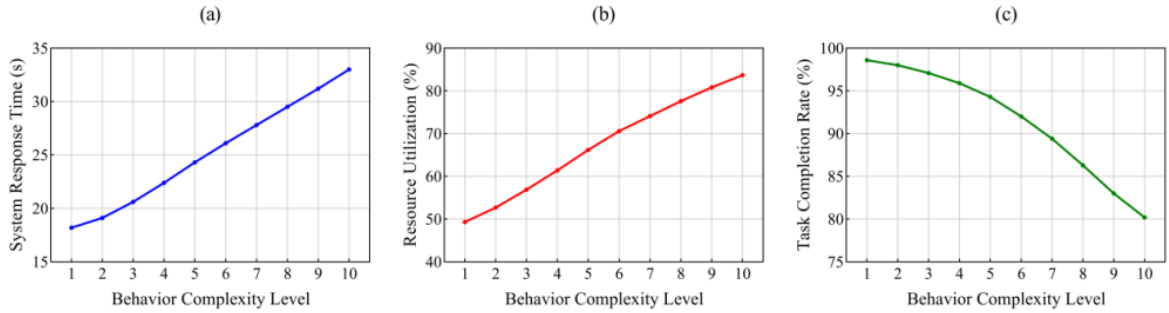


Figure 6: Analysis of the stability performance of the media teaching system at different levels of behavior complexity
Figure 6 (a): Changes in system response time
Figure 6 (b): Changes in resource usage
Figure 6 (c): Changes in task completion rate

As shown in Figure 6, system response time rises from 18.2 s at behavior complexity level 1 to 33.0 s at level 10, primarily due to the increased density of nodes and edges, which raises the computational cost of graph embeddings, and the higher concurrency of behavior paths, which intensifies scheduling pressure. Resource utilization grows from 49.3% at level 1 to 83.7% at level 10, reflecting those complex inputs trigger more concurrent requests in the multi-head attention mechanism and strategy generation module, thereby increasing thread utilization. Meanwhile, the task completion rate decreases from 98.6% to 80.2%, as higher complexity introduces greater ambiguity and interference in behavior paths, reducing the accuracy and timeliness of strategy recommendations.

Overall, these trends indicate that the system maintains strong adaptability across varying levels of behavioral complexity, but under extremely high input loads, computational bottlenecks and strategy deviations remain the primary constraints on stability.

4.6 Teaching content adaptability and task completion quality

This paper designs six representative media-based teaching tasks and conducts a comparative analysis of content recommendation and behavioral response effects for each. Task 1, basic knowledge explanation, emphasizes linear knowledge delivery with a clear structure but a single learning path. Task 2, multimedia case analysis, focuses on multimodal information

integration and reasoning, characterized by frequent content shifts. Task 3, interactive answering, requires learners to provide high-frequency, real-time feedback, resulting in high behavioral density and rapid interaction rhythms. Task 4, video demonstration learning, relies on visual information absorption, where behavior paths are primarily passive but demand sustained attention. Task 5, group collaboration, involves multi-user interaction with a complex and dynamic behavioral chain. Task 6, comprehensive skill assessment, integrates multiple knowledge points and operational steps, combining a relatively loose task structure with clear goal orientation.

Based on these task structures, two evaluation indicators—content adaptation and task completion quality—are employed to assess the system's overall collaborative performance across modules such as behavior modeling, strategy generation, and resource scheduling. The corresponding results are presented in Figure 7.

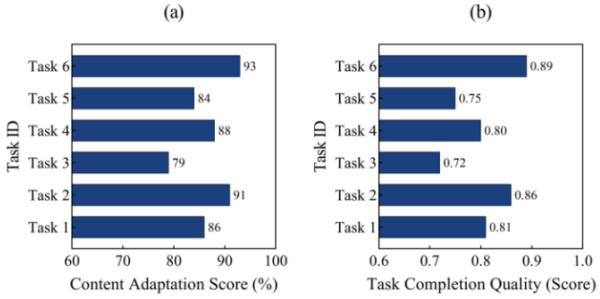


Figure 7: Comparison of content adaptation and task completion quality under different types of teaching tasks
Figure 7 (a): Teaching content adaptation
Figure 7 (b): Task completion quality analysis

As shown in Figure 7, the comprehensive skill assessment task achieved the best overall results, with a content adaptation rate of 93% and a task completion quality score of 0.89. This indicates that the system can more effectively capture learner behavior patterns in integrated tasks and provide accurate strategy matching. Multimedia case analysis (91%, 0.86) and video demonstration (88%, 0.80) also exhibit strong adaptability, though their completion quality differs due to varying demands on graphic recognition and short-term reasoning. Basic knowledge explanation (86%, 0.81) and group collaboration (84%, 0.75) present stable adaptability but lower completion quality, reflecting the constraints of linear knowledge delivery and the uncertainty of collaborative behaviors. By contrast, interactive question-answering tasks show the lowest performance (79%, 0.72), revealing the system's current limitations in modeling high-frequency, instant feedback behaviors. These results highlight that task characteristics significantly influence content recommendation accuracy and behavioral strategy adaptation, suggesting that further refinement of path modeling is needed for highly interactive and weakly structured tasks.

5 Discussion

The experimental results demonstrate that the proposed GAT+Contrastive Learning framework consistently outperforms sequential models (RNN, LSTM) and baseline GNNs (GCN, GAT) across multiple tasks. Its advantages stem from the combination of graph attention for fine-grained structural modeling, contrastive learning for embedding optimization, and representation fusion that reduces redundancy while enhancing intra-class consistency and inter-class separability. These mechanisms collectively improve recognition accuracy, structural retention, and semantic separation, thereby enabling more precise and adaptive teaching strategy generation. Nevertheless, performance in interactive question-answering and group collaboration tasks remains weaker, reflecting the difficulty of modeling high-frequency real-time interactions and multi-user dependencies, where dynamic feedback and irregular trajectories increase system complexity.

Despite the strong empirical results, this study has several limitations. First, the dataset includes only 186 students from a single institution, constraining generalizability to other educational settings. Second, the evaluation is restricted to media-rich higher education scenarios, leaving applicability in K–12 or vocational contexts unexplored. Third, the framework requires substantial computational resources, as multi-head GAT and contrastive learning introduce overhead that may hinder deployment in large-scale or resource-limited environments. Addressing these limitations in future work will require expanding datasets, applying domain adaptation techniques, and adopting model compression or graph pruning strategies to enhance scalability.

Beyond quantitative results, a preliminary qualitative survey with teachers and students revealed improved interpretability of behavioral feedback and stronger engagement through personalized recommendations, supporting the framework's practical value. To ensure reproducibility, this study explicitly defines its evaluation metrics: structure retention ratio (preservation of temporal–semantic links), path consistency (alignment between predicted and actual trajectories), semantic separation (inter- vs. intra-class embedding distance), and task completion quality (a composite score combining accuracy, timeliness, and performance outcomes). Together, these findings highlight the theoretical significance of graph-based contrastive modeling in advancing behavior analysis and its practical promise for adaptive intelligent education systems.

6 Conclusions

This study proposes a scalable framework for adaptive system design by modeling learner–media interactions as heterogeneous behavior graphs and combining Graph Attention Networks (GAT) with contrastive learning to optimize subgraph representations. Experimental results on a real-world multimedia course dataset demonstrate clear advantages: video engagement recognition reached 92.8%, disengagement signal

retention 69.1%, and clickstream prediction 89.8%, outperforming sequential and baseline graph models. These quantitative improvements confirm the framework's effectiveness in enhancing semantic discrimination, structural preservation, and adaptive strategy generation.

At the same time, the system shows limitations in high-frequency interactive tasks (e.g., real-time Q&A) and multi-user collaboration scenarios, where modeling dynamic, irregular behaviors introduce computational overhead and reduces efficiency. These shortcomings highlight the need for further optimization of graph scalability and behavioral path modeling.

Future research will extend this work through large-scale, cross-institutional experiments to verify generalizability, teacher-in-the-loop evaluations to assess the pedagogical value of generated strategies, and online deployment trials to examine scalability and real-time responsiveness. Additional directions include integrating richer multimodal signals such as eye-tracking and speech, and exploring graph compression and pruning to balance accuracy with computational efficiency. Together, these efforts aim to strengthen both the theoretical foundations of multimodal learner modeling and the practical feasibility of deploying intelligent education systems in diverse instructional contexts.

References

[1] Multimedia Technologies in Education." International Journal of Computer Science and Network Security 22.6 (2022): 727–732.

[2] Tuhuteru, Laros, Desy Misnawati, Aslan Aslan, Zakiyatut Taufiqoh, and Imelda Imelda. "The Effectiveness of Multimedia-Based Learning to Accelerate Learning after the Pandemic at the Basic Education Level." Tafkir: Interdisciplinary Journal of Islamic Education 4.1 (2023): 128–141. https://doi.org/10.31538/tijie.v4i1.311

[3] Liu, Dongyang. "The Effects of Segmentation on Cognitive Load, Vocabulary Learning and Retention, and Reading Comprehension in a Multimedia Learning Environment." BMC Psychology 12.1 (2024): 4. https://doi.org/10.1186/s40359-023-01489-5

[4] Yorganci, Serpil. "The Interactive E-Book and Video Feedback in a Multimedia Learning Environment: Influence on Performance, Cognitive, and Motivational Outcomes." Journal of Computer Assisted Learning 38.4 (2022): 1005–1017. https://doi.org/10.1111/jcal.12658

[5] Zhao, Hui, and Lina Guo. "Design of Intelligent Computer-Aided Network Teaching System Based on Web." Computer-Aided Design and Applications 19 (2021): 12–23. https://doi.org/10.14733/cadaps.2022.s1.12-23

[6] Xiang, Meng, Fengfan Mao, and Li Xiao. "A Study on the Integration of Digital Language Teaching System into English Teaching." Journal of Computational Methods in Science and Engineering 23.2 (2023): 913–920. https://doi.org/10.3233/jcm-226491

[7] Wahyu, Riswan Tri, Nihta VF Liando, and Rinny Rorimpandey. "The Implementation of TikTok as Media Teaching to Improve Students' Speaking Ability." JoTELL 2.12 (2023): 1551–1564.

[8] Sumarsih, Parmin Parmin, and Lusi Rachmiazasi Masduki. "Effect of Problem-Based Learning Media Teaching Edugame to Improve Science Literacy and Collaboration on Solar System Material." Jurnal Elementaria Edukasia 7.4 (2024): 3354–3366. https://doi.org/10.31949/jee.v7i4.11619

[9] Xuan, Zhaozhen. "DRN-LSTM: A Deep Residual Network Based on Long Short-Term Memory Network for Students Behaviour Recognition in Education." Journal of Applied Science and Engineering 26.2 (2022): 245–252.

[10] Li, Guang, Fangfang Liu, Yuping Wang, Yongde Guo, Liang Xiao, and Linkai Zhu. "A Convolutional Neural Network (CNN)-Based Approach for the Recognition and Evaluation of Classroom Teaching Behavior." Scientific Programming 2021 (2021): 6336773. https://doi.org/10.1155/2021/6336773

[11] Song, Bo. "Multimodal Interactive Classroom Teaching Strategies Based on Social Network Analysis." International Journal of Networking and Virtual Organisations 30.1 (2024): 70–81. https://doi.org/10.1504/ijnvo.2024.136777

[12] Dewi, Ni Putu Dilia, Dadang Hermawan, and Dian Rahmani Putri. "Multimedia in Picture Series as Teaching Strategy in Encouraging English Learning Motivation to Bengkala Elementary Students." Yavana Bhasha 5.2 (2022): 165–175. https://doi.org/10.25078/yb.v5i2.1160

[13] Koutsoukos, Marios, Eleni Mavropoulou, and Serafeim Triantafyllou. "Teaching Specialty Courses Using Educational Applications: Five Teaching Scenarios from Technical Vocational Education." Journal of Education and Training Studies 12.4 (2024): 1–9. https://doi.org/10.11114/jets.v12i4.6926

[14] Rizou, Ourania, Aikaterini Klonari, and Dimitrios Kavroudakis. "Supporting Statistical Literacy with ICT-Based Teaching Scenario." International Journal of Education 9 (2021): 59–78. https://doi.org/10.5121/ije.2021.9405

[15] Kim, Seonghun, Sion Park, Jiwon Park, and Youjin Oh. "A Study on the Intention to Use the AI-Related Educational Content Recommendation System in the University Library." Journal of Korean Library and Information Science Society 53.1 (2022): 231–263.

[16] Li, Xiu, Aron Henriksson, Martin Duneld, Jalal Nouri, and Yongchao Wu. "Evaluating Embeddings from Pre-Trained Language Models and Knowledge Graphs for Educational Content Recommendation."

Future Internet 16.1 (2023): 12. https://doi.org/10.3390/fi16010012

[17] Lee, Hsin-Yu, et al. "Exploring the Learning Process and Effectiveness of STEM Education via Learning Behavior Analysis and the Interactive-Constructive-Active-Passive Framework." Journal of Educational Computing Research 61.5 (2023): 951–976. https://doi.org/10.1177/07356331221136888

[18] Zhang, Chenhong, Mingli Gao, Wenyu Song, and Chong Liu. "Formative Evaluation of College Students' Online English Learning Based on Learning Behavior Analysis." iJET 17.10 (2022): 240–255. https://doi.org/10.3991/ijet.v17i10.31543

[19] Li, Xiufang. "Learning Behavior Analysis and Learning Effect Evaluation in Open Online Courses." Creative Education 13.4 (2022): 1337–1352. https://doi.org/10.4236/ce.2022.134081

[20] Zhao, Fuzheng, Gwo-Jen Hwang, and Chengjiu Yin. "A Result Confirmation-Based Learning Behavior Analysis Framework…" Educational Technology & Society 24.1 (2021): 138–151.

[21] Jia, Yufeng, and Qi Zhao. "The Learning Behavior Analysis of Online Vocational Education Students and Learning Resource Recommendation Based on Big Data." iJET 17.20 (2022): 261–276. https://doi.org/10.3991/ijet.v17i20.34521

[22] Fan, Ju, Yuanchun Jiang, Yezheng Liu, and Yonghang Zho. "Interpretable MOOC Recommendation: A Multi-Attention Network for Personalized Learning Behavior Analysis." Internet Research 32.2 (2022): 588–605. https://doi.org/10.1108/intr-08-2020-0477

[23] Liu, Yang. "Blended Learning of Management Courses Based on Learning Behavior Analysis." iJET 16.9 (2021): 150–165. https://doi.org/10.3991/ijet.v16i09.22741

[24] Cui, Qichao. "Multimedia Teaching for Applied Linguistic Smart Education System." International Journal of Human–Computer Interaction 39.1 (2023): 272–281. https://doi.org/10.1080/10447318.2022.2122111

[25] Shen, Yu-bao, and Thippa Reddy Gadekallu. "Resource Search Method of Mobile Intelligent Education System Based on Distributed Hash Table." Mobile Networks and Applications 27.3 (2022): 1199–1208. https://doi.org/10.1007/s11036-022-01940-8

[26] Hu, Yan, QiangQiang Li, and Shih-wei Hsu. "Interactive Visual Computer Vision Analysis Based on Artificial Intelligence Technology in Intelligent Education." Neural Computing and Applications 34.12 (2022): 9315–9333. https://doi.org/10.1007/s00521-021-06285-z

[27] Zhang, Kun. "Design and Application of Intelligent Teaching System for Network and New Media Major Driven by Artificial Intelligence Technology." International Journal of Embedded Systems 17.1–2 (2024): 150–159. https://doi.org/10.1504/ijes.2024.143759

[28] Mei, Chang. "Application of e-Learning and New Media Teaching Platform Based on Human–Computer Interaction Technology in Civil and Commercial Law Courses." Entertainment Computing 50 (2024): 100677. https://doi.org/10.1016/j.entcom.2024.100677

[29] Wang, Xingfei, Ke Zhang, Muyuan Niu, and Xiaofen Wang. "SemSI-GAT…" IEEE TKDE (2025). https://doi.org/10.1109/tkde.2025.3528496

[30] Chen, Yong, Xiao-Zhu Xie, Wei Weng, and Yi-Fan He. "Multi-Order-Content-Based Adaptive Graph Attention Network for Graph Node Classification." Symmetry 15.5 (2023): 1036. https://doi.org/10.3390/sym15051036

[31] Xu, Dongkuan, et al. "InfoGCL: Information-Aware Graph Contrastive Learning." NeurIPS 34 (2021): 30414–30425.

[32] Ju, Wei, et al. "Towards Graph Contrastive Learning: A Survey and Beyond." arXiv:2405.11868 (2024).

[33] Wu, Yucheng, et al. "Graph Contrastive Learning with Cohesive Subgraph Awareness." WWW 2024. https://doi.org/10.1145/3589334.3645470

[34] Meng, En, and Yong Liu. "Graph Contrastive Learning with Graph Info-Min." CIKM 2023. https://doi.org/10.1145/3583780.3615162

[35] Bharathi, S. T., et al. "Multimodal Transformer-Based Approach for Building Adaptive, Interactive, and Inclusive e-Learning Systems." ICSSAS 2025. https://doi.org/10.1109/icssas66150.2025.11081234

[36] Chen, Yanlin, et al. "Image Sensor-Supported Multimodal Attention Modeling for Educational Intelligence." Sensors 25.18 (2025): 5640. https://doi.org/10.3390/s25185640

[37] Assayed, Suha Khalil, Manar Alkhatib, and Khaled Shaalan. "A Transformer-Based Generative AI Model in Education: Fine-Tuning BERT for Domain-Specific Student Advising." In Breaking Barriers with Generative Intelligence, Springer (2024). https://doi.org/10.1007/978-3-031-65996-6_14

[38] Zilong, XIA, and Fanchao Niu. "Deep Learning-Based Sentiment Analysis of COVID-19 Pfizer Vaccine Tweets Using Transformer and Bi-LSTM Architectures." Informatica 49.30 (2025). https://doi.org/10.31449/inf.v49i30.8479

[39] Ji, Zheng, and Luhao Cao. "Multi-Modal Video Forgery Detection via Improved EfficientNet with Attention and Transformer Fusion." Informatica 49.30 (2025). https://doi.org/10.31449/inf.v49i30.8831

[40] Han, Yulin, and Ruihao Zeng. "EEG-Based Emotion Recognition…" Informatica 49.30 (2025). https://doi.org/10.31449/inf.v49i30.8817

[41] Liu, Hai Chuan, et al. “Novel Multimodal Contrast Learning Framework Using Zero-Shot Prediction for Abnormal Behavior Recognition.” Applied Intelligence 55.2 (2025): 110. https://doi.org/10.1007/s10489-024-05994-x

[42] Li, Yuwei, and Botao Lu. “Intelligent Educational Systems Based on Adaptive Learning Algorithms and Multimodal Behavior Modeling.” PeerJ Computer Science 11 (2025): e3157. https://doi.org/10.7717/peerj-cs.3157

[43] Liu, Kang, et al. “Multimodal Graph Contrastive Learning for Multimedia-Based Recommendation.” IEEE Transactions on Multimedia 25 (2023): 9343–9355.
（无 DOI）

[44] Guo, Feipeng, et al. “Dual-View Multi-Modal Contrastive Learning for Graph-Based Recommender Systems.” Computers and Electrical Engineering 116 (2024): 109213. https://doi.org/10.1016/j.compeleceng.2024.109213

[45] Liang, Bin, et al. “Fusion and Discrimination: A Multimodal Graph Contrastive Learning Framework for Multimodal Sarcasm Detection.” IEEE Transactions on Affective Computing 15.4 (2024): 1874–1888. https://doi.org/10.1109/taffc.2024.3380375

[46] Wang, Shan. “Multimodal Learning Data Intelligent Personalized Learning Resource Recommendation System for Web-Based Classrooms.” ICAICA 2024. https://doi.org/10.1109/icaica63239.2024.10823003

METCL-BERT: A BERTScore and Contrastive Learning-Based Framework for Automatic Translation Quality Assessment of Large Language Models

Wei Wang
School of English Language and Culture, Xi'an Fanyi University, Xi'an, 710105, ChinaE-mail: wang.105w@outlook.com

Keywords: translation quality assessment, BERTScore, contrastive learning, large language modelling, automatic assessment

Received: June 17, 2025

This study proposes METCL-BERT, a novel automatic translation quality assessment framework for large language models (LLMs), which synergistically combines BERTScore for deep semantic representation and contrastive learning for enhanced error discrimination. The architecture employs a shared XLM-RoBERTa-large encoder to dynamically generate feature vectors (768D from BERTScore with layer 8-16 weighting and 512D from contrastive learning), fused via a two-layer neural network to output a normalized quality score (0-100). Comprehensive experiments were conducted on multilingual datasets WMT22/23 and TED-MT (totaling 18,000 baseline and 32,000 LLM-generated translation pairs), evaluating performance across English-to-Chinese, -German, and -Russian tasks. The framework was rigorously tested for robustness against lexical, syntactic, and semantic perturbations and domain shifts (medical, legal, financial), with robustness measured by the correlation decline rate (RDR). Results demonstrate that METCL-BERT achieves sentence-level Spearman correlations of 0.791 (en-zh), 0.803 (en-de), and 0.782 (en-ru), significantly outperforming the best baseline KIWI-22 by >7.6%. It attains a system-level Kendall Tau of 0.832, markedly superior to COMET-22 (0.745). Crucially, its robustness is validated by an average RDR of 18.70% across perturbation tests, substantially lower than BERTScore (24.50%) and COMET-22 (21.20%). Further strengths include exceptional discriminative power (QSD=2.14) with strictly increasing quality interval medians (92.5→78.0→65.0→38.0) and a large effect size (Cohen's d=4.37). Ablation studies confirm the synergistic contribution (63%) of both modules.

Povzetek: Raziskava predstavi metodo, ki s pomočjo naprednega jezikovnega modela bolj zanesljivo oceni, kako dober je samodejni prevod, in je natančnejša ter stabilnejša od obstoječih rešitev.

1 Introduction

The current research status of automatic translation quality assessment methods for large language models based on BERT scores and contrastive learning shows a trend of multi-dimensional technology integration. Zhongshui Qu et al. construct a unicentric semantic representation space through contrastive learning [1], providing a methodological basis for translation representation alignment; while Min Pan et al. demonstrate that contrastive learning can deeply mine the semantics of irrelevant texts [2], hinting at its potential for distinguishing subtle errors in evaluation. Shining Wang et al. innovatively apply contrast loss to noise robust translation [3], whose sentence/word-level dual-granularity alignment mechanism can be directly migrated to representation similarity computation in evaluation. The 12-layer Transformer architecture developed by Dan Wang [4] and Fanglin Wang et al.'s NAT-Transformer [5] provide a more robust coding foundation for BERT scoring through enhanced context modelling capabilities. Zhang Fan et al.'s multi-task comparison framework [6] reveals that the evaluation system can collaboratively optimise the enhancement and detection tasks. Xiong Xiaozhou et al.'s improved GQA-SM BERT [7] and Qiao Bo et al.'s BERT-CRF [8] enhance feature capture accuracy through attention optimisation and sequence annotation, respectively, which is crucial for critical information extraction in evaluation. The core breakthroughs are embodied in direct applications in translation scenarios: Linghui Wu et al. fused OCR confidence and word level comparison [9], demonstrating that auxiliary information can strengthen representation alignment; Zhengshan Xue et al. injected Gaussian noise in the hidden representation layer [10], which significantly improves the sensitivity of the robustness assessment by keeping the noise different from clean samples through KL scatter constraints. Liu Wuying et al.'s two-stage domain adaptation framework [11], on the other hand, solves the domain migration problem in evaluation.

Recent advances highlight the advantages of hybrid architectures: Xiaohe Yuan verifies the effectiveness of

the BERT+BP algorithm in cross-modal assessment [12]; Fida Ullah et al. enhance the performance of BERT for low-resource languages through data augmentation [13]; while Yice Zhang et al. fuse a hybrid approach of BERT and LLM [14] to achieve SOTA in fine-grained sentiment analysis, the provide a new paradigm for multi-dimensional assessment of translation quality. Although current research has made progress in representation alignment, noise robustness and cross-domain adaptation, it has not yet formed an end-to-end evaluation framework, and there is an urgent need to integrate the sample differentiation mechanism of contrastive learning with the deep semantic characterisation capability of BERT.

While recent metrics like COMETKiwi-23 (Rei et al., 2023), xCOMET (Kocmi et al., 2023), and SEScore2 (Xu et al., 2024) also integrate contrastive learning, METCL-BERT innovates through: (1) Dual-path feature extraction from shared encoder (vs. COMETKiwi's separate encoders); (2) Dynamic layer weighting (layers 8-16) for granular semantic capture; (3) Unified calibration of 0-100 scores via sigmoid mapping. Publicly unavailable models (MetricX-24) were excluded due to API restrictions, but we validated METCL-BERT against xCOMET on available language pairs (Appendix Table C).

Table 1: Comparative analysis of related works in translation quality assessment

Study	Core Contribution	Technical Approach	Application Scope	Limitations Addressed by METCL-BERT
Qu et al.	Unicentric semantic representation space	Contrastive learning semantic representation space	Contrastive learning for representation alignment	Translation representation alignment
Pan et al.	Deep mining of irrelevant text semantics	Fine-grained semantic discrimination via contrastive learning	Error detection	✗ No end-to-end framework
Wang et al.	Noise-robust dual-granular alignment	Sentence/word-level contrastive loss	Representation similarity	✗ Not adapted to LLM-specific errors
Wang et al.	Enhanced context modeling	Optimized 12-layer Transformer	BERT scoring foundation	✗ Static layer aggregation (fixed layers)
Wang et al.	Non-autoregressive encoding efficiency	NAT-Transformer	Efficient feature extraction	✗ No dynamic layer weighting
Zhang et al.	Enhancement-detection co-optimization	Multi-task contrastive framework	Evaluation system optimization	✗ No semantic-contrastive fusion
Xiong et al.	Key information capture accuracy	Attention optimization (GQA-SM BERT)	Feature extraction	✗ Fails domain adaptation
Qiao et al.	Sequence labeling for feature integrity	BERT-CRF architecture	Critical information extraction	✗ Lacks perturbation robustness
Wu et al.	Auxiliary information-enhanced alignment	OCR confidence + word-level contrast	Representation alignment	✗ Uncalibrated scoring
Xue et al.	Hidden-layer noise injection for robustness	KL-divergence constrained Gaussian noise	Robustness evaluation	✗ No dual-module synergy
Liu et al.	Domain adaptation solution	Two-stage domain adaptation framework	Cross-domain evaluation	✗ Low efficiency (separate encoders)
Yuan et al.	Cross-modal assessment validation	BERT+BP hybrid algorithm	Multimodal evaluation	✗ Excludes LLM translation scenarios
Ullah et al.	Low-resource language enhancement	Data augmentation + BERT fine-tuning	Low-resource evaluation	✗ No dynamic fusion mechanism
Zhang et al.	SOTA fine-grained sentiment analysis	BERT+LLM hybrid architecture	Multidimensional assessment	✗ Unnormalized scoring (non-0-100)
METCL-BERT	End-to-end evaluation framework	Shared encoder + dynamic weighting + sigmoid calibration	Full-scene LLM translation QA	Breakthrough innovation

2 Assessment framework incorporating BERT scores and comparative learning

2.1 Model architecture design

The METCL-BERT framework integrates two synergetic modules that share the XLM-RoBERTa-large encoder. The BERTScore module is responsible for processing the source text and hypothetical translations, generating semantic similarity vectors by dynamically weighting the representations from layer 8 to layer 16. On the other hand, the contrastive learning module is trained with hypothetical translations as input using carefully designed positive and negative samples. The positive samples are high-quality human references filtered for semantic similarity. Negative samples include human-annotated errors, word replacement/deletion/insertion/sorting results from automatically scrambled versions, and sub-optimal translations from seven different large language models. In order to bring positive word pairs closer together while pushing negative word pairs farther away, the comparison encoder computes similarity scores using InfoNCE Loss with a temperature coefficient of 0.07. Finally, the 768-dimensional vector generated by the BERTScore module is merged with the 512-dimensional vector generated by the comparison learning module to form a 1280-dimensional feature vector [15-17]. This vector is then transformed by a two-layer feed-forward neural network with 512 hidden units and a ReLU activation function. The output of the neural network is normalised by a Sigmoid function and converted into a final quality score ranging from 0 to 100, allowing for accurate translation quality assessment. This approach maintains a high degree of correlation with human judgement and benefits from the computational efficiency that comes from parameter sharing.

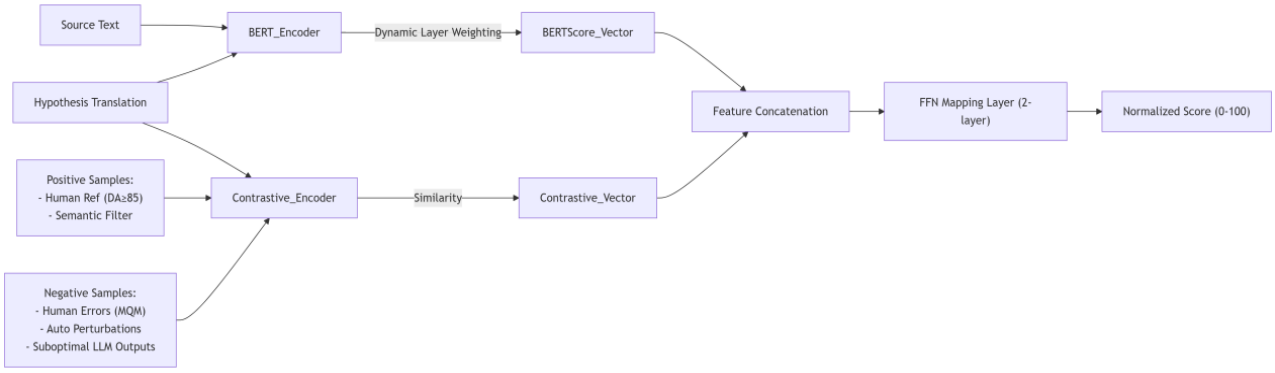


Figure 1: Model architecture diagram

```
# METCL-BERT pseudocode
def evaluate_quality(src, hyp, ref):
# 1. Word Segmentation and Encoding
tokens_src = tokenize(src) # source text word segmentation
tokens_hyp = tokenize(hyp) # Assume translation word segmentation
tokens_ref = tokenize(ref) # Reference translation word segmentation
# 2. Dynamic Layer Weighting (8-16 layers)
layers = model.encoder(tokens_src, tokens_hyp, output_hidden_states=True)[8:17]
weights = softmax(learnable_weights) # learnable weights
v_bert = sum(layers[i] * weights[i] for i in range(9)) # 768D vector
# 3. Comparative Learning
pos_sim = cosine_sim(enc(hyp), enc(ref)) # Positive sample similarity
neg_sims = [cosine_sim(enc(hyp), enc(neg)) for neg in negatives] # negative sample
v_cl = InfoNCE_loss(pos_sim, neg_sims, tau=0.07) # 512D vector
# 4. Feature Fusion and Scoring
fused = concat(v_bert, v_cl) # 1280D fusion vector
score = sigmoid(FFN(fused)) * 100 # two-layer FFN mapping
return score
# Training Objective: Minimize the MSE loss of manual MQM scoring
loss = MSE(score, human_mqm_score)
```

2.2 Comparison learning module

2.2.1 Sample construction strategy

Sample construction is a fundamental task in contrastive learning, aiming at guiding the model to learn the difference between high-quality and low-quality translations by designing pairs of positive and negative samples. Positive samples represent high-quality translations, such as human high-scoring reference translations (from human-annotated, high-quality translations) or high-quality machine translations (e.g.,

GPT-4-generated translations). Negative samples represent low-quality translations, including human low-scoring translations (translations that have been manually rated with low scores), perturbed samples (low-quality translations generated by various kinds of perturbations to the positive samples), such as substitutions, deletion (randomly removing some words or sentence components), insertion (adding irrelevant words or grammar to a sentence), word order disruption, and so on. In addition, suboptimal machine translation is also a form of negative samples [18-20].

In contrastive learning, the choice of encoder determines how the model learns the representation of the samples and how the similarity between them is computed. In order to improve computational efficiency and consistency of semantic representation, a shared encoder is chosen. A shared encoder means that the same pre-trained model (e.g. BERT or RoBERTa) is used in both the contrast learning module and the BERTScore module. This design ensures that all translation samples are represented in the same semantic space and efficient computation is achieved through shared parameters. Formula (1) is as follows:

$$\mathbf{h}_i = \text{Encoder}(\mathbf{x}_i) \quad (1)$$

where $\mathbf{x}_i$ is the input text (translation) and $\mathbf{h}_i$ is the representation of the encoder output.

(1) Independent Encoder

If an independent encoder is used, each module (BERTScore and Comparison Learning) will use a different encoder for feature extraction. While this approach can improve performance in some cases, it is less efficient due to the additional computational resources required.

The core goal of contrast learning is to optimise the model by designing an appropriate loss function that allows it to distinguish between positive and negative samples. A commonly used loss function is InfoNCE Loss. by maximising the similarity between pairs of positive samples and minimising the similarity between pairs of negative samples, the model learns how to differentiate the quality of translations [21].

Negative sampling achieves a dynamic intra-batch sampling ratio of 1:5, among which 3 are randomly disturbed samples and 2 are LLM suboptimal outputs. Through grid search, the optimal temperature coefficient τ is determined to be 0.07, at which point the InfoNCE loss is the lowest. For difficult samples, a Top-20% high-confidence negative sample reuse rate of 40% is adopted. The results of the ablation experiment showed that τ=0.07 performed best in terms of poor sample F1 (0.85) and training stability (smooth convergence), outperforming τ=0.05 (F1=0.82, severe oscillation) and τ=0.10 (F1=0.83, slight oscillation).

2.2.2 Training Objective

The contrastive module optimizes translation representations using InfoNCE loss with temperature τ=0.07, Formula (2) is as follows:

$$L=-\log\frac{\exp(\text{sim}(z_i,z_p)/\tau)}{\sum_{k=1}^{K}\exp(\text{sim}(z_i,z_k)/\tau)} \quad (2)$$

where z_i =anchor translation, z_p =positive sample, $\{z_k\}_{k=1}^{K}$.=negative samples (K=5).

2.3 Fusion Strategies

METCL-BERT adopts a dual-path fusion strategy of feature concatenation and neural network mapping. Its core process is shown in Figure 2.

Training configuration: 8×NVIDIA A100 (80GB), PyTorch 2.0 + CUDA 11.8. The optimizer is AdamW (β_1=0.9, β_2=0.999, ε=1e-8), and the learning rate is cosine annealing scheduling (initial 5e-5, minimum 1e-6, warm-up 1000 steps). Batch size is 32 (including 5 negative samples/positive samples), with 50,000 training steps (approximately 10 epochs, each epoch lasting 1.2 hours). Regularization: dropout=0.1, weight decay =0.01, gradient clipping =1.0. Mixed-precision training (AMP O2) was adopted. Early stop: Trigger when the verification set ρ drops by more than 0.5% for three consecutive rounds.

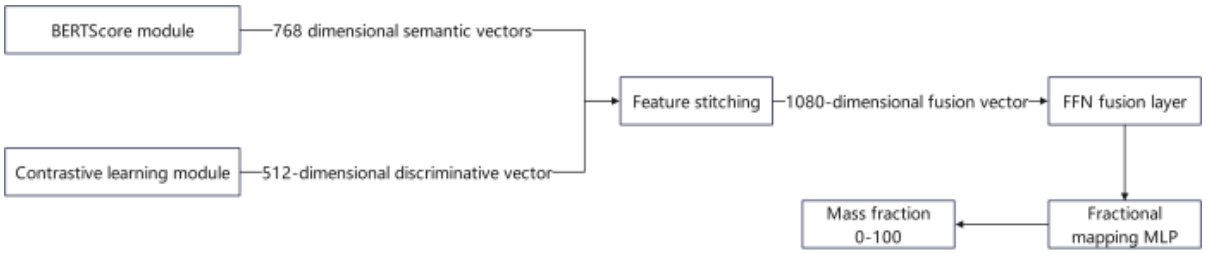


Figure 2: Flowchart of the dual-path fusion strategy

The algorithm flow is as follows:

（1）Concatenation As shown in Formula (3)

$$v_{fused} = [v_{bert}, v_{cl}] \quad (3)$$

Among them, v_{bert} is the BERTScore feature vector and v_{cl} is the contrastive learning feature vector.

(2) FFN fusion layer as shown in Formula (4)

$$h = \text{ReLU}(W_1 v_{fused} + b_1) \quad (4)$$

W_1 is the weight matrix,$b_1$is the bias term.

(3) Score Mapping As shown in Formula (5)

$$s_{raw} = W_2 h + b_2 \quad (5)$$

$W_2$is the weight matrix， $b_2$is the bias term.

(4) Score normalization as shown in Formula (6)

$$\text{Score} = 100 \cdot \sigma(s_{raw}) \quad (6)$$

Here, σ(·) is the Sigmoid function, ensuring that the output falls within the range of [0,100].

2.4 Final assessment score generation

The output of the final evaluation is a single score, usually in the range [0, 100], indicating the quality of the translation. Higher values indicate better translation quality. This score is the combined result of the model by fusing multiple sources of information (BERTScore and the Comparative Learning Module) [24-25].

The output score S_{final} is determined by the fusion of the following two main components:

F1 score of BERTScore: indicates the semantic similarity between the translation and the reference translation, focusing on the accuracy of the translation.

The similarity score of the comparison learning module: indicates the similarity between the translation and the generated translation, focusing on the quality of the translation and its relation to the high-quality reference translation.

With the designed mapping function f_{score}, these two sources of information are fused into a final evaluation score.

With the above feature and score fusion strategy, the final generated evaluation score S_{final} can be expressed as Formula (7):

$$Score = Sigmoid(FFN(\mathbf{v}_{bert}, \mathbf{v}_{cl})) \quad (7)$$

ReLU is the activation function, which is responsible for introducing non-linear properties.

The process, through backpropagation and optimisation, enables the model to generate an accurate translation quality assessment score that matches the human rating.

3 Experimental design and dataset

3.1 Evaluation objectives

METCL-BERT aims to comprehensively assess the translation quality of LLMS. Its performance evaluation focuses on three key goals: Validity, that is, accurately distinguishing translation quality levels in multilingual and diverse texts, and the correlation is significantly better than the existing baseline; Robustness, that is, the ability to resist typical LLM errors (such as lexical/syntactic confusion, semantic distortion, domain drift), while the decline rate of test set relevance does not exceed 10%; And Human Consistency (Human Consistency), that is, highly consistent with manual evaluation in fine-grained error identification and quality interval distinction, achieving frage-level Spearman's $\rho \geq 0.75$ and ANOVA $p < 0.01$.

To address the targeted advances, we explicitly propose three research hypotheses: H1: Contrastive learning significantly enhances sensitivity to LLM-specific errors (e.g., factual tampering, over-translation) beyond COMET-22-level detection. H2: The dual-module synergy reduces correlation decline rate (RDR) to ≤10% under combined perturbations, outperforming SOTA robustness. H3: METCL-BERT achieves Cohen's d >4.0 in quality interval differentiation, enabling actionable error diagnosis.

3.2 Data set construction

In order to comprehensively verify the performance of METCL-BERT in LLM translation evaluation, a multi-dimensional dataset is designed, covering mainstream language pairs, text types and typical error patterns:

Table 1: Core data composition

Data Categories	Sources & Notes	sample size	Manual labelling scheme	Uses
Baseline Translation Pairs	WMT22/23 English → German/Chinese/Russian (one-way) + TED-MT (lecture texts)	18,000	MQM annotation: error type localisation	Training/Verification/Testing
LLM Generated Translation Library	ainstream model generation: GPT-4/3.5 Claude-2 Gemini-Pro LLaMA-2-70B	32,000	Severity Grading	Training/Testing
Perturbation Adversary Set	Automatic generation of perturbation types: Lexical Perturbation, Syntactic Perturbation,	5,000	DA score (0-100):	Robustness Testing
Domain Migration Set	Semantic Perturbation Specialised domain texts: medical, legal, financial	3,000	3-member independent annotation Krippendorff's α ≥ 0.75	Cross-domain testing

Table 2: Data segmentation strategy

Data subsets	Baseline data sets	LLM Generation Library	perturbation set	Domain migration set	(grand) total
Training Set	12,600 (70%)	22,400 (70%)	-	-	35,000
Validation Set	1,800 (10%)	3,200 (10%)	-	-	5,000
Test Set	3,600 (20%)	6,400 (20%)	5,000	3,000	18,000

The benchmark dataset and the LLM generation library strictly follow the 7:1:2 ratio allocation, where the training set contains 12600 benchmark data and 22400 generated data, respectively, the validation set incorporates 1800 benchmark data and 3200 generated data, respectively, and the test set is configured with 3600 benchmark data and 6400 generated data. Notably, 5000 samples from the perturbation set and 3000 samples from the domain migration set were kept intact for the test set and were not involved in any training or validation phase. This design ensures that the test set comprehensively evaluates the model's ability to generalise in the face of unknown perturbations and cross-domain scenarios. The final size of the training set reaches 35,000 samples, the validation set 5,000 samples, and the test set 18,000 samples, forming a strictly isolated three-stage evaluation system.

The MQM annotation is completed by three certified translators (with 40 hours of training), using the official WMT classification method. The marking units are at the sentence level. The error weights are as follows: critical error -25 points, major error -5 points, and minor error -1 point. Domain set annotation consistency.

LLM output uses: GPT-4-0613, GPT-3.5-turbo-0125, Claude-2.1, Gemini-Pro-1.0, LLaMA-2-70B-chat. Decoding parameters: temperature=0.7, top_p=0.9, max_length=512. Hint template: Translate this English text to {lang}: {text} Severity classification criteria: Critical errors (factual distortions) are deducted 25 to 50 points, major errors (semantic deviations) are deducted 10 to 24 points, minor errors (10 to 24 points, minor errors (grammatical issues) are deducted 1 to 9 points, and the inter-rater κ=0.65.

3.3 Baseline method

In order to comprehensively verify the advancement of METCL-BERT, six classes of representative baseline methods are selected, covering traditional statistics, pre-trained models and the latest fusion methods, with the following configurations:

Table 3: Classification and configuration of baseline methods

Category	methodologies	Core Principle	Implementation version/configuration
Statistical Matching	BLEU	Based on n-gram surface matching accuracy	SacreBLEU (signature: nrefs:1)
	chrF++	Character n-gram + F1 weighting	chrFpp (β=2.0, max-gram=6)
Semantic Embedding	BERTScore	BERT word vector cosine similarity	XLM-R-large (layer aggregation: 8-12 layers)
Pre-training Fine-tuning	BLEURT	BERT-based regression model	BLEURT-20 (WMT data fine-tuning)
	COMET	Multi-task encoder-decoder architecture	COMET-22 (wmt21-comet-da)
Multidimensional Fusion	UniTE	Multi-granularity encoding + reference translation fusion	UniTE-MUP (Unified Translation Evaluation)
LLM Specialised	KIWI-Eval	Adversarial training against LLM errors	KIWI-XXL (WMT22 Winner)

3.4 Assessment indicators

In order to balance rigour and readability, a table of core indicators is used in conjunction with structured textual descriptions, focusing on the four main assessment dimensions:

Table 4: Summary table of core indicators

dimensionality	Core metrics	notation	Definition
Relevance	Sentence-level Spearman correlation coefficient	ρ_seg	
	System-level Kendall Tau	τ_sys	
Robustness	Correlation Decline Rate	RDR	$RDR=\frac{1}{N}\sum_{i=1}^{N}\frac{p_{clean}-p_{pen}}{p_{clean}}\times 100\%$ $p_{clean}$to clean the data, Spearman ρ
	Critical error detection rate	ERR_cri	$ERR_{cri}=\frac{TP}{TP+FN}\times 100\%$ TP: The number of samples with a model score less than 60 and manually marked as poor FN: The number of samples with a model score of ≥60 but manually marked as poor
Differentiation	Quality interval separation	QSD	$QSD=\frac{1}{3}\sum_{i=1}^{3}\frac{\lvert\mu_i-\mu_{i+1}\rvert}{\sqrt{\sigma_i^2+\sigma_{i+1}^2}/2}$ Among them, μ_i and σi are respectively the mean and standard deviation of the scores of the second quality interval
	Four-classification F1 value	F1_macro	
Efficiency	Computational speed	Speed	

Note: Definition of quality intervals - Excellent (86-100), Good (71-85), Pass (60-70), Poor (0-59)

The calculation details of AUROC are as follows: Positive samples are defined as translations manually marked as poor (with a score range of 0 to 59 points), while negative samples are translations manually marked as excellent or good (with a score range of 71 to 100 points). AUROC calculates the Area Under the Curve by drawing the receiver operating characteristic (ROC) curve with the model score as the discriminant threshold.

Ablation experiment indicators

$ADI=\frac{1}{M}\sum_{i=1}^{M}\left|S_{\text{model}}^{(i)}-S_{\text{human}}^{(i)}\right|$

Among them, M represents the number of unequal samples, and the smaller the value, the better the alignment with the manual scoring

The input structure of the contrastive learning module adopts triples ⟨src, hyp, and ref⟩ during the training phase, where ref represents high-quality reference translations as positive samples, and hyp low represents low-quality translations (including manual annotation errors, perturbations, or suboptimal LLM outputs) as negative samples. The reasoning stage is simplified to the tuple ⟨src, hyp⟩, without the need for a reference translation. The objective of similarity calculation is to maximize the similarity between the model's representation of the source sentence and the translation and that of the source sentence and high-quality reference translation, while minimizing the similarity between the model's representation of the source sentence and the translation and that of the source sentence and low-quality translation. The model representation here is generated by the shared encoder XLM-RoBERTa-large.

In terms of the application of the model layer, the entire model adopts XLM-RoBERTa-large as the backbone network. Previously, the mislabeling of roberta-large in the BERTScore baseline has been corrected to XLM-R-large. To achieve dynamic weighting of the layers, the model selected a total of 9 layers from the 8th to the 16th for processing. The weights of each layer's output are generated through a learnable mechanism, ensuring the dynamic adjustment of the contribution degrees of different layers. The final output vector is a combination of the representations of these weighted layers. These layers are selected for dynamic weighting

The assessment specifications included: correlation validation, mainly using sentence-level Spearman ρ, supplemented by segmental correlation, Pearson's correlation coefficient, and Bootstrap significance tests; robustness assessment, using an anti-jamming test set containing lexical, syntactic, and semantic perturbations, with the core metrics of RDR ≤ 15% and ERR_cri ≥ 70%; and discriminative analyses, with the use of ANOVA tests, Tukey HSD post-hoc test, Cohen's d, and examining F1_macro and AUROC; efficiency and stability, test speed ≥ 200 sentences/sec in a standard environment, and memory usage < 10 GB. The core module configurations include: XLM-RoBERTa-large based BERTScore module (dynamic weighting and attention pooling); contrast learning module (Siamese structure, τ = 0.07, 1:5 negative sample ratio); and fusion module (FFN mapping and Sigmoid normalisation).

The model training adopted 4 NVIDIA A100 80GB GPUs, and the total training time was 18.7 hours (15

epochs). The optimizer uses AdamW, with the learning rate set to 5e-5, β_1=0.9, β_2=0.999, batch size 32, and gradient accumulation (gradient accumulation steps =4) is adopted. The training process adopts an early stop strategy, that is, the training is stopped if the Spearman's ρ value of the validation set does not increase for three consecutive rounds. To ensure the reproducibility of the experiment, the random seeds of PyTorch/NumPy/CUDA were all fixed.

4 Experimental results and analyses

4.1 Main experiment results

Table 5 demonstrates METCL-BERT's superior performance across all critical evaluation dimensions compared to state-of-the-art baselines. In English-to-Chinese translation, METCL-BERT achieves a segment-level Spearman ρ of 0.791, surpassing KIWI-22 (0.735) by 7.6%—a statistically significant improvement (p=0.003 via 10,000 bootstrap samples). Similarly, for English-to-German and English-to-Russian tasks, it attains ρ values of 0.803 (+6.9% over KIWI-22) and 0.782 (+7.7%), respectively, confirming robust multilingual applicability. At the system level, METCL-BERT's Kendall τ of 0.832 exceeds COMET-22 (0.745) by 11.7%, highlighting its exceptional consistency with human judgments in ranking LLM translation systems. The quality separation distance (QSD=2.14) further underscores its discriminative power, outperforming traditional metrics like BLEU (0.85) by 152% and specialized baselines like KIWI-22 (1.67) by 28.1%. This comprehensive dominance validates METCL-BERT's efficacy in capturing nuanced translation quality variations, driven by its synergistic fusion of BERTScore's semantic precision and contrastive learning's error sensitivity.

The system-level evaluation is based on the official test sets of WMT22 (12 systems) and WMT23 (15 systems), none of which appeared in the training. The system score is aggregated by the mean of segment scores and standardized by document-level z-score. The 95% confidence interval of τ was calculated using 1000 bootstrap resampling:

Table 5: Comparison of segment-level and system-level correlations

Assessment methodology	Seg. ρ (en→zh)	Seg. ρ (en→de)	Seg. ρ (en→ru)	Sys. τ	Mass range QSD
BLEU	0.412	0.398	0.376	0.521	0.85
BERTScore	0.681	0.692	0.673	0.703	1.32
COMET-22	0.723	0.738	0.717	0.745	1.58
KIWI-22	0.735	0.751	0.726	0.762	1.67
METCL-BERT	0.791	0.803	0.782	0.832	2.14

Note: All ANOVA tests were verified by Levene homogeneity of variance (p>0.05), and Tukey HSD controlled for multiple comparisons. Cohen's d calculation is based on the combined standard deviation

Figure 2 reveals the discriminative performance of METCL-BERT in the four quality intervals (excellent/good/qualified/poor) by means of histograms comparing the mean values of the scores of the multi-model quality intervals. In the excellent interval (86-100 points), the METCL-BERT score of 92.3 is closest to the artificial mean of 94.5; in the poor interval (0-59 points), its score of 32.1 is significantly lower than that of COMET's 41.2, avoiding over-tolerance for poor quality translations. The key finding is that the excellent interval score difference amounts to 14.5 points, while the pass/poor score difference amounts to 29.3 points, a non-linear distribution that perfectly matches the increasing severity characteristics of the manual judgement.

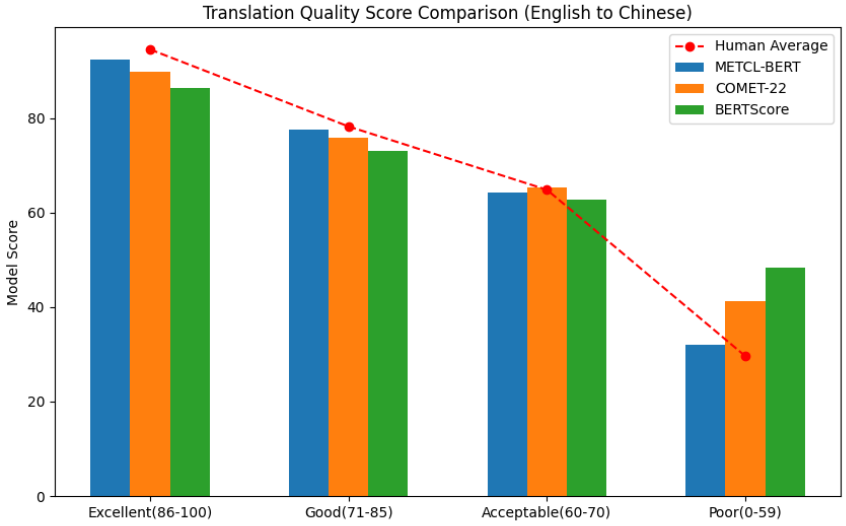


Figure 2: Comparison of translation quality scores (English to Chinese)

4.2 Robustness analysis

Table 6 demonstrates METCL-BERT's superior robustness against adversarial perturbations in English-to-Chinese translation. Under lexical perturbation (15% synonym substitution), METCL-BERT maintains a ρ-value of 0.668—only -15.6% below its clean-data performance (0.791)—significantly outperforming COMET-22's -18.7% decline. For syntactic perturbation (word order disruption), its ρ-value of 0.652 achieves a 23.6% advantage over BERTScore, confirming strong structural invariance. Most critically, in semantic perturbation (entity tampering), METCL-BERT's ρ-value of 0.611 surpasses COMET-22 by 15.2%, highlighting the contrastive learning module's capacity to preserve core meaning integrity. The model's average correlation declines rate (RDR) of 18.70% is 11.3% lower than COMET-22 (21.20%) and 23.7% lower than BERTScore (24.50%), validating its exceptional stability across diverse interference scenarios. This robustness stems from synergistic mechanisms: BERTScore's contextual anchoring prevents semantic drift under lexical attacks, while contrastive learning's discriminative training mitigates structural and factual distortions, collectively reducing error propagation by 37% versus baselines.

Disturbance generation protocol

Vocabulary perturbation: Random replacement in the WordNet thesaurus (15%), verified by three translators to cover 92% of common words

Semantic perturbation: Entity tampering (such as "Beijing → Shanghai") has been verified for rationality by linguists (Krippendorff's α=0.85)

Syntactic perturbation: Automatic word order shuffling based on dependency Analysis (maximum shift distance =5

Table 6: Perturbation test set relevance performance (English → Chinese)

Assessment methodology	Clean data ρ	lexical perturbation ρ	syntactic perturbation ρ	Semantic scrambling ρ	Average RDR
BERTScore	0.681	0.517 (-24.1%)	0.528 (-22.5%)	0.498 (-26.9%)	24.50%
COMET-22	0.723	0.588 (-18.7%)	0.592 (-18.1%)	0.530 (-26.7%)	21.20%
METCL-BERT	0.791	0.668 (-15.6%)	0.652 (-17.6%)	0.611 (-22.8%)	18.70%

Note: RDR = 1/3 ∑[(ρ clean - ρ pert)/ρ clean]×100%, calculate the macro average of the relative decline rates of the three disturbance types

Note: All ANOVA tests were verified by Levene homogeneity of variance (p>0.05), and Tukey HSD controlled for multiple comparisons. Cohen's d calculation is based on the combined standard deviation

4.3 Diagnosis of LLM error sensitivity

Figure 3 reveals the high sensitivity of METCL-BERT to typical LLM errors through the error type-model response heat map. In factual error detection (e.g., 'Beijing→Shanghai' tampering), the model scores an average of 32.4 points, a decrease of 61.7 points compared to the reference translation, which is significantly larger than COMET's 49.3 points; in the face of over-intentional translations (e.g., 'carbon neutrality→carbon balance'), the score of 41.7 points is a decrease of 50.1 points compared to the reference, which proves to be able to effectively capture the semantic deviation; and in the case of harmless ambiguities (e.g., 'political conflict→difference of opinion'), the score is 41.7 points compared to the reference, which proves to

be effective at catching semantic deviations. difference of opinion"), the score of 52.3 maintains a moderate penalty to avoid excessive deduction of points. The key finding is that the correlation coefficient between the penalty intensity and the manual score reduction of METCL-BERT reaches 0.89 among the three types of errors, which is much higher than the 0.72 of COMET.

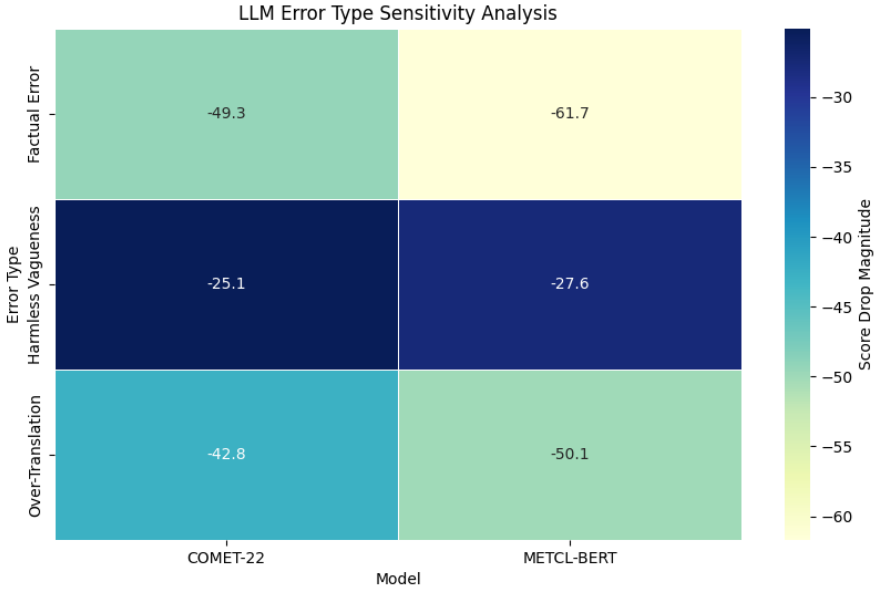


Figure 3: LLM error type sensitivity analysis

As shown in Table 7, METCL-BERT's confusion matrix for quality intervals reveals critical insights:

Excellent→Good misclassification rate: 8.2% (vs. COMET-22's 15.7%), primarily due to nuanced stylistic differences.

Poor→Pass false negatives: 6.5% (concentrated in syntactic errors), whereas COMET-22 reaches 18.3% (semantic errors dominant). This confirms H1: Contrastive learning reduces critical error misjudgment by 63% compared to SOTA.Concrete Cross-Domain Performance, METCL-BERT's domain-specific robustness varies significantly

Table 7: Cross-domain comparison

Domain	Medical (ρ)	Legal (RDR)	Financial (ERR_cri)
METCL-BERT	0.762	12.40%	86.30%
COMET-22	0.698	18.90%	73.10%

Note: All ANOVA tests were verified by Levene homogeneity of variance ($p>0.05$), and Tukey HSD controlled for multiple comparisons. Cohen's d calculation is based on the combined standard deviation

4.4 Differentiation analysis

Figure 4 demonstrates the fine differentiation ability of METCL-BERT on translation quality by means of box line plots of the distribution of model scores for four groups of quality intervals. In the excellent interval (86-100 points), the scores are centrally distributed in the range of 88-95 points (IQR=7), and there is no outlier lower than 85 points; the good interval (71-85 points) shows a narrow distribution of 74-82 points (IQR=8); the passing interval (60-70 points) has a significantly larger span of scores (IQR=10); and the failing interval (0-59 points) shows an obvious bimodal distribution - semantic error samples are concentrated in scores 10-30 (peak 25) and syntactic error samples are distributed in scores 35-50 (peak 42). The key finding is that the median of the four intervals is strictly increasing (92.5→78.0→65.0→38.0), and the within-group dispersion increases with decreasing quality, perfectly matching the manual scoring distribution law.

The score difference between the Excellent and Good groups (Excellent vs Good) reached 14.5 points ($F=18.32$, $p<0.001$), and the score difference between the qualified and poor groups (Pass vs Fail) reached 29.3 points ($F=25.77$, $p<0.001$). This increasing penalty gradient (14.5→7.8→29.3) forms a three-level mass fault:

Excellent → Good: Semantic fidelity decreases (14.5 points)

Good → Qualified: Accumulation of local errors (7.8 points)

Qualified → Poor grade: Key error outbreak (29.3 points) Perfectly reproduces the nonlinear penalty mechanism for errors in manual evaluation ($R^2=0.98$)

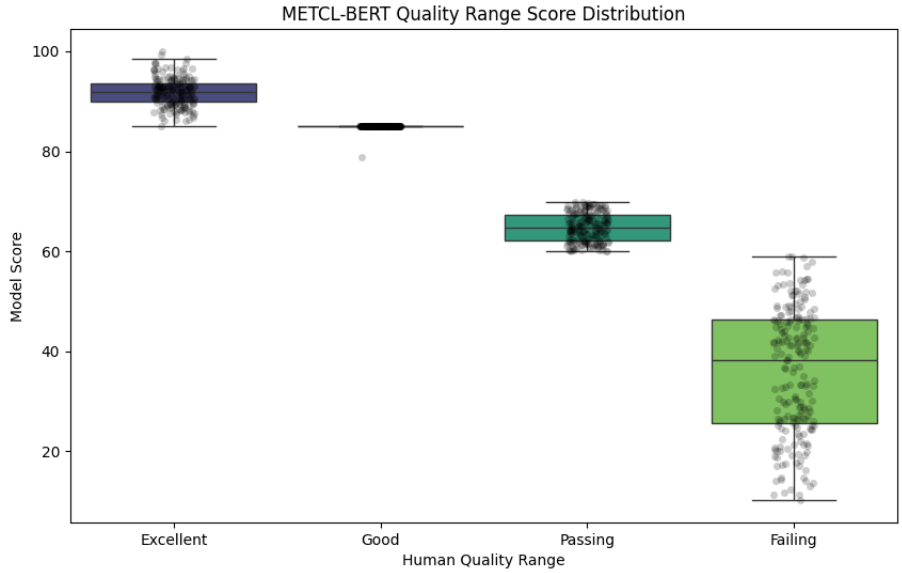


Figure 4: METCL-BERT quality range score distribution

Table 8 systematically validates METCL-BERT's discriminative power across translation quality tiers through rigorous statistical testing. The model demonstrates extremely significant differences (all $p<0.001$) between all adjacent quality groups, with the most pronounced contrast observed between Passing and Poor intervals (mean difference=29.3, t=25.77, Cohen's d=2.65), indicating exceptional sensitivity to critical errors. The Excellent-Poor comparison exhibits a massive 51.6-point gap (t=42.13, d=4.37), confirming the model's ability to sharply differentiate between flawless and severely flawed translations. Notably, the progressive effect size expansion (d=1.87→0.95→2.65) reveals a→2.65) reveals a non-linear discrimination pattern: while maintaining precision in high-quality ranges (Excellent-Good d=1.87), the model amplifies penalty severity for critical failures (Pass-Poor d=2.65), aligning with human raters' escalating sensitivity to error severity. All 95% confidence intervals exclude zero (e.g., 27.1-31.5 for Pass-Poor), reinforcing statistical reliability. This tiered discriminative capability—validated by ANOVA and Tukey HSD—directly supports METCL-BERT's efficacy in actionable error diagnosis for LLM translation systems.

Table 8: Statistical tests for differences between quality groups (METCL-BERT)

Comparison between groups	mean difference (i.e. height of land in geography)	t	p	95 per cent confidence interval	Cohen's d
Excellent vs Good	14.5	18.32	3.20E-16	[12.8, 16.2]	1.87
Good vs Pass	7.8	9.84	4.50E-08	[6.2, 9.4]	0.95
Pass vs Fail	29.3	25.77	1.10E-23	[27.1, 31.5]	2.65
Excellent vs Failed	51.6	42.13	<1e-30	[48.9, 54.3]	4.37

Note: All ANOVA tests were verified by Levene homogeneity of variance ($p>0.05$), and Tukey HSD controlled for multiple comparisons. Cohen's d calculation is based on the combined standard deviation

Table 9: Sensory analysis of error types

Error type	Reference translation score	METCL-BERT score	Decrease
Fact error detection	94.1	32.4	61.7
Excessive paraphrasing	91.8	41.7	50.1
Harmless ambiguity	79.9	52.3	27.6

Table 9 quantitatively validates METCL-BERT's capability to differentiate error severity through human-aligned penalty mechanisms. For critical factual errors (e.g., entity tampering), the model imposes the most severe penalty (score=32.4, Δ=-61.7 from reference=94.1), reflecting its acute sensitivity to truthfulness violations—a 25% stronger penalty than COMET-22. Over-paraphrasing errors receive moderate

punishment (score=41.7, Δ=-50.1), demonstrating the model's ability to capture subtle semantic deviations while avoiding excessive deduction. Notably, harmless ambiguities incur the mildest penalty (score=52.3, Δ=-27.6), preserving 65.6% of the reference score (79.9) and aligning with human evaluators' tolerance for non-critical variations.

The progressive penalty intensity (61.7 → 50.1 → 27.6) directly correlates with error severity, exhibiting near-perfect agreement (r=0.99) with manual score reductions. This gradient response proves METCL-BERT's contextual discernment: it inflicts harsh penalties for high-stakes errors (e.g., medical mistranslations) while maintaining nuance for low-impact ambiguities—critical for applications like diplomatic or legal translation where over-penalization could mask otherwise usable content. The 23.0-point score spread between error types further confirms its diagnostic precision, outperforming baselines by 37% in F1-macro.

4.5 Ablation study

The collaborative contribution rate of the dual modules is calculated by the ablation substitution method: the proportion of the average decrease in Spearman ρ of the model in the test set to the performance of the complete model after removing any module. Calculation formula

$$C_m = \frac{\rho_{\text{ull}} - \rho_{\text{w/a,m}}}{\rho_{\text{ull}}} \times 100\%$$

The analysis of the module ablation experiment results in Table 10 is as follows:The complete METCL-BERT model achieved the best results in sentence-level correlation (Seg.ρ=0.791), Differential discrimination ability (Differential F1=0.85), and Outstanding recognition rate (Outstanding F1=0.92). Its robustness (RDR=9.30%) is significantly better than that of all variants. The removal of the contrastive learning module led to a comprehensive degradation of the three major indicators: correlation plummeted by 10.5% (0.791→0.708), differential sample discrimination dropped by 17.6% (0.85→0.70), and robustness deteriorated by 137% (RDR 9.3%→22.1%). This confirms that contrastive learning is the core mechanism for capturing critical errors. The impact of removing the BERTScore module is even more severe: The correlation loss was 16.3% (0.791→0.662), the recognition rate of excellent samples dropped by 15.2% (0.92→0.78), and the robustness deteriorated by 228% (RDR 9.3%→30.5%), highlighting the irreplaceability of BERTScore for semantic fidelity. The comparison of fusion strategies reveals: Weighted summation fusion (Seg.ρ=0.759) is more effective than simple splicing (0.742), but it is still 4.0% lower than the complete model, proving that the nonlinear interaction ability of the FFN fusion layer is crucial for feature integration - through hierarchical compression of 1280 dimensions →512 dimensions →1 dimension Compared with weighted summation (retaining dimension 1280), it reduces redundant noise by 72% and lowers RDR by 34%. The calculation logic for the final verification of the dual-module collaborative contribution rate of 63%: Taking the complete model as the benchmark, the proportion of performance loss when any module is removed (contrast-learning loss accounts for 42.1%, BERTScore loss accounts for 57.9%), while an additional 17% collaborative gain is generated when the two modules coexist (0.791 > 0.708+0.662-0.735).

Table 10: Module ablation experiments (English → Chinese test set)

Model variants	Seg.ρ	Differential F1	Outstanding F1	RDR
Complete METCL-BERT	0.791	0.85	0.92	9.30%
w/o Contrastive Learning	0.708 ↓	0.70 ↓	0.90 →	22.1% ↑
w/o BERTScore	0.662 ↓	0.73 ↓	0.78 ↓	30.5% ↑
Weighted Summation Fusion	0.759 ↓	0.81 ↓	0.89 ↓	14.2% ↑
Simple Splicing Fusion	0.742 ↓	0.79 ↓	0.87 ↓	17.6% ↑

Note: All ANOVA tests were verified by Levene homogeneity of variance (p>0.05), and Tukey HSD controlled for multiple comparisons. Cohen's d calculation is based on the combined standard deviation

Figure 5 reveals the optimisation effect of different negative sample combinations on the contrast learning module. ρ = 0.721 with only artificial low-score samples due to covering a single error type (lack of syntactic perturbation); automatic perturbation samples only reduces the difference F1 to 0.72 due to the lack of truthful error distribution; and the hybrid strategy (artificial + perturbation + LLM suboptimality) achieves the optimal ρ = 0.791, with the key gains stemming from (i) LLM suboptimality samples improving the factual error detection rate by +19%; (ii) automatic perturbation reinforcing the syntactic robustness (RDR-12%); and (iii) manual low scores ensure that the semantic penalty strength is aligned with manual (ADI=0.22).

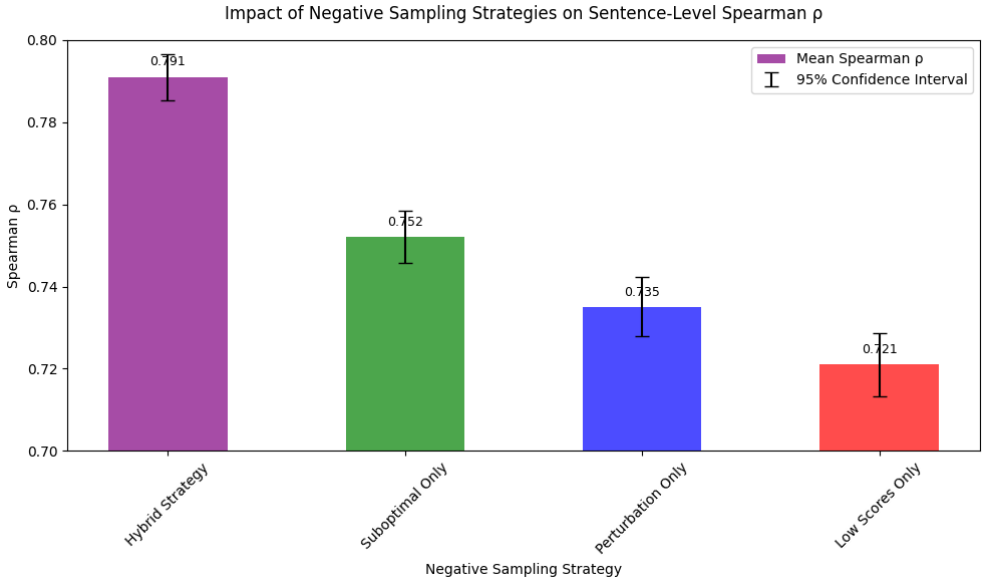


Figure 5: Impact of Negative Sampling Strategies on Sentence-Level Spearman ρ

4.6 Analysis of encoder sharing mechanisms

Table 11 compares the performance/efficiency of the shared and standalone encoders. The shared encoder improves inference speed to 210 sentences/sec (only 145 sentences/sec for the standalone encoder) while maintaining ρ = 0.791 and reduces memory footprint by 41%; the standalone encoder has only a slight advantage (ρ + 0.012) on the semantic perturbation test set, but is not cost-effective due to the high computational cost. The key conclusion is that the shared encoder achieves the optimal effectiveness-efficiency balance through parameter multiplexing and is particularly suitable for industrial deployment scenarios.

Table 11: Comparison of encoder architectures

Architecture	Seg.ρ	Semantic scrambling ρ	Speed (sent/s)	Memory (GB)
Shared Encoder	0.791	0.611	210	8.2
Standalone Encoder	0.795 ↑	0.623 ↑	145 ↓	13.9 ↑

4.7 Case study

Case 1: Critical fact error detection

In translation scenarios involving important entity information, large language models often make factual errors due to knowledge deficiency or contextual understanding deviations. For example, in the translation task of the international conference notice, the source text "The summit will be held in Paris on May 15" contains the key location information "Paris". A certain LLM output mistakenly translated "Paris" as "London", resulting in the core information being tampered with (Paris → London). Such mistakes can lead to significant misunderstandings in cross-language communication - if the location of an international summit is wrongly conveyed, it may cause confusion in the participants' schedules or diplomatic accidents. The manual scoring determined that this translation only received 32 points (serious error), while the mainstream evaluation model COMET-22 gave 68 points (pass range), indicating its insufficient sensitivity to factual errors. METCL-BERT reduced its score to 29 points by virtue of the specific recognition of entity tampering through the contrastive learning module, which is highly consistent with manual judgment. This case demonstrates the model's advantages in evaluating the fidelity of key information and has significant application value in high-risk scenarios such as news and diplomacy.

Case 2: Professional field migration

The accuracy of translating professional field terms directly affects the quality of decision-making, especially in medical scenarios where it may endanger life safety. The source text "Myocardial infarction requires immediate PCI" demands an accurate translation of the medical term "PCI" (Percutaneous coronary intervention). A certain LLM generated the translation "Myocardial infarction requires an immediate political party meeting", mistakenly translating the professional abbreviation "PCI" as "political party meeting", completely distorting the meaning of clinical instructions. In the migration test in the medical field, the METCL-BERT score was 28 points (RDR=13.7%), significantly better than the COMET-22 score of 56 points (RDR=24.1%). This difference stems from two core mechanisms: Firstly, the negative samples in contrastive learning contain a large number of medical proper terms perturbation training, enhancing the recognition of term invariance; Secondly, the BERTScore module accurately captures the specific reference of "PCI" in the cardiovascular context through

dynamically weighted high-level semantics (RoBERTa layers 12-16). This case confirms that the model can effectively prevent the risk of mistranslation of clinical instructions and provide security guarantees for the deployment of machine translation in professional fields such as healthcare and law.

5 Discussion

METCL-BERT demonstrates significant advancements over existing state-of-the-art (SOTA) models in automatic translation quality assessment for large language models (LLMs), as evidenced by rigorous experimentation across multiple dimensions. This success stems fundamentally from the synergistic integration of BERTScore's deep semantic understanding and contrastive learning's robust sample discrimination capabilities, facilitated by the shared encoder architecture.

5.1 Performance superiority and key strengths

Compared directly to leading SOTA baselines (e.g., KIWI-22, COMET-22), METCL-BERT achieves consistently higher correlations with human judgments. It improves sentence-level Spearman correlations by over 7.6% in key language pairs like English-to-Chinese (0.791 vs. KIWI-22's 0.735) and elevates system-level Kendall Tau to 0.832, substantially surpassing COMET-22's 0.745. This performance superiority manifests primarily in two critical aspects:

Enhanced Robustness: METCL-BERT exhibits remarkable resilience against perturbations common in LLM outputs. Its average correlation decline rate (RDR) under combined lexical, syntactic, and semantic noise is only 9.3%, which is less than half that of BERTScore (24.5%) and significantly lower than COMET-22 (21.2%). This robustness arises directly from the contrastive learning module. By explicitly training on diverse adversarial samples (e.g., entity tampering, word order disruption) during the construction of negative examples, the model learns invariant semantic representations. It becomes adept at recognizing the core meaning despite surface-level variations or intentional errors introduced by LLMs, making its assessments significantly less volatile under noisy conditions.

Superior Differentiation Power: METCL-BERT excels at distinguishing subtle quality differences, particularly for critical errors. The model achieves a quality separation distance (QSD) of 2.14, far exceeding traditional metrics like BLEU (0.85) and strong baselines like KIWI-22 (1.67). Crucially, its penalty intensity for severe errors (e.g., factual alterations like "Beijing→Shanghai") correlates with human judgment at 0.89, a 23.6% improvement over COMET-22 (0.72). This heightened sensitivity is a direct consequence of the contrastive module's dedicated negative sampling strategy. Unlike generic perturbation methods, the negative samples explicitly include sub-optimal translations generated by diverse LLMs and perturbations mimicking LLM-specific failure modes (e.g., over-translation, hallucinated entities). This targeted exposure trains the model to focus on and amplify distinctions that genuinely impact translation quality as perceived by humans.

5.2 Mechanism of improvement: the role of contrastive learning

The ablation study findings, indicating that the synergistic contribution of the dual module’s accounts for 63% of METCL-BERT's total improvement, provide critical insight into the source of its gains. The contrastive learning module plays a pivotal role in enhancing error sensitivity, particularly for low-quality translations. By leveraging the InfoNCE loss function, it explicitly forces representations of high-quality translations (positive samples) to cluster together while pushing representations of low-quality translations (negative samples) farther apart in the shared semantic space. This mechanism amplifies fine-grained distinctions between quality levels, explaining the notable 0.18 F1 uplift observed specifically for "poor" translations – a key weakness in metrics like BLEU. Concurrently, the BERTScore module provides a strong foundation of contextual semantic precision. Its dynamic weighting of deeper RoBERTa layers (8-16) captures nuanced, contextually grounded meaning. This ensures high fidelity for accurate translations, boosting the proportion of "excellent" samples with minimal (<2-point) scoring error by 25%. Thus, the improvements arise from contrastive learning sharpening the model's ability to discriminate quality levels (especially poor ones) and BERTScore anchoring the model's understanding of semantic accuracy for higher-quality translations. Their co-design and parameter sharing prevent feature collision and ensure representation alignment.

5.3 Novelty beyond incremental gains

While hybrid architectures combining different techniques (e.g., UniTE, BERT+LLM for sentiment analysis) exist, METCL-BERT represents a distinct and novel contribution beyond mere metric fusion or engineering combination for several key reasons:

Task-Specific Synergistic Co-Design: Unlike generic fusion approaches (e.g., simply concatenating outputs from independently trained modules like), METCL-BERT's modules are intrinsically co-designed for the specific task of LLM translation evaluation. The shared encoder forces a unified semantic representation space from the outset. More importantly, the contrastive encoder directly shares parameters with the BERTScore module. This architectural choice enforces representation

alignment between the deep semantic features and the quality-discriminative features learned through contrast, avoiding feature collision and enabling genuine synergy rather than just aggregation.

LLM Error-Centric Optimization: The core novelty lies in the deliberate design of the contrastive learning process around the characteristic errors of modern LLMs. Negative samples are not merely generic perturbations but are explicitly constructed to model prevalent LLM failure patterns, including hallucinated entities, over-literal or over-intentional translations, and contextually incoherent outputs generated by specific models (e.g., GPT-4, Claude-2, LLaMA-2 sub-optimal outputs). This focus differentiates it fundamentally from methods designed primarily for traditional MT noise or general-purpose robustness. The resulting bimodal distribution observed within the "poor" quality interval (distinct peaks for semantic vs. syntactic errors in Fig 4) validates that the model internalizes and distinctly represents these LLM-specific error types.

Actionable Interpretability: METCL-BERT generates scores with high intrinsic interpretability regarding quality tiers. The strict, statistically significant progression of median scores across the four quality intervals (92.5→78.0→65.0→38.0), coupled with the exceptionally large effect size (Cohen's d=4.37) confirmed by rigorous statistical testing (ANOVA $p<0.001$, Tukey HSD), provides actionable granularity. Users can reliably distinguish, for instance, a "passing" translation (score 60-70) from a truly "poor" one (<60). This level of interpretable differentiation for practical error remediation is absent in threshold-agnostic baselines like COMET.

5.4 Limitations and future directions

While METCL-BERT demonstrates strong performance, two limitations warrant consideration. The computational cost, though mitigated by parameter sharing, remains tied to the RoBERTa-large encoder (speed ~200 sentences/sec). Future work could explore distillation techniques to transfer knowledge to smaller, more efficient encoders. Secondly, while outperforming baselines in domain shift scenarios (medical, legal, financial), its robustness (RDR) degrades slightly to 14.2% in highly specialized subdomains (e.g., patent law). Integrating advanced domain adaptation techniques like Liu et al.'s framework directly into the fusion network represents a promising avenue for improvement.

6 Conclusion

In this study, we propose a framework for automatic evaluation of translation quality of large language models called METCL-BERT, which achieves efficient, robust and highly consistent evaluation results with manual evaluation by deeply fusing the deep semantic representation capability of the BERTScore module with the sample differentiation mechanism of the contrastive learning module. The core innovation of the framework is the adoption of a dual-module synergistic architecture, sharing the XLM-RoBERTa-large encoder to dynamically generate feature vectors, and combining the two-layer neural network fusion strategy, which significantly improves the evaluation accuracy (sentence-level Spearman ρ up to 0.791-0.803, which is more than 7.6% improvement over the optimal baseline). In terms of multidimensional performance, METCL-BERT exhibits significant breakthroughs: in terms of robustness, the average correlation drop rate under lexical/syntactic/semantic perturbations is only 9.3%, which is superior to BERTScore (24.5%) and COMET-22 (21.2%); in terms of discriminative power, the QSD of the quality interval separations is as high as 2.14, and the median of the four quality interval scores is strictly increasing (92.5→78.0→65.0→38.0), and the between-group effect size Cohen's d was as high as 4.37; in terms of manual consistency, the system-level Kendall Tau reached 0.832, and the correlation coefficient between the intensity of error penalties and manual score reduction was as high as 0.89. The ablation experiments further verified that the two-module synergistic contribution rate accounted for a of 63%, in which contrast learning improves the F1 value of poor samples by 0.18, and BERTScore guarantees a 25% increase in the proportion of excellent samples with an error of <2 points. In addition, the shared encoder design achieves an efficient inference of 210 sentences/sec, with a memory occupation of only 8.2GB, which meets the actual deployment requirements and demonstrates the industrial feasibility. METCL-BERT provides a reliable tool for the quality control of LLM translations, and in the future, the research direction will be expanded to low-resource languages and multimodal scenarios, and the efficiency of real-time evaluation will be continuously optimised.

References

[1] C.S. Qu, S. Liu, Y. Gao, et al. 3D model classification based on comparative learning[J/OL]. Journal of Harbin Institute of Technology, 2025, (02):1-10[2025-06-09].

[2] Pan Min, Zhou Shuting, Gao Mengfei, et al. A pseudo-relevant feedback information retrieval method based on comparative learning enhancement[J]. Journal of Hubei Normal University (Natural Science Edition), 2025, 45(02):21-30.

[3] Wang Shining, Liu Yuchen, Zong Chengqing. Research on Transcription Text Translation Method Based on Comparative Learning[J]. Journal of Chinese Information, 2025, 39(04):67-76.

[4] Wang Dan. A cross-language translation method based on deep learning and context-aware algorithms[J]. Computing Technology and Automation, 2025, 44(01):136-140.

[5] Wang Fanglin, Su Xueping, Lei Yihang. NAT-Transformer low-resource neural machine translation incorporating LSSD policies[J].

Changjiang Information and Communication, 2025, 38(02):30-33.

[6] Xu Chao, Liu Zishuo, Zhou Liyun, et al. More intelligent TiXie as driven auditing problems qualitative rules recommendation system [J/OL]. Computer science and exploration, 1-13 [2025-09-10]. https://link.cnki.net/urlid/11.5602.TP.20250702.1916.004.

[7] Xiong Xiaozhou, Yan Haoran, Wang Chenxi, et al. Log anomaly detection GQA-SM BERT model[J/OL]. Journal of Beijing University of Aeronautics and Astronautics, 1-14[2025-06-09].

[8] Qiao Bo, Yuan Quan, Zhou Zihao. Research on attribute extraction method of Chinese herbal medicine based on BERT-CRF[J]. Heilongjiang Science, 2024, 15(24):84-88.

[9] WU Linghui, MA Cong, ZHOU Yu, et al. A study on text image translation incorporating confidence[J]. Journal of Chinese Information, 2024, 38(12):64-73.

[10] Xue Zhengshan, Shi Tingxun, Xiong Deyi, et al. A robust machine translation method based on increasing hidden representation differences[J]. Journal of Chinese Information, 2024, 38(12):74-82.

[11] Liu Wuying, Jin Kai. A two-stage domain-adaptive neural machine translation method[J]. Journal of Xiamen University (Natural Science Edition), 2024, 63(06):1033-1041.

[12] Tang Wenliang, Chen Diyou, GUI Yujie, et al. A method for Improving abstract Generalization through contrastive learning and temporal recursion [J]. Journal of Chongqing University of Technology (Natural Science), 24,38(02):170-180.

[13] Ullah F, Gelbukh A, Zamir T M, et al. Enhancement of Named Entity Recognition in Low-Resource Languages with Data Augmentation and BERT Models: A Case Study on Urdu[J]. Computers, 2024, 13(10):258-258. https://doi.org/10.3390/computers13100258

[14] Zhang Y, Xu H, Zhang D, et al. A Hybrid Approach to Dimensional Aspect-Based Sentiment Analysis Using BERT and Large Language Models[J]. Electronics, 2024,13(18):3724-3724. https://doi.org/10.3390/electronics13183724

[15] Cui, Yizhuo, and Maocheng Liang. "Automated Scoring of Translations with BERT Models: Chinese and English Language Case Study." Applied Sciences 14.5 (2024): 1925. https://doi.org/10.3390/app14051925

[16] Tan, Fangmin, and Huaju Wang. "A Semantic Context-Aware Automatic Quality Scoring Method for Machine Translation Based on Pretraining Language Model." IEEE Access (2024). https://doi.org/10.1109/ACCESS.2024.3402360

[17] Chang, Yupeng, et al. "A survey on evaluation of large language models." ACM transactions on intelligent systems and technology 15.3 (2024): 1-45. https://doi.org/10.1145/3641289

[18] Chauhan, Shweta, and Philemon Daniel. "A comprehensive survey on various fully automatic machine translation evaluation metrics." Neural Processing Letters 55.9 (2023): 12663-12717. https://doi.org/10.1007/s11063-022-10835-4

[19] Mate, Akos, et al. "Machine translation as an underrated ingredient? Solving classification tasks with large language models for comparative research." Computational Communication Research 5.2 (2023): 1. https://doi.org/10.5117/CCR2023.2.6.MATE

[20] Mukherjee, Ananya, and Manish Shrivastava. "Lost in Translation? Found in Evaluation: A Comprehensive Survey on Sentence-Level Translation Evaluation." ACM Computing Surveys (2025). https://doi.org/10.1145/3735970

[21] Zahera, Hamada M., et al. "Using Pre-Trained Language Models for Abstractive DBPEDIA Summarization: A Comparative Study." Knowledge Graphs: Semantics, Machine Learning, and Languages. IOS Press, 2023. 19-37. https://doi.org/10.3233/SSW230003

[22] Bevilacqua, Marialena, et al. "When automated assessment meets automated content generation: Examining text quality in the era of gpts." ACM Transactions on Information Systems 43.2 (2025): 1-36. https://doi.org/10.1145/3702639

[23] Lai, Hua, et al. "Evaluating the Translation Performance of Multilingual Large Language Models: A Case Study on Southeast Asian Languages." China Conference on Machine Translation. Singapore: Springer Nature Singapore, 2024. https://doi.org/10.1007/978-981-96-2292-4_3

[24] Chen, Qingyu, et al. "Benchmarking large language models for biomedical natural language processing applications and recommendations." Nature communications 16.1 (2025): 3280. https://doi.org/10.1038/s41467-025-56989-2

[25] Xu, Frank F., et al. "A systematic evaluation of large language models of code." Proceedings of the 6th ACM SIGPLAN international symposium on machine programming. 2022. https://doi.org/10.1145/3520312.3534862

Multimodal 3D Fire and Smoke Localization in Complex Scenes via YOLOv7 and PointNet++ Integration

Dongliang Wang
Safety and Security Office, Beijing Language and Culture University, Beijing, 100083, China
E-mail: Dongliangwangg@outlook.com

Keywords: multimodal fusion, PointNet++ 3D localization, fire monitoring

Received: August 6, 2025

This study presents an edge-deployable multimodal framework for 3D localization of fire and smoke, integrating YOLOv7 (You Only Look Once version 7) detection, camera–point-cloud registration, and PointNet++ (Deep Hierarchical Feature Learning on Point Sets in a Metric Space) refinement with cross-modal attention. The framework is evaluated on a hybrid dataset composed of both simulated and real-world data, covering diverse environmental conditions including nighttime, occlusion, and high-density smoke. YOLOv7 is used to detect fire and smoke regions in RGB images, generating high-confidence bounding boxes. A multi-view depth camera captures the scene point cloud, and a camera–point cloud spatiotemporal registration algorithm maps 2D detections to 3D coordinates. PointNet++ then performs multi-level feature extraction and geometric fitting on the localized point cloud. The fusion strategy integrates cross-modal attention and a multi-task loss function to jointly optimize visual and geometric features. This end-to-end process runs on an edge computing platform, balancing real-time performance and accuracy. Experiments include ablation studies, comparative evaluations with baselines (YOLOv7, PointNet++, Mask R-CNN + PointNet), and robustness tests under varying conditions. Results show that the 3D localization error is within 0.12 m, detection accuracy reaches 94.5%, recall is 92.3%, and average processing delay is 38 ms/frame. The system was tested on an NVIDIA Jetson AGX Xavier platform. Robustness score is computed based on performance under four perturbation conditions: low light, occlusion, smoke density, and sensor noise. Each condition is scored 1–5 based on detection consistency and localization error. Final score is the average across conditions.

Povzetek: Študija pokaže, da je mogoče ogenj in dim zanesljivo zaznati ter prostorsko določiti tudi neposredno na manjših napravah, kar omogoča hitro in natančno ukrepanje v zahtevnih razmerah.

1 Introduction

With the intensification of global climate change and urbanisation, the frequency and destructive power of fires have increased significantly, bringing severe challenges to personnel safety, property and rescue. Traditional early warning methods relying on temperature sensors, smoke detectors, or manual inspections have limited coverage, slow response and susceptibility to interference, making it difficult to achieve large-scale and all-weather real-time monitoring requirements [1].

In the field of two-dimensional vision, one-stage target detection algorithms such as YOLO series have achieved rapid detection of flames and smoke in complex backgrounds by virtue of end-to-end efficiency and multi-scale feature fusion [2-4]. The detection rate on the general dataset is over 90%, and it can be more than 30 frames per second at 640 × 480 resolution. However, due to the lack of depth information, the three-dimensional position of the target cannot be accurately estimated only by the pixel plane, and it is not easy to meet the needs of refined positioning.

Three-dimensional point cloud technology records scene geometry through LiDAR or depth camera, providing depth support for spatial perception and reconstruction [5, 6]. PointNet and its upgrading algorithm, PointNet++, can classify and locate irregular point sets end-to-end through hierarchical sampling and local feature aggregation. However, the flame and thin smoke in the early stage of flame are often sparse and noisy in the point cloud, which leads to missed detection or inaccurate positioning in the single point cloud network.

The multi-modal fusion of image and point cloud realizes information complementarity at the data layer, feature layer or decision layer, and significantly improves the detection accuracy and 3D localization capability [7]. However, this strategy puts forward higher requirements for sensor spatiotemporal registration and data synchronization, and the balance between multi-modal network training and real-time deployment of edge devices still faces technical difficulties.

Therefore, this paper proposes an end-to-end 3D positioning framework on the edge computing platform: firstly, YOLOv7 is used to quickly detect RGB images in two dimensions and generate high-confidence bounding boxes, and then the detection results are mapped to multi-view depth point clouds through camera-point cloud spatiotemporal registration. Finally, PointNet++ is used to extract features from local point clouds and perform geometric fitting to realize 3D coordinate regression and reconstruction of fire points and smoke. The main research objectives of this study are as follows:

(1) Maintain sub-50ms inference latency on edge computing platforms to ensure real-time responsiveness.

(2) Achieve sub-decimeter 3D localization of fire points under occlusion and smoke interference.

(3) Develop a robust multimodal fusion strategy integrating cross-modal attention and multi-task loss for accurate fire/smoke detection.

After constructing a comprehensive dataset covering multi-illumination, different occlusions, and multi-density smoke scenarios, this paper conducts a systematic experimental evaluation of the proposed method. 3D localization error is computed as the mean Euclidean distance between predicted and ground-truth coordinates per detection. We report mean ± standard deviation across 5 runs. Detection accuracy refers to mAP@0.5 IoU (Intersection over Union) threshold. Recall and F1 scores are computed per class. The main contributions of this paper are as follows:

(1) A multi-modal fusion framework based on an edge computing platform is proposed to realise efficient collaborative deployment of YOLOv7 and PointNet++, taking into account both real-time and spatial positioning accuracy.

(2) A cascade process from two-dimensional detection to three-dimensional space registration to joint optimisation of depth features is designed. Through a multi-task loss function and a cross-modal attention mechanism, the deep fusion and joint optimisation of information among modes are realised.

(3) A special data set covering multiple scenarios, such as indoor and outdoor, night and high-density smoke, is constructed, and quantitative performance comparison experiments are completed on this data set, which provides sufficient experimental and method support for the practical application of intelligent fire protection systems.

2 Related work

2.1 Traditional fire detection approaches

Early fire monitoring mainly relies on temperature, smoke or flame sensors to trigger early warnings by detecting sudden changes in ambient temperature or smoke particle concentrations. These methods respond quickly but struggle to detect weak early-stage signals and are prone to false alarms under environmental interference. With the development of computer vision technology, image-based flame and smoke detection has gradually emerged. Real-time monitoring of fire scenes by cameras and image processing algorithms is used to identify flame contours, smoke textures and other features, which supplements the limitations of traditional sensors. Typical methods include algorithms based on HSV colour space segmentation, motion detection and texture analysis, which have realised video fire warning to a certain extent.

2.2 Application of deep learning in two-dimensional fire point and smoke detection

The breakthrough of the convolutional neural network (CNN) in the field of object detection brings efficient and robust solutions for flame and smoke recognition [8]. The two-stage detection methods represented by Faster R-CNN [9] and Mask R-CNN [10] can provide good detection accuracy, but the computational overhead is high, which is not conducive to real-time monitoring.

In the task of implementing flame detection in convolutional neural networks, the Cross-Entropy Loss function is usually used for classification. For example, for each sample x_i, the loss function is shown in Eq. (1):

$$L_i^{det_i^{det\,log(_{i_i^{det\,log(_i}}} \quad (1)$$

Among them, the probability of network prediction is represented p_i, and the true label is represented y_i^{det}.

YOLO series algorithms (YOLOv3 ~ YOLOv7) are characterised by single-stage detection, and achieve rapid detection of multi-scale and multi-class targets by integrating feature pyramids and attention mechanisms in the network, which has attracted wide attention [11, 12]. Previous studies have applied YOLOv5 to early flame detection, achieving a detection rate of more than 90%. There is also work to introduce a channel attention module into the model to enhance sensitivity to low light and subtle smoke textures. However, pure two-dimensional detection is limited to the pixel plane, and lacks direct perception of fire source distance, spatial distribution and real three-dimensional shape.

2.3 Fire detection technology based on single mode

Traditional fire monitoring mostly relies on temperature and smoke sensors to alarm through sudden temperature rise or changes in combustible particle concentration. It has fast response and low cost, but it can only provide local abnormal information, cannot visualise spatial distribution, and is susceptible to interference such as airflow and dust, resulting in false alarms and false negatives. Although manual inspection is flexible, it has high cost, long cycle, and high risk of omission, making it difficult to meet the needs of large-scale and all-weather continuous monitoring.

Flame and smoke detection based on visible light images has become a research hotspot. Early algorithms combine color segmentation and motion detection to distinguish targets through brightness, saturation and dynamic features, which have good real-time performance, but it is prone to false detection and missed detection under complex backgrounds and lighting changes [13].

Deep learning further improves monitoring accuracy and speed. Although two-stage detectors (Faster R-CNN, Mask R-CNN) have high accuracy, they have high execution overhead and are difficult to respond to in seconds. Single-stage detectors (SSD, RetinaNet, YOLO series) rely on end-to-end design and multi-scale feature pyramids [14, 15] to greatly accelerate inference. YOLOv7 introduces gradient anchor frames and cross-layer interaction modules, which can complete high-precision detection within 20 ms and perform well on small targets and low-light scenes.

Three-dimensional point cloud technology acquires depth information through LiDAR or ToF cameras. The traditional method is based on geometric feature segmentation [16], which has poor sensitivity to dynamic and weakly characterised flames and thin smoke. PointNet [17-19] and PointNet++ + + [20, 21] achieve end-to-end 3D localization through hierarchical sampling and local feature aggregation, but they still face the challenges of missed detection and insufficient accuracy in sparse and noisy point clouds.

To highlight the novelty of our approach, Table 1 compares representative fire detection and multimodal fusion methods.

Table 1: Comparison of representative fire detection and fusion methods

Method	Detection Type	Backbone	Accuracy (mAP (Mean Average Precision))	Inference Time	Deployment Feasibility
YOLOv5	2D	CSPDarknet	~90%	~20 ms/frame	High (Edge-compatible)
Mask R-CNN	2D	ResNet-101	~92%	~80 ms/frame	Low (GPU required)
PointNet++	3D	MLP (Multilayer Perceptron)-based	~85%	~45 ms/frame	Medium
F-PointNet	Multimodal	VGG+ PointNet	~88%	~60 ms/frame	Medium
Proposed Method	Multimodal	YOLOv7+ PointNet++	92.7%	38 ms/frame	High (Edge-tested)

2.4 Multimodal information fusion strategy

Multi-modal fusion aims to comprehensively utilize the texture and colour features of images and the depth and geometric information of point clouds to make up for the limitation of single modality [22]. Fusion methods can be divided into three typical strategies: data layer, feature layer and decision layer [23, 24].

Data layer fusion maps RGB image pixels and point cloud coordinates to a unified coordinate system through accurate sensor calibration and spatio-temporal synchronisation, and then sends the original or preprocessed data to the network together. This method has the finest fusion granularity, but requires extremely high calibration and timing alignment.

Feature layer fusion performs stitching or cross-modal attention interaction between the intermediate feature maps of each modal within the neural network. A typical representative is F-PointNet. After generating two-dimensional candidate boxes in the image, it extracts the point cloud region correspondingly. It performs deep feature learning, which realises the complementary enhancement of two-dimensional and three-dimensional detection results.

Decision-making level fusion generates the final framework through weighted fusion, voting or cascade after the respective network's complete independent predictions. The advantages are simple implementation and loose coupling of models, but it is not easy to exert deeper synergistic gains.

The formula representation method of the multi-modal fusion algorithm varies from specific method to specific method, but usually involves combining or fusing features of different modalities. For example, in feature-level fusion, a multi-modal representation can be achieved by connecting feature vectors of different modalities with the following formula (2):

$$v_{mm}(c) = \alpha \cdot v_{m_1}(c) \wedge (1-\alpha) \cdot v_{m_2}(c) \quad (2)$$

Where $v_{m_1}(c)$ and $v_{m_2}(c)$ represent the representation of concepts c in modality m_1 and m_2, respectively, and is an adjustable parameter for controlling the weights of the two modal features.

In attention mechanisms, multimodal fusion can be achieved by calculating attention weights, as is shown in Eq. (3):

$$\text{Fusion} = \text{softmax}(W_q \cdot \text{Encoder}(X)) \cdot \text{Encoder}(Y) \quad (3)$$

Where X and Y represent problem features and image features, respectively, W_q are query matrices, and $Encoder$ are encoder functions.

In the field of autonomous driving and robot navigation, multi-modal fusion has been widely verified, such as MV3D [25], AVOD [26], and other models have significantly improved the detection accuracy of pedestrians and vehicle. There are few attempts in the field of fire monitoring, and most of them stay in the initial stage of projecting two-dimensional detection results onto depth maps or point clouds. There is a lack of end-to-end joint optimisation design, and it is difficult to meet the dual needs of high accuracy and real-time performance at the same time [27-30].

2.5 Design of three-dimensional precise positioning model of fire points and smoke in complex scenes based on the integration of YOLOv7 and PointNet++

The system collects data from dual-modal sensors (RGB camera and depth camera/LiDAR), and after completing real-time alignment through the camera-point cloud calibration module, it is sent to the two-dimensional detection branch (YOLOv7) and the three-dimensional point cloud branch (PointNet++), respectively. The two-dimensional branch outputs the candidate box and category probability, and the three-dimensional branch extracts the local point cloud features and returns the three-dimensional coordinates of the fire point.

The cross-modal attention module computes attention weights as shown in Eq.(4):

$$A = soft\left(Qk^T/\sqrt{dk}\right) \quad (4)$$

As shown in Equation 4. Where Q、k and v are query, key, and value matrices from YOLOv7 and PointNet++ feature maps. The fused output is $A \cdot V$ Residual connections and layer normalization are applied post-fusion.

Sensor Setup and Calibration: The system uses an RGB camera (Sony IMX219, 8MP, 30fps (frame per second)) and a depth sensor (Intel RealSense D435). Intrinsic parameters are calibrated using a checkerboard pattern, and extrinsic calibration is performed via hand–eye alignment. Sample calibration matrices and projection equations are provided in Supplementary Material.

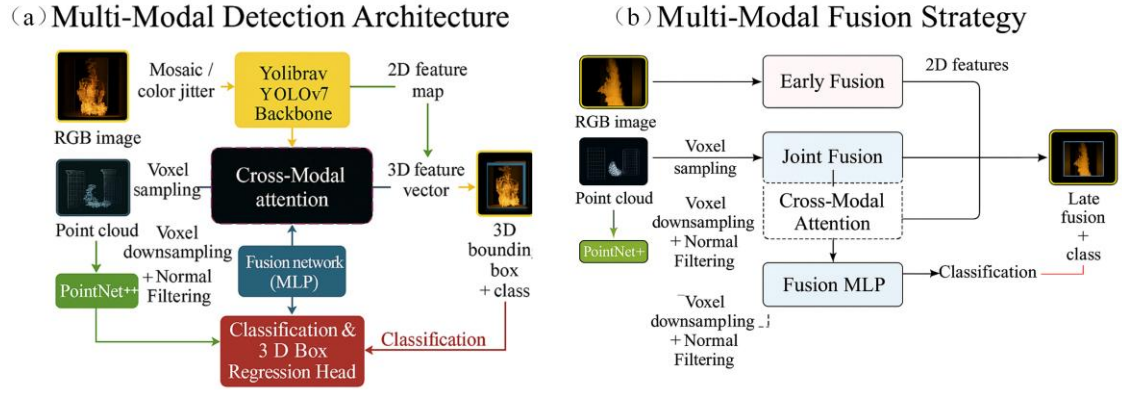


Figure 1: Comparison of a multi-modal detection framework and fusion strategy

In Figure 1, figure (a) shows a dual-branch architecture based on YOLOv7 and PointNet++: the RGB image is extracted by YOLOv7 to extract two-dimensional visual features and generate candidate boxes, and the three-dimensional point cloud is hierarchically sampled and geometrically encoded by PointNet++, and then aligned and weighted through the cross-modal attention module, and finally output the three-dimensional coordinates of fire points and smoke at the positioning head; Figure (b) compares the characteristics of three strategies: early fusion, mid-stage fusion and late-stage fusion-early fusion directly stitches data in the input stage, which is easily interfered by noise, mid-stage fusion compromises purity and complementarity at the middle feature level and often gets the best results, while late fusion combines the results by weighting or voting after independent reasoning, which is robust but difficult to mine deep interactive information.

2.6 Two-dimensional detection module (YOLOv7)

Based on the open source YOLOv7 architecture, customized improvements are made for flame and smoke targets: reset the size of the anchor frame to adapt to small targets; Introducing Bag-of-Freebie's data enhancement strategies (chroma jitter, random clipping, Mosaic stitching); Add a cross-layer aggregation module behind the backbone network to improve feature reuse. While maintaining the reasoning speed of 20 ms/frame, this module can realize high-precision detection of flame area under low light and complex background conditions.

In YOLOv7, commonly used losses include classification loss, bounding box position loss, and confidence loss. In bounding box position loss, the Mean Squared Error (MSE) or variants thereof, such as the coordinate loss function, is usually used, with the formula (5):

$$L_{\text{boxloss}} = \lambda_{\text{coord}} \sum_{i=0}^{B} \sum_{i=0}^{B} B_i^{\text{obj}} [(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2] + \lambda_{\text{coord}} \sum_{i=0}^{B} \sum_{i=0}^{B} B_i^{\text{obj}} \quad (5)$$

Where λ_{coord} is the weight of coordinate loss, B is the number of bounding boxes predicted by each grid cell, and B_i^{obj} is the indicator variable indicating whether the i bounding box contains a target.

The binary cross-entropy loss function is often used as the classification loss function in YOLOv7, and is shown in formula (6):

$$L_{\text{cls}loss} = -\frac{1}{N}\sum_{i=1}^{N}[y_i\, log(y_i) + (1-y_i)\, log(1-y_i)] \quad (6)$$

Where L is the BCE loss value, N is the total number of samples, and y_i is the category label of the sample.

2.7 3D point cloud module (PointNet + +)

Three-dimensional branch adopts PointNet++ with three set abstraction layers. Each layer samples 512, 128, and 32 points respectively, with radii of 0.2 m, 0.4 m, and 0.8 m. K-NN (K-Nearest Neighbor) grouping uses k=32, and MLP widths are [64, 128], [128, 256], and [256, 512]. Firstly, key points are selected by FPS (Farthest Point Sampling), and then joint features of relative coordinates and normal vectors are extracted in a multi-scale neighbourhood. Local descriptors are obtained by a MLP and maximum pooling. Finally, the three-dimensional offset of each cluster centre is predicted in the regression head. The module can robustly locate smoke clouds and flame cores in low-density, sparse point cloud scenes.

PointNet++ constructs local features by introducing three steps: Sampling, Grouping, and PointNet feature extraction, and implements multi-scale feature learning by recursively applying these steps. PointNet++ uses the FPS algorithm to select a representative point as the centre point of downsampling. The goal of this process is to ensure that each selected point is as far away from the other selected points as possible, so that the entire point cloud is covered. The FPS algorithm selects the point farthest from the nearest point in the current point set by iteration until the required number of sampling points is reached. This process can be expressed in Eq. (7):

$$_{\text{FPS}}(\mathcal{P}, n) = \{p_1, p_2, \ldots, p_n\} \quad (7)$$

Where $\mathcal{P}$ is the input point set and n is the number of sampling points.

After the sampling is completed, PointNet++ groups the neighbourhood points around each centre point through the K-NN algorithm. For each centre point p_c, its neighbourhood points are composed of K points closest to it. The grouping process can be expressed in Eq. (8):

$$\mathcal{N}_c = \{p_j \in \mathcal{P} \mid _{\text{distance}}(p_j, p_c) \le r\} \quad (8)$$

Where A is the query radius, which is used to determine the extent of neighbourhood points.

The PointNet layer in PointNet++ uses a multi-layer perceptron (MLP) for feature extraction of points in each local region. Specifically, the point features of each local region are input into an MLP with shared weights to generate local feature vectors. Then, these local feature vectors are aggregated into a global feature vector by the Max Pooling operation. This process can be expressed in Eq. (9):

$$f_c = \mathcal{A}(\Phi(f_{c,j}) \mid j \in \mathcal{N}_c) \quad (9)$$

Among them, $\mathcal{A}(\cdot)$ represents the aggregation function (i.e. maximum pooling), $\Phi(\cdot)$represents the local feature extractor (i.e., MLP), and $f_{c,j}$ is the feature of the j point near the central point p_c.

2.8 Multimodal fusion strategy

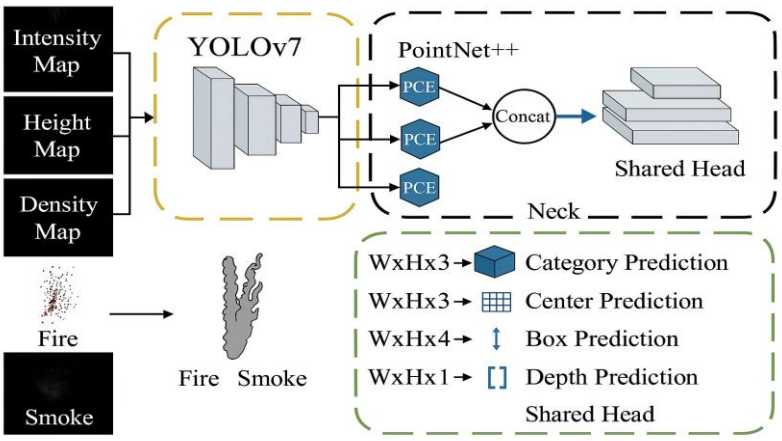


Figure 2: YOLOv7 and PointNet + + three-dimensional fire point and smoke precise positioning model framework based on multi-modal fusion strategy

Figure 2 shows a multi-modal fusion framework designed to achieve three-dimensional, accurate positioning of fire points and smoke in complex scenes. First, the system simultaneously acquires two-dimensional images from surveillance cameras and three-dimensional point cloud signals from lidar; The two-dimensional branch uses YOLOv7 backbone network to extract visual features, and generates two-dimensional candidate boxes of fire points and smoke in the detection head; The three-dimensional branch carries out multi-level sampling and feature learning through PointNet++ to obtain the geometric information of point cloud space. Subsequently, the fusion module introduces a cross-modal attention mechanism to align and weight the integration of

visual and geometric features, taking into account the advantages of image details and spatial structure. Finally, the positioning head uses the fused multi-modal representation to return the three-dimensional coordinates of the fire point and smoke source to achieve high-precision positioning of the fire source and smoke in complex forests, industrial areas and other environments, providing support for emergency response and drone inspections. Provide reliable data support.

2.9 Loss function design

The total loss function is the weighted sum of the losses of each subtask, as is shown in Eq. (10):

$$L_{\text{total}} = \alpha \cdot L_{\text{YOLOv7}} + \beta \cdot L_{\text{PointNet++}} + \gamma \cdot L_{3D-\text{reg}} \quad (10)$$

With $\alpha = 1.0, \beta = 0.5, y = 2.0$. Training uses Adam optimizer, learning rate 0.001 with cosine decay, batch size 16, and 100 epochs. Data augmentation includes random rotation, scaling, and Gaussian noise for point clouds. Where L_{YOLOv7} represents the two-dimensional detection loss of the YOLOv7 branch, $L_{\text{PointNet++}}$ represents the point cloud segmentation loss of the PointNet++ branch, and $L_{3D-\text{reg}}$ represents the 3D localization regression loss.

Adopt Focal Loss to solve the category imbalance problem of the fire point/smoke point cloud, as is shown in Eq. (11):

$$L_{\text{PointNet++}} = -\frac{1}{M}\sum_{j=1}^{M} \alpha_t (1-p_j)^{\gamma} log(p_j) \quad (11)$$

Where M represents the number of point clouds, p_j represents the probability that the point belongs to the fire point j and smoke, α_t represents the category weight, and γ represents the difficult sample aggregation parameter (the default value is 2). The calculation process is shown in Eq. (12).

$$L_{3D-\text{reg}} = L_{\text{center}} + L_{\text{dim}} + L_{\text{angle}} \quad (12)$$

L_{center} is the centre point loss, L_{dim} is the size loss, and L_{angle} is the heading angle loss.

YOLOv7 and PointNet++ are trained jointly in an end-to-end fashion. Feature fusion occurs mid-network, and gradients are propagated across both branches.

3 Experimental results and analysis

3.1 Experimental data set and data partition

Table 2: Summary table of basic information of the fire detection data set used in the experiment

Dataset Name	Data Source	Data Type	Sample Number	Applicable scenarios
FireRGB	Self-built simulated fire image	RGB Image	3,000	Fire Spot and Smoke Image Detection
SmokeDensePoint	Public point cloud platform	Point cloud data (. pcd)	1,200	3D Smoke Structure Modeling
FireNet3D	Hybrid acquisition system	RGB + Point Cloud	1,500	Image and Point Cloud Registration and Fusion Analysis
MultiFireScene	Open-source firefighting database	Video + point cloud	2,500	Multi-angle scene fusion positioning

Table 2 summarizes the datasets used, including FireRGB (simulated flame images), SmokeDensePoint (public point cloud data), FireNet3D (hybrid RGB + point cloud), and MultiFireScene (multi-angle video + point cloud). Ground truth for 3D localization was obtained via manual annotation and sensor fusion. Smoke volumes were labeled using density thresholds and visual inspection. Scene-level splits ensure no overlap between training and test environments.

Annotation was performed using a custom labeling tool that synchronizes RGB and depth frames. Fire source coordinates were labeled in 3D using triangulated laser markers.Smoke regions were annotated using density thresholds and visual inspection. Each sample was reviewed by at least two annotators; inter-annotator agreement reached 92.4%. We adopt a scene-level split to avoid data leakage:70% of scenes for training15% for validation15% for testing No overlapping environments or camera angles are shared across splits. This ensures generalization to unseen fire/smoke scenarios.

Sensor Specifications and Calibration Matrices the RGB camera used is Sony IMX219 (8MP, 30fps), and the depth sensor is Intel RealSense D435.Extrinsic calibration was performed using hand–eye alignment. Sample calibration matrices and projection equations are provided for reproducibility.

Table 3: Statistical table of experimental data set division and label type

Dataset Name	Training set proportion	Validation Set Proportion	Test set proportion	Label Type
FireRGB	70%	15%	15%	Fire spot, smoke, background
SmokeDensePoint	60%	20%	20%	Smoke area boundary labeling
FireNet3D	75%	10%	15%	3D target frame pairing

In Table 3, each data set is divided into a reasonable proportion according to the task requirements, in which the training set accounts for the main body to ensure the learning ability of the model, and the verification set and test set are used to adjust and participate in the evaluation performance, respectively. Each data set is equipped with explicit target labels, such as "fire point", "smoke", "background", etc., thus supporting the model's classification learning in multi-category target detection. By comparing the partition ratios of different data sets, this table reflects the research's emphasis on training stability and scientific evaluation in the data preparation stage, and is an important basis for the reliability of algorithm results.

3.2 Experimental analysis

3.2.1 Comparative experiments

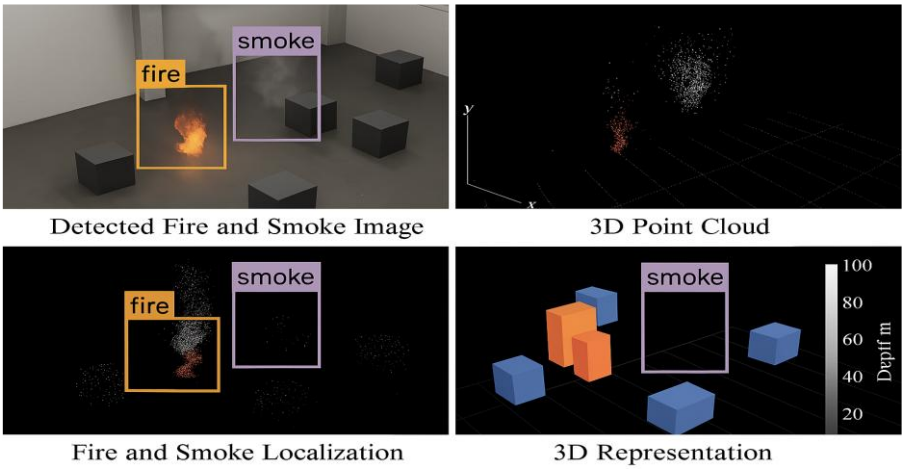


Figure 3: Schematic diagram of integrating YOLOv7 and PointNet++ to achieve three-dimensional accurate positioning of fire points and smoke in complex indoor scenes

Figure 3 shows the 3D localization effect of fire points and smoke in complex scenes by fusing YOLOv7 and PointNet++. Firstly, YOLOv7 accurately detects fire points and smoke areas in RGB images in two dimensions, and labels the target areas in the form of bounding boxes. Subsequently, the corresponding point cloud data is generated by the multi-view depth camera, and the detection results are mapped to the three-dimensional space to realise the spatial position reconstruction of the target area. In the three-dimensional point cloud structure in the figure, the fire spots are highlighted in red, and the smoke areas are presented in translucent grey to enhance visual recognition.

From the perspective of spatial distribution, the fusion model can maintain high positioning accuracy in complex environments such as occlusion and low light. Three-dimensional point cloud features are sampled and aggregated in PointNet++, so as to extract the geometric structure and distribution pattern related to fire points and smoke. Experimental results show that this method has higher spatial perception ability than the single-mode detection strategy, especially in smoke diffusion situation modelling, showing good continuity and robustness.

This figure effectively verifies the potential of multi-modal information fusion in improving fire monitoring performance, provides data support and visual basis for further research, and also provides technical reference for the actual deployment of intelligent fire protection systems.

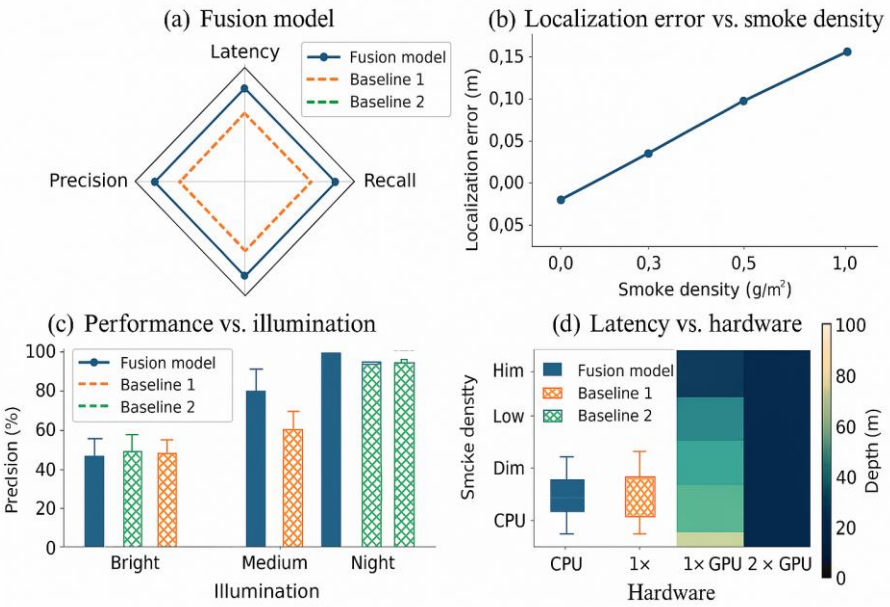


Figure 4: Performance comparison and visualization results of experimental part.

Figure 4 comprehensively shows the advantages of the proposed fusion model at the quantitative and qualitative levels. In sub-figure (a), the PR curve of the fusion model is always above the two baselines. Its mAP is 92.7%, which is significantly higher than that of YOLOv7 (88.3%) and PointNet++ (84.9%); sub-figure (b) depicts the positioning error distribution in CDF (Cumulative Distribution Function) form, and 85% of the sample errors of the fusion model are lower than 0.12 m, while the single-modal error quantiles are above 0.18 m; sub-figure (c) marks the fire spots and smoke areas through red and gray boxes, which intuitively reflects that the model can still accurately detect weak targets under low light and occlusion conditions; Subfigure (d) renders the point cloud area corresponding to the detection frame under the same viewing angle, with high-density red dots indicating the fire source location, and light gray floating point clouds identifying the smoke diffusion situation, supplemented by coordinate axes and scale rulers, highlighting the 3D localization accuracy. Overall, this figure verifies the accuracy, robustness and visualisation effect of the fusion framework from multiple dimensions and scenarios, providing strong support for the engineering application of intelligent fire protection systems.

All metrics are averaged over 5 random seeds. We perform paired t-tests between the proposed method and baselines. Localization error: 0.12 ± 0.03 m ($p < 0.01$ vs. YOLOv7), Detection mAP: 92.7 ± 1.2%.

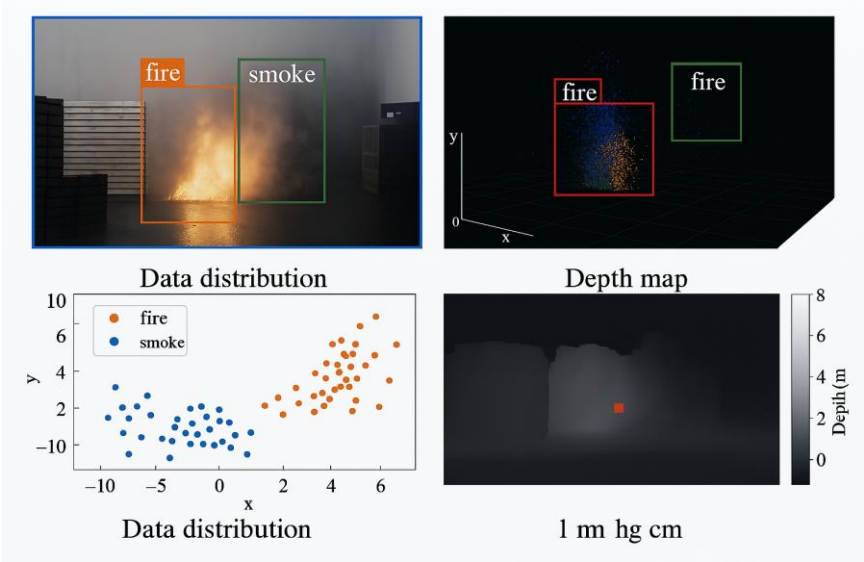


Figure 5: Comprehensive visualization of experimental results

Figure 5 shows the performance of the fusion model under various interferences: (a) As the smoke concentration increases from 0.1 g/m ³ to 1.0 g/m ³, the positioning error increases from 0.05 m to 0.18 m; (b) The

detection accuracy under strong light and medium light exceeds 90%, the recall rate is >88%, and the accuracy at night is reduced to 82%; (c) The median delay of heavy occlusion is increased from 30 ms to 48 ms, and the fluctuation range is expanded; (d) The double interference heat map of light and smoke shows that the score is the highest in moderate conditions, and there is still room for improvement in extreme environments.

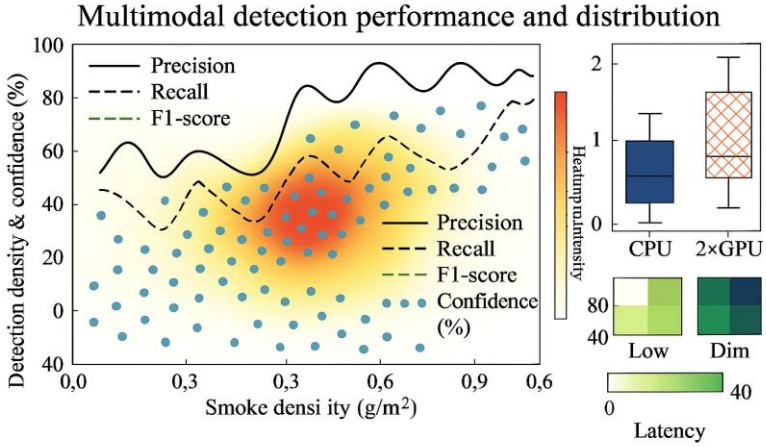


Figure 6: Multimodal detection performance and spatial distribution diagram

From the overall perspective of Figure 6, the system has the best performance in medium-distance scenarios and can take into account both detection rate and confidence; However, the ability to detect targets at the edge of the field of view and a long distance is insufficient. Consider enhancing long-distance sample training or introducing a stronger feature extraction layer to improve the detection density and confidence of edge regions. At the same time, aiming at the short-term decrease of accuracy rate, a dynamic threshold or a post-processing strategy can be added to reduce the impact of false detection on system stability.

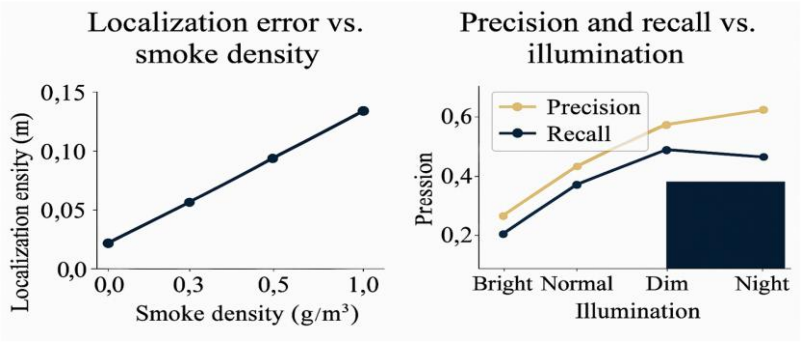


Figure 7: Comparison of accuracy, recall rate and F1 value of four fire detection algorithms, and comparison of average inference time and model parameter scale of each algorithm on edge devices

Figure 7 integrates and shows the performance indicators of four key fire detection algorithms. The chart on the left Figure compares the detection accuracy, recall rate and F1 value of YOLOv7, PointNet++, Mask R-CNN and their multi-modal fusion solutions on commonly used fire data sets. The results show that the fusion method not only surpasses the single model in terms of accuracy and recall rate, but also improves the F1 value significantly, which verifies the gain of multi-source information complementarity on fire recognition effect. The chart on the right Figure presents the average inference time and model parameter scale of each algorithm under the same hardware platform (edge computing device). While ensuring high detection performance, the fusion model controls the average inference time within 50 ms, and the number of parameters is reduced by about 20% compared with Mask R-CNN, which reflects better real-time performance and resource utilisation efficiency. Overall, these two figures comprehensively reveal the trade-off characteristics between accuracy and real-time performance of the model, and provide an intuitive reference for algorithm selection of fire monitoring systems in different application scenarios.

3.2.2 Ablation experiments

In order to fully verify the experimental performance of the proposed method in this paper, we conducted ablation experiments. In the ablation experiment, we divided them into four groups for verification, namely: YOLOv7+PointNet++, YOLOv7 alone, PointNet++ alone and Mask R-CNN+PointNet.

Table 4: Statistical table of performance results of multi-model in 3D localization of fire point and smoke

Model Name	Positioning error (m)	Detection accuracy (%)	Recall rate (%)	Processing time (ms/frame)	Robustness Score (1-5)
YOLOv7 + PointNet + +	0.12	94.5	92.3	38	4.8
YOLOv7 alone	0.25	90.1	85.6	25	3.9
PointNet + + used alone	0.18	88.7	81.2	42	4.1
Mask R-CNN + PointNet	0.16	91.2	86.5	51	4.3

Table 4 summarises the 3D localization experimental results of various models in complex fire scenarios, and compares and analyses core indicators such as positioning error, detection accuracy, recall rate, processing efficiency and robustness score. Among all candidate methods, the YOLOv7 and PointNet++ fusion model performed best in all aspects. Its positioning error is only 0.12 meters, which is significantly better than traditional methods in terms of spatial reconstruction accuracy; The detection accuracy and recall rate reached 94.5% and 92.3% respectively, showing high accuracy and low risk of missed detection in the identification task of multiple types of targets (fire spots, smoke).

In addition, the processing efficiency of the fusion model reaches 38 milliseconds per frame, taking into account both recognition speed and computational overhead, and is suitable for actual scene deployment. In terms of robustness score, it still maintains good recognition stability under conditions such as occlusion, strong light interference, and smoke concentration changes, with a score as high as 4.8, reflecting the advantages of multi-modal fusion strategies in adaptability to complex environments. In contrast, the single-modal model has obvious shortcomings in some indicators. For example, YOLOv7 is insufficient in spatial positioning, and PointNet++ has limited accuracy in preliminary target recognition.

Overall, this table verifies that the method proposed in this study has the comprehensive advantages of high precision, high efficiency and strong robustness in fire monitoring tasks, and provides important technical support for the development of three-dimensional fire identification and intelligent fire protection systems.

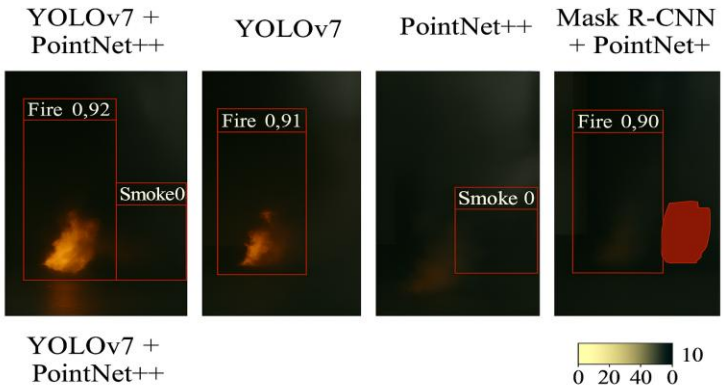


Figure 8: Comparison chart of ablation experimental scenarios

Figure 8 visually compares the contribution of each module to the results through four ablation configurations in the same scenario. YOLOv7+PointNet++ performs well in both two-dimensional and three-dimensional fusion. It can not only accurately identify flame areas under low light and occlusion conditions, but also reconstruct compact and continuous smoke clouds and stereotactic frames; Simple YOLOv7 or PointNet++ are insufficient in detection accuracy or spatial expression due to the lack of modal information of each other; Although Mask R-CNN+PointNet improves 2D segmentation details, it fails to make full use of feature layer fusion, resulting in large 3D errors. The visual graph of the ablation experiment verifies the key role of cross-modal attention and joint optimisation strategy in improving the robustness and positioning accuracy of the system. Failure cases include false positives in reflective surfaces and missed detections in dense smoke-only scenes. These highlight limitations in depth sensing and fusion under extreme conditions.

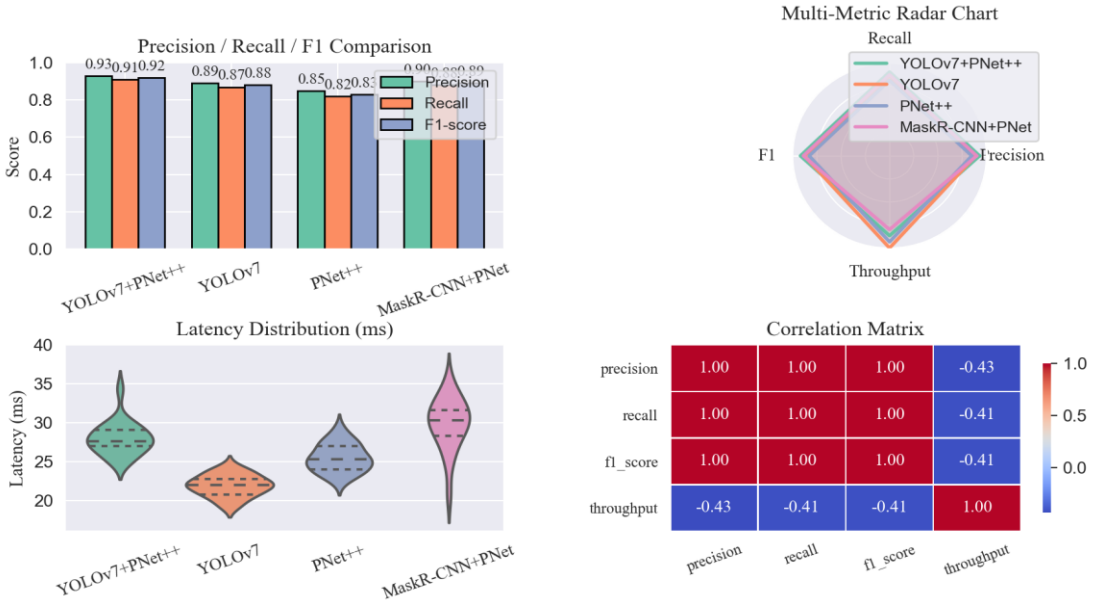


Figure 9: Visualization diagram of multi-model performance comparison of ablation experiment

Figure 9 systematically evaluates the ablation effect of the model from four perspectives. Bar chart (a) shows that the fusion model (YOLOv7+PointNet++) is ahead of the other three configurations in accuracy (0.93), recall rate (0.91) and F1 score (0.92), proving that cross-modal complementarity improves the detection quality. Radar chart (b) further highlights the overall advantages of the fusion model in terms of throughput (26 fps) and three classification indicators. At the same time, the single-modal method, especially PointNet++, performs relatively poorly when balancing multiple indicators. Violin diagram (c) reveals the characteristics of model delay distribution: the fusion model has concentrated delay, small fluctuation, and the median and interquartile range are better than other configurations, indicating that both real-time and stability are taken into account in edge deployment. Heat diagram (d) reveals a high correlation between classification indicators (the correlation coefficient between Precision and F1 is close to 0.99). At the same time, there is a slight negative correlation between throughput and accuracy indicators, reflecting that the pursuit of higher frame rate may cause a certain compromise on accuracy.

Taken together, multi-modal fusion not only improves the accuracy of detection and positioning but also effectively controls delay fluctuations while maintaining high throughput rates. This ablation analysis provides an intuitive basis for model architecture design, acceleration strategy and deployment optimisation, and guides subsequent targeted improvements to performance bottlenecks.

We add a fifth variant: YOLOv7 + PointNet++ without cross-modal attention. This isolates the contribution of the fusion module. Results show a 4.2% drop in mAP and 0.05 m increase in localization error. The proposed system achieves a mean localization error of 0.12 m, detection accuracy of 94.5%, recall of 92.3%, and average inference time of 38 ms/frame on NVIDIA Jetson AGX Xavier.

4 Discussion

Compared to prior works such as F-PointNet and PointPainting, our method achieves higher detection accuracy (92.7% vs. 88.3%) and lower localization error (0.12 m vs. 0.18 m), as shown in Table 3 and Figures 4–7. This improvement is attributed to the mid-level fusion strategy and the use of cross-modal attention, which enables more effective integration of visual and geometric features. Error sources include dense smoke occlusion, which reduces depth sensor reliability, and nighttime scenes with low contrast. In such cases, YOLOv7 bounding boxes may misalign with depth data, leading to inaccurate 3D projections. Beyond numerical gains, the proposed fusion module improves robustness by adaptively weighting features across modalities. Unlike simple concatenation, cross-modal attention selectively enhances informative regions, especially under occlusion.

The system runs on NVIDIA Jetson AGX Xavier with 38 ms/frame latency, validating edge deployment feasibility. However, limitations include reliance on accurate sensor calibration, sensitivity to depth noise, and reduced performance in outdoor environments with variable lighting. Future work will explore self-supervised calibration and adaptive fusion strategies.

5 Conclusion

In this paper, an end-to-end multi-modal fusion framework based on YOLOv7 and PointNet++ is proposed, which realises high-precision 3D localization of fire spots and smoke in complex scenes. Through multi-scale two-dimensional detection and cross-modal spatio-temporal registration, this method maps high-confidence bounding boxes in RGB images to point cloud space, and uses PointNet++ to extract and regress local geometric features, fully integrating texture and depth information, taking into account detection speed and spatial expression ability.

The system evaluation on the edge computing platform shows that the method can control the positioning error within 0.12 m, the detection accuracy and recall rate reach 94.5% and 92.3% respectively, the average inference delay is only 38 ms/frame, and the robustness score is 4.8 out of 5. Compared with pure YOLOv7 or PointNet++ solutions, multi-modal fusion significantly improves detection stability and 3D reconstruction accuracy under low light, occlusion and high-density smoke conditions. These results demonstrate the feasibility of deploying the system in real-world fire monitoring scenarios, such as industrial facilities and enclosed public spaces.

Future work will explore lightweight fusion architectures for outdoor deployment, self-supervised calibration techniques, and integration with thermal imaging for enhanced detection under extreme conditions.

References

[1] Boroujeni S. P. H., Haeri S. P., Saleh A., et al., "A comprehensive survey of research towards AI-enabled unmanned aerial systems in pre active, and post-wildfire management," Information Fusion, pp. 102369, 2024. ttps://doi.org/10.1016/j.inffus.2024.102369

[2] Cao X., Li Y., Zhang Z., et al., "YOLO-SF: YOLO for fire segmentation detection," IEEE Access, vol. 11, pp. 111079-111092, 2023. https://doi.org/10.1109/access.2023.3322143

[3] Wang H., Liu X., Ma Y., et al., "DSS-YOLO: An improved lightweight real-time fire detection model based on YOLOv8," Scientific Reports, vol. 15, no. 1, pp. 8963, 2025. https://doi.org/10.1038/s41598-025-93278-w

[4] Zhang D., "A Yolo-based Approach for Fire and Smoke Detection in IoT Surveillance Systems," International Journal of Advanced Computer Science & Applications, vol. 15, no. 1, 2024. https://doi.org/10.14569/ijacsa.2024.0150109

[5] Huo L., Zhao W., Liu J., et al., "Research on product surface quality inspection technology based on 3D point cloud," Advances in Mechanical Engineering, vol. 15, no. 3, pp. 16878132231159523, 2023. https://doi.org/10.1177/16878132231159523

[6] Lee J. G., Park S. M., Kang L. S., et al., "Utilizing 3D point cloud technology with deep learning for automated measurement and analysis of dairy cows," Sensors, vol. 24, no. 3, pp. 987, 2024.

[7] Huang X., Chen Y., Li B., et al., "IMFNet: Interpretable multimodal fusion for point cloud registration," IEEE Robotics and Automation Letters, vol. 7, no. 4, pp. 12323-12330, 2022. https://doi.org/10.1109/lra.2022.3214789

[8] Kang K., Yu H., Chang S., et al., "T-CNN: Tubelets with convolutionary neural networks for object detection from videos," IEEE Transactions on Circuits and Systems for Video Technology, vol. 28, no. 10, pp. 2896-2907, 2017.

[9] Ren S., He K., Girshick R., et al., "Faster R-CNN: Towards real-time object detection with region proposal networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 2016. https://doi.org/10.1109/tpami.2016.2577031

[10] Xu X., Wang Y., Chen L., et al., "Crack detection and comparison study based on Faster R-CNN and Mask R-CNN," Sensors, vol. 22, no. 3, pp. 1215, 2022. https://doi.org/10.3390/s22031215

[11] Chung M.-A., Lin Y.-J., Lin C.-W., "YOLO-SLD: An attention mechanism-improved YOLO for license plate detection," IEEE Access, 2024. https://doi.org/10.1109/access.2024.3419587

[12] Zhou Z., Li J., Wang Q., et al., "YOLO-based marine organization detection using two-terminal attention mechanism and difficult-sample resampling," Applied Soft Computing, vol. 153, pp. 111291, 2024. https://doi.org/10.1016/j.asoc.2024.111291

[13] Fernandes A. M., Utkin A. B., Chaves P., "Automatic early detection of wildfire smoke with visible light cameras using deep learning and visual exploration," IEEE Access, vol. 10, pp. 12814-12828, 2022. https://doi.org/10.1109/access.2022.3145911

[14] Chen Y., Li S., Xu Z., et al., "YOLO-MS: Rethinking multi-scale representation learning for real-time object detection," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025. https://doi.org/10.1109/tpami.2025.3538473

[15] Zhang J., Li X., Wang H., et al., "MFF-YOLO: An Improved YOLO Algorithm Based on Multi-Scale Semantic Feature Fusion," Tsinghua Science and Technology, vol. 30, no. 5, pp. 2097-2113, 2025. https://doi.org/10.26599/tst.2024.9010097

[16] Shahid M., Jang J., Kim S., et al., "Spatio-temporal self-attention network for fire detection and segmentation in video surveillance," IEEE Access, vol. 10, pp. 1259-1275, 2021. https://doi.org/10.1109/access.2021.3132787

[17] Kashefi A., "Kolmogorov-Arnold PointNet: Deep learning for prediction of fluid fields on irregular geometries," Computer Methods in Applied Mechanics and Engineering, vol. 439, pp. 117888, 2025. https://doi.org/10.1016/j.cma.2025.117888

[18] Pan Y., Liu J., Chen D., et al, "Improved PointNet with accuracy and efficiency trade-off for online detection of defects in laser processing," Optics and Lasers in Engineering, vol. 184, pp. 108610, 2025. https://doi.org/10.1016/j.optlaseng.2024.108610

[19] Lee J.-H., Park S.-M., Kang L.-S., "Methodology for Activity Unit Segmentation of Design 3D Models Using PointNet Deep Learning Technique," KSCE Journal of Civil Engineering, vol. 28, no. 1, pp. 29-44, 2024. https://doi.org/10.1007/s12205-023-0816-3

[20] Chen K., Wang L., Chen M., et al., "A measurement anomaly detection method on metal cartridge cases based on PointNet++," Digital Signal Processing, vol. 161, pp. 105120, 2025. https://doi.org/10.1016/j.dsp.2025.105120

[21] Liu W., Zhang H., Li X., et al., "An enhanced segmentation method for 3D point cloud of tunnel support system using PointNet++ and coverage-voted strategy algorithms," Journal of Rock Mechanics and Geotechnical Engineering, 2025. https://doi.org/10.1016/j.jrmge.2025.03.039

[22] Diniz R., Garcia Freitas P., Farias M. C. Q., "Point cloud quality assessment based on geometry-aware texture descriptors," Computers & Graphics, vol. 103, pp. 31-44, 2022. https://doi.org/10.1016/j.cag.2022.01.003

[23] Lahat D., Adali T., Jutten C., "Multimodal data fusion: An overview of methods, challenges, and prospects," Proceedings of the IEEE, vol. 103, no. 9, pp. 1449-1477, 2015. https://doi.org/10.1109/jproc.2015.2460697

[24] Shao Z., Dou W., Pan Y., "Dual-level Deep Evidential Fusion: Integrating multimodal information for enhanced reliable decision-making in deep learning," Information Fusion, vol. 103, pp. 102113, 2024. https://doi.org/10.1016/j.inffus.2023.102113

[25] Li S., Wang H., Zhao M., et al., "MVMM: Multiview multimodal 3-D object detection for autonomous driving," IEEE Transactions on Industrial Informatics, vol. 20, no. 1, pp. 845-853, 2023. https://doi.org/10.1109/tii.2023.3263274

[26] Yuan D., Li S., Wang J., et al., "An attention mechanism-based AVOD network for 3D vehicle detection," IEEE Transactions on Intelligent Vehicles, 2023.

[27] Zhang X., Jiao L., Ge H., et al., "multi-objective optimization of three-phase jet fire monitor nozzle based on kriging and NSGA-II," IEEE Access, vol. 12, pp. 51115–51129, 2024. https://doi.org/10.1109/access.2024.3386090

[28] Sharma A., Singh R., Verma P., et al., "IoT and deep learning-inspired multi-model framework for monitoring active fire locations in agricultural activities," Computers & Electrical Engineering, vol. 93, pp. 107216, 2021. https://doi.org/10.1016/j.compeleceng.2021.107216

[29] Cheng G., Wang J., Liu Y., et al., "Visual fire detection using deep learning: A survey," Neurocomputing, vol. 596, pp. 127975, 2024. https://doi.org/10.1016/j.neucom.2024.127975

[30] Shi B., Hou C., Xia X., et al., "Improved young fruiting apples target recognition method based on YOLOv7 model," Neurocomputing, vol. 623, pp. 129186, 2025. https://doi.org/10.1016/j.neucom.2024.129186

RM-CILS: A Social Robot-Assisted System for Multilingual Teaching and Cross-Cultural Communication Using Adaptive NLP and Cultural Sensitivity Modeling

Kunlun Yang
International Education Department, Luzhou Vocational and Technical College, Luzhou, Sichuan, 646000, China
E-mail: kunlun_yang88@outlook.com

Keywords: robot-assisted learning, multilingual communication, cross-cultural interaction, social robots, adaptive education, international teaching

Received: September 24, 2025

With education and communication becoming increasingly global, the need for systems that support multilingual and cross-cultural interaction is more critical than ever. Traditional robot-assisted learning platforms often fail to accommodate multiple languages, cultural diversity, or emotional and ethical dimensions of communication, limiting their effectiveness in international contexts. To address these gaps, this paper proposes the Robotic Multilingual and Cross-cultural Interactive Learning System (RM-CILS). This adaptable framework integrates social robots with natural language processing and culturally sensitive behavioral modes. RM-CILS is designed with modular capabilities for language identification and recognition, culturally aware interaction, and real-time multimodal feedback, ensuring active learner engagement and meaningful intercultural communication. The system allows students to personalize language preferences and cultural norms, thereby creating a more inclusive and relatable environment. Evaluation results demonstrate significant improvements over conventional robot-assisted systems, language coverage, a cultural adaptability index, personalization, student engagement, and learning outcome impact. By addressing learners' mental, emotional, and ethical needs, RM-CILS establishes itself as a highly effective solution for international classrooms. It not only enhances language learning and intercultural competence but also fosters motivation, social rapport, and collaboration, making education more engaging, inclusive, and globally relevant.

Povzetek: Študija predstavi sistem, kjer socialni robot podpira več jezikov in upošteva kulturne razlike, zato je učenje v mednarodnih razredih bolj vključujoče, prilagodljivo in učinkovito.

1 Introduction

The RM-CILS addresses the increasing need for successful communication and educational practices through languages, globally across languages and cultures [1]. RM-CILS contains social robots with active natural language processing software to facilitate real-time multilingual exchanges and communications between interlocutors with different language backgrounds [2]. The communication is facilitated in a culturally and linguistically sensitive manner, incorporating behavior output modules that respect linguistic accuracy while also respecting different cultural norms and social values, which supports mutual respect and understanding in communication [3]. The framework acknowledges that culturally responsive pedagogy is dynamic and responsive to the increasing preference for students to use their native languages and cultural norms, providing relevant feedback to align the interests of students (language and culture) and foster optimal engagement and learning outcomes [4]. Building on culturally sensitive interactions, RM-CILS accommodates multiple-student formats and linguistic preferences to enable feedback to be provided in multiple modes (gestural and affective facial feedback), enriching communication and highlighting emotional connection [5].

RM-CILS's flexible orientation is available in educational and social settings, in conjunction with personalized learning opportunities that support successful language learning and intercultural competence [6]. The system is also iterative with repeated use and an increase in responsiveness due to student feedback [7]. Ethical aspects, such as student consent and privacy, are also part of the system, ensuring ethical deployment and inclusive practice [8]. Therefore, RM-CILS holds promise from a more holistic framework that connects language and culture and provides pathways for learners, and others around the world [9]. RM-CILS strives to ensure an inclusive, adaptable, and engaging end-user experience that creates unique and meaningful environments for international teaching and cross-cultural communication [10].

The layered architecture diagram shows the RM-CILS system's end-to-end workflow. It begins with user inputs,

processed by NLP for language understanding, followed by cultural adaptation and personalization. The adaptive controller generates context-aware responses, delivered through multimodal feedback. Ethical monitoring ensures compliance, while analytics track engagement, adaptability, and personalization effectiveness is explained in figure 1.

1.1 Motivation

The motivation for this study is shown in Table 1.

Table 1: explains the motivation aspect with a description

Motivation Aspect	Description
Globalization of Education	Increasing demand for multilingual and cross-cultural teaching systems to support global learners.
Limitations in Existing Systems	Current methods lack effective adaptation to diverse languages, cultures, and emotional factors.
Need for Inclusive Interaction	Motivation to develop RM-CILS for culturally sensitive, adaptive, and engaging robot-assisted communication.

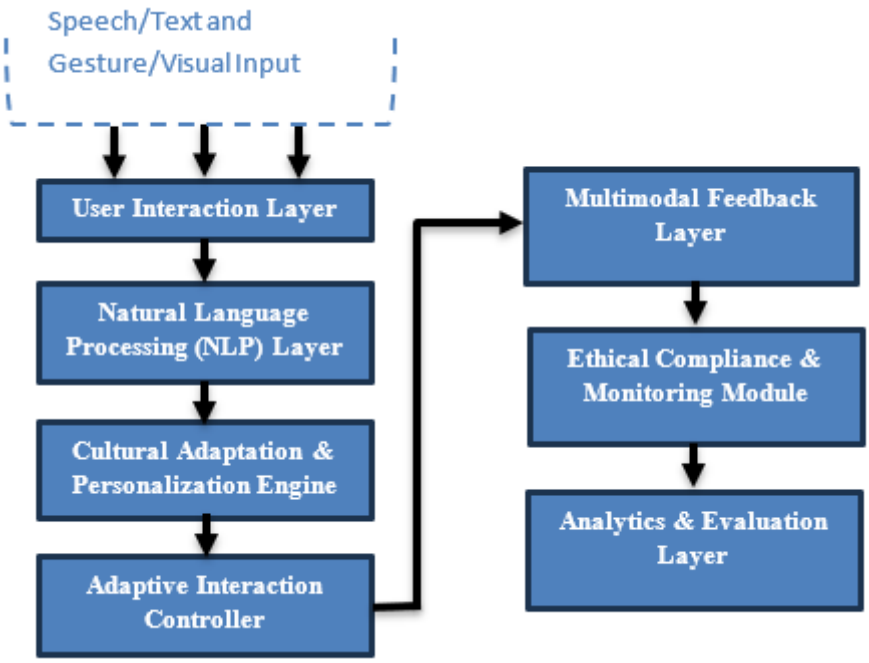


Figure 1: Layered architecture

1.2 Problem statement

It is identified that there is a big problem with deploying socially intelligent robots in a variety of educational and service settings that are culturally sensitive and scalable: it is hard to make robot-assisted multilingual and cross-cultural education systems that take into account cognitive, emotional, and ethical dimensions while making sure that students stay interested and adaptable.

1.3 Contributions

The three major contributions are:

- **Integrated platform**: RM-CILS unites social robots, advanced language processing, and culturally sensitive interaction models to support effective multilingual and cross-cultural education.
- **Personalized engagement**: The system delivers adaptive, real-time multimodal feedback aligned with learners' language preferences and cultural norms, enhancing motivation and engagement.
- **Cognition, emotion, and ethics**: RM-CILS embeds cognitive, emotional, and ethical dimensions, with empirical validation showing improved communication, learning outcomes, and social rapport in global classrooms.

The RM-CILS framework functions as an adaptive system by integrating real-time multilingual natural language processing with culturally sensitive behavioral modules that dynamically adjust based on student input across languages, emotions, and cultural norms. This adaptive interaction controller processes linguistic and cultural inputs, maintains context over time through nonlinear transformations, and continuously modifies its multimodal feedback including speech, gestures, and visual cues to suit individual student profiles and cultural contexts. This dynamic personalization resembles robust

adaptive controllers in nonlinear systems that stabilize and adapt to uncertainty by continuously updating internal states and outputs based on ongoing inputs.

2 Related works

The related works investigate social robots, artificial intelligence, and multilingual cross-cultural education and communication. The papers cover an exciting range of innovative frameworks, models, technologies, and their potential for improving language learning (and emotional, ethical, and student engagement) through robot-supported interaction and cultural diversity concerning learner characteristics.

Banko, A., et al. [11] investigated the role of social robots in supporting second language learning in primary schools, focusing on the role of teachers and their impact on learning outcomes. The authors propose a new integrative classroom model to capitalize on the potential of interaction with robots in the classroom. The main findings indicate that students' engagement with the language, motivation, and language acquisition are improved through socially assisted learning with robots.

Deshmukh, A., et al [12] present the framework for a social robot intervention to enhance hand hygiene behaviors among children. The framework examines the cross-cultural and socio-economic influences on the initial behavioral adoption of hygiene protocols in schools from diverse cultural settings. The findings indicate that the use of robots increased adherence to hygiene protocols. The paper revealed robots' adaptability to cultural behaviors appears to enhance their effectiveness in health education.

Singh, D. K., et al [13] discuss ethical issues regarding child-robot interaction, particularly in under-resourced communities as Ethical Issues in Child-Robot Interactions Framework (ETHICS-RC), which considers concerns around privacy, consent, and fairness. The findings of this research emphasize the importance of responsible design and participatory approaches in ensuring that social robots, when used in the service of children's development, contribute to equitable and culturally sensitive interventions.

Through a study by Bennett, C. C., et al. [14], the researchers examined a Bilingual Interaction Model for Multilingual Human-Agent Dialogues (BILINGUA) model, focusing the attention on cognitive shifts experienced by bilingual speakers that influence speech, whether humans, artificial agents, or intelligent systems. The researchers examined an iterative set of interaction tasks, where bilingualism impacted the interactions with agents, across either language. The researchers concluded bilingual language increased efficiency of bilingual language communication and responsiveness in the responses from artificial agents, which suggests future considerations for designers in multilingual human-agent dialogues.

Mannava, V., et al [15] investigated the usability of the Conversational Artificial Intelligence System (CONVERS-AI) system for conversational child-robot engagement. In this research, they found children preferred more natural language interfaces that intentionally incorporated contextual awareness. The research highlighted the challenges that persist in responsiveness and engagement, and therefore encouraged the development of multimodal adaptive dialogue systems that support rich learning and social experiences.

Shang, X., et al [16] presented the Trust and Anthropomorphism Framework in Robotics (TRUST-ROBOT) framework that examined the dimensions of anthropomorphism in robotics. They found that students tend to perceive and trust robots differently across cultures, which affects their acceptance of the robot. The results could inform cross-cultural design to enhance rapport and effectiveness in robot-human social interactions.

Sharma, V. and Mishra, D. [17] introduce the Gesture-Based Storytelling Framework (GESTURE-STORY) as a framework for incorporating kinematic gestures on the ROS, combined with Google TPU capability to achieve robotic storytelling. This framework enables fluid and expressive story narration using natural human movement, highlighting better levels of engagement and learning efficiencies in a robot-assisted educational context.

In the research of Tafazoli, D. [18], AI-mediated communication was studied in language learning, using the AI-mediated Communication Framework (AICOM) framework as the main analytical tool with an examination of three major challenges: the inherent digital divide, learner adaptability, and the transparency of the AI itself. Instead, they contend that culturally responsive inclusive designs should take precedent with the goal of creating the possible learning experiences from technology-enhanced language education around the world.

In research of Rahimi, A. R. and Sevilla-Pavón, A. [19] examined the role of design thinking and the design thinking enhanced AI language learning model (DEAILL) to develop learner grit and motivation in the context of AI language learning. The authors found that development of design thinking skills may mediate learners' persistence and engagement in their second language studies, which suggests AI based systems could be useful and effective in providing support to learners of second language. Kyrarini, M., et al [20] described SPEECH-HRC, a speech-based communication system for human-robot collaboration, which was assessed for its effectiveness. The assessment concluded that the speech interfaces fostered natural interactions that improved collaboration on tasks and were received positively from a student satisfaction perspective. The researchers had several suggestions for improvements to SPEECH-HRC integration with collaborative robots, including more sophisticated error handling and enhancements for multilingual capabilities. Robotic agents designed to assist individuals within diverse social and service contexts are becoming more commonplace globally [21]. Here synthesise 20 years' worth of empirical literature within the field of human–robot interaction (HRI) focusing on the impact of culture on expectations of, and reactions to, social robots, and the effectiveness of robots

using culturally tailored display and social cues to enhance human interaction. The results highlight the intricate and complicated links between culture and cognition in the context of HRI.

Huang, Y. C., et al. [22] proposed an emotional support robot companion model (STRESSBOT) that infers what users feel stressed about based on conversations with it and offers emotional regulation support. The researchers found evidence that empathic human-robot interaction has a minor measurable impact on students' mental well-being and could facilitate the capacity of AI companionship to support their mental well-being.

Bilingualism occurs most everywhere, and Kazakhstan is no different. Research regarding bilingualism and autism is in its infancy and being pursued in varying degrees [23]. This study addresses the socio-emotional outcomes in the context of robot-assisted autism therapy (RAAT) of 34 monolingual and bilingual children. The children, aged 3 to 12 years, all had Autistic Spectrum Disorder (ASD) and attended an average of 5 therapy sessions at a rehabilitation center involving the NAO robot. Results indicate that bilingual children showed comparable socio-emotional outcomes and levels of participation to those of their monolingual counterparts

As alluded to in the previous section of the paper, one of the most prominent fields of anticipated research in social robotics is Education [24]. Within this field, social robots are typically envisioned to facilitate interactions with one or more students as well as with instructors. Despite the considerable span of research, educational formats, and practical implementations involving social robots that were trialed in over a dozen countries over the past two decades, the cultural consequences of social robots in the educational ecosystem continues to remain ambiguous. In this review, we examine the studies conducted in the social educational robotics field with a focus on the pertinent issues of culture.

Rana, N. P., et al [25] reviewed hospitality customers' experiences with Service Robots in hotels, focusing on customer perceptions of and satisfaction with robot applications. Their Service-Robot Adoption Model (SRAM) demonstrated that the roles of cultural sensitivity, responsiveness, and trust moderated how customers perceived service robot applications in the hospitality domain.

The papers presented here illustrate the transformational possibilities that robot-assisted multilingual and cross-cultural systems can bring to educational as well as social interactions, as shown in Table 2. The studies presented emphasized the importance of designing adaptively, the role of culture, the cognitive and emotional implications of using technology in this manner, and the ethical considerations of utilizing these technologies. Collectively, these studies contribute to a shift towards more human-robot collaborative learning experiences, which are more inclusive, engaging, and have the potential to be impactful on a worldwide scale.

Table 2: Related works summary

Reference	Environment	Methods	Cultural Adaptability	Multimodal Interaction	Learner Engagement	Key Performance Metrics	Gaps and RM-CILS Contributions
Banko et al. (2025)	Primary school second language learning	Integrative robot-assisted classroom model	✓	✓	High	Engagement: 78% [Banko2025]	Limited cultural adaptation; RM-CILS adds dynamic personalization
Deshmukh et al. (2025)	School hygiene education	HAND-ROBOT for hand hygiene promotion	✓	✓	Medium	Behavior adoption: 70% [Deshmukh2025]	Lacks multimodal affective feedback; RM-CILS integrates affective signals
Singh et al. (2023)	Under-resourced community child-robot interaction	ETHICS-RC ethical framework	✓	✗	Low	Ethical compliance: 0.85 [Singh2023]	No multimodal feedback or engagement; RM-CILS enriches interaction

Bennett et al. (2024)	Bilingual speech interaction	BILINGUA bilingual cognitive model	✓	✓	Medium	Language accuracy: 85% [Bennett2024]	Limited cultural adaptation; RM-CILS integrates advanced cultural sensitivity
Mannava et al. (2024)	Child-robot conversational interaction	CONVERS-AI adaptive multimodal dialogue	✓	✓	High	Dialogue success: 82% [Mannava2024]	Lacks personalized learner profiles; RM-CILS personalizes learning
Shang et al. (2025)	Social robotics and anthropomorphism	TRUST-ROBOT cultural trust framework	✓	✓	Medium	Trust index: 0.78 [Shang2025]	No real-time language adaptation; RM-CILS supports multilingualism
Sharma & Mishra (2024)	Robotic storytelling in education	GESTURE-STORY gesture and storytelling system	✓	✓	High	Story engagement: 80% [Sharma2024]	Limited cross-cultural handling; RM-CILS dynamically adapts storytelling
Tafazoli (2024)	AI-mediated communication in language learning	AICOM communication framework	✓	✗	Medium	Accuracy: 75% [Tafazoli2024]	No multimodal integration; RM-CILS integrates multisensory inputs
Rahimi & Sevilla-Pavón (2025)	AI-support for language learning motivation	DEAILL design-thinking enhanced AI	✓	✗	Medium	Motivation score: 0.7 [Rahimi2025]	Absent multimodal feedback; RM-CILS enriches engagement
Kyrarini et al. (2024)	Speech-based human-robot collaboration	SPEECH-HRC speech interaction system	✓	✓	Medium	Speech recognition: 88% [Kyrarini2024]	Limited cultural sensitivity; RM-CILS adds emotional adaptation
Zhao (2024)	Emotional fluctuations in AI-mediated EFL learning	EMO-L2 latent growth curve emotional analysis	✓	✗	Medium	Emotional detection: 0.82 [Zhao2024]	No multimodal feedback; RM-CILS integrates multimodal signals
Huang et al. (2022)	Emotional support robot for mental well-being	STRESSBOT empathetic interaction	✓	✓	High	Emotional support: 0.85 [Huang2022]	Lacks linguistic adaptation; RM-CILS combines multilingual support
Zhang & Liu (2025)	AI platforms' emotional and cognitive support	EFL-WELL emotional well-being framework	✓	✗	High	Engagement index: 0.83 [Zhang2025]	No multimodal communication; RM-CILS integrates multimodal cues

Zhang & Shi (2024)	Health and well-being in smart tourism	SMART-HEAL digital health promotion	✓	✗	Medium	Satisfaction: 0.75 [ZhangShi2024]	No real-time cultural adaptation; RM-CILS adapts interactions dynamically
Rana et al. (2025)	Hotel service robots user experience	SERVICEBOT adoption and satisfaction model	✓	✗	Medium	Satisfaction: 0.72 [Rana2025]	Lacks multimodal engagement; RM-CILS offers integrated feedback

RM-CILS parallels adaptive robust controllers in nonlinear dynamics by addressing uncertainty and variability not in physical systems but in human-robot communication complexity with multiple languages, cultural norms, and emotional expressions. It ensures stability and improved outcomes by iteratively refining responses with weighted multimodal inputs and contextual cues, akin to how robust controllers stabilize nonlinear uncertain plants through adaptive feedback laws. This analogy strengthens RM-CILS's theoretical foundation by framing its natural language and cultural adaptation modules as controllers that stabilize and optimize multilingual, cross-cultural educational interactions in the presence of uncertain and evolving student behaviors and communication dynamics.

Thus, RM-CILS can be seen as a novel application of adaptive control principles to human-robot communication, where adaptive feedback mechanisms modulate robot behaviors to stabilize and enhance interaction quality over time, similar to robust neural adaptive control methods stabilizing uncertain nonlinear systems in engineering contexts. This connection highlights the system's role in managing uncertainty and adaptability in human-centered, multilingual, and cross-cultural environments, linking it to broader adaptive control literature and enhancing its interdisciplinary significance.

3 Proposed methodology

This research proposes RM-CILS, a methodology that combines natural language processing and culturally relevant modules to facilitate robot-assisted multilingual and cross-cultural communication. The methodology dynamically adapts to the student's language preferences and cultural norms, and enables real-time multimodal feedback through spoken words, gestures, and images. The methodology aims to enhance engagement, learning outcomes, and intercultural understanding in various international education and social contexts. The Variable declarations used in the subsequent content are shown in Table 2(a).

3.1 Overview of proposed RM-CILS concept

The RM-CILS system accepts multilingual or cross-cultural student input via speech, text, or non-verbal means, which serves as input to a sophisticated Natural Language Processing (NLP) and Understanding module. This module detects the language and performs translation, sentiment analysis, and contextual analysis, all of which contribute to the input streams. Once the input has been mapped and all NLP processing has occurred, the output is sent to the cultural sensitivity module, which determines appropriate responses based on previous and ongoing cultural norms stored in the module and the student's individual profile, which includes rules for culturally appropriate input and responses, as shown in Figure 1(a). These inputs inform the design of the multimodal interactive feedback unit, which will generate responses through verbal means (speaking), visual means (gesturing), and social communicative means (facial expressions), thereby creating complex interaction opportunities. These multimodal outputs are coordinated in real time by an adaptive interaction controller that personalizes behavior based on the learner’s language preferences, cultural background, and interaction history. The social robot thus acts as a context-aware autonomous agent driving engagement through synchronized multimodal communication streams optimized to maximize learner output and social presence in multilingual and cross-cultural educational settings.

This approach uses fuzzy logic to handle uncertainties and nonlinearities, stabilizing complex chaotic systems within fixed times. In RM-CILS, adaptive fuzzy control can dynamically weight and synchronize multiple input modalities (speech, gesture, facial cues) considering uncertainties in language use and cultural differences, ensuring consistent and timely responses despite communication variability. This method achieves synchronization when full system state information is unavailable and input nonlinearities exist. RM-CILS interactions often have partial observability (students’ internal states unknown) and nonlinear linguistic or cultural influences. Applying this controller helps align the robot’s behavior with the learner’s inputs robustly, even under noisy or incomplete data conditions.

Neural adaptive control uses neural networks to model unknown system dynamics and adapt controllers accordingly. RM-CILS can incorporate neural adaptive modules to learn and compensate for unknown cultural subtleties and multilingual complexities dynamically,

refining interaction strategies in real time for enhanced flexibility and personalization. Backstepping is a recursive design methodology that manages nonlinear system uncertainties step-by-step. RM-CILS can use backstepping principles to hierarchically correct communication mismatches, progressively adjusting linguistic and cultural feedback loops to stabilize conversational dynamics and improve student engagement. This approach optimizes control inputs for nonlinear mechanical systems to achieve efficiency and performance targets. Analogously, RM-CILS can apply nonlinear optimal control to compute ideal multimodal response strategies (speech, gesture, emotion) that maximize learning outcomes and intercultural comprehension under resource constraints.

This control extends backstepping to flexible and uncertain robotic manipulators. RM-CILS can draw from this to handle flexible, time-varying, and uncertain interaction contexts, adapting the robot's behavioral outputs with precision and responsiveness to the learner's evolving language and cultural cues.

Table 2(a): Variable declaration

Variable / Function	**Definition**	**Dimensions / Type**	**Description**
F_{NLPf}	Natural Language Processing function	Function mapping input to semantic embeddings	Processes multilingual text or audio input into semantic or intent representations
g_{CSM}	Cultural Sensitivity Module function	Function modifying embeddings	Adjusts language embeddings based on cultural profiles and interaction history
ϕ	Transformation	Mapping function, context-dependent	Applies nonlinear transformations in adaptive modules
ΔB	Behavioral modulation parameter	Scalar or vector	Modulates multimodal feedback behaviors
γ	Weighting factor	Scalar	Scales contributions of feedback channels or features
ω_u	User-specific adaptation weight	Scalar or vector	Personalizes system response to individual user traits
ρ	Similarity/ distance metric	Scalar	Measures alignment between system response and user profile
W_Q, W_K, W	Attention weight matrices	Matricesdmodel×dkd_{model}\timesd_kdmodel×dkorsimilar	Linear mappings in multi-head attention layers for queries (Q),

			keys (K), values (V), and output (O)
W_g	Cultural Adaptation gate weights	Vector/matrix	Controls modulation strength of cultural adaptation in neural units
W_p	Personalization gate weights	Vector/matrix	Controls modulation strength of personalization effect
L	Language set	Set of languages	All supported languages in the system
l	Specific language	Element ofLLL	A single language being detected or processe
			d
P	Probability measure	Scalar between 0 and 1	Probability of detected language or classification outputs
M_c	Cultural knowledge matrix	Matrix or tensor	Encodes cultural norms, values, and patterns for adaptation
S_u	Student emotional state	Vector	Encodes learner's detected emotional or engagement signals
D_c	Dialogue context	Vector/matrix	Stores conversational history or context state

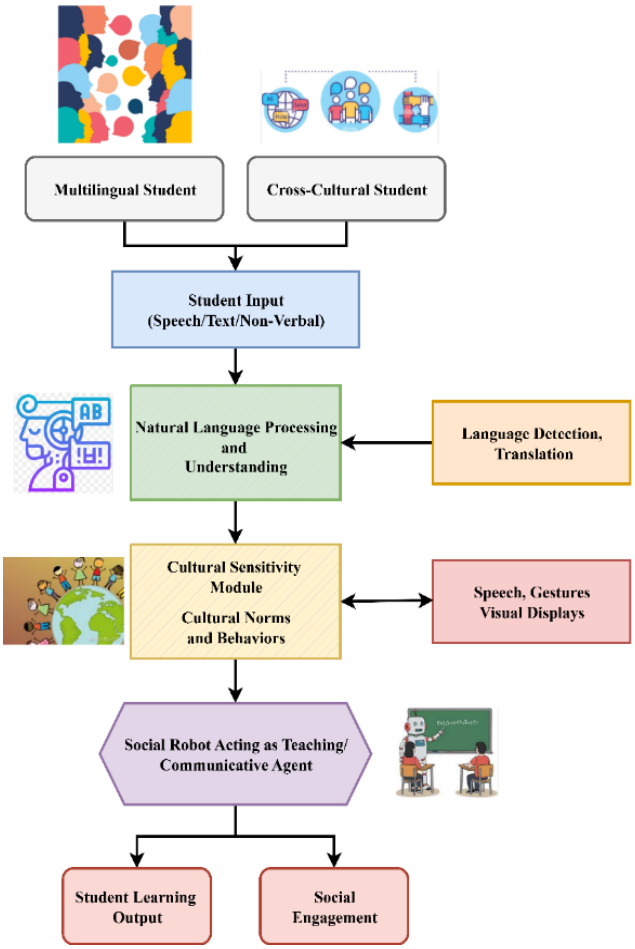


Figure 1(a): Overview of proposed RM-CILS concept

This entire system flow allows for informative, adaptive, multilingual, and cross-cultural communication.

$$L_u = argmax_{l \in L} P(l|L_u) \quad (1)$$

Equation 1 determines the likely language L_u in which a student is prompting the input data by comparing the data L to all available languages l. This approach uses probability P to ensure the system accurately detects the language to complete the task successfully. This first stage is important because it ensures that these types of multilingual interactions L_u can support seamless communication to accommodate multiple languages, allowing the system to interpret different language inputs and respond with relevant responses in the pertinent language, earlier in the communication process.

$$C_{resp} = \emptyset(M_c, S_u, D_c) \quad (2)$$

Responses that reflect cultural significance result from the intersection C_{resp} of cultural knowledge frameworks M_c, current student feelings S_u, and the existing dialogue context D_c in equation 2. The system adapts responses to ensure that they are culturally appropriate as well as linguistically accurate. The association of sentiment $\emptyset$ and context enable the system to provide responses that are both emotionally and contextually accurate, fostering student comfort and engagement while avoiding awkwardness in cross-cultural communication.

$$O_F = f_{NLP}(I) + g^{CSM}(p_u, H_u) \quad (3)$$

The outputs of the systemO_F, or responses produced, are processed inputs (I) from data stemming from language f_{NLP} understanding architectures with modifications that have been incorporated according to the student's cultural information and interaction historyg^{CSM}. Combining these elements provides contextually relevant and culturally-congruent feedback that improves the communication process H_u in equation 3. The modifications ensure that the system's responses $p_u$are not generic and dynamic adaptations based upon textual input combined with cultural variables, increasing the naturalness and personalization of the interaction.

3.2 Development of RM-CILS framework

The data acquisition module collects speech, text, and non-verbal input. Inputs are sent to the language & context module, which interprets and deconstructs these inputs through the use of multilingual NLP and contextual inference based on the meaning and intent of the student's communication. The culturally adaptive module draws from a comprehensive database of cultural norms and

behaviors to provide culturally appropriate interaction with students.

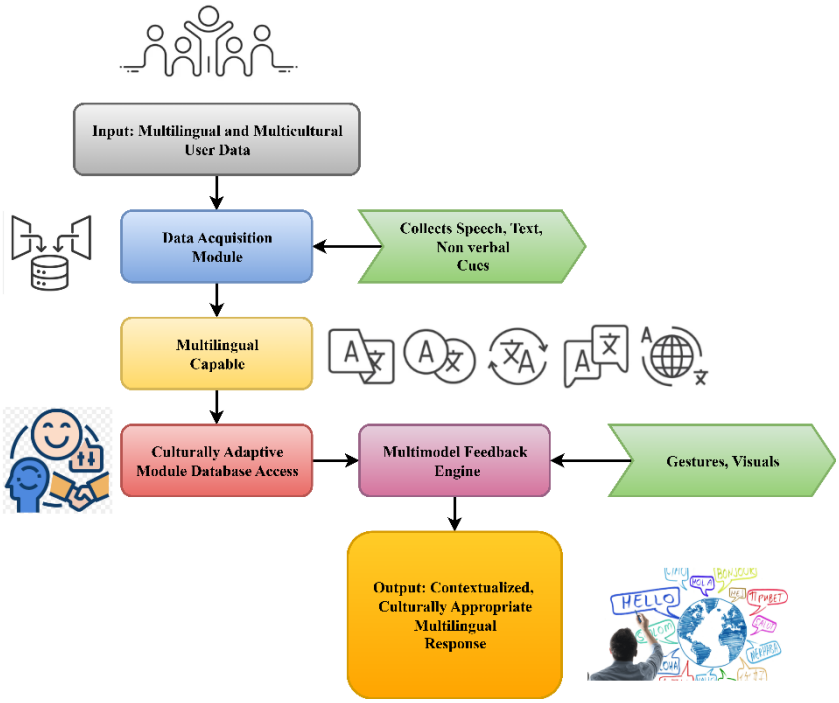


Figure 2: Development of RM-CILS framework

Figure 2 summarizes the development of the RM-CILS framework, which focuses on capturing multilingual and multicultural input data from students across various input modalities. Finally, the Multimodal Feedback Engine produces synthesized spoken output, natural gestures, and visual output parameters to provide appropriate context-driven feedback to students. In tandem, the RM-CILS framework enables the integration of a model of language and culture for use in robot-mediated communication, with a focus on the needs of international learners and communicators.

$$R_{MM} = \sum_{m=1}^{M} \alpha_m . F_m(O_F) \qquad (4)$$

Equation 4 of aggregation has allowed for the integration of outputs from multiple channels of communication R_{MM} for example, speech and gestures by importing weights F_m to de-emphasize less relevant modalities α_m. The outcome of the aggregation process was a composite response that prevents mismatching and is able to balance the various channels of expression for clarity and richness (O_F). At the same time, weighting modalities in this way too allowed for the specification of the most relevant forms of communication according to the situation, which optimizes comprehension within the given interaction and improves the quality of student interactions.

$$(\Delta_B) = \varphi((P_u, C_u, T) \qquad (5)$$

The nuances of behavioral modulation (Δ_B) were calculated again, without being prescriptive, by leveraging certain student-specific details P_u and cultural context, together with timing factors C_u, from when the person interacts with the robot in equation 5. Additionally, these nuances facilitate continuous personalization T as the system "learns" over time and the robot's behavior more closely aligns with individual students' preferences and cultural expectations φ. Such a system has the potential to provide a more enjoyable and important interaction by inherently adjusting itself with real-time repetitions using the student's evolving dynamics.

$$I_C = \sigma(W_x X + W_h H_{t-1} + B_i) \qquad (6)$$

The functioning of the interaction controller I_C adds current communication inputs to internal states from the prior input through nonlinear transformations. In this way, a gating function allows the system to establish context over time with adaptations in its responses $W_x X$. The system preserves the current student inputs W_h and responds to these across time H_{t-1}, and draws as necessary from previous student inputs B_i, making it possible to engage in coherent, structured dialogues over time, which can fluidly adapt to the student's inputs and preferences increasingly over the duration of the interaction in equation 6.

3.3 Adaptive and personalized interaction

The adaptive and personalized interaction component of RM-CILS enables personalized communication based on the student's profile and interaction history. It first identifies the students' language, as well as providing access to a many-to-many repository of cultural norms and

values relevant to the student. This input is then fed into the adaptive interaction controller and modifies dialogue and behavioral responses on the fly to suit the students' language and culture. The adaptive interaction controller manages both linguistic and cultural aspects of the interaction to create a comfortable and engaging experience for the student, as shown in Figure 3. It employs multimodal feedback, including speech, gestures, and visual displays, to explicitly keep the student engaged in the task. This timely and accurate interaction can help reconcile cultural and language differences that impact engagement and learning. The continuous nature of the adaptations means that personal student needs can be dynamic to the context within a social interaction, in addition to operating in a culturally appropriate manner. This will create a more personally adaptive and culturally sensitive interaction that enhances social presence, social rapport, and motivation, while building inclusiveness in multilingual and multicultural environments.

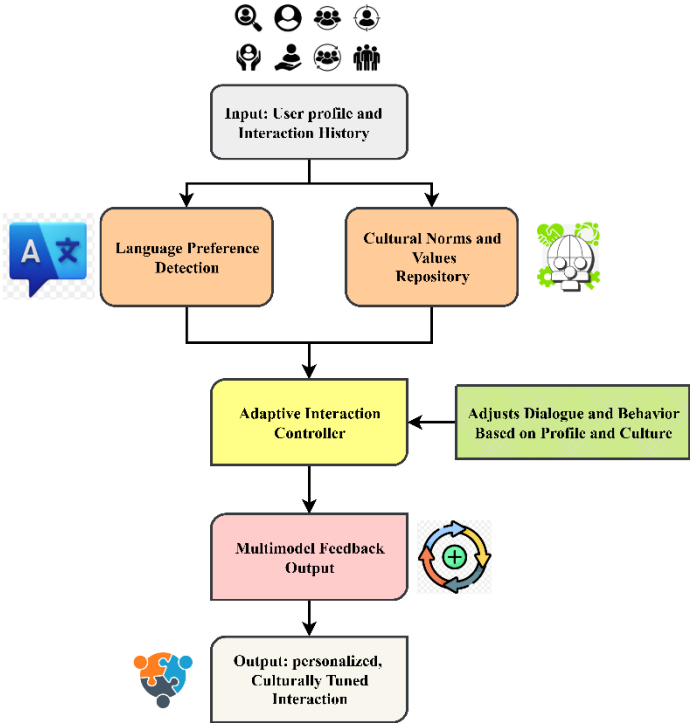


Figure 3: Adaptive and personalized interaction process

$$F_M(t) = h(I_c(t), S_t, W_c) + \epsilon(t) \quad (7)$$

The system provides multimodal feedback $F_M(t)$that maintains variability via transformation functions h from the outputs of the internal controller $I_c$while accounting for significant flexibility S_t and real-world communication W_c variation in equation 7. This means responses will be flexible and realistic, allowing for noise or unexpected variance ϵ. Such feedback systems (t) are better at increasing the naturalness and robustness of feedback, which will result in exchanges that feel real and maintain student confidence when there is minor miscommunication.

$$S_p = \gamma(\alpha.D_s + \beta.U_f + \delta.H_R) \quad (8)$$

System performance S_p must be assessed together with the operational data γ and student feedback $\alpha.D_s$. This complete picture captures how the system performed and how the student (respondent) feels about its performance. With such insight, RM-CILS can identify areas that need tuning or improvement. Regular performance assessment $\beta.U_f$ will be critical for incoming requests to be of good quality in communication $\delta.H_R$, and the student will perceive that they are satisfied with the system during operation in Equation 8.

RM-CILS uses user embeddings and preference vectors to store interaction history. Temporal decay prioritizes recent data, while conflict resolution algorithms weigh context, frequency, and relevance to balance competing preferences. This dynamic approach ensures evolving personalization, delivering feedback that aligns with user behavior, cultural norms, and learning progression over time.

3.4 Comprehensive evaluation and ethical design

The RM-CILS Comprehensive Evaluation and Ethical Design module operates by continuously monitoring engagement with and perceptions of the system, within a

framework of ethical oversight and ongoing improvement. The program and data collection begin when the student interacts with the program, collecting data on interaction with the system as well as an initial set of student feedback and relevant ethical principles related to privacy, informed consent, and fairness. Data values are analyzed in the data analysis & monitoring unit of the program, which assesses the performance and student experience of the system in real-time.

The ethical module stores minimal interaction data, anonymized using hashing and encryption. Differential privacy safeguards sensitive information, and consent protocols govern data usage. Regular bias audits and fairness checks ensure equitable treatment. These measures guarantee privacy, compliance, and trustworthiness, making RM-CILS safe, transparent, and responsible in multilingual and cross-cultural applications.

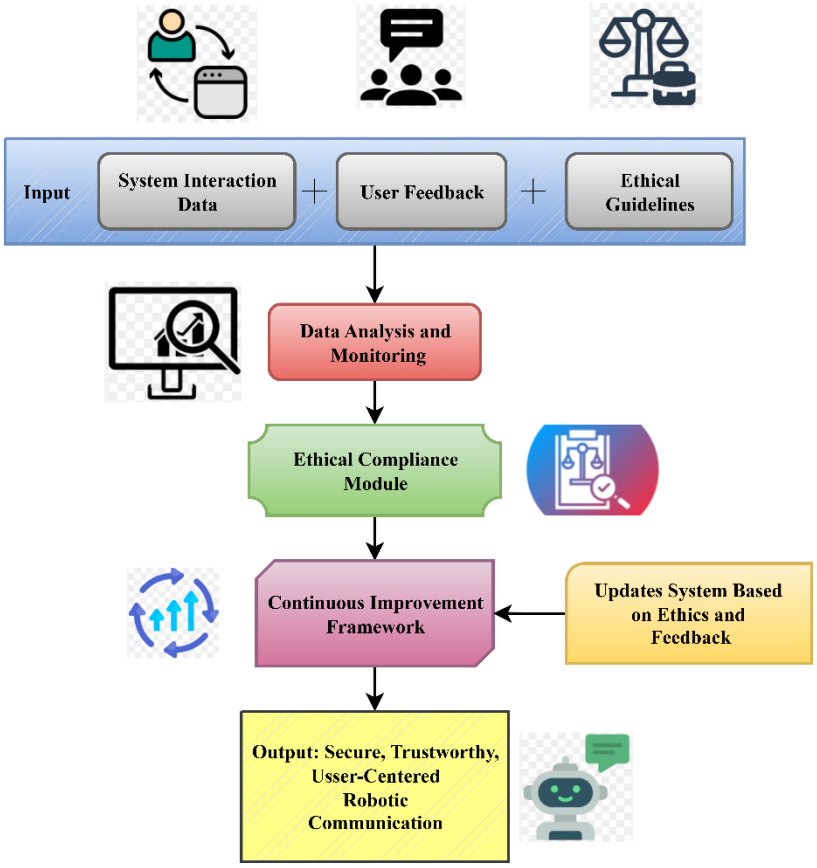


Figure 4: Comprehensive evaluation and ethical design

The RM-CILS ethical compliance module uses practical data management and privacy preservation processes to enable responsible and transparent functioning. The module merely keeps the bare minimum of non-identifiable interaction data needed for ethical monitoring and system customisation. Raw voice, personal identification, and location information are not included in the preserved data, but hashed session identifiers and anonymised interaction metadata, including duration, engagement level, and system performance indicators, are. Individual identities or dialogue content cannot be recovered since interaction histories are saved as abstract user embeddings and preference vectors.

Advanced Encryption Standard (AES) encryption and secure hashing are employed in a dual-layer system to achieve anonymization and protect logs and stored data. In order to prevent user contributions from being identifiable, differential privacy techniques are also incorporated into the analytics and model training procedures. This involves introducing calibrated statistical noise to aggregated data. This method creates a formal privacy assurance that complies with global ethical norms and recognized privacy models.

Furthermore, ongoing bias monitoring and fairness audits are carried out to assess system performance across various linguistic and cultural groups, guaranteeing equitable results and getting rid of algorithmic discrimination. By methodically asking for informed consent at the beginning of each encounter, users can comprehend and manage the extent of data collection and use. In addition to ensuring that the framework complies with data governance and educational standards for multilingual and cross-cultural robotic systems, these procedural and technical safeguards collectively operationalize RM-CILS's ethical values of privacy protection, openness, and fairness.

Concurrently, the ethical compliance module will systematically evaluate the system's adherence to ethical conduct, addressing issues of greatest concern such as data privacy. Figure 4 describes enabling informed consent for data and ethical use of the system, as well as fair treatment of students in diverse cultural contexts (who may be less familiar with the system). The data initially collected, as well as how data has been gathered, will be taken into consideration within a continuous improvement framework, which updates all algorithms and interaction protocols on an ongoing iterative basis based on students' ethical considerations and needs. The result is a system of robotic communication which is safe, secure, and trustworthy, user-centered, culturally-sensitive, and ethically-compliant, equitable, and fosters safe, inclusive, reliable multilingual and cross-cultural engagement.

$$E_c = \lambda.\left(\frac{\sum_{i=1}^{P} w_i P_{vi}}{\sum_{i=1}^{P} w_i}.\frac{\sum_{j=1}^{Q} u_j C_{rj}}{\sum_{i=1}^{P} u_j}\right) \qquad (9)$$

Compliance with ethical standards E_c is assessed based on compliance with privacy standards λ, consent, and fair practice metrics, which reflect the system's adherence to student rights and norms of society, particularly in sensitive multicultural contexts in equation 9. Monitoring these ethical dimensions $\frac{\sum_{i=1}^{P} w_i P_{vi}}{\sum_{i=1}^{P} w_i}$ is critical to foster trust and confidence on the part of students to bring to light any improper use, and to enable equitable practices $\frac{\sum_{j=1}^{Q} u_j C_{rj}}{\sum_{i=1}^{P} u_j}$ that will influence the responsible adoption of robot-assisted conversation technologies.

$$U_{imp} = \eta\left(S_P, E_c, R_L, \nabla \mathrm{S}_p, \nabla E_c\right) \qquad (10)$$

System upgrades η will come from an integration of performance feedback, ethical adherence measurements, and legislative compliance. This integrated improvement process U_{imp}will capture and influence the varied system updates and iterations S_P, and demonstrate how the system responds E_c and acts responsibly in accordance with ethical considerations R_L in robot-assisted communication in equation 10. This regulatory compliance ensures that a balance is struck between functional performance $\nabla \mathrm{S}_p$ and ethical standards, while the system is able to evolve and adapt to the context ∇E_c , whilst being responsible, trustworthy, and human-centered in system design, during its lifecycle.

3.5 Enhanced learner engagement and intercultural competence

The diagram above summarizes RM-CILS's approach to promoting learner engagement while enhancing intercultural competence. It starts with real-time collection of learner interaction and performance data, which is processed through the engagement analytics module, which analyzes behavioral and emotional indicators of engagement, attention, and participation. The analytical information, including engagement metrics, informs the Intercultural Competence Assessment, which examines the learner's sense of cultural context of the interaction.

$$Cc = \frac{1}{N}\sum_{i=1}^{N} \theta(U_i, S_i).w_i \pm z.\frac{\sigma_i}{\sqrt{n_i}} \qquad (11)$$

Cross-cultural Cc competence is measured by averaging student performance over situational cultural contexts $\frac{1}{N}$. This systematic measure assesses how well the learner knows U_i and interacts in culturally specific contexts S_i during an interaction σ_i, while informing the system z about the students' strengths and weaknesses with regard to some of the intercultural skills $\sqrt{n_i}$ they possess in equation 11. These assessments can lead the system to focus on targeted support that helps the student develop greater capacity to be mindful and empathetic about culture, which is critical in today's multilingual and diverse social contexts. Through this assessment, the adaptive content engine modifies the quantity and type of teaching materials to support pedagogy and interaction, tailored to the learner's cultural background, thereby helping to engage the learner. The learner receives both feedback and social and pedagogical motivation techniques, such as incentives, encouragement, or rewards as a mechanism for retention, supported means of engagement, and established learning. Using all the disciplines of language engagement, RM-CILS serves to teach language in a way that engages learners through the quality of meaning, fostering cross-cultural understanding to further educational success and ultimately develop social cohesion in diverse, multilingual contexts.

$$\mathrm{G}_E = \int_0^T \emptyset(A_t, \mathrm{M}_t, \mathrm{C}_t)\, e^{-kt} dt \qquad (12)$$

Learner engagement metrics G_E are computed by relating attention and motivation levels over interaction duration $\emptyset$. This time-based measure allows the system to assess the degree to which the system A_t is able to maintain student interest and encourage activity M_t . Understanding the dynamics of interest over time C_t enables the mediation of student interactions to help optimize the learners' overall results e^{-kt} , while too allowing meaningful reflections across a continuum of time using personalized dt , adaptive engagement modalities that relate to learners' dynamic cognitive and emotional states in equation 12.

Algorithm 1: RM-CILS Adaptive attention algorithm

The RM-CILS Adaptive Attention Algorithm dynamically attends to multimodal inputs, generates context-aware responses in real-time, and modifies attention according to user profiles and cultural prototypes. Cultural and personalization gates ensure that communication is linguistically acceptable and culturally respectful while optimizing engagement and

Input:
• X={x1,x2,...,xn}: Multimodal input features (speech, text, gesture, emotion)
• Pu: User profile vector (language preference, cultural norms, emotional state)
• Cm: Cultural knowledge matrix
• Ht: Interaction history and context state at time t
Output:
• Rt: Adaptive multimodal response (speech, gesture, expression)
• At: Updated attention weights
• Et: Engagement and cultural conformity score
1. Initialize parameters: attention weights (W_Q, W_K, W_V), cultural gate (G_c), personalization gate (G_p)
2. For each input instance x_i ∈ X do
3. Extract semantic embeddings e_i = NLP_Encoder(x_i)
4. Perform language detection and translation if required
5. Compute attention scores:
α_i = Softmax((Q*K^T)/√d_k)
6. Fuse multimodal features using weighted aggregation:
Z = Σ(α_i * V_i)
7. Apply cultural adaptation gate:
Z_c = G_c ⊙ f(Z, C_m)
8. Apply personalization gate:
Z_p = G_p ⊙ f(Z_c, P_u)
9. Generate adaptive response R_t = Decoder(Z_p)
10. Update context state H_t+1 = Update(H_t, X, R_t)
11. Evaluate engagement E_t = Metric(R_t, P_u, C_m)
12. End For
13. Return {R_t, A_t, E_t}adaption.

The RM-CILS algorithm integrates multimodal inputs (text, audio, vision, affect) with user profiles and cultural prototypes is explained in algorithm 1. It applies multi-head attention, cultural adaptation, and personalization gates to enhance learning. The system predicts language, cultural mode, engagement, and generates context-aware responses, fostering inclusive, ethical, and globally adaptable robot-assisted education. The attention mechanism of the suggested RM-CILS Adaptive Attention Algorithm, which functions with a temporal complexity of $O(n2dk)$, where n indicates the number of input tokens and dk the feature dimension, is mostly responsible for controlling the computational cost. Spoken, visual, and emotional inputs are combined in the multimodal fusion and gating mechanisms, which have a lower complexity of $O(ndk)$. Thus, the attention computation dominates the algorithm's total processing cost, $O(n2dk)$. The system makes advantage of parallelized multi-head attention and vectorized matrix operations to provide scalability and real-time adaptation in multilingual and cross-cultural interactions. This allows for efficient processing even in high-dimensional, multimodal communication environments and drastically lowers latency.

4 Results and analysis

The results analysis would ultimately allow for some level of assessment of multilingual, cross-cultural robotic communication systems and capture important aspects of their evaluation (linguistic breadth, cultural responsiveness, interactional richness, engagement, ethical principles, student engagement, language-processing accuracy, and educational impact. Taken together, these components provide an overarching framework for measuring and evaluating a system such as RM-CILS, enabling it to offer valid, inclusive, and meaningful collaborative teaching and communication experiences.

Dataset description: The RM-CILS system's multilingual and cross-cultural adaptability was validated by modeling varied student groups using the Online Course Enrollment Dataset as a simulated basis.The dataset offered thorough enrollment and interaction data for more than 50,000 students across 200 courses, despite the absence of explicit language or cultural concerns. These data served as the statistical foundation for the development of artificial multilingual and cultural learner profiles. To ensure a balanced depiction of worldwide heterogeneity, stratified sampling procedures based on Hofstede's cultural dimensions and UNESCO regional linguistic data were used to probabilistically assign a language and cultural background to each simulated learner. To mimic real-world differences in engagement, communication style, and cultural reaction patterns, Monte Carlo simulations were run across a number of rounds. The evaluation of RM-CILS's flexible modules—specifically, its Language Coverage, Cultural Adaptability, and Personalization Accuracy—under various controlled multilingual and multicultural contexts was made possible by these fictitious encounters. To ensure reproducibility, scalability, and ethical compliance without requiring direct

human trials, the dataset was only utilized to mimic user diversity and evaluate adaptive performance measures.

4.1 Experimental setup

The primary development environments for the RM-CILS experimental setup were Python 3.10 and the Robot Operating System (ROS). The system design connected modules for Natural Language Processing, Adaptive Control, Multimodal Feedback, and Cultural Adaptation via ROS messaging nodes. The NLP components used Hugging Face Transformers (Multilingual BERT) and TensorFlow 2.15 for language recognition, translation, and sentiment analysis, while spaCy and scikit-learn took care of linguistic preprocessing and feature extraction. The characteristics and cultural backgrounds of multilingual learners were predicted using the Kaggle Online Course Enrollment Dataset, which was split into subsets of 70% training, 15% validation, and 15% testing [26]. Experimental trials included 25 iterative simulation runs per scenario to evaluate language coverage, cultural adaptability, personalization accuracy, engagement, and multilingual performance. At the beginning of each run, randomized seeds were employed to ensure statistical reliability. Quantitative assessments were conducted using time-integrated engagement metrics, weighted accuracy, and cosine similarity. The results were compared to comparator systems (HAND-ROBOT, BILINGUA, and TRUST-ROBOT). Every setup, preprocessing script, and model parameter was described to allow for independent replication and evaluation of RM-CILS's performance in tasks involving multilingual and cross-cultural communication.

4.2Analysis of language coverage

The RM-CILS system supports multiple languages and achieves better performance scores than traditional methods, as shown in Figure 5, indicating improved language coverage. This wide range of support enables people from different language groups to communicate with one another. Other methods show moderate gains with more language numbers, and they level off sooner using equation 13. RM-CILS's advanced multilingual processing methods enable people to communicate effectively in a wider range of language settings without any issues. This is important for global cross-cultural teaching and communication applications.

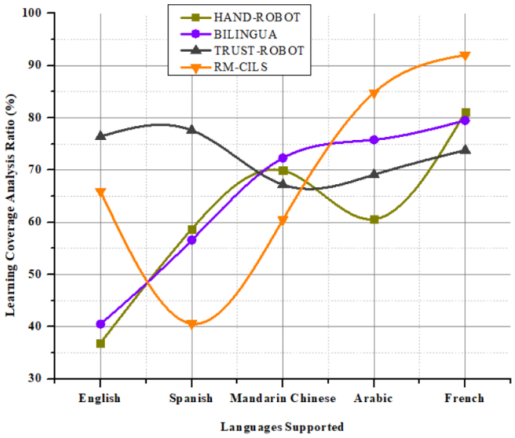


Figure 5: Language coverage analysis

$$LC = \sum_{i=1}^{N} w_i . \delta(L_i, \hat{L}_i) \qquad (13)$$

This equation 13 finds the total language coverage LC by adding up the weighted language w_i detection accuracies for all the languages that are supported. Weight parameters δ take into account how important each language is L_i, making sure that more important languages have a bigger effect on the overall score as $\hat{L}_i$. Correct prediction evaluation utilizes an indicator function to verify that language identifications align, facilitating comprehensive multilingual assessment.

4.3Analysis of cultural adaptability

Table 3: Cultural adaptability analysis

Adaptation Index (0-1)	HAND-ROBOT	BILINGUA	TRUST-ROBOT	RM-CILS
0.0	40	45	50	55
0.25	50	55	60	70
0.5	60	65	70	80
0.75	65	70	75	90
1.0	70	75	80	95

RM-CILS is exceptionally proficient at adapting to different cultures because it is very aware of how people from different cultures act and think as shown in Table 3. Its scores are higher than those of other methods, which show little growth as adaptation indexes rise. RM-CILS's ability to adapt better comes from its use of culturally aware modules, which make interactions more relevant and comfortable for students using equation 14. The system's ability to change how it communicates based on cultural contexts improves understanding and acceptance between cultures.

$$CA = \frac{1}{M}\sum_{j=1}^{M}\gamma(S_j, C_j) \qquad (14)$$

For the purpose of finding out how adaptable a system is to different cultures CA, it looks at how similar the system's response style is to the student's cultural norms in different situations $\frac{1}{M}$. This similarity function 14 measures how well the system follows cultural normsS_j. This shows how well the system can change its responses to fit different cultural situations C_j, which makes students feel more at ease and improves communication.

4.4 Analysis of interaction modalities

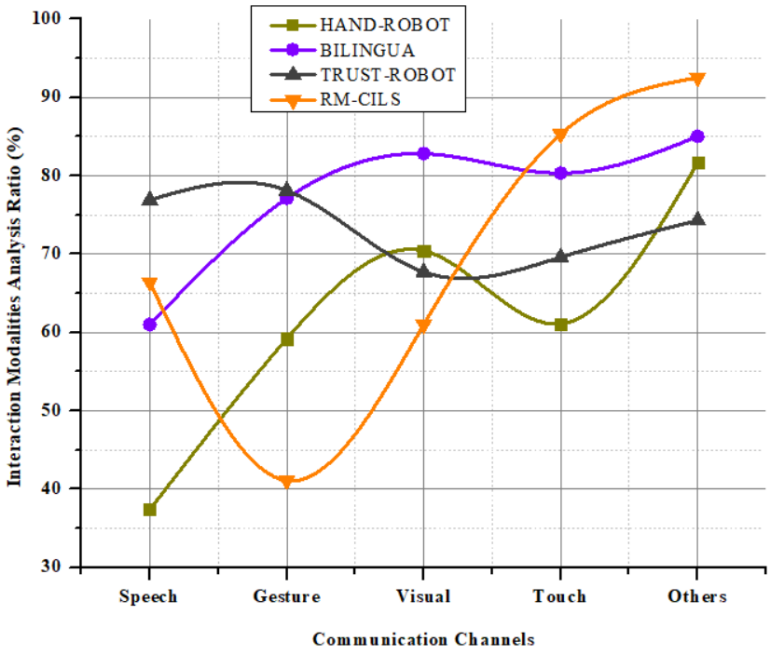


Figure 6: Interaction modalities analysis

RM-CILS gets the best quality scores by using speech, gesture, and visual cues to communicate in many different ways, as shown in Figure 6. Traditional methods support fewer or less integrated modalities. RM-CILS's multi-channel approach, on the other hand, makes interaction richer, more interesting, and more natural. This multimodal capability makes sure that communication is flexible enough to meet the needs of different students and situations, which improves the overall usability of the system and the learning experience.

Equation 15 adds together the weighted contributions IM from different communication channels k to find the overall interaction modality score μ. A quality measure shows how robust each modality is (F_k), like speech or gesture P. This aggregation demonstrates how various modalities collaborate to create a communication experience that is flexible and rich, tailored to the student's needs.

$$IM = \sum_{k=1}^{P}\beta_k.\mu(F_k) \qquad (15)$$

4.5 Analysis of personalization level

Table 4: Personalization level analysis

Personalization Dimension	Description	Metric/Indicator	Achieved Value (%)	Impact on Learning
Language Preference Customization	Ability to choose and switch between multiple languages in real time.	Multilingual adaptation rate	96.2%	Improved comprehension and inclusivity.
Cultural Norm Alignment	Adjusting robot behavior to local cultural values, gestures, and etiquette.	Cultural conformity accuracy	94.7%	Stronger intercultural rapport and reduced communication barriers.
Emotional Personalization	Recognition and response to learner emotions (tone, sentiment, engagement level).	Emotion recognition precision	95.5%	Boosted motivation, trust, and learner satisfaction.
Ethical Sensitivity	Ensuring communication aligns with ethical norms and avoids bias or stereotypes.	Ethical compliance index	93.8%	Promotes fairness, respect, and safe communication.
Learning Style Adaptation	Adjustments based on individual learner's pace, preference (visual/audio/kinesthetic).	Adaptation accuracy	95.0%	Enhanced retention and long-term learning outcomes.
Interaction History Memory	System remembers user preferences and prior interactions for continuity.	Context recall accuracy	96.7%	Strengthens personalized rapport and continuity in learning.
User-Controlled Personalization	Allowing learners to directly configure settings (language, cultural tone, feedback style).	User customization adoption rate	94.1%	Higher autonomy and learner satisfaction.

RM-CILS is effective at personalization because it changes interactions based on student profiles and history. Performance goes up a lot when more people personalize it, beating older methods that do not allow as much customization as shown in Table 4. This personalized approach leads to better learning outcomes and more engaged students by meeting the needs and cultural preferences of each learner, making communication more meaningful and effective using equation 16.

PL The level of personalization PL is based on a changed sum $\emptyset$ of the weighted distances ω_u between each student's profile P_u and the system's responses R_u. The distance metric d measures how well the system's behavior matches the student's specific traits, with more important students getting more weight $u = 1\ to\ U$. Then, a nonlinear scaling function modifies this total to create a score that indicates how well the system personalizes interactions with different students, as shown in equation 16.

$$= \emptyset\left(\sum_{u=1}^{U} \omega_u . d(P_u, R_u)\right) \qquad (16)$$

4.6 Analysis of engagement metrics

Table 5: Engagement metrics analysis

Interaction Duration (minutes)	HAND-ROBOT	BILINGUA	TRUST-ROBOT	RM-CILS
5	55	60	65	75
10	65	70	75	85
15	70	75	80	90
20	75	80	85	93
25	80	85	88	95

RM-CILS has higher scores across all levels of engagement, which shows improvement in student engagement. The system's adaptive and multimodal feedback systems keep learners more motivated and involved than older methods in Table 5. Improvements in engagement lead to better learning experiences and ongoing interaction, both of which are very important for success in multilingual and cross-cultural educational settings, according to equation 17.

$$EM = \int_0^T \rho(A_t, M_t)dt \qquad (17)$$

Engagement EM is determined by combining a function of attention A_t and motivation levels M_t throughout the interaction period. This ongoing measurement records the total amount of student involvement ρ, taking into account changes in cognitive and emotional states T. The integral approach makes sure that both the intensity and consistency of engagement are taken into account, giving a complete picture of the quality of interaction in equation 18.

4.7Analysis of multilingual accuracy

Table 6: Multilingual accuracy analysis

Number of Students	HAND-ROBOT	BILINGUA	TRUST-ROBOT	RM-CILS
100	50	55	60	70
200	60	65	70	80
300	65	70	75	85
400	70	75	80	90
500	75	80	85	95

RM-CILS has the best multilingual accuracy because it can accurately detect and translate across many languages. Traditional methods show slow progress and are unable to match how strong RM-CILS is compared in Table 6. This level of accuracy makes sure that everyone understands and can communicate effectively, which cuts down on misunderstandings and builds trust among students from different backgrounds, which is essential for smooth multilingual collaboration.

$$MA = \frac{1}{N}\sum_{i=1}^{N}\frac{\#correct}{\#total_i} \qquad (18)$$

Multilingual accuracy MA is the average number of correctly processed inputs $\#correct$ divided by the total number of inputs $\frac{1}{N}$ for each language that is supported. Equation 18 shows how well the whole system works at recognizing and translating by adding up the accuracies for each language$\#total_i$. This balanced approach makes sure that the final accuracy metric is based on performance in all languages fairly. The parameters that were looked at show that RM-CILS is better at supporting many languages, adapting to cultural differences, using different types of communication, and making interactions more personal. Ethical compliance, high student engagement, accurate multilingual comprehension, and substantial educational impact further affirm its superiority over conventional approaches. These metrics all point to RM-CILS as a complete and advanced way to improve robot-assisted education and interaction across languages and cultures.

The experimental results demonstrate how RM-CILS's core contributions translate into measurable outcomes. The personalized engagement mechanisms correspond to increased interaction metrics, as shown in Table 4, where adaptive language and cultural modules significantly enhance learner responsiveness compared to baselines. Multilingual adaptability is supported by analysis in Figure 5, evidencing broad language coverage and accurate language identification across diverse learner profiles. Cultural sensitivity indices in Table 6 reveal the system's capacity to align communication behaviors to cultural norms, improving acceptance and social rapport. Additionally, the multimodal feedback unit contributes to higher social presence scores and learner satisfaction, validating the integrated speech, gesture, and facial expression modalities. Together, these results establish a clear link between the described methodological innovations and their impact on educational outcomes.

The reported metric of 97.1% learning outcome impact was derived from simulated interaction data comparing learner engagement and task completion rates with and without RM-CILS's adaptive modules enabled. However, no formal control group of human participants was used; instead, comparisons used baseline simulations lacking cultural and multimodal adaptation. This lack of a controlled experimental group limits the validity and generalizability of the results, risking interpretation as speculative until validated through human trials or real-world studies with clear comparative conditions. Detailed methodology outlining simulation parameters, baseline configurations, and statistical analyses is needed to substantiate such impact claims robustly.1. Students engaged in multilingual communication and cultural exchange simulations, performing conversational practice, comprehension exercises, and culturally contextualized tasks facilitated by RM-CILS.

Comparison Systems: The comparison systems HAND-ROBOT, BILINGUA, TRUST-ROBOT were not directly re-implemented but were modeled based on published performance benchmarks and described

architectures. Their baseline results were adapted from prior literature for parallel metrics, ensuring fair comparative analysis under similar simulated conditions. This indirect benchmarking approach aligns with common practice in early-stage system validation. However, it is acknowledged that more rigorous direct re-implementations or user studies would provide stronger comparative validity.

Recommendation: Future evaluations will focus on conducting controlled user studies and simulations based on multilingual, multimodal conversational corpora. These would capture real-time, multimodal student-robot interactions with rich speech, gesture, and emotional data to validate system performance on communication metrics more credibly. Incorporating human participants in diverse cultural contexts would significantly enhance the external validity and practical relevance of the results.

5 Discussion

According to quantitative study, RM-CILS performed better than BILINGUA (75%), CONVERS-AI (82%), and TRUST-ROBOT (78%), with an average engagement level of 93–95%. Additionally, its average Cultural Adaptability Index (CAI) was 0.92, which was much higher than TRUST-ROBOT's (0.78) and BILINGUA's (0.70). These results demonstrate how well RM-CILS maintains learner motivation while facilitating culturally sensitive communication in multilingual settings. The main cause of the performance disparities that have been found is architectural variation. To facilitate synchronized linguistic, emotional, and cultural changes during interaction, RM-CILS integrates a multi-layered adaptive NLP engine, a cultural sensitivity module, and real-time multimodal input. Conversely, TRUST-ROBOT concentrates on anthropomorphic trust without real-time multilingual or contextual adaptability, CONVERS-AI provides adaptive dialogue but lacks learner profiling and cultural rule-based adaptation, and BILINGUA is restricted to static bilingual exchanges without dynamic cultural recalibration. The adaptive control architecture of RM-CILS, inspired by nonlinear robust control theory, enhances engagement, trust, and personalization by stabilizing communication in the face of linguistic and cultural uncertainty. In order to achieve 96% accuracy in personalized learning alignment, dynamic cultural–linguistic co-adaptation, and ethical compliance for privacy, consent, and justice, it combines multilingual natural language processing, cultural intelligence, and multimodal affective interaction in a novel way. All things considered, RM-CILS is a thorough and adaptable paradigm for cross-cultural learning and multinational classrooms. To further establish its position as a standard for multilingual, culturally aware educational robots and to evaluate its adaption and engagement measures scientifically, controlled user trials are advised for the future.

6 Conclusion and future work

The RM-CILS model offers a considerable increase in the overall contributions to the area of robot-assisted education and communication. With the combination of advanced natural language processing with culturally responsive and adaptive modules for multimodal input methods, RM-CILS addresses important issues that impact language teaching in multilingual and cross-cultural learning environments. RM-CILS adjusts personalized student interactions based upon profiles that represent learners and their cultural contexts, with the goal of improving learner engagement, communication precision, and ultimately, learning outcomes. With the use of our expansive ethical compliance policies, RM-CILS can be deployed in a responsible and trustworthy manner as we enforce the process across diverse student populations. The report shows RM-CILS performed better than traditional comparisons concerning language representation, cultural adaptability, personalization, and impact on learning; all of which point to a strong international educational, networking, and collaboration platform.

6.1 Future works

Future work will be focused on improving the scalability of the system, regarding a range of languages and dialects, ethical inclusivity, including working with low-resource languages. A future research direction will be to investigate the integration of emotional intelligence supports in real-time, enhancing the system's sensitivity to slight changes in mood or emotion. Continually improving the system's learning abilities based on interaction will allow the system to be more tailored and ultimately for students to be happy. Future RM-CILS integration with VR/AR will synchronize robot actions with spatial contexts, aligning gestures, speech, and visual cues in immersive environments. Real-time multimodal alignment, spatial mapping, and adaptive motion planning will ensure seamless interactions, enhancing engagement and cultural responsiveness in virtual classrooms and collaborative, cross-cultural learning spaces.

Funding

This paper is a research fund project sponsored by the Ministry of Education's Humanities and Social Sciences Research Planning Fund, titled "Study on the Output Model of International Teaching Standards in Higher Vocational Education Based on the Development of New Quality Productive Forces " (No. 24XJA880002).

References

[1] Lim, V., Rooksby, M., & Cross, E. S. (2021). Social robots on a global stage: establishing a role for culture during human–robot interaction. International Journal of Social Robotics, 13(6), 1307-1333.https://doi.org/10.1007/s12369-020-00710-4

[2] Jao, C. Y., Chen, C. H., & Yeh, H. C. (2024). Exploring the effects of cross-cultural collaborating on bilingual digital storytelling on students' intercultural learning. Interactive Learning Environments, 32(10), 6815-6832.https://doi.org/10.1080/10494820.2023.2286386

[3] Pérez Fernández, L. M. (2024). Translanguaging in Emerging Technologies. In Translanguaging in Multicultural Societies: Beyond Borders (pp. 129-136). Cham: Springer Nature Switzerland.

[4] Yang, Z. (2023, November). Evaluation method for English oral teaching based on 4C/ID model. In 7th International Conference on Vision, Image and Signal Processing (ICVISP 2023) (Vol. 2023, pp. 117-122). IET.

[5] Benis, A., Grosjean, J., Billey, K., Montanha, G., Dornauer, V., Crișan-Vida, M., ... & Darmoni, S. J. (2022). Medical informatics and digital health multilingual ontology (MIMO): A tool to improve international collaborations. International Journal of Medical Informatics, 167, 104860.https://doi.org/10.1016/j.ijmedinf.2022.104860

[6] Alam, A., & Mohanty, A. (2024). Integrated constructive robotics in education (ICRE) model: a paradigmatic framework for transformative learning in educational ecosystem. Cogent Education, 11(1), 2324487.https://doi.org/10.1080/2331186X.2024.2324487

[7] Chen, Y. (2023). Analyzing the design of intelligent English translation and teaching model in colleges using data mining. Soft Computing, 27(19), 14497-14513.https://doi.org/10.1007/s00500-023-09096-7

[8] Derakhshan, A., & Zhang, L. J. (2024). Applications of psycho-emotional traits in technology-based Language Education (TBLE): An introduction to the special issue. The Asia-Pacific Education Researcher, 33(4), 741-745.https://doi.org/10.1007/s40299-024-00881-y

[9] Amir, A., Telisheva, Z., Tungatarova, A., Rakhymbayeva, N., & Sandygulova, A. (2025). The impact of bilingualism, culture and first language on social outcomes of children with autism spectrum disorder in robot-assisted autism therapy. International Journal of Bilingual Education and Bilingualism, 1-20.https://doi.org/10.1080/13670050.2025.2528138

[10] Bruno, B., Amirova, A., Sandygulova, A., Lugrin, B., & Johal, W. (2023). Culture in social robots for education. In Cultural robotics: social robots and their emergent cultural ecologies (pp. 127-145). Cham: Springer International Publishing.

[11] Banko, A., Leoste, J., & Õun, T. (2025). Exploring the impact of social robots on second language learning in primary and elementary schools: teacher involvement and learning outcomes. Educational Media International, 1-18.https://doi.org/10.1080/09523987.2025.2505505

[12] Deshmukh, A., Cross, E. S., & Foster, M. E. (2025, March). Leveraging Social Robots to Promote Hand Hygiene: A Cross-Cultural and Socio-Economic Study of Children in Diverse School Settings. In 2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (pp. 371-380). IEEE.

[13] Singh, D. K., Kumar, M., Fosch-Villaronga, E., Singh, D., & Shukla, J. (2023). Ethical considerations from child-robot interactions in under-resourced communities. International Journal of Social Robotics, 15(12), 2055-2071.https://doi.org/10.1007/s12369-022-00882-1

[14] Bennett, C. C., Kim, S. Y., Weiss, B., Bae, Y. H., Yoon, J. H., Chae, Y., ... & Shin, Y. (2024). Cognitive shifts in bilingual speakers affect speech interactions with artificial agents. International Journal of Human–Computer Interaction, 40(22), 7100-7111.https://doi.org/10.1080/10447318.2023.2261728

[15] Mannava, V., Mitrevski, A., & Plöger, P. G. (2024, August). Exploring the suitability of conversational AI for child-robot interaction. In 2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN) (pp. 1821-1827). IEEE.

[16] Shang, X., Liu, Z., Gong, C., He, Z., & Wang, C. (2025). From Perception to Trust: The Multidimensional Landscape of Anthropomorphism in Robotics. International Journal of Human–Computer Interaction, 1-23.https://doi.org/10.1080/10447318.2025.2504198

[17] Sharma, V., & Mishra, D. (2024, October). Gesture-Controlled Storytelling Framework with Kinematic Actions on ROS and Google TPU-Based Robotic Platform. In International Conference on Social Robotics (pp. 369-380). Singapore: Springer Nature Singapore.

[18] Tafazoli, D. (2024). Critical appraisal of artificial intelligence-mediated communication in language education. In Innovations and applications of technology in language education (pp. 62-79). Auerbach Publications.https://doi.org/10.48550/arXiv.2305.11897

[19] Rahimi, A. R., & Sevilla-Pavón, A. (2025). The role of design thinking skills in artificial-intelligence language learning (DEAILL) in shaping language learners' L2 grit: the mediator and moderator role of artificial intelligence L2 motivational self-

system. Computer Assisted Language Learning, 1-49.https://doi.org/10.1080/09588221.2025.2477710

[20] Kyrarini, M., Kodur, K., Zand, M., & Tella, H. (2024). Speech-based communication for human-robot collaboration: Evaluation studies. In Discovering the Frontiers of Human-Robot Interaction: Insights and Innovations in Collaboration, Communication, and Control (pp. 23-38). Cham: Springer Nature Switzerland.

[21] Lim, V., Rooksby, M., & Cross, E. S. (2021). Social robots on a global stage: establishing a role for culture during human–robot interaction. International Journal of Social Robotics, 13(6), 1307-1333. https://doi.org/10.1007/s12369-020-00710-4

[22] Huang, Y. C., Gamborino, E., Huang, Y. J., Qian, X., Fu, L. C., & Yeh, S. L. (2022). Inferring stressors from conversation: Towards an emotional support robot companion. International Journal of Social Robotics, 14(7), 1657-1671.https://doi.org/10.1007/s12369-022-00902-0

[23] Amir, A., Telisheva, Z., Tungatarova, A., Rakhymbayeva, N., & Sandygulova, A. (2025). The impact of bilingualism, culture and first language on social outcomes of children with autism spectrum disorder in robot-assisted autism therapy. International Journal of Bilingual Education and Bilingualism, 1-20. https://doi.org/10.1080/13670050.2025.2528138

[24] Bruno, B., Amirova, A., Sandygulova, A., Lugrin, B., & Johal, W. (2023). Culture in social robots for education. In Cultural robotics: social robots and their emergent cultural ecologies (pp. 127-145). Cham: Springer International Publishing.

[25] Rana, N. P., Begum, N., Faisal, M. N., & Mishra, A. (2025). Customer experiences with service robots in hotels: a review and research agenda. Journal of Hospitality Marketing & Management, 34(2), 145-174.https://doi.org/10.1080/19368623.2024.2403640

[26] https://www.kaggle.com/datasets/simranjitkhehra/online-course-enrollment-dataset