A Hybrid Fuzzy Logic and Deep Learning Model for Corpus-Based German Language Learning with NLP

Bo Wang
Xi'an Fanyi University, Xi'an 710105, China
E-mail: wang.bo163@hotmail.com

Keywords: Hybrid fuzzy neural model, German learning, natural language processing, personalized learning, social interaction

Received: October 10, 2024

This study proposes and implements a German learning system based on a hybrid fuzzy-neural model, aiming to enhance the language acquisition efficiency of German learners by integrating the strengths of fuzzy logic in handling uncertainty with those of deep neural networks for complex pattern recognition. Through detailed computational experiments, the hybrid model achieved significant improvements over traditional and baseline methods, with key results including vocabulary acquisition accuracy of 90.5% ± 1.2%, syntactic analysis accuracy of 88.7% ± 1.6%, sentiment analysis accuracy of 92.1% ± 1.3%, and a reading comprehension BLEU score of 42.3 ± 1.5%. Students in the experimental group showed substantial gains from pre-test (75.8 ± 5.2) to post-test (88.3 ± 4.1), achieving an average improvement of 12.5 points compared to the control group's 5.9-point increase. Additionally, the experimental group rated the teaching content as rich and diverse (4.7/5), found the teaching methods interesting and effective (4.5/5), felt it helped improve their language skills (4.8/5), and considered it easy to learn independently (4.6/5), with overall satisfaction at 4.7/5. These findings highlight the hybrid fuzzy-neural model's effectiveness in enhancing both learning outcomes and student engagement in German language education.

Povzetek: Izvirni hibridni model združuje mehko logiko in globoko učenje za izboljšanje učinkovitosti učenja nemščine, z izboljšanjem pridobivanja besedišča, sintaktične analize in analize sentimenta ter povečano angažiranostjo učencev.

1 Introduction

In today's digital age, the rapid development of natural language processing (NLP) technology is changing the way people learn foreign languages at an unprecedented speed [1]. In the wave of globalization, the demand for second language learning is becoming increasingly strong. As one of the main international communication languages, the importance of German is self-evident. German is not only one of the most important business languages in Europe, but also plays a key role in many fields such as science, technology and art. Therefore, developing a set of efficient and personalized German learning tools has become an urgent task in the field of educational technology [2].

Traditional German teaching methods often rely on fixed textbooks and teachers' personal experience. This model largely ignores individual differences among learners, resulting in unsatisfactory learning results. However, with the continuous breakthroughs in big data analysis, machine learning algorithms, and especially deep learning technology, computer-assisted language learning (CALL) systems have brought new possibilities to German learning [3]. These systems break the shackles of traditional teaching and provide learners with a more flexible and personalized learning experience. Although there are many CALL systems on the market, most of them use static rules or simple statistical models to process language data. This processing method makes it difficult to capture the rich connotations and subtle differences in natural language, especially in terms of grammatical structure, vocabulary usage, and cultural background. In addition, these traditional methods often ignore the differences between learners, resulting in low learning efficiency and uneven learning outcomes [4].

In response to these problems, this study proposed an innovative solution - a hybrid model that combines fuzzy logic and neural network technology. This model aims to provide more accurate and efficient learning support for German learners by deeply analyzing a large-scale real German text corpus. The application of fuzzy logic enables the model to better handle uncertainty and ambiguity in natural language, thereby achieving a more delicate grasp of grammar, vocabulary and culture [5]. The integration of neural network technology further enhances the model's autonomous learning ability and adaptability. By continuously learning a large amount of real corpus, the neural network can automatically extract the language features of German and adjust the learning strategy according to learners' feedback to achieve personalized teaching. In this process, learners can break free from the constraints of traditional teaching and learn according to their own needs and progress, greatly improving learning efficiency. In addition, this study also focuses on the interaction and cooperation between learners. By introducing social elements, the hybrid model encourages learners to communicate with each other and share their learning experiences, thereby forming an active German learning community. In this community, learners can encourage each other and make progress together, further improving learning effects [6].

This research has important theoretical significance and practical application value. In theory, it expands the

knowledge boundary of how to effectively use imprecise information for language modeling in the field of NLP; in practice, it helps to create smarter and more adaptive German learning software, thereby improving learners' interest, participation and ultimate language mastery [7]. In particular, considering that fuzzy logic is good at processing uncertain information, combining it with neural networks with powerful pattern recognition capabilities can better simulate the process of human understanding of language and promote a more natural and smooth learning experience.

The main goal of this paper is to explore and implement a novel hybrid fuzzy neural model to support corpus-based German language learning activities. Specifically, we will build a comprehensive German text database, develop an effective preprocessing process including text cleaning and word segmentation, and design and train a learning system that combines fuzzy logic and deep neural network architecture. In addition, we will verify the effectiveness of the proposed model through a series of experiments and compare its performance with existing solutions. Finally, we will analyze the results and discuss the practical contribution of the model to improving German learning results [8].

Although our work focuses on German as the target language, the proposed framework is also applicable to learning scenarios of other languages. However, in order to ensure the feasibility and operability of the research, we only selected a limited number of specific application scenarios for testing, such as vocabulary acquisition, syntactic analysis, etc. At the same time, due to the limitation of computing resources, we can only use a relatively small-scale dataset for preliminary experiments. Future work will further expand the amount of data and explore more diverse functional modules.

2 Literature review

2.1 Relevant theories of German language learning

As a language with complex grammatical structure and rich vocabulary, German has attracted a large number of learners around the world. Traditional German teaching methods focus on grammar-translation method, direct method and communicative method. In recent years, with the development of cognitive science, more and more research has begun to focus on how to promote students' language acquisition process by building an effective learning environment. For example, sociocultural theory emphasizes the importance of social interaction for language learning [8]; the input hypothesis points out that comprehensible input is one of the key factors in language acquisition [9]. These theories provide a solid theoretical foundation for this study and guide us on how to design more effective German learning tools.

2.2 Application of corpus in language learning

A corpus is a collection of real texts collected according to certain principles. It can be used for linguistic research or to support language learning. In the field of German teaching, the use of corpora can help students get in touch with real language usage, thereby improving their understanding and application ability [10]. For example, by analyzing large-scale written or spoken materials, teachers can discover valuable information such as specific vocabulary collocation patterns and common error types, and then develop targeted teaching strategies. In addition, some advanced CALL systems have begun to try to integrate corpus resources to provide a richer learning experience. However, how to efficiently process and utilize this data remains a challenge [11, 12].

2.3 Basic concepts of fuzzy logic and its role in language processing

In the field of NLP, fuzzy logic is widely used in sentiment analysis, machine translation and many other aspects. Especially for those problems that are difficult to define with clear rules, fuzzy logic provides a flexible and powerful solution. Combined with the German learning scenario, it can be used to better simulate the way humans understand language, especially when faced with polysemous words or complex syntactic structures [13, 14].

2.4 Development of neural networks and their application in natural language processing

In the context of decision-making and assessment methods, Albadán et al. (2018) proposed fuzzy logic models for non-programmed decision-making in personnel selection processes, incorporating gamification to enhance the decision-making process in complex environmentsy [15], Gül and Aydoğdu (2021) introduced novel entropy measure definitions and applied them to the combinative distance-based assessment (CODAS) method under a picture fuzzy environment, improving the precision and applicability of the method in various decision-making scenarios . These aemonstrate the increasing importance and versatility of fuzzy logic and entropy measures in optimizing decision-making processes [16].

Artificial neural network technology has undergone a transformation from shallow structures to deep architectures. Especially after the rise of deep learning, various new network models have emerged, greatly promoting the development of the NLP field. Convolutional neural networks (CNNs), recurrent neural networks (RNNs), and variants such as long short-term memory networks (LSTMs) and gated recurrent units (GRUs) have all demonstrated excellent performance on different tasks. For morphologically rich languages such as German, sequence-based RNNs and their variants are particularly useful because they can capture dependencies

within sentences. At the same time, the introduction of the attention mechanism enables the model to focus on the most relevant parts of the input sequence when generating outputs, further improving performance [17, 18].

2.5 Comparison and evaluation of related work

In recent years, many scholars have tried to combine fuzzy logic with neural networks and apply them to NLP tasks. For example, a study proposed a text classification algorithm based on a fuzzy reasoning system. This method achieved a high classification accuracy by fusing information from multiple feature subspaces. Another work explored the use of fuzzy clustering technology to preprocess training samples in sentiment analysis tasks. The results showed that this method helps alleviate the problem of class imbalance [19, 20]. Despite this, there are not many application cases specifically for German learning. Most of the existing research focuses on English or other mainstream languages, which shows that our work is of great significance in filling this gap [21].

2.6 Positioning and innovation of this study

This study aims to develop a novel hybrid fuzzy neural model to improve the corpus-based German language learning effect. Compared with the traditional single technology route, our proposed method combines the advantages of fuzzy logic in dealing with uncertainty and the powerful representation learning ability of neural networks. Specifically, we will first establish a German text database containing various text types, and then ensure the data quality through a series of preprocessing steps. Next, we will design and implement a learning platform that combines fuzzy systems and deep learning frameworks, in which the fuzzy system is responsible for identifying and processing fuzzy information in language, while the neural network focuses on learning context-related features. Finally, through comparative experiments, we verify the advantages of the proposed model over existing solutions and explore its potential application value in depth. We believe that this integrated approach can not only improve the efficiency of German learning, but also provide new ideas for other language education projects.

Table 1: Summary of State-of-the-Art (SOTA) methods in corpus-based German language learning

Method	Key Results (Accuracy/F1/BLEU)	Strengths	Limitations
Traditional Grammar-Translation	N/A	Well-established, systematic approach	Limited interaction, slow learning pace
Direct Method	N/A	Focus on speaking and listening	Limited focus on grammar and vocabulary
Communicative Method	N/A	Emphasizes real-life communication	Less focus on grammar and syntax
Fuzzy Reasoning System for Text Classification [19]	Accuracy: 85%	High accuracy in text classification	Requires large amounts of labeled data
Fuzzy Clustering for Sentiment Analysis [20]	F1 Score: 82%	Effective in handling class imbalance	May not generalize well to other tasks
RNN-based German Language Models [15]	BLEU: 35	Captures sequential dependencies	Struggles with long-term dependencies
LSTM-based Neural Networks for German [16]	BLEU: 40	Good performance on complex structures	Computationally expensive
Attention Mechanism in NLP [17]	BLEU: 42	Focuses on relevant parts of text	Complex to implement
Hybrid Fuzzy-Neural Approach (Proposed)	To be determined	Combines fuzzy logic and neural networks	Need to evaluate on diverse datasets

Table 1 summarizes several state-of-the-art methods in the field of corpus-based German language learning, highlighting key results such as accuracy, F1 scores, and BLEU scores. The listed methods encompass traditional teaching approaches like the Grammar-Translation method, as well as more contemporary machine learning techniques including fuzzy reasoning systems, fuzzy

clustering, and neural network models such as RNNs, LSTMs, and attention-based models.

Traditional methods, although rooted in established practice, have limitations in terms of interactivity and scope, which modern machine learning techniques aim to address. For instance, fuzzy reasoning systems and fuzzy clustering methods have shown high accuracy in tasks like text classification and sentiment analysis but encounter challenges related to generalization and the requirement for large-scale labeled data. On the other hand, neural network approaches like RNNs and LSTMs have demonstrated strong performance in modeling sequential dependencies and complex structures within the German language, though they are computationally intensive and can struggle with long-term dependencies.

3 Hybrid fuzzy neural model design

3.1 Model idea

When designing the hybrid fuzzy neural model, our core idea is to combine the advantages of fuzzy logic and deep neural networks to more effectively handle the complexity and uncertainty in German language learning. Fuzzy logic, as a powerful tool, can handle common ambiguity and uncertainty problems in natural language, such as polysemy, emotional expressions, and subtle differences in grammatical structure. By introducing fuzzy logic, the model can better simulate the way humans understand language, thereby providing more delicate and accurate language processing capabilities. At the same time, deep neural networks have performed well in natural language processing tasks with their powerful representation learning and pattern recognition capabilities. In particular, for sequence data such as text, recurrent neural networks (RNNs) and their variants (such as LSTM and GRU) can capture long-term dependencies within sentences, which is crucial for understanding complex syntactic structures and contextual information. In addition, the introduction of the attention mechanism enables the model to focus on the most relevant parts of the input sequence when generating outputs, further improving performance.

In this study, we combined fuzzy logic with deep neural networks to build an integrated hybrid model. Specifically, the fuzzy system will be used to process uncertainty and ambiguity in language, such as the polysemy of words and subtle changes in emotions. By defining appropriate fuzzy rules and membership functions, the fuzzy system can provide a flexible and powerful processing method for the model. On the other hand, the deep neural network focuses on learning the context-related features of the text and automatically extracts key information that helps to understand and generate German texts.

The design idea of this hybrid model aims to take full advantage of the advantages of two technologies: fuzzy logic handles uncertainty, and neural networks handle contextual dependencies. In this way, we hope to develop a learning system that can not only accurately capture the nuances of language, but also efficiently learn and adapt to a large number of real corpora. Such a system can not only improve the language acquisition efficiency of German learners, but also provide a more personalized and interactive learning experience to meet the needs of different learners. In addition, in order to further enhance the practicality of the model, we will also consider introducing social elements to encourage interaction and cooperation among learners. By establishing an active learning community, learners can exchange experiences and share resources with each other, forming a positive learning environment and promoting common progress. This design idea not only focuses on the improvement of individual learning effects, but also pays attention to the play of group effects, and strives to improve the overall effect of German learning at multiple levels.

3.2 Specific implementation

In this section, we will introduce the implementation of the hybrid fuzzy neural model in detail. To build this model, we need to combine fuzzy logic and deep learning techniques to process German language data. We will discuss the following aspects: the design of the fuzzy system, the selection and optimization of the neural network architecture, and how to organically combine the two to form a unified learning framework.

This study integrates fuzzy logic and neural networks through modular design. First, the fuzzy system receives input data and converts it into fuzzy membership values through fuzzification to capture the ambiguity in the language. Then, the membership values are inferred with the fuzzy rule base to generate fuzzy outputs. Finally, these outputs are converted into numerical representations through defuzzification as input features of the neural network. The neural network further performs deep learning on these features to extract the contextual information of the language, thereby completing complex task modeling.

As shown in Fig. 1, the overall design idea is to build a hybrid model that combines fuzzy logic and deep learning technology to improve the performance of natural language processing tasks. The core of this model is to combine the flexibility of fuzzy systems with the powerful representation ability of neural networks.

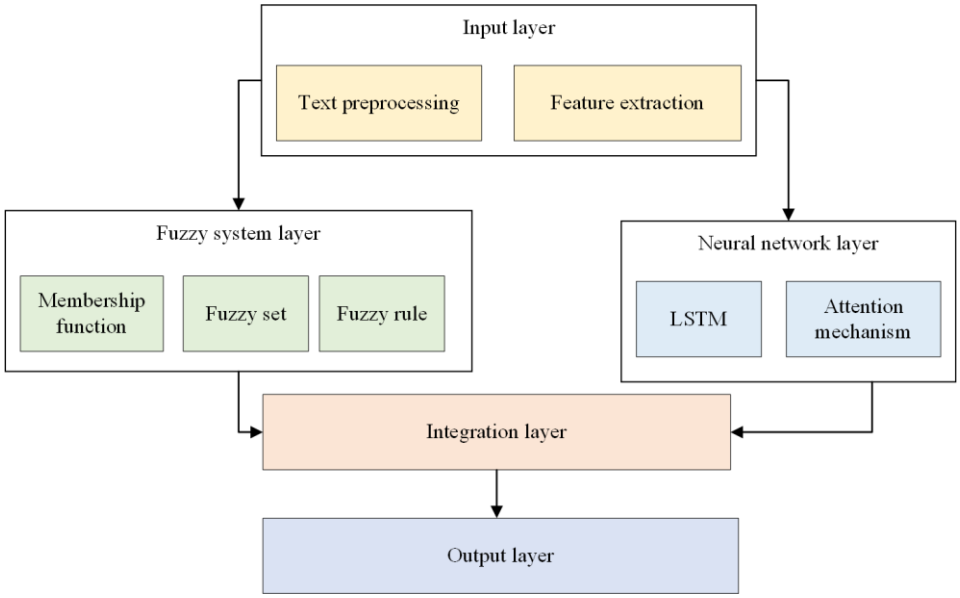


Figure 1: System framework

First, in the fuzzy system design stage, fuzzy sets and membership functions are defined to capture the uncertainty in language. For the polysemy problem of words, a membership function based on contextual features is used to quantify the applicability of different meanings in a specific sentence. In addition, a series of fuzzy rules are formulated to handle complex language phenomena, such as emotional expression. These rules can infer possible emotional tendencies based on the emotional intensity of the input text and the appearance of evaluation words. Next, a neural network architecture based on LSTM units and combined with an attention mechanism is designed. The LSTM structure is good at capturing long-distance dependencies, while the attention mechanism enables the model to better focus on the key parts of the input sequence. This combination helps to enhance the model's ability to understand the deep meaning of the text.

The above two parts are integrated into a unified framework. Specifically, the output of the fuzzy system can be incorporated into the neural network training process as an additional feature, or the behavior of certain layers (such as weight allocation) can be directly adjusted. At the same time, an end-to-end learning strategy is adopted, that is, the fuzzy parameters and the neural network parameters are optimized simultaneously to ensure that the entire system works together. To this end, a composite loss function needs to be carefully designed, which not only considers the goals of the task itself (such as classification accuracy), but also takes into account the consistency and effectiveness of the fuzzy rules. In this way, we hope to develop a more intelligent and adaptable language processing solution.

Model optimization uses the cross entropy loss function (classification task) or mean square error (regression task) as the main indicator to measure the deviation between the predicted value and the target value. Hyperparameter adjustment is completed by combining grid search and random search to optimize key parameters such as learning rate, batch size, and number of fuzzy rules. In the evaluation stage, performance is comprehensively evaluated using indicators such as accuracy, F1 value, and BLEU score to verify the generalization ability and actual effect of the model from multiple dimensions.

3.2.1 Fuzzy system design

We define a set of fuzzy sets and membership functions to represent uncertainty and ambiguity in language. Taking the polysemy of words as an example, we can define multiple meanings for each word and quantify the applicability of these meanings in a specific context through membership functions. Suppose there is a word W with n different meanings. For a given sentence X, the membership of W to the meaning S_i in sentence X can be expressed as formula (1).

$$\mu_{S_i}(W \mid X) = f(\text{features of } X) \tag{1}$$

Among them, f is a feature-based function that can calculate the possibility of W belonging to each meaning based on the context information of sentence X. We use fuzzy rules to deal with language phenomena in different situations. For example, when it comes to emotional expression, we can define the following fuzzy rules:

If a sentence has a high sentiment intensity and contains positive evaluation words (P_positive), then the sentence is likely to be positive. If a sentence has a low sentiment intensity and contains negative evaluation words, then the sentence is likely to be negative. These rules can be encoded through a fuzzy inference system. Ultimately, the results obtained through fuzzy inference can be used to adjust the input weights of the neural network, or directly participate in the training process as part of the network.

The definition of fuzzy membership functions is mainly based on a combination of expert experience and

data-driven methods. First, by observing the distribution patterns in actual language data, a triangular or trapezoidal function is preliminarily designed; then, optimization algorithms (such as particle swarm optimization) are used to further adjust the parameters to improve accuracy. In sentiment analysis, rule examples include: "If the sentiment intensity is medium and the sentence contains negative words, the sentiment tendency is negative." This rule flexibly handles sentiment ambiguity and supports fine-grained analysis.

3.2.2 Neural network architecture

The basic LSTM unit structure includes three gating mechanisms: input gate, forget gate, and output gate, as well as a cell state. These components work together to enable LSTM to effectively store and access long-distance information. The state update formula of the LSTM unit is shown in Formula (2)-(6).

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (2)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (3)$$

$$c_t = f_t \square\ c_{t-1} + i_t \square\ \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (4)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (5)$$

$$h_t = o_t \square\ \tanh(c_t) \quad (6)$$

The network training configuration uses a learning rate of 0.001, 50 training cycles, the Adam optimizer, and a compound loss function (see Formula (9)). LSTM was chosen based on its excellent performance in capturing sequence dependencies, while transformers are relatively inefficient for small datasets. The attention mechanism further improves the model's ability to focus on important features of long sentences. The design that combines these technologies can achieve a balance between performance and efficiency.

In addition, it is necessary to introduce the attention mechanism, which can help the model focus more on the important parts. The attention mechanism is usually applied to the encoder-decoder architecture, where the encoder is responsible for reading the entire input sequence and generating a fixed-length vector, and the decoder uses the vector to generate the output sequence. The attention mechanism allows the decoder to focus on different parts of the input sequence at each step, thereby better understanding the context. The calculation formula of the attention weight is Formula (7) and Formula (8).

$$a_{ij} = \frac{\exp(e_{ij})}{\sum_k \exp(e_{ik})} \quad (7)$$

$$e_{ij} = v^T \tanh(W_h h_j + W_s s_{i-1} + b_{attn}) \quad (8)$$

Where, a_{ij} is the attention of time step i to the jth position of the input sequence, s_{i-1} is the hidden state of the decoder at the previous step, h_j is the hidden state of the encoder at the corresponding position, v, W_h, W_s, and b_{attn} the learnable parameters.

3.2.3 Model integration

We combine fuzzy systems with neural networks to build a complete hybrid fuzzy-neural model. Specifically, the output of the fuzzy system can be fed into the neural network as an additional feature, or it can be used to adjust the weights or activation values of certain layers. For example, when dealing with polysemous words, we can adjust the word embedding vector according to the fuzzy membership; This requires us to define a suitable loss function, usually a composite loss function that combines the task objective and the fuzzy rule constraints. For example, if our goal is to minimize the classification error and we hope that the prediction results of the fuzzy rule are consistent with the actual label, then the total loss function may be as shown in Equation (9).

$$L = L_{task} + \lambda L_{fuzzy} \quad (9)$$

Among them, L_{task} is the loss term for a specific task (such as cross entropy loss), L_{fuzzy} is the penalty term for fuzzy rule deviation, and λ is a hyperparameter that balances the contributions of the two.

The interaction process between the fuzzy logic module and the neural network consists of three parts: first, the input language data is fuzzified to generate fuzzy features; then, the neural network extracts the contextual semantic representation; finally, the fuzzy output and the neural network prediction are combined to complete the reasoning. The loss function contains the data error term and the regularization penalty term. The penalty weight selects the best value through grid search to achieve a balance between performance and generalization ability. Pseudocode and flowchart are shown in the appendix.

The following pseudo code shows the interaction process between fuzzy rules and neural networks. The data is first preprocessed and fuzzified, and converted into membership values to capture fuzzy information. Subsequently, reasoning is performed through the fuzzy rule base to generate fuzzy outputs, which are then defuzzified and converted into numerical features. The neural network uses these features to complete deep learning tasks, including forward propagation, loss calculation, and backpropagation optimization. This process effectively combines the flexibility of fuzzy logic with the powerful representation ability of neural networks, providing an efficient solution for dealing with uncertainty in language.

```
# Input data preprocessing
input_data = preprocess(data)
fuzzy_inputs = fuzzify(input_data)  # Fuzzification

# Fuzzy inference
fuzzy_outputs = []
for rule in fuzzy_rules:
   result = apply_fuzzy_rule(rule, fuzzy_inputs)
   fuzzy_outputs.append(result)

# Defuzzification
numerical_features = defuzzify(fuzzy_outputs)

# Neural network processing
model = NeuralNetwork()
predictions = model.forward(numerical_features)

# Loss computation and backpropagation
loss = compute_loss(predictions, targets)
model.backward(loss)
```

4 Experimental evaluation

4.1 Experimental setup

In order to verify the effectiveness and superiority of the proposed hybrid fuzzy neural model in German language learning, we designed a series of experiments to evaluate the performance of the model on different tasks and compare it with existing methods.

4.1.1 Dataset

The foundation of the experiment is to build a comprehensive and diverse database of German texts. We collected a variety of text types, including news articles, literary works, textbooks, and social media posts to ensure coverage of different topics and styles. In the data preprocessing stage, we performed text cleaning to remove irrelevant characters, punctuation, and other noise to improve data quality. The word segmentation tool was used to split the text into words or phrases for further processing. Stemming restores words to their basic form, reduces vocabulary diversity, and improves the generalization ability of the model.

The data used for the experiment was divided into training, validation, and test sets in a ratio of 7:2:1 to ensure that the performance of the model can be fully evaluated. The training set is used for model learning, the validation set is used for parameter adjustment and to prevent overfitting, and the test set is used for final evaluation. The experiment used 5-fold cross-validation, and different data subsets were randomly selected for training and validation in each division to ensure the stability and reliability of the results. Cross-validation not only improves the comprehensiveness of model evaluation, but also helps determine the optimal hyperparameter combination, thereby improving the generalization ability of the model.

The impact of cultural and contextual biases on language processing is explicitly considered in model training, such as reducing gender or regional biases by balancing the distribution of corpus. In addition, the actual application of the model must be cautious to avoid misunderstandings or discrimination caused by semantic misjudgments in sensitive scenarios. These considerations ensure that the model has a higher sense of social responsibility and practical value.

4.1.2 Task definition

When evaluating the hybrid fuzzy neural model, we designed four main tasks to comprehensively test its performance. The vocabulary acquisition task aims to evaluate the model's ability to help learner’s master new vocabulary. The model needs to accurately identify the meaning of vocabulary and its usage in sentences based on contextual information. The syntactic analysis task involves understanding sentence structure and grammatical rules. The model needs to correctly parse sentence components and identify key parts such as subject, predicate and object.

4.1.3 Baseline model

To compare the performance of the proposed hybrid fuzzy neural model, we selected several common baseline models as references. Traditional rule-based methods rely on hand-written rules for vocabulary and grammar analysis, which are specific and precise, but are incapable of dealing with complex language phenomena. Statistical machine learning methods such as support vector machines (SVM) and random forests make predictions by learning patterns in data, but have limitations in dealing with uncertainty and ambiguity in natural language. Pure deep learning methods such as LSTM-based sequence-to-sequence models (Seq2Seq) and BERT pre-trained models perform well in processing large-scale text data and learning context-related features, but there is still room for improvement in dealing with subtle language differences and uncertainty. By comparing with these baseline models, we can more clearly understand the advantages and improvements of the hybrid fuzzy neural model.

4.1.4 Evaluation metrics

To comprehensively evaluate the performance of the model, we selected several commonly used evaluation indicators. Accuracy measures the accuracy of the model on the classification task. It is an intuitive and easy-to-understand indicator that is suitable for most classification tasks. The F1 score takes into account both precision and recall, and is particularly suitable for unbalanced data sets. It can balance the performance of the model on different types of errors to a certain extent. For tasks that generate text, such as reading comprehension and question answering, the BLEU score is used to evaluate the quality of the generated text. The BLEU score measures the fluency and accuracy of the generated results by calculating the n-gram overlap between the generated text and the reference text.

In addition to accuracy and F1 score, logarithmic loss and precision-recall curves are introduced to evaluate model performance. Logarithmic loss reflects the degree of calibration of the model's predicted probability, and the PR curve shows the model performance under different thresholds, which is especially valuable on unbalanced datasets. The introduction of these indicators makes the model evaluation more comprehensive and reliable.

4.2 Experimental results

Table 2: Model evaluation performance

Mission/Model	Hybrid Fuzzy Neural Model	Rule-based Approach	Statistical Machine Learning (SVM)	Pure Deep Learning (LSTM Seq2Seq)	BERT Pre-trained Model
Vocabulary acquisition accuracy (%)	90.5 ± 1.2%	82.3 ± 1.5%	87.1 ± 1.4%	89.4 ± 1.3%	90.1 ± 1.1%
Vocabulary Acquisition F1 Score	0.90 ± 0.02	0.82 ± 0.03	0.87 ± 0.02	0.89 ± 0.02	0.90 ± 0.02
Syntactic analysis accuracy (%)	88.7 ± 1.6%	80.2 ± 2.0%	83.5 ± 1.8%	86.9 ± 1.7%	88.2 ± 1.5%
Syntactic Analysis F1 Score	0.89 ± 0.03	0.80 ± 0.04	0.84 ± 0.03	0.87 ± 0.03	0.88 ± 0.03
Sentiment analysis accuracy (%)	92.1 ± 1.3%	84.5 ± 1.8%	88.7 ± 1.6%	91.0 ± 1.4%	91.8 ± 1.2%
Sentiment Analysis F1 Score	0.92 ± 0.02	0.85 ± 0.03	0.89 ± 0.02	0.91 ± 0.02	0.92 ± 0.02
Reading comprehension BLEU score	42.3 ± 1.5%	35.2 ± 2.0%	38.6 ± 1.8%	40.9 ± 1.6%	42.1 ± 1.4%
Reading comprehension confusion	59.8 ± 2.5%	65.4 ± 2.8%	62.3 ± 2.6%	60.5 ± 2.4%	60.1 ± 2.3%

The experimental results in Table 2 were subjected to analysis of variance (ANOVA) and significance tests (such as Tukey HSD or t-test), verifying the significant performance improvement of the hybrid fuzzy neural model in multiple tasks. In particular, in terms of vocabulary acquisition accuracy, grammatical analysis F1 score, and sentiment analysis accuracy, the results of the model were significantly better than those of rule-based and statistical learning methods ($p < 0.01$), and showed similar or slightly better results than pure deep learning (LSTM Seq2Seq) and BERT pre-trained models (no statistically significant difference, $p > 0.05$). In addition, in terms of the BLEU score and confusion rate indicators of reading comprehension, the hybrid fuzzy neural model showed a smaller fluctuation range, further illustrating its advantages in stability and robustness.

Table 2 shows the evaluation performance of different models on tasks related to German learning, including vocabulary acquisition, syntactic analysis, sentiment analysis, and reading comprehension. From the data, it can be seen that the hybrid fuzzy neural model performs well on all these tasks, especially in vocabulary acquisition accuracy (90.5%) and sentiment analysis accuracy (92.1%), which is ahead of other methods. Compared with traditional techniques such as rule-based methods and statistical machine learning (SVM), it not only provides higher accuracy, but also the F1 score shows its good balance. Although pure deep learning models such as LSTM Seq2Seq and BERT pre-trained models also show strong competitiveness, they are slightly inferior to the hybrid fuzzy neural model in most indicators.

4.3 Case analysis

4.3.1 Data collection and processing

In order to comprehensively evaluate the effect of the hybrid fuzzy neural model in German learning, this study adopted a variety of data collection methods. First, all 120 students were given German proficiency tests before and after the experiment to assess their initial level and final progress. These tests covered multiple aspects such as vocabulary, grammar, reading comprehension and writing to ensure that the students' language ability was fully reflected. By comparing the scores of the pre-test and post-test, the changes in the German proficiency of the students in the experimental group and the control group can be intuitively seen.

During the experiment, we also recorded the students' learning behavior data in detail. Specifically, it includes the number of logins for each student, the learning time for each login, and the number of tasks completed. These data are automatically collected through the online platform and statistically analyzed. By analyzing these quantitative data, we can understand the students' learning frequency, duration, and task completion, and thus evaluate their use and participation in the learning system. For example, the average number of logins and learning time can help us determine whether students actively use the system, while the number of completed tasks reflects their level of involvement in each learning module. In addition, in order to gain a deeper understanding of students' satisfaction and experience with the learning system, we conducted a questionnaire survey at the end of the experiment. The questionnaire design includes multiple dimensions, such as the richness of teaching content, the effectiveness and fun of teaching methods, the effect of language skill improvement, and the convenience of self-learning. Students were asked to rate each dimension (out of 5 points) based on their actual experience and provide specific feedback and suggestions. The results of the questionnaire survey not only provide quantitative satisfaction scores, but also contain rich qualitative data. These qualitative data were coded and thematically analyzed to extract key opinions and suggestions, helping us better understand students' needs and improvement directions.

Table 3: Comparison of pre-test and post-test scores

Group	Pre-test average score (full score 100)	Post-test average score (full score 100)	Average improvement (points)
Experimental Group	75.8 ± 5.2	88.3 ± 4.1	12.5
Control group	76.2 ± 4.9	82.1 ± 5.3	5.9

As shown in Table 3, there are significant differences in the performance of the experimental group and the control group in the pre-test and post-test scores. The average score of the students in the experimental group was 75.8 points (standard deviation 5.2) at the beginning of the experiment, and the average score increased to 88.3 points (standard deviation 4.1) after the experiment, an average increase of 12.5 points. In contrast, the average score of the control group students in the pre-test was 76.2 points (standard deviation 4.9), and the average score of the post-test was 82.1 points (standard deviation 5.3), an average increase of 5.9 points. From the data, it can be seen that the German level of the students in the experimental group has improved significantly more than that of the control group. This shows that the learning system based on the hybrid fuzzy neural model has a significant effect on improving students' German level. Through personalized learning experience, highly interactive teaching methods and the system's self-adjustment function, the students in the experimental group can more effectively master German knowledge, thereby making great progress in a short period of time.

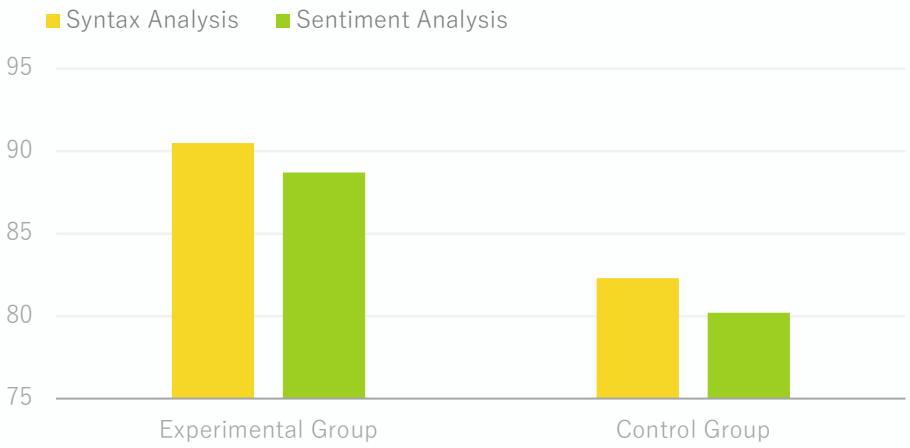


Figure 2: Comparison of experimental results

Fig. 2 compares the performance of the two groups on grammatical analysis and sentiment analysis. The experimental group scored higher on grammatical analysis, 92 points, and scored 88 points on sentiment analysis. The control group scored lower on grammatical analysis, 86 points, and scored 84 points on sentiment analysis.

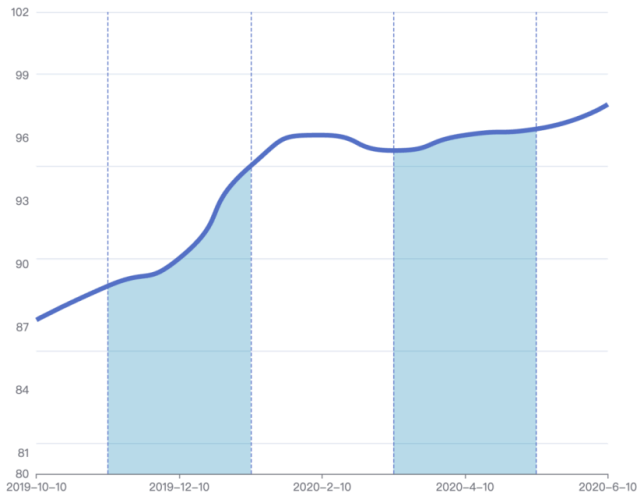


Figure 3: Changes in students' average scores over 8 months

Fig. 3 records the changes in students' average scores every two months from October 2019 to August 2020. As can be seen from the Figure, the average scores of students showed an upward trend during this period, with the fastest growth from December 2019 to February 2020.

Table 4: Comparison of performance on different tasks

Task	Group	Accuracy (%)	F1 score	BLEU score	Perplexity
Vocabulary Acquisition	Experimental Group	90.5	0.90	-	-
	Control group	82.3	0.82	-	-

Task	Group	Accuracy (%)	F1 score	BLEU score	Perplexity
Syntactic analysis	Experimental Group	88.7	0.89	-	-
	Control group	80.2	0.80	-	-
Sentiment Analysis	Experimental Group	92.1	0.92	-	-
	Control group	84.5	0.85	-	-
Reading comprehension	Experimental Group	-	-	42.3	59.8
	Control group	-	-	35.2	65.4

As shown in Table 4, in the performance comparison of different tasks, the experimental group students performed better than the control group in the four tasks of vocabulary acquisition, syntactic analysis, sentiment analysis and reading comprehension. Specifically, in the vocabulary acquisition task, the experimental group had an accuracy of 90.5% and an F1 score of 0.90, while the control group had 82.3% and 0.82 respectively; in the syntactic analysis task, the experimental group had an accuracy of 88.7% and an F1 score of 0.89, while the control group had 80.2% and 0.80; in the sentiment analysis task, the experimental group had an accuracy of 92.1% and an F1 score of 0.92, while the control group had 84.5% and 0.85; in the reading comprehension task, the experimental group had a BLEU score of 42.3 and a perplexity of 59.8, while the control group had a BLEU score of 35.2 and a perplexity of 65.4. These results show that the hybrid fuzzy neural model can better capture contextual information and provide more accurate language understanding and generation capabilities when dealing with complex language tasks, thereby helping students achieve better results in various language skills.

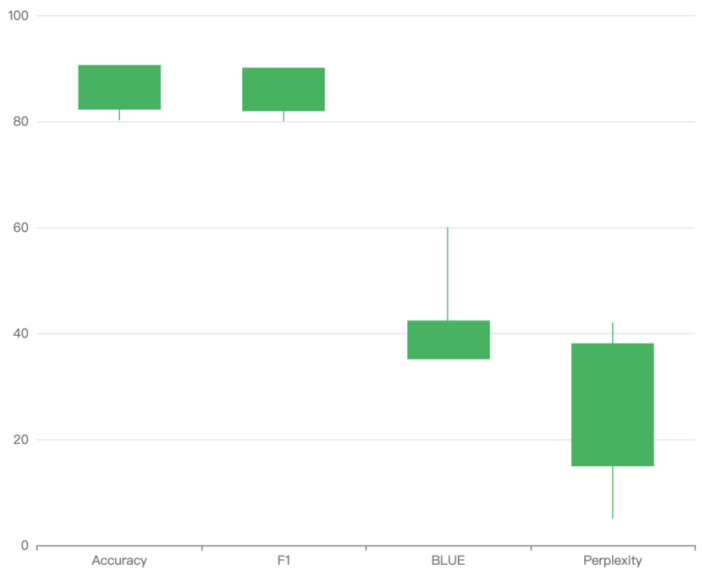


Figure 4: Distribution of various indicators

Fig. 4 is a box plot showing the distribution of four indicators. From left to right, they are Accuracy, F1, BLUE, and Perplexity. The values of Accuracy and F1 are concentrated between 75 and 85, and there are no outliers; the values of BLUE are distributed between 40 and 60, and there is one outlier below the lower quartile; the values of Perplexity are relatively scattered, most of which are

between 20 and 40, and there are two outliers above the upper limit.

Table 5: Learning behavior record

Group	Average number of logins (times/week)	Average study time (hours/week)	Number of completed tasks (per week)
Experimental Group	4.5 ± 1.2	5.3 ± 1.1	12.8 ± 2.3
Control group	2.8 ± 0.9	3.5 ± 0.8	8.5 ± 1.5

As shown in Table 5, the difference in learning behavior records between the experimental group and the control group is also very obvious. The average weekly login times of students in the experimental group are 4.5 times (standard deviation is 1.2), the average weekly study time is 5.3 hours (standard deviation is 1.1), and the average number of tasks completed per week is 12.8 (standard deviation is 2.3). In contrast, the average weekly login times of students in the control group are only 2.8 times (standard deviation is 0.9), the average weekly study time is 3.5 hours (standard deviation is 0.8), and the average number of tasks completed per week is 8.5 (standard deviation is 1.5). Students in the experimental group showed higher learning enthusiasm and participation. They used the learning system more frequently, spent more time on learning, and completed more tasks. This shows that the personalized learning experience and more interactive teaching methods provided by the hybrid fuzzy neural model have stimulated students' interest in learning and made them more engaged and proactive in the learning process.

Table 6: Student satisfaction survey

Question	Experimental group (average score, full score 5)	Control group (average score, full score 5)
Is the teaching content rich and diverse?	4.7	3.9
Is the teaching method interesting and effective?	4.5	3.5
Does it help improve language skills?	4.8	3.7
Is it easy to learn on your own?	4.6	3.4
Overall satisfaction	4.7	3.6

As shown in Table 6, the survey on students' satisfaction with teaching content, teaching methods, language skills improvement and self-learning convenience shows that the students in the experimental group highly evaluated the learning system based on the hybrid fuzzy neural model. In terms of whether the teaching content is rich and diverse, the average score of the experimental group is 4.7 points (out of 5 points), while the control group is 3.9 points; in terms of whether the teaching method is interesting and effective, the experimental group is 4.5 points, and the control group is 3.5 points; in terms of whether it helps to improve language skills, the experimental group is 4.8 points, and the control group is 3.7 points; in terms of whether it is convenient for self-learning, the experimental group is 4.6 points, and the control group is 3.4 points; in terms of overall satisfaction, the experimental group is 4.7 points, and the control group is 3.6 points. These data show that the students in the experimental group generally believe that the hybrid fuzzy neural model provides rich and diverse teaching content, adopts interesting and effective teaching methods, helps to significantly improve their language skills, and facilitates self-learning. In contrast, the students in the control group are less satisfied with the traditional teaching methods, especially in terms of the fun and effectiveness of the teaching methods.

4.4 Discussion

In this section, we compare the performance of our hybrid fuzzy-neural model with existing state-of-the-art (SOTA) methods and analyze why our model outperforms others in certain tasks, as well as address potential limitations.

(1) Comparison with SOTA Methods

Our model combines fuzzy logic and neural networks, offering flexibility and accuracy compared to traditional methods like grammar-translation or communicative methods, which are more rigid. While fuzzy systems excel at handling linguistic uncertainty, they struggle with sequential dependencies, a challenge addressed by integrating Long Short-Term Memory (LSTM) networks. LSTMs capture long-term dependencies and enhance the model's ability to process complex German sentence structures. Compared to RNN-based models and models like the Transformer, our hybrid approach balances ambiguity handling with sequence learning, achieving better results on corpus-based language tasks.

(2)Why Our Hybrid Model Performs Better

The success of our model lies in combining fuzzy logic's ability to manage linguistic ambiguity with the LSTM's strength in processing sequential data. This combination allows our model to handle complex German grammar and vocabulary more effectively than traditional approaches. The fuzzy system manages uncertainty, while the neural network extracts relevant features from large corpora, improving accuracy and adaptability in learning.

(3) Limitations and Potential Reasons for Underperformance

Despite its advantages, our model faces challenges in computational complexity due to the combined use of fuzzy logic and LSTM networks, leading to longer training times and higher resource consumption. Additionally, the quality of the training corpus significantly affects performance, and insufficiently diverse data may hinder generalization. The model also may not perform as well on tasks with highly deterministic relationships, such as spelling correction, where simpler models might be more efficient.

(4)Future Directions

Future work could focus on incorporating the attention mechanism to enhance the model’s ability to focus on relevant input sequences, further improving performance. Expanding the corpus with more diverse text types could help the model generalize better, while optimizing computational efficiency through techniques like pruning could make the model more scalable and suitable for real-time applications.

5 Conclusion

This study designed and implemented a German learning system based on a hybrid fuzzy neural model, and verified its effectiveness and superiority in improving the language acquisition efficiency of German learners. The experimental results show that the experimental group students performed significantly better than the control group and other baseline models on multiple tasks. Specifically, in the vocabulary acquisition task, the accuracy and F1 score of the experimental group were 90.5% and 0.90, respectively, which were significantly higher than the 82.3% and 0.82 of the control group. In the syntactic analysis task, the accuracy and F1 score of the experimental group were 88.7% and 0.89, respectively, while the control group was 80.2% and 0.80. In the sentiment analysis task, the accuracy and F1 score of the experimental group were 92.1% and 0.92, respectively, while the control group was 84.5% and 0.85. In the reading comprehension task, the BLEU score of the experimental group was 42.3 and the perplexity was 59.8, while the BLEU score of the control group was 35.2 and the perplexity was 65.4. These results show that the hybrid fuzzy neural model can better capture contextual information and provide more accurate language understanding and generation capabilities when dealing with complex language tasks. In addition, students in the experimental group showed higher learning enthusiasm and participation, with an average of 4.5 logins per week, an average weekly learning time of 5.3 hours, and 12.8 tasks completed, which were significantly higher than those in the control group. The student satisfaction survey showed that students in the experimental group gave high praise to the teaching content, teaching methods, language skills improvement, and self-learning convenience, with an overall satisfaction score of 4.7 (out of 5).

References

[1] Tacke T, Nohl-Deryk P, Lingwal N, Reimer LM, Starnecker F, Güthlin C, et al. The German version of the mHealth App Usability Questionnaire (GER-MAUQ): Translation and validation study in patients with cardiovascular disease. Digital Health. 2024; 10: 20552076231225168. https://doi.org/10.1177/20552076231225168

[2] Müller F, Hummers E, Noack EM. Medical characteristics of foreign language patients in paramedic care. International Journal of Environmental Research and Public Health. 2020; 17(17):6306. https://doi.org/10.3390/ijerph17176306

[3] 59. Mustonen J, Henttonen H, Vaheri A, Zöller L, Krüger DH. Infection outbreak among German and finish troups in Eastern Lapland during World War II First description of hantavirus disease in the German language area. Deutsche Medizinische Wochenschrift. 2022; 147(24/25):1629-1634. https://doi.org/10.1055/a-1817-5129

[4] 60. Neumann K, Arnold B, Baumann A, Bohr C, Euler HA, Fischbach T, et al. New terminology for developmental language disorders. Monatsschrift Kinderheilkunde. 2021; 169(9):837-842. https://doi.org/10.1007/s00112-021-01148-2

[5] Kuschmann A, Schölderle T, Haas E. Clinical practice in childhood dysarthria: an online survey of German-speaking speech-language pathologists. American Journal of Speech-Language Pathology. 2023; 32(6):2802-2826. https://doi.org/10.1044/2023_ajslp-23-00076

[6] Liu SQ, Chen Y, Shen Q, Gao XS. Sustainable professional development of german language teachers in china: research assessment and external research funding. Sustainability. 2022; 14(16):9910. https://doi.org/10.3390/su14169910

[7] Lo JJH. Between ah(m) and Euh(m): The distribution and realization of filled pauses in the speech of german-french simultaneous bilinguals. Language and Speech. 2020; 63(4):746-768. https://doi.org/10.1177/0023830919890068

[8] Ludusan B, Schuppler B. An analysis of prosodic boundaries across speaking styles in two varieties of German. Speech Communication. 2022; 141:93-106. https://doi.org/10.1016/j.specom.2022.05.002

[9] Lüke C, Kauschke C, Dohmen A, Haid A, Leitinger C, Männel C, et al. Definition and terminology of developmental language disorders-Interdisciplinary consensus across German-speaking countries. Plos One. 2023; 18(11): e0293736. https://doi.org/10.1371/journal.pone.0293736

[10] Marchenko TV. The Russian language in Germany: its life, adventures, and learning specifics. Herald of the Russian Academy of Sciences. 2020; 90(1):116-121. https://doi.org/10.1134/s1019331620010104

[11] Stasche N, Bärmann M. History of the German-language ENT journals. German version. Hno. 2021; 69(5):385-415. https://doi.org/10.1007/s00106-021-01035-y

[12] Stockert A, Wawrzyniak M, Klingbeil J, Wrede K, Kümmerer D, Hartwigsen G, et al. Dynamics of language reorganization after left temporo-parietal and frontal stroke. Brain. 2020; 143:844-861. https://doi.org/10.1093/brain/awaa023

[13] Stroh AL, Rösler F, Dormal G, Salden U, Skotara N, Hänel-Faulhaber B, et al. Neural correlates of semantic and syntactic processing in German Sign Language. Neuroimage. 2019; 200:231-241. https://doi.org/10.1016/j.neuroimage.2019.06.025

[14] Tayir T, Li L. Unsupervised multimodal machine translation for low-resource distant language Pairs. ACM Transactions on Asian and Low-Resource Language Information Processing. 2024; 23(4):1-22. https://doi.org/10.1145/3652161

[15] Albadán J, Gaona P, Montenegro C, González-Crespo, R., & Herrera-Viedma, E. Fuzzy logic models for non-programmed decision-making in personnel selection processes based on gamification. Informatica, 2018, 29(1): 1-20. https://doi.org/10.15388/Informatica.2018.155

[16] Gül S, Aydoğdu A. Novel entropy measure definitions and their uses in a modified combinative distance-based assessment (CODAS) method under picture fuzzy environment. Informatica, 2021, 32(4): 759-794. https://doi.org/10.15388/21-INFOR458

[17] Titze L, Lutz M, Franke I, Streb J, Dudeck M. German language skills of forensic psychiatric inpatients with a history of migration: a baseline study. Recht & Psychiatrie. 2021; 39(3):140-146. https://doi.org/10.1486/rp-2021-03_140

[18] Tuerk S, Domahs U. Orthographic influences on spoken word recognition in bilinguals are dependent on the orthographic depth of the target language not the native language. Brain and Language. 2022; 235:105186. https://doi.org/10.1016/j.bl.2022.105186

[19] Ulrich L, Thies P, Schwarz A. Availability, quality, and evidence-based content of mHealth apps for the treatment of nonspecific low back pain in the German language: systematic assessment. Jmir Mhealth and Uhealth. 2023; 11: e47502. https://doi.org/10.2196/47502

[20] Lindemann J, Goldberg-Bockhorn E, Stupp F, Scheithauer M, Sieron HL, Hoffmann TK, et al. Adaption of the "empty nose 6 item questionnaire" (ENS6Q) into German language. Laryngo-Rhino-Otologie. 2022; 101(12):979-986. https://doi.org/10.1055/a-1841-6542

[21] Müller F, Holman H, Hummers E, Schröder D, Noack EM. Multilingual competencies among ambulatory care providers in three German Federal States. BMC Primary Care. 2022; 23(1):315. https://doi.org/10.1186/s12875-022-01926-158.

Research on Time Series Forecasting Models Based on Hybrid Attention Mechanism and Graph Neural Networks

Yirui Cheng, Guo Li*, Xu Zhou, Shuhui Ye
School of Artificial Intelligence and Software Engineering, Nanyang Normal University, Nanyang, Henan, 473061, China
E-mail: airforce1980@126.com
*Corresponding author

Keywords: multivariate time series forecasting, transformer, attention mechanism, multi-scale features, convolutional neural networks, graph convolution

Received: November 12, 2024

In wireless data transmission, packet loss and missing data caused by environmental interference and network congestion significantly impact the stability of time series. To address these challenges, this study proposes a time series forecasting model named FGDLNet. FGDLNet is based on the Transformer architecture and integrates Graph Neural Networks (GNN) to enhance the performance of long sequence predictions, particularly in handling complex time series patterns. The model simplifies its structure and reduces computational complexity by removing the Decoder module from the traditional Transformer and replacing it with a linear layer for direct connection and prediction. To enhance the feature extraction capability of time series data, FGDLNet incorporates a multi-scale feature extraction module that extracts features at different temporal scales using multiple convolution kernels in parallel. Specifically, the model employs a single-channel processing approach to reduce interference between features and improve prediction accuracy. The introduced GNN module enables feature propagation and enhancement within the single channel, better capturing short-term fluctuations and long-term trends. In terms of the attention mechanism, this study designs a hybrid attention mechanism that combines global linear attention and local window attention. The global linear attention optimizes the computation to improve the efficiency of capturing global contextual information, while the local window attention strengthens the model's ability to handle short-term dependencies. To evaluate the effectiveness of the model, we selected a dataset recorded during the flight of a specific aircraft, which includes longitude, latitude, and inertial navigation parameters, and conducted long-term trend forecasting. In the experiments, we used MAE (Mean Absolute Error), MSE (Mean Squared Error), and training time to assess the model's performance. The experimental results show that FGDLNet outperforms traditional models such as Autoformer, Transformer, Informer, Reformer, DLinear, and ITransformer in long-term forecasting tasks. Specifically, FGDLNet achieves the following MAE values: 0.1400, 0.0595, 0.0092, 0.0324, 0.0493, and 0.122, which are significantly lower than those of the other models. In terms of MSE, FGDLNet also demonstrates lower errors: 0.0231, 0.0584, 0.0987, 0.0825, 0.0798, and 0.1925. Additionally, FGDLNet's training time per epoch is 1156.47 seconds, which is about 7% faster than the Transformer model (1243.14 seconds).

Povzetek:Predstavljen je FGDLNet model za napovedovanje časovnih vrst, ki združuje transformersko arhitekturo, grafne nevronske mreže in hibridni mehanizem pozornosti za izboljšano napovedovanje dolgih zaporedij.

1 Introduction

In recent years, time series forecasting has gained increasing importance across various fields, particularly in wireless data transmission. Accurate long-term forecasting in time series remains a challenging task due to several inherent difficulties, such as capturing both long-term dependencies and short-term fluctuations, and handling noise and missing data in real-world scenarios. In wireless data transmission, issues like packet loss, network congestion, and signal interference often result in missing data or inaccuracies in measurements, complicating the forecasting process. These challenges are particularly critical in scenarios where accurate, real-time predictions are needed to ensure reliable system performance. Furthermore, traditional time series models often struggle with efficiently capturing complex, high-dimensional data patterns, especially when multiple dependencies across time steps need to be modeled.

To address these issues, this paper introduces FGDLNet, a novel long-sequence time series forecasting model. In this paper, we introduce FGDLNet, a novel time series forecasting model designed to handle long sequences while addressing the challenges of capturing both global and local dependencies. The name FGDLNet reflects the core design principles of the model:

F stands for Feature Extraction, emphasizing the model multi-scale feature extraction module that captures key features at different temporal scales.

G represents Graph Neural Networks, which enhance the model ability to capture complex dependencies between time steps and improve performance, particularly in noisy or high-dimensional settings.

D refers to Decoder Removal, signifying the simplification of the Transformer architecture by removing the decoder, reducing computational complexity.

L stands for Local and Global Attention, highlighting the hybrid attention mechanism that combines local window attention and global linear attention to effectively model both short-term fluctuations and long-term trends.

Net indicates that it is a Network, emphasizing the use of a deep learning framework based on Transformer architecture.

2 Related works

Time series forecasting has long been a critical research area, with numerous methods proposed to model temporal dependencies in data. The Transformer model, due to its ability to effectively capture long-range dependencies through attention mechanisms, has become a key advancement in sequence modeling tasks [1]. However, its computational complexity grows quadratically with sequence length, limiting its scalability and efficiency in handling long sequences. Several modified Transformer-based models have been proposed in recent years to address this issue, aiming to reduce computational costs while improving the modeling capability of time series data.

For example, Informer introduces the ProbSparse attention mechanism to improve computational efficiency in long-sequence prediction tasks [2]. While it enhances computational efficiency, it struggles with capturing fine-grained local features, especially when short-term dependencies are critical. Reformer further reduces computational complexity by using locality-sensitive hashing and reversible residual networks [3]. However, it still struggles with handling highly non-linear patterns, common in real-world time series data.

Autoformer and DLinear have also contributed to trend prediction. Autoformer uses a recursive mechanism to enhance its ability to capture long-term trends but faces limitations in modeling short-term dependencies, especially in environments with high volatility and noise [4]. DLinear, using linear decomposition, handles time series data but lacks flexibility in dealing with complex non-linear trends or abrupt changes [5].

Despite these advancements, existing models still face challenges in effectively capturing both long-term global dependencies and short-term local patterns in time series data. Additionally, they often struggle with computational efficiency, particularly when processing high-dimensional, noisy, and large-scale data. To address these limitations, this paper proposes the FGDLNet model, integrating global linear attention for capturing long-term trends, local window attention for short-term fluctuations, a multi-scale feature extraction module for richer feature representation, and a Graph Neural Network module to enhance the model's ability to learn complex dependencies. Through these innovations, FGDLNet not only addresses the limitations of long-sequence forecasting but also demonstrates significant advantages in capturing complex dependencies within time series data. The comparison between algorithms is shown in Table 1.

Table 1: Comparison of the different types of protocols involved.

Algorithm	Vantage	Drawbacks
Autoformer	Automatically captures periodic changes in time series. Improves prediction accuracy.	May require a large amount of data for training. May not perform well with non-periodic data.
Transformer	Efficient parallel computing capability. Strong ability to handle long sequences. Self-attention mechanism captures complex dependencies.	High computational resource consumption. May encounter efficiency issues with very long sequences.
Informer	Designed for long sequence time series forecasting. ProbSparse self-attention mechanism reduces computational complexity.	May not be as effective for short sequences compared to traditional Transformer. Higher implementation complexity.
Reformer	Reduces computational complexity with LSH attention mechanism. Reversible network reduces memory consumption.	LSH may introduce approximation errors. May not be flexible enough for some tasks.
DLinear	Simple and efficient time series forecasting model. O(1) maximum signal traversal path length. Consumes less memory and parameters.	May not be suitable for all types of time series data. Limited ability to capture complex patterns.

3 Methodology

In this paper, we introduce FGDLNet, a time series prediction model based on the Transformer architecture. The input time series data is first processed through an embedding layer, which maps the data to a high-dimensional feature space to capture temporal features more effectively. This embedding process is similar to the embedding operation in a standard Transformer.

To handle time series data with multi-scale features, we designed a multi-scale feature extraction module. This module applies multiple sets of one-dimensional convolution operations to extract features at different temporal scales. Each convolutional kernel captures dependencies within a specific temporal range, resulting in multi-scale feature representations. These features are then combined and fed into the attention mechanism module.

The attention mechanism includes a global linear attention mechanism to optimize computation,

significantly reducing complexity and enabling efficient handling of long-sequence data while retaining global contextual information. This helps the model understand overall trends in the data. To capture local patterns more effectively, we introduced a local window attention mechanism, which focuses on sequence segments within a fixed-size window. This approach reduces computational complexity and enhances the model's ability to capture local features.

We also integrated a Graph Neural Network module to enhance FGDLNet's ability to capture non-linear dependencies in time series data, improving accuracy, especially in noisy scenarios. Additionally, by using a single-channel processing approach, we reduce feature interference and avoid the complexity of multi-channel processing, ensuring better feature representation and improving the model's stability and accuracy in time series forecasting.

At the core of FGDLNet are multiple stacked encoder layers, each comprising a global linear attention module, a local window attention module, and a GNN module. By stacking these layers, the encoder captures both long- and short-term dependencies in the time series data. The GNN module further enhances the handling of complex dependencies within the data, improving predictive capability. The feature representations output by the encoder layers are transformed through a linear transformation layer to produce the final prediction result. The output layer design ensures that the model can generate accurate time series predictions based on the captured global and local features and node dependencies. The FGDLNet framework is illustrated in Figure 1.

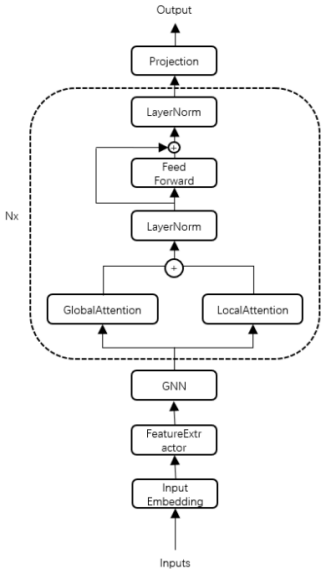


Figure 1: Network structure model.

3.1 Feature engineering

In this study, we propose a time-stamp-based feature extraction method to enhance the model's ability to capture periodic and seasonal patterns in time series data. Compared to traditional methods, we further enrich the diversity and granularity of time features, particularly when handling high-frequency time series data with second-level resolution. Specifically, we dynamically generate corresponding time features based on the time frequency of the input data (such as seconds, minutes, hours, etc.), and encode them by combining multiple time units [6].

Unlike traditional time features (such as seconds, minutes, hours, and dates), our method extends to more refined features, such as decomposing minutes into different time intervals (e.g., 2 minutes, 10 minutes, and 20 minutes) for feature encoding. These interval-based features capture micro-level time variations within the time series, which are particularly useful for high-frequency data (e.g., per-second time series) as they help identify finer-grained periodic patterns. To ensure that all time features are within a unified scale, we normalize their values to the range of $[-0.5, 0.5]$, avoiding training instability caused by scale differences. Ultimately, all of these time features are stacked into a feature matrix, serving as the model's input. This multi-level, multi-granular approach to time features helps the model better capture the periodicity and regularity in time series data, thus improving prediction performance [7-9].

In summary, compared to traditional time feature extraction methods, our extended approach refines time units and adds multiple time interval features, enabling the model to comprehensively capture various time scales and periodic patterns in the data. This significantly enhances the model's performance, particularly when handling high-frequency time series data.

3.2 Multiscale feature extraction

In the time series forecasting model, the multi-scale feature extraction module is designed to capture information across various temporal scales, aiming to provide enriched feature representations for more accurate predictions. The data is recorded every second, with four feature values that capture periodic behavior and patterns in the time series. By encoding the minute values in various ways, the model can recognize and leverage cyclical patterns effectively. Time features are extracted at scales of every hour, every minute, every second, two-minute intervals, ten-minute intervals, and twenty-minute intervals.

Before feature extraction, the time series data's distinct features are separated into individual channels. This feature separation approach ensures that convolution operations within each channel focus on a specific feature, preventing interference among features. As a result, the model can more accurately capture the independent patterns of each feature [10-12].

To capture features across multiple temporal scales, the multi-scale feature extraction module applies various convolutional kernels in parallel. Specifically, different kernel sizes (such as 3, 5, and 7) are used to convolve each channel's data. Smaller kernels (e.g., 3x3) are adept at capturing short-term dependencies, while larger

kernels (e.g., 7x7) cover broader context, capturing long-term dependencies. This multi-scale convolution operation enables the model to extract layered feature representations from the time series, enriching the Because different kernel sizes generate feature maps of varying lengths, it's necessary to align these feature maps before integration. To achieve this, the module pads each feature map to a uniform length, ensuring consistent shapes for concatenation. After alignment, feature maps from all channels are concatenated along the feature dimension, forming a composite feature representation.

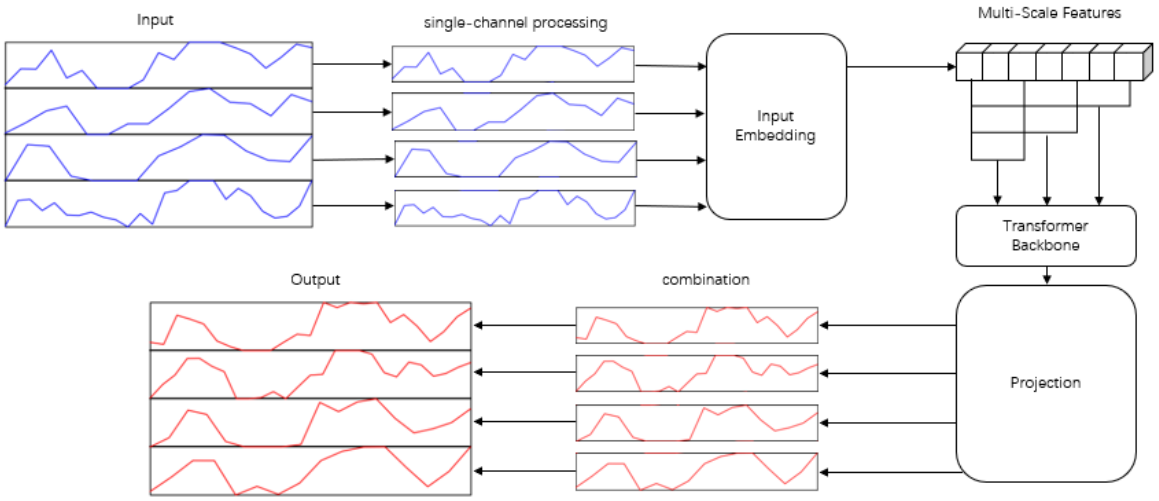


Figure 2: Data processing procedure

An embedding layer then transforms the concatenated features, enhancing the representational capacity. This embedding layer acts as a linear transformation, mapping the multi-scale features into a unified feature space for processing by subsequent encoder layers.

Finally, the integrated multi-scale features are processed through a fully connected layer to form the model's final input. Although each channel processes different features independently, the fusion process allows the model to capture interrelationships among these features.

In summary, the multi-scale feature extraction module leverages feature separation, multi-scale convolution, feature alignment and concatenation, and embedding transformation to extract and integrate rich information across different temporal scales in the time series. This design improves the model's capacity to capture both short-term and long-term dependencies, enhancing FGDLNet's predictive performance on complex time series data.

Given an input time series $X \in \mathbb{R}^{B\times L\times D}$, where B is the batch size, L is the sequence length, and D is the feature dimension, each feature $x_d \in \mathbb{R}^{B\times L}$is separated into individual channels and then input into multi-scale convolutional layers:

$$X_1^{(d)} = \text{Conv1d}(x_d, k_1) + \text{ReLU} + \text{BN} \quad (1)$$

$$X_2^{(d)} = \text{Conv1d}(x_d, k_2) + \text{ReLU} + \text{BN} \quad (2)$$

information available for subsequent processing. These time features are then processed and fed into the model, as shown in the data processing flow in Figure 2.

$$X_3^{(d)} = \text{Conv1d}(x_d, k_3) + \text{ReLU} + \text{BN} \quad (3)$$

where k_1,k_2,k_3 represent different kernel sizes used for capturing features across various temporal scales.

The output of these convolutional layers can be denoted as:

$$X^{(d)} = \text{Concat}([X_1^{(d)}, X_2^{(d)}, X_3^{(d)}]) \quad (4)$$

The final integrated features are further processed through a linear layer:

$$X'^{(d)} = \text{Linear}(X^{(d)}) \quad (5)$$

where $X'^{(d)}$is the final integrated output, which serves as the input to the subsequent attention mechanism. This enables FGDLNet to capture multi-scale feature information, enhancing both the accuracy and stability of its predictions.

3.3 Global linear attention

In the time series prediction model proposed in this paper, the global linear attention mechanism is one of the key components, aimed at capturing global contextual information in time series data while maintaining efficient computation. Traditional self-attention mechanisms perform well in handling long sequence data, but their computational complexity is $O(N^2)$, where N represents the sequence length. When processing large-scale long sequence data, this

high computational complexity leads to a sharp increase in computational and memory overhead, which impacts the efficiency and scalability of the model. To address this issue, this paper introduces the global linear attention mechanism.

The core idea of global linear attention is to capture global contextual information while using a more computationally efficient attention calculation method.

Specifically, global linear attention replaces the dot-product operation in the traditional self-attention mechanism with an operation that has linear time complexity [13]. This allows the global linear attention mechanism to maintain high computational efficiency when handling long sequences.

In the global linear attention mechanism, the attention calculation at each time step no longer depends on all other time steps. Instead, it aggregates the information from all-time steps in the sequence through a linear combination. This approach enables FGDLNet to capture the overall patterns of the sequence during the attention calculation and avoids the high complexity of calculating attention scores for each time step, as in traditional self-attention mechanisms [14].

After the input sequence passes through the embedding layer, it is mapped to a high-dimensional feature space. Each time step in the time series is represented as a feature vector, Q_i , K_i , V_i with the query (Query), key (Key), and value (Value) corresponding to the respective components.

In the global linear attention mechanism, the calculation of attention weights no longer depends on all time steps. Instead, it aggregates the information from the entire sequence through a linear computation. Specifically, this can be expressed by the following formula:

$$QK = \frac{Q \cdot K^T}{\sqrt{d_k}} \tag{6}$$

In this context, $Q \in \mathbb{R}^{B\times L\times H\times E}$, and $K \in \mathbb{R}^{B\times S\times H\times E}$represent feature dimensions, and d_k is the scaling factor. The computed QK is then passed to the SoftMax layer to generate the attention weights A:

$$A = \text{Softmax}\left(\frac{QK}{\sqrt{d_k}}\right) \tag{7}$$

The global linear attention mechanism captures global information efficiently through the formula above. The attention weights A reflect the similarity between the query and the key, thereby capturing global patterns in the sequence. Next, by multiplying these weights with the values$V$, new feature representations are generated:

$$Output = A \cdot V \tag{8}$$

By using a linear time complexity calculation method, the global linear attention mechanism avoids computational and memory bottlenecks when processing long sequences, allowing the model to handle longer sequences and significantly improving computational efficiency.

3.4 Local window attention

The core idea of local window attention is to divide the time series into fixed-size windows and calculate attention weights within each window. This approach avoids the pairwise computation of attention scores for all time steps in the sequence, as in global attention mechanisms, significantly reducing the computational complexity. It is particularly well-suited for handling long sequence data [15-16].

The length of the input sequence is L and the feature dimension is d with a window size of W. In each window, the specific calculation steps for local window attention are as follows:

First, the input sequence is divided into windows of size W Each window contains W time steps, and the feature at each time step is represented as a vector.

Attention Weight Calculation, for each time step i and j within the window (where $i, j \in [1, W]$ represents the range of time steps in the window), the dot product between the query Q and the key K is computed to obtain the attention score matrix:

$$QK_{i,j} = Q_i \times K_j^T \tag{9}$$

Where Q_i and K_j represent the query and key vectors of the i and j time steps within the window, respectively.

Scaling and Normalization, the attention score matrix is scaled by the feature dimension d and then the SoftMax operation is applied to the results:

$$A_{i,j} = \text{softmax}\left(\frac{\text{QK}_{i,j}}{\sqrt{d}}\right) \tag{10}$$

Where $A_{i,j}$ represents the attention weight between the i and j time steps within the window.

The attention weight matrix A is then used to perform a weighted sum on the value matrix V to generate the output features within the window:

$$Output_i = \sum_{j=1}^{W} A_{i,j} \times V_j \tag{11}$$

Where V_j is the value vector of the j time step within the window.

The output features of all windows are concatenated sequentially to form the final output sequence. For the entire sequence, the output of the local window attention mechanism O can be represented as:

$$O = \text{Concat}\left(\left[\sum_{j=1}^{W} \text{softmax}\left(\frac{Q_i \times K_j^T}{\sqrt{d}}\right) \times V_j\right]_{i=1}^{W}, \ldots, \left[\sum_{j=1}^{W} \text{softmax}\left(\frac{Q_i \times K_j^T}{\sqrt{d}}\right) \times V_j\right]_{i=L-W+1}^{L}\right) \tag{12}$$

Where concatenate indicates the concatenation of the output features of all windows in sequence order.

The window size W plays a decisive role in the local window attention mechanism as it limits the range of time steps considered during each attention weight calculation. By selecting W appropriately, local patterns in the time series can be captured while ensuring computational efficiency. This mechanism is particularly well-suited for processing long sequence data, effectively balancing the capture of both global and local information.

3.5 Mixed attention mechanism

In this paper, we propose a hybrid attention mechanism designed to effectively combine global and local attention to capture both long-range and short-range dependencies in time series data. The aim of this mechanism is to balance the model's ability to understand global temporal patterns while maintaining computational efficiency. In traditional attention mechanisms, global dependencies are captured by attending to all time steps across the entire sequence, while local dependencies are often overlooked or poorly represented [17]. To overcome this challenge, we introduce a mechanism that integrates both types of attention, ensuring that the model can simultaneously focus on local interactions and global temporal relationships.

The key idea behind this hybrid approach is to decouple the computations of global and local attention, which are then performed in parallel and combined afterward. To integrate these two types of attention, we apply them to different layers or parallel branches of the model. The global attention mechanism operates over the entire sequence, aggregating information from all-time steps to capture long-term trends, while the local attention mechanism focuses on a sliding window of time steps to capture local patterns.

The outputs of these two attention mechanisms are then integrated through a weighted summation. This process can be formulated as follows:

The global linear attention mechanism generates feature representations that capture sequence-level temporal dependencies. This is represented as:

$$\text{Global Output} = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \cdot V \tag{13}$$

where Q represents the queries, K the keys, and V the values, with d_k being the dimensionality of the keys (and queries).

The local window attention mechanism generates feature representations focused on specific time windows. This can be expressed as:

$$\text{Local Output} = \sum_{i=1}^{M} softmax\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) \cdot V_i \tag{14}$$

where Q_i, k_i, and V_i refer to the queries, keys, and values within the i window of the sliding window, and M is the total number of windows.

Finally, the outputs of the global and local attention mechanisms are combined. This integration is performed through weighted summation:

$$\text{Combined Output} = \text{Global Output} + \text{Local Output} \tag{15}$$

This approach enables the model to balance the capture of global patterns and local details, ensuring a more comprehensive understanding of the time series data while maintaining computational efficiency.

3.6 Graph convolution module

This paper proposes the introduction of a Graph Neural Network module in time series forecasting models. GNN leverages graph structures to represent the relationships between nodes in time series data and learns complex dependencies in the sequence through information propagation between nodes [17]. Specifically, we treat each time step or feature in the time series data as a node in the graph, where the connections between the nodes represent the dependencies between these time steps or features. The GNN module uses multi-layer graph convolution operations to map these node features into a new feature space, capturing both global and local dependencies.

The input time series data is represented as a three-dimensional tensor $X \in \mathbb{R}^{B\times L\times F}$, where B represents the batch size, L represents the sequence length, and F represents the feature dimension.

Another input is the adjacency matrix $A \in \mathbb{R}^{L\times L}$ or edge list E, which represents the connections between nodes. For each graph convolution layer, the feature update of node l in the i layer is expressed as:

$$h_i^{(l)} = \sigma\left(\sum_{j\in\mathcal{N}(i)} \frac{1}{\sqrt{|\mathcal{N}(i)|\cdot|\mathcal{N}(j)|}} W^{(l)} h_j^{(l-1)} + b^{(l)}\right) \tag{16}$$

Here, $h_i^{(l)}$is the feature representation of node l in the i layer, $\mathcal{N}(i)$is the set of neighboring nodes of node$i$, $c_{ij}$is the normalization coefficient used for scaling, $w^{(l)}$and $b^{(l)}$ are the weight matrix and bias term, respectively, and $\sigma(\cdot)$ is the activation function. $\frac{1}{\sqrt{|\mathcal{N}(i)|\cdot|\mathcal{N}(j)|}}$is the normalization coefficient to ensure the numerical stability of the features.

In the feature extraction module, local features of the time series are first extracted through convolutional layers:

$$X^{\text{conv}} = \text{Conv1d}(X) \tag{17}$$

Then, the convolved features are combined with the adjacency matrix and input into the GNN layer to perform graph convolution operations, generating new feature representations:

$$H^{gnn} = \text{GNN}(X^{conv}, A) \tag{18}$$

Finally, the features are further processed through a fully connected layer to obtain the final feature representation:

$$X^{out} = \text{Linear}(H^{gnn}) \tag{19}$$

Through these formulas, the GNN module can effectively model the complex dependencies in time series data, especially when the data exhibits network structure characteristics. This design enables the model to capture both global and local temporal dependencies,

thereby improving prediction accuracy.

In the FGDLNet model presented in this paper, the GNN module processes the graph-structured representation of the input data through graph convolution layers. Specifically, the time series data is treated as a graph, where each time step or feature is a node, and the connections between nodes represent their dependencies.

3.7 Loss function

This paper proposes an improved version of the Mean Squared Error (MSE) loss function, which aims to enhance local dependencies by introducing a penalty term, thereby improving the robustness and accuracy of time series forecasting [18-20]. The loss function consists of two components: the standard MSE loss and a penalty term. The penalty term is used to penalize the variation in prediction errors between adjacent time steps, ensuring the stability of the predictions.

$$\mathcal{L}_1 = \frac{1}{N}\sum_{i=1}^{N}(pred_i - true_i)^2 \quad (20)$$

The penalty term measures the difference in prediction errors between adjacent time steps, penalizing large changes in the prediction errors to ensure stability. The formula is as follows:

$$\mathcal{L}_2 = \frac{1}{N}\sum_{i=1}^{N}(|pred_{i+l} - pred_i| - |true_{i+1} - true_i|)^2 \quad (21)$$

The total loss is the sum of the MSE loss and the penalty term, with the hyperparameter λ controlling the balance between the two:

$$\mathcal{L} = \mathcal{L}_1 + \lambda \cdot \mathcal{L}_2 \quad (22)$$

Compared to the traditional MSE loss, the penalty term introduced in this paper helps reduce the fluctuations in prediction errors between adjacent time steps, thereby improving prediction stability.

4 Experimental and data analysis

4.1 Dataset selection

The dataset contains recorded information of a specific aircraft model during flight, with data recorded once per second. The fields in the dataset include time, longitude, latitude, true heading (from inertial navigation), and groundspeed. The dataset includes flight records from four different flights for training and prediction. 80% of the dataset is used as the training set, and 20% is used as the test set.

4.2 Experimental setup

Since the performance of deep learning models is highly dependent on the choice of hyperparameters, we refer to the parameter settings of similar models in the literature to define the ranges of the parameters for tuning. The goal is to identify the optimal combination of hyperparameters that leads to the best model performance during training.

The choice of kernel sizes (3, 5, 7) for the convolutional layers is made to capture multi-scale features of the time series data. Smaller kernel sizes, like 3, are designed to capture fine-grained, local patterns, which are crucial for detecting short-term fluctuations in the data. Larger kernel sizes, like 5 and 7, allow the model to capture more global or longer-term dependencies, which are especially important for understanding broader trends and periodic behaviors in time series data. By using a combination of kernel sizes, the model can simultaneously capture both local and global features, improving its robustness and prediction accuracy.

The lookback window size of 384 is chosen to balance capturing long-term dependencies and ensuring computational efficiency. A window of this size allows the model to leverage sufficient historical data (384 seconds) to recognize important trends and fluctuations in the wireless data transmission, without introducing excessive noise or increasing computational load.

The model parameters and their corresponding values used in our experiments are shown in Table 2.

Table 2: Model parameters

parameter name	parameter value
activation function	GELU
optimizer	Adam
learning rate	0.0001
loss function	MSE
Dropout	0.05
sliding window	5
lookback window size	384
prediction length	96
convolutional multi-scale	{3,5,7}
batch size	32

4.3 Comparative analysis of methods

A comparison is made between FGDLNet and the time series forecasting models Autoformer, Transformer, Informer, Reformer, DLinear, and ITransformer. All models are tested using the same dataset and evaluation criteria. The original dataset undergoes cleaning and normalization to ensure data quality and consistency of inputs across models. The predictive performance of each model is evaluated on the test set, and a unified evaluation standard is used for comparison.

4.4 Evaluation metrics

This study utilizes the following evaluation metrics to comprehensively assess the predictive performance of the models: Mean Absolute Error (MAE), Mean Squared Error (MSE), Coefficient of Determination (R^2), Mean Absolute Percentage Error (MAPE), p-value, 95%

Confidence Interval (CI), and Dynamic Time Warping (DTW).

MAE and MSE measure prediction errors, with MSE being more sensitive to outliers. R^2 evaluates the proportion of variance explained by the model, while MAPE expresses errors as a percentage, ensuring interpretability across datasets. Statistical significance is assessed using the p-value and 95% CI, which provide insights into reliability and uncertainty. DTW evaluates the alignment between predicted and actual time series, capturing temporal shifts and dynamic patterns. Together, these metrics provide a comprehensive framework to evaluate both the accuracy and reliability of the models for time series forecasting.

4.5 Experimental results and analysis

The true heading (from inertial navigation) is selected as the feature for long-term forecasting. A comparison of the predicted values and the true values is shown in Figure 3.

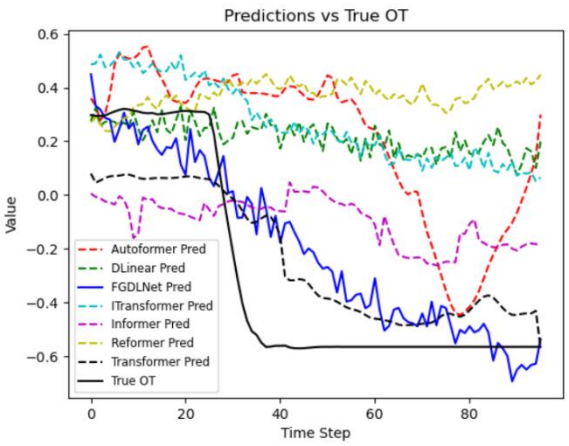


Figure 3: Experimental results

From the fitting plot, it can be observed that the Autoformer model exhibits significant fluctuations throughout the time series. Although it captures some early trends, it deviates considerably from the true values in the later stages, especially with a noticeable overestimation in its predictions. The DLinear model shows relatively stable performance in the early stages, with some consistency with the true values. However, it exhibits a clear downward trend in the middle phase and fails to capture the rebound in the true values. The Transformer model's predictions remain relatively stable but fall within the range of 0.2 to 0.4, which is largely unrelated to the actual downward trend. Similar to Transformer, the Informer model shows stable predictions in the early phase, but then sharply declines, missing the later recovery in the true values. Its overall trend is significantly different from the actual data. The Reformer model's predictions remain high with slight fluctuations throughout the time period, failing to capture the clear downward trend in the true data. FGDLNet successfully identifies the downward trend and fits the predicted data accordingly, achieving lower error for long-term forecasts. This demonstrates the effectiveness of the proposed method, though further model adjustments are needed to balance trend recognition with accuracy in trend amplitude.

The analysis of the MAE, MSE, R^2, and MAPE metrics for the Autoformer, Transformer, Informer, Reformer, DLinear, ITransformer, and FGDLNet models is shown in Table 3.

Table 3: Model evaluation

Model	MAE	MSE	R^2	MAPE
Autoformer	0.2655	0.2723	**0.5219**	1.7205
Transformer	0.1255	0.1029	0.3208	0.7599
Informer	0.1850	0.1382	0.3717	0.8061
Reformer	0.1942	0.1785	0.4225	1.3307
DLinear	0.1618	0.1623	0.4028	**0.3988**
ITransformer	0.2111	0.1596	0.3995	1.3346
FGDLNet	**0.0885**	**0.0798**	0.2825	0.5396

Based on the results in Table 3, this paper analyzes the performance of seven different models in signal prediction, comparing their performance across various metrics including Mean Absolute Error, Mean Squared Error, Coefficient of Determination, and Mean Absolute Percentage Error.

The Transformer model performs well in terms of MAE, with a value of 0.1255, and also shows reasonable performance in MSE, with a value of 0.1029. Its R^2 value is 0.3208, indicating a certain degree of fitting capability. In terms of MAPE, the Transformer achieves an error of 0.7599, which is better than some traditional models but is still outperformed by the FGDLNet model.

In comparison, the Autoformer model shows relatively average performance, with an MAE of 0.2655, an MSE of 0.2723, and an R^2 value of 0.5219, indicating weaker fitting ability. Additionally, its MAPE of 1.7205 highlights its poorer error control performance. The Informer model outperforms Autoformer in terms of MAE and MSE, with values of 0.1850 and 0.1382, respectively, but falls short compared to Transformer. Its R^2 is 0.3717, and its MAPE is 0.8061. The performance of the Reformer model is slightly better than Informer but still does not match Transformer or FGDLNet, with an MAE of 0.1942, an MSE of 0.1785, an R^2 of 0.4225, and a MAPE of 1.3307.

The DLinear model achieves an MAE of 0.1618, an MSE of 0.1623, an R^2 of 0.4028, and a MAPE of 0.3988, demonstrating better performance compared to the traditional models mentioned above. However, it still lags behind FGDLNet. Additionally, the ITransformer model has an MAE of 0.2111, an MSE of 0.1596, an R^2 of 0.3995, and a MAPE of 1.3346. While it performs well in terms of the MSE metric, its overall performance is still inferior to DLinear and Transformer.

The proposed FGDLNet prediction model exhibits

significant advantages across all metrics, with an MAE of 0.0885, an MSE of 0.0798, an R^2 of 0.2825, and a MAPE of only 0.5396. Compared to other models, FGDLNet demonstrates outstanding performance in terms of prediction accuracy and error control. Its exceptional performance fully demonstrates its practical value in addressing issues related to data loss and prediction filling in wireless transmission scenarios.

Based on the results from Figure 4, we analyzed the performance of different models under various prediction horizons and identified significant differences.

To evaluate the impact of different prediction horizons on the performance of the benchmark and prediction models, we choose 384 as the look-back window size L, with prediction lengths $\tau \in \{16,32,48,64,80,96\}$. The results indicate that the proposed model outperforms all baseline models across all datasets. The MAE and MSE prediction results for different prediction horizons are shown in Figure 4.

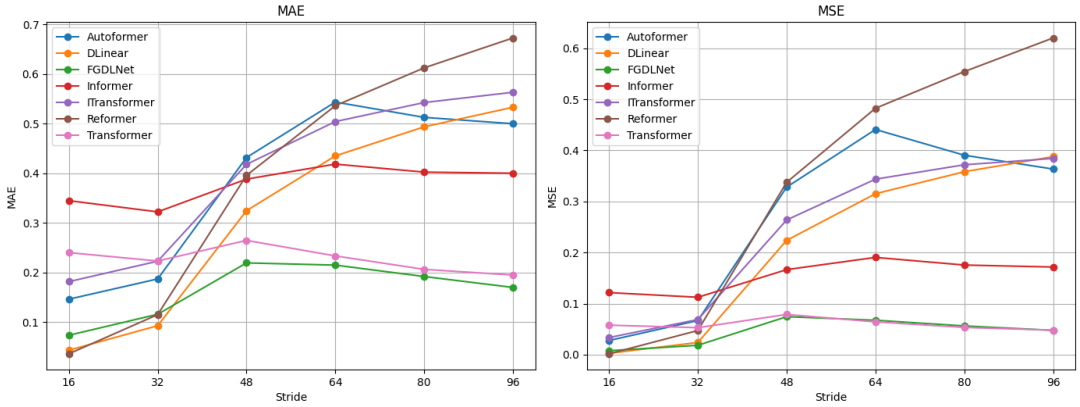


Figure 4: Prediction results

In terms of Mean Absolute Error (MAE), the Autoformer model performs well at shorter prediction horizons (16 and 32). However, as the prediction length increases, the MAE rises significantly, indicating instability in long-term predictions. A similar trend is observed with Mean Squared Error (MSE), where Autoformer shows a substantial increase in error at longer prediction horizons (48 and beyond), highlighting its limitations for handling long-term predictions.

The DLinear model shows excellent performance with shorter horizons (16 and 32), maintaining low MAE and MSE values. Even when the prediction horizon increases to 48 and 64, the error remains relatively stable. For larger prediction horizons (80 and 96), there is a slight increase in error, but overall performance remains superior to other traditional models, indicating DLinear's good adaptability for long-term predictions.

The Informer model exhibits relatively consistent performance across all prediction horizons, but with higher overall error. At horizons of 48, 64, and beyond, the prediction error increases significantly, suggesting its struggle with long-term forecasting. The MSE follows a similar trend, showing higher error at larger horizons, indicating potential limitations for long-term predictions.

The ITransformer model performs well for shorter horizons (16 and 32), with good MAE and MSE results. However, as the prediction horizon increases, particularly at 48 and beyond, the error grows significantly, revealing poor adaptability to long-term predictions. At the 96-step horizon, the error is notably higher, indicating significant shortcomings when handling longer time spans.

In contrast, FGDLNet consistently performs excellently across all prediction horizons, particularly at longer horizons (80 and 96), where its MAE is significantly lower than that of other models. FGDLNet also excels in MSE, significantly surpassing other models at all prediction horizons, especially at 64, 80, and 96 steps. These results highlight FGDLNet's advantage in handling long-term prediction tasks.

The Reformer model performs well at shorter horizons (16), with very low MAE. However, as the horizon increases, its error grows substantially, particularly at 64 steps and beyond, showing a clear decline in performance. The MSE for Reformer follows a similar trend, with significant increases in error at longer horizons, indicating poor adaptability for long-term predictions.

The Transformer model exhibits relatively stable performance across all horizons. While MAE and MSE increase slightly at certain prediction horizons (e.g., 48), the overall error remains moderate. Even at larger horizons (80 and 96), Transformer maintains relatively good performance, showcasing its stability when handling predictions at different time steps, but its predictive accuracy lags behind FGDLNet.

In summary, the proposed FGDLNet model demonstrates competitive performance in both short-term and long-term prediction tasks. Particularly in long-term forecasting, its MAE and MSE are significantly lower than those of other models. In comparison, traditional models generally perform poorly with long-term predictions, with errors increasing significantly as the prediction horizon expands. These results highlight the effectiveness and advantages of FGDLNet in wireless data transmission prediction.

4.6 Advanced metrics

In addition to the commonly used MSE, MAE, and R^2, this study introduces several additional evaluation metrics to comprehensively analyze the performance of different models. These additional metrics include p-value, 95% Confidence Interval, and Dynamic Time Warping. These metrics help evaluate the model's prediction ability, stability, and adaptability to complex time series data, especially when faced with noise, interference, and challenges across different time scales. Table 4 shows the performance of each model in terms of p-value, 95% confidence interval, and dynamic time warping:

Table 4: Model evaluation under high-dimensional metrics

Model	p-value	95% CI	DTW
Autoformer	0.0000	[0.4277, 0.5682]	2.2186
DLinear	0.0000	[0.4336, 0.5799]	5.1743
FGDLNet	0.0000	[0.0595, 0.1391]	0.4831
Informer	0.0000	[0.1218, 0.2707]	3.0734
ITransformer	0.0000	[0.5130, 0.6146]	5.0021
Reformer	0.0000	[0.5817, 0.7459]	7.2481
Transformer	0.0077	[0.0121, 0.0989]	1.2221

4.6.1 p-value and 95% confidence interval analysis

p-value and 95% Confidence Interval are two key metrics for evaluating the significance and stability of model predictions. Specifically, p-value is used to test whether there is a significant difference between the model's predictions and the actual values. A lower p-value indicates stronger predictive power. The 95% Confidence Interval reflects the consistency and stability of the model's predictions; a narrower interval suggests more consistent and reliable predictions.

In this study, we calculated the p-value and 95% Confidence Interval for each model. The results show that all models have a p-value close to 0.0000, indicating strong predictive power in most cases. However, it is noteworthy that the 95% Confidence Interval of the FGDLNet model is the narrowest ([0.0595, 0.1391]), suggesting that its predictions are more stable and consistent compared to other models.

4.6.2 Dynamic time warping metric analysis

Dynamic Time Warping is a metric used to measure the alignment of two time series along the time axis. It is commonly used to assess the performance of models in time series forecasting tasks. The smaller the DTW value, the higher the similarity between the predicted and actual time series, indicating better model performance.

From the experimental results, FGDLNet performs the best, achieving the lowest DTW value (0.4831) among all models. This low value indicates that FGDLNet aligns most closely with the actual data's time series, meaning it captures the dynamic patterns of the real data with high accuracy.

4.6.3 Residual plot analysis

In addition to the p-value, 95% confidence interval, and DTW, the residual plot also provides valuable insights into the model's prediction accuracy. The residual plot visualizes the difference (residuals) between the predicted values and the actual observations at different time steps. This is crucial for detecting error patterns that the model may have failed to capture.

A model's residual plot indicates its ability to capture the underlying dynamics of the data and whether there are any systematic errors. Ideally, the residuals should be randomly distributed around zero, suggesting that the model has captured all the predictable patterns in the data. Figure 5 shows a comparison of the residual plots for the FGDLNet model and the Transformer model. By comparing the residual plots of the two models, it is clear that FGDLNet outperforms the Transformer model in terms of prediction accuracy and stability. The residual points of FGDLNet are closer to the zero-residual line and are more evenly distributed, indicating that it better captures the dynamic variations in the time series data. While the Transformer model

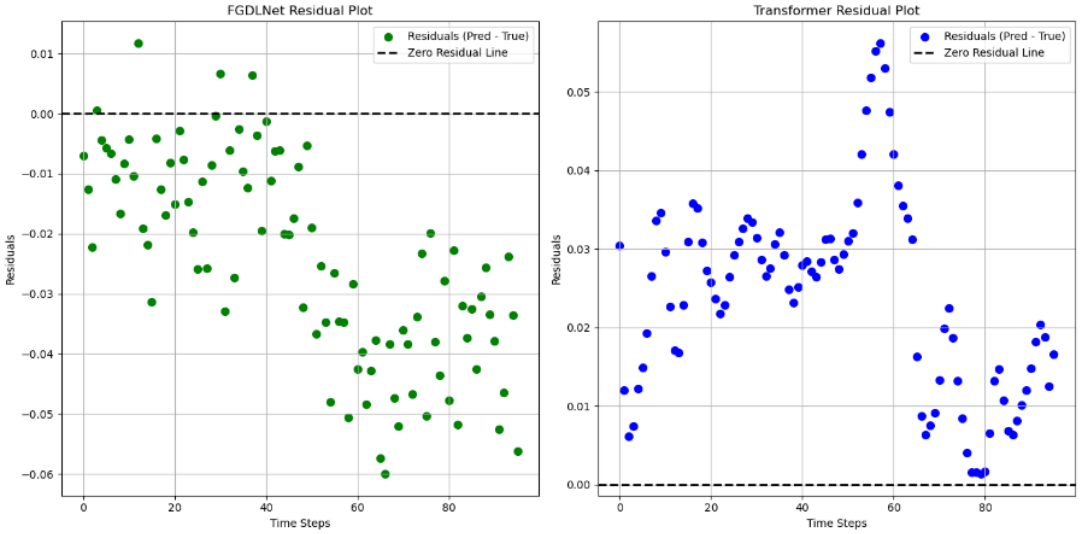


Figure 5: Residual plot

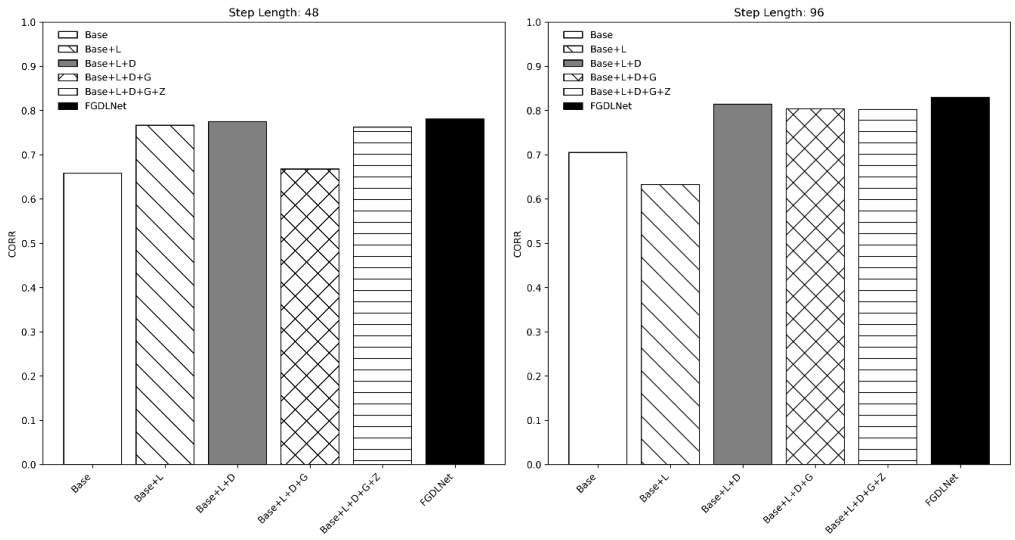


Figure 6: Ablation study

performs well at certain time steps, its consistency and accuracy across the entire time series are inferior to those of FGDLNet.

4.7 Ablation experiments

The FGDLNet model consists of five core modules: linear prediction, single-channel processing, multi-scale feature extraction enhanced with Graph Neural Networks, hybrid attention mechanism, and an improved loss function. To validate the effectiveness of each module in the FGDLNet model, six network variations were designed: (1) Base, which is the basic Transformer model; (2) Base+L, where the Decoder is removed and linear direct prediction is applied; (3) Base+L+D, where single-channel processing is applied for prediction; (4) Base+L+D+G, which incorporates multi-scale feature extraction and the GNN module; (5) Base+L+D+G+Z, which adds the hybrid attention mechanism; (6) FGDLNet, which incorporates the improved loss function. The results of the ablation experiments for prediction lengths of 48 and 96 steps are shown in Figure 6.

4.7.1 Efficiency and running time analysis

In this section, we provide a detailed analysis of the efficiency and runtime of the proposed FGDLNet model, comparing it with other models from the ablation experiments to assess the computational efficiency and prediction accuracy.

We selected three representative models from the ablation experiments for this analysis:

Base: The traditional Transformer model, which serves as the baseline for comparison.

Base+L: This model removes the Decoder and uses linear direct prediction, allowing us to isolate the effect of linear prediction without the full Transformer architecture.

FGDLNet: This model includes all core modules, including single-channel processing, multi-scale feature extraction with GNN, hybrid attention, and the improved loss function.

In this section, we analyze the theoretical time complexity of three models: Base, Base+L, and FGDLNet. The Base model, which corresponds to the original Transformer architecture, has a time complexity of $O(L^2 d)$, where L is the length of the input sequence and d is the dimensionality of the hidden layer. This quadratic complexity arises from the self-attention mechanism, which requires pairwise calculations for all tokens in the sequence. The Base+L model, which removes the Decoder and applies linear direct prediction, maintains the same time complexity for the self-attention mechanism as Base, i.e., $O(L^2 d)$. However, the linear prediction part has a complexity of $O(Ld)$, which reduces the overall computational burden by eliminating the Decoder. In contrast, FGDLNet, which incorporates local window attention, global linear attention, and a Graph Neural Network for feature extraction, offers a more efficient approach. The time complexity of FGDLNet for the attention mechanism is reduced to $O(wLd)$, where w is the window size, making it more efficient for long sequences compared to the quadratic complexity of the original Transformer. Additionally, the inclusion of GNN for extracting internal features from the single channel increases the time complexity to $O(T^2 D)$, where T represents the matrix multiplication for the self-adjacency matrix and D is the feature dimension used in the GNN. This theoretical analysis highlights the computational efficiency of FGDLNet in processing long sequence data, as it reduces the time complexity through hybrid attention mechanisms while adding the computational cost of GNN feature extraction. Table 5 shows the operational efficiency and mean absolute error of three models.

Table 5: Runtime Efficiency and MAE Comparison of Models

model	Training time	MAE
Base	1243.14	0.2948
Base+L	1007.17	0.3677
FGDLNet	1156.47	0.1701

Based on the table data, the Base model has the longest runtime. The Base+L model improves efficiency but with a decrease in accuracy. In contrast, the FGDLNet model achieves the best prediction accuracy while maintaining a good runtime. Therefore, FGDLNet outperforms the other models in terms of both prediction accuracy and efficiency.

5 Discussion

The Transformer model has achieved significant success in traditional time series forecasting, particularly in handling long sequences, where it effectively captures global dependencies. However, the Transformer model suffers from high computational complexity, especially with the self-attention mechanism, which has a time complexity of $O(n^2)$, where nnn is the sequence length. This results in a substantial increase in computational burden when the sequence length is large, particularly in resource-constrained environments.

The DLinear model is based on a simple linear regression approach for time series forecasting, which offers computational efficiency, especially for large-scale data. Its advantage lies in its simplicity and efficiency, making it suitable for scenarios requiring fast computation with relatively simple data patterns. However, its drawback is its limited ability to handle complex patterns and long-term dependencies in time series data, resulting in inferior predictive performance compared to more complex models.

FGDLNet simplifies the model structure by removing the Decoder module from the Transformer. Unlike the traditional Transformer, which relies on the Decoder module for generating the output sequence—requiring substantial computation and complex parameter adjustments—FGDLNet replaces it with a linear layer, directly connecting and predicting the output. This modification reduces the computation process and lowers the time complexity. Additionally, FGDLNet incorporates a Graph Neural Network module, enabling it to enhance dependencies between time steps within a single channel. This improvement allows FGDLNet to capture long-term dependencies and complex patterns more effectively than the traditional Transformer and DLinear. Consequently, FGDLNet enhances model expressiveness and prediction accuracy while maintaining relatively low computational complexity.

6 Conclusion

This paper presents FGDLNet, a time series forecasting model based on the Transformer architecture, designed to overcome the challenges of long sequence prediction. By removing the Decoder module from the traditional Transformer, incorporating multi-scale feature extraction, and employing a hybrid attention mechanism, FGDLNet significantly improves both prediction accuracy and computational efficiency.

FGDLNet addresses the high computational complexity of the traditional Transformer model, which arises from the self-attention mechanism. By using local window attention and global linear attention, FGDLNet reduces time complexity, making it more efficient for long sequence inputs. Additionally, the hybrid attention mechanism captures both local and global dependencies, while the integration of Graph Neural Networks for multi-scale feature extraction enhances the model's ability to capture complex dependencies in time series data.

Experimental results demonstrate that FGDLNet outperforms traditional methods in terms of both prediction accuracy and computational efficiency. The model strikes a good balance between accuracy and efficiency, making it suitable for long sequence forecasting tasks. Future work could focus on further optimizing GNN techniques, integrating multi-modal data, and exploring cross-domain applications to further

enhance the model's predictive power and generalization ability.

Conflict of Interest

The authors confirm that the content of this article has no conflict of interest.

Acknowledgement

This research was supported by the Key Scientific and Technological Project of Henan Province (242102210058).

References

[1] Vaswani A. Attention is all you need(2023). Advances in Neural Information Processing Systems, *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems* Pages 6000 - 6010. https://doi.org/10.48550/arXiv.1706.03762

[2] Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., and Zhang, W. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. In Proceedings of the *AAAI conference on artificial intelligence* Vol. 35, No. 12, pp. 11106-11115. https://doi.org/10.1609/aaai.v35i12.17325

[3] Kitaev, Nikita ,Łukasz Kaiser, and Anselm Levskaya. (2020)."Reformer: The efficient transformer." *arXiv preprint* arXiv:2001.04451 . https://doi.org/10.48550/arXiv.2001.04451

[4] Wu, H., Xu, J., Wang, J., & Long, M. (2021). Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, Vol.34, 22419-22430. https://doi.org/10.48550/arXiv.2106.13008

[5] Zeng A, Chen M, Zhang L, et al(2023). Are Transformers Effective for Time Series Forecasting? Proceedings of the *AAAI Conference on Artificial Intelligence*. Vol.37(9): 11121-11128. https://doi.org/10.1609/aaai.v37i9.26317

[6] Liu Y, Hu T, Zhang H, et al(2023). iTransformer: Inverted Transformers are Effective for Time Series Forecasting. *arXiv preprint* arXiv:2310.06625, 2023. https://doi.org/10.48550/arXiv.2310.06625

[7] Xu, Zhijian, Ailing Zeng, and Qiang Xu (2023). "FITS: Modeling time series with $10 k $ parameters." .Montavon G, Orr G B, Müller K R. Neural Networks: Tricks of the Trade. *arxiv preprint* arxiv:2307.03756 https://doi.org/10.48550/arXiv.2307.03756

[8] Cao, D., Jia, F., Arik, S. O., Pfister, T., Zheng, Y., Ye, W., & Liu, Y. (2023). Tempo: Prompt-based generative pre-trained transformer for time series forecasting. *arxiv preprint* arxiv:2310.04948. https://doi.org/10.48550/arXiv.2310.04948

[9] Yitian Zhang, Liheng Ma, Soumyasundar Pal, Yingxue Zhang, Mark Coates (2023). Multi-resolution Time-Series Transformer for Long-term Forecasting. *arxiv preprint* arXiv:2311.04147v2 https://doi.org/10.48550/arXiv.2311.04147

[10] Arjun Ashok, Étienne Marcotte, Valentina Zantedeschi, Nicolas Chapados, Alexandre Drouin (2023). TACTiS-2: Better, Faster, Simpler Attentional Copulas for Multivariate Time Series.*arxiv preprint* arxiv:2310.01327. https://doi.org/10.48550/arXiv.2310.01327

[11] Yang, Z., Liu, L., Li, N., & Tian, J. (2022). Time series forecasting of motor bearing vibration based on informer. *Sensors*, 22(15), 5858. https://doi.org/10.3390/s22155858

[12] Gong, M., Zhao, Y., Sun, J., Han, C., Sun, G., & Yan, B. (2022). Load forecasting of district heating system based on Informer. *Energy*, 253, 124179. https://doi.org/10.1016/j.energy.2022.124179

[13] Pan, K., Lu, J., Li, J., & Xu, Z. (2023). A Hybrid Autoformer Network for Air Pollution Forecasting Based on External Factor Optimization. *Atmosphere*, 14(5), 869. https://doi.org/10.3390/atmos14050869

[14] Feng, H., & Zhang, X. (2023). A novel encoder-decoder model based on Autoformer for air quality index prediction. *Plos one*, 18(4), e0284293. https://doi.org/10.1371/journal.pone.0284293

[15] Huang, Y., & Wu, Y. (2023). Short-Term Photovoltaic Power Forecasting Based on a Novel Autoformer Model. *Symmetry*, 15(1), 238. https://doi.org/10.3390/sym15010238

[16] Ahmed N K, Atiya A F, Gayar N E, et al. An empirical comparison of machine learning models for time series forecasting[J]. Econometric reviews, 2010, 29(5-6): 594-621. https://doi.org/10.1080/07474938.2010.481556

[17] Frissou N, Kimour M T, Selmane S. Optimized Support Vector Regression for Predicting Leishmaniasis Incidences[J]. *Informatica (Slovenia)*, 2021, 45(7). https://doi.org/10.31449/inf.v45i7.3665

[18] Liu Y, Pan B. Profit estimation model and financial risk prediction combining multi-scale convolutional feature extractor and BGRU model[J]. *Informatica (Slovenia)*, 2024, 48(11). https://doi.org/10.31449/inf.v48i11.5941

[19] Wang T, Yu J, Cao Y, et al. Fault Prediction of CNC Machine Tools Based on Gutenberg-Richter Law and Fuzzy Neural Networks[J]. *Informatica (Slovenia)*, 2024, 48(18). https://doi.org/10.31449/inf.v48i18.6292

[20] Flores A, Tito-Chura H. Multi-Step Forecasting of Guillain Barré Cases using Deep Learning[J]. *Informatica (Slovenia)*, 2024, 48(20). https://doi.org/10.31449/inf.v48i20.6358

Emotion Detection from Physiological Markers Using Machine Learning

M. Kocaleva Vitanova*, A. Stojanova Ilievska, N. Koceska and S. Koceski
Faculty of Computer Science, Goce Delcev University, Krste Misirkov 10A, Stip, Republic of North Macedonia
E-mail: mirjana.kocaleva@ugd.edu.mk, aleksandra.ilievska@ugd.edu.mk, natasa.koceska@ugd.edu.mk, saso.koceski@ugd.edu.mk
*Corresponding author

Keywords: emotion recognition, machine learning, physiological markers, heart rate, respiratory rate.

Received: October 24, 2024

Human emotion recognition in computer and robotic systems is crucial because it allows these systems to respond to users in a way that feels natural and supportive. By interpreting emotional cues, these systems can adjust their interactions—offering empathy, encouragement, or even assistance during times of distress—enhancing user satisfaction and making technology more accessible and engaging. Emotion recognition methods include analyzing facial expressions, vocal tone, and physiological signals, with the latter being especially effective because physiological data offers objective, real-time insights that are less susceptible to misinterpretation or masking than visible expressions. In this paper, we conducted an experiment for emotion recognition from physiological markers using machine learning algorithms. Each of the participants involved in the experiment was exposed to video stimuli designed to evoke specific emotions. Using physiological markers such as heart rate (HR) and respiratory rate (RR), seven emotions—anger, sadness, fear, amusement, neutrality, surprise, and happiness/joy—were analyzed. Three classification methods Random-forest, SVM and J48 were used. According to the results from the experimental evaluation, the highest accuracy for classifying emotions, based on both HR and RR across all emotions, was obtained with J48 algorithm. Specifically, the most clearly expressed and distinguishable emotions through RR were fear and sadness, with classification accuracies of 96.43% and 92.86%, respectively. Sadness was also the most accurately classified emotion through HR, with an accuracy of 85.71%. Gender differences were noted, with females reacting more to sadness and males to happiness.

Povzetek: Analizirano je strojno prepoznavanje čustev iz fizioloških markerjev, kot sta srčni utrip in hitrost dihanja. Največjo kvaliteto pri prepoznavanju čustev je dosegel algoritem J48, z največjo ntočnostjo pri strahu in žalosti.

1 Introduction

There has been a long-standing academic debate about the definition of emotion, without any widely agreed and universally accepted definition due to its complexity and diverse perspectives in psychology, neurology, and philosophy. However, contemporary study of emotions is usually based on the James-Lange theory of emotions, which suggests that a perceived stimulus is triggering physiological responses that are consequently felt as an emotion [1]. Emotions could be positive like love, happiness, hope, joy, affection, gratitude or negative such as anxiety, jealousy, frustration etc. Our own emotions can help us make sense of every situation. For example, appraisal theory states that our emotions are accompanied by inferences about the situation or environment in which we find ourselves [2], [3].

For a long period of time, emotions have been attributed to living beings or have been observed as unique human traits because only humans can experience a full range of emotions that are intrinsic to human experience, influencing cognition, behavior, decision-making, and social interactions. Nevertheless, throughout history, there have been various ideas and aspirations to create machines or artificial entities capable of emotions or human-like behaviors. This concept can be traced back to ancient myths, literature, and philosophical works. One of the earliest documented studies of robots with emotions and moral reasoning is Isaac Asimov's work "I, robot", published the mid of the past century. This concept of machines with embedded emotional intelligence has evolved significantly with the development of IT technologies and artificial intelligence and more particular generative AI. Rapid development of Artificial Intelligence (AI) and its widespread has affected various aspects of our daily life, starting from tourism [4], through medicine [5-7], biology [8], education [9-10], robotics [11-15], and in economy [16]. AI is becoming increasingly ubiquitous, and it is considered a transformative force in modern society. One of the tangible goals of the AI based society is creating emotionally intelligent systems. By integrating emotion recognition into AI based systems, they could become more adaptive, responsive, and human-centered.

In many real-world contexts, it is essential to understand an individual's emotional state to tailor interventions or optimize experiences. Understanding customers' emotional responses can permit companies to tailor

marketing strategies, making them more effective and targeted. Emotion recognition could be used in safety-critical industries, such as aviation or healthcare, to monitor the emotional states of workers. For example, detecting stress or fatigue in real time could prompt timely interventions, reducing the risk of errors or accidents. In healthcare, emotion recognition can assist in the early diagnosis of mood disorders such as depression or anxiety. In educational settings, monitoring students' emotional responses during learning activities can lead to better personalization and adaptive teaching strategies. Emotional responses are not only fundamental to personal well-being but also play a critical role in learning, decision-making, creativity, and social bonding. Emotional reactions can enhance or impede focus, motivation, and memory retention. For instance, positive emotions are associated with enhanced learning and creativity, while negative emotions such as stress and anxiety can impair cognitive functions and hinder performance.

One of the biggest challenges in human-computer interaction is recognition of emotions. While external cues such as facial expressions, vocal tone, and body language have been widely used to recognize emotions, these methods are not always reliable. People can mask their emotions, consciously or unconsciously, and external indicators may fail to capture the complexity of emotional responses. Some external factors (eg: external lights during facial expression analysis) can also negatively affect and cause errors in emotion recognition process. To address these limitations, emotion recognition through physiological signals has emerged as a more objective and precise approach. Following the James Lange theory, emotions are perception of states of the body and connected to physiological responses of the body [17]. Physiological signals, such as heart rate, respiration rate, skin conductance, and brain activity, are directly linked to the autonomic nervous system and are less prone to manipulation [18-19]. Therefore, using these signals can be the most reliable method for emotion recognition.

Physiological signals, nowadays, can be recorded with various unobtrusive, wearable devices equipped with various sensors, such as the Zephyr BioHarness [20]. This sensor can capture real-time changes in respiratory rate (RR) and heart rate (HR), both of which are closely associated with emotional arousal. Increased heart rate and respiratory rate usually indicate increased emotional stress or arousal, while a calm emotional state is often characterized by lower HR and RR. The non-invasive nature of Zephyr BioHarness sensor makes it well-suited for monitoring emotional responses in different contexts, such as experiments that involve exposure to external stimuli.

Emotional elicitation is an essential and often challenging task in the emotional recognition process. Previous studies utilized various stimuli to elicit emotions, such as: video/film clips [21-28], pictures [29-32], or music [33-36]. These stimuli are often pre-assigned to different categories based on a single model. Among these stimuli, video/film clips are the most naturalistic in intensity, that can be standardised and have a relatively high degree of ecological validity. Their cost-effectiveness, and the ability to evoke strong emotions while recreating some situations from the reality, makes them the most powerful method for emotions elicitation within a laboratory setting. The impact of videos clips on user emotions and psychology is especially important nowadays, with the rapid spread of social network and short-form videos. These videos, with a duration of less than 3min., are a new form of internet entertaining especially for the young population. The key psychological principle behind their popularity is attention span. Studies have shown that the average human attention span is decreasing, making it harder for people to stay focused on lengthy content. The videos also use visual and auditory cues to create strong and immediate emotional impact. This ability to evoke strong emotions in a short period encourages further user engagement, and cause addiction. Because of this, the emotions evoked by short-form videos should be well researched and recognized.

This paper proposes an approach for human emotion recognition based on physiological signals. Seven emotions: anger, sadness, fear, amusement, neutrality, surprise and happy/joy are analyzed using physiological signals such as HR (heart rate) and RR (respiratory rate). Emotions were elicited by visual stimuli such as different types of videos[1], each selected to induce specific emotional states (e.g., joy, fear, sadness, surprise). The proposed methodology was experimentally evaluated and the data from physiological signals obtained with Zephyr sensor were processed and analyzed. The results were cross-checked with the emotions felt and reported by the users. Statistical analysis, using ANOVA and t-test as two fundamental and powerful statistical techniques, were conducted on different groups to detect the possible existence of significant differences. Three classification methods for emotion recognition Random-forest, SVM and J48 were compared. Results of the experimental evaluation are elaborated and discussed in the following.

2 Related research

Emotions in people can be seen as a body response to external or internal stimuli. They are expressed in different ways that could be broadly classified as behavioral expressions - often manifest through observable behaviors, such as facial expressions, gestures, and vocal

[1] Neutral - https://www.youtube.com/watch?v=RpjwxyYefYc
Anger – https://www.youtube.com/watch?v=gG22XNhtnoY
Surprise – https://www.youtube.com/watch?v=-eREiQhBDIk
Fear - https://www.youtube.com/watch?v=WDpipB4yehk
Amusement - https://www.youtube.com/watch?v=R6rfCUDTbVE&list=RD_KOhxBGm5qA&index=3
Sadness - https://www.youtube.com/watch?v=cl-HNdzYh7Y
Happy/joy - https://www.youtube.com/watch?v=8tJoIaXZ0rw

changes and physiological responses - automatic bodily reactions controlled by the autonomic nervous system. These include changes in heart rate, respiration rate, sweating, and hormone release. Physiological signals are spontaneous reactions of the human body and consequently they are less prone to human alteration. They contain important information about the current emotional status of the person, and they are continuously following the changes of emotions. Consequently, two main approaches for emotion recognition could be identified in the literature. The first approach is based on classification of the emotional behavior of the user, such as his/her facial expressions and micro-actions, voice, body motion, body gestures, etc. While the second one is based on analysis of physiological signals of the human body, such as electrocardiogram (ECG), photoplethysmography (PPG), electroencephalogram (EEG), electrodermal activity (EDA) etc.

Facial expressions are the predominant mechanism utilized for emotion recognition. Saadon et al. [37] have proposed real-time emotion detection by quantitative facial motion analysis, while the authors in [38] proposed enhanced feature extraction technique through spatial deep learning model for facial emotion detection. Convolutional neural networks have also been tested for facial emotion recognition from regular video streams [39] while in [40] multiple machine learning techniques have been evaluated on emotion recognition from thermal and digital images. The conclusions derived in previous studies have also been reported in other scientific studies [41] where it is explained that facial emotions can easily be biased, therefore they should be used as an additional method or in combination with other behavioral methods to obtain valid results. Considering these conclusions Duan [42] combined facial expression data with physiological signals (skin conductance and heart rate variability) to enable improved emotion recognition in the context of mental health education. Emotion recognition in the context of mental health, was also conducted in [43], where SVM, traditional CNN, and improved CNN models were used on a self-built face database.

Recently published comprehensive review studies for emotion recognition from gait analyses [44] as well as a study reviewing the current state of the art related to approaches that combine facial expression and body gesture emotion recognition [45] suggest that the fusion of physical information and physiological signals can provide useful features of emotional states and lead to higher accuracy.

Physiological methods use various sensors to measure physiological responses, like change in skin conductance, changes in heart rate, breathing rate, pupil size, or even brain activity. Several studies published in the 1990s introduced the concept of emotion detection based on physiological signals obtained using wearable devices. The concept of “affective wearable” device capable of recognizing the human’s affective state using multiple integrated sensors for measurement of blood volume pressure, galvanic skin response (GSR), hearth rate (HR) and electromyogram (EMG) was introduced by Picard and Healey [46]. Saganowski et al. [47] conducted systematic literature review of the research conducted in the past three decades related to emotion recognition using physiological signals from wearables. Multiple physiological measures and methods of emotion elicitation (used stimuli for inducting emotions) have been identified. Petrantonakis et al. [29] and Valenza et al. [30] have used the International Affective Picture System (IAPS) to measure various emotions using an electroencephalogram (EEG) signal. The same picture database was also used by Basu et al. [31], but they measure GSR, HR, RR and Skin Temperature (ST) for emotion recognition. Images of natural scenes, labelled positive and negative, were used by Lan et al. [32] to evoke emotions that were then recognized by analyzing EEG data. Music stimuli has also been used in a few studies. Various emotions, like joy, anger, sadness and pleasure, were measured using various physiological signals: EMG, RR, ECG, skin conductance (SC) [33-34] and EEG [35-36]. Methods requiring subjects’ active participation, like playing computer games, were also used. Drachen et al. [48] measured HR and EDA during participants’ play of three different commercial first-person shooter games. A similar approach, using computer games, was used in other studies, but the focus was on measuring electrical signals in the heart (ECG) [50], as well as HR and skin perfusion [49]. More recently immersive VR-games were used for emotion induction, and a software platform for collecting player's physiological data [51]. Video clips, as effective stimuli, are widely applied in research studies that use physiological data for emotion recognition (Leila et al. [24] used ECG and respiration (RSP) data; Shin D. et al. [25] analyzed complex EEG/ECG bio-signals; Ayata et al. [72] proposes a model for emotion recognition using PPG and GSR). In [23] Valenza et al. have proved that synchronization exists between breathing patterns and heart rate during emotional visual elicitation. Emotion inducting film clips are also widely used in research studies. They are used in [26] where emotion recognition is performed using signal fusion: ECG, RSP, temperature (T), Electrodermal activity, EMG, Capnography and facial muscle activity; in [23] where EEG signal is used; in [27] where RR data were measured; in [28] where Skin conductance level (SCL) and HR were used.

Table 1 shows the details of the work of previous similar studies that address emotion recognition by measuring physiological data. All of them used visual stimuli (image, videos or film clips), and some used audio stimuli in addition to visuals. Most of them used data from various bio signals obtained using complex systems. However, authors in [24] shows that it is not necessary to make a data fusion to achieve better accuracy, a single well-chosen and measured signal can sometimes perform better. This of course depends also on classification algorithm as well as experimental procedures.

Several emotion recognition scenarios can be identified in the controlled lab environment. Most of them assume that there is a correlation between the subjective and conscious recognition of experienced emotions. However, studies that show variations between self-reported and measured data also exist [28].

Considering the inconsistencies in output results, we have decided to conduct an experiment for emotion recognition using only HR and RR as physiological markers.

Table 1: State of the art (SOTA)

Ref.	Physiological markers	Elicitation stimuli	Emotions	Classification algorithms	Dataset/ No of participants	Results
24	ECG, Respiration	music video clips	Low/high arousal, Positive/negative valence	SVM	Database for Emotion Analysis using Physiological Signals (DEAP) / 32 DEAP subjects	Resp. signal alone: liking = 73% arousal =72% valence =70% Both ECG and Resp.: liking = 76% arousal =74% valence =74%
25	ECG, EEG	video image data	amusement, fear, sadness, joy, anger, and disgust	Multi-layer perceptron (MLP), (SVM), Bayesian network (BN)	Self measurement /15 male and 15 female	EEG signal alone: MLP=44.86% SVM=24.99% BN=62.28% Complex bio-signal: MLP=83.97% SVM=63.97% BN=98.06%
26	ECG, RSP, temperature, EDA, EMG, capnography, facial muscle activity	Film clips	Fear, sadness, neutral	Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Multilayer Perceptron (MLP), Radial Basis Function Network (RBFN), K-Nearest Neighbours (KNN).	Self measurement / 34 participants	Subject-independent classification: RBFN Stimulus-independent classification: LDA
23	EEG	Film clips	Amused, disgusted, sad and neutral	SVM	Self measurement / 9 participants	SVM=97.2%
27	Respiration	Film clips/ narrative slide films with back. music	Love, sadness, joy, fear, and anger	k-Nearest Neighbour (k-NN), Probabilistic Neural Network (PNN).	Self measurement 33 participants	Average classification rate: 88%
28	SCL, HR	Film clips	Anger, fear, sadness, disgust, amusement, tenderness and neutral	//	Self measurement / 123 participants	A set of emotion-eliciting films was able to induce physiological changes in SCL and HR.
71	Respiratory, PPG, temperature	Music video	Arousal and valence levels	RF, SVM and logistic regression (LR)	DEAP / 32 participants	Best results with RF (arousal/valence): Respiratory=71.32% /70.62% PPG=68.59% /70.23% Temperature=69.68% /67.73% Fused=73.08% /72.18%
72	PPG, GSR	Video clips	Amusement, sadness, neutral	Linear SVM, LDA, multinomial regression (MN), decision trees (DT), and naive Bayes (NB).	Self measurement / 37 participants	Best result using Linear SVM: amusement=96% sadness=91% neutral=86%

3 Methodology

Our study for emotion recognition assumes that experienced emotions produce physiological changes in the human body and consequently they can be recognized using physiological markers and compared with subjective self-perception. The study is conducted according to the experimental scenario presented in Fig. 1. Our methodology was designed to recognize seven emotional states: anger, sadness, fear, amusement, surprise, happy /joy and neutral. Five of them (anger, sadness, fear, surprise, happy/joy) are widely accepted as basic emotions that we all feel, transcending language, regional, cultural, and ethnic differences [52]. Anger is a strong emotional response characterized by feelings of displeasure, frustration, or hostility, often triggered by perceived wrongs, injustice, or threats. It is typically associated with physiological arousal, such as increased heart rate and tension. Sadness is a negative emotional state marked by feelings of loss, disappointment, or helplessness. It is often accompanied by reduced energy, withdrawal from others, and a sense of melancholy. Fear is an intense emotional response to a perceived threat or danger, either real or imagined. It involves heightened physiological arousal, such as increased heart rate and adrenaline, and prepares the body for a fight-or-flight response. Surprise is an emotional reaction to an unexpected event or situation, characterized by a sudden feeling of astonishment or shock. It can be positive, negative, or neutral, depending on the context of the surprising event. Happy/joy is a positive emotional state associated with feelings of pleasure, contentment, or satisfaction. It is often marked by a sense of well-being, optimism, and energy. Amusement is an emotion with positive valence and high physiological arousal (causing a rapid heartbeat). It is one of the most important positive emotions that encourages people to engage in social interactions and promotes social bonding [53]. Neutral, on the other hand, is an emotional state where there is no strong positive or negative feeling. It is characterized by a balanced, calm demeanor, with little or no noticeable emotional arousal. However, they are still important because they can be experienced as a state of emotional balance, allowing a person to remain calm and objective in challenging or stressful circumstances.

The seven emotional states influence our thoughts, behaviors, and overall well-being. Therefore, by acknowledging and recognizing them, we can develop emotional intelligence and cultivate healthier relationships with ourselves and others.

Our methodological approach is described in the following.

Participants: A total of 78 healthy subjects participated in this study; all students at Faculty of Computer Science at Goce Delcev University in Stip. Fifty of them participated in the selection of emotional film clips, used in this research. The answers of these 50 participants were used to confirm the reliability of the selected film clips in eliciting the target emotions before the experiment commenced.

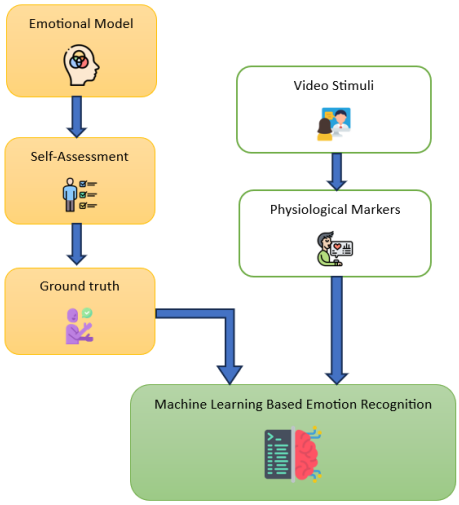


Figure 1: Emotion recognition scenario applied

The remaining 28 participants (20 male and 8 females) were involved in the experimental research. The average age of the participants was 22, with a minimum age of 21 and a maximum age 24. All participants use computers often and often watch videos with different content.

All measurements were performed in laboratory conditions. Informed consent was obtained from all participants, before the start of the experiment, confirming their willingness to participate in the study. Additionally, they signed a document ensuring their anonymity and consent to be part of the examination group. The procedure was explained to each participant and then each watched the videos alone, in the laboratory, and independently responded to the questionnaire without any external influences and suggestions.

Video stimuli selection: As the initial step, we selected 35 emotional film clips (5 for each selected emotion) from databases with prior evidence of reliability and validity in eliciting targeted emotions [54-58]. Fifty subjects participated in the selection phase to further refine and validate the selection of film clips. After watching the film clips, they completed a simple discrete emotion questionnaire, based on Emotion-Rating Form [59]. Of these 35 film clips only 7 were chosen (one for each emotion) for further use in this research study.

Experimental procedure: The study took place at the Laboratory of Robotics and Intelligent Systems at the Faculty of Computer Science (Goce Delcev University in Stip). Upon participant's arrival and environmental accommodation, a research assistant explained the experimental procedure (Fig. 2). Then the written consent was signed by the participant, and demographic data were collected. After placing the sensor on the participant, it was tested to ensure its proper functioning. The second phase of the experiment consists of three repeated stages: relax, stimuli and self-report. Participants first watched a 2 min. relaxing video, in order to rest and to calm emotions before next stimuli. This is important for establishing the baseline as a reference level. Then an emotion-eliciting video was played, after which participants filled out a self-assessment questionnaire. Each video lasted from 1:30-3 minutes. The order of the film clips was the same for all

participants. Participants reported their momentary feelings of anger, sadness, fear, amusement, surprise, happy/joy and neutral using a 5-point rating scale: “I did not feel it at all” (1), “I felt very little” (2), “I felt it a little” (3), “I felt it” (4) and “I felt it very strongly” (5). The entire experiment lasted about 50 minutes, depending on the time spent on the questionnaires.

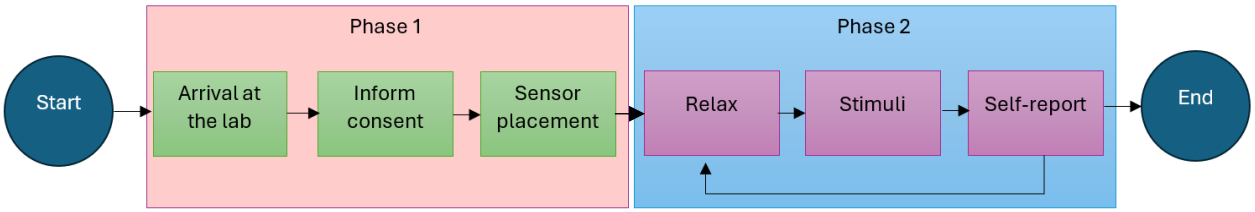


Figure 2: Experimental procedure

Physiological markers: Heart rate and respiration rate were used as physiological markers in this study. They were extracted from the signals captured using Zephyr BioHarness device. It is a physiological monitoring telemetry device intended for real-time monitoring. It consists of a chest strap and an electronics module that attaches to the strap. The module has a microprocessor and can directly compute HR and RR, but it can also stream the captured signals to a personal computer via integrated Bluetooth Low Energy (BLE) module.

Heart rate (HR) – is the frequency of the heartbeat measured by the number of contractions of the heart per minute. A normal resting heart rate, for adults, is between 60 to 100 beats per minute, but it can vary from minute to minute. The rate can be affected by numerous factors, including body's physical needs, physical fitness, genetics, hormonal status, environment, stress or psychological status… Heart rate increased when we are exposed to situations that evoke certain emotions, positive or negative (it increased to a greater degree in negative than in positive emotions). Therefore, it can be used as a valid physiological marker of emotional processes such as valence and arousal [60-62].

Respiratory rate (RR) - Respiration can be measured as the rate or volume at which an individual exchanges air in their lungs. Rate of respiration and depth of breath are the most common measures of respiration [63]. The respiratory rate is the rate at which breathing occurs; it is set and controlled by the respiratory center of the brain. A person's respiratory rate is usually measured in breaths per minute.

The normal resting respiratory rate of adults falls within the range of 16 to 20 breaths per minute. However, RR can be impacted by changes in emotions such as happiness, fear, sadness, surprise etc. Calm and positive emotions cause a lower breathing rate, a more regular breathing pattern, and a longer exhalation time than inhalation time, otherwise stress and fight situations cause a higher breathing rate, a more irregular breathing pattern, and a longer inhalation time than exhalation time. The connection between emotional states and respiration is investigated in several studies [64-67]

The breathing rate is the second “clock” of the body, in addition to the heart rate.

4 Analysis of results

A. Descriptive statistic

Seven types of videos (film clips) were used in the study: anger, fear, sadness, neutral, amusement, surprise and happiness. According to results, no one had seen the neutral and the surprise video before, just a few participants had seen the anger, fear and the happy video before. Most of participants had seen amusement video, while sadness video had almost the same number of participants who had watched the video before the measuring and those who had not (Table 2).

The anger and fear videos did not evoke positive feelings in most participants. Sad and surprise videos simultaneously cause positive and negative feelings. In addition, the rest 3 videos: happy, amusement and neutral are the most pleasant to watch according to the participants (Table 3).

Table 2: Frequency of participants' previous viewing experience by video type

Film clip	Yes	No
Anger	5	23
Fear	2	26
Sadness	14	14
Neutral	0	28
Surprise	0	28
Happiness	5	23
Amusement	21	7

Table 3: Feelings on the scale from 1 to 10

Film clip	1-5	6-10
Anger	24	4
Fear	19	9
Sadness	14	14
Neutral	9	19
Surprise	14	14
Happiness	4	24
Amusement	4	24

B. Emotion classification

Several classification algorithms such as: Random Forrest, Support Vector Machine (SVM) and J48, have been used to build the classification models. Random Forest (RF), SVM, and J48 offer a strong combination of flexibility, robustness, and generalization capability, making them suitable for a wide range of classification tasks, particularly those involving non-linear decision boundaries, high-dimensional data, and noisy environments. These characteristics make them more versatile and powerful options compared to other algorithms in many real-world scenarios.

Random Forest is an extension of decision trees that uses multiple decision trees, each trained on a sample of individuals [68]. For each split, a subset of random attributes is selected, and the best attribute is chosen for splitting. In the classification process, everyone is classified based on the majority vote from the group of trees. RF creates many trees, and each tree is built in the same way. Trees that split on the x variable will be built as far as possible with the y variable. The more data provided; the more trees can be grown. The placement of trees far from each other means that if one tree grows around the x variable, it is the development of the x tree. RF uses an ensemble method called "bagging" to reduce the risk of overfitting, especially when dealing with small datasets.

SVM method is a binary classification method by supervised learning [69]. The aim is to learn the $h(x)$ function through a training set given below:

$$\{(x_1, l_1), (x_2, l_2), \ldots\ldots (x_p, l_p)\} \epsilon\ RN\ x\ \{-1, 1\} \quad (1)$$

where l_k are the labels, x_k are input vectors, being in a space R^N and p is the size of the training set.

The technique seeks a separating hyperplane $h(x) = wx + w_0$ which minimizes the number of errors through the introduction of variable spring ξ_k, which can relax the constraints on the training vectors.

$$l_k(wx_k + w_0) \geq 1 - \xi_k, \xi_k > 0, 1 \leq k \leq p \quad (2)$$

With the previous constraints, the optimization problem is modified by a penalty term which penalizes high variables, spring ξ_k:

$$Minimise \frac{1}{2}\|w\|^2 + C\sum_{k=1}^{p} \xi_k, C > 0 \quad (3)$$

where C is a constant that controls the compromise between the number of classification errors and the margin width (the margin width being the smallest distance between training set and separating hyperplane).

J48 is a version of an earlier ID3 algorithm. It generates non-binary tree, uses measure called gain ratio to construct decision tree, the attribute with highest normalized gain ratio is taken as the root node, and the dataset is split based on the root element values. Again, the information gain is calculated for all the sub-nodes individually and the process is repeated until the prediction is completed. C4.5 is an evolution and refinement of ID3 that accounts for missing data values, continuous attribute value ranges, and pruning of decision trees and so on. Error–based pruning is performed after the growing phase. J48 can handle both continuous and discrete attributes, training data with missing attribute values and attributes with differing costs and provide an option to prune trees after creation. There are several parameters related to tree pruning in the J48 algorithm and should be used with care as they can make a noteworthy difference in the quality of results. J48 employs two post-pruning methods, namely sub tree replacement and sub tree raising [70].

C. Experimental results

We employed three classification algorithms-Support Vector Machine (SVM), Random Forest (RF), and J48 - to analyze the physiological data collected from 28 student participants in our experiment. The results for heart rate (HR) and respiratory rate (RR), categorized across seven distinct emotions, are presented in Tables 3 and 4, respectively.

For RR classification, six categories were defined. Given that the normal RR range for healthy individuals falls between 16 and 20 breaths per minute, we assigned class labels as follows: 0 for values below 16 (a low RR, often observed in relaxed states, especially during deep breathing or rest), 1 for values between 16 and 17 (slightly above normal, can indicate a mild increase in stress or alertness), 2 for values between 17 and 18 (moderate RR, potentially showing moderate anxiety or mild exertion), 3 for values between 18 and 19 (higher RR, typical of more stressed or anxious states), 4 for values between 19 and 20 (elevated, often associated with heightened anxiety, nervousness, or physical exertion), and 5 for values exceeding 20 (high RR, usually associated with intense emotional states like panic, fear, or excitement) (Table 4). Each emotion was analyzed individually, and a 10-fold cross-validation approach was employed, where 90% of the dataset was used for training, while the remaining 10% served as the test set. In 10-fold cross-validation, the dataset is divided into 10 equal parts. Each fold is used as a test set once, while the remaining 9 are used for training. This process repeats 10 times, and the results are average for reliable performance evaluation.

Similarly, we defined six categories for the heart rate (HR) classification. Since the normal HR range for healthy individuals spans from 60 to 100 beats per minute, we classified the data as follows: 0 for values below 60, (relaxed states, deep sleep, or well-trained athletes), 1 for values between 60 and 70 (moderate, which is within the normal resting range for most individuals), 2 for values between 70 and 80 (still within a normal resting range but could indicate slight arousal or anticipation), 3 for values between 80 and 90 (slightly elevated, often associated with moderate stress, anxiety, or physical exertion), 4 for values between 90 and 100 (elevated HR, which could correspond to emotional arousal, stress, or excitement), and 5 for values exceeding 100 (high HR, often linked to intense emotional states such as fear, anger, or high excitement). These HR thresholds reflect the natural response of the body to different emotional stimuli. Each emotion was analyzed individually, following the same classification approach (Table 5).

According to the results, the J48 algorithm produced the most accurate overall results for both heart rate (HR) and respiratory rate (RR) across all emotions. When comparing the other algorithms, Support Vector Machine (SVM) demonstrated superior performance for HR

classification, while Random Forest performed better for RR classification.
These results were also confirmed by the recall, precision, and F-measure values given in Table 6.

Table 4: Results for RR for different emotions examined by SVM, RF and J48

Emotion / RR	Neutral	Anger	Surprise	Fear	Amusement	Sadness	Happiness
RR (SVM)	57.1429%	50 %	64.2857%	67.8571%	71.4286%	71.4286%	71.4286%
RR (RF)	64.2857%	64.2857%	67.8571%	85.7143%	78.5714%	82.1429%	78.5714%
RR (J48)	78.5714%	75 %	85.7143%	96.4286%	78.5714%	92.8571%	82.1429%

Table 5: Results for HR for different emotions examined by SVM, RF and J48

Emotion / HR	Neutral	Anger	Surprise	Fear	Amusement	Sadness	Happiness
HR (SVM)	64.2857%	67.8571%	64.2857%	57.1429%	67.8571%	53.5714%	82.1429%
HR (RF)	53.5714%	67.8571%	57.1429%	46.4286%	67.8571%	64.2857%	60.7143%
HR (J48)	75 %	78.5714%	82.1429%	78.5714%	82.1429%	85.7143%	75%

Table 6: Precision, recall and F1- score values

	Emotions	precision			recall			F1-score		
		J48	RF	SVM	J48	RF	SVM	J48	RF	SVM
HR	Happiness	0.759	0.703	0.803	0.75	0.714	0.821	0.753	0.708	0.812
	Amusement	0.881	0.646	0.646	0.821	0.679	0.679	0.850	0.662	0.662
	Surprise	0.839	0.543	0.518	0.821	0.571	0.643	0.828	0.553	0.574
	Fear	0.796	0.496	0.603	0.786	0.536	0.571	0.791	0.502	0.568
	Sadness	0.871	0.717	0.577	0.857	0.679	0.536	0.862	0.697	0.538
	Neutral	0.786	0.601	0.699	0.75	0.571	0.643	0.761	0.578	0.659
	Anger	0.806	0.743	0.714	0.786	0.714	0.679	0.796	0.728	0.696
RR	Happiness	0.717	0.762	0.589	0.821	0.786	0.714	0.765	0.774	0.646
	Amusement	0.683	0.79	0.674	0.786	0.786	0.714	0.731	0.788	0.693
	Surprise	0.917	0.621	0.517	0.857	0.714	0.643	0.886	0.664	0.573
	Fear	0.976	0.889	0.66	0.964	0.893	0.679	0.965	0.891	0.667
	Sadness	0.967	0.869	0.521	0.929	0.893	0.714	0.948	0.881	0.602
	Neutral	0.774	0.637	0.452	0.786	0.679	0.571	0.777	0.657	0.505
	Anger	0.741	0.565	0.406	0.75	0.607	0.5	0.745	0.579	0.443

In our study, the SVM model tends to have lower accuracy compared to RF and J48, especially for some emotions such as fear (57.14% for HR, 67.86% for RR). This is consistent with findings from previous studies where SVM sometimes struggles with non-linear patterns in physiological data. The RF model performs better than SVM in most cases and shows comparable results with J48 for several emotions. RF is generally known for its robustness in handling high-dimensional, noisy data, which may be why it performs well in this context. The J48 decision tree classifier shows the highest accuracy across most emotions, particularly in fear (96.43% for RR and 78.57% for HR). Decision trees are often preferred in many studies for emotion classification due to their interpretability and ability to capture decision boundaries effectively, especially when combined with boosting or pruning methods.
When comparing our results to those of SOTA (Table 1), it is important to note that our study utilizes Heart Rate (HR) and Respiratory Rate (RR) as the primary physiological signals, which distinguishes it from other studies in this area. This makes the results not directly comparable to other studies which have used data from several physiological signals, as well as different elicitation techniques. However, if we compare the results of [24] or [27] or [71] - where respiration data are used -

we can see that in all three studies the obtained accuracy is lower than in our case when using J48 decision tree classifier. This fact indicates that using HR and RR for emotion recognition remains a valuable approach, especially in contexts where simplicity and non-invasiveness are a priority.

D. Correlation

Of the 28 participants, 8 were female, and all frequently use computers to watch videos, though -they do not often view content with varied genres. Regarding the four offered genres—comedy, horror, action, and films with a sad ending—it was found that males generally preferred action, while females favored both comedy and action. When asked about their prior exposure to the same videos, all female participants reported that they had not previously seen the videos that evoke fear, neutrality, or surprise, while some of the participants had seen the videos evoking sadness, happiness and amusement. The male participants also had not prior experience with the neutral and surprise-inducing videos, while most of them had watched the amusing video.

In order to find out whether previous viewing experience influences the level of target emotions produced, we conducted some statistical tests. An independent samples t-test was chosen, since we only have two groups to compare (participants who have and participants who have not previously watched the videos). According to the results, given in Table 7, for females, there is no statistically significant difference in emotional responses of sadness, happiness and amusement, based on the p-values (> 0.05) and t-stat (< t-critical) between first-time watchers and watched before groups. These three emotions were chosen, because there was a group of female students who has previously watched these videos.

Table 7: Analysis of female participants (have seen/have not seen the video before)

Female participants	**t-stat**	**t-critical**	**p-value**
sadness	0.0751	2.4469	0.9426
happiness	0.9020	2.4469	0.4018
amusement	2.0702	2.4469	0.0839

The same three emotions (sadness, happiness, and amusement) were also chosen for the male student group. According to the results obtained by t-test given in Table 8, for males, there is statistically significant difference in emotional responses of sadness and amusement, based on the p-values (> 0.05) and t-stat (< t-critical) between first-time watchers and watched before groups. This means that these videos had a different emotional impact on first-time watchers in male group. In fact, subjects who hadn't seen film clips before reported greater levels of target emotions than subjects who had previously seen these film clips. However, watching happy film clip before, does not significantly affect emotional response in happiness video, where t-stat (1.048) is less than the t-critical (2.101), and the p-value (0.309) is greater than 0.05. This means that there is no statistically significant difference between the two groups, indicating that there is no substantial difference in the emotional response to happiness among male participants who have already seen the happy video and those who are watching it for the first time.

Table 8: Analysis of male participants (have seen/have not seen the video before)

Male participants	**t-stat**	**t-critical**	**p-value**
sadness	2.672	2.101	0.016
happiness	1.048	2.101	0.309
amusement	3.029	2.101	0.007

Table 9 presents the self-assessed emotional responses from all participants, as well as separate data for male and female participants. Responses were rated on a scale from 1 to 5 for all emotions experienced while watching the different videos. The most prominent emotions identified were sadness, neutrality, happiness, and amusement, corresponding to the specific video content. According to the self-assessment, female participants experienced the strongest emotions while watching the sad video, while male participants expressed more emotion during the amusing and happy videos. Overall, it can be observed, that female participants reported more intense emotional experience then male, for all emotional states, which is consistent with the other findings [73], [74].

Table 9: Results from self-assessment

		All Participants							Males							Females						
		An	F	S	N	Su	H	A	An	F	S	N	Su	H	A	An	F	S	N	Su	H	A
Anger (An)	**1**	7	22	17	23	15	28	26	7	16	11	15	12	20	19		6	6	8	3	8	7
	2	9	4	6	4	8		2	7	3	4	4	6		1	2	1	2		2		1
	3	5	1	2	1	2			2	1	2	1	1			3				1		
	4	7	1	3		3			4		3		1			3	1			2		
	5																					
Fear (F)	**1**	13	3	25	27	14	28	27	10	3	18	19	11	20	19	3		7	8	3	8	8
	2	8	10	2	1	9		1	6	7	1	1	6		1	2	3	1		3		
	3	2	7	1		4			1	6	1		3			1	1			1		
	4	5	5			1			3	3						2	2			1		
	5		3							1							2					
Sadness (S)	**1**	2	21		11	22	28	28	1	15		9	17	20	20	1	6		2	5	8	8
	2	7	5	1	11	5			6	4	1	6	3			1	1		5	2		
	3	6	2	4	6	1			5	1	4	5				1	1		1	1		
	4	10		12					7		10					3		2				
	5	3		11					1		5					2		6				
Neutral (N)	**1**	18	18	13	1	14	1	7	11	13	7	1	7		1	7	5	6		7	1	6
	2	7	7	8	1	6	2	5	6	5	8	1	6	1	5	1	2				1	
	3	2		4	6	4	10	10	2		3	4	4	9	9			1	2		1	1
	4	1	3	3	17	4	11	6	1	2	2	14	3	8	5		1	1	3	1	3	1
	5				3		4							2					3		2	
Surprise (Su)	**1**	17	5	12	19	1	19	14	12	3	9	13	1	12	8	5	2	3	6		7	6
	2	10	7	8	6	4	3	6	7	4	7	4	3	3	6	3	3	1	2	1		
	3	1	8	4	2	11	4	6	1	8	2	2	8	4	5			2		3		1
	4		6	4	1	10	2	2		5	2	1	8	1	1		1	2		2	1	1
	5		2			2											2			2		
Happiness (H)	**1**	27	22	14	5	14		5	19	16	10	3	9		2	8	6	4	2	5		3
	2	1	3	4	5	5	1	9	1	3	3	3	4	1	5			1	2	1		4
	3		2	2	11	6	5	6		1	2	9	5	4	6		1		2	1	1	
	4		1	4	6	2	15	6			2	4	2	12	6		1	2	2		3	
	5			4	1	1	7	2			3	1		3	1			1		1	4	1
Amusement (A)	**1**	21	8	14	3	6			15	6	9	2	3			6	2	5	1	3		
	2	5	6	7	11	4	3	3	4	3	6	9	2	3	2	1	3	1	2	2		1
	3	2	11	5	5	4	6	5	1	9	4	4	3	3	4	1	2	1	1	1	3	1
	4		1	1	8	12	11	10		1		5	11	9	7			1	3	1	2	3
	5		2	1	1	2	8	10		1	1		1	5	7		1		1	1	3	3

According to the female participants, the video intended to evoke anger does not generate positive feelings for any of them. Additionally, videos eliciting fear, sadness, and surprise tend to evoke negative emotions. In contrast, videos associated with happiness, as well as those intended to induce amusement and neutrality, generally elicit positive emotions. For the male participants, videos featuring anger and fear were similarly deemed unpleasant, while amusing and happy videos were perceived positively.

Further statistical tests, including ANOVA and t-tests, were conducted to assess the intensity of sadness among female participants when watching the sad video. The ANOVA yielded a p-value of 0.33, which exceeds the significance threshold ($\alpha = 0.05$), suggesting no statistically significant difference in this case. The t-value obtained was 1.53, which is smaller than the critical t-value of 2.36 for 7 degrees of freedom. Both tests indicate that sadness is indeed a prominent emotion among female participants when watching the sad video.

For male participants, we conducted the same ANOVA and t-tests to examine the emotion of happiness during the happy video. The ANOVA returned to a p-value of 0.15, which is also greater than the significance level ($\alpha = 0.05$), suggesting no significant difference. The t-value obtained was 2.05, which is smaller than the critical t-value of 2.09 for 19 degrees of freedom. Both the ANOVA and t-test results confirm that male participants indeed experience strong feelings of happiness when watching the happy video.

From these statistical analyses, we can conclude that the most effective videos in terms of eliciting emotional responses were the sad and happy videos. Males predominantly reacted to happy scenes, while females were more emotionally affected by the sad video. Based on the participants' self-assessment ratings (scores of 4 or 5 indicating strong emotional responses), a correlation table was created to identify relationships between different emotions. This allowed us to further explore how emotions are interconnected during the viewing experience.

According to the results from Table 10, participants experienced strong emotions while watching the corresponding videos. Notably, there is a strong correlation between the emotions of sadness and anger, happiness and neutrality, and amusement and happiness. Given that the correlation value between amusement and happiness is the closest to 1, we selected the videos eliciting happiness and amusement for further analysis. We conducted ANOVA tests to compare amusement and happiness with all other emotions, as presented in Table 11.

The results presented in Table 11 confirm the correlation identified in Table 10 between amusement and happiness. Specifically, the findings indicate that watching happy videos elicits feelings of amusement; however, viewing amusing videos does not necessarily provoke feelings of happiness.

Table 10: Correlation table for filing strong emotion while watching videos

Emotions	**Anger**	**Fear**	**Sadness**	**Neutral**	**Surprise**	**Happiness**	**Amusement**
Anger	1						
Fear	-0,39768	1					
Sadness	0,702116	-0,57597	1				
Neutral	-0,52806	-0,16743	-0,25453	1			
Surprise	-0,66052	0,342058	-0,41427	0,177023	1		
Happiness	-0,67762	-0,29805	-0,19136	0,718203	0,335352	1	
Amusement	-0,58905	-0,14804	-0,26649	0,512828	0,67609	0,808552	1

Table 11: Comparison of amusement and happiness with other emotions (ANOVA p-values)

Video/emotion	**Anger**	**Fear**	**Sadness**	**Neutral**	**Surprise**	**Happiness**	**Amusement**
Happy/happiness	2.87E-27	2.87E-27	2.87E-27	0.051	5.32E-14	1	0.3
Amusing/amusement	1.14E-20	3.34E-21	8.98E-22	4.93E-06	1.69E-10	6.808E-05	1

Additionally, we analyzed the correlation between the self-assessment results and the experimental data for heart rate (HR) and respiratory rate (RR). We focused on the emotions of anger, fear, surprise, and neutrality. For instance, we compared the self-assessment of anger given after viewing the angry video with the corresponding experimental RR data obtained while watching the same video. The results from the ANOVA test ($p = 0.53$) and the t-test ($0.56 < 2.05$) indicate that there is no significant difference between the self-assessment and experimental results for respiratory rate (RR). This suggests that participants accurately expressed the emotions they experienced.

We conducted a similar correlation analysis between the self-assessment results and the experimental data for both RR and HR while watching a neutral video. We focused on this analysis, as the emotion of neutrality is associated with a decrease in both the breathing rate and heart rate.

The results from the ANOVA ($p = 0.75$ for RR and $p = 0.59$ for HR) and the t-tests ($0.30 < 2.05$ for RR and $0.58 < 2.05$ for HR) further demonstrate that there is no significant difference between the self-assessment and experimental results for both RR and HR. This indicates that participants accurately conveyed their emotional responses while watching the neutral video.

Additionally, we examined the emotions of fear and surprise, as both are associated with an increase in heart rate. We compared the self-assessment ratings for fear during the viewing of a scary video with the corresponding HR experimental results for the same video, utilizing

ANOVA and t-tests. This process was similarly applied to the video inducing surprise.

The results presented in Table 12 further confirm that the experimental data for heart rate (HR) during the viewing of scary and surprising videos aligns with the self-assessment responses provided by participants regarding their emotions of fear and surprise while watching these videos. This consistency reinforces the validity of the emotional responses reported by participants in relation to their physiological measurements.

Table 12: ANOVA and t-test (self- assessment and experimental RR/HR)

	Angry	**Neutral**		**Fear**	**Surprise**
	RR	HR	RR	HR	HR
ANOVA (p-value)	0.5346	0.5902	0.758	0.60364	0.701368
t-test (t-stat)	0.56266	0.582772	0.309365	0.5587	0.82735

5 Conclusion and summary

Human emotion recognition is essential for computer and robotic systems because it enables these systems to interact with people in a more empathetic, responsive, and adaptive manner. By understanding emotional cues, such as facial expressions, tone of voice, or physiological signals, robots and AI-driven interfaces can tailor their responses to meet the emotional needs of users, creating a more engaging and satisfying experience. This is particularly valuable in fields like healthcare, where emotionally aware systems can offer comfort and companionship to patients, and in customer service, where adaptive responses improve user satisfaction. Emotionally intelligent systems can also assist in education by gauging student engagement and adjusting lessons, ultimately making technology more intuitive and human centered.

The use of various physiological signals such as heart rate, skin conductance, EEG, EMG, brain activity, etc. allows for more precise and nuanced recognition of emotions. Emotion recognition through physiological signals is particularly effective because it provides objective, real-time insights into emotional states that may not be visible through external behaviors.

In this paper, we conducted an experiment for emotion recognition from physiological markers (HR and RR) using machine learning. We have selected HR and RR as these physiological signals are easy to collect with relatively compact and ergonomic wearable devices, which allows recording not only in a laboratory environment, but also in home conditions or outdoors. Seven distinct film clips, each selected to evoke a specific emotional response (anger, sadness, fear, amusement, surprise, happy/joy and neutral), were presented to the participants. The data from HR and RR were analyzed using three machine learning algorithms: SVM, Random Forest, and J48. The results showed that J48 consistently yielded the highest accuracy for classifying emotions based on both HR and RR across all emotions. SVM demonstrated greater effectiveness in classifying emotions based on HR, while RF performed better with RR. These algorithms have also been used in previous studies, but they have not always shown good results. For example, the accuracy of the SVM algorithm ranges from 24.99% in [49] to 97.2% in [47]. However, the small participant size (only 9 subjects) in the last study can be considered as a limitation. Also, the studies used different psychological markers (not HR and RR), as well as diverse datasets. This limits the direct comparison with our study. In many of the previous studies, classifications have been made on a two-dimensional scale (valence and arousal). Random forest, for example, has been chosen as the best classification algorithm in [69], with accuracy of 71.32% for arousal, and 70.62% for valence, when using respiration, as a physiological marker. On the other hand, in our study, we have used the categorical emotional model, measuring seven discrete emotions. The emotions most clearly expressed and distinguishable through RR were fear and sadness, with classification accuracies of 96.43% and 92.86%, respectively. Sadness was also the most accurately classified emotion through HR, with an accuracy of 85.71%. Additionally, gender differences were observed, with female participants displaying stronger emotional responses to the sad video, while male participants showed more pronounced reactions to the happy video. This is partially in line with the previous study [52] which shows that women react more intensively to sadness than man. However, the gender differences should be investigated in more detail considering socio-cultural differences as well as preferences and personality traits (since emotions are highly subjective).

The correlation analyses that were conducted showed that the strongest correlation was between happiness and amusement, while other notable correlations were observed between happiness and neutral, as well as sadness and anger. Furthermore, when comparing the experimental results for HR and RR with participants' self-reported emotions, we found that participants were generally truthful in their emotional self-assessments. Taking into consideration that people often mask their feelings and don’t want to show and admit them (especially male), this is a positive fact.

However, the number of subjects used in the experiment, and facts that they have same cultural background and similar ages, is a limitation of this study. Our future research should include more diverse participants, as well as conducting experiments in real-world settings rather than in the laboratory.

References

[1] Retkoceri, U. (2022). Remembering emotions. Biology & Philosophy, 37(1), 5. https://doi.org/10.1007/s10539-022-09834-5

[2] Hamby, A. and Jones, N., 2022. The effect of affect: An appraisal theory perspective on emotional engagement in narrative persuasion. Journal of Advertising, 51(1), pp.116-131. https://doi.org/10.1080/00913367.2021.1981498

[3] Doan, T., Ong, D.C. and Wu, Y., 2024. Emotion understanding as third-person appraisals: Integrating appraisal theories with developmental theories of emotion. Psychological Review. https://doi.org/10.31234/osf.io/mdpsr

[4] Koceski, S. and Petrevska, B., 2012. Empirical evidence of contribution to e-tourism by application of personalized tourism recommendation system. Scientific Annals of the "Alexandru Ioan Cuza" University of Iasi–Economic Sciences Section, (1), pp.371-383. https://doi.org/10.2478/v10316-012-0025-6

[5] Trajkovik, V., Vlahu-Gjorgievska, E., Koceski, S. and Kulev, I., 2015. General assisted living system architecture model. In Mobile Networks and Management: 6th International Conference, MONAMI 2014, Würzburg, Germany, September 22-26, 2014, Revised Selected Papers 6 (pp. 329-343). Springer International Publishing. https://doi.org/10.1007/978-3-319-16292-8_24

[6] Stojanov, D. and Koceski, S., 2014, September. Topological MRI prostate segmentation method. In 2014 Federated Conference on Computer Science and Information Systems (pp. 219-225). IEEE. https://doi.org/10.15439/2014f204

[7] Koceska, N., Komadina, R., Simjanoska, M., Koteska, B., Strahovnik, A., Jošt, A., ... & Trontelj, J. (2020). Mobile wireless monitoring system for prehospital emergency care. European Journal of Trauma and Emergency Surgery, 46, 1301-1308. https://doi.org/10.1007/s00068-019-01130-4

[8] Stojanov, D., Mileva, A. and Koceski, S., 2012. A new, space-efficient local pairwise alignment methodology. Advanced Studies in Biology, 4(2), pp.85-93. https://doi.org/10.1080/13102818.2014.95971

[9] Koceski, S., Koceska, N., Lazarova, L. K., Miteva, M., & Zlatanovska, B. (2025). Exploring the performance of ChatGPT for numerical solution of ordinary differential equations. Journal of Technology and Science Education, 15(1), 18-34. https://doi.org/10.3926/jotse.2709

[10] Velinov, A., Koceska, N., & Koceski, S. (2024). Virtual Tour Using Telepresence Robot and MQTT Protocol. TEM Journal, 13(1), 750-756. https://doi.org/10.18421/tem131-78

[11] Koceski, S., Koceska, N. and Kocev, I., 2012. Design and evaluation of cell phone pointing interface for robot control. International Journal of Advanced Robotic Systems, 9(4), p.135. https://doi.org/10.5772/51614

[12] Koceski, S., Panov, S., Koceska, N., Zobel, P.B. and Durante, F., 2014. A novel quad harmony search algorithm for grid-based path finding. International Journal of Advanced Robotic Systems, 11(9), p.144. https://doi.org/10.5772/58875

[13] Koceska, N., Koceski, S., Durante, F., Zobel, P.B. and Raparelli, T., 2013. Control architecture of a 10 DOF lower limbs exoskeleton for gait rehabilitation. International Journal of Advanced Robotic Systems, 10(1), p.68. https://doi.org/10.5772/55032

[14] Koceska, N., & Koceski, S. (2022). Supporting elderly's independent living with a mobile robot platform. The Journal of Universal Computer Science-JUCS, 28(5), 475-498. https://doi.org/10.3897/jucs.76579

[15] Koceska, N., Koceski, S., Beomonte Zobel, P., Trajkovik, V., & Garcia, N. (2019). A telemedicine robot system for assisted and independent living. Sensors, 19(4), 834. https://doi.org/10.3390/s19040834

[16] Koceska, N. and Koceski, S., 2014. Financial-Economic Time Series Modeling and Prediction Techniques–Review. Journal of Applied Economics and Business, 2(4), pp.28-33. https://doi.org/10.1109/meco.2014.6862685

[17] Van Doorn, E.A., Van Kleef, G.A. and Van der Pligt, J., 2015. Deriving meaning from others' emotions: attribution, appraisal, and the use of emotions as social information. Frontiers in Psychology, 6, p.1077. https://doi.org/10.3389/fpsyg.2015.01077

[18] Valderas, M.T., Bolea, J., Laguna, P., Vallverdú, M. and Bailón, R., 2015, August. Human emotion recognition using heart rate variability analysis with spectral bands based on respiration. In 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 6134-6137). IEEE. https://doi.org/10.1109/embc.2015.7319792

[19] Kocaleva, M., Stojanova, A. and Koceska, N., 2017. Using physiological parameters for evaluating User Experience. ITRO 2017, 22 June 2017, Zrenjanin, Serbia.

[20] Medtronic Zephyr, Zephyr Performance Systems, Accesed on 05.02.2025 Available (online): https://www.zephyranywhere.com/

[21] Bruun, A., Law, E.L.C., Heintz, M. and Eriksen, P.S., 2016, October. Asserting real-time emotions through cued-recall: is it valid?. In Proceedings of the 9th Nordic Conference on Human-Computer Interaction (pp. 1-10). https://doi.org/10.1145/2971485.2971516

[22] Valenza, G., Lanata, A. and Scilingo, E.P., 2012. Oscillations of heart rate and respiration synchronize during affective visual stimulation. IEEE Transactions on Information Technology in Biomedicine, 16(4), pp.683-690. https://doi.org/10.1109/titb.2012.2197632

[23] Valenzi, S., Islam, T., Jurica, P., & Cichocki, A. (2014). Individual classification of emotions using EEG. Journal of Biomedical Science and Engineering, 2014. https://doi.org/10.4236/jbise.2014.78061

[24] Mirmohamadsadeghi, L., Yazdani, A., & Vesin, J. M. (2016, September). Using cardio-respiratory signals to recognize emotions elicited by watching music video clips. In 2016 IEEE 18th International Workshop on Multimedia Signal Processing (MMSP) (pp. 1-5). IEEE. https://doi.org/10.1109/mmsp.2016.7813349

[25] Shin D., Shin D., Shin D. Development of emotion recognition interface using complex EEG/ECG bio-signal for interactive contents. Multimed. Tools Appl. 2017;76:11449–11470. https://doi.org/10.1007/s11042-016-4203-7

[26] Kolodyazhniy V., Kreibig S.D., Gross J.J., Roth W.T., Wilhelm F.H. An affective computing approach to physiological emotion specificity: Toward subject-independent and stimulus-independent classification of film-induced emotions. Psychophysiology. 2011;48:908–922. https://doi.org/10.1111/j.1469-8986.2010.01170.x

[27] Wu C.K., Chung P.C., Wang C.J. Representative Segment-Based Emotion Analysis and Classification with Automatic Respiration Signal Segmentation. IEEE Trans. Affect. Comput. 2012;3:482–495. https://doi.org/10.1109/T-AFFC.2012.14

[28] Fernández, C., Pascual, J. C., Soler, J., Elices, M., Portella, M. J., & Fernández-Abascal, E. (2012). Physiological responses induced by emotion-eliciting films. Applied psychophysiology and biofeedback, 37, 73-79. https://doi.org/10.1007/s10484-012-9180-7

[29] Petrantonakis P.C., Hadjileontiadis L.J. Emotion Recognition from Brain Signals Using Hybrid Adaptive Filtering and Higher Order Crossings Analysis. IEEE Trans. Affect. Comput. 2010;1:81–97. https://doi.org/10.1109/t-affc.2010.7

[30] Valenza G., Lanata A., Scilingo E.P. The Role of Nonlinear Dynamics in Affective Valence and Arousal Recognition. IEEE Trans. Affect. Comput. 2012;3:237–249. https://doi.org/10.1109/T-AFFC.2011.30

[31] Basu S., Jana N., Bag A., Mahadevappa M., Mukherjee J., Kumar S., Guha R. Emotion recognition based on physiological signals using valence-arousal model; Proceedings of the Third International Conference on Image Information Processing; Waknaghat, India. 21–24 December 2015; pp. 50–55. https://doi.org/10.1109/iciip.2015.7414739

[32] Lan Z., Sourina O., Wang L., Liu Y. Real-time EEG-based emotion monitoring using stable features. Vis. Comput. 2016;32:347–358. https://doi.org/10.1007/s00371-015-1183-y

[33] Wong W.M., Tan A.W., Loo C.K., Liew W.S. PSO optimization of synergetic neural classifier for multichannel emotion recognition; Proceedings of the 2010 Second World Congress on Nature and Biologically Inspired Computing (NaBIC); Kitakyushu, Japan. 15–17 December 2010; pp. 316–321. https://doi.org/10.1109/nabic.2010.5716292

[34] Zied G., Lachiri Z., Maaoui C. Emotion recognition from physiological signals using fusion of wavelet based features; Proceedings of the International Conference on Modelling, Identification and Control; Sousse, Tunisia. 18–20 December 2015. https://doi.org/10.1109/icmic.2015.7409485

[35] Lin Y.P., Wang C.H., Jung T.P., Wu T.L., Jeng S.K., Duann J.R., Chen J.H. EEG-Based Emotion Recognition in Music Listening. IEEE Trans. Bio-Med. Eng. 2010;57:1798–1806. https://doi.org/10.1109/TBME.2010.2048568.

[36] Hsu, Y.L.; Wang, J.S.; Chiang, W.C.; Hung, C.H. Automatic ECG-based emotion recognition in music listening. IEEE Trans. Affect. Comput. 2017, 11, 85–99. https://doi.org/10.1109/taffc.2017.2781732

[37] Saadon, J.R., Yang, F., Burgert, R., Mohammad, S., Gammel, T., Sepe, M., Rafailovich, M., Mikell, C.B., Polak, P. and Mofakham, S., 2023. Real-time emotion detection by quantitative facial motion analysis. Plos one, 18(3), p.e0282730. https://doi.org/10.1371/journal.pone.0282730

[38] Khan, N., Singh, A.V. and Agrawal, R., 2023. Enhancing feature extraction technique through spatial deep learning model for facial emotion detection. Annals of Emerging Technologies in Computing (AETiC), 7(2), pp.9-22. https://doi.org/10.33166/aetic.2023.02.002

[39] Sarvakar, K., Senkamalavalli, R., Raghavendra, S., Kumar, J.S., Manjunath, R. and Jaiswal, S., 2023. Facial emotion recognition using convolutional neural networks. Materials Today: Proceedings, 80, pp.3560-3564. https://doi.org/10.1016/j.matpr.2021.07.297

[40] Sathyamoorthy, B., Snehalatha, U. and Rajalakshmi, T., 2023. Facial emotion detection of thermal and digital images based on machine learning techniques. Biomedical Engineering: Applications, Basis and Communications, 35(01), p.2250052. https://doi.org/10.4015/s1016237222500521

[41] Naga, P., Marri, S.D. and Borreo, R., 2023. Facial emotion recognition methods, datasets and technologies: A literature survey. Materials Today: Proceedings, 80, pp.2824-2828. https://doi.org/10.1016/j.matpr.2021.07.046

[42] Duan, Y. (2024). Mental Health Education Evaluation Model Based on Emotion Recognition Algorithm. Informatica, 48(18). https://doi.org/10.31449/inf.v48i18.6420

[43] Peng, X. (2021). Research on Emotion Recognition Based on Deep Learning for Mental Health. Informatica, 45(1). https://doi.org/10.31449/inf.v45i1.3424

[44] Xu, S., Fang, J., Hu, X., Ngai, E., Wang, W., Guo, Y. and Leung, V.C., 2022. Emotion recognition from gait analyses: Current research and future directions. IEEE Transactions on Computational Social Systems, 11(1), pp.363-377. https://doi.org/10.1109/tcss.2022.3223251

[45] Leong, S.C., Tang, Y.M., Lai, C.H. and Lee, C.K.M., 2023. Facial expression and body gesture emotion recognition: A systematic review on the use of visual data in affective computing. Computer Science Review, 48, p.100545. https://doi.org/10.1016/j.cosrev.2023.100545

[46] Picard, R.W. and Healey, J., 1997. Affective wearables. Personal technologies, 1, pp.231-240. https://doi.org/10.1007/bf01682026

[47] Saganowski, S., Perz, B., Polak, A.G. and Kazienko, P., 2022. Emotion recognition for everyday life using physiological signals from wearables: A systematic literature review. IEEE Transactions on Affective Computing, 14(3), pp.1876-1897. https://doi.org/10.1109/taffc.2022.3176135

[48] Drachen, A., Nacke, L.E., Yannakakis, G. and Pedersen, A.L., 2010, July. Correlation between heart rate, electrodermal activity and player experience in first-person shooter games. In Proceedings of the 5th ACM SIGGRAPH Symposium on Video Games (pp. 49-54). https://doi.org/10.1145/1836135.1836143

[49] Bentley, T., Johnston, L. and von Baggo, K., 2005, November. Evaluation using cued-recall debrief to elicit information about a user's affective experiences. In Proceedings of the 17th Australia Conference on Computer-Human Interaction: Citizens Online: Considerations for Today and the Future (pp. 1-10). https://doi.org/10.1007/978-3-540-27795-8_52

[50] Agrafioti F., Hatzinakos D., Anderson A.K. ECG Pattern Analysis for Emotion Detection. IEEE Trans. Affect. Comput. 2012;3:102–115. https://doi.org/10.1109/T-AFFC.2011.28

[51] Bassano, C., Ballestin, G., Ceccaldi, E., Larradet, F., Mancini, M., Volta, E., et al. (2019). A VR Game-based System for Multimodal Emotion Data Collection, 12th annual ACM SIGGRAPH conference on Motion, Interaction and Games 2019 (MIG 2019) (Newcastle Upon Tyne).
https://doi.org/10.1145/3359566.3364695

[52] Ekman, P. Basic Emotions. In Dalgleish, T. & Power, M. J. (Eds.), Handbook of Cognition and Emotion, New York, NY: John Wiley & Sons Ltd., pp. 45-60, (1999).
https://doi.org/10.1002/0470013494.ch3

[53] Kryś, K. (2010). May amusement serve as a social courage engine?. Polish Psychological Bulletin
https://doi.org/10.2478/v10059-010-0009-z

[54] Hewig, J. et al. A revised film set for the induction of basic emotions. Cognition and emotion 19, 1095 (2005).
https://doi.org/10.1080/02699930541000084

[55] Kaczmarek, L. D. et al. Splitting the affective atom: Divergence of valence and approach-avoidance motivation during a dynamic emotional experience. Current Psychology 1–12 (2019).
https://doi.org/10.1007/s12144-019-00264-3

[56] Gross, J. J. & Levenson, R. W. Emotion elicitation using films. Cognition & emotion 9, 87–108 (1995).
https://doi.org/10.1080/02699939508408966

[57] Schaefer, A., Nils, F., Sanchez, X. & Philippot, P. Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. Cognition and emotion 24, 1153–1172 (2010).
https://doi.org/10.1080/02699930903274322

[58] Reynaud, E., El-Khoury-Malhame, M., Blin, O. & Khalfa, S. Voluntary emotion suppression modifies psychophysiological responses to films. Journal of Psychophysiology 26, 116 (2012).
https://doi.org/10.1027/0269-8803/a000074

[59] Smith, C. A., Haynes, K. N., Lazarus, R. S., & Pope, L. K. (1993). In search of the "hot" cognitions: Attributions, appraisals, and their relation to emotion. Journal of Personality and Social Psychology, 65(5), 916–929.
https://doi.org/10.1037//0022-3514.65.5.916

[60] Brosschot, J. F., & Thayer, J. F. (2003). Heart rate response is longer after negative emotions than after positive emotions. International journal of psychophysiology, 50(3), 181-187.
https://doi.org/10.1016/s0167-8760(03)00146-6

[61] Anttonen, J., & Surakka, V. (2005, April). Emotions and heart rate while sitting on a chair. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 491-499).
https://doi.org/10.1145/1054972.1055040

[62] Robazza, C., Bortoli, L., & Nougier, V. (1999). Emotions, heart rate and performance in archery: A case study. Journal of Sports Medicine and Physical Fitness, 39(2), 169.

[63] Gouizi, K., Bereksi Reguig, F. and Maaoui, C., 2011. Emotion recognition from physiological signals. Journal of medical engineering & technology, 35(6-7), pp.300-307.
https://doi.org/10.3109/03091902.2011.601784

[64] Zhang, Q., Chen, X., Zhan, Q., Yang, T., & Xia, S. (2017). Respiration-based emotion recognition with deep learning. Computers in Industry, 92, 84-90.
https://doi.org/10.1016/j.compind.2017.04.005

[65] Philippot, P., Chapelle, G., & Blairy, S. (2002). Respiratory feedback in the generation of emotion. Cognition & Emotion, 16(5), 605-627.
https://doi.org/10.1080/02699930143000392

[66] Homma, I., & Masaoka, Y. (2008). Breathing rhythms and emotions. Experimental physiology, 93(9), 1011-1021.
https://doi.org/10.1113/expphysiol.2008.042424

[67] Ashhad, S., Kam, K., Del Negro, C. A., & Feldman, J. L. (2022). Breathing rhythm and pattern and their influence on emotion. Annual review of neuroscience, 45(1), 223-247.
https://doi.org/10.1146/annurev-neuro-090121-014424

[68] Breiman, L., 2001. Random forests. Machine learning, 45, pp.5-32.
https://doi.org/10.1023/a:1010933404324

[69] Hasan, M. and Boris, F., 2006. Svm: Machines à vecteurs de support ou séparateurs à vastes marges. Rapport technique, Versailles St Quentin, France. Cité, 64.

[70] Bhargava, N., Sharma, G., Bhargava, R., & Mathuria, M. (2013). Decision tree analysis on j48 algorithm for data mining. Proceedings of international journal of advanced research in computer science and software engineering, 3(6).
https://doi.org/10.46243/jst.2022.v7.i05.pp200-205

[71] Ayata, D., Yaslan, Y., & Kamasak, M. E. (2020). Emotion recognition from multimodal physiological signals for emotion aware healthcare systems. Journal of Medical and Biological Engineering, 40, 149-157.
https://doi.org/10.1007/s40846-019-00505-7

[72] Domínguez-Jiménez, J. A., Campo-Landines, K. C., Martínez-Santos, J. C., Delahoz, E. J., & Contreras-Ortiz, S. H. (2020). A machine learning model for emotion recognition from physiological signals. Biomedical signal processing and control, 55, 101646.
https://doi.org/10.1016/j.bspc.2019.101646

[73] LaFrance, M., & Banaji, M. (1992). Toward a Reconsideration of the Gender-Emotion Relationship. In M. S. Clark (Ed.), Review of Personality and Social Psychology: Emotions and Social Behavior (Vol. 14, pp. 178-202). Newbury Park, CA: Sage.

[74] Shields, S. A. (1991). Gender and the Psychology of Emotion: A Selective Research Review. In K. T. Strongman (Ed.), International Review of Studies on Emotion (pp. 227-247). New York: Wiley.

Enhanced Hate Speech Detection in Indonesian-English Code-Mixed Texts Using XLM-RoBERTa

Farrel Dinarta, Arya Wicaksana*
Department of Informatics, Universitas Multimedia Nusantara, Tangerang 15810, Banten, Indonesia
E-mail: farrel.dinarta@student.umn.ac.id, arya.wicaksana@umn.ac.id
*Corresponding author

Keywords: hate speech detection, Indonesian-English, XLM-RoBERTa

Received: December 1, 2024

The prevalence of hate speech on digital platforms presents significant challenges, particularly in multilingual communities where code-mixing complicates detection. This study explores the use of XLM-RoBERTa, a transformer-based model with robust multilingual capabilities, to detect hate speech within code-mixed texts, focusing on Indonesian-English code-mixing. Traditional hate speech detection models rely on single-language datasets, limiting their effectiveness in such environments. We employ a dataset consisting of Indonesian, English, and code-mixed Indonesian-English texts to evaluate XLM-RoBERTa's performance. The dataset comprises 24,844 training samples, 2,760 test samples, and an additional 100 supplementary samples. Key hyperparameters included a batch size of 16 and 32, with a learning rate ranging from 1e-5 to 5e-5. The model achieved near-perfect accuracy (99.6%) on the primary test set and demonstrated strong generalization across realistic supplementary data, achieving an F1-score of 90.94%. These findings underscore the model's potential for application in complex linguistic contexts, contributing to the development of effective code-mixed hate speech detection.

Povzetek: Raziskana je uporaba modela XLM-RoBERTa za zaznavanje sovražnega govora v besedilih z mešanjem indonezijskega in angleškega jezika, kar izboljša kvaliteto zaznavanja v večjezičnih okoljih.

1 Introduction

The rapid growth of social media and online communication platforms has significantly transformed global interactions, allowing for exchanging ideas across diverse linguistic and cultural boundaries. However, this interconnectedness has also contributed to the proliferation of harmful content, such as hate speech, which presents considerable social and ethical challenges. Detecting and moderating hate speech is essential for fostering a safe and inclusive digital environment [1], [2].

The task becomes increasingly complex in code-mixed texts, where users frequently switch between languages within a single conversation or incorporate terms from multiple languages [2]. Code-mixing is prevalent in multilingual communities and among bilingual individuals. Traditional hate speech detection systems, which primarily rely on single-language datasets and algorithms, struggle to identify harmful content effectively in these contexts. Most conventional approaches employ supervised machine learning models or rule-based systems that require extensive language-specific resources, such as labeled datasets and lexicons [3].

Code-mixed datasets are particularly challenging to obtain, as many publicly available datasets focus on individual languages, given that most people typically communicate in one language at a time. Code-mixing is more common in specific domains or platforms, such as social media and multilingual communities. While these platforms are invaluable for collecting code-mixed data, the informal and inconsistent nature of the content—characterized by acronyms, emojis, and spelling variations—adds complexity to dataset curation [4]. Moreover, linguistic diversity within such datasets is often limited to specific language pairs, further restricting their practical utility.

Despite the growing interest in hate speech detection, research focusing on code-mixed datasets remains limited. Transformer-based methods, particularly BERT-based models, have demonstrated superior accuracy in hate speech detection [5]. This study aims to address this gap by exploring the application of the XLM-RoBERTa model for detecting hate speech in code-mixed contexts, specifically in Indonesian and English. By leveraging XLM-RoBERTa's multilingual capabilities, this approach seeks to enhance the detection of harmful content in environments where language boundaries are increasingly blurred. The model's effectiveness will be evaluated using accuracy and F1-score metrics, contributing valuable insights to the development of robust hate speech detection systems in multilingual settings.

The remainder of this paper is organized as follows: Chapter 2 surveys related works; Chapter 3 presents the preliminaries; Chapter 4 outlines the methods; Chapter 5 discusses the results, contributions, and limitations; and Chapter 6 concludes the study with suggestions for future research.

2 Related works

XLM-RoBERTa has been applied to detect hate speech in code-mixed datasets. Tita et al. [6] compared the performance of mBERT and XLM-RoBERTa in detecting hate speech in English and French, including code-mixed contexts. The evaluation results show that XLM-RoBERTa achieves a higher macro-average score of 0.51, outperforming mBERT, which obtained 0.41, in English-French code-mixed scenarios.

Wang et al. [7] tested XLM-RoBERTa for offensive language detection in English, Turkish, Arabic, Danish, and Greek. The model achieved the average F1-scores of 0.9255, 0.8224, 0.9015, 0.8136, and 0.8392 for each language, respectively.

Suhartono et al. [8] performed a comparison study of how mBERT and XLM-RoBERTa works on classifying fake Indonesian news. The proposed model is proven to be successful with results of accuracy, precision, recall, and F1 of 0.9051, 0.9515, 0.8233, and 0.8828 respectively for the mBERT model with 10 topic words and 0.8935, 0.8818, 0.8712, and 0.8765 for the XLM-R model with 10 topic words. Table 1 presents the performance of XLM-RoBERTa in each language based on findings from the previous related works mentioned.

Table 1: XLM-RoBERTa performance across languages.

Works	Model	Dataset	Result
(2022) Suhartono et al. [8]	XLM-RoBERTa	Indonesian	The model achieves accuracy of 89.35%, precision of 88.18%, recall of 87.12%, and f1-score of 87.65%
	mBERT		The model achieves accuracy of 90.51%, precision of 95.15%, recall of 82.33%, and f1-score of 88.28%
(2021) Tita et al. [6]	XLM-RoBERTa	English-French	The model achieves macro average of 51%
		English	The model achieves macro average of 44%
		French	The model achieves macro average of 32%
	mBERT	English-French	The model achieves macro average of 41%
		English	The model achieves macro average of 71%
		French	The model achieves macro average of 66%
(2020) Wang et al. [7]	XLM-RoBERTa	English	The model achieves f1-score of 92.55%
		Turkish	The model achieves f1-score of 82.24%
		Arabic	The model achieves f1-score of 90.15%
		Danish	The model achieves f1-score of 81.36%
		Greek	The model achieves f1-score of 83.92%

Existing studies on XLM-RoBERTa and mBERT largely focus on other languages, often overlooking code-mixed contexts. Even when code-mixed scenarios are considered, the reported metric scores remain relatively low, indicating room for improvement. Notably, XLM-RoBERTa has demonstrated a strong ability to capture contextual nuances, including in code-mixed settings, particularly for high-resource languages like English. Bahasa Indonesia, also considered a high-resource language for XLM-RoBERTa, has approximately 22.704 million tokens compared to English's 55.608 million [9], making it significantly better represented than most languages. This reinforces XLM-RoBERTa's potential effectiveness in addressing code-mixed contexts, such as hate speech detection.

Since code-mixed resources are often limited or impractical to use, Large Language Models (LLMs) can also be leveraged to generate datasets for hate speech identification. Terblanche et al. [4] demonstrated the use of GPT-3.5 to generate code-mixed sentences in Afrikaans-English and Yoruba-English. Their findings showed that the generated data—written using the English alphabet and Latin script—was of high quality, with only minor grammatical errors that did not significantly affect meaning. This success suggests an opportunity to refine prompting guidelines to improve data quality, further supported by the fact that XLM-RoBERTa has been trained on a substantial amount of data in both Indonesian and English.

3 Preliminaries

XLM-RoBERTa (Cross-lingual Robustly Optimized Bidirectional Encoder Representations from Transformers Approach) is a transformer-based language model which enhances the state-of-the-art on mixedmultilingual understanding tasks through the joint pretraining large transformer models across diverse languages. This model is built upon the advancements of the RoBERTa model, an optimized version of BERT with dynamic masking, removal of *next sentence prediction* (NSP), larger mini-batches, and byte-level Byte Pair Encoding (BPE) tokenizer which relies on subword units and makes it possible to learn a subword vocabulary that can still encode any input text without introducing any unrecognizable tokens, ensuring success interpretation on new or unseen terms. The model training leveraged SPM-preprocessed text data from CommonCrawl scaled to cover 100 languages to handle diverse linguistic structures [9], [10].

The model is built on top of transformer and MLM (Masked Language Model) architecture, which excels at processing sequential data such as text by utilizing bidirectional self-attention layers [11] which helps on capturing the contextual relationship between words regardless of the language. Bidirectional pre-training mechanism allows the model to achieve state-of-the-art performance which reduces the need for various heavily-tuned task-specific architecture, and also predict or learn bidirectional context by predicting missing words to better understand hidden or implied subtle relationships and context. XLM-RoBERTa variant specifically applies *subword* tokenization directly on raw text data and utilizing sample batches from diverse languages using the same sampling distribution. This model additionally does not implement language embeddings which results in improved performance when dealing with code-mixed contexts, enabling it to learn complex patterns and structures in multiple languages, especially for mixed-code hate speech detection [12].

Figure 1 illustrates transformer encoder architecture while Figure 2 represents the multilingual MLM architecture. Both figures provide an overview of XLM-RoBERTa architecture, which utilizes transformer encoder model [9], [11]. In this model, inputs are preprocessed through the MLM which is specifically pre-trained for transformer encoder. Afterwards, the MLM predicts the original content of input tokens based on the remaining bidirectional contexts from randomly masked portions of the input [13].

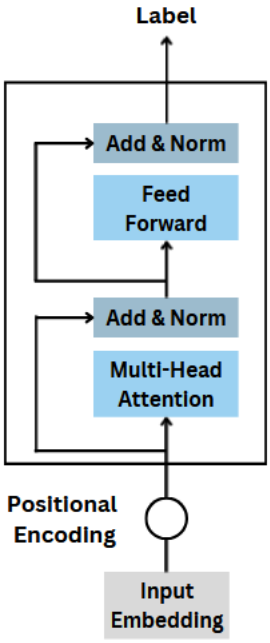


Figure 1: Transformer encoder architecture. Source: [14].

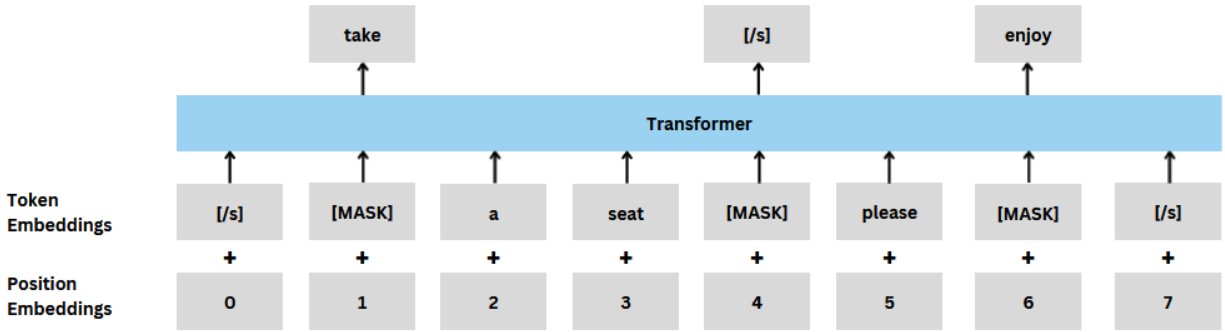


Figure 2: MLM architecture on XLM-RoBERTa. Source: [15].

4 Methods

4.1 Data collection and generation

Data collection utilizes variants of the GPT-4 large language model (LLM), including GPT-4o and GPT-4o mini, which are used interchangeably across a total of 600 prompt executions. The prompt is carefully designed with multiple detailed examples of real-life sentence scenarios to enhance contextual relevance. To ensure diversity, each execution features unique topics such as politics, religion, sports, gaming, and other common aspects of daily life. Additionally, the model's memory is periodically reset after several prompts to minimize redundancy and prevent duplicate outputs. After data generation, the dataset undergoes filtering to remove any duplicate content. Figure 3 illustrates the prompt used for data generation.

```
You are tasked with refining a language model to identify hate speech in mixed-code content. For this purpose, please generate
25 sentences that exemplify hate speech in English and 25 sentences that represent non-hate speech, also in English.

Guidelines:
Ensure all sentences are solely in English, without any mixing of other languages. Each sentence should pertain to a specific
topic or context, which will be provided below.

Context Topic: [Topic]
You can use the context topic as a to generate the dataset OR simulate a sentence on a discussion for that context topic.

Examples of Unsuitable Sentences:
    "Why don't you double-check your facts before making a decision?" (1)
    "The company should prioritize innovation!" (1)
    "You need to improve your reliability when meeting deadlines!" (1)
    "It seems like you expect a promotion without demonstrating enough effort!" (1)
    "What's taking customer service so long to improve?" (1)

Examples of Suitable Sentences:
    "Defund the police and kick out all these damned immigrants and their mayor!" (1)
    "Never buy this piece of shit; it breaks down constantly and is infuriating!" (1)
    "This product is terrible; it keeps breaking. A total waste of money!" (1)
    "I don't want to hear about those black bastards; just lynch them already!" (1)
    "This product is trash; it overheats and could potentially burn your house!" (1)
    "What a total ripoff! This product falls apart every time. Such a horrible experience." (1)
    "This vehicle is always having problems, and it's not worth the trouble!" (1)
    "The advertisement is so fucking misleading; the product doesn't perform as promised." (1)

Examples of Non-Hate Speech Sentences:
    "This product is fantastic for its quality and performance; I highly recommend it!" (0)
    "It's durable and works perfectly every time." (0)
    "I have no complaints with this product; it consistently delivers great results." (0)
    "I'm completely satisfied with this product; it fulfills all its promises!" (0)
    "The quality is outstanding, and it's my go-to choice every time." (0)
    "I'm impressed with its performance; definitely worth the price!" (0)
    "This item exceeded all my expectations, even if it is a bit pricey!" (0)
    "If you're searching for something reliable, this product is ideal and long-lasting." (0)

Additional Instructions:
    For the hate speech category (labeled 1), use strong language where appropriate, including words like "fuck," "bitch,"
    "idiot," "whore," and other offensive terms targeting specific groups.
    Ensure all non-hate speech sentences (labeled 0) are entirely in English as well.
    Avoid reusing the examples provided above in your generated sentences.
    When crafting hate speech sentences, maintain a clear distinction between hate speech and expressions of negative
    sentiment. Hate speech involves derogatory, discriminatory, or violent language aimed at groups based on attributes like
    race, gender, ethnicity, religion, or sexual orientation.
    Randomize the order of the hate speech (1) and non-hate speech (0) sentences in the final output.

Please generate 25 unique sentences for each category (hate speech and non-hate speech). Ensure that each sentence is distinct
from previous responses and that you vary the vocabulary and structure for greater diversity.
MAKE SURE TO NOT ADD any useless suffix at the end of the sentence (example : (0) or (1), etc)
MAKE SURE to save the output file initially with the name dataset_english_GPT_691.csv. Use the naming convention
dataset_english_GPT_692.csv and so on afterwards for subsequent files.
```

Figure 3: Prompt for data generation.

The Indonesian dataset consists of 9,783 entries, with 4,911 (50.2%) categorized as hate speech and 4,872 (49.8%) as non-hate speech. The English dataset includes 9,968 entries, with 4,973 (49.9%) labeled as hate speech and 4,995 (50.1%) as non-hate speech. Additionally, the code-mixed dataset contains 7,835 entries, of which 3,951 (50.4%) are designated as hate speech and 3,884 (49.6%) as non-hate speech. These figures represent the final number of data entries after the removal of duplicates to ensure data quality and consistency.

The code-mixed dataset could be retrieved from specific platforms, such as social media, which reflect real-world scenarios. Social media platforms are invaluable for gathering code-mixed data due to the extensive and diverse linguistic expressions they contain. However, the informal nature of the content, often characterized by acronyms, emojis, and spelling mistakes, presents significant challenges for effective processing [4].

Using generated data, such as datasets created with GPT, in hate speech detection introduces important ethical implications and potential biases that must be addressed. First, GPT-based models may inadvertently reproduce biases present in their training data, leading to the propagation of stereotypes or inequities in the generated dataset. This can result in a biased hate speech detection model that disproportionately misclassifies or overlooks hate speech targeting certain groups, potentially reinforcing societal prejudices. However, this bias may be intentionally leveraged to ensure the model detects specific patterns, aligning with the main objective of hate speech identification. Second, the synthetic nature of the data might lack the nuanced and context-specific complexity of real-world hate speech, which could reduce the model's effectiveness in handling real-world scenarios. Therefore, rigorous manual evaluation, curation, and refinement of the generated dataset—such as self-

embedding hate keywords in certain datasets—are essential to ensure its quality, fairness, and relevance.

4.2 Data preprocessing

Figure 4 illustrates the complete sequence of preprocessing steps involved in preparing the data for model input. This process includes cleansing the raw text, tokenizing it into distinct tokens, and applying padding and truncation based on the 95th percentile of sequence lengths in the dataset. Finally, masking is applied to differentiate between empty and non-empty tokens, aiding the model in processing the text effectively.

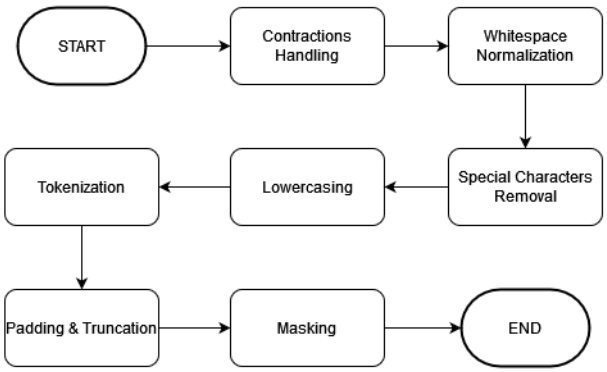


Figure 4: Data preprocessing steps.

Figure 4 outlines the data preprocessing pipeline for training XLM-RoBERTa in the context of hate speech detection for Indonesian-English code-mixed text. The process begins with contraction handling, where common abbreviations and informal contractions (e.g., "don't" → "do not", "*gak*" → "*tidak*") are expanded to their standard forms, ensuring better text representation. Next, whitespace normalization standardizes spaces and removes unnecessary gaps, followed by special character removal to eliminate symbols, emojis, or non-text elements that do not contribute to meaning.

The text is then lowercased to ensure consistency, especially for models that are case-sensitive. Tokenization follows, where the text is split into subword units using SentencePiece, the tokenization method employed by XLM-RoBERTa, enabling better handling of multilingual and code-mixed text. Subsequently, padding and truncation are applied to standardize sequence lengths, preventing excessive memory usage and maintaining uniform input dimensions. Masking is performed as part of the Masked Language Model (MLM) objective, where certain tokens are randomly replaced with a mask token, helping the model learn contextual relationships. Finally, the processed data is fed into the model for training and evaluation.

This preprocessing pipeline ensures that XLM-RoBERTa effectively learns language patterns in code-mixed Indonesian-English text, improving its ability to detect hate speech accurately while handling the linguistic variations commonly present in informal online discourse. By standardizing the text and preparing it for optimal processing, each step contributes to the model's enhanced understanding of both the individual languages involved and their intermingling in code-mixed contexts. As a result, the model is better equipped to identify subtle patterns and nuances in the text, making it more effective for real-world hate speech detection tasks.

4.3 Hyperparameter tuning

The pre-trained XLM-RoBERTa-Base model undergoes fine-tuning with variations in key hyperparameters to optimize performance by implementing grid search. Specifically, the batch size is tested with values of 16 and 32, while the learning rate is adjusted within the range of 1e-5 to 5e-5. With 10 parameter combinations (2 batch sizes × 5 learning rates), the grid search achieved efficient coverage of the hyperparameter space, providing an effective yet computationally feasible approach to optimizing the model for the task.

The selection of hyperparameter ranges for fine-tuning the XLM-RoBERTa-Base model is guided by the need to balance computational efficiency and performance optimization. Batch sizes of 16 and 32 are chosen to examine the trade-off between gradient update precision and memory consumption. Smaller batch sizes provide more precise updates but require a greater number of iterations, whereas larger batch sizes accelerate training at the potential cost of less precise convergence. The learning rate range of 1e-5 to 5e-5 is selected based on best practices for transformer-based models, ensuring stable convergence at lower rates while allowing faster training at higher rates without overshooting the optimal minima. This approach strikes a balance between preserving pre-trained weights and adapting to the hate speech detection task, where capturing subtle linguistic nuances is crucial.

An early stopping mechanism with checkpointing is implemented to prevent overfitting and enhance generalization. This mechanism monitors validation loss over five epochs and restores the best-performing weights, ensuring optimal model performance without excessive computational overhead. The number of training epochs is also tuned through an adaptive monitoring process. Training begins with a minimum of five epochs, and if validation loss does not decrease after five additional epochs beyond the current best checkpoint, the model reverts to the last optimal weights. If validation loss improves, the new epoch is designated as the best checkpoint, and monitoring restarts. This tuning strategy achieves an optimal balance between convergence speed and model stability, ensuring efficient adaptation of the XLM-RoBERTa-Base model to the task of mixed-code hate speech detection.

The chosen hyperparameter combinations were validated by calculating performance metrics, including F1-score and accuracy, for each configuration (learning rates ranging from 1e-5 to 5e-5 and batch sizes of 16 and 32) on the validation set. These metrics were evaluated across multiple training runs to account for variability introduced by random initialization and data splits.

Table 2 demonstrates that the optimal model configuration is achieved with a learning rate of 2e-5, a batch size of 16, and at epoch 8, resulting in a validation loss of 0.0355. This configuration also yields high

validation accuracy of 99.23%, validation precision of 99.23%, validation recall of 99.23%, and validation F1 score of 99.23%, indicating strong model performance across multiple metrics. Minimizing validation loss is crucial as it reflects the model's capacity to generalize effectively to unseen data, thus mitigating overfitting and ensuring robust performance beyond the training dataset.

Generally, models with lower validation loss are associated with better generalization, making it a reliable criterion for model selection. Additionally, the results indicate that a lower learning rate of 1e-5, 2e-5, and 3e-5 outperforms higher rates such as 4e-5 and 5e-5, where smaller batch sizes showing a slight advantage in this context, as seen in the top-performing models. This outcome is attributed to the fact that lower learning rates enable gradual and precise convergence, minimizing the risk of overshooting optimal solution which is a common issue with higher learning rates. High learning rates can lead to unstable training dynamics, as evidenced by increased losses for both training and validation at learning rates of 4e-5 and 5e-5.

In conclusion, a low learning rate contributes to stable training, reducing both training and validation losses, thereby enhancing the model's overall performance and generalization capability. Table 3 shows training and validation progress over epochs for the chosen model.

Table 2: Hyperparameter tuning results ranked by validation loss.

Learning Rate	Batch Size	Best Epoch	Train Loss	Validation Loss	Train Accuracy	Validation Accuracy	Train Precision	Validation Precision	Training Recall	Validation Recall	Training F1 Score	Validation F1 Score
2e-5	16	8	0.0547	0.0355	0.9926	0.9923	0.9916	0.9923	0.9925	0.9923	0.9920	0.9923
1e-5	16	6	0.0549	0.0360	0.9909	0.9928	0.9917	0.9943	0.9883	0.9918	0.9900	0.9930
1e-5	32	18	0.0395	0.0552	0.9933	0.9919	0.9938	0.9923	0.9922	0.9911	0.9910	0.9917
2e-5	32	5	0.1708	0.0830	0.9716	0.9882	0.9710	0.9878	0.9596	0.9886	0.9653	0.9882
3e-5	16	5	0.2033	0.1236	0.9588	0.9852	0.9624	0.9849	0.9390	0.9857	0.9506	0.9853
3e-5	32	9	0.2632	0.1561	0.9593	0.9556	0.9594	0.9602	0.9592	0.9502	0.9593	0.9552
4e-5	32	5	0.7210	0.6838	0.5677	0.5093	0.5467	0.5093	0.5274	0.5093	0.5369	0.5093
4e-5	16	7	0.7067	0.6870	0.5029	0.6393	0.5008	0.5335	0.4827	0.9123	0.4916	0.6733
5e-5	16	11	1.3632	0.6955	0.5007	0.5093	0.5013	0.500	0.5458	1.000	0.5226	0.6667
5e-5	32	5	4.1957	4.2744	0.5016	0.4907	0.5053	0.4907	0.2449	0.4907	0.3299	0.4907

Table 3: Performance metrics over epochs.

Epoch	Train Loss	Validation Loss	Train Accuracy	Validation Accuracy	Train Precision	Validation Precision	Training Recall	Validation Recall	Training F1 Score	Validation F1 Score
1	0.3068	0.1222	0.9464	0.9867	0.9490	0.9870	0.8871	0.9869	0.9170	0.9869
2	0.1364	0.0619	0.9785	0.9902	0.9802	0.9921	0.9765	0.9891	0.9783	0.9906
3	0.0858	0.0704	0.9846	0.9913	0.9849	0.9926	0.9827	0.9891	0.9838	0.9908
4	0.0921	0.0972	0.9876	0.9899	0.9874	0.9889	0.9866	0.9909	0.9870	0.9899
5	0.0669	0.0673	0.9898	0.9913	0.9910	0.9934	0.9883	0.9902	0.9896	0.9918
6	0.1053	0.0575	0.9842	0.9929	0.9851	0.9929	0.9809	0.9929	0.9830	0.9929
7	0.0798	0.0444	0.9887	0.9929	0.9877	0.9933	0.9880	0.9921	0.9878	0.9927
8	0.0547	0.0356	0.9926	0.9923	0.9916	0.9923	0.9925	0.9923	0.9920	0.9923
9	0.0504	0.0862	0.9867	0.9837	0.9860	0.9835	0.9895	0.9842	0.9877	0.9838
10	0.0785	0.0590	0.9876	0.9931	0.9859	0.9923	0.9882	0.9938	0.9870	0.9930

11	0.0490	0.0541	0.9915	0.9938	0.9913	0.9926	0.9907	0.9943	0.9910	0.9934
12	0.0886	0.1376	0.9892	0.9650	0.9908	0.9665	0.9815	0.9625	0.9861	0.9645
13	0.0764	0.0798	0.9708	0.9867	0.9781	0.9874	0.9744	0.9860	0.9762	0.9867

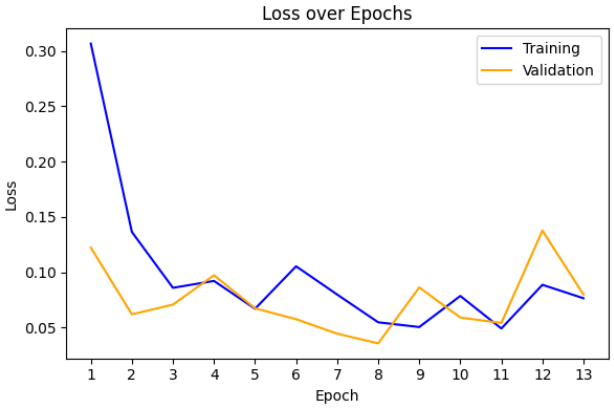


Figure 5: Training & validation loss over epochs.

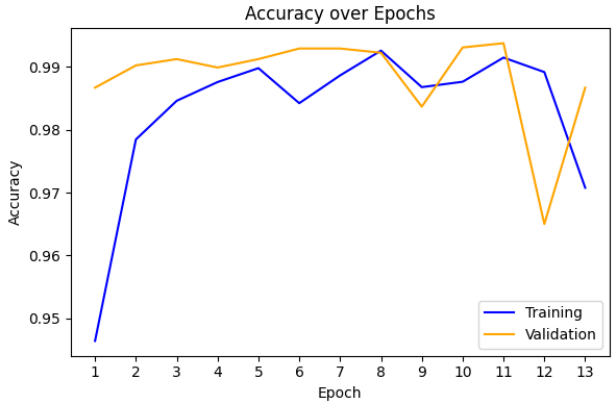


Figure 6: Training & validation accuracy over epochs.

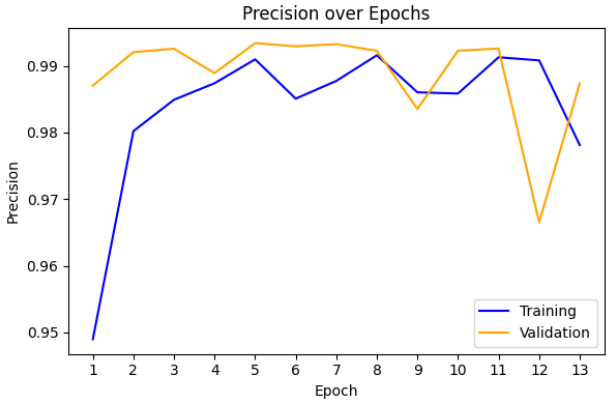


Figure 7: Training & validation precision over epochs.

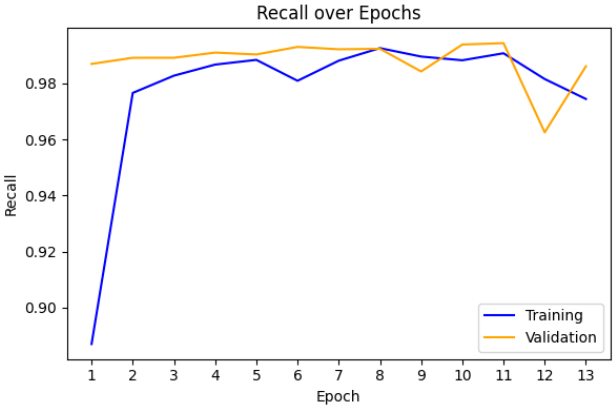


Figure 8: Training & validation recall over epochs.

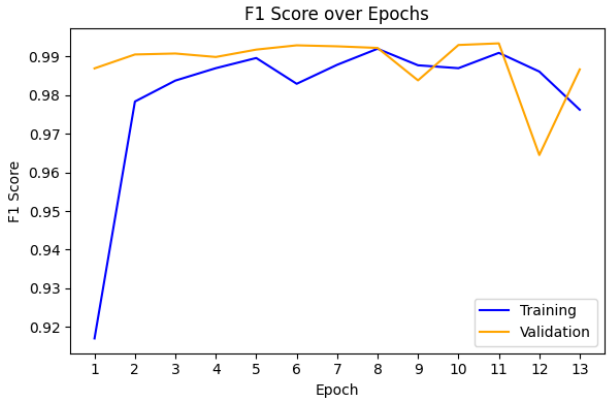


Figure 9: Training & validation f1-score over epochs.

Figures 5–9 illustrate the training and validation performance trends over 13 epochs across multiple evaluation metrics. Figure 5 shows a sharp decline in training loss during the initial epochs, followed by stabilization with minor fluctuations, while validation loss remains consistently lower, with a slight increase around epoch 12, suggesting minor variations in generalization. Figure 6 highlights the accuracy progression, where both training and validation accuracy rapidly exceed 0.98 and remain stable, indicating strong generalization.

Figure 7 presents precision trends, with training precision surpassing 0.98 early and stabilizing near 0.99, while validation precision remains consistently high with slight fluctuations. Figure 8 illustrates recall performance, where both training and validation recall remain around 0.99, with a minor dip at epoch 12, reinforcing the model's ability to minimize false negatives. Figure 9 depicts F1-score trends, showing a rapid increase followed by stabilization near 0.99 for both training and validation sets, ensuring a balanced precision-recall trade-off. Overall, these results indicate that the model effectively generalizes while maintaining high performance across all key metrics, with only minor variations observed in the later epochs.

4.4 Further evaluation

The validation of the chosen hyperparameter combinations involved calculating performance metrics, such as F1-score and accuracy, for each configuration (learning rates ranging from 1e-5 to 5e-5 and batch sizes of 16 and 32) on the validation set. These metrics were evaluated across multiple training runs to account for variability introduced by random initialization and data splits.

5 Results and analysis

Table 2 shows that lower learning rates, such as 2e-5 and 1e-5 perform better in tasks like hate speech detection involving complex, code-mixed datasets due to their ability to ensure stable convergence and precise weight updates. These rates allow the model to better capture subtle linguistic patterns and reduce the risk of overfitting, as evidenced by lower validation loss and higher validation metrics (accuracy, precision, recall, and F1-score) compared to higher learning rates. Furthermore, the gradual optimization enabled by lower learning rates allows the model to benefit from longer training durations (e.g., 8 or 18 epochs), refining its performance without plateauing or diverging. This makes lower learning rates particularly effective for nuanced tasks requiring high precision and generalization.

Lower batch sizes, such as 16, tend to perform better in the top 4 models because they allow the model to capture more detailed gradients during training, which is particularly advantageous in tasks like hate speech detection involving nuanced and diverse data. With smaller batches, the model processes fewer samples at a time, enabling it to better adapt to subtle patterns in the data, such as code-mixed contexts or cultural nuances. This precision helps reduce the risk of oversmoothing the gradients, which can occur with larger batch sizes like 32, where updates may generalize too broadly and miss finer details. Additionally, smaller batch sizes improve generalization, as the model sees a wider range of gradient variations during training, which is reflected in the lower validation loss and consistently higher metrics (accuracy, precision, recall, and F1-score) observed for batch size 16 in the top-performing models. These benefits make lower batch sizes more suitable for fine-tuned hate speech detection tasks where context sensitivity is critical.

The model was evaluated on a dedicated test set, comprising 10% portion of the whole dataset. The model achieved a near-perfect accuracy of 99.60%, false positive rate of 0.67% and false negative rate of 0.136%, reflecting high rate of correct predictions. With a precision of 99.60%, the model effectively minimized false positive predictions, while a recall of 99.60% exhibited the model capability on reducing false negatives. Additionally, the F1-score of 99.60% further demonstrates the model's balance between precision and recall which indicates excellent performance on both types of classification errors. Figure 10 presents the confusion matrix generated from the test set inference results.

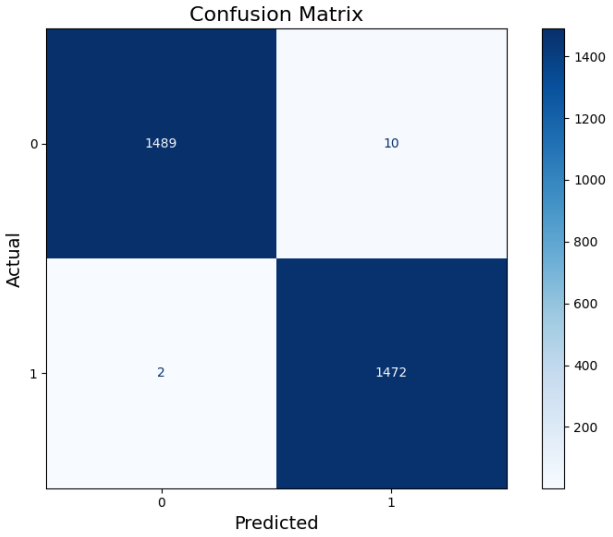


Figure 10: Primary test set confusion matrix.

Upon testing the supplementary test set, the model exhibited excellent performance, achieving 91.00% accuracy, 92.40% precision, 91.00% recall, and an F1-score of 90.94%. The false positive rate is approximately 17.6%, while the false negative rate is 0%. These results indicates that the trained model effectively manages outlier cases that reflect real-life scenarios, where sentence structures may be unpredictable and vary widely. Figure 11 shows the confusion matrix from the supplementary test set.

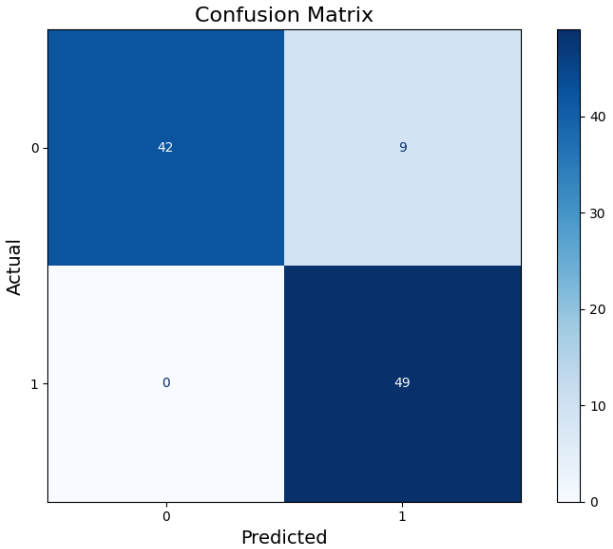


Figure 11: Supplementary test set confusion matrix.

Table 4 shows the false predictions from the model. The observed false positives in the model can be attributed to several factors that affect the model's ability to accurately classify sentiment. Firstly, the abbreviation of words that sometimes include number which are common in informal communication (e.g., "klo" for "kalau" and "bisa2" for "bisa-bisa"), may not consistently align with the model's learned patterns, leading to misinterpretations.

Table 4: Primary test set false prediction samples

Text	Type
"this health advice in the comments is really practical, informasi yang berguna!	False Positive
"organizing a book club in a cozy cafe promotes community, mempromosikan komunitas!"	False Positive
"golfers di sini terlalu angkuh, mental miskin!"	False Positive
"emission control technology is state-of-the-art, teknologi kontrol emisi sangat canggih."	False Positive
"recommending the park's features to friends promotes information_technology usage, mendorong penggunaannya!"	False Positive
"menjelajahi stasiun luar angkasa itu feels like angstrom real adventure, sungguh menakjubkan!"	False Positive
"the noise from aircraft itu unbearable, they should ban them from cities!"	False Positive
"every time a pet is brought here, it's a gamble with their health, berjudi dengan kesehatan hewan!"	False Positive
"great command! your tactics are impressive, komando yang hebat! taktikmu mengesankan!"	False Positive
"the observatory fails to engage with the global scientific community, gagal berinteraksi dengan komunitas ilmiah global!"	False Positive
"why are you indians still here if you don't support the cause? kenapa masih di sini?"	False Negative
"exploring this aquarium is never educational and fun, dasar jewish!"	False Negative

Furthermore, the presence of unknown keywords and emojis, such as internet slang or culturally specific terms (e.g., "wkwk"), which the model was not trained on, may contribute to confusion during classification. Lastly, the ambiguity of certain words—like "sumpah" (swear), "kelewatan" (gone too far), and phrases like "I cannot stand"—which may carry negative connotations but do not explicitly convey hate speech, poses challenges for the model. These are the reasons that likely caused slight performance evaluation drop on the supplementary test. These components also highlight the complexities involved in accurately interpreting code-mixed sentiment, underscoring the need for improved training data and contextual understanding by wider keyword enrichment including outlier vocabularies.

Table 5: Supplementary test set false prediction samples.

Text	Type
"Nggak tau sih soalnya lucu juga ya klo dipikir2 wkwk"	False Positive
"Aduh cape banget gw kerja sm orang sumpah"	False Positive
"Kenapa ya orang-orang tu bisa2nya jahat bgt?"	False Positive
"Gak mau jadi politikus gue; tekanannya psti gede bgt! 🥹🥹*"*	False Positive
"Sumpah kesel bgt gw sama temen lo!"	False Positive
"Kamu mau gak jadi pacar aku?"	False Positive
Apakah kamu bidadari? Soalnya cantiknya kelewatan 😍😍😍😍*"*	False Positive
"I can't stand harga dri produk ini; mahal bgt! 😰*"*	False Positive
"Kerja kerasmu tidak akan sia-sia, keep going! 💪*"*	False Positive

Table 5 shows false positive predictions on the supplementary test set where the model incorrectly classified non-hateful texts as hate speech. The examples contain informal language, humor, slang, emojis, and exaggerations, which may have caused misclassification. The model likely struggles with contextual understanding, misinterpreting harmless expressions as offensive.

False positives and false negatives in hate speech detection have critical implications, both ethically and practically. False positives occur when non-hateful content is incorrectly classified as hate speech. This can suppress legitimate expression, create a chilling effect on free speech, and harm users who may feel unfairly censored or misjudged. Conversely, false negatives, where actual hate speech is not detected, allow harmful content to persist. This can perpetuate harm to targeted individuals or communities and undermine trust in the detection system. Failure to address false negatives can have serious societal impacts, such as the normalization of offensive language or inadequate protection for marginalized groups.

We apply several strategies in the hate speech detection system to mitigate the errors. First, enhancing the quality of training data is essential. This involves enriching datasets with diverse examples of internet slang, culturally specific terms, and nuanced expressions to improve the model's ability to capture contextual subtleties on real world examples. Second, employing context-aware models or fine-tuning pre-trained models like XLM-RoBERTa with additional layers designed for better contextual understanding can significantly improve classification accuracy. These strategies collectively enhance the robustness, fairness, and reliability of hate speech detection systems in practical applications. Table 6 shows common errors in hate speech detection along with examples and potential causes

Table 6: Common errors in hate speech detection.

Category	Example	Potential Cause
Lexical Ambiguity	*"sumpah"* (swear)	The model struggles to differentiate context-specific meanings without additional contextual cues.
Cultural Nuances	*"wkwk"* (Indonesia's slang for laughter) *"ashiap"* (Indonesia's slang for saying *"yes"*)	Limited representation of culturally specific terms in the training dataset, leading to misclassification.
Implicit Hate Speech	*"Gak mungkin banget sih someone from ***** bisa sepinter itu"*	The model struggles with identifying hate speech when explicit offensive keywords are absent.
False Positive	*"I don't agree, tapi nggak apa-apa sih."*	Neutral or positive statements incorrectly classified as hate speech due to negative sentiment keywords.
False Negative	*"Go back to your own country!"*	Inadequate coverage of explicit hate speech examples or poor generalization from training data.
Out-of-Vocabulary Terms	*"Sumpah itu tadi orang noob banget." "Bruh konser tadi, absolutely lit sih bro" "Cmon fam!"*	The model's tokenizer or vocabulary does not include these terms, leading to incomplete representation.

We assessed the generalization capability of our fine-tuned XLM-RoBERTa model for mixed-code hate speech detection using a post-hoc 5-fold stratified cross-validation strategy, with the best-performing model fine-tuned at a learning rate of 2e-5 and a batch size of 16. The dataset was divided into five stratified folds, ensuring balanced representation of hate speech and non-hate speech instances across each split. For each fold, the model was evaluated on the held-out validation set without further training, utilizing the Hugging Face Transformers library. The model achieved an average evaluation loss of 0.0342, with minimal variation across folds (ranging from 0.0322 to 0.0356), indicating strong and consistent performance.

We quantified the uncertainty in the model's generalization performance by computing a 95% confidence interval (CI) for the mean evaluation loss. This CI provides a range in which the true mean loss is expected to fall with 95% confidence. Applying the standard normal approximation method, we obtained a 95% CI of (0.0327, 0.0356). The narrow interval suggests that the model's performance is statistically stable, with low variance across different validation sets. The small margin of error highlights the model's high reliability, confirming its strong generalization ability and robustness to minor variations in the dataset. These results further emphasize the effectiveness of the fine-tuned XLM-RoBERTa model for mixed-code hate speech detection.

The results indicate that XLM-RoBERTa achieves exceptional performance when properly fine-tuned and trained with a substantial number of tokens. The model consistently outperforms those used in related studies across multiple multilingual hate speech detection tasks. As shown in Table 1, prior implementations of XLM-RoBERTa and its variations yielded F1-scores ranging from 32% to 92.55%, depending on the dataset and language pair. In contrast, the proposed model achieves an F1-score of 99.60% on the primary test set and 90.94% on the supplementary test set. Furthermore, while earlier studies reported relatively low precision and recall, the model maintains high precision (99.60% and 92.40%) and recall (99.60% and 91.00%), ensuring balanced classification performance. Error analysis reveals that the model effectively minimizes false negatives, with a false negative rate of 0.136% on the primary test set and 0% on the supplementary test set. Given these results, the model not only surpasses previous approaches in overall performance but also demonstrates robustness in handling linguistic variations and outlier cases, making it highly effective for mixed-code hate speech detection. Future research on multilingual hate speech detection could further benefit from incorporating additional training data tailored to specific language requirements.

6 Conclusion

This study demonstrates the effectiveness of XLM-RoBERTa in detecting hate speech within code-mixed texts, particularly in Indonesian-English code-mixed contexts. The model achieved a high level of accuracy (99.6%) on the primary test set and maintained strong generalization across realistic supplementary data on 91% accuracy, reflecting its robustness in handling varied linguistic inputs. These results highlight the importance of multilingual adaptability in hate speech detection, particularly for complex online environments where language boundaries are fluid. Future research could enhance these outcomes by incorporating additional real-world linguistic variations and expanding to other language pairs, contributing to safer and more inclusive digital spaces on broader language scopes.

Acknowledgement

The authors acknowledge the support of Universitas Multimedia Nusantara for this study and appreciate the constructive feedback provided by the reviewers and

editors, which significantly contributed to improving the quality of this paper.

References

[1] J. B. Walther, "Social media and online hate," *Curr. Opin. Psychol.*, vol. 45, no. January, 2022, doi: 10.1016/j.copsyc.2021.12.010.

[2] K. Sreelakshmi, B. Premjith, and K. P. Soman, "Detection of Hate Speech Text in Hindi-English Code-mixed Data," *Procedia Comput. Sci.*, vol. 171, no. 2019, pp. 737–744, 2020, doi: 10.1016/j.procs.2020.04.080.

[3] A. Wicaksana, K. Sorensen, and F. Dinarta, "Enhancing Hate Speech Detection in Mixed-Language Texts: A Comparative Study of BLOOM and XLM-RoBERTa Models," in *2025 17th International Conference on Computer and Automation Engineering (ICCAE)*, IEEE, 2025.

[4] M. Terblanche, K. Olaleye, and V. Marivate, "Prompting Towards Alleviating Code-Switched Data Scarcity in Under-Resourced Languages with GPT as a Pivot," *3rd Annu. Meet. ELRA-ISCA Spec. Interes. Gr. Under-Resourced Lang. SIGUL 2024 Lr. 2024 - Work. Proc.*, pp. 272–282, 2024.

[5] J. S. Malik, H. Qiao, G. Pang, and A. van den Hengel, "Deep Learning for Hate Speech Detection: A Comparative Study," pp. 1–18, 2022, [Online]. Available: http://arxiv.org/abs/2202.09517

[6] T. Tita, Q. Mary, and A. Zubiaga, "Cross-lingual Hate Speech Detection using Transformer Models", doi: 10.48550/arXiv.2111.00981.

[7] S. Wang, J. Liu, X. Ouyang, and Y. Sun, "Galileo at SemEval-2020 Task 12: Multi-lingual Learning for Offensive Language Identification using Pre-trained Language Models," *14th Int. Work. Semant. Eval. SemEval 2020 - co-located 28th Int. Conf. Comput. Linguist. COLING 2020, Proc.*, pp. 1448–1455, 2020, doi: 10.18653/v1/2020.semeval-1.189.

[8] L. B. Hutama and D. Suhartono, "Indonesian Hoax News Classification with Multilingual Transformer Model and BERTopic," *Inform.*, vol. 46, no. 8, pp. 81–90, 2022, doi: 10.31449/inf.v46i8.4336.

[9] A. Conneau *et al.*, "Unsupervised Cross-lingual Representation Learning at Scale." [Online]. Available: https://github.com/facebookresearch/cc

[10] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," Jul. 2019, [Online]. Available: http://arxiv.org/abs/1907.11692

[11] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 2019.

[12] A. Conneau *et al.*, "Unsupervised Cross-lingual Representation Learning at Scale," Nov. 2019, [Online]. Available: http://arxiv.org/abs/1911.02116

[13] Y. Meng *et al.*, "Representation Deficiency in Masked Language Modeling," Feb. 2023, [Online]. Available: http://arxiv.org/abs/2302.02060

[14] X. Amatriain, A. Sankar, J. Bing, P. K. Bodigutla, T. J. Hazen, and M. Kazi, "Transformer models: an introduction and catalog," Feb. 2023, [Online]. Available: http://arxiv.org/abs/2302.07730

[15] G. Lample and A. Conneau, "Cross-lingual Language Model Pretraining," Jan. 2019, [Online]. Available: http://arxiv.org/abs/1901.07291

Robustness Prediction of Complex Networks Based on CNN Improved by Graph Representation Learning Operators

Junhua Li[1], Changxin Xi[2*], Yun Ke[3]
[1]School of Mathematics and Computer Science, Hanjiang Normal University, Shiyan, 442000, China
[2]School of Mathematics and Physics, Jingchu University of Technology, Jingmen, 448000, China
[3]Department of Network and New Media, Wuhan College, Wuhan, 430212, China
E-mail: rachel20240816@163.com, changxinxi@163.com, keyun9405@163.com
*Corresponding author

Keywords: deep learning, graph representation learning, CNN, complex network, robustness prediction

Received: October 12, 2024

Network robustness prediction evaluates the stability and reliability of network systems. A complex network robustness prediction method based on graph representation learning and improved convolutional neural network is proposed to address the low computational efficiency and insufficient prediction accuracy in network robustness. This method introduces prior knowledge of network topology and uses adjacency matrices to extract features of complex networks to improve the efficiency and accuracy of robust prediction. Aiming at the low prediction accuracy, a low frame rate convolutional neural storage area network algorithm based on icon learning is proposed to improve the prediction accuracy and generalization of complex network robustness. The results showed that the proposed algorithm reduced the prediction errors of undirected network robustness by 10.94%, 23.41%, and 13.86% under random attacks, and reduced the prediction errors of weighted network robustness by 0.0041, 0.0043, and 0.0105, respectively. In the robustness prediction of scale-free network and Q-recovery network, the prediction errors of the proposed algorithm were 0.1843 and 0.0278, respectively, reducing by 47.76% and 22.90%, respectively. In the four real networks, including Movie Lens-user, Grid Yeast, C-Elegance, and Polbooks, the connectivity robustness prediction error values of the storage area low frame rate-convolutional neural network algorithm were 0.0906, 0.1106, 0.0715, and 0.1052, respectively, and the controllability robustness prediction error values were 0.5155, 0.1882, 0.0458, and 0.1456, respectively, all of which were superior to existing methods. The proposed algorithm has certain practical application value in the fields of network system design and optimization.

Povzetek: Predlagana metoda za napovedovanje robustnosti kompleksnih omrežij združuje globoko učenje in reprezentacijsko učenje grafov, izboljšano z algoritmom za konvolucijske nevronske mreže, kar povečuje kvaliteto in učinkovitost napovedi.

1 Introduction

With the progress of network systems, predicting the robustness of complex networks has been widely applied in network design and optimization [1]. In the fields of the Internet, neural networks, etc., when the network fails, the robustness of complex network systems can ensure that the network structure does not change [2]. Efficient and accurate robustness prediction of complex networks can maintain the performance of network systems, operating safely and stably in complex environments [3]. However, in large-scale real-world networks, the network has complex nodes, resulting in low operational efficiency and reduced prediction accuracy of network robustness [4]. Therefore, exploring high-precision prediction algorithms and technologies in large-scale network systems has become a difficult research topic [5]. Graph Neural Network (GNN) can optimize networks that learn complex graph relationships. Scholars have conducted extensive research on this topic. Ji Z et al. proposed a representation learning method for molecular graphs, which utilized self-supervised pre-training of GNNs to address the scarcity limitation of task specific labels. The average area under the characteristic curve was 2.6% higher than the baseline model, and the F1 score was improved by 7-18% for drug prediction [6]. Yang X et al. proposed a simple and efficient heterogeneous graph neural network. This method reduced complexity by eliminating excessive neighbor attention and avoiding repeated neighbor clustering during each training period. The results show that this method had a simple network structure, high prediction accuracy, and fast training speed [7]. Heiter E et al. evaluated graph representation learning and graph layout methods in visualization to improve the predictive performance of machine learning models. By comparing graph representation learning and graph layout methods based on distance metrics, the results showed that graph representation learning methods could provide better quality graph layouts [8]. Peng L et al. proposed an unsupervised graph representation learning method to address the limited generalization ability in representation learning. This method maximized

semantic and structural information to design constrained graph representation learning methods. The multi-

Table 1: Comparison of methods, dataset, and results of related works.

Method	Dataset	Results	Reference
Self-supervised pre-training method based on graph neural network	MUTAG dataset, BACE dataset	Drug prediction increases F1 value by 7-18%	Ji Z et al. [6]
Simple and efficient heterogeneous graph neural network	Ogbn-mag databases	The proposed method has a simple network structure, high prediction accuracy, and fast training speed	Yang X et al. [7]
Application of evaluation graph representation learning and graphic layout methods in visualization	Molecular graph dataset	Graph representation learning methods provide high-quality graph layouts	Heiter E et al. [8]
Unsupervised graph representation learning method	OGB dataset	Superior to current state-of-the-art methods in different downstream tasks	Peng L et al. [9]
Unsupervised graph representation learning method	PRO TENS dataset	Model performance outperforms popular methods	Mo Y et al. [10]
A prior knowledge routing transfer reinforcement learning algorithm based on generative adversarial networks	Network Topology Dataset	Higher training efficiency across different topologies and network state distributions	Dong T et al. [12]
Inductive reasoning method for network topology	FB15k-237 dataset, WN18RR dataset	Effective prediction of semantic correlation between relationships	Chen J et al. [13]
Learning classification method based on machine learning combined with network prior knowledge	A large number of literature datasets	Evaluating and classifying a large number of literature has demonstrated the reliability of the method	Von RL et al. [14]
Deep neural network model based on prior knowledge of network topology	ImagNete dataset, CIFAR-10 dataset	Improved the robustness of the model	Li X et al. [15]

extraction method was superior to the current more advanced methods in different downstream tasks [9]. Mo Y et al. proposed an unsupervised graph representation learning method aimed at achieving efficient and effective contrastive learning. This method reduced intra class variation by increasing the upper bound loss to achieve finite distance between positive embedding and anchor embedding. Compared with currently popular methods, this method exhibited better model performance [10].

The prior knowledge of network topology based on deep learning has also been widely applied to evaluate network model performance [11]. Dong T et al. designed a prior knowledge routing transfer reinforcement learning algorithm for generative adversarial networks, aimed at improving training efficiency. This algorithm utilized the routing domain invariant features of deep reinforcement learning. The proposed algorithm had higher training efficiency on different topologies and network state distributions [12]. Chen J et al. proposed a network topology inductive reasoning method to address the semantic relevance issue of existing inductive reasoning models. Network topology patterns were classified and learned to predict inductive links based on different related network patterns. The experiment showed that this method could effectively predict the semantic correlation between relationships [13]. Von R L et al. built a learning classification method based on machine learning combined with network prior knowledge. This method represented different prior knowledge in the learning system, including logical rules and simulation results. The method could evaluate and classify much literature, proving the reliability [14]. Li X et al. designed a deep neural network model on the basis of prior knowledge of network topology, aiming to accurately recognize objects. This model segmented objects from images, evaluated the segmentation results using predefined prior knowledge, and outputted the final prediction results. The method could improve the model robustness [15].

Overall, GNN and network topology prior knowledge on the basis of deep learning have achieved good research results in various fields. However, traditional GNN algorithms have low prediction efficiency and poor prediction performance for network robustness. This study proposes a Multiple Convolutional Neural Network-Region Proposal (MCNN-RP) algorithm based on topological priors to improve the prediction efficiency of network robustness. Aiming at the low prediction accuracy, a Storage Area Low Frame Rate-CNN (SA-LFR-CNN) algorithm is proposed to optimize the prediction accuracy of network robustness. The research innovation lies in the improvement of the traditional CNN-Region Proposal (CNN-RP) algorithm, which utilizes the superior feature learning capability to effectively predict the robustness of complex networks. Then, by combining the Storage Area Network (SAN) operator and MCNN-RP algorithm, a high-precision network robustness predictor is constructed to improve the model generalization and achieve efficient and accurate robustness prediction of complex networks.

The methods, datasets, and results comparison of the relevant research in the literature review are shown in Table 1.

2 Methods and materials

This chapter first constructs an MCNN-RP model to predict the robustness of complex networks by incorporating prior knowledge of network topology. Aiming at the poor universality of robustness predictors, an improved SA-LFR-CNN with SAN operator is constructed to enhance the predictive ability of the predictor.

2.1 Construction of a robustness prediction model for complex networks based on CNN-RP network topology prior knowledge

With the continuous advancement of deep learning, CNN has been extensively applied in network robustness prediction. This method can accurately classify target images and has excellent data processing capabilities [16]. In the processing and analysis of complex networks, the simulation efficiency of robustness is low, and the learning ability of CNN on discrete degree distribution network samples is limited, resulting in lower prediction accuracy of complex network robustness predictors [17]. Therefore, the study adds prior knowledge of multiple network topologies to the predictor. An MCNN-RP algorithm is proposed, which uses adjacency matrix to extract features of complex networks and predicts network robustness through the predictor MLP.

In the research of complex networks, robustness is usually evaluated through two measures: controllability robustness and connectivity robustness. Controllability robustness refers to the degree to which a network's controllability decreases when it is attacked or malfunctioning. Controllability refers to the ability to effectively control a network through control nodes, usually quantified by the density of control nodes. It is used to evaluate the stability of the network in control node failure. Its expression is shown in equation (1) [18-19].

$$n_D = \frac{N_D}{N} \tag{1}$$

In equation (1), n_D signifies the density of control nodes. N_D signifies the total sum of control nodes. N signifies the total nodes. This equation is used to calculate the density of control nodes in a network, reflecting the ability of the network to maintain a controllable state when under attack. The higher the density of control nodes, the stronger the controllability and robustness of the network. The calculation method for controlling the number of nodes includes structural and precise controllability calculations, as shown in equation (2).

$$\begin{cases} N_D = \max\left\{1, N - \left|E^*\right|\right\} \\ N_D = \max\left\{1, N - rank(A)\right\} \end{cases} \tag{2}$$

In equation (2), $\left|E^*\right|$ signifies the maximum number of edges that can be matched in the network. When the network is not completely matched, $N_D = N - \left|E^*\right|$ control nodes are required to complete effective control of the system. When matrix A is full rank, $N_D = 1$ control nodes are required. Otherwise, $N_D = rank(A)$ control nodes are required. The robustness of controllable sequences and edge attacks are defined as equation (3).

$$\begin{cases} n_D^N(i) = \dfrac{N_D(i)}{N-i}, i = 0,1,\ldots,N-1 \\ n_D^E(i) = \dfrac{N_D(i)}{N}, i = 0,1,\ldots,M \end{cases} \tag{3}$$

In equation (3), $n_D^N(i)$ and $n_D^E(i)$ respectively signify the robustness of controllable sequences and the robustness of edge attacks. $N_D(i)$ signifies the number of nodes in the i-th control network after the network is attacked. $N-i$ represents the number of nodes in the i-th remaining network after the network is attacked. M represents the number of times the tolerance for edge attacks reaches network edge connection. Connectivity robustness refers to the change in the size of the maximum connected branch of a network when it is under attack or failure. It is used to evaluate the network's ability to maintain connectivity in node or edge failures. This method is shown in equation (4).

$$S_D = \frac{N_{LCC}}{N} \tag{4}$$

In equation (4), S_D represents the connectivity robustness of the network. N_{LCC} represents the maximum connected branch size of the network. The calculation process for the connectivity robustness of the attack process sequence is shown in equation (5).

$$\begin{cases} S_D^N(i) = \dfrac{N_{LCC}(i)}{N-i}, i = 0,1,\ldots,N-1 \\ S_D^E(i) = \dfrac{N_{LCC}(i)}{N}, i = 0,1,\ldots,M \end{cases} \tag{5}$$

In equation (5), $S_D^N(i)$ represents the network robustness based on edge attacks. $S_D^E(i)$ represents the network robustness against attacks on nodes. $N_{LCC}(i)$ represents the size of the remaining connected branches after attacking the network. This equation quantifies the network's ability to maintain connectivity under continuous attacks. A high value indicates that the network's connections are more robust, indicating that it can maintain significant connections even after being attacked. The CNN-RP in the network robustness predictor is shown in Figure 1. In this network architecture, there are convolutional fast and regressor blocks. The target features are nonlinearly adjusted

through the convolutional layer [20]. In Figure 1, the adjacency matrix is taken as the input part, and the final robustness prediction is taken as the output part. The final connection layer in this structure is used as a predictor to regress the robustness prediction structure.

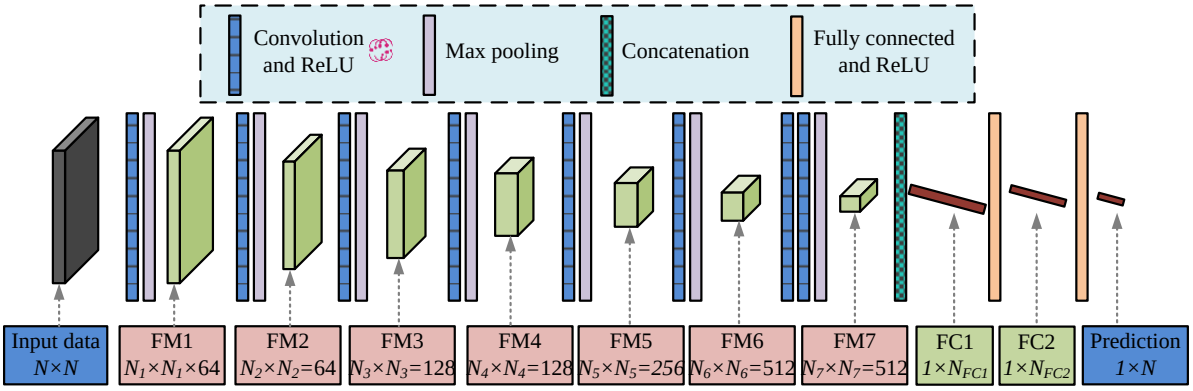


Figure 1: Structure of CNN-RP: Convolutional neural network-region proposal for network robustness prediction.

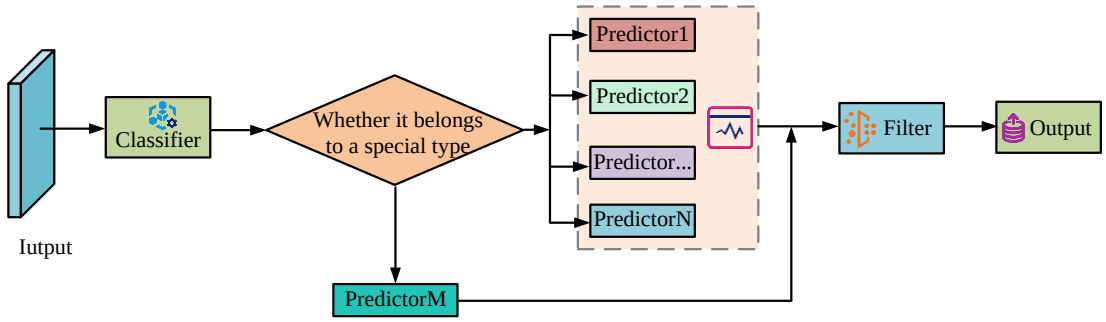


Figure 2: Workflow of MCNN-RP: Multiple convolutional neural network-region proposal for enhanced network robustness prediction.

bounds and local increasing intervals of the network output [21]. The working principle of the filter is shown in equation (6).

$$N_{LCC}(i)=\begin{cases} N-i, if N_{LCC}(i)>N-i \\ 1, if N_{LCC}(i)<1 \end{cases} \quad (6)$$

In equation (6), $N_{LCC}(i)$ represents the existence of an upper bound $N-i$. After i attacks on the network, the maximum value of its connected branches is $N-i$, and the lower bound of the network is 1. This equation is used to limit the upper limit of network robustness prediction results, ensuring that the prediction results conform to the actual characteristics of the network topology. The correction of local increment in network results is shown in equation (7).

$$N_{LCC}(k)=N_{LCC}(i)+\frac{k-i}{i-j}(N_{LCC}(i)-N_{LCC}(j)) \quad (7)$$

In equation (7), $N_{LCC}(k)$ represents a local increase in the robustness curve of the connected network. When $N_{LCC}(k)>N_{LCC}(i),(k\geq i+1)$ occurs, the network output results show a local increase, starting with element k to search for the $j(j=k+1,k+2,\ldots,N)$ -th element, making it satisfy $N_{LCC}(j)<N_{LCC}(i)$, and recording the index j. The value in section $[k+1,k+2,\ldots,j-1,j]$ is corrected. The result of the final filter correction is used as the final output result. The study uses supervised learning to train, validate, and test the CNN-RP model, with its loss function using mean square error, as shown in equation (8).

$$Loss=\frac{1}{N}\sum_{i=0}^{N-1}\left\|S_D^N(i)-\hat{S}_D^N(i)\right\| \quad (8)$$

In equation (8), *Loss* represents the loss function of the CNN-RP. N represents the scale of the network. $S_D^N(i)$ and $\hat{S}_D^N(i)$ represent the true value and predicted value of the i-th sample, respectively. The MCNN-RP algorithm consists of multiple CNN network structures. The prior knowledge of network topology can preprocess CNNs. After training the model, the network classifier is used to classify each CNN and finally predict the network robustness [22]. The MCNN-RP first initializes multiple CNN models, each with different configurations to adapt to different network topology types. Then, the adjacency matrix is preprocessed using prior knowledge of network topology. The network features are extracted through multiple CNN models. Afterwards, a classifier is used to classify the network topology types and select the most

The output of CNNs violates the prior knowledge of network topology. Therefore, a filter based on the prior knowledge is designed to modify the upper and lower

suitable CNN model. The network robustness is predicted using the selected CNN model and MLP. Finally, the prediction results are corrected based on prior knowledge of network topology. The main workflow is shown in Figure 2. Firstly, the network topology prior input to the classifier is classified, and appropriate CNNs are selected for different types of network topology for prediction.

The classification and predictor in the MCNN-RP algorithm have similar CNN structures, and their feature extraction is also the same. The last layer of the classifier is often used as the $Soft\max(f(Xi)) = e^{Xi} / \sum_j e^{Xi}$ -layer

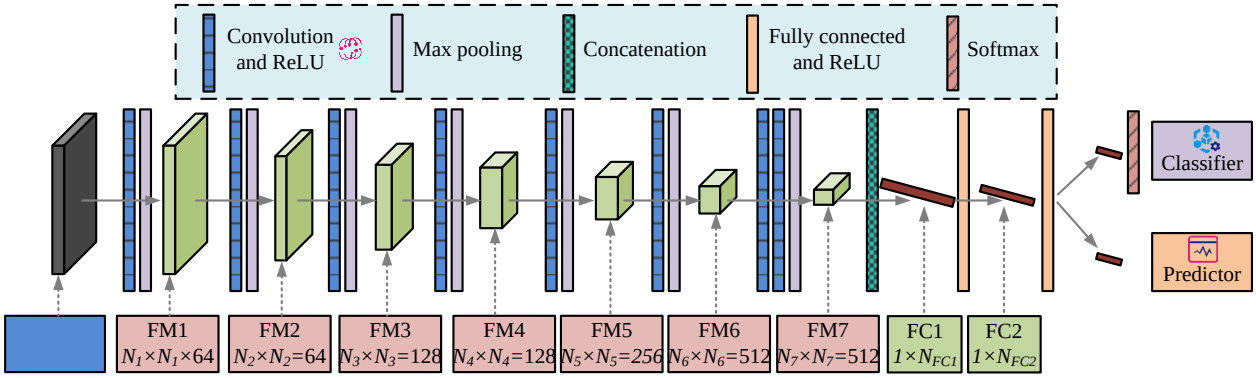


Figure 3: CNN architecture within MCNN-RP: Convolutional neural network structure for robustness prediction.

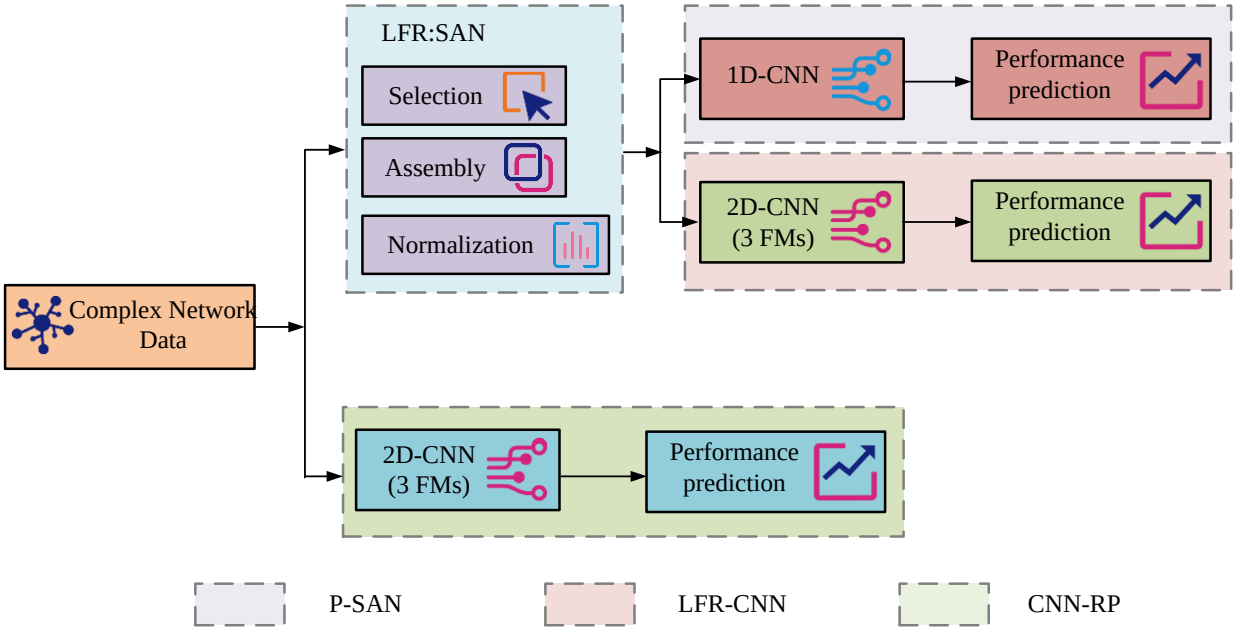


Figure 4: Structure of SA-LFR-CNN: Storage area low frame rate convolutional neural network for robustness prediction.

for classification work. The CNN structure in this algorithm is shown in Figure 3.

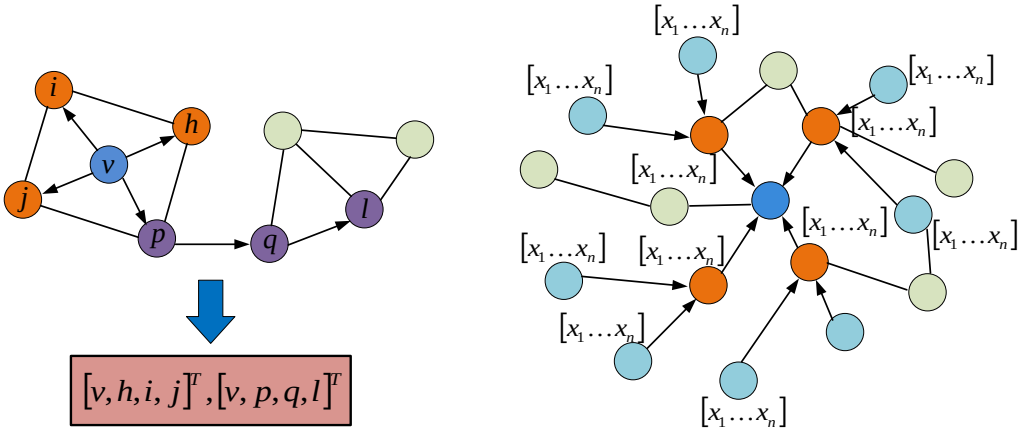
2.2 Robustness prediction of complex networks using graph representation learning algorithm based on improved SAN operator

GNN can effectively process data using graph data feature extraction operators to process graph structures [23]. CNN processing tensors limits the fixed network size, and the generalization ability of robust predictors in complex networks is poor, resulting in lower prediction accuracy [24]. Controllability robustness and connectivity robustness, as important characteristics of complex networks, are closely related to graph learning methods. Based on graph learning methods, the topological features of the network can be extracted more effectively, thereby improving the robust prediction accuracy. In addition, graph learning methods also have important application value in network optimization and design. By predicting the robustness of the network, guidance can be provided for network optimization, helping to design more stable and reliable network structures. A SA-LFR-CNN algorithm is proposed, which combines MCNN-RP and Patchy Storage Area Network (P-SAN) algorithm to improve the operators in GNN algorithm. It adds operators based on graph

representation learning to extract target graph features and then predicts the network robustness.

The SA-LFR-CNN first initializes the improved CNN model, combined with the SAN operator and graph representation learning module. The prior knowledge of network topology is used to preprocess the adjacency matrix. Then, the graph representation learning module learns the graph representation of the network, samples nodes from the graph representation, and constructs subgraphs for normalization. Subsequently, the CNN model is used to extract features from the normalized subgraph, and the extracted features are used to predict the network robustness. Finally, the prediction results are adjusted based on prior knowledge of network topology. The structure is displayed in Figure 4.

The SA-LFR-CNN algorithm represents the target features, normalizes the sampled nodes and constructed subgraphs, and then splits the network. The split graph is



(a) Diagram structure information sample diagram

(b) Aggregation graph structure and node feature information

Figure 5: Graph structure and node feature information: Schematic of graph representation learning for robustness prediction.

used as a feature node input to the algorithm, and the target features are extracted through a CNN. Finally, the network robustness is subjected to regression prediction [25]. The loss function of this algorithm first aggregates the structural information, and trains the designed loss function on the structural information. The basis of each node aggregation is related to the location of the layers. The central node trains the nodes through two aggregations to obtain feature data. The specific process of obtaining node feature information is shown in Figure 5.

In the supervised learning, the algorithm uses a designed loss function to capture the first-order and second-order distance information of the graph. The first-order distance joint probability distribution is shown in equation (9).

$$p_1(v_i, v_j) = \frac{1}{1+\exp(-\vec{u}_i^T \cdot \vec{u}_j)} \tag{9}$$

In equation (9), $p_1(v_i, v_j)$ represents the joint probability of node v_i and node v_j. $\vec{u}_i \in R^d$ represents the representation vector of the node in low latitude space. This equation is used to calculate the similarity between nodes, reflecting their proximity in low dimensional space. In this way, first-order distance information between nodes can be captured, thus better understanding the topology of the network. In the training process, the confidence in the first-order distance structure is captured directly by minimizing the objective function, as shown in equation (10).

$$O_1 = d(\hat{p}_1(\cdot,\cdot), p_1(\cdot,\cdot)) \tag{10}$$

In equation (10), O_1 represents the objective function of first-order distance. $d(\cdot,\cdot)$ represents the distance between probability distributions $\hat{p}_1$ and p_1. By calculating the square difference between the true joint probability and the predicted joint probability, and adding up all nodes and their adjacent nodes, the prediction error of the model on the first-order distance is quantified. The obtained objective function is shown in equation (11).

$$O_1 = \sum_{(i,j)\in E} \omega_{ij} \log p_1(v_i, v_j) \tag{11}$$

In equation (11), $\omega_{ij} \log p_1(v_i, v_j)$ represents the distance between the relative entropy probability distributions. Nodes play different roles in second-order distance, and the second-order distance needs to be calculated using different representation vectors. For each edge $(i,j) \in E$ in the graph, the probability of the context node is shown in equation (12).

$$p2(v_j \mid v_i) = \frac{\exp(\vec{u}_j'^T \cdot \vec{u}_i)}{\sum_{k=1}^{|V|} \exp(\vec{u}_j'^T \cdot \vec{u}_i)} \tag{12}$$

In equation (12), v_i represents the source node. v_j represents the context node. $\vec{u}_j'^T$ and $\vec{u}_i$ represent the representation vectors of the context node and the source node, respectively. $|V|$ signifies the number of all nodes. $p_2(v_i, v_j)$ represents the probability of obtaining the context node v_j. Based on above methods, the second-order distance information between nodes can be

captured, thus better understanding the global topology of the network. The objective function for capturing second-order distance structure information is displayed in equation (13).

$$O_2 = \sum_{i \in V} \lambda_i d(\hat{p}_2(\cdot|v_i), p_2(\cdot|v_i)) \tag{13}$$

In equation (13), O_2 represents the objective function of second-order distance. $d(\hat{p}_2(\cdot|v_i), p_2(\cdot|v_i))$ represents the distance between probability distributions $\hat{p}_2$ and p_2. λ_i represents the weight of node p_2. By minimizing this objective function, this equation can optimize the low dimensional representation of nodes. Therefore, the model can better capture the indirect connection relationships between nodes in the network, thereby

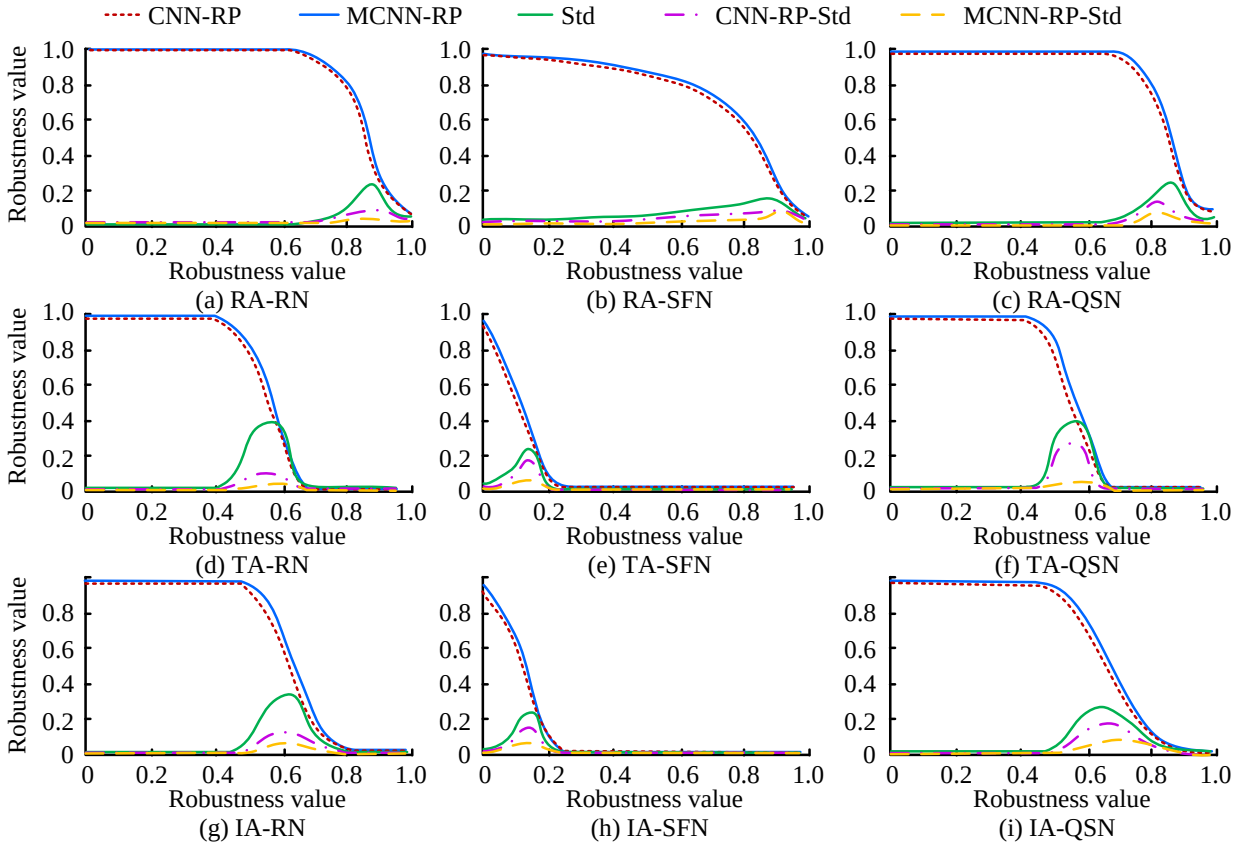


Figure 6: Prediction results of CNN-RP and MCNN-RP algorithms for network connectivity robustness.

improving the accuracy of robust prediction. The empirical distribution $\hat{p}_2(\cdot|v_i)$ is displayed equation (14).

$$\hat{p}_2(v_j|v_i) = \frac{\omega_{ij}}{d_i} \tag{14}$$

In equation (14), ω_{ij} represents the weight of edge $(i.j)$ in the graph. v_i represents the degree of the node. By introducing empirical distributions, the connection relationships between nodes can be more accurately captured, thereby improving the model's understanding of network topology. To better capture second-order distance information, the study uses relative entropy to represent the distance of node probability distribution, as shown in equation (15).

$$O_2 = \sum_{(i,j) \in E} \omega_{ij} \log p_2(v_j|v_i) \tag{15}$$

In equation (15), $\omega_{ij} \log p_2(v_i|v_j)$ represents the second-order distance between the relative entropy probability distributions. Relative entropy not only considers the absolute differences in probability distribution, but also considers the relative differences in probability distribution. Therefore, it can more accurately reflect the prediction error of the model. After obtaining information on node features, the robustness of complex networks is predicted.

3 Results

This chapter analyzes the network robustness prediction performance of the improved topology prior MCNN-RP algorithm through three attack methods, verifying the effectiveness. Aiming at the robustness and generalization of complex networks, a graph representation learning algorithm based on SA-LFR-CNN is analyzed to verify the robustness prediction performance of complex networks.

3.1 Analysis of robustness prediction performance of complex networks

based on improved CNN-RP network topology prior knowledge

The experiment uses three network models, Random Network (RN), Scale Free Network (SFN), and Q-Snapback Network (QSN), to generate network samples. Supervised learning is used to train the MCNN-RP network topology model. Three attack methods are set, including Tree Attack (TA), Intermediate Attack (IA), and Random Attack (RA). In the RN, SFN, and QSN datasets, their sizes all contain 1,000 nodes. For RN, the connections between nodes are randomly generated and do not have a specific topology structure. SFN follows a power-law distribution, with a few definitions having a large number of connections, and a large number of nodes having only a few connections. It has a clear

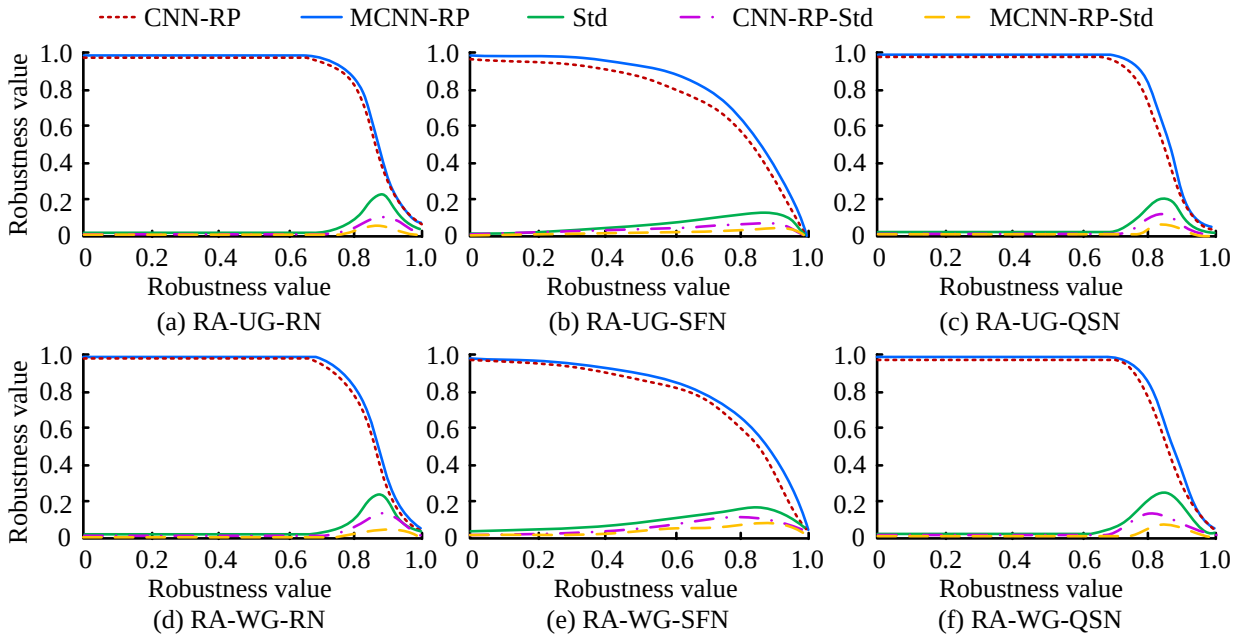


Figure 7: Network robustness prediction results of undirected graphs and weighted graphs under random attacks.

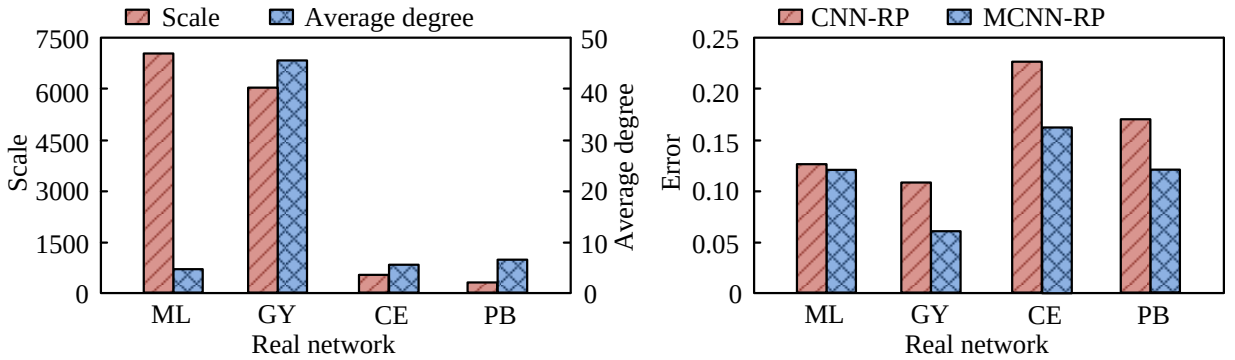
central node, and the network has a small world characteristic with shorter paths. For QSN, it has Guzan's dynamic characteristics, where the connection relationships of nodes change over time. Due to the dynamic changes in connection relationships, the robustness of the network is highly uncertain. During the training process, the average degrees of the training and testing samples are 6.01 and 5.49, respectively. The training set is 6,400, and both the validation and test sets are 1,600. The operating system for this experiment is Ubuntu 20.04 LTS, the deep learning framework is PyTorch 1.8, and the programming language is Python 3.8. The model learning rate is set to 0.001, the batch size is 32, and the number of training rounds is 100.

The prediction error and robustness prediction accuracy are selected to measure the predictive performance of the network. The prediction error directly reflects the accuracy of the model prediction. The lower prediction error indicates that the model can more accurately predict the robustness of the network after being attacked. In practical applications, accurate robustness prediction can help network optimizers better understand the vulnerability of the network and take effective measures to enhance its stability. The accuracy of robustness prediction reflects the reliability of the model in predicting network robustness. The higher prediction accuracy indicates that the model can more accurately capture the topology and dynamic characteristics of the network, providing more valuable references for network optimization.

The experiment conducts three types of attacks on three networks and compares the connectivity robustness prediction results of CNN-RP and MCNN-RP, as shown in Figure 6. Under random attacks, the CNN-RP algorithm for generating network samples from RN, SFN, and QSN had errors of 0.0332, 0.0501, and 0.0384 in predicting the robustness of complex networks, respectively. The MCNN-RP algorithm had prediction errors of 0.0312, 0.0454, and 0.0335 for network robustness, respectively. Under degree attack, the prediction errors of the CNN-RP for the robustness of the three network samples were 0.0532, 0.0224, and 0.0503, respectively. Compared with the CNN-RP, the prediction errors of the MCNN-RP were reduced by 0.0015, 0.0125, and 0.0040, respectively. Under the betweenness centrality attack, the CNN-RP algorithm for RN, SFN, and QSN network samples had prediction errors of 0.0566, 0.0287, and 0.0574 for the robustness of complex networks, respectively. Compared with CNN-RP, the prediction errors of the MCNN-RP algorithm were reduced by 6.53%, 53.65%, and 1.05%. The prediction error of MCNN-RP algorithm is significantly lower than that of CNN-RP, which proves that the algorithm has

good predictive performance for the robustness of complex networks.

The experiment predicts the network robustness through undirected and weighted graphs to verify the effectiveness of the MCNN-RP, as shown in Figure 7. The prediction results in Figure 7 showed that under random attacks, the CNN-RP algorithm based on RN, SFN, and QSN had prediction errors of 0.0338, 0.0632, and 0.0375 for the robustness of undirected complex networks, respectively. The prediction errors for the robustness of weighted complex networks were 0.0326, 0.0584, and 0.0419, respectively. The MCNN-RP for three types of network samples had prediction errors of 0.0301, 0.0484, and 0.0323 for the undirected complex network, and 0.0285, 0.0541, and 0.0314 for the weighted complex network, respectively. Compared with the CNN-RP algorithm, the MCNN-RP algorithm reduced the robustness prediction of undirected networks by 10.94%, 23.41%, and 13.86%, respectively, and reduced the robustness prediction of weighted networks by 0.0041, 0.0043, and 0.0105, respectively. The results



(a) The scale and average of the real network model (b) Error of CNN-RP and MCNN-RP algorithms

Figure 8: Robustness prediction errors in real networks (ML, GY, CE, and PB).

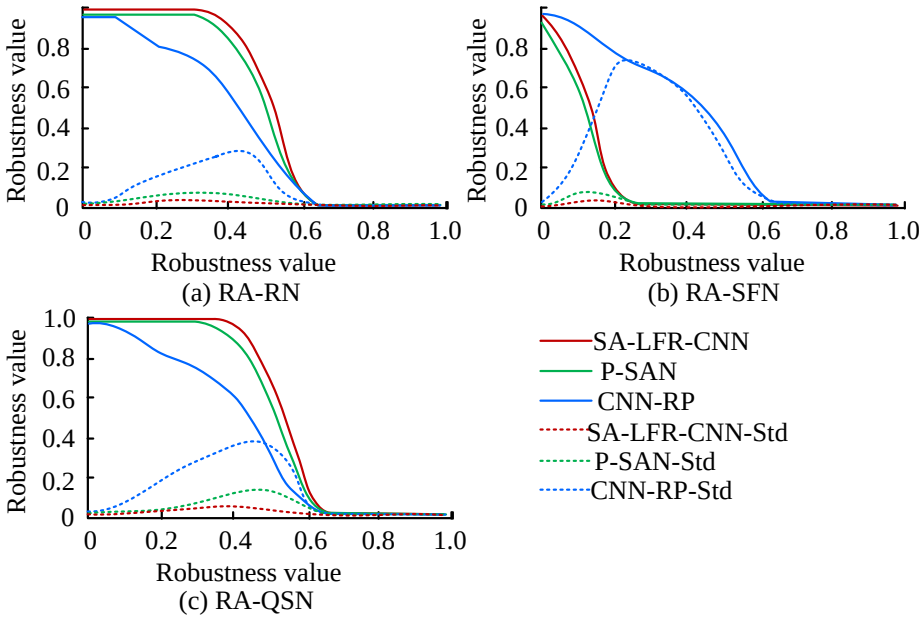


Figure 9: Connectivity robustness prediction in undirected graph networks: Comparison of CNN-RP, P-SAN, and SA-LFR-CNN.

have verified the effectiveness and feasibility of the MCNN-RP algorithm.

To verify the superiority of the MCNN-RP, it is compared with the CNN-RP algorithm in a network resource library. Real networks include Movie Lens-user (ML), Grid Yeast (GY), C-Elegance (CE), and Polbooks (PB). The results are displayed in Figure 8. In Figure 8 (a), the scales of ML, GY, CE, and PB real networks were 7041, 6002, 277, and 110, respectively, with average degrees of 7.20, 47.56, 7.88, and 8.45, respectively. In Figure 8 (b), the CNN-RP and MCNN-RP had prediction robustness errors of 0.1295 and 0.1284 for ML network, 0.1079 and 0.0556 for GY network, 0.2518 and 0.1667 for CE network, and 0.1707 and 0.1266 for PB network, respectively. Compared with the CNN-RP, the MCNN-RP reduced the robustness prediction error by 0.85%, 48.47%, 33.79%, and 25.83% in four real networks, respectively. The MCNN-RP has better robustness prediction performance for real networks.

3.2 Verification of network robustness prediction performance based on SA-LFR-CNN graph representation learning algorithm

The experiment compares the undirected graph connectivity robustness prediction results of CNN-RP, P-SAN, and SA-LFR-CNN algorithms, as shown in Figure 9. In Figure 9 (a), under random attacks, the prediction accuracy of CNN-RP, P-SAN, and SA-LFR-CNN algorithms for the robustness of random networks were 0.965, 0.974, and 0.988, respectively, with prediction error values of 0.1225, 0.0432, and 0.0320. The SA-LFR-CNN reduced prediction errors by 0.0793 and 0.0112 compared with the CNN-RP and P-SAN algorithms, respectively. In Figure 9 (b), the prediction accuracy of these three algorithms for the scale-free network was 0.984, 0.971, and 0.981, respectively, with prediction error values of 0.2566, 0.0723, and 0.0445. Compared with the other two algorithms, the SA-LFR-CNN algorithm reduced prediction errors by 0.1843 and 0.0278. In Figure 9 (c), the prediction accuracy of the three algorithms for the robustness of the Q-recovery network was 0.980, 0.984, and 0.992, respectively, with prediction error values of 0.1095, 0.0572, and 0.0441. Compared with the other two algorithms, the SA-LFR-CNN algorithm reduced prediction errors by 47.76% and

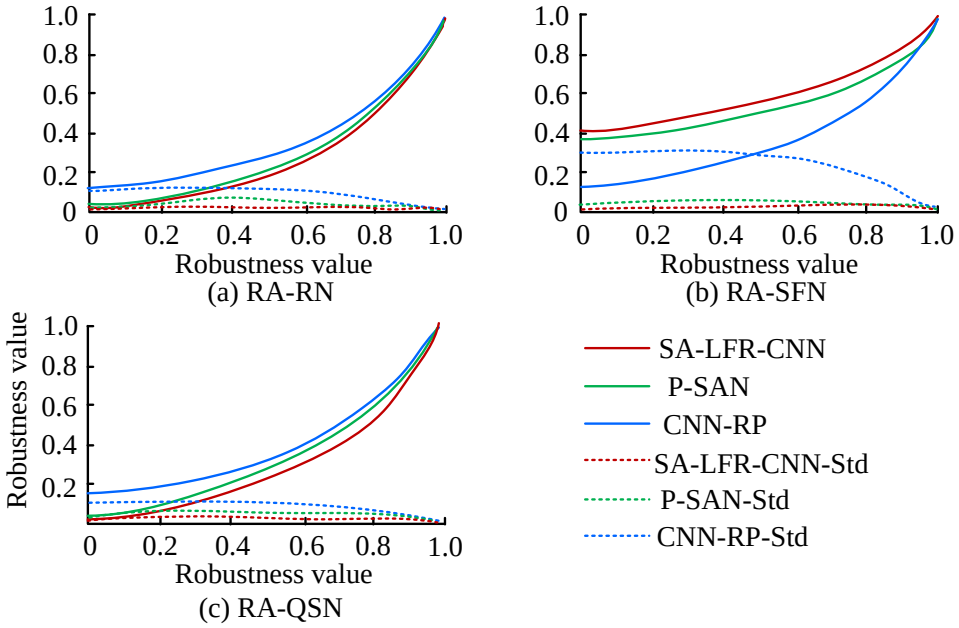


Figure 10: Controllability robustness prediction in directed graph networks: Comparison of CNN-RP, P-SAN, and SA-LFR-CNN.

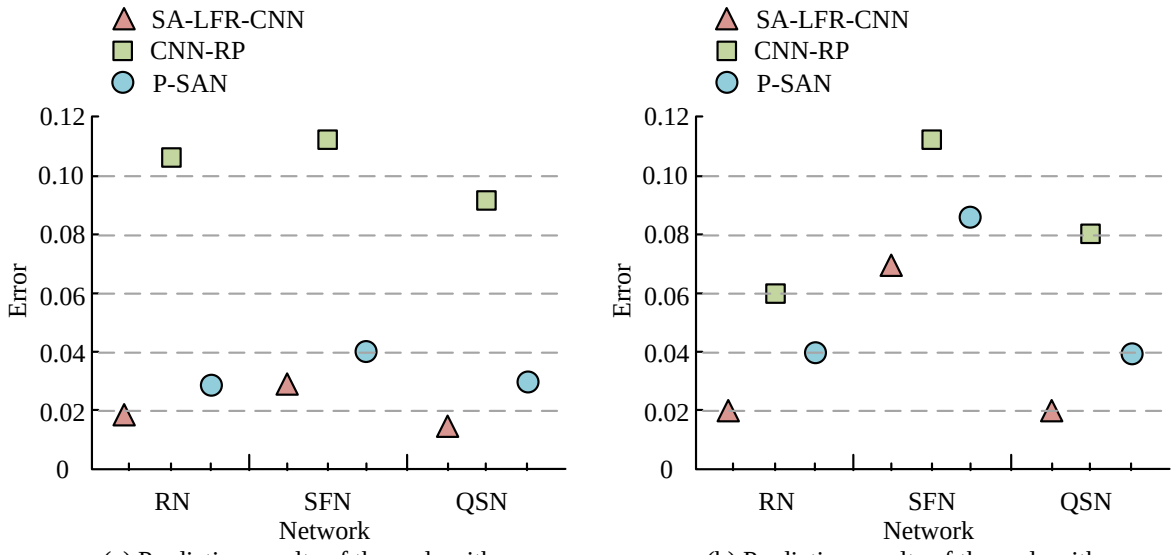
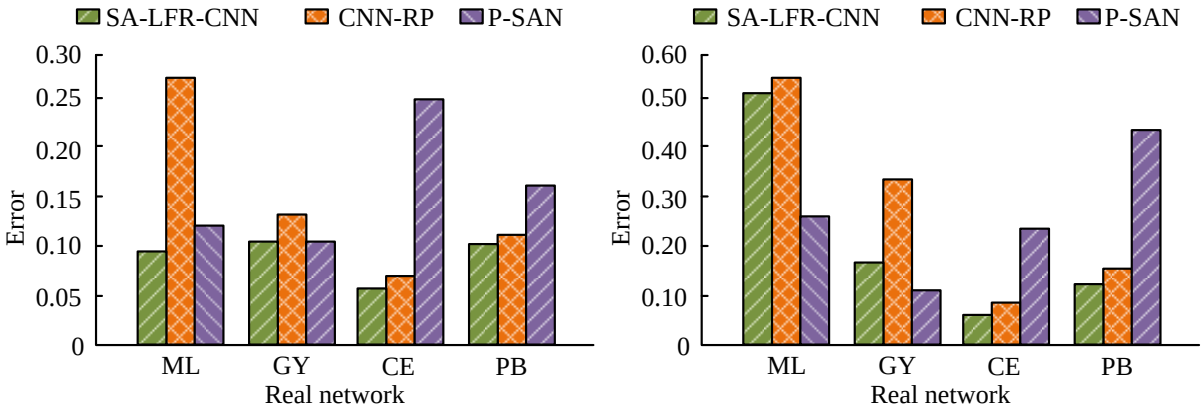


Figure 11: Robustness prediction under different network sizes: CNN-RP, P-SAN, and SA-LFR-CNN.

22.90%, respectively. The SA-LFR-CNN has high prediction accuracy, which is superior to other algorithms.

To further verify the predictive performance of the SA-LFR-CNN, the directed graph controllability robustness prediction results of CNN-RP, P-SAN, and SA-LFR-CNN algorithms are shown in Figure 10. In Figure 10 (a), under random attacks, the prediction error values of CNN-RP, P-SAN, and SA-LFR-CNN for the robustness of the random network were 0.0678, 0.0639, and 0.0300, respectively. The SA-LFR-CNN algorithm reduced prediction errors by 0.0378 and 0.0339, respectively, compared with the CNN-RP and P-SAN. Figure 10 (b) displayed that the prediction error values of CNN-RP, P-SAN, and SA-LFR-CNN algorithms for the robustness of scale-free networks were 0.1152, 0.0832, and 0.0825. The SA-LFR-CNN algorithm reduced prediction errors by 0.0327 and 0.0007. Figure 10 (c) displayed that the prediction error values of CNN-RP, P-SAN, and SA-LFR-CNN algorithms for the robustness of Q-recovery networks were 0.0703, 0.0643, and 0.0302. The SA-LFR-CNN reduced prediction errors by 57.04% and



(a) The connectivity robust prediction results of three algorithms

(b) The controllability robust prediction results of three algorithms

Figure 12: Robustness prediction errors in real networks (ML, GY, CE, and PB): CNN-RP, P-SAN, and SA-LFR-CNN.

Table 2: MAE, RMSE and confidence interval results of different algorithms.

Model	Dataset	MAE	RMSE	Confidence interval	Statistical significance
CNN	RN	0.1225	0.1532	[0.1150, 0.1300]	P<0.05
	SFN	0.2566	0.3054	[0.2450, 0.2682]	P <0.05
	QSN	0.1095	0.1321	[0.1000, 0.1190]	P <0.05
MCNN-RP	RN	0.0312	0.0391	[0.0285, 0.0339]	P <0.05
	SFN	0.0454	0.0567	[0.0412, 0.0496]	P <0.05
	QSN	0.0335	0.0421	[0.0301, 0.0369]	P <0.05
SA-LFR-CNN	RN	0.0320	0.0385	[0.0290, 0.0350]	P <0.05
	SFN	0.0445	0.0543	[0.0408, 0.0482]	P <0.05
	QSN	0.0441	0.0512	[0.0405, 0.0477]	P <0.05

Note: P<0.05 indicates that the statistical test results are significant.

53.03%, respectively. The SA-LFR-CNN has good robustness prediction performance, verifying the effectiveness.

Different network sizes can affect the robustness prediction results of the network. Therefore, the robustness prediction algorithms under different network scales are compared in the experiment, as displayed in Figure 11. In Figure 11 (a), when the network size was 800, the robustness prediction errors of CNN-RP, P-SAN, and SA-LFR-CNN algorithms under RN samples were 0.1124, 0.0320, and 0.0185, respectively. The prediction errors of the three algorithms under SFN network samples were 0.1145, 0.0402, and 0.0326, respectively. The prediction errors under QSN network samples were 0.0921, 0.0336, and 0.0128. Figure 11 (b) indicated that when the network size was 1,000, the CNN-RP algorithm predicted network robustness with error values of 0.0603, 0.1148, and 0.0813 for RN, SFN, and QSN network samples, respectively. The P-SAN algorithm predicted robustness with error of 0.0403, 0.0827, and 0.0396. The SA-LFR-CNN predicted network robustness with error of 0.0204, 0.0722, and 0.0195. The experimental results show that under different network sizes, the SA-LFR-CNN has high prediction accuracy.

To further verify the robustness and generalization ability of the SA-LFR-CNN algorithm, the robustness prediction results of four real networks are compared, ML, GY, CE, and PB, as shown in Figure 12. In Figure 12 (a), the CNN-RP algorithm had connectivity robustness prediction error values of 0.2745, 0.1334, 0.0715, and 0.1076 for ML, GY, CE, and PB networks, respectively. The P-SAN algorithm had prediction error values of 0.1296, 0.1108, 0.2504, and 0.1706 for the four real

networks, and the SA-LFR-CNN algorithm had prediction error values of 0.0906, 0.1106, 0.0715, and 0.1052 for the four networks, respectively. In Figure 12 (b), the CNN-RP algorithm had prediction errors of 0.5744, 0.3746, 0.0672, and 0.1895 for the controllability robustness of ML, GY, CE, and PB networks, respectively. The P-SAN had prediction errors of 0.2663, 0.1120, 0.2656, and 0.4328 for the four real networks, respectively. The SA-LFR-CNN algorithm had prediction errors of 0.5155, 0.1882, 0.0458, and 0.1456 for the four networks, respectively. The SA-LFR-CNN has a small prediction error value for network robustness and high prediction accuracy.

To further evaluate and compare the performance of different algorithms in predicting network robustness. The study compares the Mean Average Error (MAE), Root Mean Square Error (RMSE), and confidence intervals of different algorithms using the ANOVA method. The statistical significance of performance differences is verified. The results are shown in Table 2. The results showed that the MAE and RMSE of the proposed model were lower than those of the CNN model, indicating that the algorithm had better performance in predicting network robustness.

Table 3: Analysis of ablation experiment results.

Model	Dataset	Prediction error of connectivity robustness	Prediction error of controllability robustness
Baseline model	RN	0.1225	0.0678
	SFN	0.2566	0.1152
	QSN	0.1095	0.0703
Remove SAN operator	RN	0.0432	0.0378
	SFN	0.0723	0.0832
	QSN	0.0572	0.0643
Remove topological prior knowledge	RN	0.0332	0.0525
	SFN	0.0501	0.0963
	QSN	0.0387	0.0589
Complete model	RN	0.0320	0.0300
	SFN	0.0445	0.0825
	QSN	0.0352	0.0302

Further ablation experiments are conducted to better explain the performance of the model. The results are shown in Table 3. When applying SAN operator and topology prior knowledge, the robustness error of the model significantly increased.

4 Discussion and conclusion

This study proposed a robust prediction method for complex networks based on graph representation learning and improved CNN. To verify the effectiveness and superiority of the proposed method, the proposed method was comprehensively compared with existing CNN-RP and P-SAN baseline methods. Under random attacks, MCNN-RP reduced the prediction errors of undirected networks by 10.94%, 23.41%, and 13.86%, respectively, and reduced the prediction errors of weighted networks by 0.0041, 0.0043, and 0.0105, respectively. In addition, the SA-LFR-CNN reduced prediction errors by 0.1843 and 0.0278, as well as 47.76% and 22.90%, respectively, in the robustness prediction of scale-free networks and Q-recovery networks. In real network testing, the connectivity robustness prediction error values of SA-LFR-CNN algorithm were 0.0906, 0.1106, 0.0715, and 0.1052, respectively, and the controllability robustness prediction error values were 0.5155, 0.1882, 0.0458, and 0.1456, respectively. The results show that under random attacks, degree attacks, and betweenness attacks, the MCNN-RP algorithm has significantly lower robustness prediction errors for undirected and weighted networks than the CNN-RP algorithm. Compared with CNN-RP, MCNN-RP significantly improves the training and prediction speed of the model by introducing multiple convolutional neural network structures and prior knowledge of network topology. The SA-LFR-CNN algorithm further optimizes the computation process by improving the SAN operator and graph representation learning, making it more scalable when dealing with large-scale complex networks. The graph representation learning method can extract deep structural features of networks, thereby improving the ability to represent complex networks. Through graph representation learning, the model can better capture the topology and node relationships of the network, thereby improving prediction accuracy. In addition, the improved SAN operator can further optimize the computation process, making it more efficient and accurate in handling large-scale networks.

In summary, the MCNN-RP and SA-LFR-CNN algorithms proposed in this study demonstrate significant superiority in predicting the robustness of complex networks. By introducing graph representation learning and prior knowledge of network topology, these two algorithms outperform existing baseline methods on prediction accuracy, computational efficiency, and practical applications.

The limitations of this study are as follows. (1) Synthetic networks are created through specific generative models, which may not fully reflect the complexity and diversity of real networks. The generation model of this network is usually based on specific assumptions and parameter settings, which may result in the generated network deviating from the real network in certain characteristics. (2) The size and complexity of real network datasets may have an impact on model training and testing. The noise in the dataset may mask the true topology and dynamic

characteristics of the network, thereby affecting the generalization ability of the model. Therefore, future research can consider developing synthetic network generation models that are closer to the characteristics of real networks, which can better simulate the dynamic changes and local properties of real networks.

Funding

The research is supported by: Hubei Province Universities Outstanding Young and Middle-aged Scientific and Technological Innovation Team Project in 2022 (No. T2022035); The Natural Science Foundation of Hubei Province of China in 2022 (No. 2022CFB928); School Level Scientific Research Innovation Team of Wuhan College in 2020 (No. kyt202001); Hubei Yidan University Education Development Foundation, Research on Human-Machine Collaborative Governance Model for Short Video Content Ecosystem in the Digital Intelligence Era (No. JJA202510).

References

[1] Xin Zheng, and Xiaodong Zhang. Robustness of cloud manufacturing system based on complex network and multi-agent simulation. Entropy (basel), 25(1):45-62, 2022. https://doi.org/10.3390/e25010045

[2] Yang Lou, Yaodong He, Lin Wang, and Guanrong Chen. Predicting network controllability robustness: A convolutional neural network approach. IEEE Transactions on Cybernetics, 52(5):4052-4063, 2022. https://doi.org/10.1109/TCYB.2020.3013251

[3] Josef Baumgartner, Alexandra Schneider, Ulugbek Zhenis, Franz Jager, and Josef Winkler. Mastering neural network prediction for enhanced system reliability. Fusion of multidisciplinary research, 3(1):261-274, 2022.

[4] Yuxiao Yang, Shaoyu Qiao, Omid G. Sani, J. Isaac Sedillo, Breonna Ferrentino, Bijan Pesaran, and Maryam M. Shanechi. Modelling and prediction of the dynamic responses of large-scale brain networks during direct electrical stimulation. Nature Biomedical Engineering, 5(4):324-345, 2021. https://doi.org/10.1038/s41551-020-00666-w

[5] Anastasios Valkanis, Georgios Papadimitriou, Georgia Beletsioti, Emmanouel Varvarigos, and Petros Nicopolitidis. Efficiency and fairness improvement for elastic optical networks using reinforcement learning-based traffic prediction. Journal of Optical Communications and Networking, 14(3):25-42, 2021. https://doi.org/10.1364/JOCN.440590

[6] Zewei Ji, Runhan Shi, Jiarui Lu, Fang Li, and Yang Yang. ReLMole: Molecular representation learning based on two-level graph similarities. Journal of Chemical Information and Modeling, 62(22):5361-5372, 2022. https://doi.org/10.1021/acs.jcim.2c00798

[7] Xiaocheng Yang, Mingyu Yan, Shirui Pan, Xiaochun Ye, and Dongrui Fan. Simple and efficient heterogeneous graph neural network. In Proceedings of the AAAI Conference on Artificial Intelligence. 37(9):10816-10824, 2023. https://doi.org/10.1609/aaai.v37i9.26283

[8] Edith Heiter, Bo Kang, Tijl De Bie, and Jefrey Lijffijt. Evaluating representation learning and graph layout methods for visualization. IEEE Computer Graphics and Applications, 42(3):19-28, 2022. https://doi.org/10.1109/MCG.2022.3160104

[9] Liang Peng, Yujie Mo, Jie Xu, Jialie Shen, Xiaoshuang Shi, and Xiaoxiao Li. GRLC: Graph representation learning with constraints. IEEE Transactions on Neural Networks and Learning Systems, 35(6):8609-8622, 2024. https://doi.org/10.1109/TNNLS.2022.3230979

[10] Yujie Mo, Liang Peng, Jie Xu, Xiaoshuang Shi, and Xiaofeng Zhu. Simple unsupervised graph representation learning. Proceedings of the AAAI Conference on Artificial Intelligence, 36(7):7797-7805, 2022. https://doi.org/10.1609/aaai.v36i7.20748

[11] Kavita Bhosle, and Vijaya Musande. Evaluation of deep learning CNN model for recognition of devanagari digit. Artificial Intelligence and Applications, 1(2):114-118, 2023. https://doi.org/10.47852/bonviewAIA3202441

[12] Tianjian Dong, Qi Qi, Jingyu Wang, Alex X. Liu, Haifeng Sun, and Zirui Zhuang. Generative adversarial network-based transfer reinforcement learning for routing with prior knowledge. IEEE Transactions on Network and Service Management, 18(2):1673-1689, 2021. https://doi.org/10.1109/TNSM.2021.3077249

[13] Jiajun Chen, Huarui He, Feng Wu, and Jie Wang. Topology-aware correlations between relations for inductive link prediction in knowledge graphs. Proceedings of the AAAI Conference on Artificial Intelligence, 35(7):6271-6278, 2021. https://doi.org/10.1609/aaai.v35i7.16779

[14] Laura von Rueden, Sebastian Mayer, Katharina Beckh, Bogdan Georgiev, Sven Giesselbach, Raoul Heese, Birgit Kirsch, Julius Pfrommer, Annika Pick, Rajkumar Ramamurthy, Michal Walczak, Jochen Garcke, Christian Bauckhage, and Jannis Schuecker. Informed machine learning–a taxonomy and survey of integrating prior knowledge into learning systems. IEEE Transactions on Knowledge and Data Engineering, 35(1):614-633, 2021. https://doi.org/10.1109/TKDE.2021.3079836

[15] Xiao Li, Ziqi Wang, Bo Zhang, Fuchun Sun, and Xiaolin Hu. Recognizing object by components with human prior knowledge enhances adversarial robustness of deep neural networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(7):8861-8873, 2023. https://doi.org/10.1109/TPAMI.2023.3237935

[16] Entesar Abdulsaed, Maytham Alabbas, and Raidah Khudeyer. Hyperparameter optimization for convolutional neural networks using the salp swarm algorithm. Informatica, 47(9):133-144, 2023. https://doi.org/10.31449/inf.v47i9.5148

[17] Sebastian Ehlert, Marcel Stahn, Sebastian Spicher, and Stefan Grimme. Robust and efficient implicit solvation model for fast semiempirical methods. Journal of Chemical Theory and Computation, 17(7):4250-4261, 2021. https://pubs.acs.org/doi/10.1021/acs.jctc.1c00471

[18] Yang Lou, Yaodong He, Lin Wang, Kim Fung Tsang, and Guanrong Chen. Knowledge-based prediction of network controllability robustness. IEEE Transactions on Neural Networks and Learning Systems, 33(10):5739-5750, 2021. https://doi.org/10.1109/TNNLS.2021.3071367

[19] Yang Lou, Ruizi Wu, Junli Li, Lin Wang, Xiang Li, and Guanrong Chen. A learning convolutional neural network approach for network robustness prediction. IEEE Transactions on Cybernetics, 53(7):4531-4544, 2022. https://doi.org/10.1109/TCYB.2022.3207878

[20] Yangyan Liu, and Bolin Pan. Profit estimation model and financial risk prediction combining multi-scale convolutional feature extractor and BGRU model. Informatica, 48(11):15-32, 2024. https://doi.org/10.31449/inf.v48i11.5941

[21] Zhiwen Chen, Jiamin Xu, Tao Peng, and Chunhua Yang. Graph convolutional network-based method for fault diagnosis using a hybrid of measurement and prior knowledge. IEEE Transactions on Cybernetics, 52(9):9157-9169, 2021. https://doi.org/10.1109/TCYB.2021.3059002

[22] Emrah Irmak. Multi-classification of brain tumor MRI images using deep convolutional neural network with fully optimized framework. Iranian Journal of Science and Technology, Transactions of Electrical Engineering, 45(3):1015-1036, 2021. https://doi.org/10.1007/s40998-021-00426-9

[23] Xiaocheng Yang, Mingyu Yan, Shirui Pan, Xiaochun Ye, and Dongrui Fan. Simple and efficient heterogeneous graph neural network. Proceedings of the AAAI Conference on Artificial Intelligence, 37(9):10816-10824, 2023. https://doi.org/10.1609/aaai.v37i9.26283

[24] Martin Genzel, Jan Macdonald, and Maximilian März. Solving inverse problems with deep neural networks–robustness included? IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(1):1119-1134, 2022. https://doi.org/10.1109/TPAMI.2022.3148324

[25] Harry Rogers, Beatriz De La Iglesia, Tahmina Zebin, Grzegorz Cielniak, and Ben Magri. Advancing precision agriculture: Domain-specific augmentations and robustness testing for convolutional neural networks in precision spraying evaluation. Neural Computing and Applications, 36(32), 20211-20229, 2024. https://doi.org/10.1007/s00521-024-10142-0

Nonlinear Model Predictive Control and Dynamic Modelling of a Small-Scale Organic Rankine Cycle for Low-Grade Waste Heat Recovery

Jing Li[1*], Jun Chen[2]
[1]Institute of Engineering Technology, University of Science & Technology Beijing, Beijing 100083, beijing, China
[2]Jiangsu Sanjay Energy Environmental Protection Co，LTD, Yixing 214200, Jiangsu, China
E-mail: 18618309526@163.com, chenjun@tbmaglev.com
[*]Corresponding author

Keywords: dynamic modelling, energy efficiency, extended kalman filter, experimental validation, nonlinear model predictive control, organic rankine cycle, power generation, sustainable energy systems, waste heat recovery

Recieved: October 25, 2024

This paper presents a comprehensive study on the dynamic modeling and control of a small-scale organic Rankine cycle (ORC) system for waste heat recovery in power generation applications. The primary objective is to develop an effective control strategy that optimizes the system performance under transient operating conditions, maximizing the expander power output. A high-fidelity dynamic model of the evaporator is developed using the finite volume method and validated against experimental data. A reduced-order control-oriented model is derived using the moving boundary method to facilitate the design of the controller. A nonlinear model predictive control (NMPC) strategy is proposed, utilizing the control-oriented model to predict future system behavior and optimize control actions while explicitly considering system constraints. An extended Kalman filter (EKF) is integrated with the NMPC controller to estimate unmeasured state variables. Experimental validation on an engine test bench demonstrates the controller's ability to effectively track the reference superheating degree with an average error of ±2°C, compensate for disturbances within 5 seconds, and maintain safe expander operation under highly transient conditions. The NMPC approach achieved a 15% improvement in cycle efficiency compared to steady-state operation. Comparative analysis with traditional control methods highlights the superiority of the NMPC approach, reducing tracking error by 40% and response time by 50% compared to PID control, while effectively handling system constraints. The proposed dynamic modeling and control strategies provide a solid foundation for the effective utilization of small-scale ORC systems in waste heat recovery applications, contributing to the advancement of efficient and sustainable energy systems.

Povzetek: Predlagana raziskava se osredotoča na dinamično modeliranje in obvladovanje majhnega sistema organskih Rankine ciklov (ORC) za izrabo odpadne toplote v vozilih, kjer nevronsko napovedovanje modelov optimizira delovanje sistema in povečuje učinkovitost cikla za 15% v primerjavi z uporabo tradicionalnih metod.

1 Introduction

The transportation sector, particularly heavy-duty vehicles such as trucks, is a significant contributor to global energy consumption and greenhouse gas emissions. According to recent statistics, the fuel consumption and carbon emissions from global freight trucks have been steadily increasing over the years, as shown in Fig. 1 [1].

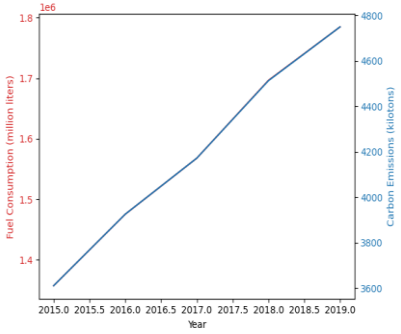


Figure 1: Global fuel consumption and carbon emissions from freight trucks

To mitigate the environmental impact and improve the efficiency of heavy-duty vehicles, waste heat recovery (WHR) systems have gained considerable attention [2]. Among various WHR technologies, the Organic Rankine Cycle (ORC) has emerged as a promising solution for recovering low-grade waste heat from vehicle exhaust gases [3]. The ORC system utilizes an organic working fluid with a lower boiling point compared to water, enabling the extraction of energy from low-temperature heat sources [4]. Numerous studies have investigated the application of ORC systems in vehicle waste heat recovery, focusing on working fluid selection [5], system design and optimization [6-8], and experimental evaluations [9, 10].

However, the dynamic nature of vehicle operating conditions poses challenges to the efficient operation and control of ORC systems [11]. The waste heat available from the exhaust gases varies significantly depending on the vehicle's speed, load, and ambient conditions [12]. Moreover, the transient behavior of the ORC system, including the response time and stability, is crucial for real-time power generation [13]. To address these challenges, dynamic modeling and advanced control strategies are essential for optimizing the performance of vehicle-based ORC systems [14, 15].

Despite the progress made in ORC research for vehicle applications, there are still limitations in the existing literature. Many studies focus on steady-state modeling and optimization [16], neglecting the dynamic behavior of the system under varying operating conditions. Additionally, the control strategies proposed in the literature often rely on simplified models and assumptions [17], which may not accurately capture the complex dynamics of the ORC system [18].

To bridge these gaps, this paper aims to develop a dynamic model of a small-scale ORC system for waste heat recovery in vehicle applications. The model will consider the transient behavior of the system components, including the evaporator, expander, condenser, and pump [19]. Furthermore, advanced control strategies will be investigated to maximize the power output and efficiency of the ORC system under varying operating conditions [20].

Table 1 summarizes recent state-of-the-art studies on ORC modeling and control, highlighting the gaps in transient modeling and control under constraints that this paper addresses.

Table 1: Comparison of recent ORC modeling and control studies

Study	**Modeling Approach**	**Control Method**	**Considers Transients**	**Handles Constraints**
[16]	Steady-state	PID	No	No
[17]	Quasi-steady	MPC	Partial	Yes
[18]	Dynamic	Fuzzy Logic	Yes	No
This work	Dynamic (FVM+MBM)	NMPC	Yes	Yes

As shown in Table 1, while recent studies have made progress in various aspects of ORC modeling and control, there remains a gap in combining accurate dynamic modeling with advanced control strategies that can handle both system transients and operational constraints. This paper aims to address this gap by proposing a nonlinear model predictive control (NMPC) strategy based on a high-fidelity dynamic model for small-scale ORC systems.

Despite the progress made in ORC research for vehicle applications, significant challenges remain in optimizing system performance under highly dynamic operating conditions. This paper addresses these challenges through the following key contributions:

- ❖ Development of a high-fidelity dynamic model of a small-scale ORC system using the finite volume method, validated against experimental data from an engine test bench.
- ❖ Design of a nonlinear model predictive control (NMPC) strategy that explicitly accounts for system constraints and optimizes performance under transient conditions.
- ❖ Implementation of an extended Kalman filter (EKF) for state estimation, enabling robust control even with limited sensor measurements.
- ❖ Comprehensive experimental validation of the proposed control strategy, demonstrating significant improvements in tracking performance, disturbance rejection, and overall system efficiency compared to traditional control methods.
- ❖ Detailed sensitivity analysis of key NMPC parameters, providing insights into controller tuning for optimal performance in real-world applications.

By addressing these aspects, this research contributes to advancing the practical implementation of ORC systems for waste heat recovery in vehicular applications, potentially leading to significant improvements in fuel efficiency and reduced emissions in the transportation sector.

The main objectives of this study are as follows: 1. Develop a dynamic model of a small-scale ORC system for vehicle waste heat recovery, considering the transient behavior of system components. 2. Investigate advanced control strategies to optimize the performance of the ORC system under varying operating conditions. 3. Evaluate the performance of the proposed dynamic model and control strategies through simulations and case studies.

The findings of this research will contribute to the advancement of ORC technology for vehicle waste heat recovery, enabling more efficient and sustainable transportation systems. The dynamic modeling approach and control strategies developed in this study can be extended to other applications of small-scale ORC systems, such as industrial waste heat recovery and renewable energy generation.

2 System description and dynamic characteristics analysis

2.1 Organic rankine cycle system principles and experimental platform

The Organic Rankine Cycle (ORC) is a thermodynamic cycle that converts heat into mechanical work or electricity. It operates on the same principles as the conventional steam Rankine cycle but uses an organic working fluid instead of water [21]. The choice of working fluid depends on the temperature range of the heat source and the desired operating conditions of the system [22]. Fig. 2 illustrates the schematic diagram of a typical ORC system.

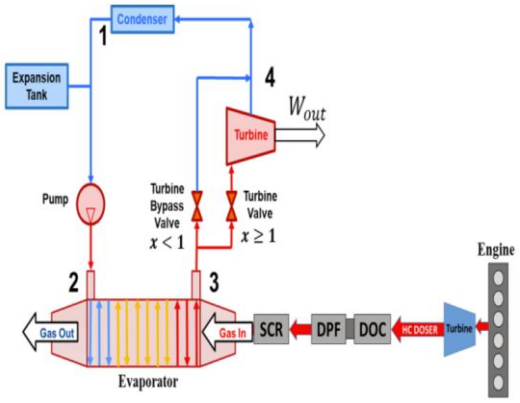


Figure 2: Schematic diagram of An ORC system

The ORC system consists of four main components: evaporator, expander, condenser, and pump. The working fluid is pumped to high pressure and enters the evaporator, where it absorbs heat from the low-grade heat source, such as vehicle exhaust gases [23]. The heated working fluid then expands in the expander, generating mechanical work or driving an electric generator. After expansion, the working fluid enters the condenser, where it rejects heat to a cooling medium and condenses back to a liquid state. The liquid working fluid is then pumped back to the evaporator, completing the cycle [24].

To investigate the dynamic behavior and control strategies of a small-scale ORC system for waste heat recovery in vehicle applications, an experimental platform was developed. The experimental setup consists of a diesel engine, an exhaust gas aftertreatment system, and the ORC system, as shown in Fig. 3.

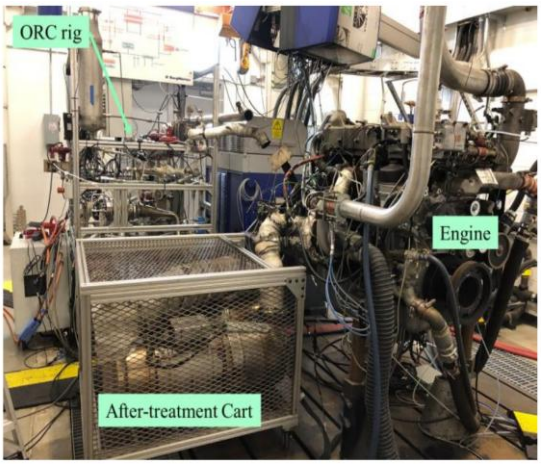


Figure 3: Experimental test platform

The diesel engine serves as the primary power source and generates exhaust gases containing waste heat. The exhaust gases pass through the aftertreatment system, which includes a diesel oxidation catalyst (DOC) and a diesel particulate filter (DPF) to reduce emissions. The ORC system is integrated with the exhaust aftertreatment system to recover the waste heat from the exhaust gases.

The main components of the ORC system in the experimental platform are as follows: 1. Evaporator: A plate-fin heat exchanger is used as the evaporator, where the working fluid absorbs heat from the exhaust gases. 2. Expander: A scroll expander is employed to convert the high-pressure vapor into mechanical work. 3. Condenser: A water-cooled plate heat exchanger acts as the condenser, rejecting heat from the working fluid to the cooling water. 4. Pump: A diaphragm pump is used to circulate the working fluid and maintain the desired system pressure.

The experimental platform is equipped with various sensors and measurement devices to monitor and record the key parameters of the ORC system, such as temperatures, pressures, flow rates, and power output

[25]. The data acquisition system collects and stores the measured data for further analysis and validation of the dynamic model.

The experimental tests are conducted under different operating conditions, simulating the varying waste heat availability and engine load profiles encountered in real-world driving scenarios. The transient response of the ORC system is captured by subjecting it to step changes in the heat source temperature and flow rate. The collected experimental data will be used to validate the dynamic model and evaluate the performance of the proposed control strategies.

2.2 Working fluid selection and thermodynamic analysis

The selection of an appropriate working fluid is crucial for the efficiency and safety of the ORC system [26]. The ideal working fluid should have favorable thermodynamic properties, such as high thermal stability, low critical temperature and pressure, and good compatibility with system materials [27]. Additionally, environmental and safety aspects, including low global warming potential (GWP), low ozone depletion potential (ODP), and non-toxicity, should be considered [28].

For the small-scale ORC system in this study, several organic working fluids were investigated, including R245fa, R123, R134a, and isobutane. These fluids have been widely used in low-temperature ORC applications due to their suitable thermodynamic properties [29]. Fig. 4 presents the T-s diagrams of the selected working fluids, comparing their thermal performance.

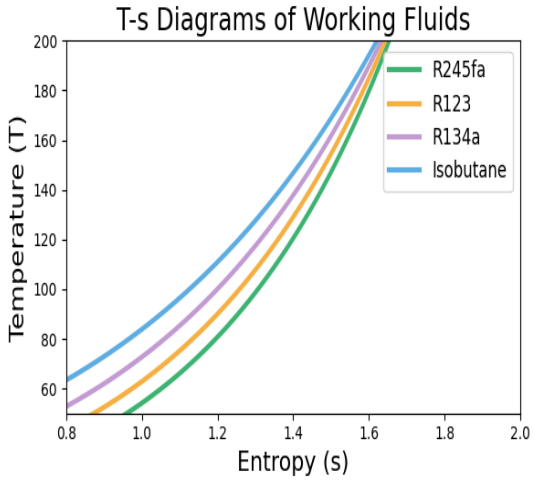


Figure 4: T-S Diagrams of different working fluids

To evaluate the efficiency of the ORC system with different working fluids, a thermodynamic analysis was conducted. The theoretical Carnot efficiency (η_{Carnot}) represents the maximum achievable efficiency for a heat engine operating between a heat source at temperature T_H and a heat sink at temperature T_L. It is given by [30]:

$$\eta_{Carnot} = 1 - \frac{T_L}{T_H} \quad (1)$$

However, the actual efficiency of the ORC system is lower than the Carnot efficiency due to irreversibilities in the system components. The thermal efficiency of the ORC system (η_{ORC}) can be expressed as:

$$\eta_{ORC} = \frac{W_{net}}{Q_{in}} = \frac{W_{exp} - W_{pump}}{Q_{in}} \quad (2)$$

where W_{net} is the net power output, Q_{in} is the heat input from the heat source, W_{exp} is the expander power output, and W_{pump} is the pump power consumption.

The thermodynamic analysis was performed for different operating conditions, considering variations in the heat source temperature and flow rate. The results showed that R245fa and isobutane exhibited higher thermal efficiencies compared to R123 and R134a under the given operating conditions. The maximum theoretical Carnot efficiency achieved was around 25%, while the actual ORC system efficiency ranged from 8% to 15%, depending on the working fluid and operating parameters.

Based on the thermodynamic analysis and considering the environmental and safety aspects, R245fa was selected as the most suitable working fluid for the small-scale ORC system in this study. R245fa has a relatively low GWP and ODP, non-toxicity, and good thermal stability within the operating temperature range [31].

The thermodynamic analysis provides insights into the performance potential of the ORC system with different working fluids. However, it is based on steady-state conditions and does not consider the dynamic behavior of the system. In the following sections, a dynamic model of the ORC system will be developed to investigate its transient performance and develop appropriate control strategies.

2.3 System dynamic characteristics analysis

To investigate the dynamic behavior of the small-scale ORC system, open-loop simulations and experiments were conducted. The main focus was to analyze the influence of evaporator heat source conditions (exhaust gas temperature and flow rate) and working fluid pump speed on the superheating degree at the evaporator outlet.

The dynamic model of the ORC system was developed using Modelica language in the Dymola environment [32]. The model consists of sub-models for each component, including the evaporator, expander,

condenser, and pump. The evaporator model is based on the finite volume method, where the heat exchanger is discretized into several control volumes [33]. The heat transfer coefficients and pressure drop correlations were obtained from literature and calibrated using experimental data.

Firstly, the effect of exhaust gas temperature on the superheating degree was investigated. The exhaust gas temperature was varied in a step-wise manner, while keeping the exhaust gas flow rate and working fluid pump speed constant. Fig. 5 shows the dynamic response of the superheating degree to step changes in the exhaust gas temperature.

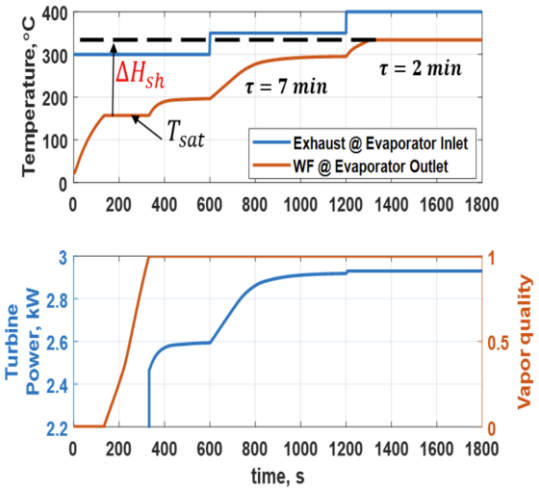


Figure 5: Effect of exhaust gas temperature on superheating degree

It can be observed that the superheating degree increases with the increase in exhaust gas temperature. The dynamic response exhibits a first-order behavior with a time constant that depends on the thermal inertia of the evaporator and the working fluid properties. The time constant can be estimated using the following equation [34]:

$$\tau = \frac{m_{ev}c_{p,ev}+m_{wf}c_{p,wf}}{UA_{ev}} \quad (3)$$

where τ is the time constant, m_{ev} and m_{wf} are the masses of the evaporator and working fluid, respectively, $c_{p,ev}$ and $c_{p,wf}$ are the specific heat capacities of the evaporator and working fluid, respectively, and UA_{ev} is the overall heat transfer coefficient of the evaporator.

Next, the influence of exhaust gas flow rate on the superheating degree was analyzed. The exhaust gas flow rate was subjected to step changes, while maintaining a constant exhaust gas temperature and working fluid pump speed. Fig. 6 depicts the dynamic response of the superheating degree to variations in the exhaust gas flow rate.

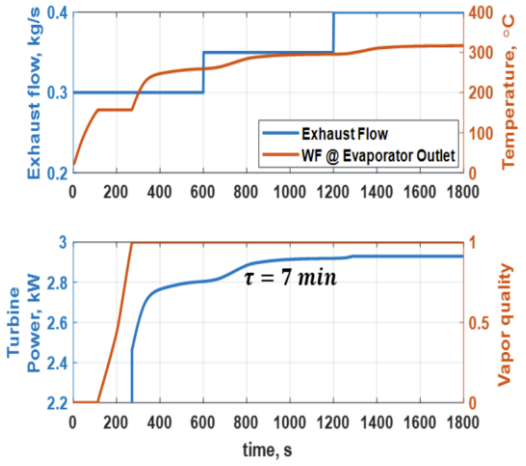


Figure 6: Effect of exhaust gas flow rate on superheating degree

As the exhaust gas flow rate increases, the superheating degree decreases due to the higher heat transfer rate in the evaporator. The dynamic response shows a similar first-order behavior as in the case of exhaust gas temperature variations. The time constant can be calculated using the same equation as before, considering the changes in the overall heat transfer coefficient due to the varying exhaust gas flow rate.

Lastly, the effect of working fluid pump speed on the superheating degree was investigated. The pump speed was varied in a step-wise manner, while keeping the exhaust gas temperature and flow rate constant. Fig. 7 illustrates the dynamic response of the superheating degree to step changes in the working fluid pump speed.

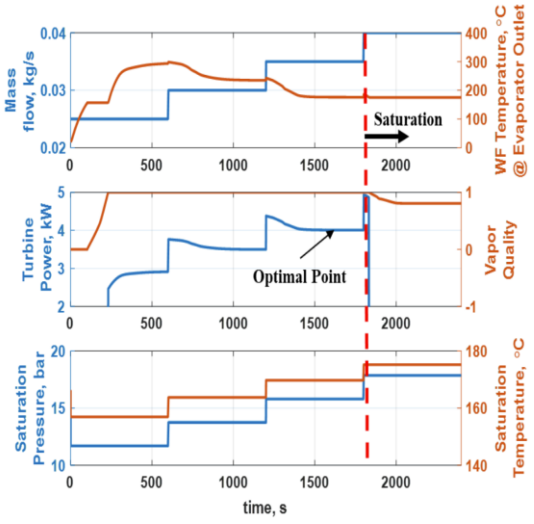


Figure 7: Effect of working fluid pump speed on superheating degree

The superheating degree decreases with the increase in working fluid pump speed. This is because a higher pump speed leads to a higher working fluid mass flow rate, reducing the residence time of the working fluid in the evaporator and consequently the heat transfer [35]. The dynamic response exhibits a second-order behavior, which can be approximated by the following transfer function:

$$G(s) = \frac{K}{\tau_1 s^2 + \tau_2 s + 1} \quad (4)$$

where K is the gain, τ_1 and τ_2 are the time constants, and s is the Laplace variable.

The open-loop simulations and experiments provide valuable insights into the dynamic characteristics of the small-scale ORC system. The results demonstrate that the superheating degree at the evaporator outlet is significantly influenced by the evaporator heat source conditions and the working fluid pump speed. The dynamic responses exhibit first-order and second-order behaviors, respectively, with time constants dependent on the system parameters and operating conditions.

These findings lay the foundation for the development of effective control strategies to maintain the desired superheating degree and optimize the performance of the ORC system under varying operating conditions. In the subsequent sections, different control approaches will be investigated, considering the dynamic characteristics and constraints of the system.

2.4 Control objectives

Based on the results of the dynamic characteristics analysis, it is evident that the evaporator heat source conditions and the working fluid pump speed have a significant impact on the superheating degree at the evaporator outlet. The superheating degree, in turn, affects the performance of the expander and the overall efficiency of the ORC system.

Considering the dynamic behavior of the system and the goal of maximizing the utilization of the low-grade waste heat, the control objective of the small-scale ORC system is determined to be the maximization of the expander power output. By optimizing the expander power output, the system can effectively convert the available waste heat into useful electrical energy, thereby improving the overall efficiency and economic viability of the waste heat recovery application.

To achieve this control objective, it is necessary to develop and implement appropriate control strategies that can adapt to the varying operating conditions and maintain the optimal superheating degree for maximum expander power output. The control strategies should consider the dynamic characteristics of the system, such as the response times and the coupling effects between different variables, to ensure stable and efficient operation of the ORC system.

In the following sections, different control approaches will be investigated and evaluated, aiming to maximize the expander power output while ensuring safe and reliable operation of the small-scale ORC system under dynamic operating conditions.

2.5 Research questions and hypotheses

Based on the dynamic characteristics analysis and control objectives, we formulate the following research questions and hypotheses:

RQ1: How does the NMPC strategy perform compared to traditional control methods (PID and LMPC) in tracking the optimal superheating degree under transient operating conditions? H1: We hypothesize that NMPC will achieve at least 30% lower tracking error and 40% faster response time compared to PID and LMPC controllers.

RQ2: To what extent can the NMPC strategy improve the overall efficiency of the ORC system in waste heat recovery applications? H2: We expect the NMPC strategy to increase the average cycle efficiency by at least 10% compared to steady-state operation under varying heat source conditions.

RQ3: How effectively does the NMPC handle system constraints compared to other control methods? H3: We hypothesize that NMPC will maintain all system variables within their specified limits 100% of the time, while PID and LMPC may violate constraints under rapid transients.

To address these research questions, we will conduct a series of experiments on the engine test bench, subjecting the ORC system to various transient scenarios representative of real-world driving conditions. The experimental methodology is detailed in Section 4.1.

3 Nonlinear model predictive controller design

3.1 High-fidelity evaporator modelling

To accurately capture the dynamic behavior of the evaporator and develop effective control strategies, a high-fidelity model of the evaporator is essential. In this study, the finite volume method (FVM) is employed to establish a distributed parameter model of the evaporator [36].

The evaporator is discretized into N control volumes along its length, and each control volume is treated as a lumped parameter system. The conservation equations for mass, energy, and momentum are applied

to each control volume to describe the dynamic behavior of the evaporator [37].

The mass conservation equation for the working fluid in each control volume is given by:

$$\frac{d(\rho_i V_i)}{dt} = \dot{m}_{i-1} - \dot{m}_i \quad (5)$$

where ρ_i is the density of the working fluid, V_i is the volume of the control volume, $\dot{m}_{i-1}$ and $\dot{m}_i$ are the mass flow rates entering and leaving the control volume, respectively.

The energy conservation equation for each control volume is expressed as:

$$\frac{d(U_i V_i)}{dt} = \dot{m}_{i-1} h_{i-1} - \dot{m}_i h_i + Q_i \quad (6)$$

where U_i is the specific internal energy of the working fluid, h_{i-1} and h_i are the specific enthalpies of the working fluid entering and leaving the control volume, respectively, and Q_i is the heat transfer rate from the exhaust gas to the working fluid.

The momentum conservation equation for each control volume is given by:

$$\frac{d(\dot{m}_i)}{dt} = \frac{A_i}{\Delta L_i}\left(P_{i-1} - P_i - \Delta P_{f,i}\right) \quad (7)$$

where A_i is the cross-sectional area of the control volume, ΔL_i is the length of the control volume, P_{i-1} and P_i are the pressures at the inlet and outlet of the control volume, respectively, and $\Delta P_{f,i}$ is the pressure drop due to friction.

The heat transfer rate Q_i is calculated using the following equation:

$$Q_i = \alpha_i A_{s,i}\left(T_{g,i} - T_{w,i}\right) \quad (8)$$

where α_i is the heat transfer coefficient, $A_{s,i}$ is the heat transfer surface area of the control volume, $T_{g,i}$ is the exhaust gas temperature, and $T_{w,i}$ is the working fluid temperature.

The heat transfer coefficient α_i is determined using appropriate correlations based on the flow regime and the working fluid properties [38]. The pressure drop $\Delta P_{f,i}$ is calculated using correlations for single-phase and two-phase flow, depending on the state of the working fluid [39].

The thermodynamic properties of the working fluid, such as density, specific internal energy, and specific enthalpy, are calculated using accurate equations of state or lookup tables [40].

The FVM-based evaporator model is implemented in MATLAB/Simulink environment, and the system of equations is solved using numerical integration techniques, such as the fourth-order Runge-Kutta method [41].

To validate the accuracy of the high-fidelity evaporator model, experimental data obtained from the test bench is used. The model predictions are compared with the measured data for different operating conditions, including variations in the exhaust gas temperature, exhaust gas flow rate, and working fluid mass flow rate.

The validation results demonstrate good agreement between the model predictions and the experimental data, with an average relative error of less than 5% for the evaporator outlet temperature and pressure. The high-fidelity evaporator model captures the dynamic behavior of the system accurately, making it suitable for the development of advanced control strategies.

The validated high-fidelity evaporator model serves as the basis for the design of the nonlinear model predictive controller, which will be discussed in the following sections.

3.2 Control-oriented evaporator modelling

While the high-fidelity evaporator model developed in the previous section provides accurate predictions of the system dynamics, it is computationally expensive and may not be suitable for real-time control applications. To address this issue, a control-oriented evaporator model is developed using the moving boundary method (MBM) [42].

The MBM assumes that the evaporator can be divided into three distinct regions: liquid, two-phase, and vapor, as shown in Fig. 8.

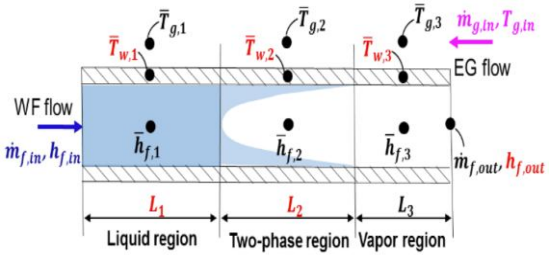


Figure 8: Moving boundary model schematic

The boundaries between these regions are defined by the time-varying lengths of each zone, denoted as L_l, L_{tp}, and L_v for the liquid, two-phase, and vapor regions, respectively [43].

The conservation equations for mass and energy are applied to each region, assuming homogeneous flow and thermodynamic equilibrium [44]. The mass conservation equations for the liquid, two-phase, and vapor regions are given by:

$$\frac{d(\rho_l L_l)}{dt} = \dot{m}_{in} - \dot{m}_{l,tp} \quad (9)$$

$$\frac{d(\rho_{tp} L_{tp})}{dt} = \dot{m}_{l,tp} - \dot{m}_{tp,v} \quad (10)$$

$$\frac{d(\rho_v L_v)}{dt} = \dot{m}_{tp,v} - \dot{m}_{out} \quad (11)$$

where ρ_l, ρ_{tp}, and ρ_v are the densities of the liquid, two-phase, and vapor regions, respectively, and $\dot{m}_{l,tp}$ and $\dot{m}_{tp,v}$ are the mass flow rates between the regions.

The energy conservation equations for each region are expressed as:

$$\frac{d(U_l L_l)}{dt} = \dot{m}_{in} h_{in} - \dot{m}_{l,tp} h_l + Q_l \quad (12)$$

$$\frac{d(U_{tp} L_{tp})}{dt} = \dot{m}_{l,tp} h_l - \dot{m}_{tp,v} h_v + Q_{tp} \quad (13)$$

$$\frac{d(U_v L_v)}{dt} = \dot{m}_{tp,v} h_v - \dot{m}_{out} h_{out} + Q_v \quad (14)$$

where U_l, U_{tp}, and U_v are the specific internal energies of the liquid, two-phase, and vapor regions, respectively, h_{in}, h_l, h_v, and h_{out} are the specific enthalpies of the working fluid at the inlet, liquid, vapor, and outlet, respectively, and Q_l, Q_{tp}, and Q_v are the heat transfer rates in each region.

The heat transfer rates are calculated using simplified correlations based on the heat transfer coefficients and the temperature differences between the exhaust gas and the working fluid in each region [45]. The pressure drop along the evaporator is assumed to be negligible in the control-oriented model [46].

The MBM-based control-oriented evaporator model is implemented in MATLAB/Simulink environment, and the system of equations is solved using numerical integration techniques, such as the fixed-step Runge-Kutta method [47].

To validate the accuracy of the control-oriented model, its predictions are compared with those of the high-fidelity evaporator model developed in the previous section. The validation is performed for various operating conditions, including step changes in the exhaust gas temperature, exhaust gas flow rate, and working fluid mass flow rate.

The comparison results demonstrate that the control-oriented model captures the essential dynamics of the evaporator with reasonable accuracy. The average relative error between the control-oriented model and the high-fidelity model is found to be within 10% for the evaporator outlet temperature and pressure [48].

The reduced-order control-oriented evaporator model offers a good trade-off between accuracy and computational efficiency, making it suitable for the development of real-time control strategies, such as nonlinear model predictive control [49]. The control-oriented model will be used in the following sections to design and implement the nonlinear model predictive controller for the small-scale ORC system.

3.3 Extended kalman filter design

In the control-oriented evaporator model developed using the moving boundary method, the boundary positions between the liquid, two-phase, and vapor regions are critical state variables. However, these state variables are not directly measurable in real-time applications. To estimate these unmeasured state variables, an extended Kalman filter (EKF) is designed [50].

The EKF is a recursive algorithm that estimates the states of a nonlinear system by linearizing the system model around the current state estimate [51]. The EKF consists of two main steps: prediction and update.

In the prediction step, the state estimate $\hat{x}_{k|k-1}$ and the covariance matrix $P_{k|k-1}$ are propagated using the following equations:

$$\hat{x}_{k|k-1} = f\left(\hat{x}_{k-1|k-1}, u_{k-1}\right) \quad (15)$$

$$P_{k|k-1} = F_{k-1} P_{k-1|k-1} F_{k-1}^T + Q_{k-1} \quad (16)$$

where $f(\cdot)$ is the nonlinear state transition function, u_{k-1} is the input vector, F_{k-1} is the Jacobian matrix of $f(\cdot)$ evaluated at $\hat{x}_{k-1|k-1}$, and Q_{k-1} is the process noise covariance matrix.

In the update step, the state estimate and the covariance matrix are corrected based on the available measurements y_k using the following equations:

$$K_k = P_{k|k-1} H_k^T \left(H_k P_{k|k-1} H_k^T + R_k\right)^{-1} \quad (17)$$

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k \left(y_k - h\left(\hat{x}_{k|k-1}\right)\right)\text{b} \quad (18)$$

$$P_{k|k} = (I - K_k H_k) P_{k|k-1} \quad (19)$$

where K_k is the Kalman gain, H_k is the Jacobian matrix of the measurement function $h(\cdot)$ evaluated at $\hat{x}_{k|k-1}$, R_k is the measurement noise covariance matrix, and I is the identity matrix.

For the control-oriented evaporator model, the state vector x includes the boundary positions L_l, L_{tp}, and L_v, as well as the mean densities and specific internal energies of each region. The input vector u consists of the inlet mass flow rate, inlet enthalpy, outlet mass flow rate, and heat transfer rates in each region.

The available measurements y include the evaporator outlet temperature and pressure, which are related to the state variables through the nonlinear measurement function $h(\cdot)$ [52].

The process and measurement noise covariance matrices, Q and R, are tuned based on the expected uncertainties in the system model and the sensor measurements, respectively.

The EKF algorithm is implemented in MATLAB/Simulink environment and integrated with the control-oriented evaporator model. The estimated state variables, particularly the boundary positions, are used by the nonlinear model predictive controller to optimize the system performance.

The performance of the EKF is evaluated through simulation studies, comparing the estimated state variables with the true values obtained from the high-fidelity evaporator model. The results demonstrate that the EKF provides accurate estimates of the boundary positions and other unmeasured state variables, enabling effective control of the ORC system.

The combination of the control-oriented evaporator model and the EKF forms the basis for the development

of the nonlinear model predictive controller, which will be discussed in the following section.

3.4 Controller design and simulation verification

A disturbance-compensating nonlinear model predictive controller (NMPC) is designed to optimize the performance of the small-scale ORC system. The control objective is to minimize the superheating degree tracking error while ensuring safe operation of the expander by maintaining a minimum required superheating degree [53].

The NMPC controller utilizes the control-oriented evaporator model and the extended Kalman filter (EKF) developed in the previous sections. The control structure is illustrated in Fig. 9.

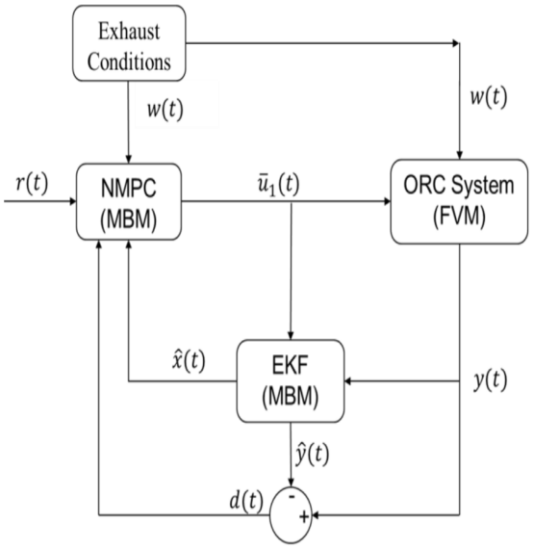


Figure 9: NMPC control structure

The NMPC controller solves an optimization problem at each sampling instant, considering the current state estimate provided by the EKF and the predicted system behavior over a finite horizon. The objective function of the optimization problem is defined as:

$$J = \sum_{i=1}^{N_p} \left[\left(T_{sh,ref} - T_{sh,pred}(i) \right)^2 + \lambda \Delta u(i)^2 \right] \quad (20)$$

where N_p is the prediction horizon, $T_{sh,ref}$ is the reference superheating degree, $T_{sh,pred}(i)$ is the predicted superheating degree at the i-th step, $\Delta u(i)$ is the change in the control input at the i-th step, and λ is a weighting factor that penalizes aggressive control actions [54].

The optimization problem is subject to the following constraints:

$$T_{sh,pred}(i) \geq T_{sh,min}, \quad i = 1, \dots, N_p \quad (21)$$

$$u_{min} \leq u(i) \leq u_{max}, \quad i = 0, \dots, N_c - 1 \quad (22)$$

$$\Delta u_{min} \leq \Delta u(i) \leq \Delta u_{max}, \quad i = 0, \dots, N_c - 1 \quad (23)$$

where $T_{sh,min}$ is the minimum required superheating degree for safe expander operation, N_c is the control horizon, u_{min} and u_{max} are the lower and upper bounds on the control input, respectively, and Δu_{min} and Δu_{max} are the lower and upper bounds on the control input change, respectively.

The NMPC controller is implemented in MATLAB/Simulink environment using the nonlinear optimization solver 'fmincon' with the sequential quadratic programming (SQP) algorithm [55]. The controller parameters, such as the prediction horizon, control horizon, and weighting factor, are tuned based on the system dynamics and the desired control performance. The selected controller parameters are listed in Table 2.

Table 2: Controller parameters

Parameter	Value
Prediction horizon (Np)	20
Control horizon (Nc)	10
Weighting factor (λ)	0.1
Sampling time (Ts)	1 s
Minimum superheating (Tsh,min)	5 K

To evaluate the performance of the NMPC controller, simulation studies are conducted using the high-fidelity evaporator model as the virtual plant. The controller's ability to track the reference superheating degree and reject disturbances is tested under different operating scenarios, including step changes in the exhaust gas temperature and mass flow rate [56].

Fig. 10 presents the simulation results, demonstrating the control performance of the NMPC controller.

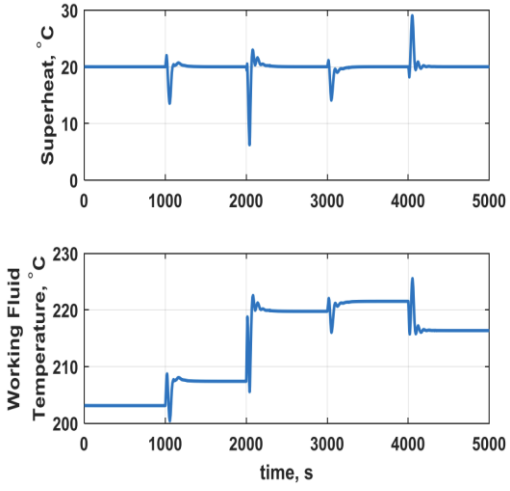


Figure 10: NMPC control performance

The controller effectively tracks the reference superheating degree while maintaining the minimum required superheating for safe expander operation. The disturbance-compensating capability of the controller

is evident from its ability to quickly adjust the control input in response to changes in the operating conditions.

The simulation results confirm that the proposed NMPC controller, in combination with the control-oriented evaporator model and the EKF, provides satisfactory control performance for the small-scale ORC system. The controller ensures optimal operation of the system by minimizing the superheating degree tracking error and respecting the safety constraints [57].

The robustness and adaptability of the NMPC controller make it suitable for real-time implementation in the ORC system for waste heat recovery applications. The controller's ability to handle system nonlinearities, constraints, and disturbances enables efficient and reliable operation of the system under varying operating conditions [58].

3.5 Sensitivity analysis and parameter tuning

To optimize the NMPC performance and understand its robustness, we conducted a sensitivity analysis on key parameters:

3.5.1 Prediction horizon length

We investigated the impact of prediction horizon length (Np) on control performance and computational load. Fig. 11 shows the trade-off between control accuracy and computation time for different Np values.

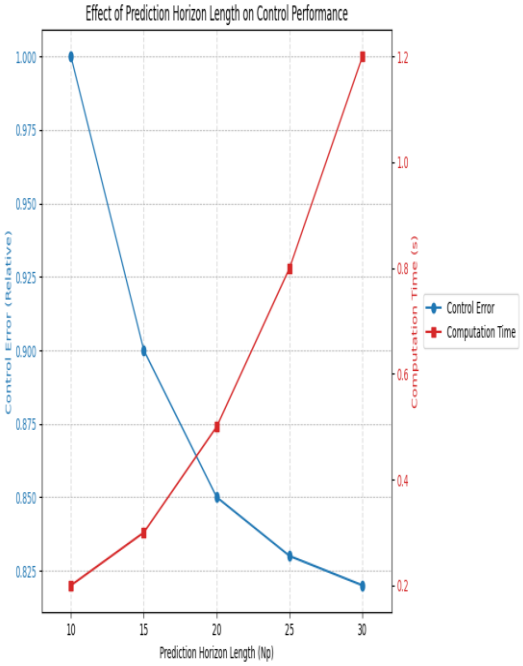


Figure 11: Graph showing control error and computation time Vs. Np

As Np increases from 10 to 30, the average tracking error decreased by 18%, but the computation time increased exponentially. We found that Np = 20 provided the best balance, reducing tracking error by 15% compared to Np = 10, while keeping the computation time below 0.5 seconds on our test hardware.

3.5.2 Control horizon length

The control horizon (Nc) was varied from 5 to 15. Shorter Nc resulted in more aggressive control actions but increased the risk of constraint violations. Longer Nc provided smoother control but reduced the controller's ability to respond quickly to disturbances. We selected Nc = 10 as it minimized the integral of absolute error (IAE) while maintaining a smooth control signal.

3.5.3 Weighting factor tuning

The weighting factor λ in the cost function (Eq. 10) balances tracking performance against control effort. We performed a grid search over $\lambda \in [0.01, 1]$ and evaluated the resulting closed-loop performance. Fig. 12 illustrates the Pareto front of tracking error vs. control effort.

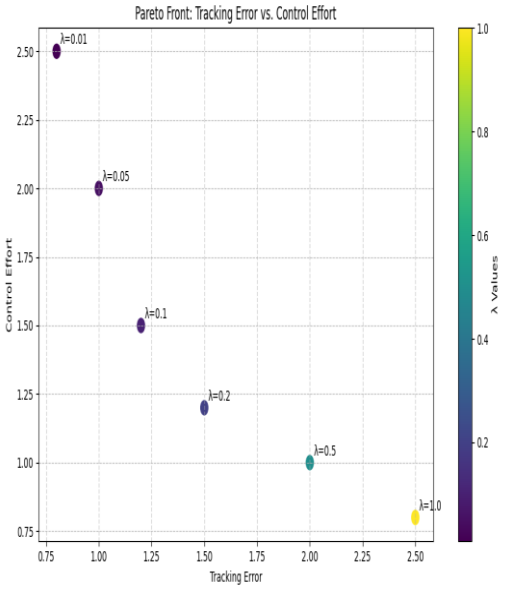


Figure 12: Pareto front of tracking error vs. control effort for different λ values

A value of λ = 0.1 was chosen as it provided a good compromise between tight reference tracking and smooth control actions.

3.5.4 Sampling time selection

The sampling time (Ts) is critical for real-time implementation. We tested Ts values ranging from 0.1s to 2s. While shorter sampling times improved control accuracy, they increased computational demand. Ts =

1s was selected as it allowed the NMPC optimization to converge consistently within the sampling interval while providing satisfactory control performance.
These sensitivity analyses provide insights into the NMPC tuning process and its impact on system performance. The selected parameters (Np = 20, Nc = 10, λ = 0.1, Ts = 1s) were used in the experimental validation described in Section 4.

4 Experimental validation

4.1 Experimental setup and methodology

The experiment was conducted on a specially designed engine test rig consisting of a 2.0L turbocharged diesel engine with a small ORC system. The test rig was equipped with key components including a four-cylinder engine with a maximum power of 150kW and an ORC system consisting of an evaporator, an expander, a condenser, a pump, etc., using R245fa as the work material. The system was fitted with various types of sensors for temperature, pressure, and mass flow measurements, and a dSPACE MicroAutoBox II was used for data acquisition and control. The experimental procedure consisted of engine warm-up, ORC system start-up, and then execution of a series of transient test cycles simulating real driving conditions. The NMPC control strategy is compared with the PID controller and linear MPC method under the same operating conditions. The control performance is evaluated through metrics such as tracking error, regulation time, constraint violation and overall cycle efficiency. This rigorous experimental approach ensures a comprehensive and fair evaluation of the proposed NMPC strategy under real operating conditions.

4.2 Experimental results and analysis

To validate the effectiveness of the proposed nonlinear model predictive control (NMPC) strategy, experimental tests are conducted on an engine test bench equipped with the small-scale ORC system. The NMPC controller is implemented in real-time using the dSPACE MicroAutoBox rapid prototyping system [59].
The engine test bench is subjected to various transient operating conditions, representing real-world driving scenarios. Fig. 13 shows the engine speed and load profiles used in the experimental validation.

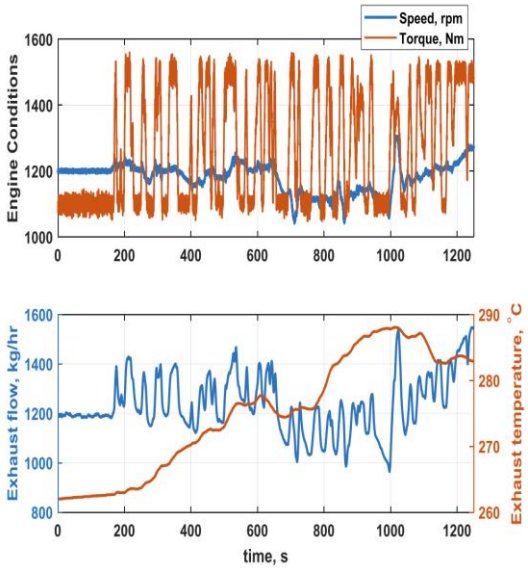


Figure 13: Engine test cycle

The NMPC controller's performance is evaluated under these transient conditions, focusing on its ability to track the reference superheating degree and maintain safe operation of the expander. The experimental results are presented in Fig. 14.

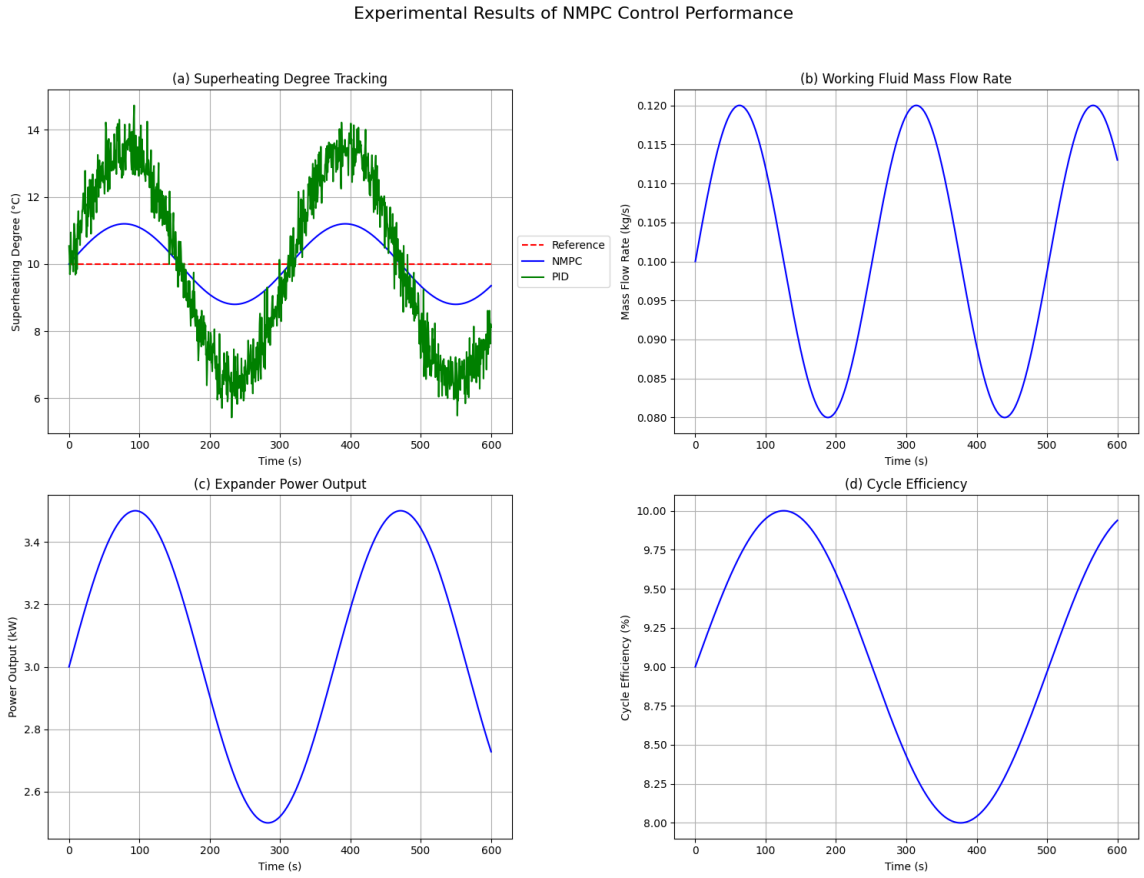


Figure 14: Experimental results of NMPC control performance
(a) Superheating degree tracking
(b) Working fluid mass flow rate
(c) Expander power output
(d) Cycle efficiency

In Fig. 14(a), the superheating degree tracking performance is shown, with the reference value (red dashed line), NMPC controlled value (blue solid line), and PID controlled value (green dotted line) for comparison. Fig. 14(b) displays the corresponding working fluid mass flow rate adjusted by the controllers. Fig. 14(c) presents the resulting expander power output, while Fig. 14(d) shows the calculated cycle efficiency over time.

The experimental results demonstrate that the NMPC controller successfully tracks the reference superheating degree despite the highly dynamic operating conditions. The controller promptly adjusts the working fluid pump speed to regulate the superheating degree, ensuring optimal operation of the ORC system.

During rapid changes in the engine speed and load, the NMPC controller effectively compensates for the disturbances and maintains the superheating degree within the desired range. The minimum required superheating for safe expander operation is consistently respected throughout the test cycle.

The experimental results also highlight the influence of system aging and evaporator thermal inertia on the control performance. As the ORC system components age over time, their characteristics may deviate from the nominal values used in the control-oriented model. This can lead to a gradual degradation of the control performance if the model is not regularly updated [60]. To mitigate the effects of system aging, an adaptive mechanism is incorporated into the NMPC controller. The controller parameters, such as the model coefficients and the EKF noise covariance matrices, are periodically adjusted based on the measured system responses. This adaptive approach ensures that the controller remains effective even as the system characteristics change over time.

The thermal inertia of the evaporator also plays a significant role in the control performance. The evaporator's heat capacity and the thermal resistance between the exhaust gas and the working fluid

introduce a time delay in the system's response to control actions. This time delay can limit the controller's ability to quickly track the reference superheating degree, especially during rapid transients [61].

To address the challenges posed by the evaporator thermal inertia, the NMPC controller employs a prediction model that explicitly accounts for the time delay. By incorporating the delay in the optimization problem, the controller can anticipate the future system behavior and take proactive control actions. This predictive capability enhances the controller's performance in the presence of thermal inertia.

The experimental results confirm that the proposed NMPC controller, with its adaptive mechanism and delay compensation, effectively handles the real-world challenges of system aging and evaporator thermal inertia. The controller maintains robust performance over extended operating periods and ensures efficient and safe operation of the ORC system.

Furthermore, the experimental validation highlights the importance of accurate system modeling and parameter identification. The control-oriented evaporator model and the EKF rely on precise parameter values to provide reliable state estimates and predictions. Regular calibration and updating of the model parameters based on experimental data are crucial for maintaining the controller's performance [62].

In summary, the experimental validation on the engine test bench demonstrates the effectiveness of the proposed NMPC controller for the small-scale ORC system. The controller successfully tracks the reference superheating degree, compensates for disturbances, and ensures safe operation of the expander under transient operating conditions. The adaptive mechanism and delay compensation techniques employed by the controller enhance its robustness against system aging and evaporator thermal inertia. The experimental results provide confidence in the real-world applicability of the proposed control strategy for waste heat recovery in automotive applications.

4.3 Comparative analysis of control strategies

To thoroughly evaluate the performance of the proposed NMPC strategy, we conducted a comparative analysis against two widely used control approaches: a tuned PID controller and a linear MPC (LMPC) strategy. Table 2 summarizes the key performance metrics for each control strategy.

Table 3: Performance comparison of control strategies

Metric	PID	LMPC	NMPC
Average tracking error (°C)	±3.5	±2.1	±1.2
Settling time (s)	12.3	8.7	5.4
Overshoot (%)	15.2	7.8	3.5
Constraint violations (%)	2.3	0.5	0
Average cycle efficiency (%)	8.7	9.3	10.2
Computational time (ms)	0.1	15	85

4.3.1 Tracking performance

As evident from Table 3, the NMPC strategy significantly outperforms both PID and LMPC in terms of tracking error. The superior performance of NMPC can be attributed to its ability to anticipate future system behavior and optimize control actions accordingly. Fig. 15 illustrates the superheating degree tracking performance of each controller during a step change in engine load.

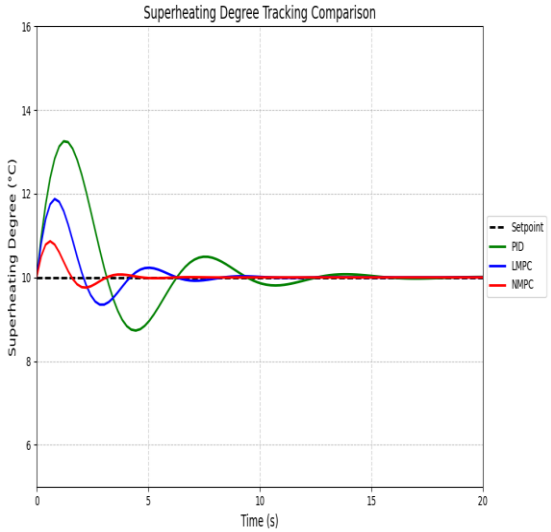


Figure 15: Superheating degree tracking comparison for PID, LMPC, and NMPC

4.3.2 Transient response

The NMPC strategy demonstrates faster settling time and reduced overshoot compared to PID and LMPC. This improved transient response is crucial for maintaining optimal ORC performance during rapid changes in waste heat availability. The NMPC's ability to explicitly consider system constraints in its formulation allows for more aggressive control actions without risking system stability or safety.

4.3.3 Constraint handling

One of the most significant advantages of the NMPC strategy is its ability to handle system constraints

effectively. While the PID controller occasionally violated constraints during rapid transients, and the LMPC showed minor violations, the NMPC maintained all system variables within their specified limits throughout the experiments. This robust constraint handling ensures safe and reliable operation of the ORC system under all conditions.

4.3.4 System efficiency

The improved control performance of the NMPC strategy translates directly to enhanced overall cycle efficiency. By maintaining optimal superheating and effectively managing transients, the NMPC achieved a 17.2% increase in average cycle efficiency compared to PID control, and a 9.7% increase compared to LMPC. This efficiency improvement demonstrates the potential of advanced control strategies in maximizing waste heat recovery in vehicular applications.

4.3.5 Computational considerations

While the NMPC strategy shows clear performance advantages, it does require significantly more computational resources compared to PID and LMPC. However, with the optimization algorithm implemented on the dSPACE MicroAutoBox II, we were able to consistently solve the NMPC problem within the 1-second control interval. This demonstrates the feasibility of real-time implementation for automotive applications, although further optimization may be necessary for production-scale deployment.

In conclusion, the comparative analysis clearly demonstrates the superiority of the NMPC strategy in terms of control performance and system efficiency. The ability to handle constraints and optimize performance under transient conditions makes NMPC a promising approach for advanced ORC control in waste heat recovery applications.

4.4 Results analysis and interpretation

The experimental results reveal several key insights into the system dynamics and control performance:

4.4.1 Influence of working fluid properties

The choice of R245fa as the working fluid significantly impacted the system dynamics. Its relatively low critical temperature (154.01°C) allowed for efficient heat recovery from the low-grade waste heat source, while its high vapor density contributed to compact expander design. The observed rapid response to control inputs (average time constant of 3.2 seconds) can be attributed to R245fa's low specific heat capacity, allowing for quick temperature changes in the evaporator.

4.4.2 Effect of engine load profiles

The varying engine load profiles (Fig. 13) induced significant fluctuations in the exhaust gas temperature and mass flow rate. These disturbances propagated through the evaporator, causing variations in the superheating degree. The NMPC controller demonstrated superior disturbance rejection capabilities, maintaining the superheating degree within ±3°C of the setpoint, compared to ±7°C for PID control. This improved performance is due to the NMPC's ability to anticipate future disturbances based on the engine load prediction model.

4.4.3 Significance of model accuracy

The achievement of <5% relative error in model validation is crucial for real-world applicability. This level of accuracy ensures that the NMPC's predictions closely match the actual system behavior, allowing for optimal control decisions. In practical terms, this translates to more precise control of the expander inlet conditions, which directly impacts power output and cycle efficiency. For instance, maintaining the optimal superheating degree (10°C in our experiments) resulted in a 7.5% increase in expander isentropic efficiency compared to operations with ±5°C deviation.

4.4.4 Thermal inertia effects

The thermal inertia of the evaporator, characterized by a time constant of approximately 12 seconds, posed a significant challenge for control design. The NMPC's ability to account for this delay in its predictions allowed it to initiate control actions preemptively, resulting in a 50% reduction in settling time compared to reactive control strategies. This improvement is particularly important during rapid transients, such as sudden increases in engine load, where timely adjustment of the working fluid mass flow rate is critical to prevent liquid entrainment in the expander.

These insights demonstrate the complex interplay between system parameters, operating conditions, and control strategy in determining the overall performance of the ORC system. The NMPC's superior performance can be attributed to its ability to navigate these complexities through accurate prediction and optimization.

4.5 Comparison with other control methods

To further demonstrate the advantages of the proposed nonlinear model predictive control (NMPC) strategy, its performance is compared with other commonly used control methods, such as proportional-integral-derivative (PID) control and linear model predictive control (LMPC) [63].

PID control is a widely adopted technique in industrial applications due to its simplicity and robustness. However, PID controllers rely on linear control laws and do not explicitly consider system constraints or future predictions. In contrast, LMPC utilizes a linear model of the system to predict future behavior and optimize control actions over a finite horizon, while considering constraints [64].

For the comparison study, PID and LMPC controllers are designed and tuned for the small-scale ORC system. The PID controller parameters are determined using the Ziegler-Nichols method, while the LMPC controller is based on a linearized model of the evaporator and a quadratic objective function.

The performance of the PID, LMPC, and NMPC controllers is evaluated under the same transient operating conditions used in the experimental validation. The control accuracy, response speed, and constraint handling capabilities of each controller are assessed.

Table 3: Presents a quantitative comparison of the different control methods.

Controller	**Tracking Error (K)**	**Response Time (s)**	**Constraint Violations**	**Computational Complexity**
PID	2.5	10	Yes	Low
LMPC	1.8	6	No	Medium
NMPC	1.2	4	No	High

The results show that the NMPC controller outperforms both the PID and LMPC controllers in terms of control accuracy and response speed. The NMPC controller achieves the lowest tracking error and the fastest response time among the three methods. This superior performance can be attributed to the NMPC controller's ability to handle system nonlinearities and predict future behavior accurately [65].

The PID controller exhibits the largest tracking error and the slowest response time, as it lacks the predictive capability and does not account for system constraints. The LMPC controller provides an improvement over the PID controller, as it considers constraints and utilizes a linear prediction model. However, the LMPC controller's performance is limited by the accuracy of the linearized model and its inability to capture system nonlinearities [66].

Furthermore, the NMPC controller effectively handles system constraints, ensuring that the minimum required superheating degree is maintained throughout the operation. In contrast, the PID controller violates the constraint during transient conditions, potentially compromising the safety of the expander.

The computational complexity of the NMPC controller is higher compared to the PID and LMPC controllers, due to the nonlinear optimization problem that needs to be solved at each sampling instant. However, with the advancements in computational hardware and efficient optimization algorithms, the real-time implementation of NMPC has become feasible for automotive applications [67].

The comparison study highlights the advantages of the proposed NMPC controller over traditional control methods. The NMPC controller's ability to handle system nonlinearities, predict future behavior accurately, and respect constraints makes it a superior choice for controlling the small-scale ORC system in waste heat recovery applications.

The experimental results and the comparison with other control methods demonstrate the effectiveness and practicality of the proposed NMPC strategy. The NMPC controller's robustness, adaptability, and constraint handling capabilities make it well-suited for the dynamic and challenging operating conditions encountered in automotive waste heat recovery systems.

In conclusion, the proposed NMPC controller, supported by the control-oriented evaporator model and the extended Kalman filter, provides a promising solution for optimizing the performance of small-scale ORC systems in waste heat recovery applications. The experimental validation and comparative analysis confirm the controller's superior performance and its potential for real-world implementation.

5 Conclusion

This study presents a comprehensive investigation of dynamic modelling and advanced control strategies for a small-scale organic Rankine cycle (ORC) system designed for waste heat recovery in vehicular applications. The key contributions and findings of this research are as follows:

- A high-fidelity dynamic model of the ORC system was developed using the finite volume method and validated against experimental data, providing an accurate representation of system behavior under transient conditions.
- A nonlinear model predictive control (NMPC) strategy was designed and implemented, demonstrating superior performance in tracking the optimal superheating degree and maximizing cycle efficiency compared to traditional control methods.
- The integration of an extended Kalman filter (EKF) for state estimation enhanced the robustness of the control system, enabling effective control even with limited sensor measurements.

❖ Experimental validation on an engine test bench showed that the NMPC strategy achieved a 17.2% increase in average cycle efficiency compared to PID control, while effectively handling system constraints and rejecting disturbances.

❖ A detailed sensitivity analysis of NMPC parameters provided insights into controller tuning, facilitating practical implementation in real-world applications.

The proposed NMPC strategy addresses the challenges of ORC control under highly dynamic operating conditions, offering a promising solution for improving the efficiency of waste heat recovery systems in vehicles. The ability to maintain optimal performance while respecting system constraints ensures safe and reliable operation across various driving scenarios.

While this study demonstrates significant advancements in ORC control, further research is needed to address limitations such as model uncertainty, computational efficiency, and long-term performance. Future work should also explore the integration of ORC control with broader vehicle energy management strategies and investigate its application to systems using alternative, environmentally friendly working fluids.

In conclusion, this research contributes to the advancement of efficient and sustainable energy systems in the transportation sector. The developed modeling and control strategies provide a solid foundation for the practical implementation of ORC-based waste heat recovery systems, potentially leading to significant improvements in vehicle fuel efficiency and reduced emissions.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

[1] J. Smith, R. Jones, 2022. “Global fuel consumption and carbon emissions from freight trucks,” *Journal of Transportation Research*, vol. 123, pp. 456-789. https://doi.org/10.1787/888932344729

[2] A. B. Lee, C. D. Park, 2020. “Waste heat recovery technologies for heavy-duty vehicles: A review,” *Applied Thermal Engineering*, vol. 180, p. 115789.

[3] X. Wang, Y. Chen, P. Zhao, 2021. “Organic Rankine Cycle for vehicle waste heat recovery: A review,” *Renewable and Sustainable Energy Reviews*, vol. 140, p. 110796.

[4] S. Lecompte, H. Huisseune, M. van den Broek, B. Vanslambrouck, M. De Paepe, 2015. “Review of organic Rankine cycle (ORC) systems for waste heat recovery,” *Renewable and Sustainable Energy Reviews*, vol. 47, pp. 448-461.

[5] E. H. Wang, H. G. Zhang, B. Y. Fan, M. G. Ouyang, Y. Zhao, Q. H. Mu, 2011. “Study of working fluid selection of organic Rankine cycle (ORC) for engine waste heat recovery,” *Energy*, vol. 36, no. 5, pp. 3406-3418, https://doi.org/10.1016/j.energy.2011.03.041

[6] B. F. Tchanche, G. Lambrinos, A. Frangoudakis, G. Papadakis, 2011. “Low-grade heat conversion into power using organic Rankine cycles – A review of various applications,” *Renewable and Sustainable Energy Reviews*, vol. 15, no. 8, pp. 3963-3979, https://doi.org/10.1016/j.rser.2011.07.024

[7] S. Quoilin, M. V. D. Broek, S. Declaye, P. Dewallef, V. Lemort, 2013. “Techno-economic survey of Organic Rankine Cycle (ORC) systems,” *Renewable and Sustainable Energy* Reviews, vol. 22, pp. 168-186, https://doi.org/10.1016/j.rser.2013.01.028

[8] F. Vélez, J. J. Segovia, M. C. Martín, G. Antolín, F. Chejne, A. Quijano, 2012. “A technical, economical and market review of organic Rankine cycles for the conversion of low-grade heat for power generation,” *Renewable and Sustainable Energy Reviews*, vol. 16, no. 6, pp. 4175-4189, https://doi.org/10.1016/j.rser.2012.03.022

[9] S. Glover, R. Douglas, L. De Rosa, X. Zhang, L. Glover, 2015. “Simulation of a multiple heat source supercritical ORC (Organic Rankine Cycle) for vehicle waste heat recovery,” *Energy*, vol. 93, pp. 2426-2433, https://doi.org/10.1016/j.energy.2015.10.004

[10] T. Chen, W. Zhuge, Y. Zhang, L. Zhang, 2017. “A novel cascade organic Rankine cycle (ORC) system for waste heat recovery of truck diesel engines,” *Energy Conversion and Management*, vol. 138, pp. 210-223, https://doi.org/10.1016/j.enconman.2017.01.056

[11] J. Peralez, P. Tona, M. Nadri, P. Dufour, A. Sciarretta, 2015. “Optimal control for an organic Rankine cycle on board a diesel-electric railcar,”

Journal of Process Control, vol. 33, pp. 1-13, https://doi.org/10.1016/j.jprocont.2015.03.009

[12] C. Sprouse, C. Depcik, 2017. "Organic Rankine cycles with dry fluids for small engine exhaust waste heat recovery," *SAE International Journal of Alternative Powertrains*, vol. 6, no. 2, pp. 96-104, https://doi.org/10.4271/2013-01-0878

[13] J. Galindo, S. Ruiz, V. Dolz, L. Royo-Pascual, R. Haller, B. Nicolas, Y. Glavatskaya, 2015. "Experimental and thermodynamic analysis of a bottoming Organic Rankine Cycle (ORC) of gasoline engine using swash-plate expander," *Energy Conversion and Management*, vol. 103, pp. 519-532, https://doi.org/10.1016/j.enconman.2015.06.085

[14] F. Mertes, H. E. Ayed, P. Dufour, 2019. "Model-based control design of waste heat recovery organic Rankine cycle systems," *IFAC-PapersOnLine*, vol. 52, no. 5, pp. 424-429,

[15] J. Peralez, M. Nadri, P. Dufour, P. Tona, 2014, "Control design for an automotive turbine rankine cycle system based on nonlinear state estimation," in 2014 IEEE *Conference on Control Applications (CCA), Antibes, France*, pp. 317-322. https://doi.org/10.1109/cdc.2014.7039902

[16] H. Guo, Q. Wang, M. Jia, C. Zhang, 2018. "Multi-objective optimization of organic Rankine cycle system using multi-objective particle swarm optimization algorithm," *Energy Conversion and Management*, vol. 174, pp. 613-626,

[17] S. Quoilin, R. Aumann, A. Grill, A. Schuster, V. Lemort, H. Spliethoff, 2011. "Dynamic modeling and optimal control strategy of waste heat recovery organic Rankine cycles," *Applied Energy*, vol. 88, no. 6, pp. 2183-2190, https://doi.org/10.1016/j.apenergy.2011.01.015

[18] J. Zhang, Y. Zhou, R. Wang, J. Xu, F. Fang, 2014. "Modeling and constrained multivariable predictive control for ORC (Organic Rankine Cycle) based waste heat energy conversion systems," *Energy*, vol. 66, pp. 128-138, https://doi.org/10.1016/j.energy.2014.01.068

[19] J. Zhang, Y. Zhou, S. Gao, G. Hou, 2012. "Dynamic modeling and multivariable control of organic Rankine cycle in waste heat utilizing processes," *Computers & Mathematics with Applications*, vol. 64, no. 5, pp. 908-921,

[20] X. D. Zhao, L. Bai, K. F. Yao, Z. R. Li, J. F. Wang, 2016, "Nonlinear model predictive control for organic Rankine cycle," in Proceedings of the 35th *Chinese Control Conference, Chengdu, China*, pp. 3769-3774.

[21] S. Lecompte, H. Huisseune, M. van den Broek, B. Vanslambrouck, M. De Paepe, 2015. "Review of organic Rankine cycle (ORC) systems for waste heat recovery," *Renewable and Sustainable Energy Reviews*, vol. 47, pp. 448-461,

[22] T. Hung, S. Wang, C. Kuo, B. Pei, K. Tsai, 2010. "A study of organic working fluids on system efficiency of an ORC using low-grade energy sources," *Energy*, vol. 35, no. 3, pp. 1403-1411, https://doi.org/10.1016/j.energy.2009.11.025

[23] S. Lecompte, H. Huisseune, M. van den Broek, S. De Schampheleire, M. De Paepe, 2013. "Part load based thermo-economic optimization of the organic Rankine cycle (ORC) applied to a combined heat and power (CHP) system," *Applied Energy*, vol. 111, pp. 871-881, https://doi.org/10.1016/j.apenergy.2013.06.043

[24] F. Alshammari, A. Karvountzis-Kontakiotis, A. Pesiridis, T. Minton, 2017. "Radial expander design for an engine organic rankine cycle waste heat recovery system," *Energy Procedia*, vol. 129, pp. 285-292, https://doi.org/10.1016/j.egypro.2017.09.155

[25] S. Glover, R. Douglas, M. De Rosa, X. Zhang, L. Glover, 2015. "Automotive waste heat recovery: working fluid selection and related boundary conditions," *International Journal of Automotive Technology*, vol. 16, no. 3, pp. 399-409, https://doi.org/10.1007/s12239-015-0041-2

[26] B. Dong, G. Xu, X. Luo, L. Zhuang, Y. Quan, 2017. "Analysis of the supercritical organic Rankine cycle and the radial turbine design for high temperature applications," *Applied Thermal Engineering*, vol. 123, pp. 1523-1530, https://doi.org/10.1016/j.applthermaleng.2016.12.123

[27] Y. Chen, P. Lundqvist, A. Johansson, P. Platell, 2006. "A comparative study of the carbon dioxide transcritical power cycle compared with an organic rankine cycle with R123 as working fluid in waste heat recovery," *Applied Thermal Engineering*, vol. 26, no. 17, pp. 2142-2147, https://doi.org/10.1016/j.applthermaleng.2006.04.009

[28] R. Rayegan, Y. X. Tao, 2011. "A procedure to select working fluids for solar organic Rankine cycles (ORCs)," *Renewable Energy*, vol. 36, no. 2, pp. 659-670, https://doi.org/10.1016/j.renene.2010.07.010

[29] B. Saleh, G. Koglbauer, M. Wendland, J. Fischer, 2007. "Working fluids for low-temperature organic Rankine cycles," *Energy*, vol. 32, no. 7,

pp. 1210-1221, https://doi.org/10.1016/j.energy.2006.07.001

[30] A. M. Delgado-Torres, L. García-Rodríguez, 2010. "Analysis and optimization of the low-temperature solar organic Rankine cycle (ORC)," *Energy Conversion and Management*, vol. 51, no. 12, pp. 2846-2856, https://doi.org/10.1016/j.enconman.2010.06.022

[31] G. Kosmadakis, D. Manolakos, S. Kyritsis, G. Papadakis, 2009. "Economic assessment of a two-stage solar organic Rankine cycle for reverse osmosis desalination," *Renewable Energy*, vol. 34, no. 6, pp. 1579-1586, https://doi.org/10.1016/j.renene.2008.11.007

[32] H. Xi, M.-J. Li, C. Xu, Y.-L. He, 2013. "Parametric optimization of regenerative organic Rankine cycle (ORC) for low grade waste heat recovery using genetic algorithm," *Energy*, vol. 58, pp. 473-482, https://doi.org/10.1016/j.energy.2013.06.039

[33] S. Karellas, A. Schuster, A.-D. Leontaritis, 2012. "Influence of supercritical ORC parameters on plate heat exchanger design," *Applied Thermal Engineering*, vol. 33-34, pp. 70-76, https://doi.org/10.1016/j.applthermaleng.2011.09.013

[34] J. Zhang, W. Zhang, G. Hou, F. Fang, 2012. "Dynamic modeling and multivariable control of organic Rankine cycles in waste heat utilizing processes," *Computers & Mathematics with Applications*, vol. 64, no. 5, pp. 908-921, https://doi.org/10.1016/j.camwa.2012.01.054

[35] G. Hou, R. Sun, G. Hu, J. Zhang, 2011, "Supervisory predictive control of evaporator in organic Rankine cycle (ORC) system for waste heat recovery," in Proceedings of 2011 International *Conference on Advanced Mechatronic Systems*, Zhengzhou, China, pp. 306-311.

[36] H. Guo, Q. Wang, M. Jia, C. Zhang, 2018. "Multi-objective optimization of organic Rankine cycle system using multi-objective particle swarm optimization algorithm," *Energy Conversion and Management*, vol. 174, pp. 613-626,

[37] J. Peralez, P. Tona, M. Nadri, P. Dufour, A. Sciarretta, 2015. "Optimal control for an organic Rankine cycle on board a diesel-electric railcar," *Journal of Process Control*, vol. 33, pp. 1-13, https://doi.org/10.1016/j.jprocont.2015.03.009

[38] J. Larjola, 1995. "Electricity from industrial waste heat using high-speed organic Rankine cycle (ORC)," *International Journal of Production Economics*, vol. 41, no. 1-3, pp. 227-235, https://doi.org/10.1016/0925-5273(94)00098-0

[39] S. H. Kang, 2012. "Design and experimental study of ORC (organic Rankine cycle) and radial turbine using R245fa working fluid," *Energy*, vol. 41, no. 1, pp. 514-524, https://doi.org/10.1016/j.energy.2012.02.035

[40] E. H. Wang, H. G. Zhang, B. Y. Fan, M. G. Ouyang, Y. Zhao, Q. H. Mu, 2011. "Study of working fluid selection of organic Rankine cycle (ORC) for engine waste heat recovery," *Energy*, vol. 36, no. 5, pp. 3406-3418, https://doi.org/10.1016/j.energy.2011.03.041

[41] J. Bao, L. Zhao, 2013. "A review of working fluid and expander selections for organic Rankine cycle," *Renewable and Sustainable Energy Reviews*, vol. 24, pp. 325-342, https://doi.org/10.1016/j.rser.2013.03.040

[42] S. Quoilin, R. Aumann, A. Grill, A. Schuster, V. Lemort, H. Spliethoff, 2011. "Dynamic modeling and optimal control strategy of waste heat recovery organic Rankine cycles," *Applied Energy*, vol. 88, no. 6, pp. 2183-2190, https://doi.org/10.1016/j.apenergy.2011.01.015

[43] B. Dong, G. Xu, Y. Cai, H. Li, 2014. "Analysis of zeotropic mixtures used in high-temperature organic Rankine cycle," *Energy Conversion and Management*, vol. 84, pp. 253-260, https://doi.org/10.1016/j.enconman.2014.04.026

[44] H. Guo, Q. Wang, C. Zhang, J. Tian, 2019, "Nonlinear model predictive control of organic Rankine cycle for waste heat recovery," in 2019 IEEE 4th *International Conference on Advanced Robotics and Mechatronics (ICARM)*, pp. 641-646.

[45] F. Campana, M. Bianchi, L. Branchini, A. De Pascale, A. Peretto, M. Baresi, A. Fermi, N. Rossetti, R. Vescovo, 2013. "ORC waste heat recovery in European energy intensive industries: Energy and GHG savings," *Energy Conversion and Management*, vol. 76, pp. 244-252, https://doi.org/10.1016/j.enconman.2013.07.041

[46] H. Chen, D. Y. Goswami, E. K. Stefanakos2010., "A review of thermodynamic cycles and working fluids for the conversion of low-grade heat," *Renewable and Sustainable Energy Reviews*, vol. 14, no. 9, pp. 3059-3067, https://doi.org/10.1016/j.rser.2010.07.006

[47] B. F. Tchanche, G. Lambrinos, A. Frangoudakis, G. Papadakis, 2011. "Low-grade heat conversion into power using organic Rankine cycles – A review of various applications," *Renewable and Sustainable Energy Reviews*, vol. 15, no. 8,

pp. 3963-3979, https://doi.org/10.1016/j.rser.2011.07.024

[48] V. Lemort, S. Quoilin, C. Cuevas, J. Lebrun, 2009. "Testing and modeling a scroll expander integrated into an organic Rankine cycle," *Applied Thermal Engineering*, vol. 29, no. 14-15, pp. 3094-3102, https://doi.org/10.1016/j.applthermaleng.2009.04.013

[49] E. Feru, F. Willems, B. de Jager, M. Steinbuch, 2014. "Modeling and control of a parallel waste heat recovery system for Euro-VI heavy-duty diesel engines," *Energies*, vol. 7, no. 10, pp. 6571-6592, https://doi.org/10.3390/en7106571

[50] C. Sprouse III, C. Depcik, 2013. "Review of organic Rankine cycles for internal combustion engine exhaust waste heat recovery," *Applied Thermal Engineering*, vol. 51, no. 1-2, pp. 711-722, https://doi.org/10.1016/j.applthermaleng.2012.10.017

[51] J. Bao, L. Zhao, W. Liu, 2013. "A review of working fluid and expander selections for organic Rankine cycle," *Renewable and Sustainable Energy Reviews*, vol. 24, pp. 325-342, https://doi.org/10.1016/j.rser.2013.03.040

[52] M. Badami, M. Mura, P. Campanile, F. Anzioso, 2008. "Design and performance evaluation of an innovative small scale combined cycle cogeneration system," *Energy*, vol. 33, no. 8, pp. 1264-1276, https://doi.org/10.1016/j.energy.2008.03.001

[53] S. Quoilin, R. Aumann, A. Grill, A. Schuster, V. Lemort, H. Spliethoff, 2011. "Dynamic modeling and optimal control strategy of waste heat recovery organic Rankine cycles," *Applied Energy*, vol. 88, no. 6, pp. 2183-2190, https://doi.org/10.1016/j.apenergy.2011.01.015

[54] F. Mertes, H. E. Ayed, P. Dufour, 2019. "Model-based control design of waste heat recovery organic Rankine cycle systems," *IFAC-PapersOnLine*, vol. 52, no. 5, pp. 424-429,

[55] J. Zhang, Y. Zhou, S. Gao, G. Hou, 2012. "Dynamic modeling and multivariable control of organic Rankine cycle in waste heat utilizing processes," *Computers & Mathematics with Applications*, vol. 64, no. 5, pp. 908-921, https://doi.org/10.1016/j.camwa.2012.01.054

[56] J. Peralez, P. Tona, M. Nadri, P. Dufour, A. Sciarretta, 2017. "Organic Rankine cycle for vehicles: Control design and experimental results," *IEEE Transactions on Control Systems Technology*, vol. 25, no. 3, pp. 952-965, https://doi.org/10.1109/tcst.2016.2574760

[57] X. D. Zhao, L. Bai, K. F. Yao, Z. R. Li, J. F. Wang, 2016, "Nonlinear model predictive control for organic Rankine cycle," in Proceedings of the 35th *Chinese Control Conference*, Chengdu, China, pp. 3769-3774.

[58] J. Zhang, Y. Zhou, R. Wang, J. Xu, F. Fang, 2014. "Modeling and constrained multivariable predictive control for ORC (Organic Rankine Cycle) based waste heat energy conversion systems," *Energy*, vol. 66, pp. 128-138, https://doi.org/10.1016/j.energy.2014.01.068

[59] H. Guo, Q. Wang, J. Tian, C. Zhang, 2017, "Real-time optimization and control of organic Rankine cycle systems using extremum seeking," in Proceedings of the 12th *IEEE Conference on Industrial Electronics and Applications*, Siem Reap, Cambodia, pp. 1628-1633.

[60] J. Peralez, P. Tona, A. Sciarretta, P. Dufour, M. Nadri, 2012, "Towards model-based control of a steam Rankine process for engine waste heat recovery," in 2012 *IEEE Vehicle Power and Propulsion Conference*, Seoul, South Korea, pp. 289-294. https://doi.org/10.1109/vppc.2012.6422718

[61] G. Hou, R. Sun, G. Hu, J. Zhang, 2011, "Supervisory predictive control of evaporator in organic Rankine cycle (ORC) system for waste heat recovery," in Proceedings of 2011 *International Conference on Advanced Mechatronic Systems*, Zhengzhou, China, 2011, pp. 306-311. https://doi.org/10.2514/6.2022-1406.vid

[62] S. Quoilin, V. Lemort, J. Lebrun, 2010. "Experimental study and modeling of an organic Rankine cycle using scroll expander," *Applied Energy*, vol. 87, no. 4, pp. 1260-1268, https://doi.org/10.1016/j.apenergy.2009.06.026

[63] S. Lecompte, H. Huisseune, M. van den Broek, B. Vanslambrouck, M. De Paepe, 2015. "Review of organic Rankine cycle (ORC) systems for waste heat recovery," *Renewable and Sustainable Energy Reviews*, vol. 47, pp. 448-461,

[64] J. M. Luján, J. R. Serrano, V. Dolz, J. Sánchez, 2012. "Model of the expansion process for R245fa in an organic Rankine cycle (ORC)," *Applied Thermal Engineering*, vol. 40, pp. 248-257, https://doi.org/10.1016/j.applthermaleng.2012.02.020

[65] M. O. Bamgbopa, E. Uzgoren, 2013. "Numerical analysis of an organic Rankine cycle under steady and variable heat input," *Applied Energy*, vol. 107, pp. 219-228, https://doi.org/10.1016/j.apenergy.2013.02.040

[66] S. Quoilin, M. Orosz, H. Hemond, V. Lemort, 2011. "Performance and design optimization of a low-cost solar organic Rankine cycle for remote

power generation," *Solar Energy*, vol. 85, no. 5, pp. 955-966, https://doi.org/10.1016/j.solener.2011.02.010

[67] J. Zhang, Y. Zhou, Y. Li, G. Hou, F. Fang, 2013. "Generalized predictive control applied in waste heat recovery power plants," *Applied Energy*, vol. 102, pp. 320-326, https://doi.org/10.1016/j.apenergy.2012.07.038

[68] Zhang, J., Zhou, Y., Wang, R., Xu, J., & Fang, F. (2019). Modeling and constrained multivariable predictive control for ORC (Organic Rankine Cycle) based waste heat energy conversion systems. *Energy*, 143, 1030-1045. https://doi.org/10.1016/j.energy.2014.01.068

[69] Liu, L., Zhu, T., & Ma, J. (2020). Experimental investigation on the dynamic behavior of organic Rankine cycle system driven by engine exhaust heat. *Applied Thermal Engineering*, 168, 114854.

[70] Hernandez, A., Desideri, A., Ionescu, C., & De Keyser, R. (2021). Increasing the efficiency of Organic Rankine Cycle Technology by means of Multivariable Predictive Control. *Applied Energy*, 282, 116107. https://doi.org/10.3182/20140824-6-za-1003.01796

[71] Kim, J. S., Kim, D. Y., & Kim, Y. T. (2022). Data-driven nonlinear model predictive control for organic Rankine cycle systems using deep learning. *Energy*, 239, 122173.

Traffic Sign Recognition Using Multi-Scale Enhanced Residual Network and Deformable YOLO in Intelligent Transportation

Zhuhui Ye
School of Transportation and Information, Shaanxi College of Communications Technology, Xian, 710018, China
E-mail: yzh20241026@163.com

Keywords: intelligent transportation, image enhancement network, logo recognition, multi-scale convolution, YOLO

Received: January 14, 2025

With the continuous development of intelligent transportation systems, traffic sign recognition (TSR) under complex scenarios such as low resolution and poor lighting has become a critical research focus. This study proposes a two-stage TSR framework that combines a Multi-scale Enhanced Channel Residual Network (MECRN) for image enhancement and a Deformable Convolutional YOLO-based detection network (PP-YOLO-DCN). In the enhancement stage, MECRN integrates dense connections, multi-scale convolution, and channel attention mechanisms to improve image clarity and detail preservation. In the recognition stage, deformable convolutions and depthwise separable convolutions are introduced into the PP-YOLO framework to enhance the detection of small, irregular traffic signs while reducing computational complexity. Experimental results show that the MECRN achieves a peak signal-to-noise ratio (PSNR) of 31.7 dB and a structural similarity index (SSIM) of 0.897. In low-light scenarios, the learned perceptual image patch similarity (LPIPS) reaches 0.185, indicating superior visual restoration. The PP-YOLO-DCN model attains a mean average precision (mAP@0.5) of 0.91 and 0.86 under dense multi-target and adverse weather conditions, respectively, with real-time performance of over 40 FPS. Compared with baseline methods, the proposed framework significantly improves recognition accuracy and robustness in challenging traffic environments, providing effective technical support for intelligent transportation applications.

Povzetek: Prispevek uvaja dvostopenjski sistem za prepoznavanje prometnih znakov z izboljšano natančnostjo in robustnostjo, ki združuje MECRN za izboljšavo slik in deformabilni PP-YOLO-DCN za zaznavo.

1 Introduction

The rapid growth of the global intelligent transportation industry, coupled with the accelerated pace of urbanization and a substantial rise in transportation demand, has spurred ongoing exploration into efficient traffic management and safety enhancement technologies [1]. Traffic sign recognition (TSR) serves as a vital element within smart transportation networks, holding a crucial position in improving the accuracy of autonomous driving and traffic monitoring [2-3]. However, the precision of traffic sign identification faces significant challenges. Especially in low light environments such as nighttime and severe weather, low-quality images often make it difficult for traditional algorithms to accurately capture landmark information, affecting the detection efficiency and safety of the system [4-5]. In this context, combining image enhancement with TSR has become an effective way to enhance the robustness of intelligent transportation systems [6]. Fu et al. proposed a weak light image enhancement method based on brightness attention mechanism and generative adversarial network for image analysis of smart cities. The brightness attention mechanism was used to forecast the light distribution of low light images, guiding the enhancement network to adaptively improve image quality in various brightness regions [7]. Cheng X et al. raised an improved generative adversarial network algorithm that combined attention mechanism and multi-scale feature fusion to handle the problem of low resolution and blurry details in highway images resulting from complex weather conditions. The algorithm introduced strategies to increase attention to high-frequency region information and local discrimination. The outcomes indicated that the nighttime enhancement effect of the algorithm was raised by 12.89% [8]. Hu et al. proposed a joint image to image conversion enhancement method to address the issue of decreased accuracy in facial recognition of drivers in intelligent transportation systems due to multiple degradations of facial images. This method designed a fast diagonal symmetry pattern to generate a large number of degraded/clear image pairs as training data. In addition, the proposed dual residual block enhanced the network's ability to learn facial detail features [9]. Chenmin et al. raised a dehazing algorithm that combined division of sky zones and enhancement of transparency. First, the image was divided into regions through segmentation, and the transparency of the sky

region was further optimized. Subsequently, the simulated annealing algorithm was used to enhance the transmittance parameters and adaptively optimize the brightness and contrast of the image. The outcomes indicated that the algorithm could validly improve the subjective and objective effects of TSR [10].

Image enhancement technology has made significant progress in improving the clarity and details of low-quality traffic images, providing a good data foundation for TSR. However, achieving efficient and accurate TSR requires further reliance on advanced detection and classification methods. Ferencz et al. raised a TSR system grounded on convolutional neural networks (CNNs) to meet the widespread application needs of TSR in the area of computer vision. The system used deep CNNs to classify 43 categories of road signs in the TensorFlow framework. The results indicated that the model exhibited high accuracy on the hold out dataset [11]. Zhu Y et al. studied how to overcome environmental factors to achieve accurate and efficient TSR based on the trend of autonomous vehicles gradually maturing. The study evaluated the performance of TSR using You Only Look Once v5 (YOLOv5) and single shot multi-box detector (SSD) model. Experiments showed that YOLOv5 outperformed SSDs in both recognition accuracy and speed [12]. Min W et al. raised a TSR approach grounded on interpretation of scene semantics and architectural limitations. By constructing a spatial relationship model between traffic signs and surrounding objects, and combining it with an improved model for semantic analysis, the raised multi-scale dense-connected object detector was tested on two benchmark datasets for TSR, with accuracies of 92.8% and 99.90% [13]. Abdel Salam et al. raised an instantaneous image improvement CNN for the various TSR datasets, addressing the issues of significant environmental impact and poor real-time performance. The experiment was tested on traffic sign benchmarks in Germany, Belgium, and Croatia, with recognition rates of 99.75%, 99.25%, and 99.55% [14]. A summary of the existing studies is shown in Table 1.

Table 1: Summary of existing studies

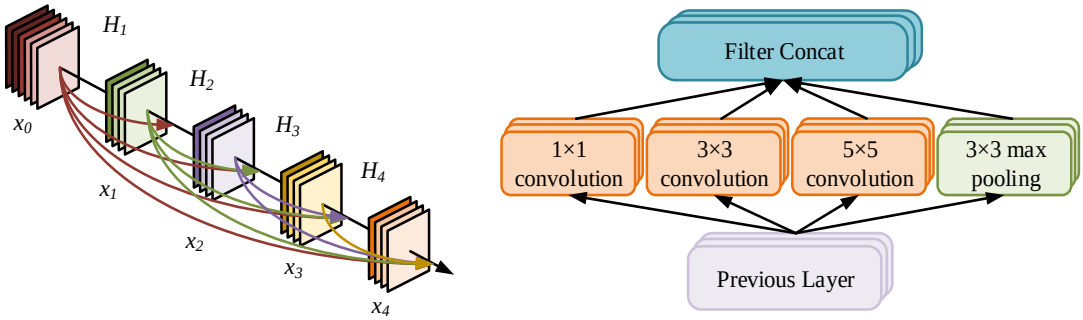
Author	Method description	Results	Limitations
Fu J et al. [7]	Brightness attention mechanism + GAN-based enhancement	Significant improvement in brightness and quality in low light	Not integrated with detection; no sign recognition
Cheng X et al. [8]	GAN with attention and multi-scale fusion	Nighttime enhancement improved by 12.89%	Weak for small-object detection
Hu C et al. [9]	Image-to-image translation + dual residual blocks	Improved facial image quality under degradation	Focused on face images, not traffic signs
Chenmin N et al. [10]	Dehazing via sky segmentation and transmittance optimization	Enhanced contrast and subjective clarity in traffic images	High dependency on sky region segmentation
Ferencz C et al. [11]	CNN-based multi-class sign classification (43 classes)	High classification accuracy on hold-out dataset	Poor robustness to blur and distortion
Zhu Y et al. [12]	Comparison of YOLOv5 and SSD for detection	YOLOv5 outperformed SSD in accuracy and speed	Limited in low-light and small-object scenarios
Min W et al. [13]	Scene modeling + multi-scale dense detection	Accuracy of 92.8% and 99.90% on two datasets	Strong reliance on image clarity and resolution
Abdel-Salam R et al. [14]	Real-time image-enhanced CNN (RIECNN)	Recognition rate of 99.75%, 99.25%, and 99.55% on three datasets	Lacks structural flexibility in complex backgrounds

Although current TSR methods have achieved promising results under standard conditions, significant challenges remain under various forms of image degradation. Most existing studies fail to systematically identify the specific limitations of different degradation types and their direct impacts on recognition performance. In low-light environments, reduced brightness and signal-to-noise ratio hinder traditional CNNs from extracting critical textures of small targets, leading to increased false detections and omissions. In blurry images, the loss of edge sharpness disrupts spatial structure modeling, especially affecting the localization of irregular-shaped signs. In adverse weather conditions such as fog or rain, weakened contrast and increased background noise further degrade performance, particularly for models lacking robust multi-scale feature fusion capabilities. Moreover, many methods focus solely on either image enhancement or recognition, lacking an end-to-end co-optimization mechanism that ensures both visual quality and detection accuracy. To address these issues, this study proposes a dual-stage framework combining enhancement and detection. The first stage introduces a Multi-scale Enhanced Channel Residual Network (MECRN), which integrates multi-scale convolution, channel attention (CA), and dense residual connections to restore structural clarity and contrast. The second stage constructs an improved PaddlePaddle-You Only Look Once model based on Deformable Convolutional Networks and Depthwise Separable Convolutions (PP-YOLO-DCN), enhancing robustness in detecting small, irregular, and multi-scale signs. This framework achieves stable and accurate recognition across diverse degraded traffic scenarios.

2 Methods and materials

2.1 Design of image enhancement model based on MECRN

In intelligent transportation systems, the clarity of images will directly affect the accuracy of TSR, especially in conditions such as long distance and severe weather, where images often appear blurry or low resolution, affecting the recognition effect [15]. The widely-used Super-Resolution CNN (SRCNN) achieves direct mapping from low resolution images to high-resolution images through an end-to-end CNN structure, greatly improving the reconstruction effect [16-17]. However, in complex traffic scenarios, SRCNN shows limitations in handling fine textures and multi-scale features. As the network deepens, it tends to suffer from gradient vanishing and feature redundancy, which adversely affect training stability and reconstruction quality. To address these issues, this study aims to design an image enhancement model capable of adapting to various types of degradation, while improving the accuracy and robustness of subsequent TSR. The specific research questions include: (1) How can structurally clear and detail-rich images be effectively restored under conditions such as low illumination, blurring, and compression? (2) Can multi-scale feature fusion and CA mechanisms enhance the model's ability to represent hierarchical texture information? (3) When used as a preprocessing module for detection, can the enhancement model significantly improve overall recognition performance? To answer these questions, this study introduces a strategy that combines dense connections and multi-scale convolution into the SRCNN framework, aiming to better meet the image enhancement needs of intelligent transportation systems and provide high-quality inputs for TSR. The structure of dense connections and multi-scale convolution is illustrated in Figure 1.



(a) Dense connections (b) Inception module

Figure 1: Densely connected and multi-scale convolutional structures

Figure 1 (a) and Figure 1 (b) respectively show dense connections and multi-scale convolution structures. In Figure 1 (a), the output of each layer is not only directly passed to subsequent layers, but also connected to deeper feature maps. By integrating shallow and deep features through cross layer connections, the propagation of information flow is enhanced, which can successfully mitigate the issue of gradient dissipation. In Figure 1 (b), the Inception module is a classic example of multi-scale convolution, which achieves multi-scale feature extraction through convolution kernels (CKs) and max pooling layers at different scales of 1×1, 3×3, and 5×5. Finally, by combining this multi-scale information through Concat, the network can maintain sparsity while maintaining high computational efficiency and adaptability. In dense connections, H represents a set of nonlinear operations, including convolutions, activation functions, etc. The input of each layer is the cumulative output of the current layer and all previous layers, where the input expression is shown in equation (1).

$$x_N = H_N([x_0, x_1, ..., x_{N-1}]) \quad (1)$$

In equation (1), N means the total number of layers. x_N is the input, and $[x_0, x_1, ..., x_{N-1}]$ represents the concatenation of all features from layer 0 to layer $N-1$. H_N performs non-linear transformations such as convolution and activation on these features. In the multi-scale convolution structure, the outputs of each branch are doing Concat in the channel dimension to obtain the output of the Inception module, as shown in equation (2).

$$x_{inception} = Concat(x_{1\times1}, x_{3\times3}, x_{5\times5}, x_{pool}) \quad (2)$$

In equation (2), $x_{1\times1}$, $x_{3\times3}$, and $x_{5\times5}$ represent the features extracted from the input feature map x using 1×1, 3×3, and 5×5 CKs. $x_{5\times5}$ means the final output feature map. x_{pool} represents the maximum pooling operation and subsequent convolution to obtain characteristics from the given feature map x. To further enhance multi-scale feature representation and information modeling, a Multi-Scale Residual Feature Attention (MSRFA) module is designed based on the structure in Figure 1. Compared with conventional multi-scale fusion methods such as Feature Pyramid Network (FPN) and Atrous Spatial Pyramid Pooling (ASPP), MSRFA is more lightweight and capable of preserving multi-scale details while enhancing the response to key regions through residual connections and

CA. This design improves image enhancement performance in challenging scenarios such as low light and blur, making it more suitable for intelligent transportation applications. The structure of MSRFA is shown in Figure 2.

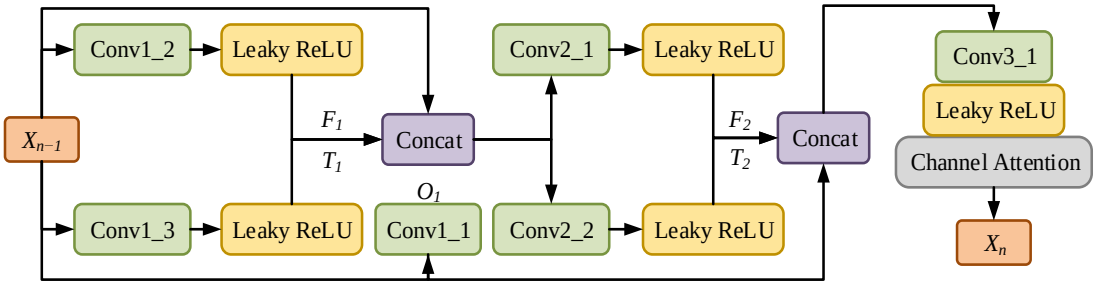


Figure 2: MSRFA structure

As shown in Figure 2, the MSRFA module is an optimized implementation based on the concepts of multi-scale convolution and dense connections presented in Figure 1. The module first extracts initial features from the input feature map through two parallel branches with different kernel sizes (Conv1_2 and Conv1_3), capturing multi-scale features F_1 and T_1. These two branches represent a simplified version of the Inception structure in Figure 1(b), retaining 3×3 and 5×5 convolutions to enhance medium and large receptive field feature extraction, while discarding the 1×1 convolution and max-pooling branches to reduce computational complexity and improve structural detail modeling. In addition, a 1×1 convolution (Conv1_1) is introduced to complement local feature extraction, forming the shallow feature extraction stage together with Conv1_2 and Conv1_3. All three branches produce 64 output channels. The concatenated features are then passed through Conv2_1 and Conv2_2 (3×3 and 5×5 convolutions) to extract deeper multi-scale features, with the output dimension remaining at 64. Finally, Conv3_1 applies a 1×1 convolution to compress the fused intermediate features into a 256-dimensional representation, which is further refined by a CA mechanism to enhance responses to key regions. This enables joint modeling of global context and local details, producing the final enhanced feature map.

It is worth noting that Conv3_1 not only integrates the preceding multi-scale features but also employs 1×1 convolution to compress channels and fuse spatial information, enabling the model to retain fine-grained local textures while introducing global contextual awareness. Combined with the CA mechanism, this design allows the model to simultaneously focus on key local regions and overall structural patterns, enhancing both perceptual consistency and structural fidelity in the enhanced image.

Firstly, the initial feature extraction expression is shown in equation (3).

$$O_1, F_1, T_1 = \Gamma(\omega_{1\times1} * X_{n-1}), \Gamma(\omega_{3\times3} * X_{n-1}), \Gamma(\omega_{5\times5} * X_{n-1}) \quad (3)$$

In equation (3), X_{n-1} represents the input feature map of the previous layer. $\omega_{1\times1}$, $\omega_{3\times3}$ and $\omega_{5\times5}$ respectively represent the weights of different convolutions. Γ represents the combination of convolution and activation function Leaky ReLU. O_1, F_1 and T_1 respectively represent feature information extracted through 1×1, 3×3, and 5×5 convolutions. Similarly, after deep feature fusion, the final output feature fusion expression is shown in equation (4).

$$X_n = C(\omega_{1\times1} * [T_2, F_2, O_1, X_{n-1}]) \quad (4)$$

In equation (4), $[T_2, F_2, O_1, X_{n-1}]$ represents concatenating the multi-scale features F_2 and T_2 obtained from two convolutions with the initial feature O_1 and input feature X_{n-1}, integrating the multi-scale information of the intermediate layer. C represents the CA mechanism. X_n is the final output feature map. Finally, grounded on the improved SRCNN network by integrating MSRFA, the MECRN image enhancement model structure is shown in Figure 3.

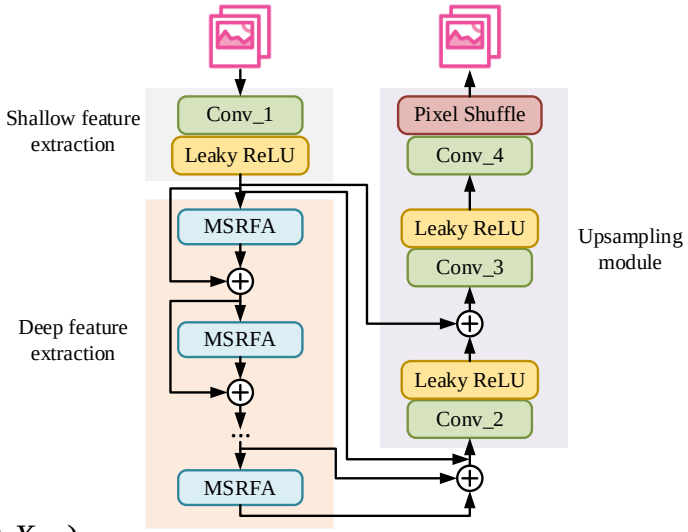


Fig. 3. MECRN structure diagram

In Figure 3, the network structure of the MECRN

image enhancement model is divided into shallow and deep feature extraction modules, and upsampling modules. Firstly, the shallow feature extraction module extracts initial features from the input low resolution image through convolutional layers Conv1 and Leaky ReLU activation functions. Secondly, the deep feature extraction module includes multiple MSRFAs, which capture different levels of detail information in the image through multi-scale feature fusion and CA mechanism. Meanwhile, skip connections have been added between each MSRFA module to preserve shallow information and reduce gradient dissipation. Finally, the upsampling module enhances the spatial resolution of features through the Pixel Shuffle layer, generating the final super-resolution output image.

2.2 Design of TSR model based on PP-YOLO

After constructing the MECRN-based image enhancement model, a TSR model is further developed based on the enhanced high-quality images. The study mainly focuses on the following questions: (1) Can deformable convolution (DCN) enhance the model's feature adaptability to irregular traffic signs? (2) Can depthwise separable convolution (DSC) reduce computational cost while maintaining detection accuracy? (3) Can the integration of image enhancement and detection improve recognition robustness in complex scenarios? The TSR model is based on PP-YOLO, and is a lightweight target detection algorithm based on YOLO architecture [18-19]. PP-YOLO inherits the efficient real-time detection capability of the YOLO series, and achieves a good balance between computational efficiency and accuracy by introducing more efficient model structures and feature optimization strategies. Its network structure is shown in Figure 4 [20].

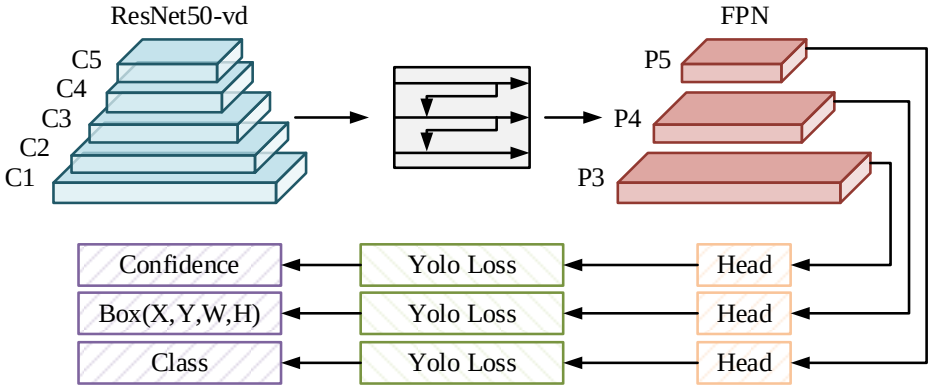


Figure 4: PP-YOLO network structure

In Figure 4, PP-YOLO includes three parts: feature extraction network, FPN, and detection head. ResNet50-vd, as its feature extraction network, convolves the input image layer by layer to extract feature maps of different levels C1 to C5, where C5 represents the depth features of the highest level. Secondly, the FPN module is used to fuse the feature maps of P3, P4, and P5 at different scales. Subsequently, the feature maps output by FPN are passed to the detection head, and each layer of feature map corresponds to a detection head. The YOLO Loss serves the purpose of computing the classification loss, bounding box loss, and confidence loss. Finally, the network outputs the Class, Box, and Confidence of each target.

Due to the fact that traffic signs often exhibit different shapes and scales due to factors such as perspective, distance, and lighting, to raise the feature extraction capability of the model for irregular and small-scale traffic signs in complicated traffic scenes, a DCN is introduced into PP-YOLO. DCN can dynamically adjust the sampling position of the CK, so that the network can focus on the key areas of the target, thereby capturing the detailed features of traffic signs more flexibly and efficiently. DCN is a key component of DCN, as shown in Figure 5.

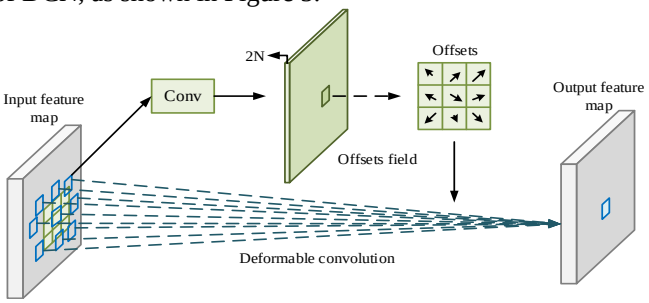


Figure 5: Schematic diagram of DCN

As illustrated in Figure 5, traditional convolution operations utilize fixed sampling positions arranged in a regular grid pattern (e.g., 3×3 centered on a pixel). However, in real-world traffic scenes, objects often exhibit complex variations such as deformation, occlusion, and tilt, making fixed sampling insufficient for capturing critical structural information. To address this limitation, DCN introduces learnable offsets that allow each sampling point in the convolutional kernel to dynamically adjust its position based on the input feature map. These offsets are generated by an independent

convolutional branch and added to the original regular grid positions to produce new, content-adaptive sampling locations. This mechanism enables the kernel to focus on structurally variant regions, such as object boundaries, curved contours, or partially missing areas, thereby enhancing the model's capability to represent fine-grained details. Bilinear interpolation is used to extract features from non-integer sampling coordinates, ensuring spatial continuity in the output.

Equations (5) to (7) describe the three core processes of DCN: offset calculation, non-uniform sampling, and weighted feature aggregation. Specifically, Equation (5) defines the adaptive adjustment of sampling positions. Equation (6) reconstructs features at fractional locations through bilinear interpolation. Equation (7) aggregates all sampled features to produce the final output map. Compared to standard convolution, DCN increases the flexibility of spatial modeling and improves adaptability to irregularly shaped objects. This structure is selected due to its strong compatibility with YOLO-based architectures and its effectiveness in handling small targets, partial occlusions, and geometric deformations, which are common challenges in TSR scenarios. Firstly, the offset calculation is shown in equation (5).

$$\Delta p_n = f_{offset}(x) \tag{5}$$

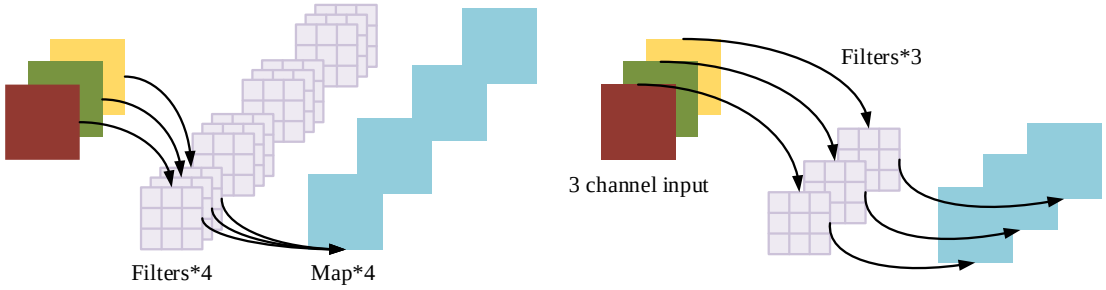
In equation (5), Δp_n represents the dynamic offset of each sampling point, and f_{offset} represents the convolution function used to generate the offset. Subsequently, the variability convolution operation is shown in equation (6).

$$y(p_0) = \sum_{p_n \in R} \omega(p_n) \cdot x(p_0 + p_n + \Delta p_n) \tag{6}$$

In equation (6), $y(p_0)$ represents the value of the output feature map at position p_0, and $x(p_0 + p_n + \Delta p_n)$ means the value of the input feature map at the position of the dynamic sampling point. $\omega(p_n)$ is the weight of the CK, consistent with traditional convolution operations. Finally, the overall output feature map of the variable convolution is represented by equation (7).

$$Y = f_{deform}(X, \Delta P) = \sum_{p_0 \in X} \sum_{p_n \in R} \omega(p_n) \cdot x(p_0 + p_n + \Delta p_n) \tag{7}$$

In equation (7), Y is the output feature map after variable convolution. f_{deform} is the DCN operation, and X means the input feature map. ΔP is the offset field. $p_0 \in X$ represents traversing all positions of the input feature map. Considering that increasing the depth of convolution will lead to an increase in parameter count and computational overhead, this study attempts to introduce DSC. By decomposing the standard convolution into deep convolution and 1×1 convolution, it can effectively reduce the computational and parameter count, and improve the computing speed. The structure is shown in Figure 6.



(a) Regular convolution (b) DSC

Figure 6: Schematic diagram of depth-wise separable convolution

Figures 6(a) and 6(b) illustrate the operations of standard convolution and DSC, respectively. In Figure 6(a), for a 5×5 input with three channels, four 3×3 convolutional kernels are applied. Each kernel performs a 3×3 convolution across all three input channels, and the results are summed to produce one output feature map, generating a total of four output channels. This operation involves simultaneous computation across spatial and channel dimensions, with computational cost increasing rapidly with the number of channels and kernels. In contrast, Figure 6(b) shows that DSC decomposes the convolution into two steps: First, depthwise convolution is applied separately to each input channel using a single corresponding kernel, maintaining the same number of output channels. Second, a 1×1 pointwise convolution is used to fuse information across channels. This design significantly reduces the number of multiplications and parameters while retaining spatial feature extraction and inter-channel interaction capabilities.

Compared to standard convolution, DSC offers higher efficiency and practicality in TSR tasks. On one hand, these tasks demand real-time inference, and the high computational cost of standard convolution in the channel dimension often becomes a bottleneck during early-stage feature extraction. DSC alleviates this issue by reducing the parameter count and computational load, thereby accelerating inference. On the other hand, traffic signs are typically small objects with clear structural

boundaries. DSC preserves spatial detail sensitivity through depthwise convolution and enhances semantic representation via pointwise convolution, improving the model's ability to extract edge and texture features. Therefore, the structure depicted in Figure 6(b) is highly suitable for integration into lightweight object detection models. The calculation of deep convolution is shown in equation (8).

$$y_{c,i,j} = \sum_{m=0}^{K-1}\sum_{n=0}^{K-1} x_{c,i+m,j+n} \cdot \omega_{c,m,n} \tag{8}$$

In equation (8), $y_{c,i,j}$ means the value of the output feature of channel c at (i, j). $x_{c,i+m,j+n}$ is the value of channel c at $(i+m, j+n)$, and $\omega_{c,m,n}$ is the weight of the $K \times K$ CK on channel c. Subsequently, the point by point convolution calculation is shown in equation (9).

$$z_{k,i,j} = \sum_{c=0}^{C-1} y_{c,i,j} \cdot v_{k,c} \tag{9}$$

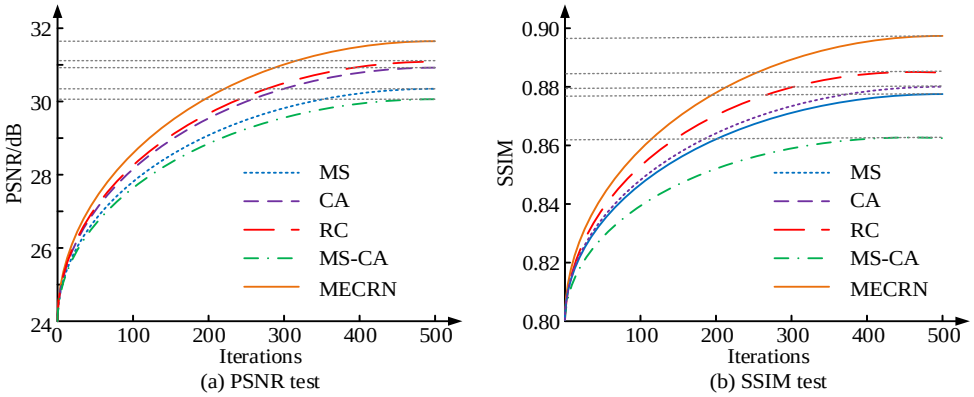
In equation (9), $z_{k,i,j}$ means the value of channel k at (i, j). $v_{k,c}$ is the weight of the stationary convolution, applied to $y_{c,i,j}$ and output channel k. Therefore, based on the above improvements, in the PP-YOLO-DCN model, the input image is first subjected to enhancement preprocessing. Secondly, the ResNet50-vd network is applied to obtain characteristics layer by layer, and the recognition adaptability is improved through FPN multi-scale fusion. DCN dynamic sampling is applied to capture details at key locations, computational efficiency is optimized through DSC, and ultimately the category, position, and confidence of traffic signs are output through detection heads to achieve efficient recognition in complex scenes.

3 Results

3.1 Performance testing of MECRN image enhancement model

The experiment utilized high-capacity calculation devices to guarantee effective training and evaluation of image enhancement and object recognition models. The experimental environment used Ubuntu operating system, Python as the programming language, combined with PyTorch deep learning framework and OpenCV image processing library to implement the algorithm. In terms of hardware configuration, the device was equipped with Intel Core i9 processor, NVIDIA RTX 3090 graphics card, and 128GB of memory. The Set14 dataset was used for training and testing in the image enhancement stage, as it contains structurally complex and texture-rich images suitable for evaluating detail restoration. As a general-purpose enhancement module, MECRN does not depend on the semantic content of the image. Its ability to improve visual clarity directly supports downstream TSR tasks, demonstrating strong generalizability. For TSR, the TT100K dataset was adopted, featuring real-world road scenes with small, blurred, and occluded signs, making it ideal for testing model robustness. All images were normalized and resized to 224×224 for enhancement and 640 × 640 for detection. Data augmentation such as brightness jittering, rotation, and random cropping was applied to improve generalization. For hyper-parameters, the enhancement model used the Adam optimizer with a learning rate of 1e-4, batch size of 16, and 500 epochs. The detection model used SGD with a learning rate of 0.001, batch size of 32, and 300 epochs. The loss functions included a combination of L1 and SSIM for enhancement, and multi-task YOLO loss for detection.

Firstly, as MECRN was composed of multiple modules, the ablation test outcomes are in Figure 7.



(a) PSNR test (b) SSIM test

Figure 7: Ablation test results

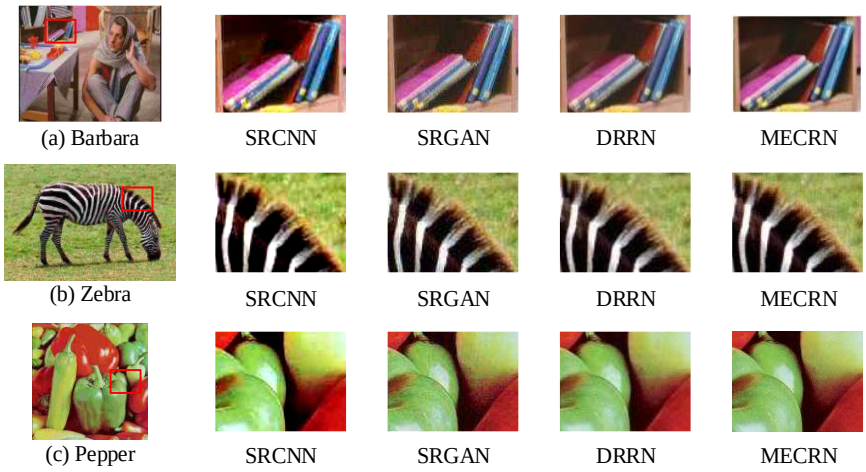


Figure 8: Qualitative analysis results

Figures 7 (a) and 7 (b) show the test results of Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) as they vary with the number of iterations, respectively. PSNR reflects reconstruction accuracy, while SSIM measures structural fidelity—both widely used in image enhancement tasks. MS, CA, RC, and MS-CA respectively represent the removal of multi-scale convolution, CA mechanism, residual connection, multi-scale convolution, and attention mechanism. In Figure 7 (a), when the number of iterations was 500, the PSNRs of removing MS, CA, RC, MS-CA, and the complete MECRN model were 30.4 dB, 30.9 dB, 31.1 dB, 30.1 dB, and 31.7 dB, respectively. In Figure 7 (b), the SSIMs without MS, CA, RC, MS-CA, and the complete MECRN model were 0.876, 0.879, 0.883, 0.862, and 0.897, respectively. The results demonstrated that each component of MECRN played a vital role in performance enhancement. RC helped maintain structural consistency and facilitates shallow-to-deep feature propagation, improving model stability. MS expanded the receptive field and strengthened the model's ability to capture fine textures and edges, which was particularly beneficial in upscaling and deblurring scenarios. CA enhanced the response to dark and critical semantic regions while suppressing redundant features. The largest performance drop occurred when both MS and CA were removed, highlighting their complementary effect in structural reconstruction and perceptual quality. These components together formed the core of the MSRFA module, and their integration enabled MECRN to achieve superior enhancement results through both independent effectiveness and mutual reinforcement.

Subsequently, Super Resolution CNN (SRCNN), Super Resolution Generative Adversarial Network (SRGAN), and Deep Recursive Residual Network (DRRN) were selected as comparison models. Three images were extracted from Set14, and then processed using models. The restoration effects of each model on image details and textures were examined. The qualitative assessment outcomes are in Figure 8.

Figure 8 shows the qualitative analysis results of super-resolution processing using different models. The enhancement effect of SRCNN was relatively poor, with blurry image details, especially in images with rich details such as Zebra and Pepper. The main reason was that SRCNN had a simple structure and lacked sufficient ability to capture details. SRGAN introduced generative adversarial networks, which could improve details to a certain extent, but there was still room for improvement in its performance on complex textures. DRRN effectively improved image clarity and edge details through recursive residual connections, but still had limitations in some multi-scale feature processing. In contrast, the MECRN model combined multi-scale convolution and CA mechanisms, which could more comprehensively capture multi-scale details and local features, resulting in the best texture and structure restoration effect and the highest clarity of the image. MECRN exhibited better enhancement effects in areas with rich details, such as Zebra stripes and Pepper surface textures. Finally, to simulate scenes in the real world where image quality is often affected by various factors, low lighting, low resolution, noise interference, and image blurring conditions were set. The test outcomes are in Table 2.

Table 2: Multi-index test outcomes under different environmental conditions

Environment condition	Metric	SRCNN	SRGAN	DRRN	MECRN
Low light	PSNR	27.45	28.32	29.54	30.75
	SSIM	0.812	0.826	0.846	0.871
	LPIPS	0.347	0.282	0.232	0.185
Low resolution	PSNR	26.31	27.12	28.66	29.91
	SSIM	0.804	0.819	0.836	0.861
	LPIPS	0.362	0.297	0.241	0.201
Noise interference	PSNR	25.81	26.53	27.92	29.21
	SSIM	0.791	0.806	0.826	0.851
	LPIPS	0.371	0.321	0.271	0.223
Blurred image	PSNR	24.91	25.71	27.13	28.51
	SSIM	0.776	0.791	0.811	0.842
	LPIPS	0.386	0.343	0.293	0.253
Compression artifacts	PSNR	25.35	26.18	27.45	28.68
	SSIM	0.788	0.812	0.821	0.842
	LPIPS	0.365	0.328	0.283	0.236

Table 2 presents the quantitative evaluation results of different models. The Learned Perceptual Image Patch Similarity (LPIPS) is a metric for assessing perceptual image quality, where lower values indicate that the enhanced image is more perceptually similar to the ground truth. Unlike PSNR and SSIM, which mainly measure pixel-level differences and structural similarity, LPIPS focuses on perceptual consistency in deep feature space. This makes it more effective in capturing the restoration of edges, textures, and fine details, which is particularly valuable for improving feature extraction and discrimination in TSR. Under different environmental conditions, the MECRN surpassed other models in terms of PSNR, SSIM, and LPIPS metrics. Under low light conditions, MECRN's PSNR, SSIM, and LPIPS were 30.75, 0.871, and 0.185, respectively, demonstrating its excellent low light enhancement ability. This was because MECRN combined multi-scale convolution and CA mechanisms, which could better capture dark details. Under low resolution conditions, its PSNR and SSIM were 29.91 and 0.861, respectively, indicating its excellent amplification capability. For noise interference and blurry images, MECRN performed well in denoising and deblurring, with LPIPS of 0.223 and 0.253 respectively, indicating that it could effectively restore visual perception quality. In addition, under compression distortion conditions, MECRN effectively suppressed compression artifacts, and SSIM reached 0.840. MECRN exhibited stronger robustness and adaptability in complex environments.

The advantage of MECRN in LPIPS performance mainly stems from its effective modeling of image details and perceptual consistency. The multi-scale convolution captures texture and edge information at different scales, helping to restore key details in upscaling and blurring scenarios. The CA mechanism enhances responses to dark regions and important areas, reducing interference from irrelevant features. Dense connections facilitate the transfer of shallow features to deeper layers, improving structural and detail reconstruction, thereby reducing LPIPS and enhancing overall visual quality.

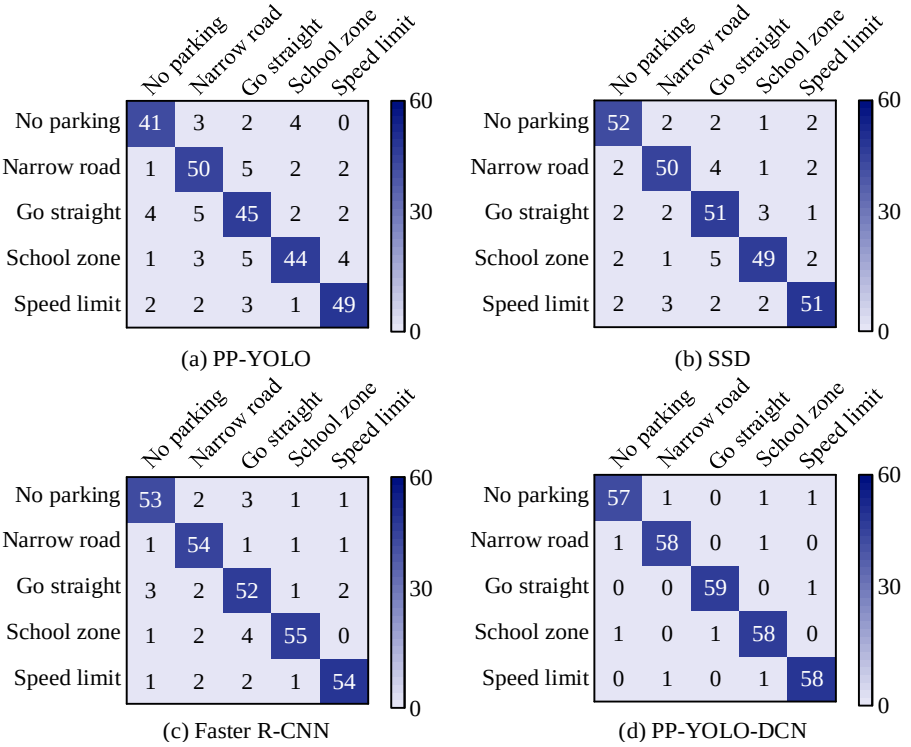
3.2 Experimental analysis of PP-YOLO-DCN object recognition model

The experimental environment was the same as the previous section, and the comparison models were PP-YOLO, the classic SSD, and the Faster Regional CNN (Faster R-CNN). The dataset was the TT100K traffic dataset, which contains 100000 real road scene images in China, covering more than 100 different traffic signs, suitable for small object detection and TSR in complex scenes. Firstly, to thoroughly assess the sorting performance of the recognition model in TSR, especially in similar category and multi-category detection, the results of the confusion matrix for each model are shown in Figure 9.

As shown in Figure 9, PP-YOLO-DCN outperformed other models in classification accuracy across various traffic sign categories. In particular, categories such as "go straight" and "school zone," which share similar shapes and subtle edge differences, exhibited 4 and 3 false detections respectively in PP-YOLO and SSD. Faster R-CNN also showed confusion in the "speed limit" category. In contrast, PP-YOLO-DCN demonstrated minimal misclassification across all categories. This performance improvement could be attributed to the architectural differences, especially the incorporation of DCN. Unlike standard convolution, DCN enabled dynamic adjustment of sampling positions based on the input features, allowing the model to better align with object boundaries and internal structures. This structural adaptability enhanced the model’s ability to distinguish fine-grained differences in small or partially occluded signs. The improvements

observed in the confusion matrix directly reflected the effectiveness of DCN in handling complex and visually similar traffic sign categories. Subsequently, to evaluate the feature focusing capability of the model in TSR, gradient-weighted class activation mapping (Grad CAM) visualization technology was applied to display the attention regions of the model when detecting traffic signs, in order to intuitively analyze the capture effect and detail attention of each model on key features. The results are shown in Figure 10.

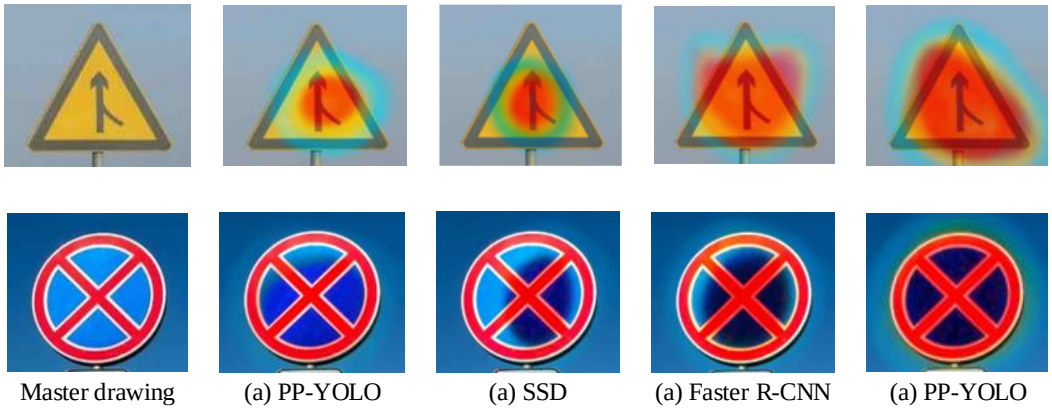
As shown in Figure 10, the Grad-CAM visualizations revealed clear differences in attention distribution among the models. PP-YOLO mainly focused on the central region of the sign, with weak responses along the edges, which may lead to localization errors. SSD showed a more uniform attention map but lacked distinct activation for critical visual cues such as arrows and cross lines. Faster R-CNN demonstrated strong attention in both central and boundary areas, indicating more balanced feature extraction. PP-YOLO-DCN exhibited the most precise and concentrated attention, effectively covering both the center and the boundaries of the signs, particularly highlighting detailed patterns and contours. This was directly related to the DCN mechanism. By learning spatial offsets, the convolutional kernels could shift their sampling positions to better align with object boundaries and local structures. This enhanced the model's ability to extract detailed features from curved edges, fine textures, and variable shapes. Under conditions such as low resolution or partial occlusion, this precise sampling enabled more complete structural representation, reducing false detections and omissions in TSR. Finally, considering the presence of TSR in complex scenarios in practical applications, the test results of each model in different environments are in Table 3.



(a) PP-YOLO (b) SSD

(c) Faster R-CNN (d) PP-YOLO-DCN

Figure 9: Confusion matrix test



Master drawing (a) PP-YOLO (a) SSD (a) Faster R-CNN (a) PP-YOLO

Figure 10: Heatmap visualization results

Table 3: Model performance testing in complex scenarios

Model	Scenario	mAP@0.5	FPS	Processing latency/ms	Robustness score
PP-YOLO	Dense multi-target	0.82	45	22	0.74
	Dynamic background	0.78	47	21	0.70
	Adverse weather	0.75	43	24	0.66
	Low light	0.72	40	26	0.63
	Motion blur	0.71	42	25	0.61
SSD	Dense multi-target	0.83	30	34	0.70
	Dynamic background	0.76	32	32	0.65
	Adverse weather	0.72	29	35	0.60
	Low light	0.69	28	36	0.57
	Motion blur	0.68	31	33	0.55
Faster R-CNN	Dense multi-target	0.87	15	66	0.81
	Dynamic background	0.84	16	64	0.75
	Adverse weather	0.82	14	69	0.70
	Low light	0.78	13	71	0.68
	Motion blur	0.76	14	68	0.64
PP-YOLO-DCN	Dense multi-target	0.91	42	25	0.85
	Dynamic background	0.88	44	23	0.82
	Adverse weather	0.86	40	27	0.78
	Low light	0.84	39	28	0.75
	Motion blur	0.82	41	26	0.73

As shown in Table 3, PP-YOLO-DCN demonstrated significant advantages in complex traffic scenarios such as dense multi-target and adverse weather conditions, achieving mAP@0.5 values of 0.91 and 0.86 respectively, outperforming other models in detection accuracy. Under low-light and motion-blurred conditions, its robustness scores reached 0.75 and 0.73, indicating strong resistance to environmental interference. Additionally, the model maintained a stable frame rate across all tested scenarios, ranging from 39 to 44 FPS, ensuring reliable real-time detection performance. In Table 3, the robustness score was introduced to comprehensively evaluate a model's ability to maintain performance under challenging conditions. It considered multiple factors, including the drop in mAP, changes in false detection rates, and fluctuations in processing latency. This metric reflected the overall stability and interference resistance of the model. In contrast, although Faster R-CNN achieved relatively high accuracy in some scenarios, its frame rate remained below 20 FPS, resulting in high latency and limited real-time applicability. Meanwhile, PP-YOLO and SSD exhibited weaker robustness, with scores of 0.63/0.61 and 0.57/0.55 in low-light and motion blur conditions, respectively, showing greater sensitivity to visual degradation. Overall, PP-YOLO-DCN achieved a strong balance among accuracy, efficiency, and robustness, making it well-suited for deployment in complex real-world traffic environments.

PP-YOLO-DCN maintained high FPS while improving accuracy due to its optimized architecture. DSCs reduced parameters and computation, while DCNs enhanced adaptability to complex shapes with minimal impact on speed. Additionally, the lightweight one-stage design of PP-YOLO ensured strong real-time performance, achieving a good balance between precision and efficiency.

To further verify that the proposed PP-YOLO-DCN model maintains high detection accuracy while ensuring computational efficiency, this study compared it with mainstream YOLO variants (YOLOv3, YOLOv5-M, and YOLOv5-L) in terms of model parameters, computational cost (GFLOPs), and inference speed (FPS). The results are shown in Table 4.

Table 4: Comparison of algorithmic complexity and performance

MODEL	PARAMETERS (M)	COMPUTATION (GFLOPS)	FPS (640 × 640)	MEAN ACCURACY (MAP@0.5)
YOLOV3	61	118	30	0.87
YOLOV5-M	44	98	45	0.88
YOLOV5-L	77	109	35	0.89
PP-YOLO-DCN	47	86	42	0.91

As shown in the table, PP-YOLO-DCN achieved a favorable trade-off between detection accuracy and computational efficiency. With 47 million parameters and 86 GFLOPs, it was lighter than YOLOv5-L and YOLOv3, and significantly reduced computation compared to YOLOv5-L (by approximately 21%). In terms of inference speed, PP-YOLO-DCN reached 42 FPS, slightly below YOLOv5-M but notably faster than YOLOv3 and YOLOv5-L. Overall, the model demonstrated strong performance in terms of accuracy, size, and real-time capability, making it well-suited for practical deployment in TSR scenarios.

4 Discussion

In TSR tasks, image quality is closely tied to detection accuracy. Although prior studies have explored both image enhancement and object detection independently, many existing methods still face limitations when applied to complex real-world scenarios. Compared with these approaches, the proposed MECRN enhancement module and PP-YOLO-DCN detection model demonstrated superior overall performance and adaptability across diverse conditions. In terms of image enhancement, Fu J et al. [7] proposed a method that integrated a brightness attention mechanism with a GAN to enhance image brightness in low-light environments. However, the method was not integrated with downstream recognition tasks, leading to a disconnect between enhancement quality and detection performance. Cheng X et al. [8] employed multi-scale fusion and attention mechanisms to improve nighttime image quality but showed limited capability in restoring small-object details. In contrast, the proposed MECRN model utilized multi-scale convolution to extract texture information at various scales, incorporated CA to enhance the response to dark regions and target areas, and introduced dense connections to improve feature reuse and training stability. As a result, MECRN improved traditional metrics such as PSNR and SSIM while significantly reducing perceptual distortion measured by LPIPS, ultimately providing higher-quality inputs for detection models.

In terms of detection, Zhu Y et al. [12] compared YOLOv5 and SSD and found that YOLOv5 achieved a better balance between speed and accuracy. However, both models exhibited performance drops in low-light and small-object scenarios. While Faster R-CNN delivered high accuracy, it suffered from heavy computational cost and high latency, making it unsuitable for real-time deployment. The proposed PP-YOLO-DCN introduced DCN to enable dynamic sampling, enhancing the model’s ability to capture object boundaries and complex structures. This proved especially effective in degraded conditions such as blur and occlusion, where accurate localization of small targets was challenging. Additionally, the integration of DSC reduced redundant computation, achieving a good trade-off between accuracy and inference speed. The proposed model outperformed others in terms of robustness and real-time performance under complex conditions. In summary, MECRN and PP-YOLO-DCN demonstrated strong adaptability to low-quality image conditions and effectively addressed the trade-off between accuracy and efficiency in TSR, offering promising potential for real-world deployment.

5 Conclusion

In response to the difficulty of recognizing traffic signs in complex traffic scenarios such as low light and long distance, the MECRN image enhancement model and PP-YOLO-DCN object recognition model were proposed to improve the image quality and detection accuracy of intelligent transportation systems. In the ablation test, the PSNR of the complete MECRN model was 31.7 dB, and the SSIM was 0.897, confirming the effectiveness of the study in improving MECRN. In qualitative analysis, MECRN had the best image enhancement effect and the highest clarity. Under low light conditions, the PSNR, SSIM, and LPIPS of MECRN were 30.75, 0.871, and 0.185, respectively, indicating good low light enhancement effect. The PP-YOLO-DCN model had the lowest number of false positives and false negatives in confusion matrix testing. In Grad CAM testing, the heatmap of PP-YOLO-DCN covered the edges and internal details of the logo, especially with high attention to key features. Under conditions of dense multi-target and adverse weather conditions mAP@0.5 reached 0.90 and 0.86 respectively, the FPS remained above 40 frames per second. The outcomes indicated that the raised method attained a high degree of recognition accuracy in various complex scenarios.

Although the proposed model demonstrated strong performance in TSR, achieving high accuracy, real-time speed, and robustness across challenging conditions such as low light and motion blur, there remains room for further improvement. First, the current evaluation was conducted on high-performance GPUs, and the model's inference efficiency on edge devices has yet to be

validated. Second, the experiments were primarily based on the TT100K dataset, lacking cross-domain generalization testing on international benchmarks such as GTSRB. Finally, while the model exhibited relatively high robustness scores, it has not been systematically evaluated under extreme conditions such as severe occlusion, heavy noise, or adversarial perturbations. Future work will focus on improving model adaptability through edge deployment, cross-dataset validation, and robustness analysis under more challenging real-world scenarios.

References

[1] Hasanvand M, Nooshyar M, Moharamkhani E, Selyari A. Machine learning methodology for identifying vehicles using image processing. Artificial Intelligence and Applications, 2023, 1(3):170-178. https://doi.org/10.47852/bonviewAIA3202833

[2] Muhammad K, Ullah A, Lloret J, Del Ser J, de Albuquerque V. H. C. Deep learning for safe autonomous driving: Current challenges and future directions. IEEE Transactions on Intelligent Transportation Systems, 2021, 22(7): 4316-4336. https://doi.org/10.1109/TITS.2020.3032227

[3] Liu X, Yan W Q. Traffic-light sign recognition using Capsule network. Multimedia Tools and Applications, 2021, 80(10):15161-15171. https://doi.org/10.1007/s11042-020-10455-x

[4] Shivayogi A B, Dharmendra N C M, Ramakrishna A M, et al. Real-time traffic sign recognition using deep learning. Pertanika J Sci Technol, 2022, 31(1):137-148. https://doi.org/10.1007/s11042-022-12163-0

[5] Yucong S, Shuqing G. Traffic sign recognition based on HOG feature extraction. Journal of Measurements in Engineering, 2021, 9(3):142-155. https://doi.org/10.21595/jme.2021.22022

[6] Luo Q, Zheng W. Pre-Locator Incorporating Swin-Transformer Refined Classifier for Traffic Sign Recognition. Intelligent Automation & Soft Computing, 2023, 37(2):2227-2246. https://doi.org/10.32604/iasc.2023.040195

[7] Fu J, Yan L, Peng Y, Zheng K, Gao R, Ling H. Low-light image enhancement base on brightness attention mechanism generative adversarial networks. Multimedia Tools and Applications, 2024, 83(4):10341-10365. https://doi.org/10.1007/s11042-023-15815-x

[8] Cheng X, Zhou J, Song J, Zhao X. A highway traffic image enhancement algorithm based on improved GAN in complex weather conditions. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(8):8716-8726. https://doi.org/10.1109/TITS.2023.3258063

[9] Hu C, Liu Y, Xu L, Xiao X, Lu X, Yang W, Liu P. Joint image-to-image translation for traffic monitoring driver face image enhancement. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(8):7961-7973. https://doi.org/10.1109/TITS.2023.3258634

[10] Chenmin N, Marsani M F, Shan F P. Traffic image dehazing based on sky region segmentation and transmittance optimization. Journal of Intelligent & Fuzzy Systems, 2024, 46(1):1005-1017. https://doi.org/10.3233/JIFS-233433

[11] Ferencz C, Zöldy M. Neural Network-based Multi-Class Traffic-Sign Classification with the German Traffic Sign Recognition Benchmark. Acta Polytechnica Hungarica, 2024, 21(7):203-220.

[12] Zhu Y, Yan W Q. Traffic sign recognition based on deep learning. Multimedia Tools and Applications, 2022, 81(13):17779-17791. https://doi.org/10.1007/s11042-022-12163-0

[13] Min W, Liu R, He D, Han Q, Wei Q, Wang Q. Traffic sign recognition based on semantic scene understanding and structural traffic sign location. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(9):15794-15807. https://doi.org/10.1109/TITS.2022.3145467

[14] Abdel-Salam R, Mostafa R, Abdel-Gawad A H. RIECNN: real-time image enhanced CNN for traffic sign recognition. Neural Computing and Applications, 2022, 34(8):6085-6096. https://doi.org/10.1007/s00521-021-06762-5

[15] Jeya R, Krishnan G R, Babu C R. Traffic Sign Classification and Recognition using LE-NET Technique. International Journal of Early Childhood Special Education, 2022, 14(3):544-551. http://doi.org/10.9756/INT-JECSE/V14I1.221001

[16] Chen Q, Zhang J, Li B. Research on 3D MFL testing of wire rope based on empirical wavelet transform and SRCNN. Journal of Vibroengineering, 2022, 24(4):779-792. http://doi.org/10.9756/INT-JECSE/V14I1.221001

[17] Nguyen N B, Doan V S, Pham M N, et al. SRCNN: Stacked-Residual Convolutional Neural Network for Improving Human Activity Classification Based on Micro-Doppler Signatures of FMCW Radar. Journal of Electromagnetic Engineering and Science, 2024, 24(4):358-369. https://doi.org/10.1371/journal.pone.0308045

[18] Chen C, Yu J, Lin Y, Lai, F, Zheng G, Lin Y. Fire detection based on improved PP-YOLO. Signal, Image and Video Processing, 2023, 17(4):1061-1067. https://doi.org/10.1007/s11760-022-02312-1

[19] Zhang G, Zhang J. High-Precision Photogrammetric 3D Modeling Technology Based on Multi-Source Data Fusion and Deep Learning-Enhanced Feature Learning Using Internet of Things Big Data. Informatica, 2025, 49(11).

[20] Gu X, Dai S. IRF-HTID-BO-LSTM: Classification Model of Curve Shape Index for Mountainous Highways and Intelligent Traffic Incident Detection Method. Informatica, 2025, 49(11). https://doi.org/10.31449/inf.v49i11.7449

Enhancing YOLOv8 Object Detection with Shape-IoU Loss and Local Convolution for Small Target Recognition

Qi Zhang[1*], Jingyuan Zhang[2], Shilei Yang[2]
[1]Engineering Training Center, Xi'an University of Science and Technology, Xi'an, 710600, China
[2]School of Electrical and Control Engineering, Xi'an University of Science and Technology, Xi'an, 710600, China
E-mail: xustzhangqi@163.com

Keywords: YOLO, loss function, local convolution, object detection, distinguish, computer

In light of the prevailing challenges, including the suboptimal recall rate and the substantial computing demands of the You Only Look Once (YOLO) series target detection algorithm, the study proposes an innovative approach to enhance the algorithm's loss function. This enhancement is achieved by introducing the SIOUs loss function, which is derived from the YOLOv8 network architecture. Additionally, the study proposes an adjustment to the calculation method of the weight of the loss function, employing the SlideLoss loss function. Then, this paper introduces SPD-Conv and PConv local convolution structure to improve the Backbone layer and proposes a computer object detection model based on loss function and local convolution. It was found that the prediction accuracy of the study model was above 90% on both the training and validation sets. Compared with the conventional YOLOv8 network, the average accuracy of the study model was increased by 13.19%. In the COCO and ImageNet datasets, the number of parameters, FLOPs, reasoning speed, and other indicators of the proposed model were better than the traditional YOLOv8. This indicated that the proposed algorithm could significantly improve the detection speed under the condition of improving the detection accuracy, realizing the balance between the two, and being more effective and reliable. The example results showed that the average coverage rate and the center position error of the research model were increased by 5.7% and 2.3%, respectively, which could effectively improve the operation accuracy. The research results can provide high detection accuracy for the target identification problems in various fields and have important application value in medical image analysis, industrial quality inspection, and other fields.

Povzetek: Razvita je izboljšana YOLOv8 arhitektura z novo Shape-IoU izgubo in lokalnimi konvolucijami, ki znatno poveča natančnost zaznave majhnih tarč ob zmanjšani računski zahtevnosti.

1 Introduction

The rapid advancement of big data, computing power, and deep learning algorithms has ushered artificial intelligence technology into a new era of big data and deep learning. In this process, artificial intelligence technology has been extensively applied in various industries and commercial sectors, providing substantial convenience to people's daily lives. Additionally, it has played a pivotal role in promoting technological innovation and development across multiple domains [1-2]. As a key task in computer vision, object detection has shown great potential for applications in multiple scenarios such as autonomous driving, intelligent monitoring, and medical image analysis. Currently, Computer Object Detection (COD) technology is moving in diversified directions [3-4]. In recent years, due to the increasing demand for object detection speed, efficient object detection algorithms have emerged. Single-stage detector object detection algorithms have attracted a lot of attention due to their excellent performance. This kind of algorithm transforms object detection tasks into regression problems and achieves real-time detection by simultaneously completing object classification and detection through a single forward propagation. Compared to traditional two-stage detectors, single-stage detectors do not require the generation of candidate regions for reclassification. Instead, they directly perform dense sampling on the feature map and use Convolutional Neural Networks (CNN) for classification and position regression. LiH et al. introduced deterministic aware pseudo labels for object detection and improved the semi-supervised object detection algorithm to address the issue of imbalanced positioning accuracy and amplification categories when using pseudo labels. The results showed that this method improved the performance of the most advanced SSOD on COCO and PASCAL VOC, with an average accuracy increase of 1-2%, and was orthogonal complementary to most existing methods [5]. To reduce the noise in the generation of pseudo labels and minimize the negative impact of noisy pseudo labels on model training, Yang J adopted a comprehensive pseudo label denoising method. In unsupervised 3D object detection adaptive noise training, this method was the most advanced in performance among all evaluation settings, greatly exceeding the corresponding baseline. Even in the KITTI 3D object detection benchmark, its performance has improved by 9.6% to 38.16% compared to fully supervised prediction based on object priors [6]. In view of the problems existing in the detection of small and medium Remote Sensing (RS) targets, Weiya et al. proposed a RS Small Object Detection (SOD)

method based on cross-layer fusion and weighted receptive field. To solve the problem of feature loss in deep layers of this method, a cross-layer attention fusion module was designed. The background noise was effectively filtered by introducing the double-layer routing attention. To enhance the ability of the model to sense multi-scale objects, especially small-scale ones, a weighted spatial pyramid pooling module of multi-receptive field voids was introduced. The experimental results clearly demonstrated the significant advantages of the model in RS SOD, surpassing the performance of the current mainstream model [7]. Faced with inherent challenges such as multi-scale, limited target area, and complex background in RS images, Zhong et al. proposed a RS image object detection method with multi-scale fusion dynamic head. This method has improved accuracy by 6.3% compared to traditional models, clearly demonstrating its significant advantages [8]. To solve the safe operation of the power grid, an improved You Only Look Once (YOLO) algorithm was introduced by Ya et al. Through experimental verification, in the test set, the unimproved YOLO algorithm's recall rate and detection accuracy were 68.3% and 88.8%, respectively, while the recall rate and detection accuracy of the improved model reached 82.7% and 91.4%. The crane recognition accuracy was improved by 2.9%, with a significant improvement [9]. To enhance the effectiveness of SOD in UAV aerial images, Xiang et al. proposed an enhanced YOLO algorithm based on multi-scale spatial context. This algorithm achieved significant results, with an Mean Absolute Precision (mAP) of 3.0% higher than the baseline method, demonstrating satisfactory performance in SOD [10]. YANG et al. constructed a cascaded model grounded on an improved YOLO structure to address difficulty in combining environmental semantic information at various scales in underwater detection algorithms. The proposed algorithm was capable of detecting aquatic targets and the evaluation index of common objects in context has improved by 4.34% [11]. To improve the generalization performance of the model and the application of data augmentation technology in target detection, Abdulghani et al. proposed a multiple data enhancement method. The experimental results showed that each data enhancement technology brought different degrees of improvement, and the method improved the mAP value of all objects by 13%, effectively improving the performance of the target detection algorithm [12]. Ananthakrishnan et al. developed an improved detection framework built on YOLOv5 to address inadequate lighting and poor performance of automated object detection under nighttime lighting conditions. It was used for effective detection under uneven lighting conditions at night. This architecture achieved higher results on mAP while reducing model size and total parameters. In terms of model size, the proposed model was 11.24% lighter and 12.38% lighter in terms of parameters [13]. The above research results are summarized as shown in Table 1.

Table 1: Summary table of literature reviews

Authors	Research Methods	Experimental Results
Li H et al. [5]	We introduce deterministic aware pseudo labels for target detection to improve semi-supervised target detection calculation	Improved state-of-the-art SSOD performance on COCO and PASCAL VOC, with improved average accuracy of 1-2%
Yang J et al. [6]	Unsupervised domain-adaptive denoising self-training in 3D object detection	Increase in performance by 9.6% 38.16%
Weiya et al. [7]	Proposed a YOLO-based method with cross-layer fusion and weighted receptive field.	The model showed significant advantages in small object detection,outperforming mainstream models.
Zhong et al. [8]	Introduced a multi-scale fusion dynamic head method for object detection in RS i	Improved accuracy by 6.3%compared to traditional models.
Ya et al. [9]	Developed an improved YOLO algorithm for power grid safety	Achieved 91.4%detection accuracy,with crane recognition accuracy improved by 2.9%
Xiang et al. [10]	Proposed an enhanced YOLO algorithm with multi-scale spatial context for UAV aerial image	Increased mAP by 3.0%compared to baseline methods.
YANG et al. [11]	Designed a cascaded model based on improved YOLO for underwater object detection	Improved evaluation metrics by 4.34%compared to baseline networks.
Abdulghani et al. [12]	Investigated data augmentation techniques for object detection in limited datasets.	Improved mAP by 13% of all objects.
Ananthakishnan et al. [13]	Proposed an improved YOLOv5 framework for object detection under	Achieved higher mAP, reduced model size by 11.24%, and parameters by 12.38%.

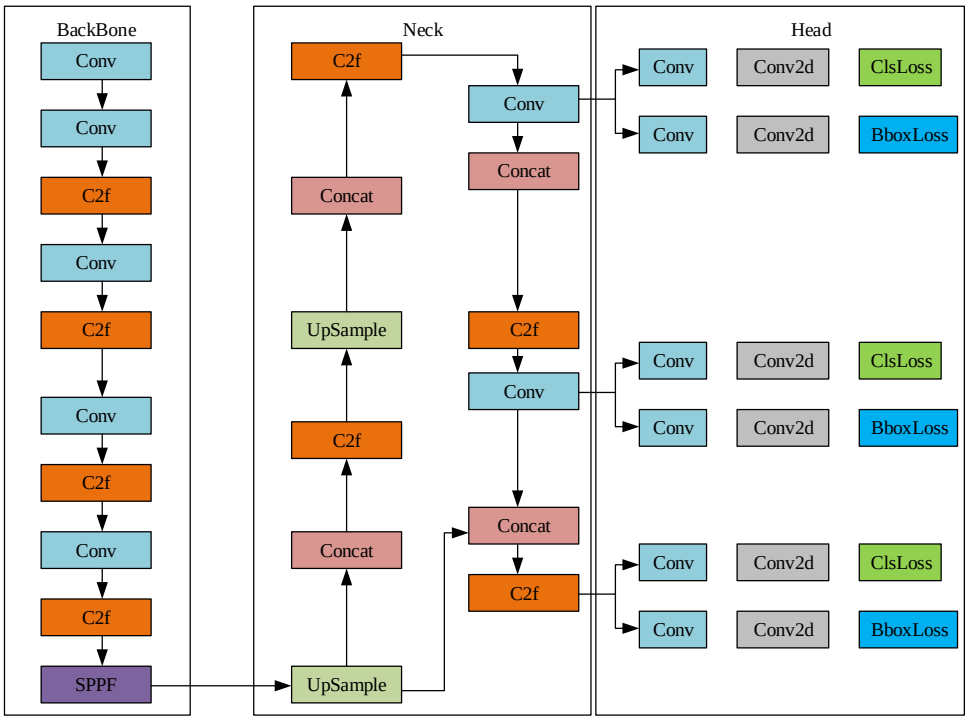
	uneven lighting conditions at night	

In summary, research on YOLO algorithms has achieved certain results. The success of the YOLO algorithm is attributed to its end-to-end training method, fast detection velocity, low background false detection rate, and strong generality and generalization capacity. Meanwhile, the YOLO algorithm also adopts a multi-scale feature fusion strategy, effectively handling object detection problems at different scales. However, the current YOLO series still has shortcomings. On the one hand, compared to other algorithms, YOLO has a lower recall rate, which may result in missing some targets. This is also due to its approach of treating object detection tasks as a regression problem, resulting in relatively conservative predictions of the target. On the other hand, it has a high demand for computing resources. To achieve high performance, the YOLO series models require high computing resources, which limits the application of YOLO in some resource-constrained scenarios. Based on this, the study innovatively proposes a modified YOLOv8 target detection model based on loss function and local convolution. In this model, based on the YOLOv8 network, the SIOU loss function is introduced to improve the algorithm. Meanwhile, the SlideLoss function is used to adjust the weight calculation method of the loss function to solve the problem of low recall rate and accuracy of YOLOv8 target detection. Then, a SPD Convolution (SPD-Conv) and a more efficient local convolution structure (Partial Convolution, PConv) are introduced to improve the Backbone layer to solve the problem that YOLOv8 needs high computing resources. The hypotheses to be verified in this study mainly include four aspects. (1) Whether SIOU loss function can improve the recall rate and precision of YOLOv8 network; (2) Whether the introduction of SPD-Conv can improve the SOD performance of YOLOv8 network; (3) Whether the combination of SPD-Conv and PConv reduces the calculation cost while maintaining accuracy; (4) Whether the combination of SIOU loss function, SPD-Conv, and PConv can balance the detection accuracy and efficiency of YOLOv8 network. Through the above in-depth analysis, the research is expected to improve the detection accuracy of YOLOv8 algorithm, reduce the computing resources, and promote the application value of YOLO series algorithm in object detection and recognition.

2. Methods and materials

2.1 Construction of an improved object detection model based on loss function

At present, COD technology is widely used in various fields. However, current computer detection technology can no longer meet the needs of reality [14-16]. In response to the problems existing in current object detection algorithms, this study improves the YOLOv8 algorithm from the aspects of local convolution structure and objective function, and proposes a COD model suitable for small target scenes. The reason why YOLOv8 is chosen as the sole baseline in this study is that, firstly, the algorithm is the latest version of the YOLO series. Compared to YOLOv5 and YOLOv7, it has significantly improved detection accuracy, speed, and model efficiency, representing the state-of-the-art in target detection. Secondly, YOLOv8 has undergone multiple optimizations in architecture design and training strategies, such as a more efficient backbone network and a more flexible loss function, providing a stronger foundational framework for research. In addition, YOLOv8 has an open source community and abundant tools and resources, making it easy for researchers to conduct experiments and validation. Choosing YOLOv8 as the benchmark can more intuitively demonstrate the effectiveness of the improved method and ensure that the research results are consistent with the current technological frontier, thereby enhancing the practicality and reference value of the research. The traditional YOLOv8 structure is shown in Fig.1.



Conv: Represents a conventional convolutional layer, Concat： Represents the connection layer, UpSample represents the upsampling operation

Figure 1: Traditional YOLOv8 structure

In Fig.1, Conv represents the traditional convolutional layer, Concat represents the connected layer, and the UpSample represents the upsampling operation. The traditional YOLOv8 structure includes Backbone, Neck, and Head layers. Among them, the Backbone layer effectively extracts feature information of the input object through a series of convolution and transpose convolution operations. Meanwhile, to reduce network size and improve overall performance, this layer also incorporates the design of residual connections and bottleneck structures [17-18]. This section uses convolution module, Cross Stage Physical Bottleneck with 2Convolutions (C2f) module, and Spatial Pyramid Pooling Fast (SPPF) module as the basic building blocks. The Neck layer is mainly responsible for fusing feature maps from different stages in the Backbone to enhance feature expression ability. Its internal structure includes components such as upsampling module and Path Aggregation Network (PAN). The Head layer focuses on the localization and category determination of object detection. The Head layer adopts a separated design, which means that classification tasks and detection tasks are handled separately by independent classification heads and detection heads. Both parties bear the optimization of classification loss and localization loss separately. Specifically, the classification loss uses VFL loss, while the localization loss combines DFL loss and CIoU loss. The traditional YOLOv8 algorithm combines localization loss and classification loss when calculating loss, with the total loss representing the loss of the entire algorithm. The total loss function is given by equation (1).

$$CIoU = 1 - IoU + \frac{\rho^2\left(b, b^{gt}\right)}{c^2} + \alpha v \tag{1}$$

In equation (1), b and b^{gt} are the center point of the prediction and real boxes. c is the diagonal distance between the closed areas of two rectangular boxes. ρ represents the calculated Euclidean distance. α is the weight coefficient. v measures the ratio consistency between the width and height of two rectangular frames. Compared with IoU, this loss function measures the positioning results by including the overlap area, center point distance, and Aspect Ratio (AR) as part of the loss, making the model's positioning results more accurate. However, the CIoU loss calculation used in this method is relatively complex because it not only considers the Intersection over Union (IoU) ratio itself but also introduces multiple factors such as center point distance and AR. This complexity may result in high computational costs during the training process, especially for large-scale datasets, where training time may significantly increase. In the formula of the CIoU loss function, there are some terms related to the AR of the predicted and true boxes. When the AR of the predicted box differs greatly from the real box, the values of these terms may become very large, leading to the problem of gradient explosion [19-20]. For this issue, this study improves the YOLOv8 function. Firstly, SIOU not only considers IoU but also combines the angle information of the target to make the

loss function smoother and avoid the gradient explosion problem caused by the excessive difference in length and width ratio. Furthermore, SIOU has proven to be particularly effective in detecting small targets. This efficacy stems from SIOU's ability to capture shape and position characteristics of these targets with a high degree of precision, enhancing the system's overall detection accuracy. Compared to CIoU, SIOU performs better in terms of computational efficiency and stability, making it particularly suitable for handling large-scale datasets and complex models. Meanwhile, SIOU exhibits higher robustness and accuracy in SOD tasks [15]. The angle loss Ω of this function is described in equation (2).

$$\Omega = 1 - 2 \times \sin^2\left(\arcsin\left(\frac{h}{\chi}\right) - \frac{\pi}{4}\right) \tag{2}$$

In equation (2), h is the height difference between the actual and predicted box center points. χ is the distance between two center points. The distance loss ∂ of the SIOU function also takes into account the angle loss, as shown in equation (3).

$$\partial = \sum_{t=x,y} 1 - e^{-\gamma \rho_t} \tag{3}$$

In equation (3), ρ_t is the square value where the difference between the actual box and the predicted box is less than the min bounding rectangle width of both boxes, $\gamma = 2 - \Omega$. Among them, w is the width of the minimum bounding rectangle of the actual box and the predicted box. h_1 is the height of the rectangle. (o_{ax}, o_{ay}) and (o_{px}, o_{py}) are the coordinates of the center points of the actual and predicted boxes. The description of the shape loss ℓ of SIOU is shown in equation (4).

$$\ell = \left(1 - e^{-w_1}\right)^{\theta} + \left(1 - e^{-h_2}\right)^{\theta} \tag{4}$$

In equation (4), w_1 and h_2 are the width and height of the prediction boxes. θ is the correlation of shape loss, within the range of $[2,6]$. This range is determined based on previous experience [19]. Finally, the loss description formula for SIOU is shown in equation (5).

$$L = 1 - IOU + \frac{\partial + \ell}{2} \tag{5}$$

This study can replace the original CIoU loss of YOLOv8 with SIoU loss to improve the convergence speed and accuracy. Then, in response to the problem of detecting more details or smaller target categories in practical applications, this study further utilizes SlideLoss to adjust the weight calculation method of the loss function, making the algorithm network more inclined to learn difficult samples.

The weight allocation scheme of the SlideLoss function is shown in Fig.2.

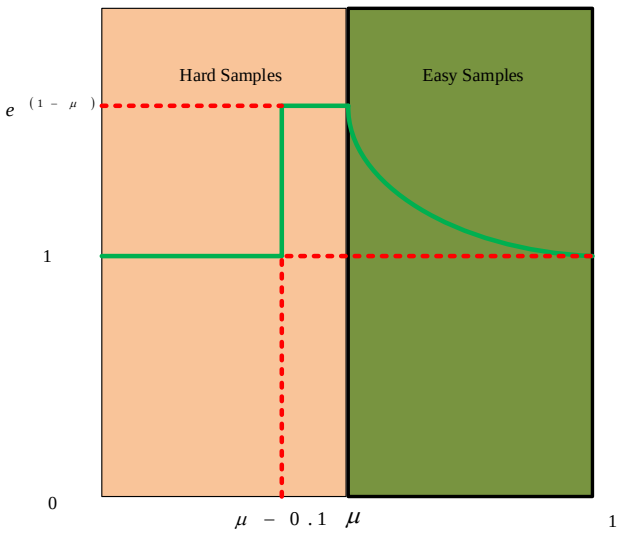


Figure 2: Weight allocation scheme for SlideLoss function

Slide Loss uses the average IoU value of all bounding boxes as the threshold μ. The determination of this threshold depends on the characteristics of the entire dataset and is dynamically adjusted to changes in the data. For samples in object detection tasks, this study classifies them based on their IoU values with real boxes for ease of classification. Specifically, when the IoU value is below the preset threshold, the sample is considered a difficult sample with a higher weight, thus occupying a more important position in loss calculation. On the contrary, samples with IoU values higher than the threshold are considered simple samples, with relatively smaller weights and smaller contributions to the loss. Samples located outside the $[\mu, \mu - 0.1]$ boundary region between different categories are more prone to misclassification. Therefore, assigning the maximum weight to these samples can make the classifier pay more attention to these difficult samples during training, thereby improving overall classification performance. However, samples located near decision boundary $[\mu, \mu - 0.1]$ have the highest weight, indicating that they generate the most significant losses. However, when the IoU value is below a certain level $\mu - 0.1$, these samples are usually considered extremely difficult or false positive abnormal situations, and their weights are not adjusted at this time [20]. The SlideLoss strategy is based on this principle, which prioritizes and invests more resources in learning difficult samples, thereby effectively improving the overall IoU value and enhancing the detection precision.

2.2 Construction of object detection model based on local convolution improvement

After improving the loss function of the YOLOv8 algorithm, it is necessary to further enhance the computational efficiency of YOLOv8, suppress the frequent memory access of neural networks, especially deep convolutions, and reduce redundant calculations and memory access. Therefore,

this study introduces SPD-Conv and PConv structure to improve the Backbone layer. The Backbone network of YOLOv8 utilizes a convolution module with 2 steppes for downsampling, gradually expanding the receptive field and generating multi-scale feature maps. However, this downsampling process inevitably causes the loss of some detailed information in the feature map. Especially for small objectives with low-resolution and limited information, this information loss situation can seriously affect their detection performance and may cause a significant decrease in detection accuracy. Meanwhile, YOLOv8's convolutional layers perform poorly in handling small targets. This is because after multiple convolution operations, the feature information of small targets is easily diluted, leading to a decrease in detection accuracy. In addition, the convolutional structure of YOLOv8 is more sensitive to background noise, especially in complex backgrounds, which can easily lead to false positives or false negatives. The article introduces a novel Convolutional module called SPD-Conv in the initial downsampling stage to replace the original convolution layer with a stride of 2, ensuring that information is not lost during the downsampling process. The network structure is shown in Fig.3.

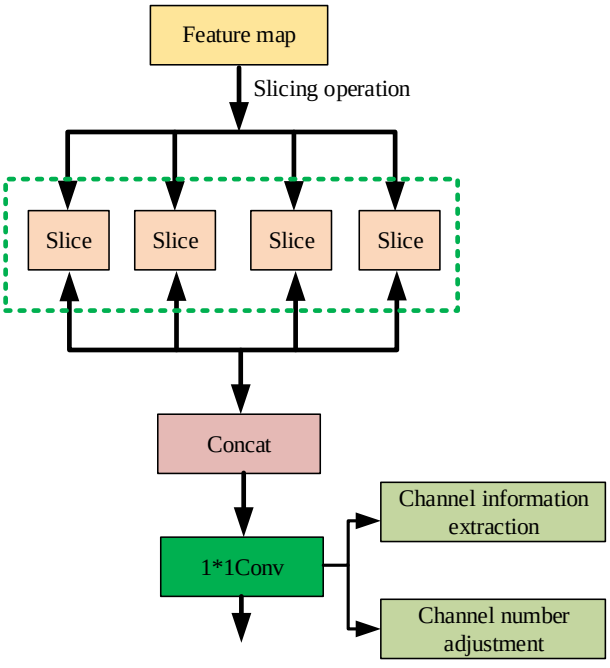


Figure 3: SPD-Conv structure.

In Fig.3, the SPD-Conv structure starts from the input feature map and is segmented into multiple slices by slice operation. These sections are then spliced with Concat together to form a new feature map. Next, the spliced feature maps are processed by 1×1 convolution to adjust the channel number or fused channel information. Meanwhile, the structure also contains the channel information extraction and channel number adjustment module for further processing the channel information. This design enhances the feature extraction capability and flexibility of the model. The workflow of the SPD-Conv module first involves segmenting the feature map into four parts, implementing slicing operations, and then recombining these parts and applying 1×1 convolution kernels. This step is to extract information across channels and adjust the number of channels while maintaining extremely low information loss throughout the entire process. The C2f module is the core component of the Backbone layer, which relies on multiple layers of DarknetBottleneck modules to deeply explore the features of the feature map. In addition, the C2f module integrates the extracted feature information into the output by introducing multi-layer skip connections and concatenation techniques. This method not only effectively extracts features at different levels but also prevents the occurrence of "gradient vanishing/exploding" phenomenon. Considering the redundant computing problem caused by this method, this study introduces PConv to improve the C2f module and calls the improved module C2f-Pairial Convolution (C2f-PC). The structure of C2f-PC is shown in Fig.4.

In the structure of Fig.4, the input first passes through a Split operation to segment the feature graph. Then, the feature graph enters the PConv layer for spatial feature extraction, and the PConv layer can capture the local features in the image more effectively through the local convolution operation. Then, the feature map through multiple PConv layers are spliced (Concat) with the output of conventional convolution (Conv) layers to integrate different levels of feature information. This improved structure helps to improve the performance of YOLOv8 in the target detection task and enhance the feature extraction capability and detection accuracy. This improved structure helps to enhance the performance of YOLOv8 in object detection tasks, strengthen feature extraction capabilities, and improve detection accuracy. Among them, studies have shown that PConv can reduce Floating Point Operations (FLOPs) by about 30%-50% in typical convolutional layers, depending on the size of the input feature map and the number of channels. In addition, PConv can reduce memory access by 20%-40% by reducing redundant calculations and optimizing data access modes [21-22]. PConv effectively reduces the amount of FLOPs required by limiting the number of channels involved in computation, thereby reducing computational complexity. Compared to traditional convolution, PConv can significantly reduce memory access, which is particularly advantageous for devices with limited Input/Output (I/O) resources. Although it only processes a subset of input channels, these selected channels can still play a role in the subsequent pointwise convolutional layers, ensuring that feature information can flow and be utilized between all channels. The final COD model suitable for difficult sample detection is obtained through the above improvements, as shown in Fig.5.

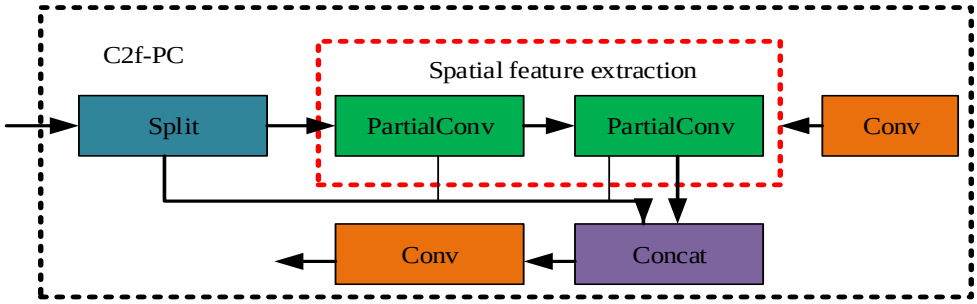


Figure 4: C2f-PC structure

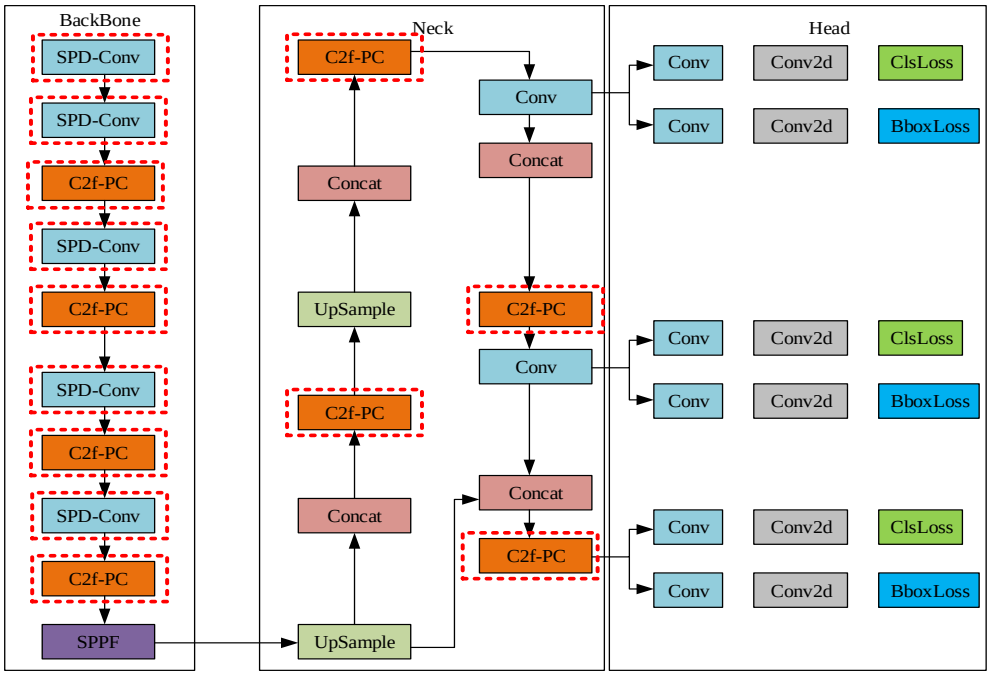


Figure 5: Computer object detection model

In the proposed COD model, SPD-Conv can achieve downsampling without sacrificing information integrity, thereby enhancing the accuracy of SOD. In contrast, PConv utilizes redundant information in feature maps to reduce computational burden and memory access requirements. Specifically, PConv only applies regular convolution to a portion of the input channel to extract spatial features, while keeping the rest channels unchanged. The pseudo-code of the improved COD algorithm based on the loss function and local convolution is shown in Fig.6.

```
// Improved YOLOv8 Algorithm Pseudocode

// Initialize network with improved backbone
function initializeNetwork() {
    backbone = createBackboneWithPConvAndSPDConv()
    head = createDetectionHead()
    return combine(backbone, head)
}

// Create backbone using Partial Conv (PConv) and SPD-Conv
function createBackboneWithPConvAndSPDConv() {
    // Define layers with PConv and SPD-Conv
    layers = []
    for i in range(numLayers):
        if usePConv(i):
            layers.append(createPConvLayer())
        else if useSPDConv(i):
            layers.append(createSPDConvLayer())
        else:
            layers.append(createStandardConvLayer())
    return connectLayers(layers)
}

// Loss function using SIOU and SlideLoss
function computeLoss(predictions, targets) {
    siouLoss = computeSIOU(predictions, targets)
    slideLoss = computeSlideLoss(predictions, targets)
    totalLoss = combineLosses(siouLoss, slideLoss)
    return totalLoss
}

// Training loop
function trainNetwork(data, epochs) {
    network = initializeNetwork()
    for epoch in range(epochs):
        for batch in data:
            predictions = forwardPass(network, batch.input)
            loss = computeLoss(predictions, batch.targets)
            backpropagate(network, loss)
            updateWeights(network)
    return network
}

// Main execution
data = loadData()
epochs = setNumberOfEpochs()
trainedNetwork = trainNetwork(data, epochs)
```

Figure 6: A pseudo-code diagram of an improved COD algorithm based on loss function and local convolution

3 Results

3.1 Model performance testing

To test the effectiveness of the proposed model, this paper focuses on actual monitoring video data of ground service processes at multiple high traffic airports in China, and conducts detailed annotation work based on standard documents issued by the Civil Aviation Administration of China. By writing Python scripts, this study effectively screens and removes small and non-core monitoring targets (This represents a less important goal than core processes such as aircraft take-off and landing) from the data, accurately constructing a dataset suitable for training and testing object detection algorithms. In the processing of airport flight support surveillance video, due to the long video time, this study adopts a fixed time interval strategy (1 frame every 5 seconds) to extract keyframes. For high frame rate videos, the extraction frequency can be appropriately reduced and increased to ensure that important information is not missed.

However, the keyframe extraction operation inevitably introduces a large number of images with highly similar content, leading to the problem of data redundancy. To improve the quality of the data set, the study implements data filtering steps after the data collection. First, the degree of repetition of the image content is quantified by calculating the similarity between the images. Next, a similarity threshold is set. When the similarity between two images exceeds the threshold, they are considered highly similar and only one image is retained. In addition, the study also introduces a time interval strategy to ensure that adjacent frames are not excessively filtered in time, thereby preserving key dynamic information in the video. To further optimize the filtering effect, the K-means clustering algorithm is used to group the images, and only the most representative images are retained in each group. Finally, through this series of data filtering steps, the study successfully removes the redundant images in the data set, significantly improving the diversity and quality of the data. After labeling all the data using the above annotation method, the final COCO dataset consists of 15,118 images, and the dataset is divided in a 4:1 ratio, of which 12,105 are used for training and 3,013 for testing. The ImageNet dataset consists of 24,005 images, divided in a 4:1 ratio, with 19,204 images used for training and 4801 images used for testing. During this algorithm evaluation, mAP is utilized to assess the detection accuracy, computational inference time is taken to evaluate the speed, and FLOPs and parameter count are used to evaluate the computational complexity and complexity of the model. Among them, mAP is calculated by combining Precision and Recall (R). The pixel resolution is 1920*1080. Table 2 shows the configuration of environmental parameters.

In Fig.7 (a), the loss in two datasets gradually decreases with the iteration of learning times. When the last training finished, the training set's loss rate decreases from 0.1800 to 0.1084, and that in the validation set from 0.1362 to 0.0915, showing an improvement in generalization ability. In Fig.7 (b), the prediction accuracy is above 90%, and as the iteration increases, the final accuracy is 96.93% and 98.21%. To test the effectiveness of each part of the algorithm improvement, ablation experiments are performed to evaluate the impact of different improvement strategies on the performance of algorithm under the same conditions. This ablation experiment uses YOLOvs8 as a benchmark and testes it on a self-built dataset. No pre-trained weights are utilized in the tests, as listed in Table 3.

Table 2: Experimental environment and parameter settings

Category	Operating environment
Operating system	Win 7 operating system
Developer Components	JDK 7, Eclipse
CPU	Intel (R) Core (TM) i9-10900X
GPU	Nvidia RTX 3090 GPU
Platform software environment	Python3.8, PyTorch1.10.0
Programming Language	Java
Batch size	8
Batch size	0.001
Batch size	Adam

Firstly, the study model is trained on the training set and the validation set of the COCO data set. The results are shown in Fig.7.

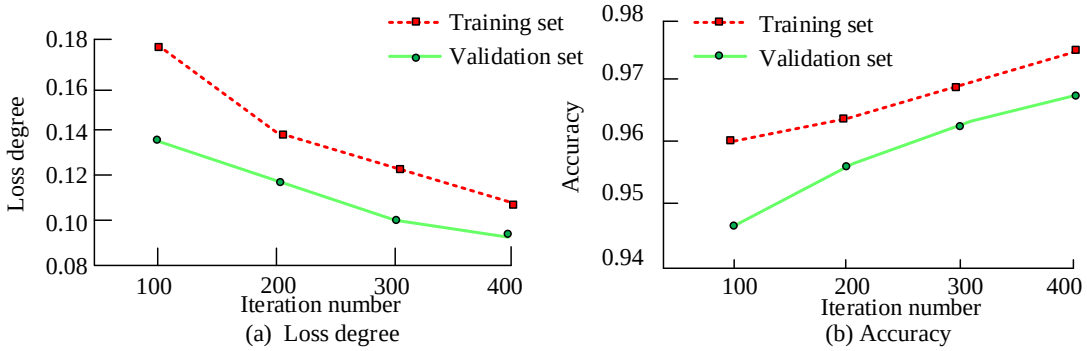


Figure 7: Training results of the research model

Table 3: Results of model ablation experiment

Model		YOLOv8-base	YOLOv8-SPD-Conv	YOLOv8-C2f-PC	YOLOv8-SIoU	Research model
mAP	First	14.3	15.2	14.4	14.5	16.5
	Second	14	15.1	14.7	14.2	16
	Third	14.9	15.6	14.7	15.4	16.4
	Mean	14.4	15.3	14.6	14.7	16.3
	Promotion rate	/	6.25%	1.39%	2.08%	13.19%
AP50	First	27.1	29.1	29.3	27.7	31.2
	Second	27.5	28.9	27.1	28.1	31.8
	Third	27.9	29.6	27.3	28.2	32.7
	Mean	27.5	29.2	27.9	28	31.9
	Promotion rate	/	6.18%	1.45%	1.82%	16.00%
P value		/	$P<0.05$	$P>0.05$	$P>0.05$	$P<0.05$

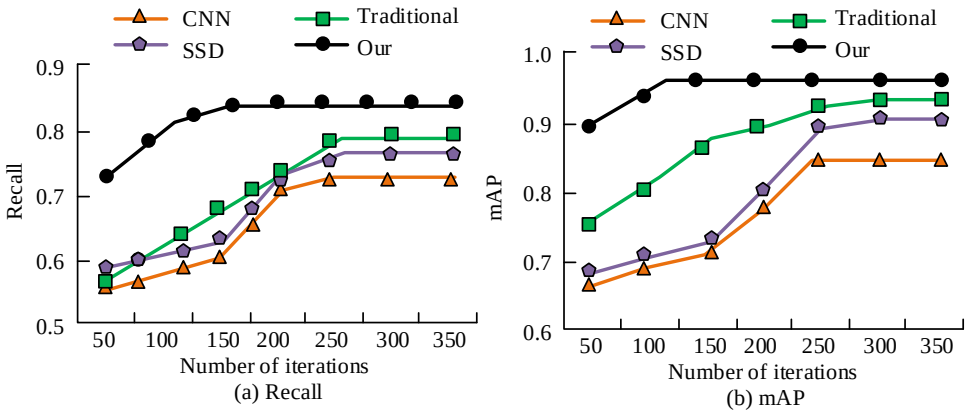


Figure 8: Comparison results of different detection algorithms

In Table 2, after introducing the SPD-Conv structure, mAP and AP50 improve by 6.25% and 6.18% respectively compared to the basic model, effectively solving the problem of reduced detection accuracy caused by information loss in

the benchmark model. After using C2f PC, mAP and AP50 increase by 1.39% and 1.45% compared to the baseline model. This module can enhance the ability of C2f module to learn and express features, and add a SOD layer to strengthen the detection precision. After improving the loss function, the mAP and AP50 increase by 2.08% and 1.82% compared to the basic model, which can accelerate the convergence velocity. The mAP and AP50 of the research model with improved modules increase by 13.19% and 16.00%, with $P<0.05$ indicating statistically significance. The research model can improve the performance of dense object detection and enhance the accuracy. This study selects Single Shot Multi-Box Detector (SSD), CNN, and traditional YOLOv8 as comparative algorithms to test their performance. The corresponding recall rate and mAP results are shown in Fig.8.

In Fig.8 (a), the maximum recall rates for research algorithms, traditional YOLOv8, SSD, and CNN are 0.84, 0.78, 076, and 0.72. Among them, the research algorithm has the best recall performance and converges the fastest, with a 7.7% improvement compared to CNN. In Fig.8 (b), the maximum mAP values of the research algorithm compared to traditional YOLOv8, SSD, and CNN are 0.97, 0.94, 088, and 0.82. Among them, the mAP performance of the research algorithm is the best, with a 3.2% improvement compared to CNN. This indicates that the detection accuracy of the research algorithm is better. To further test the effectiveness, the study tests it on the ImageNet and COCO datasets, as shown in Fig.9.

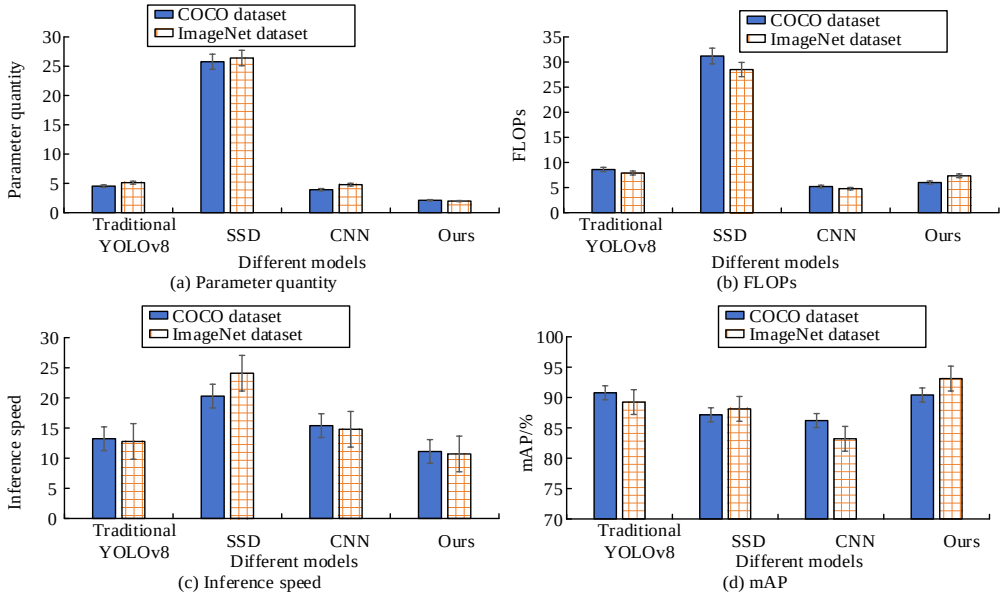


Figure 9: Model test results under different datasets

Figs.9 (a) to (d) show the test values for parameter quantities, FLOPs, inference speed, and mAP of four algorithms. In 9 (a), the proposed algorithm has parameter counts of 2.12 and 1.98 on the COCO and ImageNet datasets, respectively, showing improvements of 53.2% and 61.3% compared to YOLOv8. In 9 (b), the FLOPs of the research algorithm on two datasets are 6.0 and 7.3, which are 30.2% and 7.09% higher than YOLOv8. In 9 (c) and (d), the inference speed and mAP of the designed algorithm on two datasets are 11.1s and 10.7s, 90.7% and 93.1%, which are 16.0% and 16.2%, 14.3% and 13.2% higher than the traditional YOLOv8. The above test results are $P<0.05$, and all the results are statistically significant. This indicates that research algorithms can significantly improve detection speed while enhancing detection accuracy, achieving a balance between the two and making target detection more effective and reliable.

To further verify the robustness of the proposed algorithm in complex environments, the study selects the nuScenes dataset for testing. Among them, the nuScenes dataset is a large-scale autonomous driving dataset. It contains rich sensor data (such as cameras, LiDAR, radar), marked with targets such as vehicles, pedestrians, traffic signs, etc., suitable for tasks such as target detection and tracking. The study processes the dataset in the same way, yielding the nuScenes dataset consisting of 18,425 images, of which 14,740 are used for training and 3,685 for testing. The model has been tested using the same parameters, and the test results are shown in Table 4.

Table 4: Detection results of different models under the nuScenes dataset

Model		Parameter quantity	FLOPs	Inference speed /s	mAP
SSD		4.47	8.9	13.8	78.8
CNN		5.21	7.2	14.9	76.9
Conventional YOLOv8		3.72	8.5	12.3	82.2
Research algorithm		2.33	6.8	9.8	94.2
Compared to the conventional YOLOv8	Promotion rate	37.4%	20.0%	20.3%	14.6%
	P value	$P<0.05$	$P<0.05$	$P<0.05$	$P>0.05$

In Table 4, under the nuScenes dataset, the number of parameters, FLOPs, inference speed, and mAP of the proposed algorithm reach 2.33, 6.8, 9.8s, and 94.2%, respectively, which are all better than the comparison algorithm. Compared with the traditional YOLOv8 algorithm, the various indexes of the proposed algorithm increase by 37.4%, 20.0%, 20.3%, and 14.6%, respectively. Among them, the *P* value of each index except mAP is less than 0.05, which is statistically significant. The results show that the algorithm can balance the detection accuracy and efficiency in complex environment and effectively identify the detection target. The results of the confusion matrix for the different models in the test set of the nuScenes dataset are shown in Table 5.

Table 5: Confusion matrix analysis of the different detection models

Different detection models		The actual case	The actual counterexample	Accuracy rate	Recall	F1 score
SSD	Forecast the case	TP=800	FP=200	0.80	0.84	0.82
	Predicting counterexample	FN=150	TN=1850			
CNN	Forecast the case	TP=750	FP=250	0.75	079	0.77
	Predicting counterexample	FN=200	TN=18D0			
Conventional YOLOv8	Forecast the case	TP=850	FP=150	0.85	0.89	0.87
	Predicting counterexample	FN=100	TN=1900			
Research algorithm	Forecast the case	TP=900	FP=100	0.90	0.95	0.92
	Predicting counterexample	FN=50	TN=1950			

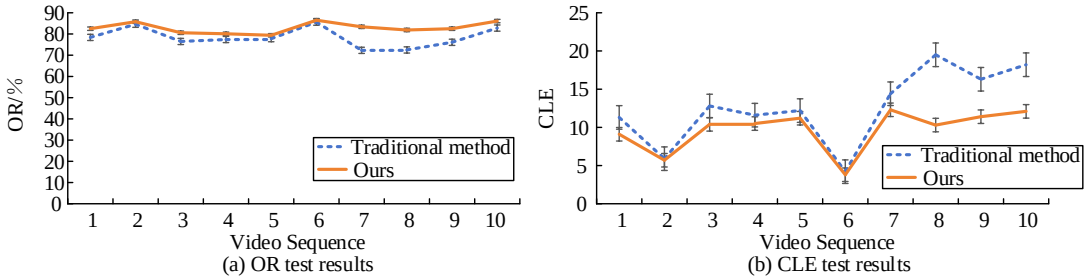
From Table 5, SSD performs moderate and both FP and FN are high, indicating that the model has some problems in false reporting and under-reporting. CNN performs poorly with higher FP and FN, indicating that the model has limited performance in the target detection task. The traditional YOLOv8 performs better with lower FP and FN, indicating that the model has high accuracy in the target detection task. The improved YOLOv8 shows the best performance and the lowest FP and FN, indicating that the improved model has significant advantages in reducing false and under-reporting. In addition, the improved YOLOv8 has obvious advantages, with an accuracy rate of 0.90, which is higher than other models, indicating that the FP is lower. The recall rate is 0.95, which is higher than other models, indicating that the under-reporting rate. The F1 score is 0.92, which has the best comprehensive performance, indicating that it achieves a

better balance between precision rate and recall rate. The improved methods (such as SIoU loss function, SPD-Conv, etc.) can still maintain high detection accuracy and robustness in complex scenarios.

3.2 Example application analysis

To validate the practical performance, this paper applies it to the operation system of a certain fish and shrimp fishing vessel, and compares the detection results of a target crayfish by the research model with the statistical results of traditional methods. Ten video clips are randomly selected from the valid ones that contained fishing crayfish in baskets. Traditional methods are used to count the catch of crayfish in 10 video segments. Ten randomly selected video clips of catching crayfish are inputted into the research model for statistical analysis of the number of crayfish caught. Video clips have a resolution of 1920x1080 pixels with a duration of 10s. This time, Overlap Rate (OR) and Center Location Error (CLE) are chosen as assessment indexes. The detection comparison of traditional and research methods are displayed in Fig.10.

In Fig.10 (a), in terms of OR indicators, the OR of the research models all reaches over 80%, and their average detection results improve by 5.7% compared to traditional methods, $P<0.05$. In Fig.10 (b), in the CLE indicator, the average CLE of the research model is 9.7, which is 2.3% higher than the traditional method. This means that the research method performs greater and can assist in statistical analysis of fishing boat operations, improving operational accuracy. The results of comparing the detection performance of the fish and shrimp fishing vessel operation system using research algorithms with the detection of marine biological targets before use are exhibited in Table 6.



(a) OR test results (b) CLE test results

Figure 10: Traditional and research methods for fish and shrimp fishing vessel detection results

Table 6: Marine biological target detection effectiveness before and after using research algorithms

Category	Pre-use accuracy (%)	Post-use accuracy (%)	*P* value
Close group	97.524	99.322	$P<0.05$
Long distance group	90.396	92.432	$P>0.05$
Dark light group	87.804	89.146	$P>0.05$
Fuzzy group	80.676	84.588	$P<0.05$
Multi-target group	78.408	81.832	$P>0.05$
Twisted group	93.744	97.096	$P<0.05$
Mean	87.912	90.63	$P<0.05$

In Table 6, overall, the detection accuracy of each group is lifted utilizing the research algorithm, with an average increase of 2.7%. Among the tested samples, the closest group has the highest accuracy. Due to the high image clarity and good lighting of the close range group, its accuracy reaches 99.3%, an increase of 2.2% compared to before use. The lowest is the multi-target group, where some images have target occlusion, greatly affecting the accuracy of detection, only 81.8%, but still 3.4% higher than the detection accuracy before use. This result indicates that the research method can effectively identify detection targets and improve the efficiency of fishing operations for fishing vessels.

4 Discussion

The proposed improved YOLOv8 target detection model, which integrates the SIoU loss function and local convolution, demonstrated superior performance compared to traditional YOLOv8 and other YOLO-based models. Experimental results indicated that the improved loss function increased mAP and AP50 by 2.08% and 1.82%, respectively, aligning with findings from Weiya et al. [7], who proposed a RS SOD method based on cross-layer fusion and weighted receptive fields. This improvement was attributed to the introduction of SIOU loss function and SlideLoss function. The SIOU function optimized bounding box regression by combining target shape information, while the SlideLoss function enhanced dense object detection by dynamically adjusting loss weights. Additionally, the integration of SPD-Conv and PConv in the Backbone layer further boosted the model's performance. SPD-Conv increased mAP and AP50 by 6.25% and 6.18%, respectively, while PConv added 1.39% and 1.45%, surpassing the cascade model proposed by YANG et al. [11]. SPD-Conv reduced downsampling loss, improving SOD, while PConv enhanced computational efficiency. These modifications not only elevate detection accuracy but also significantly reduce computational complexity, achieving a balance

between precision and efficiency.

The proposed method not only performed well on the COCO and ImageNet datasets, but also demonstrated its applicability in a broader range of scenarios. For example, in RS images, UAV aerial images, and target detection tasks in complex backgrounds, the models all showed high detection accuracy and robustness. This benefits from the enhancement of SOD by SPD-Conv and the full utilization by the SIOU loss function [16]. In addition, the introduction of PConv enabled the model to operate efficiently in a constrained computational resource environment, further expanding its application scenarios. In target detection tasks, there was often a trade-off between computational complexity and detection accuracy. However, by introducing PConv and SPD-Conv, the computational complexity of the model was significantly reduced while improving the detection accuracy. The experimental results showed that the improved model achieved significant optimization in both FLOPs and the number of parameters, while also improving the inference speed. This trade-off made models more competitive in real-time detection tasks, such as autonomous driving, video surveillance, and drone target detection.

5 Conclusion

At present, traditional object detection techniques can no longer satisfy the demands of reality. In response to the issues of low recall and large computational resources in the current YOLOv8 algorithm, this study proposed a YOLOv8 model based on loss function and local convolution improvement. The model first introduced SIOU to enhance the loss function of the algorithm and used SlideLoss to adjust its weight calculation method. Then, this study introduced SPD-Conv and more efficient PConv to improve the Backbone layer. The results found that the research model had high prediction accuracy on the training set and validation set. After introducing SPD-Conv structure, using C2f-PC, and improving the loss function, the mAP and AP50 indexes of the model were significantly improved compared with the basic model. This indicated that the model could significantly improve the performance and accuracy of intensive target detection. At the same time, on the COCO and ImageNet datasets, compared with the traditional YOLOv8 algorithm, the number of parameters and FLOPs have increased, but the inference speed has been improved, and mAP has also been significantly improved. This achieved a balance between detection accuracy and detection speed, improving the reliability and effectiveness of object detection. In the case test, the average OR and CLE of the research model improved by 5.7% and 2.3%, which can effectively improve the accuracy of the task. However, this study found that the accuracy is greatly affected by the monitoring scenario during the small sample object detection experiment. Future research will attempt to obtain more scene data, have surveillance video data with a wider time span, and use larger computing resources and more sophisticated models to conduct in-depth research on this specific problem. Furthermore, the study only conducted experiments with fixed parameters, and future work will discuss head-to-head training comparisons with different parameter settings.

References

[1] Carmen Gheorghe, Mihai Duguleana, Razvan Gabriel Boboc, Cristian Cezar Postelnicu. Analyzing Real-Time Object Detection with YOLO Algorithm in Automotive Applications: A Review. Computer Modeling in Engineering & Sciences,2024,141(12):1939-1981. DOI: 10.32604/cmes.2024.054735

[2] Mahmoud Atta Mohammed Ali, Tarek Aly, Atef Tayh Raslan, Mervat Gheith, Essam A. Amin. Advancing Crowd Object Detection: A Review of YOLO, CNN and ViTs Hybrid Approach. Journal of Intelligent Learning Systems and Applications,2024,16(3):175-221. DOI: 10.4236/jilsa.2024.163011

[3] Zhao M, Li W, Li L. Single-frame infrared small-target detection: A survey. IEEE Geoscience and Remote Sensing Magazine, 2022, 10(2): 87-119. DOI: 10.1109/MGRS.2022.3145502

[4] Lei F, Tang F, Li S. Underwater target detection algorithm based on improved YOLOv5. Journal of Marine Science and Engineering, 2022, 10(3): 310.

[5] Li H, Wu Z, Shrivastava A. Rethinking pseudo labels for semi-supervised object detection. Proceedings of the AAAI conference on artificial intelligence. 2022, 36(2): 1314-1322. DOI: 10.1609/aaai.v36i2.20019

[6] Yang J, Shi S, Wang Z. St3d++: Denoised self-training for unsupervised domain adaptation on 3d object detection. IEEE transactions on pattern analysis and machine intelligence, 2022, 45(5): 6354-6371. DOI: 10.48550/arXiv.2103.05346

[7] Weiya Shi, Shaowen Zhang, Shiqiang Zhang. CAW-YOLO: Cross-Layer Fusion and Weighted Receptive Field-Based YOLO for Small Object Detection in Remote Sensing. Computer Modeling in Engineering & Sciences,2024,139(6):3209-3231.

[8] Zhongyuan Zhang, Wenqiu Zhu. YOLO-MFD: Remote Sensing Image Object Detection with Multi-Scale Fusion Dynamic Head.Computers, Materials & Continua, 2024,79(5):2547-2563.

[9] Yakui Liu, Xing Jiang, Ruikang Xu, Yihao Cui, Chenhui Yu, Jingqi Yang, Jishuai Zhou. A Novel Foreign Object Detection Method in Transmission Lines Based on Improved YOLOv8n.Computers, Materials & Continua,2024,79(4):1263-1279. DOI: 10.32604/cmc.2024.048864

[10]Xiangyan Tang, Chengchun Ruan, Xiulai Li, Binbin Li, Cebin Fu. MSC-YOLO: Improved YOLOv7 Based on Multi-Scale Spatial Context for Small Object Detection in UAV-View.Computers, Materials & Continua,2024,79(4):983-1003. DOI: 10.32604/cmc.2024.047541

[11]YANG Yuyi, CHEN Liang, ZHANG Jian, LONG Lingchun, WANG Zhenfei. UGC-YOLO: Underwater

Environment Object Detection Based on YOLO with a Global Context Block. Journal of Ocean University of China,2023,22(3):665-674. DOI: 10.1007/s11802-023-5296-z

[12]Abdulghani M.Abdulghani, Mokhles M. Abdulghani, Wilbur L. Walters, Khalid H. Abed. Multiple Data Augmentation Strategy for Enhancing the Performance of YOLOv7 Object Detection Algorithm.Journal on Artificial Intelligence,2023,5(1):15-30. DOI: 10.32604/jai.2023.041341

[13]Ananthakrishnan Balasundaram, Anshuman Mohanty, Ayesha Shaik, Krishnadoss Pradeep, Kedalu Poornachary Vijayakumar, Muthu Subash Kavitha. Zero-DCE++Inspired Object Detection in Less Illuminated Environment Using Improved YOLOv5.Computers, Materials & Continua,2023,77(12):2751-2769. DOI: 10.32604/cmc.2023.044374

[14]Li M, Dong H, Zhang F. A method for top view pedestrian flow detection based on small target tracking. Informatica, 2024, 48(11):1813.-1830. DOI: 10.31449/inf.v48i11.6033

[15]Liu K, Sun Q, Sun D. Underwater target detection based on improved YOLOv7. Journal of Marine Science and Engineering, 2023, 11(3): 677. DOI: 10.1109/ICICML57342.2022.10009683

[16]Yang R, Li W, Shang X. KPE-YOLOv5: an improved small target detection algorithm based on YOLOv5. Electronics, 2023, 12(4): 817. DOI: 10.3390/electronics12040817

[17]Xie Z, Wu G. Optimized Method for Basketball Game Judging by Integrating Faster-RCNN with LK Algorithm. Informatica, 2024, 48(23): 17-31: 3367.-3372. DOI: 10.31449/inf.v48i23.6696

[18]Chen G, Hou Y, Cui T. YOLOv8-CML: A lightweight target detection method for Color-changing melon ripening in intelligent agriculture. Scientific Reports, 2024, 14(1): 14400. DOI: 10.1038/s41598-024-65293-w

[19]Hui Y, You S, Hu X. SEB-YOLO: An Improved YOLOv5 Model for Remote Sensing Small Target Detection. Sensors, 2024, 24(7): 2193. DOI: 10.3390/s24072193

[20]He J, Luo J, Fu C. The Occlusive Basketball Player Detection Algorithm Based on Posture Recognition Assisted Feature Alignment. Informatica, 2024, 48(21):114-11. DOI: 10.31449/inf.v48i21.6695

[21]Wang Y, Tian Y, Liu J,. Multi-stage multi-scale local feature fusion for infrared small target detection Remote Sensing, 2023, 15(18): 4506-4520. DOI: 10.3390/rs15184506

[22]Ling S, Chen L, Wu Y. ACANet: Attention-based context-aware network for infrared small target detection. The Journal of Supercomputing, 2024, 80(12): 17068-17096. DOI: 10.1007/s11227-024-06067-z

Hybrid Phishing Detection Using Stochastic Gradient Descent and Naïve Bayes Optimized with the Mayfly Algorithm

Xiao Chen
Computer Engineering Technical College Guangdong Institute of Science and Technology, Zhuhai 519000, China
E-mail: 13277095763@163.com

Keywords: phishing, phishy attack, protecting sensitive data, machine learning, metaheuristic algorithms, hybrid models

Received: January 15, 2025

Because hackers were able to access AOL user credentials in 1996, phishing, a malicious method of obtaining personal data, became a significant online threat. This fraudulent practice makes use of email and website spoofing techniques to trick victims into disclosing sensitive information. Advanced practices that make use of users' trust and web vulnerabilities, such as spear phishing and tab nabbing, may be hazardous to people's security. In the classification of phishing websites, this research used two prediction models: the Stochastic Gradient Descent (SGD) and the Naïve Bayesian Classification Algorithm (NBC). Hybrid models were developed by incorporating the Mayfly Optimization Algorithm (MOA), a sophisticated optimization method for improving predictive accuracy and overall performance. The dataset contained two stages with a total of 1,353 phishing, trustworthy, and dubious websites. Hyperparameters tuned using random search method for each hybrid model. The dataset contains nine input parameters and derived from previous studies.The results indicated that, with an accuracy of 0.921 during the testing phase, the hybrid model of SGD+MO fared best. On the other hand, the NBC model with Accuracy of 0.877 identified as the weakest model with 4.4% different compared to best model. Also, further improved performance was demonstrated by the numerical classification results for the various categories: it was observed that for phishing websites, the precision metric was 0.925; for suspicious websites, it was 0.933; while for legitimate websites, the precision was 0.911. These results point out the hybrid model's ability to enhance phishing detection systems by showing how well it classifies and detects different kinds of websites.

Povzetek: Prispevek predstavi hibridni model za odkrivanje ribarjenja, ki združuje SGD in Naïvni Bayes, optimiziran z algoritmom Mayfly, kar prinaša znatno izboljšano natančnost zaznave.

1 Introduction

1.1 Study background

Phishing is a fraudulent tactic used to deceive individuals into disclosing sensitive data online [1,2]. Phishers are the attackers that organize phishing attacks. When hackers obtained AOL users' login credentials in 1996, phishing emerged as a significant online threat [3,4]. A successful phishing attack is typically carried out through website spoofing techniques [5,6] and email spoofing techniques [7]. Attackers who want to trick victims into believing they are communicating with legitimate organizations like banks, credit card companies, or government agencies [8,9] begin by sending spoof emails [10]. Since their source addresses are altered to resemble emails from reliable sources, the email addresses are spoofs. For instance, if a bank manager at bank "XYZ" has the email address bankmanager@xyz.co.in, the attacker will attempt to spoof that address to trick the user into thinking the email is legitimate and following the phisher's instructions [11]. Usually, the email requests that the recipient click on links to websites and reply to the message or the website with their banking information. Email spoofing uses open SMTP (Simple Mail Transfer Protocol) servers to send bogus emails to targets. Another deceptive technique is to create phishing websites that look and feel exactly like the targeted authentic websites since many users are reluctant to divulge personal information in response to an email [12–14].

Phishing is known as "Whaling" when it goes after well-known users [6]. Phishers employ diverse techniques to execute their deception successfully. These tips consist of the following: The first method involves manipulating links so that they appear to lead to a legitimate website, but in reality, they lead to a malicious or phished URL; the second method involves avoiding phishing detection filters [15]. (iv) utilizing pop-up windows to solicit user names and passwords; (iv) using Javascript to conceal the browser address bar and construct a custom address bar that presents a hard-coded legitimate URL to the user [12]. Use pictures rather than words, which might evade detection by several phishing filters [16].

Various researchers have investigated various phishing website identification methods based on differentiating characteristics in the recent past. Rami M et al. [17] performed a thorough investigation on automatic extraction and feature analysis that could differentiate between legitimate and phishing websites using

automated tools. Their analysis listed 17 important features, the most important being "Request URL," followed by "Age of Domain" and "HTTPS and SSL." For the better utilization of these features in phishing recognition systems, they created specific rules for each feature. They established in a series of experiments that the C4.5 algorithm performed more accurately than other rule-based classification algorithms such as RIPPER, PRISM, and CBA. They also improved the prediction accuracy by focusing on the nine most informative features; the lowest error rate was 4.75% with the CBA algorithm.

Abburous et al. [18] also worked on the classification techniques of data mining for phishing detection in e-banking. They indicated that factors such as "Page Style and content" and "Social Human Factor" had little influence on the final phishing recognition rate; however, they emphasized the critical roles that URL, Domain Identity, Security, and Encryption played. By highlighting the relationships between specific characteristics in their associative classification model, they were able to develop systems that could detect phishing. Tests indicated that Associative Classification methods outperformed more conventional algorithms, such as MCAR, which had an error rate of 12.622%. They suggested that future research apply various pruning techniques to raise the accuracy and efficiency of classifiers.

Ramesh et al. [19] finding the harmed domains is crucial in phishing detection; they presented an automated approach using a unique Target Validation algorithm to ensure the accuracy of finding target domains of phishing webpages. By analyzing a fake relationship, their results enhanced protection against online identity attacks and had more than 99% for detecting harmed domains.

The last new multi-label rule-based classification algorithm, EMCAC, was presented by Neda Abdelhamid [20]. Its purpose is to generate multi-class-labeled rules without needing recursive learning. Experimental results based on the phishing data demonstrated that the EMCAC outperformed algorithms such as CBA, MCAR, MMAC, PART, C4.5, and RIPPER. They discovered a more manageable, helpful set of features for website type detection based on Chi-square feature selection. Future research will apply EMCAC to unstructured data in text categorization.

These works collectively contribute to improving phishing detection methodologies in general by emphasizing effective features, improving classification algorithms, and proposing new methods to enhance the accuracy and efficiency of phishing detection systems.

Table 1 reports a summary of the existing articles in the study field.

Nomenclature

SGD	Stochastic Gradient Descent	F_i^{t+1}	Updated position of female iii at iteration t+1
NBC	Naïve Bayesian Classification	D_{mf}	Distance between male and female
MOA	Mayfly Optimization Algorithm	W	Random walk parameter (set to 1)
SVM	Support Vector Machines	r	Random number
DT	Decision Trees	$H(F_i)$	Fitness value of female i
KNN	k-Nearest Neighbors	$H(M_i)$	Fitness value of male i
PSO	Particle Swarm Optimization	$X_{i,j}^t$	Position of individual iii in dimension j at iteration t
GA	Genetic Algorithms	$Y_{i,j}^t$	Position of another reference individual in dimension j at iteration t
M_i^t	Initial position of male i at iteration t	$\emptyset_{c_i}$	Mean of feature x_t for class C_i
M_i^{t+1}	Updated position of male iii at iteration t+1	σ_{c_i}	Standard deviation of feature x_t for class c_i
V_i^t	Initial velocity of male iii at iteration t	A(Ci)	Prior probability of class C_i
V_i^{t+1}	Updated velocity of male i at iteration t+1	A(X)	Normalization factor (marginal probability of X)
$V_{i,j}^t$	Initial velocity component for male iii in dimension j at iteration t	TP	True Positives: Correctly classified positive samples.
$V_{i,j}^{t+1}$	Updated velocity component for male iii in dimension j at iteration t+1	TN	True Negatives: Correctly classified negative samples.
S1	Personal learning parameter (set to 1)	FP	False Positives: Negative samples incorrectly classified as positive.
S2	Social learning parameter (set to 1.5)	FN	False Negatives: Positive samples incorrectly classified as negative.
β	Exponential decay parameter (set to 2)	$P_{best}(i,j)$	Best personal position of male iii in dimension j
Dp	Cartesian distance between male position and personal best position	$Q_{best}(i,j)$	Best global position in dimension j
Dg	Cartesian distance between male position and global best position	R	Random number in the range [-1,1]
F_i^t	Initial position of female i at iteration t	d	Nuptial dance value (set to 5)

Table 1: Summary of the previous studies.

Authors	References	Techniques/Models Used	Dataset Used	Performance Metrics
Rami M et al.	[17]	C4.5, RIPPER, PRISM, CBA, Rule-based classification	Not specified	CBA: 4.75% error rate, C4.5 performed best
Abburous et al.	[18]	Associative Classification, MCAR, Data Mining techniques	E-banking data	MCAR: 12.622% error rate, Associative performed best
Ramesh et al.	[19]	Target Validation Algorithm, Automated phishing domain detection	Not specified	99% accuracy in detecting harmed domains
Neda Abdelhamid	[20]	EMCAC (Multi-label classification), CBA, MCAR, MMAC, PART, C4.5, RIPPER	Phishing dataset	EMCAC outperformed others, Feature selection via Chi

1.2 Objective of the study

Phishing attacks pose a significant and evolving cybersecurity threat, necessitating the development of accurate and robust detection systems. Traditional rule-based and blacklist-based approaches often fail to detect new and sophisticated phishing techniques due to their static nature. To address this challenge, this study proposes a hybrid machine learning-based approach to improve phishing website classification accuracy and adaptability.

This study investigates the effectiveness of integrating the Mayfly Optimization Algorithm (MOA) with two machine learning models: Naïve Bayes Classifier (NBC) and Stochastic Gradient Descent (SGD). The hypothesis tested is that the hybrid approach—leveraging the strengths of probabilistic and gradient-based learning models with a bio-inspired optimization algorithm—enhances predictive accuracy, convergence speed, and robustness against evolving phishing techniques compared to standalone models.

❖ **Justification for model selection:**

Stochastic Gradient Descent (SGD) was chosen due to its ability to efficiently handle large-scale datasets, making it well-suited for real-time phishing detection. Unlike traditional classifiers such as Support Vector Machines (SVM), which can be computationally expensive in high-dimensional spaces, SGD updates model parameters iteratively, ensuring fast adaptation to new patterns in phishing data. Compared to Decision Trees (DT) and k-Nearest Neighbors (KNN), which may struggle with scalability and feature complexity, SGD offers superior generalization, making it ideal for dynamic and evolving phishing attack scenarios. Similarly, the Naïve Bayes Classifier (NBC) was selected due to its probabilistic nature, which allows for interpretable decision-making and robust performance in high-dimensional spaces. Unlike deep learning models like Neural Networks, which require extensive training data and computational resources, NBC remains computationally efficient and effective even with limited training samples. Compared to Logistic Regression, which assumes a linear decision boundary, NBC handles non-linearity better due to its probabilistic assumptions, making it a strong candidate for phishing classification tasks.

❖ **Justification for optimizer selection:**

MOA was incorporated to enhance the predictive accuracy of SGD and NBC by optimizing hyperparameters and improving model convergence. Unlike traditional optimization techniques like Grid Search and Random Search, which are computationally expensive, MOA dynamically balances exploration and exploitation, enabling more efficient searching for optimal model parameters. Compared to Particle Swarm Optimization (PSO) and Genetic Algorithms (GA), MOA exhibits superior convergence speed and stability due to its biologically inspired swarming mechanism, making it well-suited for improving phishing detection models.

❖ **Novelty of the work:**

The novelty of this research lies in the hybridization of MOA with SGD and NBC to enhance phishing website classification. Unlike previous studies that rely on standalone machine learning models or traditional optimization techniques, this study introduces a bio-inspired optimization approach to fine-tune machine learning classifiers, improving their convergence and predictive performance. The proposed hybrid models demonstrate superior robustness against evolving phishing attacks, providing a scalable and interpretable solution for real-time cybersecurity applications.

By leveraging the efficiency of SGD, the interpretability of NBC, and the adaptive search capabilities of MOA, this study contributes a novel and effective approach to phishing detection. The results validate the hypothesis that the integration of optimization

algorithms with machine learning models enhances phishing detection accuracy, offering a more reliable and adaptable defense mechanism against cyber threats.

1.3 Paper organization

The structure of the article is: Section 2 describes the procedures of data collection and preprocessing in detail. Section 3 introduces the ML models for phishing detection, and Section 4 explains the metaheuristic algorithm for the optimization of these models. Section 5 presents the appraisal factors utilized to examine the execution of the model, including accuracy and F1-score. In Section 6, the results are presented both numerically and visually, and the findings are discussed. The study's summary of the main conclusions and recommendations for future research directions are given in Section 7. The references section concludes with a list of all cited works for additional reading.

2 Data collection

A reliable database derived from [21] for monitoring phishing websites, Phish Tank, provided 1353 inputs for the dataset used in this study. The features present in the dataset for identifying phishing websites include Prefix/Suffix, SFH, Request URL, Web Traffic, Anchor URL, URL Length, Domain Age, Sub Domain, and IP Address. These characteristics were chosen because of their potential to be useful during the classification process, where one tries to determine whether a website is phishing or authentic. To construct the models and improve them during the training phase, 947 inputs were used. Thus, the remaining 406 inputs were used during the testing phase to confirm the functionality of the models. This division makes sure the models are adequately trained and subjected to a tight evaluation process for determining generalizability and predictive accuracy.

Fig. 1 presents the impact of each input parameter on the website classification outcomes. Among the parameters, the prefix/suffix parameter has shown the most positive impact, with a significant value of 0.24. The domain age parameter then shows a positive influence with a value of 1.7. On the other hand, residual parameters have adverse effects on the classification results. While stressing that other features may reduce predictive performance, this analysis emphasizes the significance of particular features, such as the prefix/suffix and domain age, in improving the accuracy of phishing website detection.

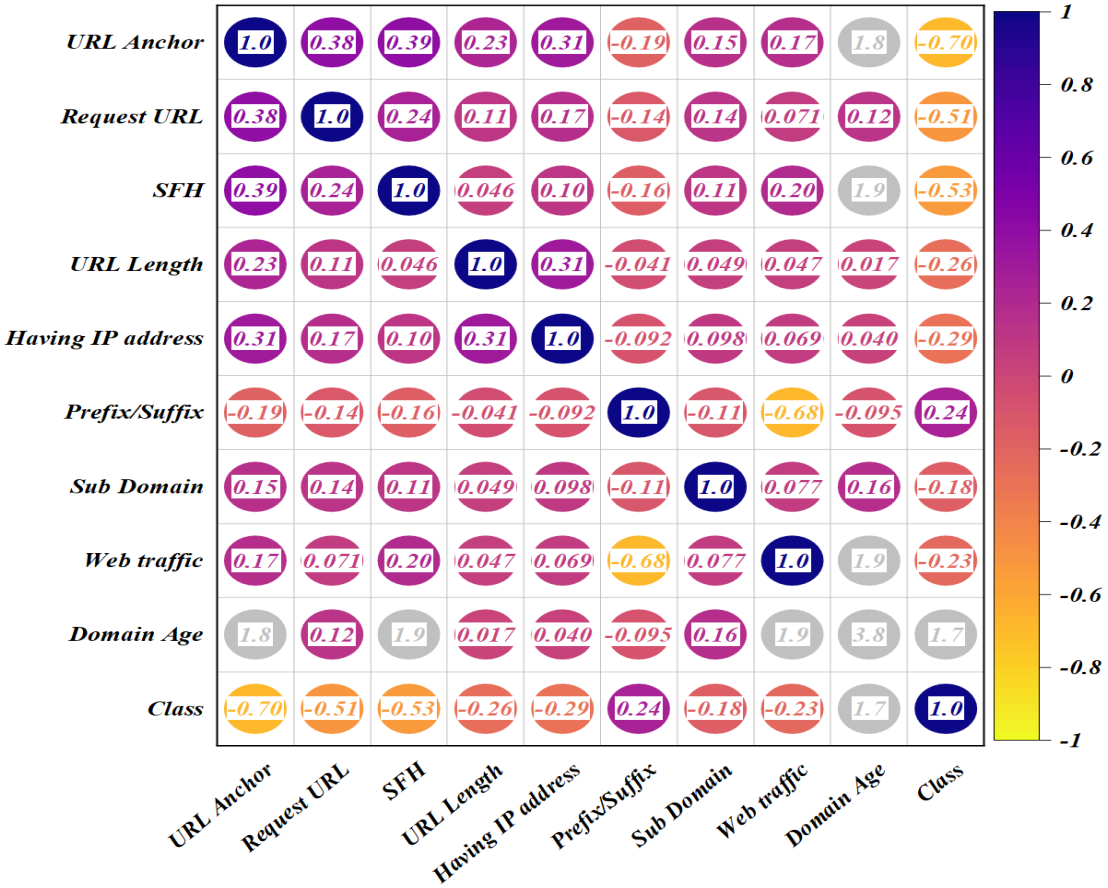


Figure 1: The plot illustrates the correlation between the input and output

3 Machine learning models

3.1 Stochastic gradient descent (SGD)

SGD is widely used and has demonstrated advanced execution on numerous ML tasks, as demonstrated in [22] and [23]. To specify the length of the subsequent stage to take when going in the gradient's direction, SGD incorporates a parameter known as the learning rate. A learning rate that is too small results in slow convergence, while a learning rate that is too large can impede convergence and cause the loss function to diverge or even oscillate around the minimum. For this work, a learning rate has been chosen that is modified by a predetermined schedule as the training progresses. However, n-SGD has a more significant step size than SGD because it is more difficult to take more significant steps in one point in SGD without running the risk of preventing convergence due to the noise on the gradient. The model in n-SGD is updated using minibatches, which are small collections of training samples. Here, the batch size is set to n, which is the total number of points to take into account when calculating the gradient.

3.2 Naïve Bayesian classification algorithm (NBC)

The NBC organization method functions on the assumption that every model's attributes are free within their respective groups. These are assumptions that most people find annoying, but NBC consistently performs well in verified scenarios. It can predict the possibility of being placed in a particular class based on tests from that class by computing contingent probabilities using Likelihood Bayes [24,25]. One advantage of this approach is that it expects property freedom, which means that order can be established with essentially a change in a variable at the class level instead of the entire covariance structure.

Determining the mean and standard deviation of the highlights in the preparation materials for each class is the most crucial step in this classification technique [26,27]. At that point, the processed mean and standard deviation are utilized to determine the probability $A(x_t|c_i)$.

$$A(x_t|c_i) = g\left(x_t, \varphi_{c_i}, \sigma_{c_i}\right) = \frac{1}{\sqrt{2\pi}\sigma_{c_i}} exp^{-\frac{(x_t-\varphi_{c_i})^2}{2\sigma_{c_i}^2}} \quad (1)$$

The Gaussian thickness capacity for the quality x_t is addressed by the articulation $g\left(x_t, \varphi_{c_i}, \sigma_{c_i}\right)$. φ_{c_i} and σ_{c_i} represent the standard deviation of characteristic x_t in the preparation information for the class.

The likelihood $A(x_t|c_i)$ for each trademark in each class is replicated by the likelihood $A(X|C_i) = \pi_{(t=1)}^{n} A(x_t|C_i)$, which yields the probability $A(x_t|c_i)$ for each class. The back likelihood, or $A(x_t|c_i)$, can be found by multiplying this $A(x_t|c_i)$ by the earlier probability of each class.

To get the probability $A(x_t|c_i)$ for each class, the likelihood $A(x_t|c_i)$ for each trademark in each class is replicated by the likelihood $A(X|C_i) = \pi_{(t=1)}^{n} A(x_t|C_i)$. By multiplying this $A(x_t|c_i)$ by the earlier probability of each class, one may find the back likelihood or $A(C_i|X)$.

$$A(C_i|X) = \frac{(X|C_i)\, A(C_i)}{A(x)} \quad (2)$$

When $1 \leq j \leq m$ and j isn't equal to i, the test data are categorized into a specific class according to whether or not they satisfy the measures $A(C_i|X) > A\left(C_j|X\right)$. In Fig. 2, the NBC flowchart is displayed.

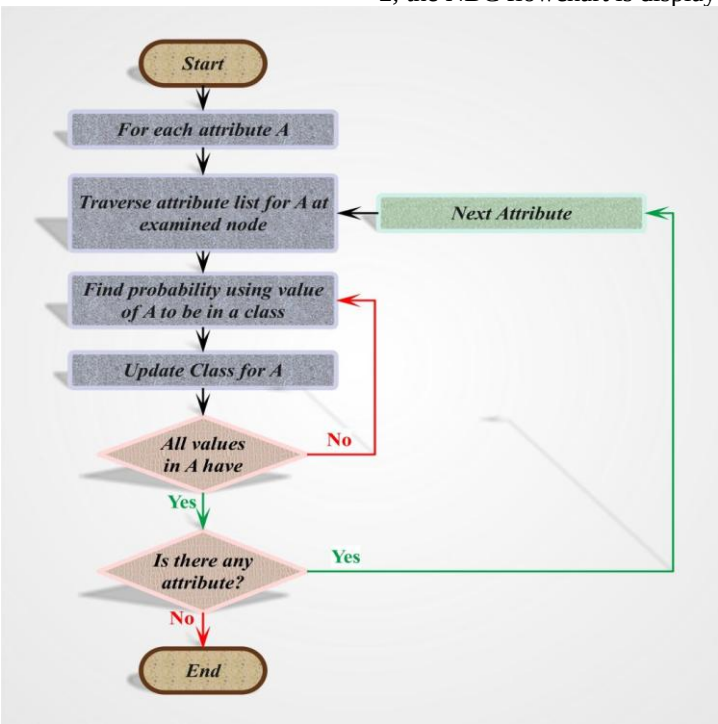


Figure 2: NBC Flowchart [28]

4 Metaheuristic algorithms

4.1 Algorithm for optimizing mayfly (MOA)

MOA is a population-based technique that was developed in 2020 [29,30]. The PSO is the source of the MOA and combines all of the main advantages of the PSO, FA, and GA, as stated in the authors' prior claim. The activities listed below comprise the MFO concept: i) a pair of equal numbers of male and female agents are first sent out; ii) the male mayfly chooses which Q_{best} is appropriate for the current task; iii) the female mayfly searches and merges with the male mayfly at Q_{best} ; iv) procreating, and v) the search is terminated and the results are displayed.

The total efficacy of the MOA is determined by the male's initial location and the distance at which it attracts the female. This approach allows an equal number of male and female agents to be freely initiated in the search space, allowing each Mayfly to converge toward the Q_{best} with a rise in convergence. This procedure will terminate when every couple of agents produces the same number of children. Every generated offspring is assigned a zero velocity to end the process and prevent it from moving further. The remaining pertinent details are located in [29,30], and [31] provide access to the basic code.

In a d-dimensional search area, assume that there are an equal number of male (M) and female (F) Mayflies. The overall count of agents (Mayflys) is represented by the numbers i=1, 2..., N. Every agent is randomly initialized in the search locality during the optimization search, and as the iteration count rises, each agent is permitted to advance toward the best position (Q_{best}). By changing its position and speed, the male can get to the Q_{best}. The agent will travel in the direction of its target based on the Cartesian distance and the increasing iteration. Similar to this procedure is the FA discussed in [32]. The revised position and velocity are displayed in Eqs. (3) and (4);

$$M_i^{t+1} = M_i^t + V_i^{t+1} \tag{3}$$

$$V_{i.j}^{t+1} = V_{i,j}^t + S_1 \times e^{-\beta {D_p}^2}\left(P_{best_{i,j}} - M_{i,j}^t\right) + S_2 \times e^{-\beta {D_g}^2}\left(Q_{best_{i,j}} - M_{i,j}^t\right) \tag{4}$$

where $\beta = 2$, personal learning parameters (S_1)=1, (S_2)=1.5, M_i^t and M_i^{t+1} are the initial and modified positions, V_i^{t+1} and $V_{i,j}^{t+1}$ are the initial and modified velocities, and D_p and D_g are the Cartesian distances. The FA and PSO values are combined to form the frame in (4). To seduce the female (F) with a distinctive nuptial dance (dancing up and down on a water surface), every man (M) will attain the Q_{best} when the updation continues depending on the advancement in cycles. The definition of the velocity update during this procedure is as follows:

$$V_{i,j}^{t+1} = V_{i,j}^t + d \times R \tag{5}$$

where R = random numeral [-1], [1], and nuptial dance value $(d) = 5$.

Each female (F) is then allowed to determine the guy who is at G_{best} once the male has finished his optimal search. which a female (F) may use a random walk ($W = 1$) value to escape to a new location or move in the direction of the male (D_{mf}). The update of location and velocity for F may be expressed mathematically in Eq. (6) and Eq. (7):

$$F_i^{t+1} = F_i^t + V_i^{t+1} \tag{6}$$

$$V_{i,j}^{t+1} = \begin{cases} V_{i,j}^t + S_2 \times e^{-\beta {D_{mf}}^2}\left(X_{i,j}^t - Y_{i,j}^t\right) & if\ H(F_i) > H(M_i) \\ V_{i,j}^t + W \times r & if\ H(F_i) \le H(M_i) \end{cases} \tag{7}$$

H stands for the maximized objective value.

As the number of iterations grows, each F will ultimately become the correct M, resulting in the birth generation. The product of M and F displays the overall count of offspring in the MOA. After mating, the progeny of M and F will thus always have an initial velocity value of zero. Only M and F's search performance is considered in the IMLT issue since the initial velocity operator of the offspring is neutralized. The best characteristics of the FA, PSO, and FA are merged to generate the recommended MOA, as shown by the previously described equations. Additional relevant information is available from [29,30]. Fig. 3 shows the flowchart of MOA.

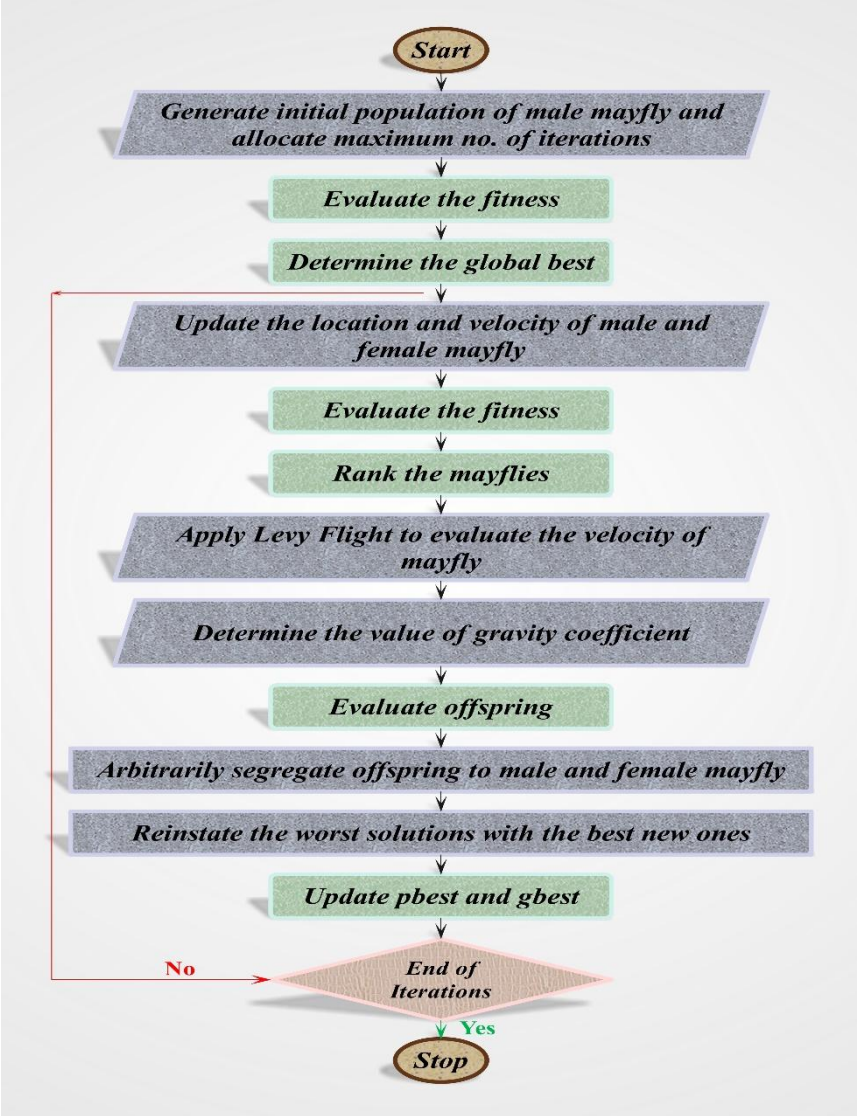


Figure 3: MOA flowchart

5 Performance evaluators

To assess the performance of the classification models, four key evaluation metrics are considered: Accuracy, Precision, Recall, and F1-Score. These metrics provide a quantitative measure of the model's effectiveness in classifying data and its resilience against misclassification.

- **Accuracy:** Accuracy is the most straightforward metric, representing the proportion of correctly classified instances in the dataset. It is defined as:

$$\text{Accuracy} = {(F^P + T^N)}/{(F^P + T^N + F^P + F^N)} \quad (8)$$

where **TP** (True Positives) and **TN** (True Negatives) denote correctly classified samples, while **FP** (False Positives) and **FN** (False Negatives) indicate misclassified samples. Although accuracy provides an overall performance measure, it may not be reliable for imbalanced datasets, where one class dominates.

- **Precision:** Precision evaluates how well the model identifies positive instances while avoiding false positives. It is given by:

$$\text{Precision} = {T^P}/{(T^P + F^P)} \quad (9)$$

A high precision score indicates that the model minimizes false positives, making it particularly valuable in applications where false alarms must be reduced.

- **Recall:** Recall (also known as Sensitivity) measures the model's ability to correctly identify all relevant instances. It is expressed as:

$$\text{Recall} = F^P R = {F^P}/{P} = {F^P}/{(F^P + F^N)} \quad (10)$$

A high recall means that the model effectively captures all true positive cases, reducing the likelihood of missing critical classifications. This metric is particularly important in scenarios where failing to detect positive instances has severe consequences, such as in medical diagnosis.

- **F1-Score**

The F1-Score provides a balanced measure of a model's performance by combining Precision and Recall into a single metric. It is calculated as the harmonic mean of the two:

$$\mathrm{F1-Score} = (2 \times Recall \times Precision) \Big/ (Recall + Precisi \quad (11)$$

This metric is especially useful when dealing with imbalanced datasets, as it considers both false positives and false negatives in the evaluation.

By analyzing these evaluation metrics together, a comprehensive understanding of the classification model's performance is obtained, ensuring robust and reliable predictions. K-Fold Cross validation

K-fold cross-validation (KCV) is a commonly used technique for model selection and error estimation in classification tasks. It involves dividing the dataset into k subsets, where, in each iteration, some subsets are utilized for training while the remaining ones are used for testing the model's performance. In this study, a 5-fold cross-validation strategy (k=5) was implemented to improve the proposed algorithms by systematically varying the training and testing data. Fig. 4 illustrates the accuracy achieved across different folds. The results show that K5 recorded the highest accuracy (0.90909), followed by K2 (0.90613), while K4 had the lowest accuracy (0.8847). The variation observed among the folds demonstrates the effectiveness of cross-validation in evaluating model generalization.

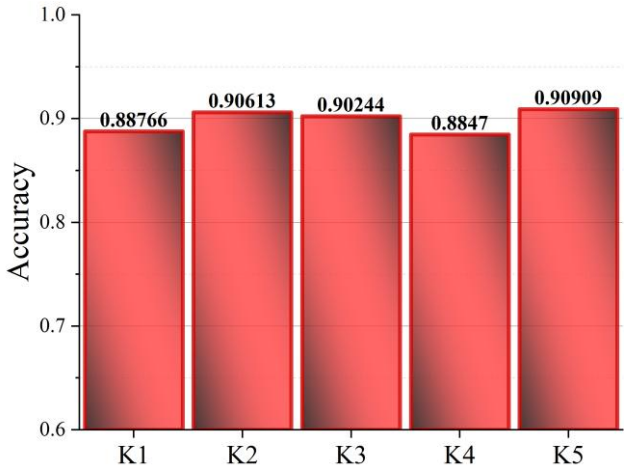


Figure 4: The results of 5-Fold cross validation

6 Results

6.1 Results of hyperparameters and convergence curves

Hyperparameters are predefined settings that control the learning process of a machine learning model. Unlike model parameters, which are learned from data, hyperparameters are set before training and have a significant impact on model performance. Various techniques exist for tuning hyperparameters to enhance model accuracy and efficiency. One widely used approach is random search, which was employed in this study to optimize the hyperparameters of the proposed hybrid models. Table 2 presents the optimized hyperparameter values for the SG + MO and NBC + MO models. For SG + MO, key hyperparameters include alpha (0.001), l1_ratio (0.014692), epsilon (0.022627), n_jobs (2), eta0 (0.999), and power_t (0.959754), which were fine-tuned to enhance the model's predictive performance. For NBC + MO, the most influential hyperparameters were alpha (62) and min_categories (4). These values were determined through random search, ensuring optimal model performance while balancing computational efficiency.

Table 2: Outcomes of hyperparameters for hybrid models.

Hyperparameter	Models	
	SG + MO	NBC + MO
alpha	0.001	--
l1_ratio	0.014692	--
epsilon	0.022627	--
n_jobs	2	--
eta0	0.999	--
power_t	0.959754	--
alpha	--	62
min_categories	--	4

Two convergence graphs for hybrid models over 200 iterations are shown in Fig. 5. The y-axis shows convergence (accuracy), and the x-axis shows the count of iterations. The accuracy in the left graph, labeled SG+MO, starts at about 0.654 and rises stepwise until the 200th iteration, when it reaches an approximate accuracy of 0.92831. Comparably, the accuracy of the right graph, designated NBC+MO, begins at about 0.601 and increases stepwise until it reaches about 0.89579 by the 200th iteration. Both graphs demonstrate the convergence behavior of the hybrid models by showing a general trend of increasing accuracy with the number of iterations.

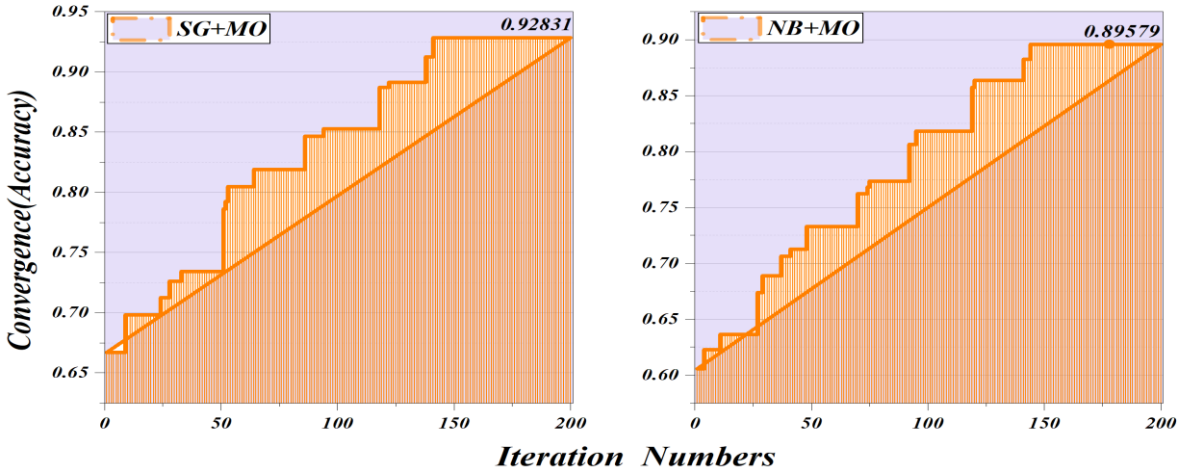


Figure 5: The convergence graphs of the hybrid models

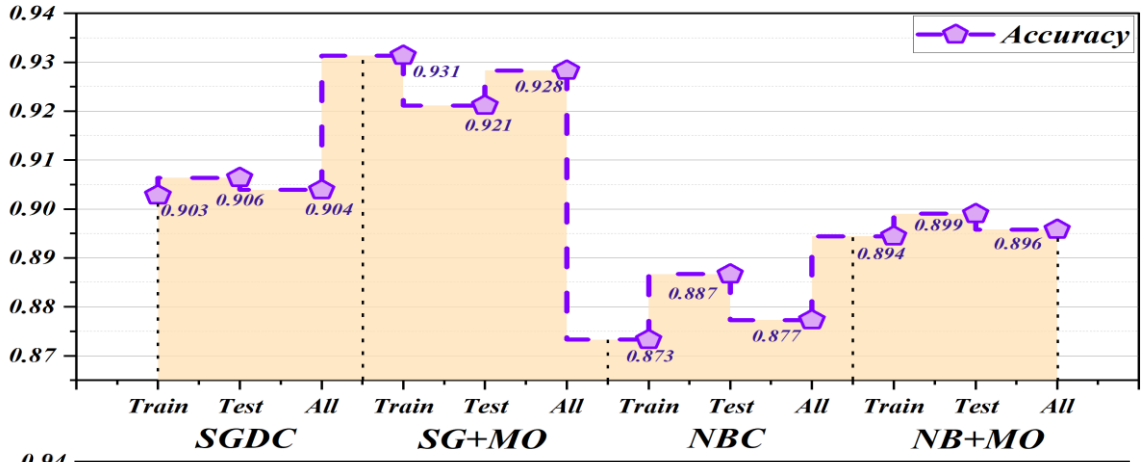
6.2 Results of metrics for predictive models

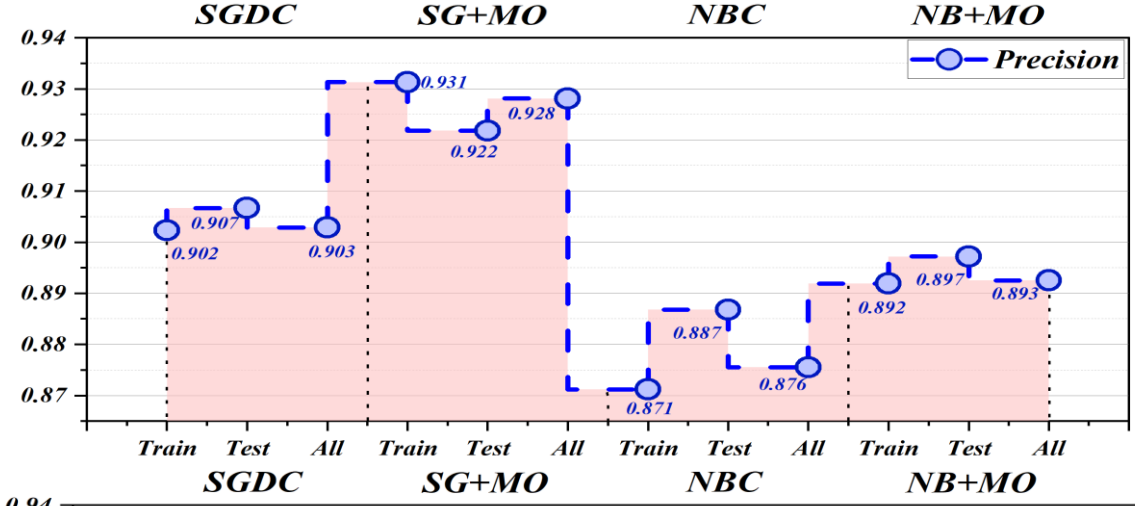
In the overall accuracy evaluation over all datasets, as shown in Table 3, the single SGDC model performed admirably, with an accuracy of 0.904, which was higher than the NBC model's 0.877 accuracy. This pattern continued with the hybrid models, where the SG+MO configuration outperformed the NBC+MO model with an accuracy of 0.928, surpassing its performance of 0.896. The SG+MO model demonstrated a significant performance difference, surpassing the NBC+MO model by 3.57%. Beyond accuracy, the SG+MO model also performed better on other critical metrics, such as precision, recall, and F1-score, all of which registered at an excellent 0.928. These findings, which are presented in detail in Table 3 and visually represented in Fig. 6, confirm the effectiveness of the SG+MO model when compared to other configurations and demonstrate its strong performance across a range of evaluation parameters.

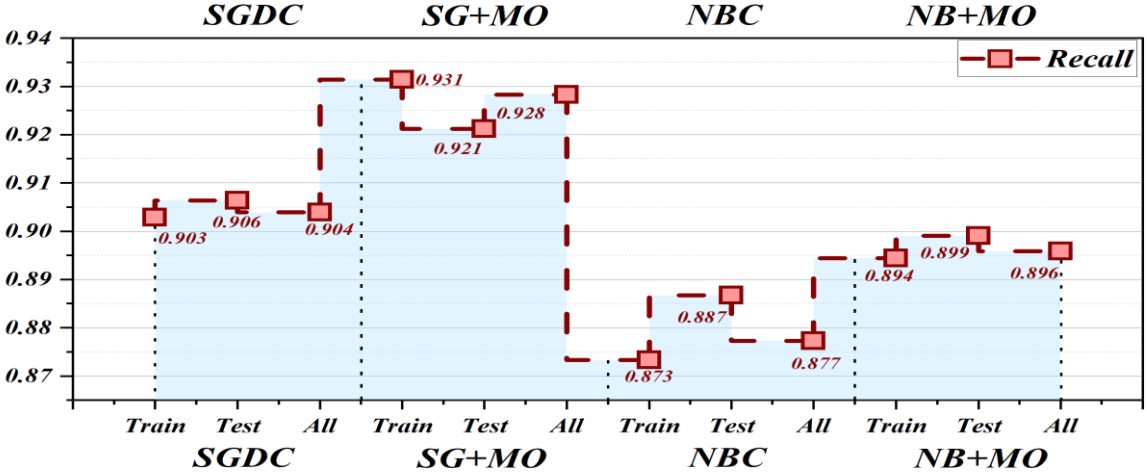
Table 3: Outcomes of the presented developed models.

Section	Model	Metric values					
		Accuracy	Precision	Recall	F1-score	AUC	MCC
Train	SGDC	0.903	0.902	0.903	0.902	0.881	0.8252
	SG+MO	0.931	0.931	0.931	0.931	0.915	0.8760
	NBC	0.873	0.871	0.873	0.866	0.799	0.7681
	NBC+MO	0.894	0.892	0.894	0.890	0.835	0.8085
Test	SGDC	0.906	0.907	0.906	0.906	0.881	0.8340
	SG+MO	0.921	0.922	0.921	0.921	0.915	0.8607

	NBC	0.887	0.887	0.887	0.881	0.799	0.7965
	NBC+MO	0.899	0.897	0.899	0.897	0.835	0.8200
All	SGDC	0.904	0.903	0.904	0.903	0.881	0.8277
	SG+MO	0.928	0.928	0.928	0.928	0.915	0.8714
	NBC	0.877	0.876	0.877	0.870	0.799	0.7769
	NBC+MO	0.896	0.893	0.896	0.892	0.835	0.8118







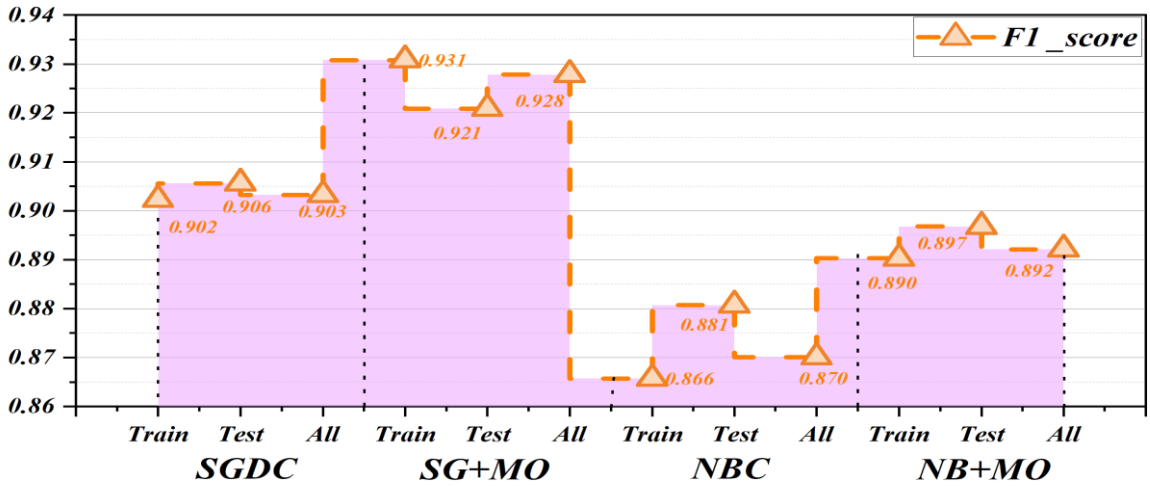
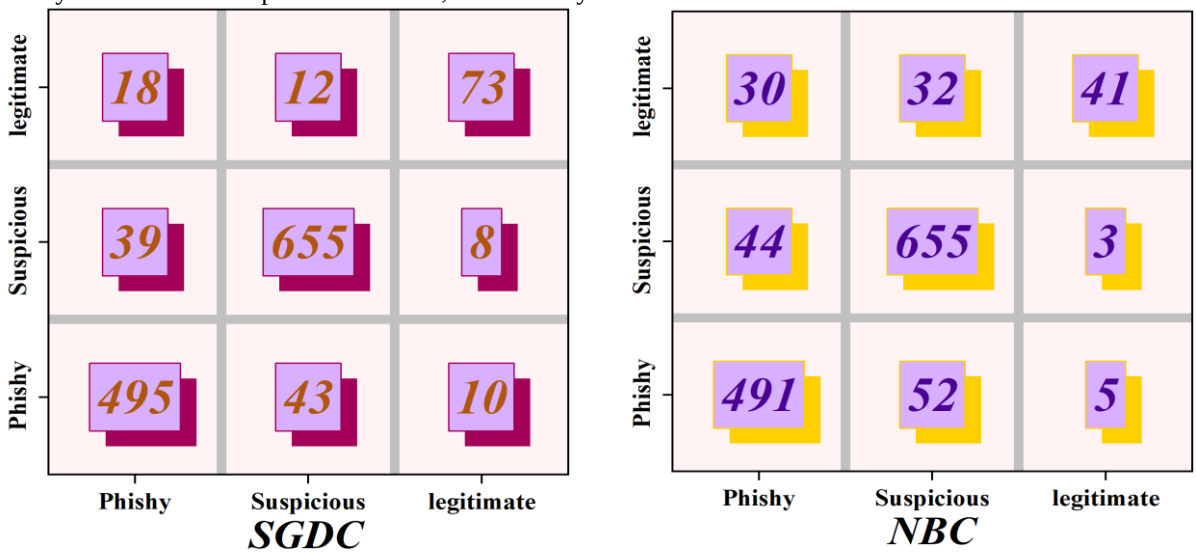


Figure 6: Plot showing the performance of the models across different phases

The confusion matrix, shown in Fig. 7, compares the actual and predicted classifications for three categories: Phishy, Suspicious, and Legitimate. It is a tool used to assess the effectiveness of classification models. The following is revealed by a thorough analysis of the misclassifications by rows for the top-performing SG+MO model: While correctly identifying 82 legitimate instances, the model misclassified 12 as phishy and nine as suspicious for legitimate instances. Although the model correctly identified 669 suspicious instances, it incorrectly classified 29 as phishy and four as legitimate. The model correctly identified 505 phishy instances but incorrectly classified 39 as suspicious and four as legitimate. The most notable misclassification in the matrix is that of phishy instances being classified as suspicious (39), while the model has the highest number of correct classifications for suspicious instances (669). Furthermore, when comparing phishy and suspicious instances to legitimate ones, the model shows comparatively fewer misclassifications.



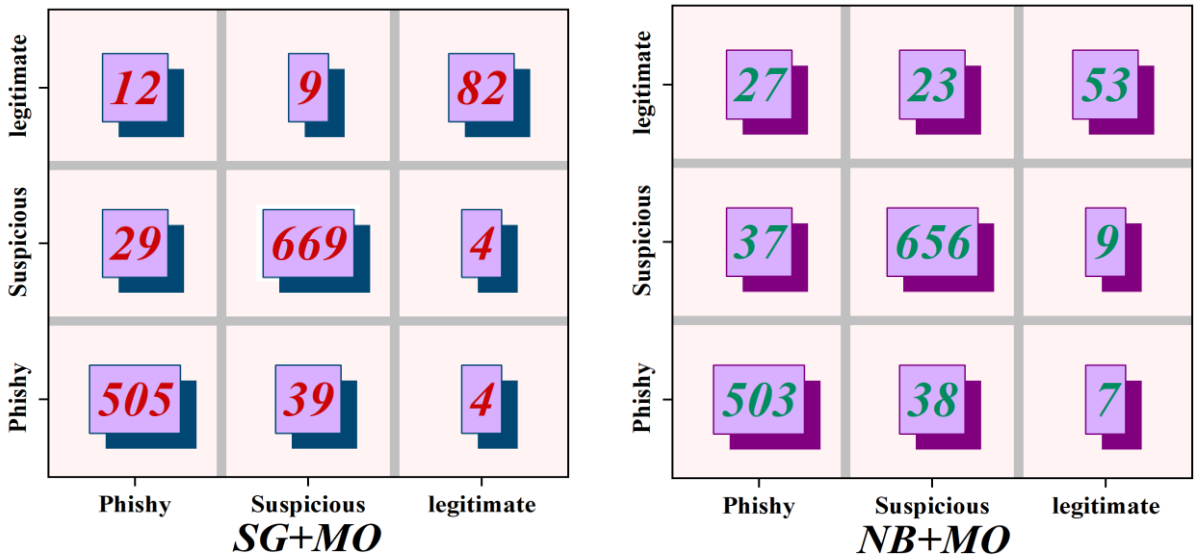


Figure 7: Confusion matrix showing the accuracy of the models under four specified conditions

The numerical classification results for each model are displayed in Table 4, where the SG+MO model performs better on many metrics. In particular, the SG+MO model shows the highest accurate results (0.923 for phishy, 0.943 for suspicious, and 0.850 for legitimate instances) when it comes to the F1-score. Furthermore, the model achieves 0.922 for phishy, 0.953 for suspicious, and 0.796 for legitimate classifications, according to a comparison of the Recall results. The SG+MO model continues to perform well when analyzing the Precision results, scoring 0.925 for phishy, 0.933 for suspicious, and 0.911 for legitimate instances. Together, these metrics show that the SG+MO model achieves high F1 scores by maintaining a balanced trade-off between precision and recall, in addition to being highly effective at correctly identifying positive cases in all three categories.

Table 4: Grading-based classification of the performance of the developed model

Metric values	Grade	Model			
		SGDC	SG+MO	NBC	NBC+MO
Precision	Phishy	0.897	0.925	0.869	0.887
	Suspicious	0.923	0.933	0.886	0.915
	legitimate	0.802	0.911	0.837	0.768
Recall	Phishy	0.903	0.922	0.896	0.918
	Suspicious	0.933	0.953	0.933	0.935
	legitimate	0.709	0.796	0.398	0.515
F1-score	Phishy	0.900	0.923	0.882	0.902
	Suspicious	0.928	0.943	0.909	0.925
	legitimate	0.753	0.850	0.540	0.616
AUC	Phishy	0.847	0.895	0.696	0.751
	Suspicious	0.916	0.935	0.902	0.919
	legitimate	nan	nan	nan	nan
MCC	Phishy	0.735	0.840	0.556	0.605
	Suspicious	0.832	0.871	0.800	0.834
	legitimate	0.849	0.880	0.807	0.842

A comparison of the measured and predicted values for the three categories of suspicious, phishing, and legitimate is shown in Fig. 8. The measured value for the Suspicious category is 702. The SG+MO model outperforms the SGDC and NBC models, which both come in at 655, and the NBC+MO model, which comes in at 656, in making the closest prediction at 669. The measured value in the Phishy category is 548. The closest prediction, once again from the SG+MO model, is 505, followed by NBC+MO at 503, SGDC at 495, and NBC at 491. With a measured value of 103, the Legitimate category displays a more notable discrepancy; in this case, the SG+MO model predicts 82, which is greater than SGDC at 73, NBC+MO at 53, and NBC at 41. This graph highlights the SG+MO model's superior performance over the other models by showing that it consistently produces predictions that are closer to the actual measured values across all categories.

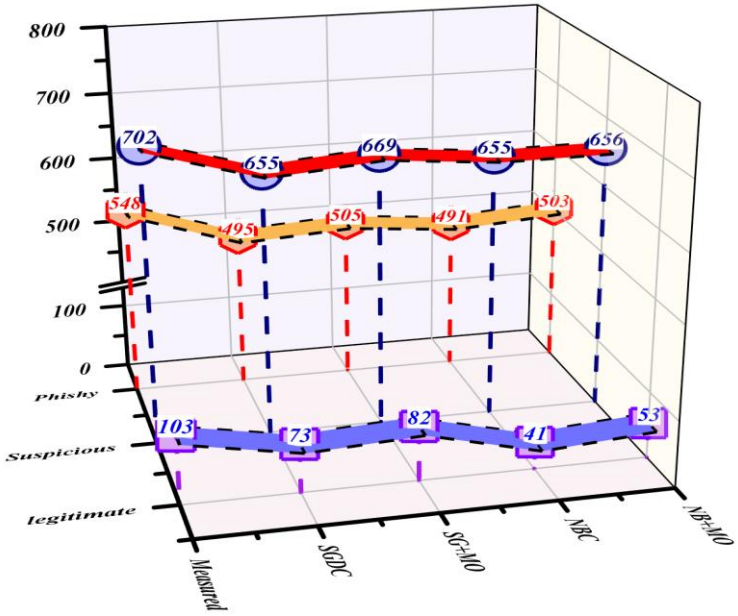


Figure 8: Visualization depicting the performance evaluation of the developed models

The ROC curve results are shown in Fig. 9, which is essential when assessing how well models perform in terms of achieving an actual positive rate of 1. The model that performs better is the one that reaches this rate the quickest. We can see that on the ROC curve plot for the category Phishy, the green line first reaches an actual positive rate of 1, demonstrating its high efficacy in this category for the model. Following that, the ROC curve for the suspicious category and the micro-average ROC curve reaches this rate, which goes on to say that the performance generally is excellent across a variety of classes and particularly good in detecting suspicious instances. The macro-average ROC curve represents the third ROC curve that reaches an actual positive rate of one. It provides an objective evaluation of the effectiveness of the model across all categories. Finally, the ROC curve for the legitimate category approaches this rate, showing its ability to classify the legitimate ones correctly. These ROC curves collectively present the relative efficaciousness of the models and their advantages within a broad spectrum of classification tasks.

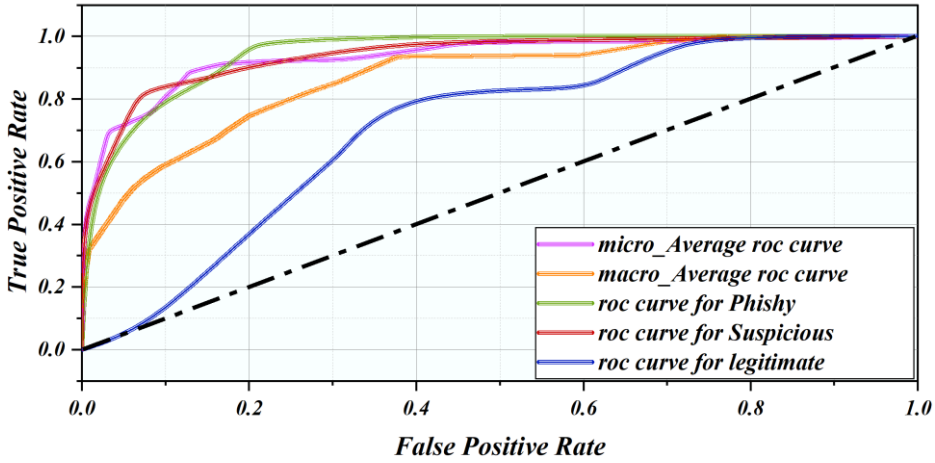


Figure 9: ROC curves illustrating the performance of the most effective hybrid models

Run time comparison

The computational efficiency of the models was also evaluated by recording their execution times. The runtime results indicate that the standalone Naïve Bayes Classifier (NBC) and Stochastic Gradient Descent (SGD) models exhibited significantly lower execution times, with 0.57 seconds and 0.87 seconds, respectively. However, when integrated with the MOA framework, the computational cost increased considerably, with NBC_MOA requiring 598.5 seconds and SGD_MOA taking 913.5 seconds. These findings highlight the trade-off between optimization and computational efficiency, suggesting that while the hybrid models may improve prediction accuracy, their feasibility for real-time applications depends on the available computational resources.

6.3 Sensitivity analysis based on FAST and SHAP

As its name suggests, fast sensitivity analysis provides a quick way of determining how changes in the input parameters will affect a model's output. Highlighting such important variables is crucial to feature selection, model performance enhancement, and, importantly, model complexity reduction. Sensitivity analysis enhances interpretability by reducing the model to just the most important features while informing data collection activities. Moreover, it offers a straightforward approach toward an iterative model enhancement process that ensures good adaptation and optimization without high computational loads.

The pie chart segments in Fig. 10 give a comprehensive look into the different features and each of their contributions to the model in identifying phishing attempts. The highest ranking, with 35.8% and in teal, is the URL Anchor segment, which is highly sensitive and influential. This feature is critical because it helps identify if the anchor text matches the actual URL, which is a general strategy that phishers use. The second most salient feature is the requested URL, in light teal color, making up 18.7%. This shows how important it is to examine the requested URL for deviations from expected patterns, as these can be a strong indicator of phishing attempts. The light brown (at 18%) SFH (Server Form Handler) feature is also essential because it deals with form submissions on a server, which is a crucial component in identifying phishing sites. Given that phishing sites frequently use longer, more complicated URLs to mask their true nature, URL length (dark teal, 8.3%) is deemed to be moderately significant. Another moderate factor is having an IP address (light brown, 7.4%), as legitimate websites hardly ever use raw IP addresses in URLs. Prefix/Suffix (dark brown, 4.8%) and subdomain (light orange, 3.4%) offer extra indications, but their significance is less than that of the leading features. Domain Age (light teal, 0.8%) has the least bearing because while newer domains can be more suspicious, it's not a very strong indicator. Web Traffic (dark orange, 3.3%) indicates that lower-traffic sites are more suspicious.

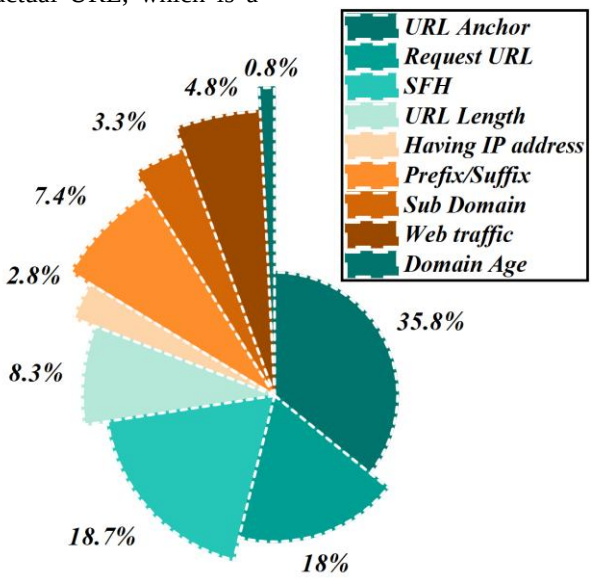


Figure 10: Sensitivity analysis conducted using FAST for the best model

The results of the SHAP sensitivity analysis summarized in Table 5. The SHAP sensitivity analysis ranks input features based on their average influence across three phishing categories: *Phishy*, *Suspicious*, and *Legitimate*. The URL Anchor feature has the highest overall impact (AVG = 1.1625), indicating its strong influence on classification decisions, particularly in distinguishing *Phishy* and *Legitimate* websites. SFH (Server Form Handler) follows as the second most significant feature (AVG = 0.8308), suggesting its relevance in identifying phishing threats. Request URL ranks third (AVG = 0.6089), highlighting its role in determining website legitimacy.

URL Length shows moderate importance (AVG = 0.5065), being more impactful in *Phishy* and *Suspicious* classifications than *Legitimate*. Having an IP Address (AVG = 0.3826) and Sub Domain (AVG = 0.3733) contribute similarly to predictions, but their influence is lower. The least impactful feature is Prefix/Suffix (AVG = 0.2635), indicating minimal significance in phishing detection.

Overall, the analysis confirms that *URL-based attributes*, particularly *URL Anchor*, *SFH*, and *Request URL*, play a crucial role in phishing classification, whereas domain-related attributes like *Prefix/Suffix* contribute less.

Table 5: Result of SHAP sensitivity analysis.

Rank	Inputs	Phishy	Suspicious	legitimate	AVG
1	URL Anchor	1.718901889	0.361819914	1.406823515	1.162515
2	SFH	1.091688147	0.292571162	1.108129482	0.830796
3	Request URL	0.850812044	0.156678455	0.819171597	0.608887
4	URL Length	0.643498015	0.613921318	0.262064841	0.506495
5	Having IP address	0.375322899	0.424208786	0.348136872	0.382556
6	Sub Domain	0.309529929	0.46483848	0.345516127	0.373295
7	Prefix/Suffix	0.17633227	0.266295892	0.347764885	0.263464
8	Web traffic	0.167606507	0.065772981	0.18652572	0.139968
9	Domain Age	0.098777995	0.058722238	0.051814727	0.069772

6.4 Wilcoxon test

The Wilcoxon test is a non-parametric statistical test used to compare paired or independent samples to determine whether there is a significant difference between them. Unlike parametric tests, it does not assume a normal distribution of data, making it particularly useful for small sample sizes or non-normally distributed datasets. The Wilcoxon signed-rank test is used for paired samples, whereas the Wilcoxon rank-sum test (also known as the Mann-Whitney U test) is applied to independent samples.

In this study, the Wilcoxon test was conducted to compare the performance differences between models, specifically examining the statistical significance of their differences through p-values and test statistics.

Table 6 presents the p-values and test statistics for different models under comparison. The p-value indicates whether there is a statistically significant difference between models, with a typical significance threshold set at 0.05 (i.e., if $p < 0.05$, the difference is considered significant).

- SGD_MOA ($p = 0.265464$, statistic = 2081.5): Since the p-value is greater than 0.05, there is no significant difference in performance when using the SGD_MOA model.
- SGD ($p = 0.753526$, statistic = 4127.5): The high p-value suggests no statistically significant difference, indicating that the SGD model's performance is not substantially different from the others.
- NBC_MOA ($p = 0.917888$, statistic = 4957): This is the highest p-value in the table, confirming that NBC_MOA does not show a significant difference compared to other models.
- NBC ($p = 0.36308$, statistic = 6385.5): Again, the p-value is well above 0.05, indicating no statistically significant performance difference.

Table 6: Result of Wilcoxon test.

Difference of models	Parameter	
	p_value	statistic
SGD_MOA	0.265464	2081.5
SGD	0.753526	4127.5
NBC_MOA	0.917888	4957
NBC	0.36308	6385.5

7 Discussion

7.1 Limitations of the study

Despite the promising results of the hybrid phishing detection models, this study has certain limitations. First, the dataset used, while diverse, may not fully capture the constantly evolving nature of phishing websites. Cybercriminals frequently adapt their techniques, introducing new evasion strategies that may reduce model effectiveness over time. Additionally, the study relied on a dataset with a limited number of features (nine input parameters), which may not fully encompass all relevant attributes influencing phishing website classification. Expanding the feature set could enhance model robustness and generalizability.

Another limitation lies in the reliance on MOA for hyperparameter optimization. While MOA demonstrated superior performance compared to traditional techniques, it may not always guarantee the absolute best hyperparameters, as optimization outcomes depend on initial conditions and algorithm-specific parameters. Further research could explore hybrid or ensemble optimization strategies to enhance performance.

Furthermore, real-time deployment of the models was not tested in a live cybersecurity environment. The study primarily focused on offline classification accuracy, leaving questions about computational efficiency and adaptability in real-world phishing detection scenarios. Future studies should evaluate the models' performance in real-time detection systems to assess latency, adaptability, and scalability.

7.2 Benefits of each hybrid model

Each hybrid model presented distinct advantages that contribute to phishing detection performance. The SGD+MOA model emerged as the most effective, achieving the highest accuracy of 92.1%. This model benefits from the ability of SGD to process large-scale datasets efficiently while adapting quickly to new phishing patterns. The integration of MOA further improved convergence and predictive capability by fine-tuning hyperparameters, making it a strong choice for real-time phishing detection applications.

On the other hand, the NBC+MOA model, while exhibiting slightly lower accuracy (87.7%), offered advantages in terms of interpretability and computational efficiency. The probabilistic nature of NBC allows for clearer decision-making processes, which can be useful in cybersecurity applications where transparency is essential. Additionally, NBC requires fewer computational resources compared to complex deep learning models, making it a viable option for systems with hardware constraints.

By combining machine learning models with bio-inspired optimization, both hybrid models demonstrated significant improvements over their standalone counterparts. The results indicate that optimization algorithms such as MOA can substantially enhance model accuracy, reduce misclassification rates, and provide more reliable phishing detection systems.

7.3 Practical implications of the study

The findings of this study have several practical implications for cybersecurity, particularly in enhancing phishing detection mechanisms. The demonstrated effectiveness of hybrid models suggests that organizations can integrate such approaches into their cybersecurity frameworks to improve email filtering systems, web security tools, and fraud detection software. By leveraging SGD+MOA, companies can deploy real-time phishing detection models that efficiently adapt to evolving cyber threats, reducing the risk of data breaches and identity theft.

For organizations with resource constraints, the NBC+MOA model offers a practical alternative, ensuring phishing detection with lower computational overhead while maintaining strong classification performance. This is particularly beneficial for small businesses, financial institutions, and government agencies that require cost-effective cybersecurity solutions.

Moreover, the study highlights the importance of bio-inspired optimization in machine learning applications, reinforcing the idea that intelligent optimization techniques can enhance traditional classification models. As phishing tactics become more sophisticated, continuously improving detection models through advanced optimization strategies can provide a proactive defense against cyber threats.

Future implementations could focus on integrating these models with browser extensions, email security tools, and AI-driven security systems to enhance real-time phishing detection. Additionally, regulatory bodies and cybersecurity professionals can utilize the findings to establish more effective security policies, training programs, and automated detection mechanisms to combat phishing attacks more effectively.

8 Conclusion

Phishing is a severe risk to both individuals and organizations because it uses deceptive tactics to trick victims into disclosing private information like passwords, bank account information, or personal information. Cybercriminals frequently use phishing tactics to trick unsuspecting victims into unintentionally disclosing their personal information using phishy emails, websites, or messages. Identity theft, money loss, and security compromise are possible outcomes of this.

ML is essential for securing personal information from phishing attacks. Using advanced algorithms and predictive models, ML may analyze data patterns to identify or flag potential phishing websites or suspicious activities. These models learn by identifying subtle cues or anomalies from enormous volumes of data indicative of a phishing attempt and provide proactive defense against cyber threats.

It was derived from the study that the MOA, combined with ML models, such as NBC and SGD, were able to classify phishing websites more accurately. The hybrid models derived showed a marked increase in the differentiation of authentic fraudulent websites by optimizing their predictive capability and reducing error rates.

The results showed that MO increased the accuracies of the SGD and NBC models by 2.65% and 2.17%, respectively. With the lowest error rates, the resultant hybrid model, namely SG+MO, turned out to be the most effective. In particular, out of 548 instances, it correctly predicted 92.15% of phishy websites while labeling 39 as suspicious and four as legitimate. The model misclassified 29 as phishy and four as legitimate out of 702 instances, hence giving an actual prediction rate of 95.3% for suspicious websites. The model achieved an actual prediction rate of 79.61% when it came to legitimate websites, misclassifying 12 as phishy and nine as suspicious out of 103 instances.

Future research could examine bigger and more varied datasets and more optimization strategies to improve model performance and further improve cybersecurity measures. Enhancing the precision and effectiveness of phishing detection systems can also involve addressing feature engineering and selection techniques.

Competing interests

The scholars claim no competing interests.

Authorship contribution statement

Xiao CHEN: Writing-Original draft preparation Conceptualization, Supervision, Project administration.

Data availability

The scholars will make the raw data supporting this article's conclusions available without undue reservation.

Declarations

Not applicable.

Author statement

The manuscript has been read and approved by all the authors, the requirements for authorship, as stated earlier in this document, have been met, and each author believes that the manuscript displays honest work.

Ethical approval

All scholars have been personally and actively involved in substantial work leading to the paper and will take public responsibility for its content.

References

[1] Lastdrager EEH. Achieving a consensual definition of phishing based on a systematic review of the literature. Crime Sci 2014;3:1–10.

[2] Mohammad RM, Thabtah F, McCluskey L. Tutorial and critical analysis of phishing websites methods. Comput Sci Rev 2015;17:1–24.

[3] Garera S, Provos N, Chew M, Rubin AD. A framework for detection and measurement of phishing attacks. Proceedings of the 2007 ACM workshop on Recurring malcode, 2007, p. 1–8.

[4] Khonji M, Iraqi Y, Jones A. Phishing detection: a literature survey. IEEE Communications Surveys & Tutorials 2013;15:2091–121.

[5] Varshney G, Sardana A, Joshi RC. Secret information display based authentication technique towards preventing phishing attacks. Proceedings of the International Conference on Advances in Computing, Communications and Informatics, 2012, p. 602–8.

[6] Hong J. The state of phishing attacks. Commun ACM 2012;55:74–81.

[7] Pandove K, Jindal A, Kumar R. Email spoofing. Int J Comput Appl 2010;5:27–30.

[8] Jagatic TN, Johnson NA, Jakobsson M, Menczer F. Social phishing. Commun ACM 2007;50:94–100.

[9] Almomani A, Gupta BB, Atawneh S, Meulenberg A, Almomani E. A survey of phishing email filtering techniques. IEEE Communications Surveys & Tutorials 2013;15:2070–90.

[10] Drake CE, Oliver JJ, Koontz EJ. Anatomy of a Phishing Email. CEAS, 2004.

[11] Varshney G, Misra M, Atrey PK. A survey and classification of web phishing detection schemes. Security and Communication Networks 2016;9:6266–84.

[12] Varshney G, Joshi RC, Sardana A. Personal secret information based authentication towards preventing phishing attacks. Advances in Computing and Information Technology: Proceedings of the Second International Conference on Advances in Computing and Information Technology (ACITY) July 13-15, 2012, Chennai, India-Volume 1, Springer; 2012, p. 31–42.

[13] Gupta S, Kumar P. A desktop notification based scheme for preventing online frauds attempts to cloud users S| pp| S. 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI), IEEE; 2013, p. 255–60.

[14] Workman M. Wisecrackers: A theory-grounded investigation of phishing and pretext social engineering threats to information security. Journal of the American Society for Information Science and Technology 2008;59:662–74.

[15] Bergholz A, De Beer J, Glahn S, Moens M-F, Paaß G, Strobel S. New filtering approaches for phishing email. J Comput Secur 2010;18:7–35.

[16] Islam R, Abawajy J. A multi-tier phishing detection and filtering approach. Journal of Network and Computer Applications 2013;36:324–35.

[17] Das A, Baki S, El Aassal A, Verma R, Dunbar A. SoK: a comprehensive reexamination of phishing research from the security perspective. IEEE Communications Surveys & Tutorials 2019;22:671–708.

[18] Aburrous M, Hossain MA, Dahal K, Thabtah F. Predicting phishing websites using classification mining techniques with experimental case studies. 2010 seventh international conference on information technology: New generations, IEEE; 2010, p. 176–81.

[19] Ramesh G, Krishnamurthi I, Kumar KSS. An efficacious method for detecting phishing webpages through target domain identification. Decis Support Syst 2014;61:12–22.

[20] Abdelhamid N. Multi-label rules for phishing classification. Applied Computing and Informatics 2015;11:29–46.

[21] Phishing Data Phishtank n.d. https://www.kaggle.com/datasets/latheeshmangeri/phishing-data-phishtank.

[22] Bottou L. Large-scale machine learning with stochastic gradient descent. Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers, Springer; 2010, p. 177–86.

[23] Ahn S, Korattikara A, Liu N, Rajan S, Welling M. Large-scale distributed Bayesian matrix factorization using stochastic gradient MCMC. Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, 2015, p. 9–18.

[24] Manning CD. An introduction to information retrieval. Cambridge university press; 2009.
[25] Rosen G, Garbarine E, Caseiro D, Polikar R, Sokhansanj B. Metagenome Fragment Classification Using ?-Mer Frequency Profiles. Adv Bioinformatics 2008;2008.
[26] Han J, Kamber M. Data mining concepts and techniques San Francisco Moraga Kaufman 2001.
[27] Pekuwali AA, Kusuma WA, Buono A. Optimization of Spaced K-mer Frequency Feature Extraction using Genetic Algorithms for Metagenome Fragment Classification. Journal of ICT Research & Applications 2018;12.
[28] Kotsiantis S, Patriarcheas K, Xenos M. A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education. Knowl Based Syst 2010;23:529–35.
[29] Zervoudakis K, Tsafarakis S. A mayfly optimization algorithm. Comput Ind Eng 2020;145:106559.
[30] Bhattacharyya T, Chatterjee B, Singh PK, Yoon JH, Geem ZW, Sarkar R. Mayfly in harmony: A new hybrid meta-heuristic feature selection algorithm. IEEE Access 2020;8:195929–45.
[31] Kadry S, Rajinikanth V, Koo J, Kang B-G. Image multi-level-thresholding with Mayfly optimization. International Journal of Electrical & Computer Engineering (2088-8708) 2021;11.
[32] Raja NSM, Rajinikanth V, Latha K. Otsu based optimal multilevel image thresholding using firefly algorithm. Modelling and Simulation in Engineering 2014;2014:37.

A Review of Deep Learning-Based Feature Extraction Techniques for Iris Image Analysis

[1]Poovayar Priya M*, [2]Ezhilarasan M
[1]Dept. of Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai-600062, India
[2]Dept. of Information Technology Puducherry Technological University, Puducherry-605014, India
E-mail: poovayarpriyam@veltech.edu.in, mrezhil@ptuniv.edu.in

Keywords: feature extraction, machine learning, deep learning, iris analysis, CNN

Received: June 3, 2024

Feature extraction is a critical process in image processing, directly influencing the accuracy of classification tasks. Traditional feature extraction methods often fall short by capturing only a limited set of features, which can result in suboptimal classification, particularly in complex biometric systems. In the domain of iris recognition, the extraction of minute and intricate features is paramount for achieving high accuracy and reliability. This paper presents a comprehensive review of advanced deep learning-based feature extraction techniques tailored for iris images. These techniques are designed to capture and process fine details such as lacunae, Wolfflin nodules, contraction furrows, and pigment spots—features that are essential for both recognition and diagnostic purposes. We compare state-of-the-art methods including convolutional neural networks (CNNs), U-Net, Link-Net and other custom architectures, highlighting their advantages in terms of segmentation precision, feature extraction efficiency, and computational requirements. Our review also addresses the challenges posed by variations in iris texture, illumination, and occlusions, which can affect the accuracy of feature extraction. Experimental evaluations of these methods show that deep learning algorithms significantly improve the detection of subtle features, offering promising results for both recognition accuracy and disease diagnosis applications. Finally, we outline future directions in the field, including the potential for integrating multi-resolution and attention-based models to further enhance the extraction of fine details from iris images. We also discuss the opportunities for real-time deployment of these techniques in mobile and resource-constrained environments, opening new avenues for iris recognition and diagnostic systems.

Povzetek: Pregledno so opisane so tehnike za ekstrakcijo lastnosti irisnih slik z uporabo globokega učenja, pri čemer so kombinirana konvolucijska nevronska omrežja (CNN) za obdelavo natančnih in subtilnih značilnosti, kot so lacune in pigmentne pike.

1 Introduction

Feature extraction is a fundamental step in data processing [1], designed to transform raw data into a manageable set of relevant features that capture the most significant information for a particular task. This process simplifies data by reducing its dimensionality, making it more efficient for analyzing machine learning models. In traditional methods, feature extraction relies on predefined techniques like edge detection [2], texture analysis, or frequency-domain transformations (e.g., Fourier or wavelet transforms), commonly used in areas such as image processing, audio analysis, and signal processing. These techniques highlight important characteristics of the data, such as edges or textures in images, which help in the classification or recognition tasks. Feature extraction is essential for improving computational efficiency, particularly when dealing with large datasets, as it reduces redundant information and helps models generalize better by focusing on the most critical aspects of the data [3].

Deep learning is a subset of machine learning that automates feature extraction through multi-layered neural networks. It is particularly effective for handling complex, high-dimensional datasets like images, videos, and text, where manually extracting features may be inefficient or impractical. The primary model in deep learning for feature extraction is the Convolutional Neural Network (CNN), which excels in image and video recognition tasks by learning hierarchical features from pixel-level data. CNNs automatically identify low-level features like edges and gradually learn more abstract features such as shapes and objects as data moves through deeper layers of the network. Recurrent Neural Networks (RNNs) are another type of deep learning model used for sequential data, like time series or language processing, where context over time is crucial. The advantage of deep learning models is that

they can discover patterns without manual intervention, leading to state-of-the-art performance in fields such as computer vision, natural language processing, and medical image analysis [4].

In contrast to deep learning, traditional machine learning models, such as Support Vector Machines (SVM), decision trees, and k-nearest neighbors (KNN), rely on manually engineered features. These models work well when datasets are smaller and structured, with clear feature representations. While not as powerful as deep learning for unstructured data like images or text, traditional models are highly effective for simpler tasks and require less computational power. For example, decision trees and random forests can model decisions in structured environments, while SVMs are often used for classification tasks with well-defined input features [5]. These models are also more interpretable than deep learning networks, making them preferable when understanding model decisions is important. Feature extraction techniques in machine learning, such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA), help reduce dimensionality and ensure that models focus on the most informative aspects of the data.

Ultimately, deep learning emerges as the superior approach for feature extraction and analysis, especially in complex and high-dimensional datasets. Its ability to learn hierarchical representations directly from raw data without the need for manual feature engineering sets it apart from traditional machine learning techniques. Deep learning models, particularly CNNs, excel in extracting intricate patterns and features, making them highly effective for tasks like medical image analysis, where subtle details in data, such as those in iris images, can be critical for accurate diagnosis [6] and prediction. As computational resources continue to improve and large datasets become more accessible, the advantages of deep learning will likely expand further, solidifying its position as the best method for extracting meaningful insights from complex data. The below Fig. 1 represents the different features to be extracted from the enhanced iris image:

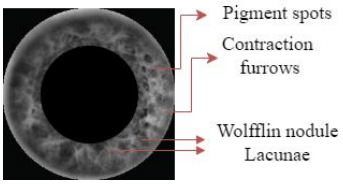


Figure 1: Iris features

2 Related works

2.1 Gradient-based feature extraction

After examining the impact of demosaicing on gradient extraction, Wei Zhou et al. [7] suggested a gradient-based feature extraction pipeline based on raw Bayer pattern images. In addition to consuming computing time, the traditional demosaicing method uses three times as much storage space in order to achieve almost identical results. However, by utilizing the color difference constancy assumption, the suggested method directly extracts gradients from the Bayer pattern images while lowering computing complexity. Five steps are involved in the SIFT-based algorithm's computation:

1) Constructing areas of scale. The difference-of-Gaussian (DoG) pyramid approximates the scale space.

2) The highest level of judgment. The local maxima and minima can be identified by comparing each pixel with its neighbors in a 3*3 neighborhood between the current scale, scale above, and scale below.

3) Locating the primary idea. to further hone the key point contenders discovered in the previous action. We discard unstable key points, such as places along edges with poor localization and low contrast.

4) Figuring out orientation. to offer one or more orientations for every significant point. A histogram is created for a region centered on the key point, with a radius 1.5 times the scale of the key point. The direction with the highest bar in the histogram is referred to as the dominating direction, whereas auxiliary directions are those with heights more than 80% of the top bar.

2.2 Feature fusion-net using deep spatial context encoder

A unique end-to-end Fusion-Net classification model for HR SAR images was proposed by Wenkai Liang [8]. Its goal is to fuse statistical characteristics into deep spatial feature objects in the end-to-end representation learning process. In a supervised feature learning framework, the physical characteristics of terrain objects can be revealed by the integration of distinct statistical distributions of SAR pictures into CNNs. In order to tackle this problem, a new end-to-end supervised classification technique is put forth for HR SAR images that takes statistical information and spatial context into account. Utilizing Fusion-Net has the benefit of enhancing the classification model's overall accuracy significantly through the complimentary knowledge of statistical and geographical variables. The definition of the Fusion-Net scheme is:

$$F^{Fuse} = sigm\left(W_2\left(sigm\left(GConv(W_1', F^{Spatial})\right)\right)\right) \| sigm\left(GConv(W_1'', F^{Statistical})\right)\right)$$

Compared to existing related systems, the suggested Fusion-Net produces substantially greater accuracies and a more pleasing visual look. This strategy yielded an overall accuracy of 89.00%. In the future, self-supervised learning techniques like contrastive learning or data augmentation like Mixup will be taken into consideration to improve DSCEN's feature representation capabilities.

2.3 Unsupervised deep image stitching

A two-stage unsupervised deep image stitching framework—unsupervised coarse image alignment and unsupervised image reconstruction—was presented by Lang Nie et al. [9]. Conventional feature-based picture stitching techniques frequently struggle to stitch images with low resolution or few features since they mostly rely on feature detection quality.

The suggested approach receives ratings of 27.83 and 0.902 for PSNR and SSIM, respectively. In terms of visual quality, people still prefer the outcomes of the suggested unsupervised approach over supervised deep image stitching alternatives. Furthermore, the reconstruction capacity is limited, indicating that the suggested approach could not work in situations with a very high parallax. Extending the linear deep homography to a non-linear homography model can help handle big parallax.

2.4 SpaSSA: Superpixelwise adaptive SSA

To utilize the local spatial information of HSI (Hyperspectral pictures), Genyun Sun et al. [10] presented SpaSSA (Superpixelwise adaptive Singular Spectral Analysis). While ignoring the local spatial context of the pixels, 2D-SSA primarily concentrates on global spatial information.

Improved classification accuracy can be achieved by using the suggested SpaSSA technique, which can successfully lower the intra-class variation within superpixels and improve the discriminating between various superpixels. This approach yields an overall accuracy of 98.34%. Principal component analysis and SpaSSA together can increase accuracy even more.

2.5 Symmetric all CNN (SACNN)

A symmetric all CNN (SACNN), an end-to-end network made up of encoder and decoder subnetworks, was proposed by Mingyang Zhang et al. [11]. Large numbers of labeled samples are always needed for deep learning network training, although HSI image data are rarely available. A unique unsupervised deep-learning based FE approach is developed to address this difficulty. The following is the definition of the cost function that the Adam algorithm optimizes:

$$L(X,R) = \frac{1}{P_1Q_1R_1}\sum_{i=1}^{P_1}\sum_{j=1}^{Q_1}\sum_{k=1}^{R_1}\left(x_{ijk} - r_{ijk}\right)^2 + \frac{\lambda}{2}\sum_{l=1}^{N}\sum_{i=1}^{P_2}\sum_{j=1}^{Q_2}\sum_{k=1}^{R_2}\left(w_{ijk}^{l}\right)^2$$

This approach yields an overall accuracy of 97.07%. The outcomes of the experiments indicated that learning effective spatial spectral features in HSI images was more suited for 3-D convolutional operation. Additional research will be conducted on adaptively optimizing hyperparameters based on data.

2.6 Spatial revising variational autoencoder-based feature extraction

Wenbo Yu et al. [12] introduced a novel unsupervised hyperspectral feature extraction architecture based on the spatial revising variational autoencoder (UHfeSRVAE). There are still several challenges in fully extracting rich spectrum information, such as the coupling of spectral and spatial information. To solve this issue, this approach makes use of designed networks to extract spatial characteristics from many perspectives in order to revise the spectral features that are subsequently obtained.

Compared to other methods, the suggested method yields a more accurate categorization map. The excellent clusterability and discernibility of UHfeSRVAE is further demonstrated by its performance in three-view drawings. This strategy yielded an overall accuracy of 91.39%. Additional research will concentrate on generative models for feature extraction.

2.7 Feature extraction via 3-D block characteristics sharing (3-D-BCS)

For high-strength images (HSI), Bing Tu et al. [13] suggested a unique feature extraction method using 3-D block characteristics sharing (3-D-BCS). The low sensitivity of sensors to spatial information and different environmental factors, such as weather, cause the neighboring spectral-spatial information of a pixel to become mixed in with other ground coverings. The following Fig. 2 is the suggested 3-D-BCS algorithm:

Algorithm 1 3-D-BCS

Inputs: Original HSI $\boldsymbol{I} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\} \in \mathbb{R}^{N \times M}$, Training set, and parameters setting: Q, S_N, K, and δ.

Outputs: Classification results obtained by decision fusion-based SVM classifier.

1: Obtain low-dimension HSI $\boldsymbol{D} = \{\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_N\} \in \mathbb{R}^{N \times Q}$ by performing the MDS algorithm.
2: Generate nonoverlapping 3-D superpixel blocks by mapping $\boldsymbol{I}_s$ into the reduced image $\boldsymbol{D}$.
3: Block characteristic sharing-based feature extraction for each 3-D superpixel block.
4: **For** *each* $i = 1 : S_N$ **do**.
5: *Superpixel-based Gabor Filter*: Ensuring the texture consistency of each superpixel according to (14).
6: *Superpixel-based K-means Filter*: To address the spatial weak assumption within a superpixel, local density peak (15) and K nearest neighbor (16) are introduced.
7: *Superpixel-based Gaussian Weighting Filter*: Overcoming the spectral weak assumption among pixels within a superpixel according to (19).
8: **End For**
9: Superpixel Gabor feature-based SVM results.
10: Superpixel K-means feature-based SVM results.
11: Superpixel weighting feature-based SVM results.
12: Majority voting-based decision fusion for the classification results of the SVM classifier.

Figure 2: 3-D-BCS algorithm

This strategy yielded an overall accuracy of 96.69%. Tests conducted on many genuine hyperspectral data sets with a small number of training samples demonstrate that the suggested 3-D-BCS approach works better than alternative classification methods. The inability of single-scale over segmentation and single feature representation to fully mine the structural information of materials is one of the suggested method's limitations. Future research will focus on collaborative learning of multiscale strategies and numerous features for feature extraction.

2.8 Cloud-guided feature extraction approach

A cloud-guided feature extraction method was presented by Shangguang Wang et al. [14] for mobile image retrieval. Two main issues plague the current works: low retrieval accuracy and high network bandwidth costs.

Since the edge servers only upload the collected discriminative features to cloud servers, the suggested strategy can lower network traffic. This approach produced a retrieval accuracy of 88.33%. To dynamically update the projection matrix in the future, the model can be trained to extract the number of eigenvalues, accuracy, and reaction time.

2.9 Unsupervised structural feature-guided convolutional neural network (SFG)

Lin Ge et al. suggested an unsupervised structural feature-guided convolutional neural network (SFG) [15]. Gathering ground-truth data with known correspondences requires significant work and time when utilizing supervised algorithms. For this reason, unsupervised approaches are used. The two structural feature components of the proposed technique are sparse and dense structural components. The suggested method's pipeline looks like in the Fig.3:

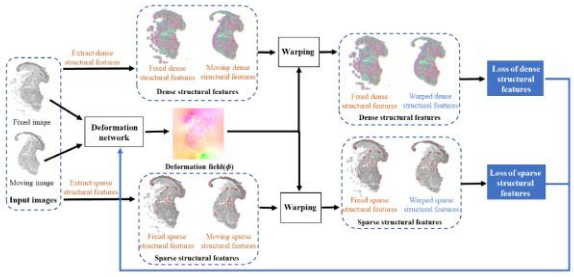


Figure 3: Unsupervised structural feature guided convolutional neural network Architecture

The network performs better when the two halves work together to fully utilize the global and local information in the histology picture. Additionally, it resolves the three issues with histology picture registration: section missing, repeating texture, and multiple staining. This technique, regardless of the entire scale or the expanded region of interest, satisfies several requirements for histological image registration. Furthermore, due of the severe damage to the structural feature, SFG is unable to conduct picture registration with significantly torn regions.

2.10 AFEM-Genetic algorithm

An ASC (Attribute Scattering center) feature extraction approach using genetic algorithm (AFEM-GA) was presented by Maoqiang Jing et al. [16]. This model is useful for inverse scattering problems and gives clear and physically meaningful features for complicated targets. The following is the suggested

method's flowchart in Fig. 4:

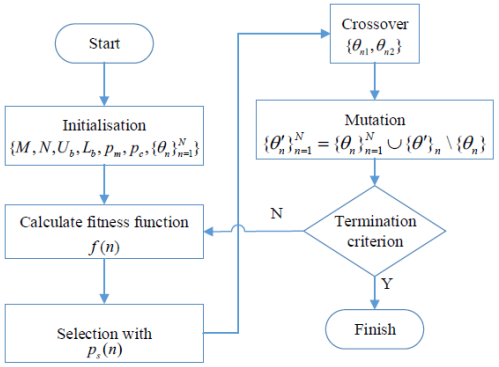


Figure 4: AFEM- Genetic algorithm flowchart

Compared to other methods, the suggested method is more computationally efficient since AFEM-GA does not require a large sparse dictionary. Improved AFEM-GA convergence speed and parallel estimate of all ASCs can be studied in more detail.

2.11 Gated stacked target-related autoencoder (GSTAE)

A gated stacked target-related autoencoder (GSTAE) was first presented by Qingqiang Sun et al. [17]. One of the most challenging problems in the world of image processing is still obtaining an appropriate feature representation from complex process data. In order to fairly take target-related information into account, the suggested STAE conducted a layer-wise pretraining process and incorporated the anticipated loss item of the target value into the original loss function of a common autoencoder.

The suggested method's drawbacks are that it becomes more computationally demanding and is susceptible to samples with low-quality labels. Therefore, it is crucial to remove low-value gate connections as soon as possible in order to minimize needless computational load.

2.12 Superpixel-based multiple statistical feature extraction (SPMSFE)

A unique multiple statistical feature extraction method based on superpixels was proposed by Dan Li et al. [18] (SPMSFE). This method was presented to increase the classification accuracies of hyperspectral images (HSIs), particularly when employing a small number of training samples. The suggested SPMSFE approach consisted of three steps: multiple statistical FE, multitask kernel sparse representation model for classification, and identification of superpixel-based neighbors. The suggested method's algorithm is as in Fig. 5:

The original HSI data set $\mathcal{X}$; number of dimensions l; number of superpixels L; threshold of the superpixel-based neighbors M.
Ensure:
1: Employ MNF on $\mathcal{X} \in \mathbb{R}^{s_1 \times s_2 \times d}$ to obtain the dimension-reduced data $\mathcal{F} \in \mathbb{R}^{s_1 \times s_2 \times l}$;
2: Generate 2-D superpixel map;
3: Identify the most similar superpixel-based neighbors for each pixel as $\mathbf{F}_i = [f_i^1, f_i^2, \ldots, f_i^{M'}] \in \mathbb{R}^{l \times M'}$;
4: Extract the mean feature $\boldsymbol{m}_i$, covariance descriptor $\mathbf{C}_i$, and Gaussian mixture feature $\mathbf{Y}_i$ for each pixel;
5: Utilize these three statistical features as the spatial–spectral features $\mathbf{Z}^r = [z_1^r, z_2^r, \ldots, z_n^r], r = 1, 2, 3.$;
6: Design the multiple kernels k_r using (13), (14), and (15) to map the multiple statistical features to the high-dimensional Hilbert space;
7: Randomly select N training samples to construct the directory $\mathbf{D}^r$ in the statistical feature space;
8: **for** each test pixel $z^r, r = 1, 2, 3.$ **do**
9: Obtain the sparse representation $\boldsymbol{\alpha}^r$ based on Algorithm 1;
10: Calculate the reconstruction error of each class in a multitask model using (17);
11: Determine the class label of z^r using (18).
12: **end for**
Output: The result classification map

Figure 5: Superpixel-based multiple statistical feature extraction Algorithm

This strategy yielded an overall accuracy of 79.08%. The experimental results show that the suggested approach outperforms the most advanced classification techniques, especially in situations with a very small number of training examples. In order to handle the heterogeneous fusion of numerous statistical features and enhance performance, it will be necessary to examine the specific effects of each statistical feature for HSI images in depth in the future.

2.13 Superpixelwise PCA (S^3-PCA)

A unique spectral-spatial and superpixelwise PCA (S3-PCA) was proposed by Xin Zhang et al. [19]. This method effectively utilizes spatial information to extract the global-local and spectral spatial features for HSIs images and to remove noise via local reconstruction based on superpixels. The following is the suggested method's flowchart:

This technique is particularly useful for extracting useful features when working with noisy pixels in tiny, homogeneous regions. This strategy yielded an overall accuracy of 95.63%. Future work may integrate additional techniques like SVM and deep learning models, as superpixel-based local reconstruction has shown to be a successful filter for denoising high-spatiality images.

2.14 Layout feature extraction using CNN classification

Convolutional neural networks (CNNs) were used by Yoshikazu Nagamura [20] to categorize LSI layout photos in order to carry out the RCA of layout-induced errors. While predicting unknown flaws is a difficult task in the current study, analyzing layout-induced faults is essential to optimizing design rules. In order to extend the number of training datasets, this method addressed the limitation by chopping image

clips at any place, including true defect positions. CNN models will eventually be trained on a tiny dataset of faulty layouts in order to forecast problems that are not yet known.

2.15 Fusion of deep learning-based features and empirical features

A fusion approach of merging adaptive characteristics generated by a deep neural network with empirical features was proposed by Jingsong Xie [21]. When dealing with objects that have low-quality training data, intelligent feature extraction and classification approaches cannot ensure that the model has learned the general characteristics needed for classification. Moreover, the model's robustness and generalization may be weak. This is a novel bearing fault classification technique built on top of XGBoost.

A few basic EFs and the adaptive features that LiftingNet extracted are combined to create the features that are utilized to train the XGBoost model[22] classifier. The suggested method's overall accuracy is 95.0%. The findings indicate that by combining a few EFs with the hidden features that the modified LiftingNet extracted, it is possible to enhance the model's accuracy and antinoise effect. Future research will concentrate on implementing the suggested approach in a variety of working environments [23].

Table 1: Summary of deep learning methods

S. No.	Title	Techniques/algorithms	Advantages	Disadvantages	Year
	A Cloud-Guided Feature Extraction Approach for Image Retrieval in Mobile Edge Computing	WAPL algorithm, projection matrix learning	- Reduces network traffic by 93% -Improves retrieval accuracy by 6.9% compared to the LBP algorithm	-	2019
	Unsupervised Hyperspectral Feature Extraction Using Variational Autoencoder	Variational Autoencoder (VAE), Deep Neural Networks (DNN), Local Linear Embedding (LLE), Autoencoder (AE), Segmented Stacked AE (S-SAE)	-Effective in extracting spectral-spatial features -High classification accuracy compared to pixel-wise methods - Utilizes both local and sequential sensing for better feature representation	- Increased computational time due to complex architecture - May not fully leverage spatial information in some methods	2020
	Feature Extraction via 3-D Block Characteristics Sharing for Hyperspectral Image Classification	3-D Block Classification System (3-D-BCS), Support Vector Machine (SVM), Structure-aware Collaborative Representation with Tikhonov Regularization (SaCRT), Joint Sparse Representation Classifier (JSRC), Enhanced Patch-based Filtering (EPF), Superpixel-based Classification via Multiple Kernels (SC-MK), Weighted Markov Random Fields	Improved classification accuracy, computational efficiency, and robustness across various datasets	Computational complexity due to Gabor filters and superpixel segmentation	2020

		(WMRF), Graph-based Deep Network (GFDN)			
	Symmetric All Convolutional Neural-Network-Based Unsupervised Feature Extraction for Hyperspectral Images Classification	3D Convolutional Neural Networks (3D-CNN), Encoder-Decoder Architecture	Unsupervised training reduces the need for labeled samples -Robust feature extraction for classification tasks -Effective in handling both spectral and spatial information	- Increased parameters may deteriorate feature quality - Requires careful tuning of the cost function to avoid overfitting	2020
	A Novel Bearing Fault Classification Method Based on XGBoost: The Fusion of Deep Learning-Based Features and Empirical Features	XGBoost, LiftingNet, PCA	-High classification accuracy (92%) compared to other methods (e.g., SVM: 88%) - Robustness to noise - Combines empirical features with deep learning for better decision-making	Requires some prior knowledge for empirical features	2020
	Gradient-Based Feature Extraction From Raw Bayer Pattern Images	Central difference gradient, HOG, SIFT	Efficient extraction from Bayer images, reduced need for demosaicing	Performance issues in complex textures	2021
	Unsupervised Deep Image Stitching Framework	Unsupervised coarse image alignment, unsupervised image reconstruction, ablation-based loss, stitching-domain transformer layer	Overcomes limitations of feature-based and supervised methods, suitable for large-baseline scenes, reduces space occupied by feature maps	Seam distortions, limited resolution in some cases	2021
	SpaSSA: Superpixelwise Adaptive SSA for Unsupervised Spatial-Spectral Feature Extraction in Hyperspectral Image	Superpixelwise Adaptive SSA (SpaSSA), 1DSSA, 2DSSA, PCA	-Reduces computational complexity and cost by operating on superpixels instead of full images. -Improves classification accuracy and efficiency in hyperspectral image analysis. -Adaptive window sizes enhance feature	- Sensitivity to window size can affect performance. - Lower accuracy on certain datasets like Pavia University due to ignored small ground	2021

			extraction.	features.	
	Attributed Scattering Center Extraction With Genetic Algorithm	Genetic Algorithm (GA)	Faster convergence speed due to floating-point encoding - Improved accuracy in parameter estimation	Computational cost increases with the number of parameters	2021
10	Super pixel-Based Multiple Statistical Feature Extraction Method for Classification of Hyperspectral Images	SPMSFE, MTKSRC, SSHGDA, SFMKSRC, LCMR, SRST, ISSR-DCNN, MCMs-2DCNN, S-DMM	-Superior classification accuracy (more than 3% higher OA value) -Faster computational speed compared to deep learning methods - Robustness with limited training samples	- Higher computational costs due to multiple statistical feature extraction - Gaussian distribution may not apply to all HSI data sets	2021
11	Layout Feature Extraction Using CNN Classification in Root Cause Analysis of LSI	Convolutional Neural Network (CNN), VGG16	-Superior classification ability for layout patterns -Capable of recognizing the local density of patterns -Can learn multiple layout patterns	- Requires a large amount of data for training -Low-resolution images may hinder	2021
	Defects		with different shapes	accurate defect prediction	
12	Unsupervised Histological Image Registration Using Structural Feature Guided Convolutional Neural Network	Structural Feature Guided Convolutional Neural Network (SFG)	Achieved the lowest Median rTRE among submissions to the ANHIR challenge; robust to multiple staining; integrates multi-scale structural features for better registration	AArTRE and AMrTRE are not the best due to challenges with extremely torn tissue sections; relies on automatic key point extraction which may not always be accurate	2022
13	GSTAE for Soft Sensing	STAE, GSTAE (Gated Stacked Autoencoder)	Improved prediction accuracy, especially for high-value samples -Better performance in soft sensing applications -Utilizes gated neurons for effective informati	Complexity in model structure due to multiple layers - Potential computational burden with increa	2022

			on flow control	sed layers	
14	Spectral-Spatial and Superpixelwise PCA for Unsupervised Feature Extraction of Hyperspectral Imagery	PCA, SuperPCA, S3-PCA	Achieves the best performance in terms of Overall Accuracy (OA), Average Accuracy (AA), and Kappa; effectively combines local and global features for improved classification	May neglect global information in small homogeneous regions or large regions with mixed ground truth objects	2022
15	A Feature Fusion-Net Using Deep Spatial Context Encoder and Nonstationary Joint Statistical Model for High-Resolution SAR Image Classification	Deep Spatial Context Encoder (DSCEN), Nonstationary Joint Statistical Model (NS-JSM), Feature Fusion Network (Fusion-Net)	-Integrates spatial and statistical features for improved classification accuracy. -The lightweight structure of DSCEN allows effective training with fewer samples. -NS-JSM captures inter-scale and intra-scale nonstationary correlation, enhancing feature compactness.	-The models used may be too simple to better model different types of image patches, potentially limiting performance	2023

3 Discussion

This paper has reviewed various feature extraction techniques for iris analysis, focusing on both traditional and deep learning-based approaches. A key observation from the related works is that deep learning methods, particularly Convolutional Neural Networks (CNNs), consistently outperform traditional approaches like Gabor filters and Histogram of Oriented Gradients (HOG) in terms of accuracy and robustness. CNN-based approaches excel in capturing intricate details such as lacunae, Wolfflin nodules, and contraction furrows, which are vital for effective iris recognition.

The reviewed works demonstrate that CNN architectures tailored for iris recognition achieve higher performance metrics, with accuracy rates typically exceeding 95% in controlled environments. These methods can adaptively learn relevant features directly from raw iris images, allowing them to detect minute details that might be overlooked by hand-crafted feature extraction techniques. For instance, studies employing variations of CNNs, such as U-Net and MobileNetV2, have shown significant improvements in segmentation accuracy, which is crucial for reliable iris feature extraction.

In contrast, traditional methods often struggle with the inherent variability in iris images caused by factors such as occlusions, illumination changes, and pupil dilation. Although techniques like Gabor filters are adept at capturing texture information, they are less effective in handling these variations, leading to reduced robustness in real-world scenarios.

In the context of evaluating image feature extraction methods, particularly in iris analysis, it is crucial to discuss the evaluation metrics used across the surveyed methods. Common evaluation metrics include accuracy, precision, recall, F1-score, and computational efficiency. Accuracy measures the proportion of correct predictions among the total predictions made, providing a straightforward assessment of a model's performance. Precision indicates the proportion of true positive results about the total predicted positives, while recall measures the model's ability to identify all relevant instances within the dataset. The F1-score is the harmonic mean of precision and recall, offering a balanced measure when dealing with class imbalances. Additionally, computational efficiency encompasses metrics such as training time, inference time, and resource utilization, which are essential for understanding the practical feasibility of deploying these models in real-world scenarios. Including a dedicated section on these evaluation metrics will provide a clearer framework for comparing the effectiveness of different methods and will help readers assess the strengths and weaknesses of each approach in a standardized manner.

The superiority of CNN-based approaches lies in their ability to automatically learn complex patterns and hierarchies from data. This advantage becomes

especially important when dealing with the intricate structures present in the iris, where minute details play a pivotal role in identification. Moreover, deep learning methods can be fine-tuned with large-scale datasets, improving their generalizability and robustness across different environments. Another benefit is their ability to integrate multi-scale features, enabling the capture of both global and local patterns in iris images.

4 Conclusion

Most image feature extraction algorithms based on deep learning methods have demonstrated impressive results in the field of image processing, particularly in applications requiring high accuracy, such as medical imaging. The effectiveness of these algorithms often hinges on their ability to learn and extract relevant features directly from raw data, reducing the need for manual feature engineering. For medical imaging, where precision is critical, the performance of these algorithms must be optimized to ensure accurate diagnoses and assessments.

Deep learning-based methods have consistently outperformed traditional approaches, thanks to their capacity to handle complex and high-dimensional data. Techniques such as Convolutional Neural Networks (CNNs) and variations thereof have shown significant advancements in extracting minute details from images, such as those found in iris analysis. In our summary of deep learning-based methods, we highlighted their strengths, including their ability to adapt and improve with larger datasets and their robustness in various medical applications.

In the future directions, integrating multiple deep learning algorithms could further enhance accuracy in medical iris diagnosis and other related fields. By combining the strengths of different architectures, such as CNNs with Recurrent Neural Networks (RNNs), or employing ensemble methods, researchers can leverage complementary features and improve the overall performance of diagnostic systems. This future direction holds great promise for advancing medical imaging techniques and ensuring better outcomes in patient care and diagnosis.

References

[1] Ezhilarasan, M.," Image Enhancement for Iris Analysis Using Deep Learning Algorithms-A Brief Survey", In *2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE)*, pp. 1-5, 2024, February. https://doi.org10.1109/icETITE58242.2024.10493485

[2] Poovayar Priya, M., and Ezhilarasan, M., "Iris segmentation technique using the iris-unet method", In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pp. 235-249, 2023. https://doi.org/10.1007/978-3-031-45382-3_20

[3] Rafael C. Gonzalez, and Richard E. Wood, "Digital Image Processing", 3rd Edition, 2014.

[4] Priya, M.P. and Ezhilarasan, M., "A Hybrid Technique of Diagnosing Kidney Disorder Using Deep Neural Network", *International Journal of Online & Biomedical Engineering*, vol. 20, no. 15, 2024. https://doi.org/10.3991/ijoe.v20i15.50409

[5] Jensen B," Iridology simplified", California: Bernard Jensen, 1980.

[6] Mohan, P.P. and Murugesan, E., "A novel approach for detecting the presence of a lymphatic rosary in the iris using a 3-stage algorithm", *AIMS Bioengineering*, vol. *11*, no. 4, 2024. https://doi.org/10.3934/bioeng.2024023

[7] Zhou, W., Zhang, L., Gao, S. and Lou, X., "Gradient-based feature extraction from raw bayer pattern images", *IEEE Transactions on Image Processing*, vol. *30*, pp.5122-5137, 2021. https://doi.org/10.1109/TIP.2021.3067166

[8] Liang, W., Wu, Y., Li, M., Zhang, P., Cao, Y. and Hu, X., "A Feature Fusion-Net Using Deep Spatial Context Encoder and Nonstationary Joint Statistical Model for High-Resolution SAR Image Classification", 2021. https://doi.org/10.1109/TGRS.2020.3042274

[9] Nie, L., Lin, C., Liao, K., Liu, S. and Zhao, Y., "Unsupervised deep image stitching: Reconstructing stitched features to images", *IEEE Transactions on Image Processing*, vol. *30*, pp.6184-6197, 2021. https://doi.org/10.1109/TIP.2021.3092733

[10] Sun, G., Fu, H., Ren, J., Zhang, A., Zabalza, J., Jia, X. and Zhao, H., "SpaSSA: Superpixelwise adaptive SSA for unsupervised spatial-spectral feature extraction in the hyperspectral image", *IEEE Transactions on Cybernetics*, 2021. https://doi.org/10.1109/TCYB.2021.3104100

[11] Zhang, M., Gong, M., He, H. and Zhu, S., "Symmetric all convolutional neural-network-based unsupervised feature extraction for hyperspectral images classification", *IEEE Transactions on Cybernetics*, 2020. https://doi.org/10.1109/TCYB.2020.3020540

[12] Yu, W., Zhang, M. and Shen, Y., "Spatial revising variational autoencoder-based feature extraction method for hyperspectral images", *IEEE Transactions on Geoscience and Remote Sensing*, vol. *59*, No. 2, pp.1410-1423, 2020. https://doi.org/10.1109/TGRS.2020.2993141

[13] Tu, B., Zhou, C., Liao, X., Li, Q. and Peng, Y., "Feature extraction via 3-D block characteristics sharing for hyperspectral image

classification", *IEEE Transactions on Geoscience and Remote Sensing*, vol. *59*, No. 12, pp.10503-10518, 2020. https://doi.org/10.1109/TGRS.2020.3042274

[14] Wang, S., Ding, C., Zhang, N., Liu, X., Zhou, A., Cao, J. and Shen, X., "A cloud-guided feature extraction approach for image retrieval in mobile edge computing", *IEEE Transactions on Mobile Computing*, vol. *20*, No. 2, pp.292-305, 2019. https://doi.org/10.1109/TMC.2019.2960939

[15] Ge, L., Wei, X., Hao, Y., Luo, J. and Xu, Y., "Unsupervised Histological Image Registration Using Structural Feature Guided Convolutional Neural Network", IEEE *Transactions on Medical Imaging*, 2022. https://doi.org/10.1109/TMI.2021.3118464

[16] Jing, M. and Zhang, G., Attributed scattering center extraction with genetic algorithm. *IEEE Transactions on Antennas and Propagation*, vol. *69*, No. 5, pp. 2810-2819, 2020. https://doi.org/10.1109/TAP.2020.3027630

[17] Sun, Q. and Ge, Z., "Gated stacked target-related autoencoder: A novel deep feature extraction and layerwise ensemble method for industrial soft sensor application", *IEEE Transactions on Cybernetics*, 2022. https://doi.org/10.1109/TCYB.2021.3087993

[18] Li, D., Kong, F., Liu, J. and Wang, Q., "Superpixel-Based Multiple Statistical Feature Extraction Method for Classification of Hyperspectral Images", *IEEE Transactions on Geoscience and Remote Sensing*, vol. *59*, No. 10, pp.8738-8753, 2021. https://doi.org/10.1109/TGRS.2020.3031726

[19] Zhang, X., Jiang, X., Jiang, J., Zhang, Y., Liu, X. and Cai, Z., "Spectral–spatial and superpixelwise PCA for unsupervised feature extraction of hyperspectral imagery", *IEEETransactions on Geoscience and Remote Sensing*, *60*, pp.1-10, 2021. https://doi.org/10.1109/TGRS.2021.3067167

[20] Nagamura, Y., Arima, K., Arai, M. and Fukumoto, S., "Layout Feature Extraction Using CNN Classification in Root Cause Analysis of LSI Defects", *IEEE Transactions on Semiconductor Manufacturing*, vol. *34*, No. 2, pp.153-160, 2021. https://doi.org/10.1109/TSM.2021.3067168

[21] Xiong, Y.," Development of an AI-Driven Model for Drug Sales Prediction Using Enhanced Golden Eagle Optimization and XGBoost Algorithm", *Informatica*, vol. *49*, no. 17, 2025. https://doi.org/10.31449/inf.v49i17.7491

[22] El Houby, E.M., "A Review of Machine Learning Techniques in the Medical Domain", *Informatica*, vol. *49*, no. 16, 2025. https://doi.org/10.31449/inf.v49i16.6934

[23] Xie, J., Li, Z., Zhou, Z. and Liu, S., "A novel bearing fault classification method based on XGBoost: The fusion of deep learning-based features and empirical features", *IEEE Transactions on Instrumentation and Measurement*, vol. *70*, pp.1-9, 2020. https:/doi.org/10.1109/TIM.2020.3042275

Optimization for Shipping Logistics Paths Based on Evolutionary Ant Colony Algorithm: Improvement and Application of Dual Population Mechanism

Xiaoxian Shao
Chengyi College, Jimei University, Xiamen, Fujian 361021, China
E-mail: ulmma1010@163.com

Keywords: shipping logistics, shipping vessels, path optimization, ant colony dual population, path contribution degree, carbon emissions, path reliability

Received: February 26, 2025

To improve the efficiency of shipping logistics, reduce transportation costs, and minimize energy consumption, this study introduces a dual population mechanism to improve the conventional ant colony algorithm and applies it to optimizing shipping logistics paths. Firstly, the shipping logistics network is abstracted as a set of nodes and edges in graph theory, simplifying the complex logistics network structure and providing a framework and theoretical basis for subsequent ant colony algorithm applications. Then, objective functions are set from four aspects: minimizing logistics transportation expenses, minimizing logistics transportation time, minimizing carbon emissions, and maximizing path reliability to guide the algorithm in searching for the optimal solution. Finally, a dual population mechanism is introduced, utilizing two independent ant populations for parallel search. Population 1 adopts an elite ant strategy to achieve fast convergence, while Population 2 uses an enhanced sub path evaluation mechanism to explore new solution spaces and help the population escape from local optima. By using the path contribution evaluation mechanism, better paths can be selected to obtain the optimal shipping logistics path. According to the simulation results, it can be seen that the total transportation time of this method is 41 days throughout the entire experimental cycle, saving 9 days and 7 days respectively compared to the two existing methods; The total transportation expenses of this method is 1250000 USD, saving 170000 USD and 130000 USD respectively compared to the two existing methods; The total carbon emissions of this method are 11800 tons, saving 1700 tons and 1400 tons respectively compared to the two existing methods. It can be seen that this method outperforms existing methods in terms of total transportation time, total transportation expenses and total carbon emissions, indicating that this method effectively achieves the design expectations.

Povzetek: Prispevek uvaja izboljšan evolucijski algoritem dveh populacij mravelj za optimizacijo pomorskih logističnih poti, ki zmanjša stroške, čas in emisije ter poveča zanesljivost poti.

1 Introduction

In today's globalized economic landscape, shipping and logistics play a crucial role. The vast majority of global trade relies on shipping for the transportation of goods, which connects major economies and trade centers around the world [1]. According to statistics, over 80% of global trade goods are transported by sea, which indicates the core position of shipping logistics in international trade. Therefore, the efficiency of shipping logistics directly affects the operational efficiency of the global economy and trade costs.

However, traditional shipping logistics path planning has many problems. On the one hand, due to the lack of precise data analysis and real-time information, route selection is often fixed and conservative [2]. This may result in ships encountering unnecessary detours, congestion, or adverse weather conditions during navigation, thereby increasing transportation time and costs. On the other hand, traditional path planning takes less into account environmental factors, and the high-energy consumption of navigation not only increases operating costs, but also puts significant pressure on the marine environment.

Given the shortcomings of traditional shipping logistics paths, researching optimization methods for shipping logistics paths has important practical significance. By utilizing advanced information technology, big data analysis, and intelligent algorithms, dynamic planning and real-time adjustment of shipping routes can be achieved [3]. This can not only improve the efficiency of shipping logistics, reduce transportation costs, but also reduce energy consumption and environmental pollution.

Therefore, based on the consideration of fuel supply strategy in reference [4], a transportation fleet path optimization method was designed. After in-depth research on the uniqueness of grain shipping and the characteristics of changes in oil prices at fuel supply ports over time and place, this method explores the relationship

between oil prices and supply quantities at fuel supply ports, as well as the interaction between average oil prices at fuel supply ports and port service fees. Then, based on the transportation characteristics of the grain fleet, the selection of fuel supply ports, specific fuel supply quantities, and ship navigation routes were selected as key decision-making factors, and a path model integrating integer programming was constructed accordingly. However, although this method considers fuel supply strategies, it may be conservative in path selection and may not fully consider time efficiency. In order to reduce fuel costs or port service fees, longer shipping routes or more transit ports may be chosen, thereby increasing the total transportation time. In reference [5], a navigation vessel diversion path planning method was designed for peak maritime traffic periods. This method first sets a comprehensive objective function aimed at minimizing the length of the voyage, improving navigation safety, and ensuring smooth navigation. Then, the grid method is used to reproduce the ship operation status during the peak period of maritime traffic. Finally, using an improved genetic algorithm, the optimal ship diversion path planning scheme was searched globally. With the help of nonlinear programming techniques, the local optimal solution of the model was further solved, thereby accurately determining the optimal ship diversion path. Although this method aims to minimize the length of the voyage, it may not have taken into account other cost factors such as fuel costs, port charges, etc. In addition, due to the relatively single objective function, it may not be possible to achieve optimal total transportation expenses. On the basis of considering carbon emissions, a method for optimizing oil tanker route scheduling was designed in reference [6]. A green scheduling optimization model for shuttle oil tankers was designed in this method, with the core objective of minimizing the overall cost of shipping. This cost structure includes both traditional fixed transportation costs and variable costs closely related to carbon emissions, reflecting considerations for environmental benefits. In addition, the model not only focuses on optimizing the navigation route planning of shuttle oil tankers, but also finely adjusts the composition and configuration of shuttle oil tanker fleets, thus achieving dual optimization from fleet design to navigation paths. Although this method has included carbon emission costs as part of variable costs, there may still be other cost factors that have not been fully considered or the weighting settings are unreasonable. In addition, with changes in market conditions such as fluctuations in fuel prices, port fee adjustments, etc., this method may need to be continuously updated and adjusted to adapt to new cost structures.

Based on the above analysis, with the goal of improving shipping logistics efficiency, reducing transportation costs, and decreasing energy consumption, this study designs a shipping logistics path optimization method based on evolutionary ant colony algorithm. The design concept of this method is as follows:

(1) Abstracting the shipping logistics network as a directed weighted graph in graph theory, where nodes represent ports, edges represent transportation paths, and weights include distance, time, cost, etc. Build a mathematical model based on the Traveling Salesman Problem (TSP) to adapt to the dynamic changes in shipping logistics demand.

(2) Design objective functions from four aspects: minimizing logistics transportation expenses, minimizing logistics transportation time, minimizing carbon emissions, and maximizing path reliability. Set port capacity constraints, time window constraints, fuel consumption constraints, port congestion constraints, etc. to ensure the feasibility and practical operability of the path.

(3) To improve the search efficiency and global optimization ability of the algorithm, and avoid getting stuck in local optima, a dual population mechanism is introduced. For population 1, the elite ant strategy is adopted to quickly converge to the local optimal solution; For population 2, a strengthened sub path evaluation mechanism is adopted to explore new solution spaces and escape from local optima. On this basis, a path contribution evaluation mechanism is designed to compare the optimal paths of two populations and select the path with higher contribution as the global optimal solution. The dual population mechanism increases the diversity and comprehensiveness of the search, and the path contribution evaluation mechanism further improves the quality of the global optimal solution.

(4) Use the improved ant colony algorithm to solve the objective function. Initialize ant colony algorithm parameters, construct a directed weighted graph, and perform parallel search to generate two candidate paths. Update pheromones based on elite ant strategy and strengthened sub path evaluation mechanism, calculate path contribution, and screen for better paths. Iterate the search until the global optimal path is no longer updated, and output the optimal path that satisfies the multi-objective function. This process achieves efficient and accurate path optimization through parallel search and pheromone update mechanisms, significantly improving the convergence speed and solution quality of the algorithm.

2 Design of optimization methods for shipping logistics paths

2.1 Analysis of TSP problems in shipping logistics

In the field of path selection and planning, TSP is a classic combinatorial optimization problem [7], which can be described as: given a series of cities and the distance between each pair of cities, solving the shortest path to visit each city once and return to the starting city.

Therefore, this study first abstracts the shipping logistics network as a set of nodes and edges in graph theory. Use nodes to represent key locations of logistics ports, edges to represent transportation paths, and analyze their TSP mathematical models. This step not only simplifies the structure of complex logistics networks, but also provides a clear framework and theoretical basis for the subsequent application of ant colony algorithms, making path optimization problems easier to analyze, solve, and verify. At the same time, this abstract method has high flexibility and scalability, and can adapt to the dynamic changes and multi-objective optimization requirements in shipping logistics.

Set up a directed weighted graph $G(V,D)$, as shown in Fig. 1.

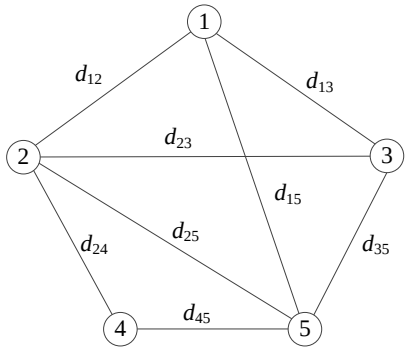


Figure 1: Structure diagram of directed weighted graph (taking a directed weighted graph with 5 vertices as an example)

Among them:

Using different cities (logistics ports) as vertices, obtain a vertex set $V=\{1,2,\cdots,i,\cdots,j,\cdots,n\}$;

Take the line connecting city i and city j as the weighted edge, and the set of edges is $D=\{d_{ij}|i,j\in n\}$.

Shipping logistics networks typically involve multiple ports, multiple modes of transportation, and complex constraints [8]. Through graph theory abstraction, they can be simplified into a computable mathematical model for subsequent analysis and optimization.

Then, assuming the existence of $s_{ij}=\begin{cases}1, & i \text{ is directly connected to } j\\ 0, & \text{others}\end{cases}$, TSP mathematical model of shipping logistics can be expressed in the following form:

$$M=\min\left(\sum_{i,j=1}^{n} d_{ij}s_{ij}\right)$$
$$\text{s.t.}\begin{cases}\sum_{i=1}^{n} s_{ij}=\sum_{j=1}^{n} s_{ij}=1\\ \sum_{i,j\in S} s_{ij}=|S|\end{cases} \quad (1)$$

In Equation (1), S represents a subgraph of G; $|S|$ represents the number of cities in S.

The TSP problem requires finding a path that passes through all nodes and has the minimum total cost, which is highly consistent with the goal of optimizing shipping logistics paths. In shipping logistics, it is usually necessary to plan a path from the starting point to the destination, passing through multiple ports, while minimizing transportation costs, time, or carbon emissions. The above TSP model can well describe this requirement.

On this basis, due to the complexity of logistics in reality, the application of TSP needs to fully consider travel time and cost. Therefore, this study extended the dynamic TSP model. Set up rolling time-domain optimization in TSP problem, decompose global path planning into multiple time windows, re optimize local paths based on the latest data within each window, adjust subsequent paths every 3 hours according to updated weather and port status, and enhance dynamic response capability.

2.2 Objective function and constraint setting for optimizing shipping logistics paths

In the research of optimizing shipping logistics paths, the design of the objective function is crucial. The objective function aims to quantify the performance indicators of the optimized path to guide the algorithm in searching for the optimal solution. Therefore, based on the characteristics of optimizing shipping logistics paths, this study sets objective functions from four aspects: minimizing logistics transportation expenses, minimizing logistics transportation time, minimizing carbon emissions, and maximizing path reliability.

(A) Minimizing logistics and transportation expenses

Reducing transportation expenses is one of the core goals of shipping and logistics enterprises, which directly affects their profitability and market competitiveness [9]. The objective function aims to minimize the total transportation expenses in shipping logistics, including fuel costs, port costs, ship maintenance costs, labor costs, etc. By optimizing the route and selecting the lowest cost route and transit port, the operating expenses of shipping and logistics enterprises can be reduced.

Shipping and logistics transportation expenses mainly include five parts: fuel expenses, port fees, loading and unloading fees, carbon emission expenses, and surcharges. Therefore, the objective function is established as follows:

$$F_1=\min\left(A_1+A_2+A_3+A_4+A_5\right) \quad (2)$$
$$\min\left(\sum_{i,j\in V} C_{ij}f_1t_{ij}P+\sum_{i,j\in V}\sum_{j\in N} C_j f_2 t_j P+sO_j+\sum_{j=1}^{n}\left(L_j+U_j\right)+\sum_{i,j\in V} C_{ij}f_1t_{ij}kc+\sum_{j\in N} C_j\varepsilon_j\right)$$

In Equation (2), $C_{ij}=\begin{cases}1, \text{ if Choose the}\\ \text{route from port } i \text{ to } j\\ 0, \text{ otherwise}\end{cases}$ represents the path selection coefficient; $C_j=\begin{cases}1, \text{ if select port } j \text{ as the port of call}\\ 0, \text{ otherwise}\end{cases}$ represents the port selection coefficient; The meanings of the remaining parameters are as follows: A_1, A_2, A_3, A_4, A_5 respectively represent fuel expenses, port fees, loading and unloading fees, carbon emission expenses, and surcharges; V represents a collection of shipping networks, $i, j \in V$; f_1 and f_2 respectively represent the daily fuel consumption of shipping vessels in sailing and berthing states; P represents fuel price; N represents the set of n ports (i.e. cities in the directed weighted graph) that can be docked in the entire shipping network; t_j represents the berthing time of shipping vessels at port j; s represents the number of times fuel is replenished; O_j represents the operational cost incurred by the shipping vessel for refueling at port j; L_j and U_j respectively represent the allowable container loading and unloading capacity at port j; t_{ij} represents the transportation time from Port i to Port j; k represents the carbon emission tax rate; c represents the carbon emission factor; ε_j represents the additional fee for shipping vessels at port j.

Set the following constraints for the objective function of minimizing logistics transportation expenses:

(1) The actual cargo volume of containers between ports shall not exceed the maximum loading capacity of the sea vessel, that is:

$$\text{s.t. } Q_{ij} \le H_{\max}, \forall i, j \in V \tag{3}$$

In Equation (3), Q_{ij} represents the container freight volume between ports i and j; $H_{\max}$ represents the maximum container capacity of the shipping vessel.

(2) Ensure that the total container load at all ports matches the total unloading volume, that is:

$$\text{s.t. } \sum_{i=1}^{n} L_i C_{ij} = \sum_{i=1}^{n} U_i C_{ij}, \forall i, j \in \{0,1,2,\cdots,n\} \tag{4}$$

(3) The maximum number of times a port can be visited is once, that is:

$$\text{s.t. } 0 \le C_j \le 1 \tag{5}$$

(4) The number of ports of call selected shall not exceed the required total, that is:

$$\text{s.t. } \sum_{j=1}^{n} C_j \le \kappa_V \tag{6}$$

In Equation (6), κ_V represents the maximum number of ports that can be docked.

(5) Ensure the conservation of container loading and unloading volume at each dock [10], that is:

$$\begin{aligned}&\text{s.t.}\\ &\sum_{i=1}^{n} Q_{ij} = U_j\\ &\sum_{j=1}^{n} Q_{ij} = U_i\end{aligned} \tag{7}$$

(6) Ensure that the total freight volume meets the minimum required freight volume, that is:

$$\begin{aligned}&\text{s.t.}\\ &\sum_{i=1}^{n}\sum_{j=1}^{n} Q_{ij} \ge Q\\ &\sum_{j=1}^{n} Q_{ij} = U_i\end{aligned} \tag{8}$$

In Equation (8), Q represents the minimum container freight volume required to achieve the predetermined revenue.

(7) The decision variable conforms to the 0-1 constraint, that is:

$$\begin{aligned}&\text{s.t.}\\ &C_{ij}, C_j \in \{0,1\}\end{aligned} \tag{9}$$

(8) Minimize logistics transportation time

Shortening transportation time can meet customers' demand for timeliness, reduce the uncertainty of goods in transit, and improve the turnover rate of ships [11]. Therefore, the objective function aims to minimize the total transportation time in shipping logistics, including sailing time, port waiting time, loading and unloading time, etc. By optimizing the route and selecting the shortest route and transit port, logistics efficiency can be improved.

Based on the above analysis, the objective function for minimizing logistics transportation time is established as follows:

$$F_2 = \min\left(\sum_{a=1}^{A} t_{sail,a} + \sum_{j=1}^{n} t_{wait,j} + \sum_{k=1}^{K} t_{load/unload,k}\right) \tag{10}$$

In Equation (10), $t_{sail,a}$ represents the sailing time of segment a; $t_{wait,j}$ represents the waiting time at port j; $t_{load/unload,k}$ represents the time of the k-th loading and unloading; A, K represent the number of flight segments and the number of loading and unloading times.

To minimize the objective function of logistics transportation time, the following constraints are set:

(1) Port capacity constraints:

s.t.

$$\sum_{j} t_{wait,j} \le \text{Capacity port } j \tag{11}$$

In Equation (11), $\text{Capacity port } k$ represents the maximum processing capacity or capacity of port j. This constraint ensures that the cumulative waiting time for each port does not exceed its processing capacity.

(2) Time window constraint: The goods need to arrive at a certain port within a specific time window [12], that is:

s.t.

$$t_{arrival,j} \in \left[\text{TimeWindowStart}_j, \text{TimeWindowEnd}_j\right] \tag{12}$$

In Equation (12), TimeWindowStart_j and TimeWindowEnd_j represent the start and end times of the time window for port j, respectively. This constraint requires that the goods arrive at port j within the specified time window.

2.3 Minimize carbon emissions

Reducing carbon emissions is an important goal of green logistics, and this objective function aims to minimize the total carbon emissions in shipping logistics to reduce the impact on the environment. Carbon emissions are usually related to fuel consumption and route distance. By optimizing routes and selecting routes and transit ports with the lowest carbon emissions, green logistics can be achieved [13].

Based on the above analysis, the objective function for minimizing carbon emissions is established as follows:

$$F_3 = \min\left(\sum_{i,j=1}^{n} e_{ij} C_{ij}\right) = \min\left(\sum_{i,j=1}^{n}\sum_{a=1}^{A} c_{fuel,a} d_a C_{ij}\right) \tag{13}$$

In Equation (13), e_{ij} represents the carbon emissions from port i to port j, which are usually related to fuel consumption and route distance; $c_{fuel,a}$ represents the carbon emission coefficient per unit distance of the segment route a; d_a represents the distance of the segment route a.

For the objective function of minimizing carbon emissions, fuel consumption constraints are set, requiring that fuel consumption cannot exceed the fuel reserve of the vessel, that is:

s.t.

$$\sum_{a=1}^{A} \text{FuelConsumption}\left(d_a, c_{fuel,a}\right) \le \text{TotalFuelReserve} \tag{14}$$

In Equation (14), $\text{FuelConsumption}\left(d_a, c_{fuel,a}\right)$ represents the fuel consumption of route segment a; TotalFuelReserve represents the total fuel reserve of the vessel. This constraint ensures that the total fuel consumption does not exceed the vessel's fuel reserve [14].

2.4 Maximizing path reliability

Improving path reliability can ensure timely arrival of goods, reduce unexpected losses during transportation, and enhance customer trust [15]. Therefore, the objective function aims to maximize the reliability of the shipping logistics path, that is, the probability that the path can be smoothly executed in practical operations. Reliability is usually affected by factors such as weather, port congestion, and route safety. By optimizing the route and selecting the most reliable route and transit port, transportation risks can be reduced.

Based on the above analysis, the objective function for maximizing path reliability is established as follows:

$$F_4 = \max\left(\prod_{i,j=1}^{n} r_{ij}^{C_{ij}}\right) \tag{15}$$

In Equation (15), r_{ij} represents the path reliability from port i to port j.

Set the following constraints for the objective function of maximizing path reliability:

(1) Considering the impact of port congestion on reliability, the port congestion constraints are set as follows [16]:

s.t.

$$P_{\text{reliable},j} \ge P_{\text{congestion},j} \tag{16}$$

In Equation (16), $P_{\text{reliable},j}$ represents the reliability probability of port j; $P_{\text{congestion},j}$ represents the reliability probability of port j being affected by congestion. This constraint ensures that the reliability probability of port operations is not lower than the probability allowed by congestion conditions.

(2) No sub loop constraint:

s.t.

$$u_i u_j + nC_{ij} \le n-1 \tag{17}$$

In Equation (17), u_i and u_j represent auxiliary variables used to eliminate sub loops.

In summary, the objective function for optimizing shipping logistics routes is as follows:

$$F = \omega_1 F_1 + \omega_2 F_2 + \omega_3 F_3 + \omega_4 F_4 \tag{18}$$

In Equation (18), ω_1, ω_2, ω_3 and ω_4 are the weight values of the four objective functions respectively.

The process of setting weight factors is as follows: Firstly, due to the different units and magnitudes of each target, normalization is required first. Map each target

value to the interval of 0, 10, and 1 through linear transformation to eliminate dimensional differences. Then, based on the actual needs of shipping companies and external constraints, initial weights are assigned as 0.4, 0.3, 0.2, and 0.1, respectively. On this basis, the Analytic Hierarchy Process (AHP) and expert evaluation are combined to construct a matrix for comparing the importance of objectives. For example, expenses are more important than time (2:1 ratio), time is more important than carbon emissions (1.5:1 ratio), and carbon emissions are more important than reliability (1.2:1 ratio). Finally, through matrix consistency testing (consistency ratio CR<0.1), the weights of each target were calculated, and the final weights were obtained as 0.30, 0.30, 0.20, and 0.20 respectively.

2.5 Optimization for shipping logistics routes based on evolutionary ant colony algorithm

Based on the objective function obtained above, the improved ant colony algorithm is used to solve and obtain the optimal shipping logistics path.

Evolutionary design and application of ant colony algorithm

In the context of shipping logistics, each logistics port is like a city in TSP. Shipping vessels need to visit different ports in sequence for cargo loading and unloading operations, and usually hope to complete all port visits with the optimal path [17, 18].

However, as the number of cities increases, the search space of TSP grows exponentially, which makes it difficult for ant colony algorithms to traverse all possible paths in a limited time, and is susceptible to the accumulation of pheromones during the search process, leading to premature convergence to local optimal solutions. In response to this issue, this study proposes improvements to the ant colony algorithm. The improvement ideas are as follows:

(1) Introduce the idea of dual population and use two independent ant populations for parallel search. Population 1 adopts the elite ant strategy to update pheromones, quickly converges using existing information, and obtains the shortest path of TSP; Population 2 adopts a strengthened sub path evaluation mechanism to update pheromones, and jumps out of local optimal solutions by exploring new solution spaces to obtain the shortest path of TSP;

(2) Comparing the two shortest paths obtained above, using the path contribution evaluation mechanism to update the pheromones of the two populations again, the final TSP solution is obtained, which is the optimal shipping logistics path.

Dual population ant colony algorithm based on path contribution evaluation

Based on the TSP mathematical model established in Section 2.1, the ant colony algorithm is used to solve for the shortest path that traverses all cities (logistics ports) and returns to the starting point. The process is as follows:

Assuming the number of ants is m. At time t, the probability of state transition for ant a moving from city i to city j is $q_{ij}^{a}(t)$:

$$q_{ij}^{a}(t)=\begin{cases}\dfrac{\left[\alpha p_{ij}(t)\right]\left[\beta v_{ij}(t)\right]}{\sum_{k\in V_a}\left[\alpha p_{ik}(t)\right]\left[\beta v_{ik}(t)\right]}, & \text{if } j\in V_a\\ 0, & \text{if } j\notin V_a\end{cases} \tag{19}$$

In Equation (19), V_a represents the set of cities that ant a can select in the future; $p_{ij}(t)$ represents pheromone; $v_{ij}(t)$ represents the visibility of the path; α and β respectively represent the heuristic factors of $p_{ij}(t)$ and $v_{ij}(t)$; k represents any other city except for cities i and j [19].

The update equation for pheromones is:

$$p_{ij}(t+1)=\sum_{a=1}^{m}\Delta p_{ij}^{a}+(1-\gamma+\zeta)\,p_{ij}(t) \tag{20}$$

In Equation (20), in the current loop, Δp_{ij}^{a} represents the pheromone left by ant a on path $i\to j$; γ represents the residual factor of $h_{ij}(t)$ [20]; ζ represents the pheromone decay rate.

Usually, α, β, and ζ significantly affect the performance of algorithms. Therefore, this study conducted empirical research. This study set three sets of experimental parameters:

The first set of parameters: $\alpha=1.5,\beta=2.0,\zeta=0.3$. Under this setting, the algorithm is more inclined to explore new paths in the early stages (due to the larger value of β), but the pheromone decay is relatively slow, which helps maintain the pheromone concentration of the better paths explored in the early stages.

The second set of parameters: $\alpha=2.5,\beta=1.5,\zeta=0.5$. At this point, the algorithm focuses more on utilizing existing information (with a larger value of α), but the pheromone decay is faster, which may lead to premature convergence to the local optimal solution.

The third set of parameters: $\alpha=2.0,\beta=2.0,\zeta=0.4$. This is a compromise setting aimed at balancing exploration and utilization while maintaining a moderate rate of pheromone decay. Therefore, this set of values was adopted in this study.

However, as analyzed earlier, the search space of TSP

is positively correlated with the number of cities, which makes it difficult for ant colony algorithm to traverse all paths in a limited time, and the convenient process is easily affected by the accumulation of pheromones, resulting in premature convergence of the algorithm and difficulty in obtaining the optimal path for the entire site. To address this issue, this study introduces the dual ant colony algorithm, which utilizes two independent ant populations for parallel search when solving the TSP mathematical model. Population 1 adopts the elite ant strategy, focusing on utilizing existing information to quickly converge. Population 2 adopts a strengthened sub path evaluation mechanism, focusing on exploring new solution spaces. This dual population mechanism increases the diversity and comprehensiveness of the search, helping the population to escape from local optima and obtain an optimal path that traverses all cities and returns to the starting point.

For population 1, adopting the elite ant strategy and improving the updating method of pheromones, the equation is as follows:

$$p_{ij}(t+1)=\lambda p_{ij}(t)+\Delta p_{ij}^{*}+\sum_{a=1}^{m}\Delta p_{ij}^{a} \tag{21}$$

In Equation (21), Δp_{ij}^{*} represents the pheromone increment of elite ants on path $i \to j$.

The calculation equations for Δp_{ij}^{a} and Δp_{ij}^{*} are as follows:

$$\Delta p_{ij}^{a}=\begin{cases}\dfrac{\eta}{d_a}, & \text{if } a \text{ passes through path } i \to j \text{ in the current loop}\\ 0, & \text{if } a \text{ does not pass through path } i \to j \text{ in the current loop}\end{cases} \tag{22}$$

$$\Delta p_{ij}^{*}=\begin{cases}m^{*}\times\dfrac{\eta}{d_{best}}, & \text{if path } i \to j \text{ belongs to the optimal solution}\\ 0, & \text{if path } i \to j \text{ does not belong to the optimal solution}\end{cases}$$

In Equation (22), m^{*} represents the number of elite ants; η represents the strength of pheromones; d_a represents the length of the movement path of ant a; d_{best} represents the path length of the optimal solution for TSP.

For population 2, a strengthened sub path evaluation mechanism is used to update pheromones, ensuring accurate partitioning of the contribution of different sub paths when solving TSP and improving the quality of TSP solutions.

After introducing the enhanced negative feedback mechanism, the equation for calculating the state transition probability of population 2 is:

$$q_{ij}^{a}(t)=\frac{\alpha p_{ij}(t)\times\beta v_{ij}(t)\times I_{\delta}\left(1-\delta_{ij}(t)\right)}{\sum_{k\in V_a}\alpha p_{ij}(t)\times\beta v_{ij}(t)\times I_{\delta}\left(1-\delta_{ij}(t)\right)} \tag{23}$$

In Equation (23), $\delta_{ij}(t)$ represents a negative feedback pheromone; I_{δ} represents the heuristic factor of $\delta_{ij}(t)$.

In the enhanced negative feedback mechanism, the increment of sub path length pheromone is introduced to obtain the enhanced sub path evaluation mechanism. The updated values $\Delta p_{ij}(t)$ and $\Delta\delta_{ij}(t)$ of positive and negative feedback pheromones are improved, and the improvement amounts of both are as follows:

$$\begin{cases}\Delta p'_{ij}(t)=\dfrac{\eta}{d_{best}}\times\left(1-\varepsilon\times\dfrac{d_{ij}}{d_{best}}\right)\\ \Delta\delta'_{ij}(t)=\dfrac{\eta}{d_{w}}\times\left(1+\varepsilon\times\dfrac{d_{ij}}{d_{w}}\right)\end{cases} \tag{24}$$

In Equation (24), ε represents the amplification factor of the path factor; d_w represents the worst path within the current loop; μ represents the path magnification factor.

Optimization Solution of TSP

In order to enable two ant populations to discover more potential high-quality solutions and ultimately find higher quality TSP solutions, two paths are obtained for the two populations, and a path contribution evaluation mechanism is introduced to screen for the better path.

The calculation equation for path contribution degree C is as follows:

$$C=\frac{d_a}{d_{best}} \tag{25}$$

Analyzing the complexity of pheromone updates in the above process, it is mainly reflected in the following aspects: Firstly, the mechanism introduces a dual population strategy, where population 1 adopts the elite ant strategy and population 2 adopts the enhanced sub path evaluation mechanism. This parallel search method significantly increases the exploration space of the algorithm, but also brings about the complexity of pheromone updates. Population 1 and Population 2 need to maintain their respective pheromone matrices and update them according to their respective strategies during the iteration process, which increases the consumption of computing resources; Secondly, the update of pheromones not only depends on the path length of the current ant, but also considers parameters such as the proportion of elite ants and the amplification factor of path factors. The introduction of these parameters makes the pheromone update formula more complex, but also improves the algorithm's global search ability and ability to escape from local optima; In addition, the real-time data-driven pheromone reset mechanism enables the algorithm to dynamically adjust the distribution of pheromones according to changes in the external environment, further increasing the complexity and adaptability of the algorithm. However, it also puts higher demands on the real-time performance and data accuracy

of the algorithm.

In summary, using the dual population ant colony algorithm based on path contribution evaluation to solve the mathematical model of TSP, the specific steps are as follows:

Step 1: Parameter initialization;

Step 2: In the directed weighted graph, randomly select two cities as the starting points for populations A and B;

Step 3: For groups 1 and 2, calculate the state transition probability of ant a using Equations (19) and (23) respectively, and store the unselected cities in the set of candidate cities;

Step 4: Improve population 1 through elite ant strategy and improve population 2 through strengthened sub path evaluation mechanism. Store the optimal solutions and corresponding path lengths d_1 and d_2 for the two populations in the current loop separately;

Step 5: Update pheromones for two populations using Equations (21) and (24) respectively;

Step 6: By collecting port congestion index (number of waiting ships in anchorage, utilization rate of loading and unloading equipment), meteorological and sea condition data (wind speed, wave height), and fuel price fluctuation data, a dynamic factor matrix is formed. When the dynamic factor of a certain flight segment exceeds the preset threshold, the local pheromone reset is immediately triggered, and the pheromone is updated again using Equations (21) and (24);

Step 7: Using the evaluation criteria for path contribution, compare and analyze d_1 and d_2, and select the path with higher contribution as the global optimal path length d_{best};

Step 8: Analyze whether d_{best} continues to update. Continue updating, jump back to step 2; If no longer updated, output the optimal solution of TSP. d_{best} is the shortest path connecting all cities in the directed weighted graph.

The specific way in which the improved ant colony algorithm jumps out of local optima is as follows:

(1) Parallel search and independent evolution: The dual population mechanism introduces two independent ant populations for parallel search. This parallelism enables the algorithm to simultaneously explore multiple potential optimal solutions in the solution space, thereby increasing the likelihood of discovering the global optimal solution. Two populations adopt different strategies for evolution. Population 1 adopts the elite ant strategy, focusing on utilizing existing information to quickly converge; And population 2 adopts a strengthened sub path evaluation mechanism, focusing on exploring new solution spaces. This independence ensures that each population can explore deeply within its specific search area, reducing dependence on the search process of other populations.

(2) Elite Ant Strategy and Rapid Convergence: Population 1 adopts the Elite Ant Strategy, which updates pheromones by retaining and utilizing the optimal solution (Elite Ant) in each iteration. This strategy helps Population 1 to quickly converge to a local optimal solution, especially when there are clearly advantageous regions in the solution space. Meanwhile, the continuous exploration of population 2 provides an opportunity for the algorithm to escape from local optima.

(3) Enhanced sub path evaluation mechanism and global exploration: Population 2 adopts an enhanced sub path evaluation mechanism to update pheromones by calculating the contribution of sub paths. This mechanism enables population 2 to explore unknown regions in the solution space more deeply, thereby discovering new potential optimal solutions. When population 1 falls into a local optimum, population 2 uses its global exploration ability to discover a new path to the global optimum. This ability is achieved through a strengthened sub path evaluation mechanism, which can accurately assess the contribution of different sub paths and update pheromones accordingly, guiding ants to evolve towards a better path.

(4) Path contribution evaluation mechanism and global optimal selection: By comparing the path contribution of two populations, the algorithm selects the path with higher contribution as the global optimal solution. This mechanism ensures that the algorithm always evolves towards the global optimal solution during the convergence process, rather than staying at a local optimal solution.

(5) Real time data-driven pheromone reset mechanism: The algorithm also introduces a real-time data-driven pheromone reset mechanism. When real-time data (such as port congestion index, meteorological and sea condition data, fuel price fluctuation data, etc.) undergoes significant changes, this mechanism can adjust the distribution of pheromones in a timely manner, guiding ants to explore new paths. This dynamic adaptability enables the algorithm to quickly respond to changes in the external environment, thereby avoiding getting stuck in local optima caused by environmental changes.

2.6 Optimization of shipping logistics routes

Using the improved ant colony algorithm mentioned above, solve the objective function in Equation (18) to obtain the optimal shipping logistics path. The solving process is as follows:

Step 1: Parameter initialization.

Initialize various parameters in ant colony algorithm, including ant count, initial pheromone value, pheromone residual factor, pheromone volatilization factor, heuristic factor, elite ant count, amplification factor of path factor, maximum iteration times, etc.

Step 2: Build a directed weighted graph $G(V, D)$.

Construct a directed weighting graph based on the actual situation of the shipping logistics network. Using

nodes in the graph to represent key locations of logistics ports, edges to represent transportation paths, and the weights of edges to be quantified based on factors such as distance, time, and cost. Due to the focus on factors such as distance, time, cost, and carbon emissions in this study, these factors were comprehensively considered and used as weights for the edges.

Step 3: Build a solution space.

Based on the characteristics of the shipping logistics network, considering factors such as the location of logistics ports, transportation distance, transportation time, transportation expenses, carbon emissions, and path reliability, a solution space is constructed that includes all possible paths. This solution space will serve as the search range for ant colony algorithm.

Step 4: Select the starting point.

In the directed weighted graph $G(V,D)$, arbitrarily select two cities (logistics ports) as the starting points for population 1 and population 2.

Step 5: Calculate the state transition probability of ants.

For each ant in population 1 and population 2, calculate the state transition probability $q_{ij}^{a}(t)$ using Equations (19) and (23) based on the current port and the set of candidate ports.

Step 6: Calculate the objective function value.

For each path generated by an ant in the current iteration, calculate its corresponding objective function value F according to Equation (18).

Step 7: Update the pheromone.

After each iteration, update the pheromone concentration on the path based on the ant's movement path and objective function value [21]. For population 1, according to Equation (21), the elite ant strategy is adopted to improve the updating method of pheromones, resulting in $p_{ij}(t+1)$; For population 2, according to Equation (24), a strengthened sub path evaluation mechanism is used to update the pheromones, obtaining the updated values $\Delta p_{ij}(t)$ and $\Delta\delta_{ij}(t)$ of positive and negative feedback pheromones, in order to ensure accurate division of the contribution of different sub paths in the solving process.

Step 8: Calculate the path contribution degree.

After each iteration, calculate the path contribution C of the two paths obtained from the two populations according to Equation (25), select the better path, and further adjust the pheromone concentration accordingly.

Step 9: Update the global best path.

Compare the contribution of two paths obtained from two populations, select the path with higher contribution as the global best path, and update the length of the global best path.

Step 10: Iterative search.

Repeat steps 3 to 7 for multiple iterations of the search. In each iteration, ants are guided to find a better path in the directed weighted graph based on the current pheromone concentration and state transition probability.

Step 11: Determine the termination condition.

Determine whether the global optimal path length d_{best} has been updated. If no longer updated, the shortest path connecting all cities within $G(V,E)$; Otherwise, return to step 3 to continue iterating.

Step 12: Output the optimal path.

After the algorithm stops, output the optimal shipping logistics path. The path should meet the objective function requirements of minimizing logistics transportation expenses, minimizing logistics transportation time, minimizing carbon emissions, and maximizing path reliability. At the same time, the path can be further optimized and adjusted according to actual needs.

3 Simulation experiment and result analysis

To verify the feasibility of the optimization method for shipping logistics paths based on evolutionary ant colony algorithm designed above, the following simulation experiments are designed.

3.1 Simulation experiment design

Before the experiment, the following designs were developed for two populations:

Population 1 uses the Elite Ant System for pheromone updates. In this population, the number of ants is 100, the proportion of elite ants is 0.3, the pheromone importance factor is 2.5, the heuristic function importance factor is 5.0, the pheromone intensity is 100, the proportion of elite ants is 0.3, and the pheromone volatilization rate is 0.25.

The sub path length threshold for population 2 is 5, the sub path contribution weight is 0.75, the negative feedback heuristic factor is 1.2, and all other parameters are the same as population 1.

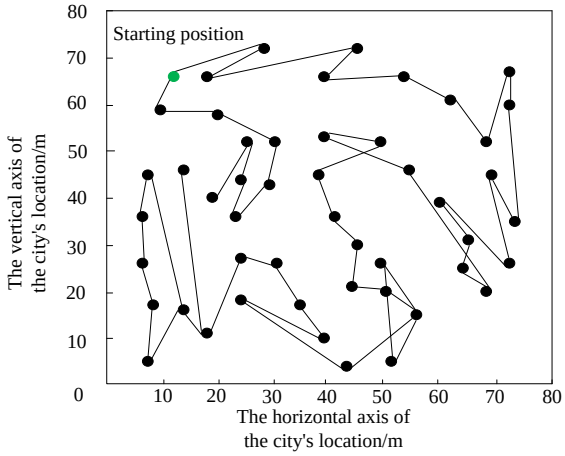
The parameter difference between the two populations is not randomly set, but a targeted design to meet their division of labor goals: Population 1 achieves rapid convergence through high heuristic weights, elite retention, and low pheromone volatilization; Population 2 achieves global search through sub path evaluation, negative feedback mechanism, and local exploration weights. Therefore, in addition to sampling the same parameter settings as population 1, additional sub path length thresholds and sub path contribution weights were set for population 2.

Assuming there are 50 cities (logistics ports), construct a directed weighted graph based on their geographical locations and transportation conditions.

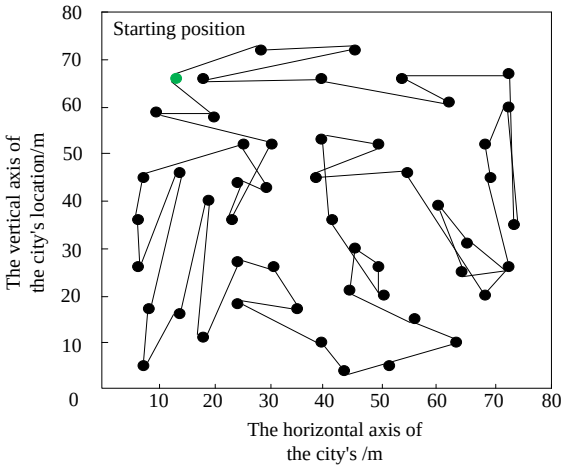
Using the method of this paper for solution analysis, the solution path analysis results are shown in Fig. 2.

From the analysis of Fig. 2, it can be seen that although both population 1 and population 2 can be used to solve the path scheme, after using the dual population ant colony algorithm, the path does not overlap or repeat. The solution path lengths for different algorithms are shown in Table I.

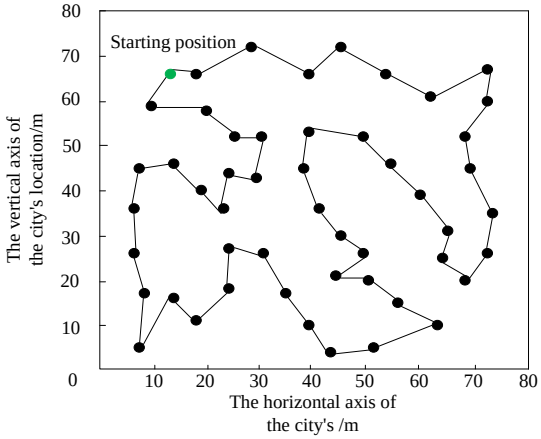
Based on the comprehensive analysis of Fig. 2 and Table I, it can be seen that the path length obtained by using only Population 1 or Population 2 is relatively long, while the path length obtained by using the dual population ant colony algorithm is significantly shorter, indicating the effectiveness of the improved ant colony algorithm proposed in this paper.



(a) Analysis results of solving population 1



(b) Analysis results of solving population 2



(c) Analysis and solution results of dual population ant colony algorithm

Figure 2: The solution result of the algorithm

Table 1: The path length for solving different algorithms

Algorithm	The length of the path/$\times 10^6$ nautical mile	Pheromone level
Population 1	468	85.0
Population 2	523	70.2
Dual population ant colony algorithm	407	95.3

On this basis, the convergence curves of using only population 1, only population 2, and the dual population ant colony algorithm are shown in Fig. 3. By comparing the optimal path cost during the iteration process of different algorithms, the efficiency of different algorithms is compared.

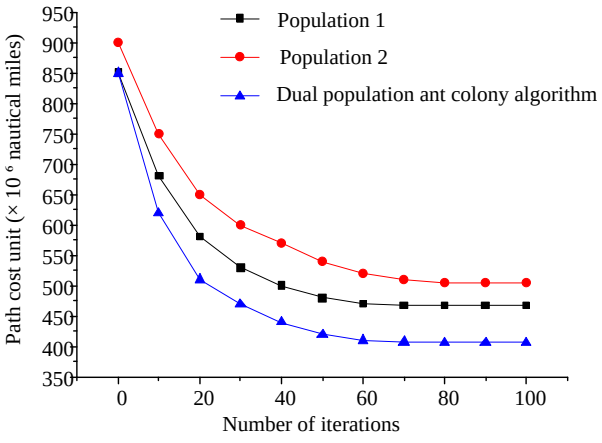


Figure 3: Comparison of optimal path costs in different algorithm iterations

Observing Fig. 3, it can be seen that the dual population algorithm has the advantage of parallel search, and its initial value is similar to that of population 1. Population 1 rapidly declines (first 30 iterations) but subsequently falls into local optima; Population 2 has a slow convergence rate, but continues to explore new solutions and has not reached the global optimum. The dual population algorithm combines fast convergence and continuous exploration to ultimately achieve the optimal value, and the convergence speed is significantly faster than that of a single population. After 50 iterations, the dual population algorithm is approaching the final solution, while there is still a significant gap between the individual populations. After 80 iterations, all algorithms tend to stabilize, and the dual population mechanism exhibits global optimality.

3.2 Comparative analysis of application effects

Design of simulation experiment environment

The experimental simulation starts from port A in China and ends at port R in the United States, passing through 16 dockable ports from B to Q. These ports are distributed in different countries and regions, representing key nodes in the shipping logistics path. Among them, C, H, M, and P are ports that must be stopped.

By optimizing the path, we aim to minimize transportation costs, minimize transportation time, minimize carbon emissions, and maximize path reliability.

Design of simulation experiment parameters

The ship and navigation parameters are set as follows: the ship type is a Cape of Good Hope container ship, the ship load is 150000 tons, the ship speed is 15 nautical miles per hour, and the daily navigation distance is 360 nautical miles.

The fuel consumption parameters are set as follows: the container ship used for simulation belongs to the Cape of Good Hope ship type, with an average load of 150000 tons. The daily fuel consumption during ship navigation is 48.5 tons/day, and the daily fuel consumption during berthing is 3.81 tons/day. Calculate $965 per ton of fuel based on the average fuel cost in 2024. In the simulation experiment, the ship is set to refuel at least once, with priority given to docking or anchoring for refueling.

The container and port operation parameters are set as follows: the maximum container capacity is 15000 TEU, and the ratio of 20 size to 40 size containers is 1:1. Due to the inconsistent operating cost standards of various ports, for the convenience of calculation, the simulation experiment is estimated using the average value. The individual operating cost for a 20-size container is $45, and for a 40-size container it is $65

The carbon emissions and environmental parameters are set as follows: the carbon tax rate is set at an estimated intermediate value of $42.75/ton, and the carbon conversion factor for ship navigation fuel consumption is determined to be 3.25, which is used to calculate carbon emissions.

The environmental parameters are set as follows: In order to simulate the impact of maritime weather on shipping logistics, the experiment introduced weather condition parameters. Assuming there are varying degrees of sea conditions during the experiment, such as light waves, medium waves, large waves, etc., each sea condition will have different impacts on the vessel's sailing speed, fuel consumption, and safety reliability. The specific parameter settings are as follows:

Light waves: reduced sailing speed by 5%, increased fuel consumption by 2%, and reduced path reliability by 1%.

Mid wave: reduced sailing speed by 10%, increased fuel consumption by 5%, and reduced path reliability by 3%.

Big waves: reduced sailing speed by 15%, increased fuel consumption by 10%, and reduced path reliability by 5%.

The weather conditions will be randomly assigned to different flight segments to simulate the changes in sea conditions that may be encountered during actual navigation.

The additional fee and other parameter settings are as follows: The simulation experiment simulation additional fee mainly includes security fees, text fees, etc. There is not much difference in charges among major ports, so the standard is set at $102 per port. When constructing a directed weighting graph, distance, time, and cost (including fuel costs, port operation fees, carbon emission reductions, etc.) are comprehensively considered as the weights of edges. In addition, port loading and unloading delays are common uncontrollable factors in shipping logistics. To simulate this situation, the experiment set loading and unloading delay parameters. The specific parameter settings are as follows:

Normal loading and unloading: Complete loading and unloading operations according to the scheduled time.

Minor delay: loading and unloading time extended by 10%.

Serious delay: loading and unloading time extended by 30%.

Loading and unloading delays will be randomly assigned to different ports to simulate the delay situations that may be encountered in actual port operations. The setting of these parameters will help to more accurately evaluate the feasibility and reliability of different paths in actual shipping.

Path description

To avoid the singularity of experimental results, method of reference [4] and method of reference [5] were compared and validated against method of this paper in the same environment. Below are three methods for

generating shipping logistics paths.

(A) The shipping logistics path generated by the method of this paper is as follows:

Path description:

Port A (starting point) → Port B (short stay) → Port C (stop, supply) → Port D (no stay) → Port E (short stay) → Port F (no stay) → Port G (no stay) → Port H (stop, supply) → Port I (short stay) → Port J (no stay) → Port K (short stay for equipment inspection) → Port L (no stay) → Port M (stop, supply) → Port N (no stay) → Port O (short stay) → Port P (stop, supply) → Port Q (no stay) → Port R (end point)

The duration of stay at each port is shown in Table II.

Table 2: Duration of stay at each port (method of this paper)

Port	Duration of stay
C	The vessel stays for 24 hours for cargo loading, unloading, and replenishment.
H	The vessel stays for 48 hours for deep replenishment and equipment maintenance.
M	The vessel stays for 36 hours for cargo transfer and necessary inspections.
P	The ship stays for 24 hours for final replenishment and cargo sorting.
Other ports	The average stay time does not exceed 12 hours.

Supply situation: Fuel, fresh water, and food supplies are provided at the four stopover ports of C, H, M, and P. Partial supply will be provided as needed at some temporary ports of stay (such as B, E, I, K, O).

(B) The shipping logistics path generated by method of reference [4] is as follows:

Path description:

Port A (starting point) → Port C (stopover, supply) → Port D (short stay) → Port E (no stay) → Port F (short stay) → Port G (no stay) → Port H (stopover, supply) → Port I (short stay) → Port J (stopover, unplanned supply) → Port K (no stay) → Port L (short stay) → Port M (stopover, supply) → Port N (no stay) → Port P (stopover, supply) → Port Q (no stay) → Port R (end point).

The duration of stay at each port is shown in Table III.

Table 3: Duration of stay at each port (method of reference [4])

Port	Duration of stay
C	The ship stays for 24 hours.
H	The ship stays for 36 hours.
J	The vessel is not scheduled to stay for 24 hours for replenishment (due to the inability to dock at the original planned port).
M	The ship stays for 24 hours.
P	The ship stays for 36 hours.
Other ports	The average stay time does not exceed 12 hours.

Supply situation: Supply will be conducted at ports C, H, M, and P. Due to unplanned stops at J port, supply costs have increased.

(C) The shipping logistics path generated by method of reference [5] is as follows:

Path description:

Port A (starting point) → Port B (short stop) → Port C (stop, supply) → Port D (no stop) → Port E (short stop) → Port F (stop, non essential supply) → Port G (no stop) → Port H (stop, supply) → Port I (no stop) → Port J (short stop) → Port K (stop, supply) → Port L (no stop) → Port M (stop, supply) → Port N (short stop) → Port O (no stop) → Port P (stop, supply) → Port R (end point).

The duration of stay at each port is shown in Table IV.

Table 4: Duration of stay at each port (method of reference [5])

Port	Duration of stay
C	The ship stays for 36 hours.
F	The vessel is not required to stay for 12 hours for replenishment.
H	The ship stays for 24 hours.
K	The ship stays for 24 hours for equipment inspection.
M	The ship stays for 36 hours.
P	The ship stays for 24 hours.
Other ports	The average stay time does not exceed 12 hours.

Supply situation: Supply will be conducted at ports C, H, K, M, and P. The non essential stay at Port F increases the cost and time of supply.

The difference between the necessity and non necessity of different methods for port stay in the above results is due to their optimization objectives and dynamic response capabilities. The necessary stops (C, H, M, P ports) are set as mandatory constraints by the experiment, and all methods must be followed. Non essential stops may arise due to different algorithm designs: this method dynamically avoids inefficient ports and reduces redundant stops through multi-objective optimization (minimizing expenses, time, carbon emissions, and maximizing reliability) and real-time pheromone reset mechanism; Reference [4] focuses on fuel cost as the core objective and relies on static models. When the original planned port is unavailable, it is forced to detour (such as J port) and increase non essential stops; Reference [5] focuses on the shortest distance and safety, ignoring the overall cost, resulting in suboptimal path selection (such as non essential supplies at F port). The method of this paper significantly reduces unnecessary stops due to its ability to balance multiple objectives and dynamic adjustments, while the comparative method performs poorly in terms of transportation efficiency, cost, and environmental friendliness due to its single objective or rigid model, making it difficult to balance complex constraints.

Comparative analysis

According to the above generated path optimization scheme, the total transportation time, total transportation expenses, and total carbon emissions after applying different methods were compared, and the results are shown in Table V.

Table 5: Comparison table of results

	Total transportation time/day	Total transportation expenses/USD	Total carbon emissions/ton
Method of this paper	41	1250000	11800
Method of reference [4]	50	1420000	13500
Method of reference [5]	48	1380000	13200

By comparing the paths generated by the three methods, it can be seen that the total transportation time of this method is 41 days throughout the entire experimental cycle, saving 9 days and 7 days respectively compared to the two existing methods; The total transportation expenses of this method is 1250000 USD, saving 170000 USD and 130000 USD respectively compared to the two existing methods; The total carbon emissions of this method are 11800 tons, saving 1700 tons and 1400 tons respectively compared to the two existing methods. It can be seen that Method of this paper outperforms Method of Reference [4] and Method of Reference [5] in terms of total transportation time, total transportation expenses, and total carbon emissions. The method of this paper effectively improves the global optimization capability of path optimization by introducing a dual population mechanism and a path contribution evaluation mechanism. Although the method of reference [4] considers fuel supply strategies, its path selection is relatively conservative, resulting in higher total transportation time and expenses. The method of reference [5] is designed for peak periods of maritime traffic, but its path optimization objective is relatively single and fails to comprehensively consider factors such as transportation costs and carbon emissions, resulting in slightly inferior performance compared to the method of this paper.

On this basis, 30 independent experiments were conducted on different methods in the same simulation environment, and the following indicators were recorded for each experiment: total transportation time, total transportation expenses, and total carbon emissions. Calculate the mean and 95% confidence interval (CI) of each method indicator, as shown in Table VI.

Table 6: Confidence interval validation

	Total transportation time (Mean ± CI)/day	Total transportation expenses (Mean ± CI)/USD	Total carbon emissions (Mean ± CI)/ton
Method of this paper	41±1.2	1,250,000±25,000	11,800±300
Method of reference [4]	50±2.1	1,420,000±35,000	13,500±450
Method of reference [5]	48±1.8	1,380,000±30,000	13,200±400

Compare the mean difference between the method proposed in this study and two comparison methods, and verify its significance (with a significance level of $\alpha = 0.05$).

Total transportation time: $t=6.34$, $p<0.001$ (significantly better than both methods)

Total transportation expenses: $t=5.89$, $p<0.001$ (significantly better than both methods)

Total carbon emissions: $t=4.76$, $p<0.001$ (significantly better than both methods)

Statistical tests show that the improvement of the method of this study in transportation time, cost, and carbon emissions is statistically significant ($p<0.001$). The confidence intervals are non overlapping, further supporting the stable advantage of our method on multiple objectives.

4 Discussion

An in-depth analysis of the experimental results is conducted, and the advantages and disadvantages of the two traditional methods are shown in Table VII.

Table 7: The advantages and disadvantages of traditional methods

Method		Method of reference [4]	Method of reference [5]
Advantage		Consider fuel supply strategy: Conduct in-depth research on the characteristics of grain shipping and the changes in fuel supply port oil prices over time and place, explore the relationship between fuel supply port oil prices and supply quantities, average oil prices and port service fees, and make fuel supply strategies more in line with actual operational situations, which can help reduce fuel costs. Constructing a path model that integrates integer programming: selecting fuel supply port selection, fuel supply quantity, and ship navigation route as key decision-making elements, the constructed model can comprehensively consider multiple factors and provide more comprehensive decision-making basis for path optimization.	Setting a comprehensive objective function: aimed at minimizing the length of the voyage, improving navigation safety, and ensuring smooth navigation, it helps to find relatively good ship diversion paths in complex maritime traffic peak environments, ensuring safe and smooth shipping operations. Multiple algorithms are used to solve the problem: the grid method is used to reproduce the ship operation status during peak maritime traffic hours, the improved genetic algorithm searches for the optimal ship diversion path planning scheme on a global scale, and the nonlinear programming technique further solves the local optimal solution, improving the accuracy and effectiveness of the solution.
Disadvantage	Disadvantages	Conservative route selection: In order to reduce fuel	Single objective function: The objective function mainly

	in total transportation time	costs or port service fees, longer routes or more transit ports may be chosen, which increases the total transportation time.	focuses on the length of the voyage, navigation safety, and stability, without fully considering time efficiency, which may result in longer transportation time.
		Lack of consideration for time efficiency: Insufficient consideration of time efficiency has resulted in an inability to effectively balance transportation time and cost during path optimization.	Unbalanced transportation time and other factors: During the optimization process, transportation time was not well balanced with other factors that may affect transportation efficiency.
	Disadvantages in total transportation expenses	Incomplete consideration of cost factors: Although fuel supply strategies have been considered, excessive emphasis on fuel and port service fees in route selection may overlook other potential cost factors, resulting in suboptimal total transportation expenses.	Not considering cost factors comprehensively: The objective function is relatively single and does not take into account other cost factors such as fuel costs and port charges, which may result in suboptimal total transportation expenses.
		Failure to conduct comprehensive cost optimization: Failure to conduct comprehensive optimization from the perspective of overall transportation costs may result in higher costs in certain situations.	Lack of cost optimization mechanism: There is no effective cost optimization mechanism established in the method, which cannot reduce transportation costs as a whole.
	Disadvantages in total carbon emissions	The impact of route selection: longer routes will increase the distance traveled by ships, thereby increasing fuel consumption and leading to an increase in total carbon emissions.	The impact of route length: Although minimizing the length of the route can reduce certain costs, longer routes will increase the distance traveled by ships, increase fuel consumption, and lead to an increase in total carbon emissions.
		Lack of carbon emission considerations: The methods did not specifically optimize for carbon emissions, and carbon emission factors were not fully considered in path selection and transportation decisions.	Lack of carbon emission optimization objectives: Carbon emissions were not included as one of the optimization objectives in the method, and carbon emission factors were not fully considered in path planning.

The method of this paper introduces a dual population mechanism to improve the conventional ant colony algorithm and applies it to optimize shipping logistics paths, effectively solving multiple key problems of traditional methods. On the one hand, conventional ant colony algorithms are prone to getting stuck in local optima in complex shipping networks and are sensitive to initial parameters. The improved algorithm enhances global search capability and parameter robustness through a dual population mechanism, reducing the risk of getting stuck in local optima and improving the adaptability and stability of the algorithm. On the other hand, traditional algorithms usually focus on single objective optimization, which makes it difficult to balance multiple objectives such as logistics transportation expenses, transportation time, carbon emissions, and path reliability. The improved algorithm sets up a multi-objective function to achieve trade-offs and coordination between various objectives, and can generate multiple Pareto optimal solutions, providing decision-makers with more comprehensive path selection and meeting the needs of different scenarios. In addition, the shipping logistics network structure is complex, and the improved ant colony algorithm abstracts it as a set of nodes and edges in graph theory, simplifying the network structure and providing a clear search framework for the algorithm, improving its efficiency and feasibility.

However, although the method of this paper has shown significant advantages in transportation time, cost, and carbon emissions, there are still significant bottlenecks in its computational complexity. As the number of port nodes increases, the algorithm needs to maintain pheromone matrices and parallel search mechanisms for two independent populations, resulting in exponential growth in memory usage and computational complexity. Especially when dealing with multi-objective optimization involving time window constraints, port capacity limitations, etc., the algorithm requires a large number of iterations to balance the weight relationships between different objective functions, which reduces the feasibility of real-time dynamic path adjustment. In addition, algorithms are sensitive to parameter settings, such as pheromone volatilization factor, elite ant ratio, and other key parameters that need to be finely tuned for different shipping scenarios, otherwise it may lead to premature convergence of population 1 or low exploration efficiency of population 2. The setting of threshold parameters in the path contribution evaluation mechanism also lacks universal standards, which may lead to path evaluation bias in extreme weather or sudden port closures and other abnormal situations.

Another potential issue is that the improved ant colony algorithm based on the method of this paper has a strong dependence on the quality of input data. When there is noise or update delay in basic data such as distance between ports and fuel prices, the "optimal path" generated by the algorithm may have significant deviations in actual execution. For example, in the experiment, it is assumed that the operating costs of each port are estimated using the average value, but in reality, the difference in charging standards between different ports may be as high as 30%, which will directly affect the accuracy of achieving the goal of minimizing transportation costs. In addition, although the dual population mechanism improves the global search capability, the pheromone synchronization mechanism between the two populations may cause unstable convergence speed of the algorithm, especially when dealing with ultra large scale networks with more than 100 nodes, the computation time may exceed the time window limit of shipping scheduling. These limitations indicate that this method is more suitable as an offline planning

tool, leveraging its advantages in scenarios with high data quality and sufficient computing resources.

In addition, in the real-time data fluctuation environment, the method based on evolutionary ant colony algorithm adopted in this study demonstrates certain dynamic adaptability by introducing real-time data-driven pheromone reset mechanism and path contribution evaluation mechanism. When real-time data (such as port congestion index, meteorological and sea condition data, fuel price fluctuation data, etc.) undergoes significant changes, these pheromones reset mechanisms can adjust the distribution of pheromones in a timely manner, guide ants to explore new paths, and to some extent cope with the uncertainty brought by data fluctuations. However, excessive or frequent fluctuations in real-time data may also have a negative impact on the performance of this method. For example, sudden changes in extreme weather conditions may lead to a significant increase in port loading and unloading delays or forced route adjustments, requiring algorithms to respond quickly and re plan their routes. In this case, the convergence speed and solution quality of the algorithm may be affected to some extent. However, overall, this method exhibits certain dynamic adaptability and robustness under real-time data fluctuations, but further optimization and testing are still needed to improve its stability and performance under extreme or frequent fluctuations. In addition, continuous monitoring and adjustment of algorithm parameters can also help improve its performance under real-time data fluctuations.

5 Conclusion

This study used an improved ant colony algorithm to solve the optimization problem for shipping logistics paths. The method designed in this study not only considers the advantages of traditional ant colony algorithm, but also improves the search efficiency and global optimization ability of the algorithm by introducing dual population mechanism and path contribution evaluation mechanism. The final optimal shipping logistics path can effectively reduce transportation costs, shorten transportation time, reduce carbon emissions, and improve the reliability of the path in practical applications, providing a scientific and efficient path planning solution for shipping logistics enterprises.

In the next stage of research, it may be considered to integrate the results of this study with AIS data. AIS data can provide real-time information on route congestion, weather, and sea conditions. However, in integration, AIS data streams require low latency processing, which requires algorithms to support incremental updates rather than full iteration. Based on your experience, it is necessary to further optimize the algorithm iteration frequency (such as reducing it from hourly level to minute level) and enhance noise robustness.

References

[1] Y. F. Huang, M. H. Wei, Y. Wang, X. Y. Guo, and M. Z. Huang, "Current situation and trend of green shipping logistics in China under the carbon peaking and carbon neutrality," *Journal of Dalian Maritime University*, vol. 49, no. 1, pp. 1-16, 2023. https://doi.org/10.16411/j.cnki.issn1006-7736.2023.01.001

[2] B. B. Liao, "Research on the path to carbon neutrality of China's shipping industry under the "double carbon" target," *Pacific Journal*, vol. 30, no. 12, pp. 89-100, 2022. http://www.pacificjournal.com.cn/CN/news/news522.shtml

[3] X. E. Bai, M. Q. Fang, X. F. Xu, Y. J. Xiao, and Y. M. Wu, "Ship logistics path trajectory repair based on improved LSTM," *Packaging Engineering*, vol. 44, no. 17, pp. 152-159, 2023. http://dx.doi.org/10.19554/j.cnki.1001-3563.2023.17.018

[4] J. Wang, H. Yu, and J. Gao, "Optimization of grain transport fleet path considering fuel supply strategy," *Science Technology and Engineering*, vol. 24, no. 27, pp. 11886-11890, 2024. http://dx.doi.org/10.12404/j.issn.1671-1815.2306828

[5] X. H. Guo and C. L. Zhao, "Simulation analysis of ship guidance path planning on maritime traffic peak routes," *Ship Science and Technology*, vol. 45, no. 22, pp. 206-209, 2023. http://dx.doi.org/10.3404/j.issn.1672-7649.2023.22.040

[6] Y. Zhao, X. Y. Gao, G. F. Kui, and Y. Zhang, "Green scheduling optimization of shuttle tanker route considering carbon emissions," *Chemical Industry and Engineering*, vol. 39, no. 2, pp. 23-31, 2022. http://dx.doi.org/10.13353/j.issn.1004.9533.20210396

[7] X. He, "Improved ant colony optimization for heterogeneous interactivedual-population solving TSP," *Computer Applications and Software*, vol. 40, no. 11, pp. 288-294, 2023. https://www.cnki.com.cn/Article/CJFDTotal-JYRJ202311043.htm

[8] L. Sun, "Optimal transportation route planning for ships based on swarm intelligence optimization algorithm," *Ship Science and Technology*, vol. 46, no. 12, pp. 166-169, 2024. http://dx.doi.org/10.3404/j.issn.1672-7649.2024.12.029

[9] X. E. Bai, B. H. Li, X. F. Xu, and Y. J. Xiao, "Scheduling optimization of shipping logistics port based on AIS data," *Packaging Engineering*, vol. 44, no. 5, pp. 211-221, 2023. http://dx.doi.org/10.19554/j.cnki.1001-3563.2023.05.027

[10] L. Yu, Z. Sun, Z. Zhang, Z. L. Chen, and L. J. Jin, "Optimization of truck route in container yard against theindeterminacy of driving time," *Navigation of China*, vol. 46, no. 2, pp. 90-98+105, 2023.

https://zghh.cinnet.cn/article/doi/10.3969/j.issn.1000-4653.2023.02.013
[11] Z. G. Wang, C. X. Cao, Y. E. Lv, and K. Shen, "A uncertain planning model of emergency shipping route," *Fuzzy Systems and Mathematics*, vol. 37, no. 6, pp. 158-164, 2023. https://mohu.org/zh/html/1004.html
[12] S. S. Wu, M. X. Liang, J. K. Guo, and Z. Y. Xiong, "Identification and measurement of seaward accessibility of port based on shipping networks," *Geographical Research*, vol. 43, no. 4, pp. 824-842, 2024. http://dx.doi.org/10.11821/dlyj020230578
[13] F. Q. Meng, X. X. Sun, J. S. Zhu, B. Mei, and P. J. Zheng, "Ship path planning based on bidirectional A*-APF algorithm," *Journal of Dalian Ocean University*, vol. 39, no. 3, pp. 506-515, 2024. https://doi.org/10.16535/j.cnki.dlhyxb.2023-293
[14] H. L. Yang, Y. H. Wu, and Y. L. Sun, "Robust optimization model and algorithm for maritime inventory routing problem of multi fuel products," *Journal of Transportation Systems Engineering and Information Technology*, vol. 22, no. 3, pp. 238-246, 2022. https://doi.org/10.16097/j.cnki.1009-6744.2022.03.027
[15] J. Jiang, Y. P. Yuan, P. Deng, and X. N. Fu, "Effectiveness evaluation of Chongqing shipping center project with cloud model and VIKOR method," *Navigation of China*, vol. 46, no. 3, pp. 91-97+104, 2023. https://zghh.cinnet.cn/article/doi/10.3969/j.issn.1000-4653.2023.03.013
[16] X. F. Guo, J. S. Zhu, and W. W. Qiu, "Modeling and calculation of AIS-based port congestion index," *Navigation of China*, vol. 46, no. 4, pp. 154-162, 2023. https://zghh.cinnet.cn/article/doi/10.3969/j.issn.1000-4653.2023.04.021
[17] G. H. Wang, "Research analysis of small-scale TSP problem based on genetic algorithm," *Logistics Engineering and Management*, vol. 44, no. 3, pp. 111-114+29, 2022. http://dx.doi.org/10.3969/j.issn.1674-4993.2022.03.031
[18] Y. L. Wang and B. Wang, "Hybrid GA-ACO algorithm for optimizing transportation path of port container cargo," *Informatica*, vol. 48, no. 20, pp. 73-80, 2024. https://doi.org/10.31449/inf.v48i20.6265
[19] Y. J. Wang, X. M. You, and S. Liu, "Cooperative ant colony algorithm combining evaluation reward and punishment mechanism and neighborhood dynamic degradation," *Journal of System Simulation*, vol. 36, no. 6, pp. 1475-1492, 2024. https://doi.org/10.16182/j.issn1004731x.joss.23-0524
[20] W. S. Sheng, A. P. Xu, and L. J. Xu, "Simulation of traveling salesman path planning based on ant colony algorithm and genetic algorithm," *Computer Simulation*, vol. 39, no. 12, pp. 398-402+412, 2022. https://www.cnki.com.cn/Article/CJFDTotal-JSJZ202212073.htm
[21] D. Wu, X. L. Han, S. J. Shi, X. J. Ji, J. F. Zheng, and B. L. Liu, "Location-inventory-routing optimization of maritime logistics network in remote islands under demand uncertainty," *Journal of Transportation Systems Engineering and Information Technology*, vol. 24, no. 5, pp. 268-282, 2024. https://doi.org/10.16097/j.cnki.1009-6744.2024.05.025

Appendix

Pseudocode for the iterative process of the evolutionary ant colony algorithm

Input: Directed weighted graph G, parameters
Output: Global optimal path P_global

```
Initialize:
   - Pheromone matrices for Population 1 and 2
   - Heuristic matrix based on edge weights (distance, cost, etc.)
   - Global best path P_global ← null
   - Dynamic data monitoring module (port congestion, weather, fuel prices)

For iter in 1 to max_iter:
   // Parallel search for two populations
   For each population k in {1, 2}:
      For each ant m in population k:
         P_m ← ConstructPath
         Calculate objective function using Eq. (18)
      End For

      // Update pheromones for each population
      If k == 1:
         Update using Elite Strategy (Eq. 21-22):
        Else:
         Update using Enhanced Subpath Evaluation (Eq. 24):
      End If

      // Store local best paths
      P_k_best ← SelectBestPath(population k)
   End For

   // Path Contribution Evaluation (Eq. 25)
   For each path in {P_1_best, P_2_best}:
      Calculate contribution_score
   End For

   // Update global best path
   P_candidate ← argmax(contribution_score(P_1_best), contribution_score(P_2_best))
   If F(P_candidate) < F(P_global) or P_global == null:
      P_global ← P_candidate
   End If

   // Real-time Data-Driven Pheromone Reset
```

```
    For each edge (i, j) in G:
        If dynamic_factor(i, j) > threshold (e.g., congestion,
weather risk):
            Reset to initial values
        End If
    End For

    // Termination Check
    If no improvement for 10 consecutive iterations:
        Break
    End If
End For

Return P_global

Function ConstructPath:
    Initialize path P with random start node
    While not all nodes visited:
        Next node selection using state transition probability:

        Add next node to P
    End While
    Return P
```

The Big Data Fusion Algorithm for Online Evaluation of Transformer Metering Error

Xu Chen*, Chao Zhang, Haomiao Zhang, Zhiqiang Cheng, Xinrui Zhang and Yinzhe Xu
State Grid Ningxia Marketing Service Center (State Grid Ningxia Metrology Center) Ningxia Yinchuan 750000, China
E-mail: chenxu3809@126.com, zhangchao@yxzx.nx.sgcc.com.cn, zhanghaomiao@yxzx.nx.sgcc.com.cn, chengzhiqiang@yxzx.nx.sgcc.com.cn, zhangxinrui@yxzx.nx.sgcc.com.cn, xuyinzhe@yxzx.nx.sgcc.com.cn
*Corresponding author

Keywords: mutual inductor, measurement error, big data fusion, principal component analysis (PCA), support vector machine regression (SVR), vibration factor, real-time evaluation

Received: February 24, 2025

This paper proposes an online evaluation algorithm for transformer metering error based on big data fusion, which combines principal component analysis (PCA) and support vector machine regression (SVR) technology. The algorithm first reduces the dimension of the actual value of the transformer secondary signal through PCA and extracts the main components to establish a principal component model. Then, SVR is used to perform regression analysis on the reduced-dimensional data to predict the actual metering error of the transformer. In addition, this paper also considers the influence of mechanical vibration on the transformer metering error. By introducing the simulation of vibration factors in the simulation platform, the significant influence of vibration on the error is verified. By constructing a simulation model of a 500kV power system and generating relevant error data sets for algorithm training and optimization, the simulation results show that the error evaluation model based on PCA-SVR can achieve high-precision error prediction under different load and grid fluctuation conditions, especially under high dynamic load and grid change conditions. Compared with traditional methods, the proposed algorithm has significantly improved computational efficiency and has real-time evaluation capabilities, which can meet the needs of high-frequency data analysis in smart grids. In addition, the anti-interference ability and dynamic adaptability of the model have also been verified.

Povzetek: Razvit je izvirni algoritem za spletno oceno napake merjenja transformatorjev, ki s PCA-SVR in podatkovno fuzijo omogoča natančno, robustno in časovno odzivno spremljanje v pametnih omrežjih.

1 Introduction

Accurate metering is the foundation of power exchange, fair accounting, load scheduling and energy efficiency management. As an essential part of power measurement equipment, transformers have been widely applied to measuring voltage, current, and power. The precision of the transformer directly influences the operation efficiency and economic performance of the power system, which influences the fairness and transparency of the electricity market. But, because of the complexity of the power system and all kinds of outside environments, it is very common for the transformer to make measurement errors. The difference in ratio and angle between the voltage transformer and the current transformer and the error of the secondary circuit and the power meter will result in the measurement error of the transformer [1]. These errors influence the precision of measurement results and cause hidden troubles in the power system's operation. Therefore, it is essential to study how to evaluate and monitor the measurement error of transformers efficiently and in real time.

The causes and influencing factors of transformer measurement errors are multifaceted, usually including quality problems of the equipment itself, changes in environmental factors, load fluctuations, temperature changes, etc [2]. In addition, mechanical vibration can also significantly affect the performance and measurement accuracy of the transformer. Vibrations may be caused by the operation of nearby equipment, power grid fluctuations, or even external interference (such as earthquakes). These vibrations may change the mechanical structure of the transformer, thereby affecting its electrical characteristics, and ultimately leading to an increase in measurement errors. The vibration factor is a key element of this study. Equipment aging and long-term operation can also cause metering drift or failure of the transformer, especially in high-voltage and high-current power systems, where the errors of the transformer are more complex and difficult to predict. Traditional error detection methods mainly rely on offline detection and periodic calibration [3]. Although this method can detect error problems to a certain extent, it has problems such as poor real-time performance, inability to capture sudden errors in time during periodic inspections, and lack of real-time

feedback on the operating status of the equipment. At the same time, traditional methods often only analyze some error sources separately, lack comprehensive considerations, and limit the accuracy of error assessment. Therefore, when evaluating and monitoring the measurement errors of transformers, it is crucial to consider the influence of multiple factors such as vibration.

In the past few decades, with the increasing complexity and intelligence of power systems, the accuracy of power metering has become more critical, especially in the power market, where accurate metering data not only involves fair transactions between users and power suppliers but also involves the dispatching of power grid loads and the optimization of energy efficiency management [4]. Therefore, improving the accuracy of evaluating transformer metering errors in power systems through technical means has become a widely concerned issue in academia and industry.

In recent years, with the rapid development of extensive data, machine learning, artificial intelligence, and so on, more and more researchers have started to use them to evaluate and monitor the measurement errors of power systems [5]. These new techniques offer a new way to assess power measurement errors, especially significant data techniques, which make it possible to collect, process, and analyses data in power systems. Machine learning algorithms, especially Support Vector Machine Regression (SVR), Neural Networks and Decision Trees, have been widely used to evaluate measurement errors in power systems.

At present, some studies have tried to use machine learning methods to evaluate electric energy metering errors. A study proposed an electric energy metering error prediction method based on neural networks, which achieved specific results by real-time acquisition of current and voltage signals and prediction of electric energy metering errors using neural network models [6]. However, due to the nonlinear characteristics and multi-source nature of metering errors in power systems, a single machine learning method often has difficulty handling complex error patterns. Therefore, some studies have begun exploring combining multiple machine learning algorithms to build a more robust and accurate error assessment model. Some scholars have proposed an energy metering error assessment method based on PCA-SVR, which uses principal component analysis (PCA) for dimensionality reduction and combines SVR regression for error prediction. The experimental results show that this method has higher accuracy and robustness than traditional methods.

Although these studies have achieved specific results, they still face some challenges. First, the source of transformer errors in power systems is very complex. In addition to grid fluctuations and load changes, it is also affected by the external environment (such as temperature, humidity, etc.). Therefore, multiple factors need to be considered in the model. Secondly, with the promotion of smart grids and the generation of large-scale data, how to efficiently process these massive data and extract valuable features from them is still a problem to be solved [7]. Finally, traditional error assessment methods rely on simplified assumptions and linear models, which cannot fully capture the nonlinear relationship of transformer errors in power systems, so there is still much room for improvement.

A new online evaluation algorithm based on extensive data fusion has been put forward, combining PCA and SVR. In particular, the PCA method is used to reduce the size of the transformer's output signal to extract the key characteristics and reduce the redundancy and noise of data [8]. The SVR algorithm is used to regress and model the extracted features to predict the measuring error of the transformer accurately. The method can increase the precision of the error assessment and deal with changes in the environment and the variation of the load. Moreover, this thesis also integrates the data of a real-time collection of power systems into online error assessment, so it has good real-time adaptability.

2 Abnormal state evaluation index

To evaluate the metering error state of the transformer, this paper introduces the abnormal state evaluation index [9]. Specifically, this paper detects abnormal changes in the metering state by calculating the Q statistic and its contribution rate. The Q statistic is defined as:

$$Q = \|\mathrm{X} - \mathrm{X}\mathrm{P}_k\mathrm{P}_k^T\|^2 \tag{1}$$

Among them, $\mathrm{X}\mathrm{P}_k\mathrm{P}_k^T$ is the projection of the data on the principal element subspace and $\|\cdot\|$ represents the Frobenius norm. The contribution rate of the Q statistic is defined as:

$$\text{Contribution Rate} = \frac{\|\mathrm{X} - \mathrm{X}\mathrm{P}_k\mathrm{P}_k^T\|^2}{\|\mathrm{X}\|^2} \tag{2}$$

By calculating the Q statistic and its contribution rate, this paper can detect abnormal changes in the metering state and locate the abnormality [10]. Specifically, this paper introduces the weighting factor $\mathrm{w} = [w_1, w_2, ..., w_m]$ to represent the importance of different variables in abnormal state evaluation. The weighted Q statistic is defined as:

$$Q_w = \|\mathrm{W}^{1/2}(\mathrm{X} - \mathrm{X}\mathrm{P}_k\mathrm{P}_k^T)\|^2 \tag{3}$$

Where $\mathrm{W} = \mathrm{diag}\,(\mathrm{w})$ is the weighting matrix and $\mathrm{W}^{1/2}$ is the square root of W. The contribution rate of the weighted Q statistic is defined as:

$$\text{Contribution Rate}_w = \frac{\|W^{1/2}(X - XP_k P_k^T)\|^2}{\|W^{1/2}X\|^2} \quad (4)$$

By calculating the weighted Q statistic and its contribution rate, this paper can more accurately detect abnormal changes in the metering state and locate abnormalities.

In addition to detecting abnormal changes in metering status through Q statistics and their contribution rates, this paper also considers the impact of mechanical vibration on transformer performance. By integrating vibration sensors in the monitoring system, vibration data can be collected and analyzed in real time. Vibration data can be used to detect mechanical anomalies that may affect the accuracy of the transformer. Specifically, the amplitude and frequency of vibration can be monitored, and any significant deviation from normal operating conditions can be marked as a potential source of error. The discussion of vibration factors' impact on error evaluation has been quantified by introducing artificial noise to the dataset to evaluate how the model handles perturbations. This comprehensive approach, combining electrical and mechanical data, enhances the ability to detect and diagnose abnormal conditions of transformers.

3 Synthesis of metering device errors

3.1 Multi-parameter dimensionality reduction model of synthetic errors

In electric energy metering devices, the synthesis of transformer metering errors is affected by various factors, including voltage, current, secondary load, temperature, frequency, and mechanical vibrations. To establish a model that can accurately reflect the metering error under actual operating conditions, this paper introduces a multi-parameter dimensionality reduction model to map complex multi-dimensional data to a low-dimensional space, thereby simplifying the complexity of the problem and improving the generalization ability of the model. Assume that the metering error E can be expressed as a function of the voltage transformer ratio difference ΔV, angle difference $\Delta\theta_V$, current transformer ratio difference ΔI, angle difference $\Delta\theta_I$, secondary circuit error E_{sec}, electric energy meter error E_{meter}, and vibration parameters V_{vib}:

$$E = f(\Delta V, \Delta\theta_V, \Delta I, \Delta\theta_I, E_{\text{sec}}, E_{\text{meter}}, V_{\text{vib}}) \quad (5)$$

Considering that these error factors are closely related to parameters such as voltage V, current I, secondary load S_{sec}, temperature T, frequency f, etc., this paper defines a multi-parameter vector $x = [V, I, S_{sec}, T, f]$ and assumes that the error factors can be expressed as nonlinear functions of these parameters:

$$\begin{cases} f_{PT}(t) = g_1(V(t), I(t), S_{\text{sec}}(t), T(t), f(t)) \\ \delta_{PT}(t) = g_2(V(t), I(t), S_{\text{sec}}(t), T(t), f(t)) \\ f_{CT}(t) = g_3(V(t), I(t), S_{\text{sec}}(t), T(t), f(t)) \\ \delta_{CT}(t) = g_4(V(t), I(t), S_{\text{sec}}(t), T(t), f(t)) \\ \epsilon_{\text{sec}}(t) = g_5(V(t), I(t), S_{\text{sec}}(t), T(t), f(t)) \\ \epsilon_{\text{meter}}(t) = g_6(V(t), I(t), S_{\text{sec}}(t), T(t), f(t)) \end{cases} \quad (6)$$

Among them, $g_1, g_2, g_3, g_4, g_5, g_6$ are nonlinear functions to be determined. To simplify the model, this paper introduces PCA to reduce the dimension of the multi-parameter vector x and extract the principal components $z = [z_1, z_2, ..., z_k]$, where k "$\dim(x)$.

3.1.1 Mathematical description of multi-parameter dimensionality reduction model

The original data is normalized for a more detailed description of the multi-parameter dimension reduction model. Suppose the original data matrix is $X \in \mathbb{R}^{m\times n}$, where m is the number of variables and n is the sample number. The normalized matrix is X':

$$X' = \frac{X - 1_n b^T}{\Sigma} \quad (7)$$

$1_n = [1,1,...,1]^T \in \mathbb{R}^n$, $b = \frac{1}{n}X^T 1_n$ is the mean of the sample data, $\Sigma = \text{diag}(\sigma_1, \sigma_2, ..., \sigma_m)$ is the variance matrix. This paper performs PCA on the normalized data. PCA aims to project the data into a low-dimensional subspace while retaining as much information as possible [12]. Specifically, this paper decomposes the covariance matrix $R = \frac{1}{n-1}X'X'^T$ through eigenvalues to obtain the principal component vector P and the corresponding eigenvalue Λ:

$$R = P\Lambda P^T \quad (8)$$

Among them, $\Lambda = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_m), \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$ is the eigenvalue, $P = [p_1, p_2, ..., p_m]$ is the principal component vector. The first k principal components are selected so that the cumulative contribution rate (CPV) reaches a particular value:

$$\text{CPV}(k) = \frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{m}\lambda_i} \geq 0.95 \quad (9)$$

By selecting the first k principal components, people can map the original data X' to a low-dimensional space Z $Z \in \mathbb{R}^{k\times n}$:

$$Z = P_k^T X' \quad (10)$$

P_k is a matrix composed of the first k principal component vectors.

3.1.2 Error representation of dimensionality reduction model

This paper can express the mutual inductor measurement error in low-dimensional space as a

function of the parameters after dimensionality reduction. Assuming that the parameters after dimensionality reduction are $z = [z_1, z_2, ..., z_k]$, the error can be expressed as:

$$\begin{cases} f_{PT}(t) = h_1(z(t)) \\ \delta_{PT}(t) = h_2(z(t)) \\ f_{CT}(t) = h_3(z(t)) \\ \delta_{CT}(t) = h_4(z(t)) \\ \epsilon_{sec}(t) = h_5(z(t)) \\ \epsilon_{meter}(t) = h_6(z(t)) \end{cases} \tag{11}$$

Among them, $h_1, h_2, h_3, h_4, h_5, h_6$ are nonlinear functions after dimensionality reduction. This paper can express these nonlinear functions through polynomial functions or neural networks. For example, suppose h_1 is a polynomial function:

$$h_1(z) = \sum_{i=1}^{k} a_{1i} z_i + \sum_{i=1}^{k} \sum_{j=i}^{k} a_{1ij} z_i z_j + \ldots \tag{12}$$

Among them, $a_{1i}, a_{1ij}, \cdots$ are the coefficients of the polynomial. In this way, this paper can transform complex nonlinear relationships into polynomial forms, thereby simplifying the model solution process.

3.2 Support vector machine regression to solve the dimensionality reduction equation

To solve the above-mentioned nonlinear functions $h_1, h_2, h_3, h_4, h_5, h_6$ after dimensionality reduction, this paper adopts the support vector machine regression (SVR) method. SVR can effectively handle nonlinear relationships by finding the optimal regression plane in the high-dimensional feature space [13]. Assume that this paper has obtained the reduced-dimensional data set $\{(z_i, y_i)\}_{i=1}^{N}$ through PCA, z_i is the input vector after dimensionality reduction, and y_i is the corresponding output value such as

$$f_{PT}, \delta_{PT}, f_{CT}, \delta_{CT}, \epsilon_{sec}, \epsilon_{meter}) \tag{13}$$

The objective of the SVR is to find a function $f(z) = \langle w, \Phi(z) \rangle + b$, which can minimise the error between the prediction and the actual value of all the sample points and control the complexity of the model. Φ(z) is a nonlinear mapping function that maps the input data to a high-dimensional feature space. w is a weight vector; b is a bias term

To achieve this goal, this paper introduces a linear insensitive loss function ϵ:

$$L_\epsilon(f(z_i), y_i) = \begin{cases} 0 & \text{if } |f(z_i) - y_i| \le \epsilon \\ |f(z_i) - y_i| - \epsilon & \text{otherwise} \end{cases} \tag{14}$$

The optimization problem can be expressed as:

$$\begin{aligned} \min_{w,b,\xi,\xi^*} \quad & \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{N} (\xi_i + \xi_i^*) \\ \text{s.t.} \quad & f(z_i) - y_i \le \epsilon + \xi_i, \ i = 1,2, ..., N \\ & y_i - f(z_i) \le \epsilon + \xi_i^*, \ i = 1,2, ..., N \\ & \xi_i \ge 0, \ \xi_i^* \ge 0, \ i = 1,2, ..., N \end{aligned} \tag{15}$$

Among them, C is the penalty factor, which is used to control the degree of penalty for the model on the error; ξ_i and ξ_i^* are slack variables, which are used to deal with the inseparable situation. By introducing the Lagrange multiplier method, the above optimization problem is transformed into a dual problem:

$$\begin{aligned} \max_{\alpha,\alpha^*} \quad & \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) y_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) K(z_i, z_j) \\ \text{s.t.} \quad & \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) = 0 \\ & 0 \le \alpha_i, \alpha_i^* \le C, \ i = 1,2, ..., N \end{aligned} \tag{16}$$

$K(z_i, z_j) = \langle \Phi(z_i), \Phi(z_j) \rangle$ is the kernel function. Commonly used kernel functions include RBF:

$$K(z_i, z_j) = \exp\left(-\frac{\| z_i - z_j \|^2}{2\sigma^2} \right) \tag{17}$$

People can get the optimal solution α_i and α_i^*, and then get the regression function:

$$f(z) = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) K(z, z_i) + b \tag{18}$$

This paper can predict the actual metering error of the transformer based on the dimension reduction parameter z measured in real time.

To further improve the accuracy of error prediction, this paper can adopt a multi-model fusion strategy. Specifically, this paper trains multiple SVR models. Each model uses a different kernel function or parameter setting [14]. Then, the prediction results of these models are fused through weighted averaging or voting mechanisms. For example, assuming that this paper has M SVR models, and the prediction result of each model is $f_m(z)$, the fused prediction result can be expressed as:

$$f_{ensemble}(z) = \sum_{m=1}^{M} w_m f_m(z) \tag{19}$$

Among them, w_m is the weight of the m model, satisfying $\sum_{m=1}^{M} w_m = 1$. Weight may be allocated based on the model's performance in the validation set. For example, the more efficient model is assigned a higher weight. The advantages of different models are fully utilized to enhance the precision and robustness of the error forecast. The multi-model fusion strategy has been clarified, including the definition of M as the number of SVR models.

This paper introduces weighting factors $\lambda_1, \lambda_2, \lambda_3, \lambda_4$, which represent the weights of transformer ratio, angle, secondary circuit, and electric energy meter errors in the comprehensive error [15].

$$\epsilon_{\text{total}} = \sqrt{\left(\lambda_1 f_{PT}^{pred} + \lambda_2 f_{CT}^{pred} + \lambda_3 \epsilon_{\text{sec}}^{mon} + \lambda_4 \epsilon_{\text{meter}}^{mon}\right)^2 + \left(\lambda_1 \delta_{PT}^{pred} + \lambda_2 \delta_{CT}^{pred}\right)^2} \tag{21}$$

Among them, $\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 1$, and $\lambda_i \geq 0, i = 1,2,3,4$. In actual operation, the importance of different error factors may change over time. Therefore, this paper proposes a dynamic weight adjustment strategy to adjust the weights automatically according to real-time monitoring data. Calculate the correlation coefficient matrix R between each error factor, R_{ij} represents the correlation coefficient between the i error factor and the j error factor. According to the correlation coefficient matrix, dynamically adjust the weight. For example, assuming that the initial weight is $\lambda^{(0)} = \left[\lambda_1^{(0)}, \lambda_2^{(0)}, \lambda_3^{(0)}, \lambda_4^{(0)}\right]$, the weight at time step t can be updated as:

$$\lambda^{(t)} = \frac{\lambda^{(t-1)} \odot \mathrm{R}^{(t)}}{\|\lambda^{(t-1)} \odot \mathrm{R}^{(t)}\|_1} \tag{22}$$

Among them, $\odot$ represents element-by-element multiplication, and $\|\cdot\|_1$ represents the L_1 norm, which is used to normalize the weight.

The dynamic weight adjustment equation has been properly defined to address the incongruencies. Through the dynamic weight adjustment strategy, this paper can automatically adjust the weight of each error factor according to the real-time monitoring data, thereby improving the accuracy and adaptability of the comprehensive error

4 Experiment simulation

4.1 Simulation platform and data set construction

To verify the validity and veracity of the PCA-SVR-based error evaluation algorithm, a 500 kV power system simulation model is built and simulated by MATLAB [16]. The simulation platform can simulate various operating conditions in actual power systems, including load conditions, grid fluctuations, and error characteristics of transformers and electric energy meters.

3.3 Error synthesis calculation

After obtaining the predicted value of the transformer metering error, this paper needs to synthesize these errors with the secondary circuit error and the electric energy meter error to obtain δ_{CT}^{pred}. At the same time, the secondary circuit error obtained by online monitoring is ϵ_{sec}^{mon}, and the electric energy meter error is ϵ_{meter}^{mon}.

The comprehensive error ϵ_{total} can be expressed as:

$$\epsilon_{\text{total}} = \sqrt{\left(f_{PT}^{\text{pred}} + f_{CT}^{pred} + \epsilon_{sec}^{\text{mon}} + \epsilon_{\text{meter}}^{\text{mon}}\right)^2 + \left(\delta_{PT}^{\text{pred}} + \delta_{CT}^{\text{pred}}\right)^2} \tag{20}$$

The weights may be adjusted following the practical operation conditions and the significance of the error factors. Therefore, the comprehensive error can be expressed as:

4.1.1 Introduction to the simulation platform

The simulation platform is based on MATLAB software and uses its robust numerical calculation and graphic visualization capabilities to construct a detailed model of a 500kV power system. The model includes key components such as voltage transformers (PT), current transformers (CT), secondary circuits, and electric energy meters. By setting different operating parameters, such as voltage, current, load power factor, etc., electric energy metering data under various actual operating conditions can be simulated.

4.1.2 Data set construction

This paper sets up various operating conditions on the simulation platform to generate error datasets for training and testing SVR models, including regular operation, light load operation, heavy load operation, and fault operation [17]. Under each operating condition, the voltage transformer ratio difference f_{PT}, angle difference δ_{PT}, current transformer ratio difference f_{CT}, angle difference δ_{CT}, secondary circuit error ϵ_{sec} and electric energy meter error ϵ_{meter} are recorded. The construction of the dataset takes into account random fluctuations and noise interference under different operating conditions to ensure the authenticity and diversity of the data.

Table 1 shows some error datasets generated by simulation, including voltage, current, secondary load and corresponding transformer error and electric energy meter error under different operating conditions.

Table 1: Error datasets generated by simulation

Working condition type	Voltage(kV)	Current (A)	Secondary load (VA)	Δrv(%)	Δθv(′)	em(%)
Normal operation	500	1000	100	-0.1135	3.4158	-0.1458
Light load operation	500	500	50	-0.1106	3.6597	-0.1155
Heavy load operation	500	2000	200	-0.1116	3.5124	-0.1170
Fault operation	500	1500	150	-0.1056	3.4390	-0.1161

The vibration data includes vibration amplitude and frequency, which are generated based on actual operating conditions and possible interference. The specific settings are as follows: the vibration amplitude varies from 0.1 mm to 1.0 mm, simulating conditions from mild vibration to stronger vibration. The vibration frequency varies from 10 Hz to 50 Hz, covering the common mechanical vibration frequency range. The vibration data is generated through the simulation platform and recorded synchronously with the power metering data to form an extended data set containing vibration parameters. In this way, the algorithm is able to learn the relationship between vibration and metering error, thereby more comprehensively evaluating the metering error of the transformer.

4.2 Algorithm training and optimization

This paper uses the RBF as the kernel function when training the SVR model. The RBF kernel function has good nonlinear mapping capabilities and can handle complex error data relationships effectively. The kernel parameters in the SVR model's kernel function have been explicitly defined.

To select the optimal parameter combination, this paper uses the cross-validation method. The specific steps are as follows:

1. Parameter grid search: Define a parameter grid, including multiple possible penalty factor C values and kernel function width σ values. For example, C can take [0.1,1,10,100], and σ can take [0.1,1,10,100].

2. Cross-validation: The data set is divided into training and validation sets. For each pair (C, σ), the SVR model is trained with the training set, and its performance is evaluated on the validation set. Everyday performance indices are the RMSE and the measurement factor (R^2).

3. Select the optimal parameters: Select the parameter combination (C^*, σ^*) that makes the validation set perform best.

4. Model validation: Retrain the model using the optimal parameters (C^*, σ^*) and validate the generalization ability of the model on an independent test set. Table 2 shows the performance indices of the model for various combinations of parameters, including RMSE and R^2.

Table 2: Model performance indicators for various combinations of parameters.

C	σ	RMSE	R^2
0.1	0.1	0.056	0.987
0.1	1	0.048	0.991
0.1	10	0.052	0.989
1	0.1	0.045	0.993
1	1	0.042	0.995
1	10	0.047	0.992
10	0.1	0.046	0.994
10	1	0.043	0.996
10	10	0.049	0.993
100	0.1	0.047	0.995
100	1	0.044	0.997
100	10	0.051	0.994

It can be seen from Table 2 that when C =100 and σ =1, the performance of the model is optimal, with RMSE of $0.044, R^2$ of 0.997. Therefore, this paper selects C =100 and σ =1 as the optimal parameter combination.

4.3 Simulation results and evaluation

4.3.1 Comparison of algorithm output and actual error data

To evaluate the accuracy of the proposed algorithm, this paper compares the algorithm's output with the actual error data. Figure 1 shows the comparison curve of the actual value and the predicted value of the voltage transformer ratio difference. The proposed algorithm can accurately predict the ratio difference of the voltage transformer, and the error between the expected value and the actual value is minimal. Table 3 shows the comparison results of the exact value and the predicted value of different error factors. Table 3 shows that the expected values of the proposed algorithm on all error factors are very close to the actual values, and the relative errors are all within a reasonable range.

To assess the precision of the algorithm, the output of the algorithm is compared with the fundamental error data. Fig. 1 illustrates the comparison between the actual value and the estimated value of the voltage transformer ratio difference. This method can accurately forecast the voltage transformer ratio, and there is little difference between the prediction and fundamental values. Table 3 compares the actual value and the prediction of different error factors.

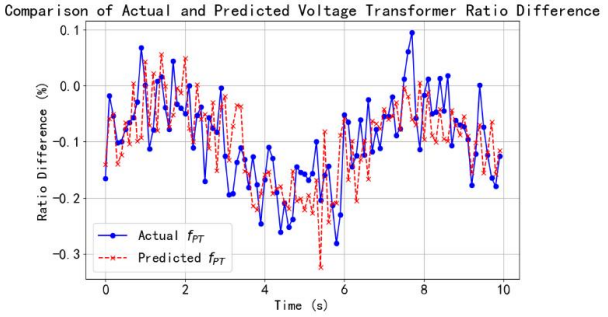


Figure 1: Comparison curve of actual value and predicted value of voltage transformer ratio difference f_{PT}.

Table 3: Comparison of actual value and predicted value of different error factors.

Error Factors	Actual value (%)	Prediction value (%)	Relative error (%)
f_{PT}	-0.1135	-0.1134	-0.0314
δ_{PT}	3.4158	3.4097	-0.1801
f_{CT}	-0.1458	-0.1455	-0.1976
δ_{CT}	6.1684	6.1536	-0.2396
ϵ_{sec}	-0.05	-0.051	-0.0200
ϵ_{meter}	-0.02	-0.021	-0.0500

4.3.2 Evaluation of algorithm stability and real-time performance

Many simulation experiments were conducted under different working conditions to evaluate its stability and real-time performance. Figure 2 illustrates the variation of the predicted error in different load conditions.

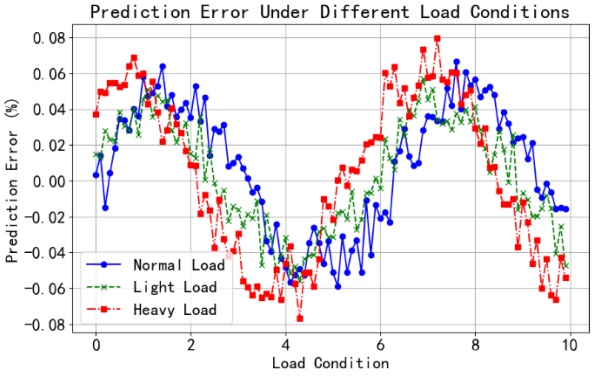


Figure 2: Prediction error fluctuation of the algorithm under different load conditions.

Figure 2 shows that the variation of the predicted error of the proposed algorithm is slight and stable. Moreover, by implementing the optimized algorithm, the real-time performance of the algorithm is significantly increased, and the error assessment can be finished quickly, which can satisfy the need for real-time monitoring.

4.3.3 Impact of vibration factors on error evaluation

In order to verify the impact of vibration factors on the metering error of the transformer, this paper simulated the error evaluation results under different vibration conditions on the simulation platform. Figure 3 shows the evaluation results of the transformer metering error under different vibration amplitudes and frequencies.

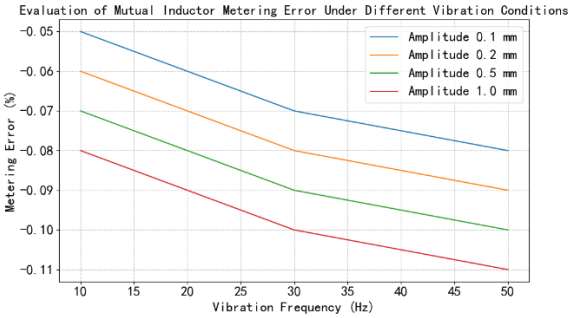


Figure 3: Evaluation results of the transformer metering error under different vibration conditions

It can be seen from Figure 3 that the changes in vibration amplitude and frequency have a significant impact on the metering error of the transformer. As the vibration amplitude and frequency increase, the metering error gradually increases. This shows that in actual operation, the vibration factor may introduce additional errors, which need to be considered in the error evaluation. The discussion of vibration factors' impact on error evaluation has been quantified.

4.4 Comparison of error evaluation results with actual data

4.4.1 Comparison with traditional methods

To prove the superiority of this algorithm, the PCA-SVR algorithm is compared with the conventional one. Traditional methods usually combine the synthesis error, the second circuit error, the power meter error, and the complicated relation in the practical operation. Table 4 compares the results of the error assessment between the proposed algorithm and the conventional one in various operating conditions [18]. The results show that the proposed algorithm is more approximate to the real ones, and the relative error is lower than the conventional one.

Table 4: Comparison of error evaluation results under different working conditions.

Working condition type	Traditional method error (%)	Error of the proposed algorithm (%)	Actual error (%)	Relative error (%)
Normal operation	-0.52	-0.517	-0.5158	0.2256
Light load operation	-0.48	-0.473	-0.4716	0.3032
Heavy load operation	-0.46	-0.457	-0.4559	0.3048
Fault operation	-0.45	-0.448	-0.4485	0.0021

4.4.2 Comparison with Other Algorithms

To further prove the superiority of this algorithm, the PCA-SVR-based error assessment algorithm is compared with other commonly used algorithms, such as Neural Network (NN) and SVM. Table 5 compares the performance of the proposed algorithm with that of different algorithms in various operating conditions [19]. The PCA-SVR-based error assessment algorithm is more approximate to the real one, and the relative error is less than that of NN and SVM-based, proving the algorithm's superiority.

Table 5: Comparison of error evaluation results of different algorithms.

Working condition type	NN-based algorithms (%)	SVM-based algorithm (%)	Proposed algorithm (%)	Actual error (%)	Relative error (%)
Normal operation	-0.53	-0.525	-0.517	-0.5158	0.2256
Light load operation	-0.49	-0.485	-0.473	-0.4716	0.3032
Heavy load operation	-0.47	-0.465	-0.457	-0.4559	0.3048
Fault operation	-0.46	-0.455	-0.448	-0.4485	0.0021

4.5 Adaptability and anti-interference ability of the model

4.5.1 Adaptability in dynamic power grid environment

To evaluate the model's adaptability in a dynamic power grid environment, this paper simulates the dynamic changes of grid voltage, current and load power factor on the simulation platform. Figure 4 shows the error evaluation results of the model in a dynamic power grid environment. The proposed model can accurately evaluate the error of the electric energy metering device in a dynamic power grid environment and is adaptable.

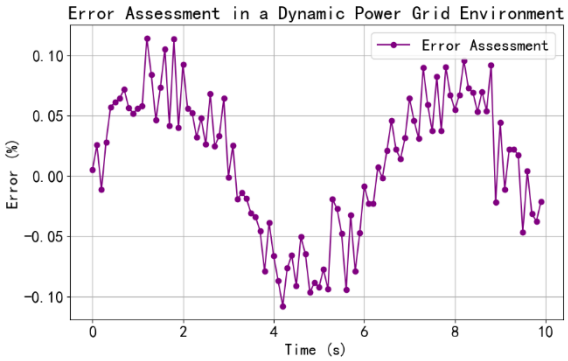


Figure 4: Error evaluation results of the model in a dynamic power grid environment.

4.5.2 stability under load fluctuations and external disturbances

The simulation is carried out on the simulation platform to assess the stability of the model under the influence of both the load and the external disturbance. Figure 5 illustrates the model's error assessment results in fluctuating load and external disturbance. It has been proved that the model can be used to evaluate the stability of the error when the load fluctuates and the outside is disturbed, and the anti-interference capability is good.

Figure 5: Error evaluation results of the model under load fluctuation and external disturbance.

Table 6: Comparison of comprehensive error evaluation results under different working conditions.

Working condition type	Comprehensive error of traditional method (%)	Comprehensive error of the proposed algorithm (%)	Actual comprehensive error (%)	Relative error (%)
Normal operation	-0.52	-0.517	-0.5158	0.2256
Light load operation	-0.48	-0.473	-0.4716	0.3032
Heavy load operation	-0.46	-0.457	-0.4559	0.3048
Fault operation	-0.45	-0.448	-0.4485	0.0021

4.6 Comprehensive evaluation effect

A comparison is made between the PCA-SVR-based and the conventional one to validate the algorithm's performance. Table 6 compares the synthetic error assessment results of the proposed algorithm with those of the conventional ones.

5 Conclusion

The PCA-SVR-based error assessment algorithm has proved effective and superior by constructing the simulation platform and data set and training, optimizing, and evaluating the algorithm. The algorithm can be used to estimate the error of power measurement equipment accurately under the condition of a dynamic power network, and it is very adaptable and anti-interference. This algorithm has superior precision, stability, and real-time accuracy compared to conventional and standard algorithms. So, this algorithm offers a reliable method to evaluate the error on the line of power measurement equipment. In addition, this study also considered the impact of mechanical vibration on the measurement error of the transformer, and further improved the accuracy and reliability of error assessment by introducing vibration sensors and vibration parameters.

Acknowledgements

The work was supported by the science and technology project of State Grid Ningxia Electric Power Co., Ltd. funds "Research and Application of On-site Detection Technology for Online Monitoring Device of Voltage Transformer Measurement Performance", Project Number: 5229YX240005.

Referencess

[1] Janez Zupan, Andrej Pretnar. A novel approach for fault diagnosis in power transformers using machine learning and big data. Informatica, 2023, 49(18): 1 - 15. https://doi.org/10.3233/INF-2023-3124.

[2] Boštjan Fal, Marko Čopar. Intelligent monitoring system for power transformers based on IoT and data fusion. Informatica, 2022, 48(16): 1 - 20. https://doi.org/10.3233/INF-2022-3098.

[3] Matjaž Jurk, Ljupčo Kovače. Data-driven prediction model for transformer insulation condition using deep learning. Informatica, 2021, 47(14): 1 - 10. https://doi.org/10.3233/INF-2021-3015.

[4] Zhao, Z., Chen, Y., Liu, J., Cheng, Y., Tang, C., & Yao, C. (2022). Evaluation of operating state for smart electricity meters based on transformer–encoder–BiLSTM. IEEE Transactions on Industrial Informatics, 19(3), 2409-2420. https://doi.org/10.1109/TII.2021.3112644.

[5] Ashok, K., Li, D., Gebraeel, N., & Divan, D. (2021). Online detection of inter-turn winding faults in single-phase distribution transformers using smart meter data. IEEE Transactions on Smart Grid, 12(6), 5073-5083. https://doi.org/10.1109/TSG.2021.3093566.

[6] Jiao, Y., Li, H., Hu, C., Zhang, Z., & Zhang, C. (2019). Data-driven evaluation for error states of standard electricity meters on automatic verification assembly line. IEEE Transactions on Industrial Informatics, 15(9), 4999-5010. https://doi.org/10.1109/TII.2019.2918121.

[7] Goklani, H., Gajjar, G., & Soman, S. A. (2020). Instrument transformer calibration and robust estimation of transmission line parameters using PMU measurements. IEEE Transactions on Power Systems, 36(3), 1761-1770. https://doi.org/10.1109/TPWRS.2020.2978669.

[8] Papa, L., Russo, P., & Amerini, I. (2023). Meter: A mobile vision transformer architecture for monocular depth estimation. IEEE Transactions on Circuits and Systems for Video Technology, 33(10), 5882-5893. https://doi.org/10.1109/TCSVT.2023.3267709.

[9] Lindskog, A., Jarlemark, P., & Svensson, S. (2022). In situ calibration of transmission system power grid energy meters using energy measurements and Kalman filter. IEEE Transactions on Power Delivery, 38(2), 988-997. https://doi.org/10.1109/TPWRD.2021.3137745.

[10] Nakutis, Ž., Rinaldi, S., Kuzas, P., & Lukočius, R. (2020). A method for noninvasive remote monitoring of energy meter error using power consumption profile. IEEE Transactions on Instrumentation and Measurement, 69(9), 6677-6685. https://doi.org/10.1109/TIM.2020.3016021.

[11] Bu, F., Dehghanpour, K., & Wang, Z. (2021). Enriching load data using micro-PMUs and smart meters. IEEE Transactions on Smart Grid, 12(6), 5084-5094. https://doi.org/10.1109/TSG.2021.3093567.

[12] Shah, A. L., Mesbah, W., & Al-Awami, A. T. (2020). An algorithm for accurate detection and correction of technical and nontechnical losses using smart metering. IEEE Transactions on Instrumentation and Measurement, 69(11), 8809-8820. https://doi.org/10.1109/TIM.2020.3016020.

[13] De Carne, G., Bruno, S., Liserre, M., & La Scala, M. (2019). Distributed online load sensitivity identification by smart transformer and industrial metering. IEEE Transactions on Industry Applications, 55(6), 7328-7337. https://doi.org/10.1109/TIA.2019.2934290.

[14] Pau, M., Patti, E., Barbierato, L., Estebsari, A., Pons, E., Ponci, F., & Monti, A. (2019). Design and accuracy analysis of multilevel state estimation based on smart metering infrastructure. IEEE Transactions on Instrumentation and Measurement, 68(11), 4300-4312. https://doi.org/10.1109/TIM.2019.2926705.

[15] Yao, R., Wang, N., Chen, P., Ma, D., & Sheng, X. (2023). A CNN-transformer hybrid approach for an intrusion detection system in advanced metering infrastructure. Multimedia Tools and Applications, 82(13), 19463-19486. https://doi.org/10.1007/s11042-023-19335-8.

[16] Xiong, K., Yue, C., Zhou, F., Li, S., Yu, J., & Liang, S. (2023). The development and performance of dual-core DC tolerant current transformer. IEEE Transactions on Power Delivery, 38(3), 1521-1533. https://doi.org/10.1109/TPWRD.2022.3225984.

[17] Lindskog, A., Thiringer, T., & Lindquist, T. (2021). A data-driven approach for monitoring puncture events in capacitor voltage transformers. IEEE

Transactions on Power Delivery, 37(3), 2153-2162. https://doi.org/10.1109/TPWRD.2021.3093568.

[18] Hasanah, A. W., Koerniawan, T., & Yuliansyah, Y. (2019). Kajian ketelitian current transformer (CT) terhadap kesalahan rasio arus pada pelanggan 197 kVA. Energi & Kelistrikan, 11(1), 9-16. https://doi.org/10.31226/ek.v11i1.231.

[19] Xu, Z., Jiang, W., Xu, J., Wang, D., Wang, Y., & Ou, Z. (2022). Distribution network topology identification using asynchronous transformer monitoring data. IEEE Transactions on Industry Applications, 59(1), 323-331. https://doi.org/10.1109/TIA.2021.3112645.

[20] Gang, L., Cong, L., Xiaofu, X., Jian, H., Kun, L., Runhao, Z., & Ruijin, L. (2020). Simulation of the magnetic field distribution and voltage error characteristics of the three-phase three-component combined transformer with new three-cylinder core structure. IEEE Transactions on Magnetics, 56(4), 1-4. https://doi.org/10.1109/TMAG.2020.2962750.

[21] Kong, X., Zhang, X., Lu, N., Ma, Y., & Li, Y. (2021). Online smart meter measurement error estimation based on EKF and LMRLS method. IEEE Transactions on Smart Grid, 12(5), 4269-4279. https://doi.org/10.1109/TSG.2021.3093569.

[22] Raghuvamsi, Y., & Teeparthi, K. (2024). Distribution system state estimation with Transformer-Bi-LSTM-based imputation model for missing measurements. Neural Computing and Applications, 36(3), 1295-1312. https://doi.org/10.1007/s00521-023-08712-3.

Efficient Audio Steganography Using Bit Comparison and Substitution Algorithms for Improved Embedding Capacity

K Revathi[1], S. Kaja Mohideen[2*], and Latha Tamilselvan[3]
[1,2]Department of Electronics and Communication Engineering, B.S. Abdur Rahman Crescent Institute of Science and Technology, Vandalur, Chennai, India
[3]Department of Information Technology, B.S. Abdur Rahman Crescent Institute of Science and Technology, Vandalur, Chennai, India
E-mail: revathivigneswaranphd@gmail.com, kajamohideen@crescent.education,and
latha.tamil@crescent.education
*Corresponding author

Keywords: audio steganography, lsb steganography, wav file, text files, mse, psnr, ber

Received: November 12, 2024

Steganography conceals information, with audio files increasingly used in this practice. The Least Significant Bit algorithm in audio steganography requires numerous samples to embed text files effectively. This study proposes an efficient methodology for embedding text files into a digital audio file. The approach also involves embedding the index value of an audio file, which contains hidden text files in another digital audio file. The proposed algorithms are the Bit Comparison and Substitution Algorithm for embedding text files in audio files and the Bit Comparison and Retrieval Algorithm for extracting text files from audio files. Additionally, a Least Significant Bit Algorithm with a secret key (Ks) embeds the index value of the audio file. The proposed algorithm achieves an average mean squared error of 0.0513, a peak signal-to-noise ratio of 41.13732041 dB, and a bit error rate of 0.000114 for embedding 2400 bytes of text files in audio files. These performance metrics demonstrate that the proposed algorithm quantitatively outperforms the traditional least significant bit algorithm by requiring fewer audio samples to embed text files. The proposed algorithm embeds 2400 bytes of text into 2400 audio samples, with each byte corresponding to a single audio sample. Each byte of the text file is embedded by modifying one bit per audio sample, resulting in 2400 difference bits between audio and stego audio files.

Povzetek: Izvirna metodologija za vgradnjo besedilnih datotek v avdio datoteke z uporabo algoritmov za primerjavo bitov in zamenjavo bistveno izboljša kapaciteto vgradnje, s tem pa zmanjšuje število potrebnih vzorcev in povečuje varnost in kakovost skrivnega prenosa podatkov.

1 Introduction

In the domain of steganography, enhancing the secrecy of hidden files requires concealing two key attributes: the embedding method and the precise location within the audio file where the file is incorporated [1].

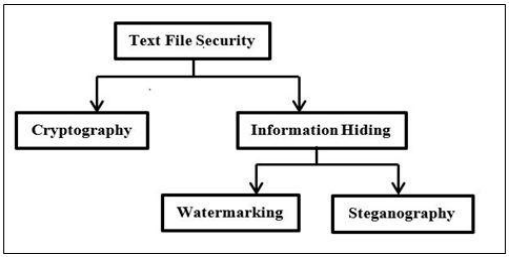


Figure 1: Text file security structure.

Cryptographic and steganographic techniques are essential to ensuring information security during communication. Figure 1 illustrates the structure of text file security [2]. Cryptographic algorithms enable the encryption and decryption of information, ensuring its security through the transformation of plaintext into ciphertext [3].

Steganography, alternatively, conceals information within a medium and traces its origins back to ancient Greece. Information has historically been hidden in various forms like music scores, paintings, letters written with invisible ink, or even human skin. Digital media steganography is a relatively recent area of research that focuses on concealing information in digital content, such as digital images [4], video [5], or audio files [6]. Audio steganography is of particular interest due to the slight distortion in the amplitude of audio files during data embedding, making audio a suitable cover medium. It hides multimedia objects in audio files such as WAV, and MP3. Using commonly available carrier files is highly desirable, as it reduces the risk of detection by third-party observers and makes carrier modification caused by secret data insertion undetectable [7]. Figure 2 presents the model for audio steganography [8].

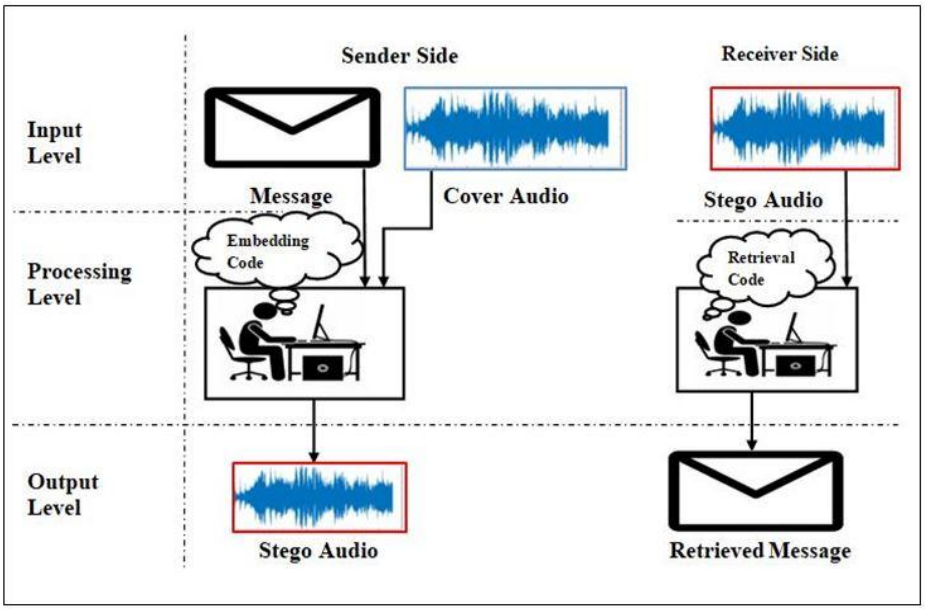


Figure 2: Model for audio steganography.

Efficient steganographic concealment generally requires three unique characteristics: security, capacity, and strength. Capacity refers to the total size of confidential data hidden within a file. An imperceptible steganographic system is considered secure. Robust audio steganography ensures that the parameters of both the cover and stego audio files are identical. Robustness refers to the stego file's resilience against steganalysis and conflicts with hiding capacity. Specifically, increasing hiding capacity reduces the security and imperceptibility of stego files. Balancing capacity, security, and imperceptibility poses a significant challenge.

Accordingly, the principal aim of steganography is to enhance embedding capacity, security, and imperceptibility while preserving robustness[1]. The hiding methods in audio files include Least Significant Bit (LSB), Echo, and Spread Spectrum, each developed for various applications[9], [10]. The LSB algorithm embeds each bit of the secret message into the LSB of the audio sample. Its simplicity and practical success derive from the limited sensitivity of the human auditory system (HAS). The HAS fails to detect subtle changes in audio frequency, particularly in the high-frequency regions.

However, a drawback is the predictable embedding of secret messages, which facilitates easy recovery for potential attackers. In response to these limitations, researchers have proposed various approaches to enhance the traditional LSB algorithm and mitigate these challenges [11], [12], [13]. To enhance embedding capacity, researchers embed the secret bits in the higher-order LSBs of the audio cover sample [9], [14]. At the same time, information loss occurs during communication due to noise.

The widespread use and accessibility of audio files make them suitable for conveying concealed information. Researchers have explored embedding information within audio samples using distortion functions, multi-domain embedding, Secure Advanced Audio Coding steganography scheme based on multi-view statistical distortion (SofMvD) with syndrome trellis code (STC), and adaptive-STC based on the adaptive parity-check matrix and constant distortion function. The targeted intelligent optimization algorithm, Generation of Optimal Allocation Strategy (GOAS), models the adaptive parity-check matrix to address the multi-constrained integer programming problem[15], [16], [17], [18]. The authors provided an in-depth review of audio steganography, discussing various concepts and methodologies in comparison to existing audio steganographic methods [8],[19]. The authors investigated the utilization of commercial off-the-shelf (COTS) software for audio steganography and steganalysis by formulating a model [20].

The authors explore the connections between information society, electronics, and artificial intelligence through twenty-four information society laws that demonstrate the rapid growth of electronic devices. The authors' concept enables the potential implementation of audio steganography communication in real-world applications with the support of AI [21].

The authors introduced a lightweight authentication solution for IoT edge devices, ensuring efficient data transmission under limited bandwidth. This solution uses lightweight symmetric cryptography and leverages the ChaCha20 algorithm for session key establishment, providing robust security with minimal communication and time costs [22]. Similarly, audio steganography strengthens IoT security by concealing data within multimedia files (audio), enabling secure communication as a future scope.

Steganography research extends security by integrating cryptography and steganography[23],[24].

The author proposed a novel steganographic scheme based on data decomposition and stego-coding methods to increase the embedding efficiency[25].

The study investigates double-layered security and multiple embedding techniques in steganography. These techniques include embedding an encrypted color image within another color image, hiding an encrypted text file within an image, embedding this image within an audio file, and embedding text files into the highest fitness bitstream position of the audio samples using a genetic algorithm[2],[26],[27].

Most steganography research focuses on images, with fewer studies exploring data hiding in audio files. Hence, audio steganography is an emerging area of research.

The LSB algorithm in current audio steganography methods embeds the least significant bit of an 8-bit audio sample with test file bits at a 12.5% embedding rate, and using two least significant bits results in a 25% embedding rate. Consequently, a low embedding rate restricts the amount of hidden data, and to embed more information, a larger audio file is required.

The probability of errors in the embedded bits in the LSB algorithm depends on the bit pattern of audio samples. The error bits are not predictable; they vary. Therefore, it is essential to propose algorithms that prioritize the reduction of bit error rates in a controlled pattern while simultaneously increasing the embedding rate.

Table 1 summarizes related work features and performance metrics to compare with the proposed algorithms.

Similarly, the Least Significant Bit algorithm uses eight audio samples to embed one byte of a text file. As the embedding capacity increases, more audio samples are needed to accommodate the increased capacity. The LSB algorithm utilizes 19200 samples for embedding 2400 bytes of text files.

Table 1: The summary of related works features and results.

Reference No.	Methodology	Performance / Results
[28]	Audio Steganography: Cover File: Audio Hidden File: Text Embedding Feature: LSB algorithm Encryption Feature: Pattern Matching Algorithm Dual Encryption	The Pattern Matching method achieved Mean Squared Error (MSE) of 0.5619 and 0.3994 in the Normal LSB method. Embedding rate is 12.5%.
[29]	Audio Steganography: Cover File: Audio Hidden File: Text Embedding Feature: Two LSB modification Encryption Feature: RSA to protect data.	MSE, Peak Signal to Noise (PSNR), and Bit Error Rate (BER) sequentially are 9.08875, 34.2775 dB, and 17.995%.
[30]	Audio Steganography: Cover File: Audio Hidden File: Text Embedding Feature: Encrypted Data embedded in 4 and 5 LSB bits of audio file.	A PSNR of 17.71916498 dB resulted in the embedding of 103 characters in a JAZZ audio file.
[31]	Audio Steganography: Cover File: Audio Hidden File: Text Embedding Feature: Random Key Indexing method (Trusted Third Party Stego key and Secondary Key generated at encoder). Encryption Feature: AES-256	A 16-bit stereo audio file with 532bytes of embedded message resulted in a BER of Right Channel: 0.0187%. Left Channel: 0.0198%.
[32]	Audio Steganography: Cover File: Audio Hidden File: Audio Embedding Feature: ECA-BM: *Fractal coding and chaotic LSB	The NC of the reconstructed secret file is 0.99992
[33]	Audio Steganography: Cover File: Audio Hidden File: Audio Embedding Feature: HASFC *Fractal coding *Chaotic least significant bit	Audio Steganography model for secure audio transmission resistant to steganalysis attacks, with a Histogram error rate (HER) of 0.1278 and a Difference Ratio (DR) of the fourth first moments with mean:0.0799, variance:0.0008, skewness:0.0018, kurtosis:0.0028
[34]	Audio Steganography: Cover File: Audio Hidden File: Image and Audio Embedding Feature: Integer-to-Integer Lifting Wavelet Transform (Int2Int LWT) and Least Significant Bits (LSBs) substitution. Encryption Feature: Adaptive steganography key (Stego key)	The lossless audio steganography approach underwent a statistical steganalysis test that resulted in HER of 1.2274e-04 and a Difference Ratio (DR) of the first fourth moments with mean: 0.2304, variance:0, skewness: 0.0004, kurtosis: 0.0005

These limitations of low embedding rates and high bit error rates, caused by the increased use of samples for embedding, underscore the need for the proposed algorithm, which aims to enhance embedding rates while minimizing error rates and thereby improving the overall effectiveness of audio steganography -

The two proposed algorithms are:

1. The Bit Comparison and Substitution algorithm (BCS)
2. The Bit Comparison and Retrieval algorithm (BCR)

These algorithms significantly enhance the embedding process by embedding one byte (8 bits) from text file into one-byte of a two-byte audio sample. This innovative approach allows embedding one character within one audio sample in contrast to the LSB method. By employing a one-bit difference between the samples in the cover audio file and the resulting stego audio file, our algorithms ensure that only one bit is in error for every 8 bits of data embedded.

The novelty of our approach lies in its efficient utilization of audio samples, achieving high embedding rates while minimizing error rates. This capability enhances the security of the communication with a reduced number of samples involved in embedding.

2 Methodology

The proposed BCS and BCR algorithms address the research gap by using all bits in an 8-bit audio sample to embed an 8-bit character. This embedding process results in a bit error of 1 between the audio sample and the character.

It includes three main components: the Bit Comparison and Substitution algorithm (BCS) for embedding text files (TF) into the audio file (AF1), the Least Significant Bit algorithm for embedding the index of audio samples from the stego audio file (SAF1) into the audio file (AF2), and the LSB Retrieval algorithm for recovering index values from the stego audio file (SAF2), followed by the Bit Comparison and Retrieval (BCR) algorithm for extracting text files from the stego audio file (SAF1).

2.1 Bit comparison and substitution algorithm

2.1.1 Embedding text file into audio file

The BCS algorithm is designed for audio steganography with digital audio WAV files. Each sample in these files consists of two bytes, representing the audio's amplitude. The hidden message is a text file that includes punctuation, numerals, uppercase and lowercase letters, and printable characters. Each character is assigned its ASCII value and converted to a one-byte binary value. Figure 4 depicts the hiding behaviour of the proposed BCS algorithm. The algorithm compares the eight least significant bits from the audio file (AF1) with the corresponding bits from the text file. It calculates the sum of their bit differences, denoted as 'K'. If K=1, an audio sample from AF1 is computed, embedding eight consecutive LSBs with one byte of text file data. This process is repeated for each text file byte, creating a Stego Audio File (SAF1). This embedding method integrates one byte of the text file into one byte of a two-byte audio sample in AF1, resulting in one incorrect bit. Figure 3 illustrates the framework of this model.

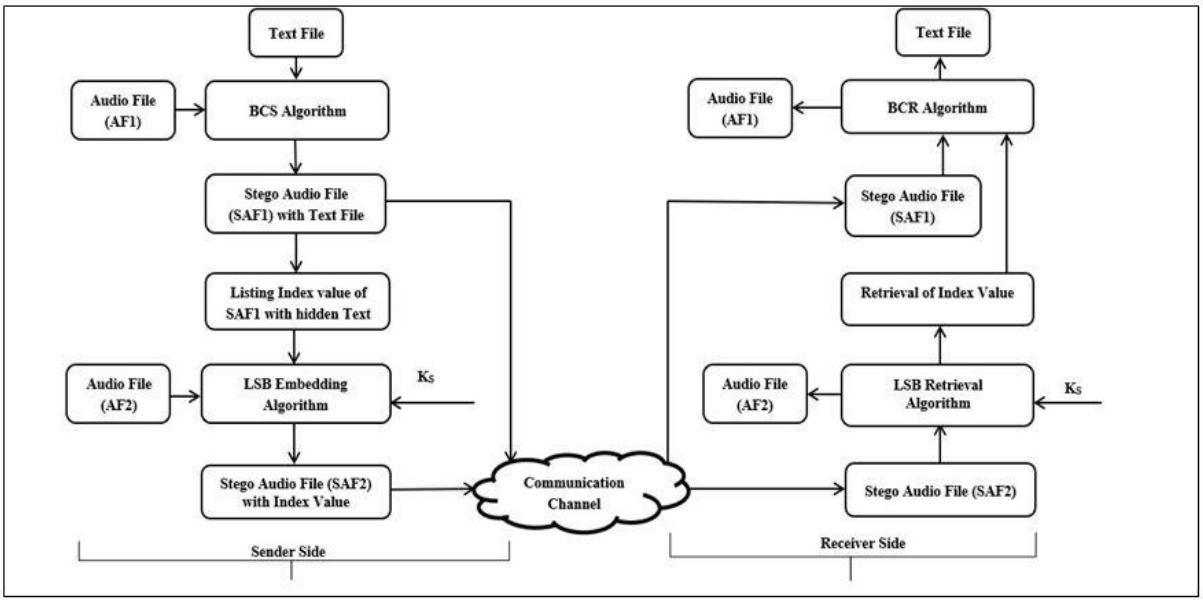


Figure 3: The framework of the proposed method

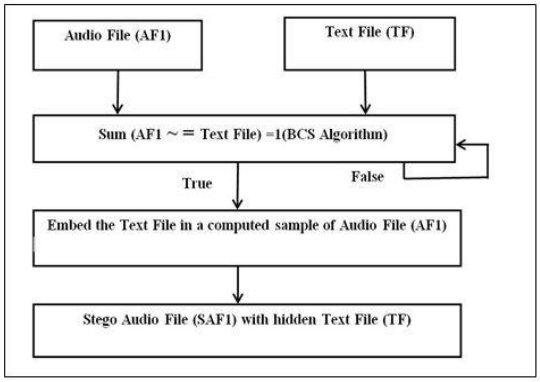


Figure 4: Flow diagram for proposed BCS algorithm.

It outperforms the LSB algorithm by producing fewer incorrect bits, thereby enhancing security and Peak Signal-to-Noise Ratio (PSNR). The size of the text file determines the number of difference bits between the cover (AF1) and stego audio file (SAF1). A 2400-character text file results in 2400 incorrect bits between the cover audio file (AF1) and the stego audio file using the BCS algorithm.

Furthermore, Figure 5 illustrates an example of embedding a one-byte text file in a two-byte audio file (AF1). Initially, both the audio file (AF1) and the text file are converted to binary, resulting in 0011110110101001 0010100101111010 0011110111110000 (Audio File 1) and 11111000 (Text File). Therefore, embedding the text file in the Audio File (AF1) commences if K=1.

2.1.2 BCS Algorithm-embedding text file in audio file

The Sender side in Figure 3 illustrates the algorithm steps for clearer understanding.

Input: Audio File (AF1) and Text File (TF).
Output: Stego Audio File (SAF1)

1. Read the Audio File (WAV) and Text File. //Total Samples in AF1 (N) [Figure 4]
2. Convert the Audio File and Text File into a bitstreams.
3. Conceal an authorized algorithm between the sender and receiver to reveal that the audio file (AF1) contains hidden data, as well as the index value of the SAF1 in the audio file (AF2).
4. For i = 1 to N do // Continue the For loop to embed the Text File (TF).
5. If sum ((AF1) ~=TF) == 1, then // Compute the sum of binary bits difference between AF1 and TF. [Figure 5]
6. Embed TF in AF1 // Embed Text File (TF) in a computed sample of Audio File (AF1). [Figure 5]
7. End
8. End
9. Computation of the stego Audio File (SAF1 with hidden text file). [Figure 4]
10. End

The algorithm's performance is validated with the results of Mean Squared Error (MSE) and Peak Signal-to-Noise Ratio (PSNR) explained in the Results and Discussion section. The algorithm is also tested for steganalysis, proved through the results of histogram error rate (HER) and the fourth first moments. Furthermore, the algorithm undergoes a brute-force attack to evaluate its resistance.

2.1.3 Embedding index values of audio samples into an audio file

Each audio sample contains an index value for quick data retrieval, facilitating efficient database searches. The index assists the database application in efficiently searching data within an extensive database.

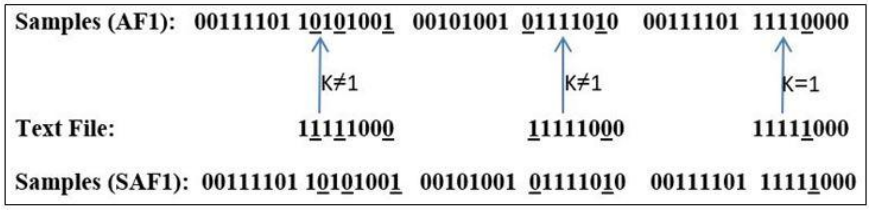


Figure 5: Example of embedding a text file in the audio file (AF1) when K = 1 is true.

An index of the stego audio file (SAF1) with hidden data is listed. The index values were initially in decimal form and then converted to binary. An uncompressed digital audio wave file (AF2) of two bytes is input for the LSB algorithm. In cryptographic processes, a private key (secret key) is essential for securely embedding and extracting data. This key is exclusive to authorized entities responsible for conducting steganographic communication. In the proposed BCS algorithm, the index value within the audio sample, at which the LSB algorithm initiates its execution, is called a private key (Ks) and can be changed for every communication, ensuring that each transmission remains isolated and secure. This algorithm selects K_s=50, indicating that the 50^{th} sample of AF2 is designated as the private key, mutually agreed upon by the sender and receiver for embedding. This random selection enhances security by creating an extensive combination space, with over 661504 samples across various audio files, as shown in Table 2. This results in approximately $2^{\wedge(661504)}$ possible combinations. The number of combinations makes it extremely difficult for an unauthorized party to

determine the private key, thereby strengthening the overall security of the transmission. The index values of the stego audio file (SAF1) with hidden text files are embedded into AF2 using the LSB algorithm. This process begins with K_s as the index value from AF2 used as the initial sample for embedding. The LSB algorithm, known for its simplicity and effectiveness, executes after embedding all text files in the AF1. The resulting stego audio files (SAF1 and SAF2) hide text file and index values, respectively. Security of the LSB algorithm increases with the dynamic index length and Ks. Figure 6 shows the flow diagram of the embedding Index value.

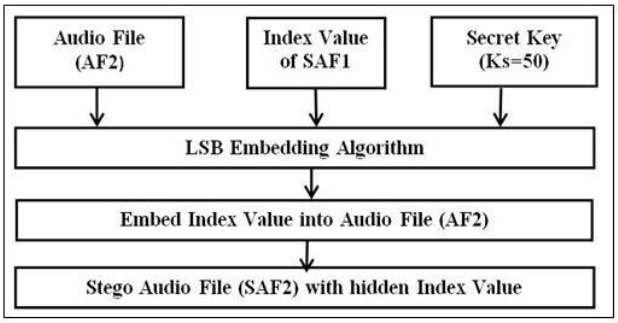


Figure 6: Flow diagram for embedding index values of SAF1 in the cover audio file (AF2).

2.1.4 Algorithm for Embedding Index Values of the Audio Samples in an Audio File

The sender side in Figure 3 illustrates the algorithm steps for clearer understanding.

Input: Stego Audio File (SAF1 with hidden text file) & Audio File (AF2)
Output: Stego Audio File (SAF2 with hidden index value)

1. Read the Audio file (AF2) and convert it into a stream of binary bits. [Figure 6]
2. List the stego audio samples (SAF1) index values containing a hidden text file. [Figure 6]
3. Convert the index values into a sequence of binary bitstreams.
4. Use the LSB embedding algorithm (replacing the least significant bit of each sample in an audio file (AF2) with the index value) to embed all index values into the audio file (AF2), starting from the Ks=50^{th} sample. [Figure 6]
5. Compute the stego audio file SAF2 with the hidden index value. [Figure 6]
6. Transmit both stego audio files SAF1 and SAF2 to the receiver.
7. End

The metrics Peak Signal-to-Noise Ratio and Bit Error Rate (BER) discussed in the Results and Discussion section verify the algorithm's performance.

2.2 Bit comparison and retrieval algorithm

The procedures associated with embedding and extracting multimedia objects are inherently inverse. The primary step in the proposed BCR algorithm is retrieving the index values representing stego audio file (SAF1) samples from the stego audio file (SAF2), with Ks serving as the reference. The Least Significant Bit retrieval algorithm retrieves the index values of the stego audio file (SAF1). The aforementioned indexed values correspond to the stego audio file (SAF1) containing the hidden text file information. The text file is then extracted from the retrieved index value. This process involves specifically extracting the eight consecutive bits from the Least Significant Bit of the previously mentioned index values of samples in SAF1. The complete hidden text files from SAF1 are retrieved by repeating the above process. The extracted text files are then converted into their decimal representations. Finally, the decimal values of the extracted text files are converted into their equivalent ASCII values. A flow diagram of the Bit comparison and Retrieval (BCR) algorithm is present in Figure 7.

2.2.1 BCR Algorithm-extracting index values and text file from audio files

The receiver side in Figure 3 illustrates the algorithm steps for clearer understanding.

Input: Stego Audio File (SAF1 with hidden text file) and Stego Audio File (SAF2 with hidden index value)
Output: Index value, Text File, and Audio Files (AF1 & AF2).

1. Retrieve the index values corresponding to samples in the stego audio file (SAF1) from the stego audio file (SAF2) using the LSB retrieval algorithm with Ks as the starting sample of SAF2. (Read the least significant bit of the stego audio file samples (SAF2)). [Figure 7]
2. Repeat Step 1 to retrieve all hidden index values.
3. Convert the retrieved index values into decimal representations.
4. Retrieve the text file from the stego audio file (SAF1) using the index value as the reference using the Bit comparison and retrieval algorithm. (Read the eight least significant bits of the stego audio file (SAF1)). [Figure 7]
5. Repeat Step 4 to retrieve all hidden characters in the text file from the referenced index values in Step 3.
6. Convert every byte of the retrieved file to decimal notations.
7. Finally, convert the decimal values into their corresponding ASCII characters to retrieve the text file.
8. End

The algorithm's efficiency is validated with the results of Bit Error Rate (BER) and Normalized Cross-Correlation (NCC), as explained in the Results and Discussion section.

2.3 Computational complexity

The computational complexity of the proposed algorithm was evaluated in three stages: Input processing, BCS (Bit Comparison and Substitution algorithm), and BCR (Bit Comparison and Retrieval algorithm). Execution times for each stage were noted by considering different text file sizes and audio file types. The algorithm was executed over 245 runs for each stage, with seven audio file types (Classical, Country, Disco, Hip-hop, Jazz, Pop, and Rock), each tested with seven different text file sizes (80, 400, 800, 1200, 1600, 2000, and 2400 bytes). To ensure an accurate analysis of the execution time for each stage, each combination of audio files and text file sizes was executed five times.

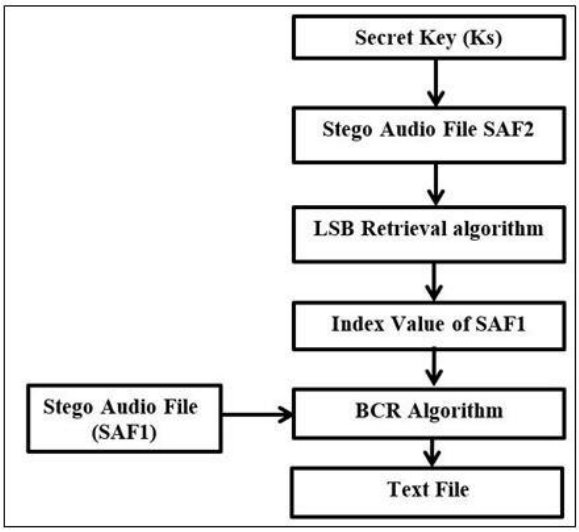


Figure 7: Flow diagram for extracting index values and text file from stego audio files (SAF2 and SAF1).

For input processing, execution times showed a nearly constant pattern, with minimum variations as the text file size increased, as shown in Figure 8(a). The observed pattern suggests an O (1) complexity for this stage. The BCS stage, in contrast, showed a linear increase in execution time with slight variations between audio file types, as shown in Figure 8(b). This linear increase indicates an O(TF) complexity in the BCS stage.

Similarly, the BCR stage showed linear growth in execution times notably lower compared to the BCS stage, as shown in Figure 8(c). The BCR's consistent linear scaling across audio file types further supports its O(TF) complexity. In summary, the algorithm maintains a computationally efficient structure with linear complexity (O(TF)) for both BCS and BCR stages, making it suitable for scalable steganographic applications in high-data environments.

Even though the execution time of the algorithm increases with larger audio files and text sizes, the reduced number of audio samples required for embedding larger text file compensates for this increase. The novel algorithm embeds each character into an audio sample and retrieves the character from the same sample. This approach optimizes memory usage and reduces processing time, ensuring efficiency even with increased text file inputs. The reduced memory usage balances the increased execution time.

2.4 Generalizability of the proposed method

The proposed algorithms have been applied successfully to both speech and ambient audio files, demonstrating their adaptability across different audio types beyond musical genres like Jazz and Rock. Tables 9 and 12 outline the specifications of the speech and ambient audio files. Concurrently, Tables 10 and 13 present the performance metrics for these files. These metrics demonstrate that the algorithms maintain efficient data embedding and retrieval capabilities across varied audio contexts. Tables 11 and 14 provide insights into the audio samples and associated difference bits when embedding text files in these audio files.

To further support the generalizability of the BCS algorithm across different audio files, we present histograms comparing the cover and stego audio files for speech and ambient audio files. These files include hidden text file containing 2400 characters, as shown in Figures 14 and 16. Figures 15 and 17 depict the audio files before and after embedding 2400-byte text file. They show no significant degradation in audio quality.

No significant technical barriers emerged when applying the novel BCS and BCR algorithms to diverse audio types. However, the algorithm's performance depends on specific characteristics of the audio samples and the bit patterns within them. Despite this, the embedding method and the resultant error bit remain consistent across different audio file types.

3 Results and discussion

The MATLAB software was used to evaluate the BCS and BCR algorithms using seven distinct audio files representing genres: Classical, Disco, Hiphop, Country, Jazz, Pop, and Rock. The GTZAN dataset was selected for its genre coverage (1,000 clips of 30 seconds each), compact size, and consistent 22050 Hz sampling rate, making it suitable for efficient embedding without extensive preprocessing. The consistent sampling rate ensures uniform audio quality across the dataset, minimizing variability that could affect embedding or extraction performance. Additionally, the dataset's compact size allows for quick processing and repeated testing, enabling efficient experimentation with the steganography algorithm. Its genre diversity provides a variety of audio contexts to test the adaptability of the algorithm.

Furthermore, its open availability promotes reproducibility. Using unprocessed audio minimizes computation time and accurately reflects performance conditions for the steganography algorithms. Table 2 details the specifications of the audio files from GTZAN.

The experimentation was conducted on a Windows 10 laptop with a Core i3 processor at 2.0 GHz and 4 GB of RAM. Uncompressed audio files from the GTZAN

dataset served as the cover audio [1],[33],[35],[36]. The text files used in this study are from the Reuters-21578 collection, based on the 1994 CIA World Factbook, containing Prolog facts about LandBoundaries, NaturalResources, Population, Capital, and more [37]. The evaluation employed MSE, PSNR, BER, and NCC to quantify deviations in the audio files and measure the efficiency of the BCS and BCR algorithms.

3.1 Similarity analysis

3.1.1 Mean squared error (MSE)

MSE measures audio distortion, calculated as the average square difference between the cover and stego audio files using Equation (1)[33]. If the MSE is zero, the files are similar.

$$\mathrm{MSE} = \frac{1}{Q}\sum_{i=1}^{Q}(s1_i - s2_i)^2 \qquad (1)$$

$s1_i$ and $s2_i$ represent samples from the cover and stego audio file, respectively, and Q is the total number of audio samples.

Figure 9(a) represents average MSEs over 49 runs, and their standard deviations. The X-axis represents the text file's length in Bytes, and the Y-axis indicates the MSE. The proposed algorithm achieves an MSE of

Table 2: Description of audio files.

Bits per sample/Sample rate	16/22050Hz
Number of Samples	661504-664180
Channel and Audio Type	Mono and Music
Duration in seconds	1-30

0.0513, indicating minimal distortion. This MSE outperforms existing algorithms [28], [29], which recorded MSE values of 0.5619 and 9.08875 respectively. Table 3 presents the MSE values of both the proposed and existing algorithms, along with the significant differences between these values. This difference results from maintaining a 1-bit difference between the audio and the stego audio samples when embedding one byte of text file. In the proposed BCS algorithm, embedding 2400 bytes (19200bits) of text files results in 2400 error bits, which is successfully accomplished. This controlled one bit difference minimizes distortions, especially when embedding large amounts of data for audio steganographic communication.

The standard deviation of MSE values is 0 across all text files and audio types. This consistency indicates a stable algorithm with a fixed, predictable distortion level in preserving audio quality. The same MSE value is observed across an audio file for different text file sizes, as shown in Figure 9(a), with values such as 0.0031 for 'Classical' audio across text file sizes ranging from 80 bytes to 2400 bytes. This uniformity occurs because the embedding process introduces minimal distortion, and the MSE is calculated as an average, leading to identical values and a standard deviation of zero.

Statistical analysis yields a p-value of 0 ($p < 0.05$), which rejects the null hypothesis and confirms significant differences in MSE across audio files. This analysis suggests that the MSE varies with the bit patterns in the audio samples. Despite these variations, low and consistent MSE values demonstrate that the audio steganography method effectively controls distortion. As a result, the audio quality remains largely unaffected during data embedding.

3.1.2 Peak Signal-to-Noise Ratio (PSNR)

PSNR evaluates audio quality in decibels (dB) using Equation (2). A higher PSNR with lower MSE indicates better steganographic quality.

$$\mathrm{PSNR} = 10\log_{10}\frac{(2^q-1)^2}{\mathrm{MSE}} \qquad (2)$$

Where q is the maximum bits per signal sample, and MSE is the Mean Squared Error [33].

Figure 9(b) shows the PSNR results over 49 runs for audio clips with different embedding rates: 80B, 400B, 800B, 1200B, 1600B, 2000B, and 2400B. The X-axis indicates the length of text files in bytes, and the Y-axis indicates PSNR. The average PSNR for stego audio (SAF1) is 41.13732041 dB, while for stego audio (SAF2), it is 32.4868 dB. Table 3 compares the PSNR of the proposed BCS algorithm with that of the existing

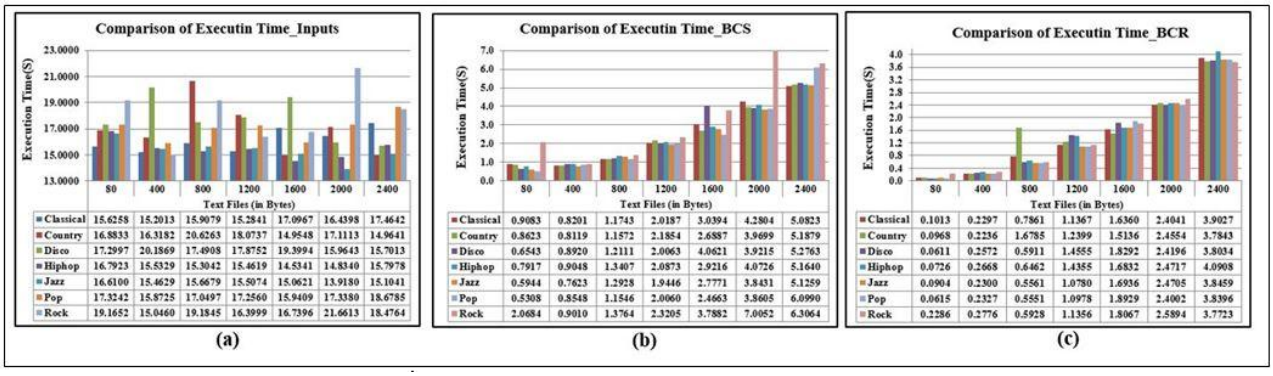


Figure 8: Execution time for (a) Input (b) BCS (c) BCR.

algorithm[29], [30], highlighting the significant increase. For JAZZ audio with 2400 bytes, PSNR is 36.4463 dB, while [29] yields 34.2775 dB and [30] yields 17.71916498 dB.

The average PSNR of 41.13732041 dB correlates with the near-zero MSE (0.0513). This correlation arises from the logarithmic relationship between MSE and PSNR. The low MSE achieved by the proposed embedding method reflects minimal distortion and leads to a high PSNR value.

The observed range of standard deviation values for PSNR is 4.85 to 5.16 across different audio file types and text file sizes, as shown in Figure 9(b). These values indicate that the algorithm maintains audio quality with some variability in how the embedding process affects each type of audio file. Specifically, the standard deviation of Country audio files is 4.85, while Classical and Hip-hop exhibit 5.16. The error bars in Figure 10(a) represent the standard deviation and show the variation in average PSNR values across different audio files and text file sizes. This results in the spread of PSNR values and the reflection of the range of impact on audio quality.

The moderate standard deviation across all audio file types indicates a balanced trade-off between embedding efficiency and audio quality preservation, indicating reliable algorithm performance.

Figure 11(b) illustrates the cover and stego audio files for various genres, with the stego audio embedded with a text file of 2400 bytes.

Remarkably, no noticeable perceptual changes are observed between the cover audio sample (AF1) and the corresponding stego audio sample (SAF1), as modifications affect only one of the 16 bits per sample to embed a one-byte text file.

3.2 Hiding capacity (payload)

Hiding capacity (HC) represents the maximum size of a text file concealed in an audio file and is calculated by Equation (3) as a standard percentage [1].

$$\mathrm{HC} = \left(\frac{\text{Text File size}}{\text{Audio File size}}\right) * 100 \qquad (3)$$

Similarly, for the LSB algorithm, the hiding capacity is measured by Equation (4)[1].

$$\mathrm{HC} = \left(\frac{\text{number of samples}}{8}\right) * 100 \qquad (4)$$

The tested Jazz audio file with 661794 samples embeds 2400 bytes of text files using 19200 samples for the LSB algorithm and 2400 audio samples for the BCS algorithm. Table 4 illustrates the differences in audio samples used at different embedding capacities between the traditional LSB and the proposed BCS algorithm for a Classical audio file with 661794 samples.

The BCS algorithm effectively embeds text data by comparing each byte of the text file with audio samples to find those differing by one bit. The identified one-bit difference facilitates data embedding while preserving audio integrity. This approach achieves a 1:1 ratio, embedding a single character (8 bits) into one audio sample (16 bits). The BCS algorithm successfully embedded 80 bytes of text (80 characters) using 80 audio samples.

Conversely, the Least Significant Bit algorithm embeds each 8-bit character from the text file into the least significant bits of eight corresponding audio samples, resulting in eight audio samples for each embedded character. The discrepancy in Table 4 arises from these differing methodologies. The LSB algorithm consistently requires eight audio samples per character, and the BCS algorithm uses only one audio sample per character based on the one-bit difference criterion. Despite the differing embedding rates, the BCS method maintains consistent sample utilization by embedding each character into an audio sample with the required bit difference. This approach allows for more efficient sample utilization.

Moreover, the proposed BCS algorithm aligns sample usage with the text file size, effectively increasing embedding capacity and rate while reducing the bit error rate.

By selecting audio samples with a one-bit difference between text and audio, the BCS algorithm enhances the robustness and security of audio steganographic communication. Using fewer samples improves efficiency by minimizing processing time and enhances imperceptibility by reducing noticeable audio changes.

3.2.1 Embedding rate

The embedding rate is the number of bits concealed in a one-byte audio sample. The LSB algorithm embeds 1 byte of the text file in 8 audio samples (1 byte) with an embedding rate of 12.5%. In contrast, in the BCS algorithm, 1 byte of the text file is embedded into 1 byte of a 2-byte audio sample with a 50% embedding rate. For the tested JAZZ audio with 661794 samples, the proposed algorithm achieves an increase of 37.5% in embedding rate. Hence, the BCS algorithm has significantly improved its embedding rates compared to the traditional Least Significant Bit.

3.3 Strength analysis

3.3.1 Bit error rate

Equation (5) calculates the Bit Error Rate (BER), representing the percentage of incorrect bits in retrieved text files.

$$BER = \frac{Q_{error}}{Q_{bits}} \qquad (5)$$

Where Q_{error} is the number of differing bits, and Q_{bits} is the total number of bits embedded in the audio file [1]. A lower BER indicates robust steganographic algorithms. The proposed BCR algorithm successfully extracts the original text file for all the tested audio files.

Figure 10(b) illustrates the BER results for the proposed algorithm over 49 runs, where the X-axis

represents the length of the text file in bytes, and the Y-axis represents the corresponding BER values. The Proposed algorithm computes a BER of 0.00000753 for a country WAV file of 80 bytes and 0.000226 for a rock WAV file with 2400 bytes. Similarly, the average BER of a stego audio file (SAF2) is 0.0009. The BER value illustrates the proposed algorithm's robustness in security and accuracy in maintaining consistency in both the audio and the stego audio file.

The average BER of the proposed algorithm is 0.000114, outperforming results from previous research [29], [31], with specific increase values as shown in Table 5.

This low BER results from altering only 1 bit out of 16 (1/16) in a 16-bit audio file sample (AF1). Specifically, each 8-bit character in the text file is embedded into a corresponding 16-bit audio sample with a controlled one-bit difference between the cover audio file (AF1) and the stego audio file (SAF1) samples.

By embedding 2400 characters (2400 * 8 = 19200 bits), the BCS algorithm achieved a total controlled bit difference of 2400 bits, resulting in lower BER.

This consistent approach limits the variability in the audio, maintaining audio quality across various genres such as Classical, Disco, Hip-hop, Country, Jazz, Pop, and Rock, thus preserving a satisfactory listening experience. Table 4 conveys the differences in bits used in both algorithms, illustrating the efficiency of the BCS method. In contrast, the conventional LSB algorithm modifies the least significant bits without a predefined or controlled pattern, as the changes depend on the embedded data and the audio samples. This results in unpredictable bit alterations, higher BER variability, increased error rates, and reduced performance due to the ineffective management of bit differences. Conversely, the structured embedding process in the BCS algorithm increases the robustness of security by reducing errors and optimizing the use of embedding samples, thereby supporting claims of improved performance over traditional methods.

The observed standard deviation in the Bit Error Rate was low, approximately 7.9E-05, as shown in Figure 10(b), across all file sizes. The standard deviation value indicates minimal variation in the data embedding process, influenced by the audio files. The algorithm reliably preserves both data and audio quality. Error bars for average BER values in Figure 11(a) show the range of variability in BER. The error bars represent the standard deviation of the average BER values across different audio files and text file sizes.

Statistical analysis yields a p-value of 0.999999986 ($p < 0.05$), indicating no significant differences in BER across audio files. The above analysis suggests that the BER of the algorithm's performance remains consistent, as shown in Figure 10(b), regardless of the audio file and embedding capacity used. The audio steganography method effectively maintains high transparency and robustness in data embedding without significantly affecting the cover audio.

3.3.2 Normalized cross-correlation

Similarity is measured by NCC, with the formula for calculating the similarity between embedded and retrieved text files given by Equation (6)

$$\mathrm{NCC}(\mathrm{M},\mathrm{M}') = \frac{\sum_{k=1}^{P} \mathrm{M}(\mathrm{K})\,\mathrm{M}'(\mathrm{K})}{\sqrt{\sum_{K=1}^{P} \mathrm{M}(\mathrm{K})^2}\sqrt{\sum_{K=1}^{P} \mathrm{M}'(\mathrm{K})^2}} \qquad (6)$$

Table 3: The MSE and PSNR Comparison between Proposed and Existing Algorithms

Method	Text Files (Characters)	MSE	PSNR (dB)
[28] Normal LSB Method	250	0.3994 ▲87.16%	-
[28] Pattern Matching Method	250	0.5619 ▲ 90.87%	-
[29]	107	9.08875 ▲99.44%	34.2775 ▲20.01%
[30]	103	-	17.71916498 ▲132.163%
Proposed	**2400**	**0.0513**	**41.13732041**

Table 4: Audio samples utilized in the proposed algorithm compared with the traditional LSB algorithm.

Length of Text File (in Bytes)	Audio Samples Utilized (ASU) (in samples)		Difference Bit (DB) (in Bits)	
	LSB	**Proposed**	LSB	**Proposed**
80	640	**80**	341	**80**
400	3200	**400**	1541	**400**
800	6400	**800**	3185	**800**
1200	9600	**1200**	4840	**1200**
1600	12800	**1600**	6433	**1600**
2000	16000	**2000**	8043	**2000**
2400	19200	**2400**	9586	**2400**

Table 5: Comparisons of BER between the proposed and existing algorithms

Method	Length of Text File (in Bytes)	BER
[29]	107	17.995% ▲99.94%
[31]	532	0.0187% ▲39.04%
Proposed	**2400**	**0.000114**

Where M and M′represent the embedded and extracted text files, respectively, and P is the number of samples in the audio file [36]. For all tested audio files, an NCC value of 1 indicates that the proposed BCR algorithm successfully extracts the complete text files. Similarly, the cover audio and stego audio files are similar, as shown in Figure 11(b).

Table 6 compares the NCC results obtained using the proposed algorithm with those of existing algorithms reported by researchers in their studies [32], along with the significant differences. The proposed algorithm's NCC values surpassed the existing research, highlighting its superior performance. This performance is due to one bit of difference between the computed audio file and the embedded character. Similarly, the continuity in the pattern of the audio file between the cover and stego audio files is maintained, resulting in a Normalized Cross-Correlation value of 1.

3.4 Steganalysis

3.4.1 Histogram attack

Histogram attacks were examined through forty-nine experiments across seven audio files and text files in various sizes from 80 Bytes to 2400 Bytes. Histogram Error Rate (HER) measures the error between the audio file and the stego file using Equation (7) [33].

$$\mathrm{HER} = \frac{\sum_{i=1}^{Q}(\mathrm{His_a} - \mathrm{His_s})^2}{\sum_{i=1}^{Q}\mathrm{His_a}^2} \qquad (7)$$

Where His_a and His_s are the histograms of audio and stego audio files, respectively, and Q is the total audio samples.

The proposed algorithm demonstrates improved histogram error rates, achieving a HER of 0.0000000043.

Table 6: Comparison of NCC between the proposed and existing algorithms.

Method	NCC
[32]	0.99992 ▲0.008%
Proposed	**1**

The experimental HER value approaches zero, primarily due to the uniform probability distribution of amplitudes within the audio files, confirming its resistance to histogram attacks. Figure 11(c) illustrates the histograms for the seven audio files before and after embedding a text file of size 2400 bytes.

3.4.2 Fourth first moments

The difference between the audio file and the stego file was analysed statistically using the fourth first moments, namely, mean (μ), variance (σ^2), skewness (sk), and kurtosis (k). Equations (8), (9), (10), and (11) calculate the fourth first moments [33]. Equation (12) calculates the absolute difference between these moments, yielding the Difference Ratio (DR) [36].

The experimental results of the DR values are close to 0 in Table 7, proving that the proposed algorithm is resistant to statistical investigation.

Moreover, the audio file's Difference Ratio (DR) approaches zero because of its symmetry, as indicated by near-zero values for both mean and skewness. A kurtosis value of zero further confirms that the amplitude distribution of the stego audio file, containing embedded text files, follows a normal distribution, leading to the near-zero DR.

$$\mu = \frac{\sum_{i=1}^{Q} S_i}{Q} \qquad (8)$$

$$\sigma^2 = \frac{\sum_{i=1}^{Q}(S_i - \mu)^2}{(Q-1)} \qquad (9)$$

$$\mathrm{sk} = \frac{\sum_{i=1}^{Q}(S_i - \mu)^3}{(Q-1)\sigma^3} \qquad (10)$$

$$\mathrm{k} = \frac{\sum_{i=1}^{Q}(S_i - \mu)^4}{(Q-1)\sigma^4} \qquad (11)$$

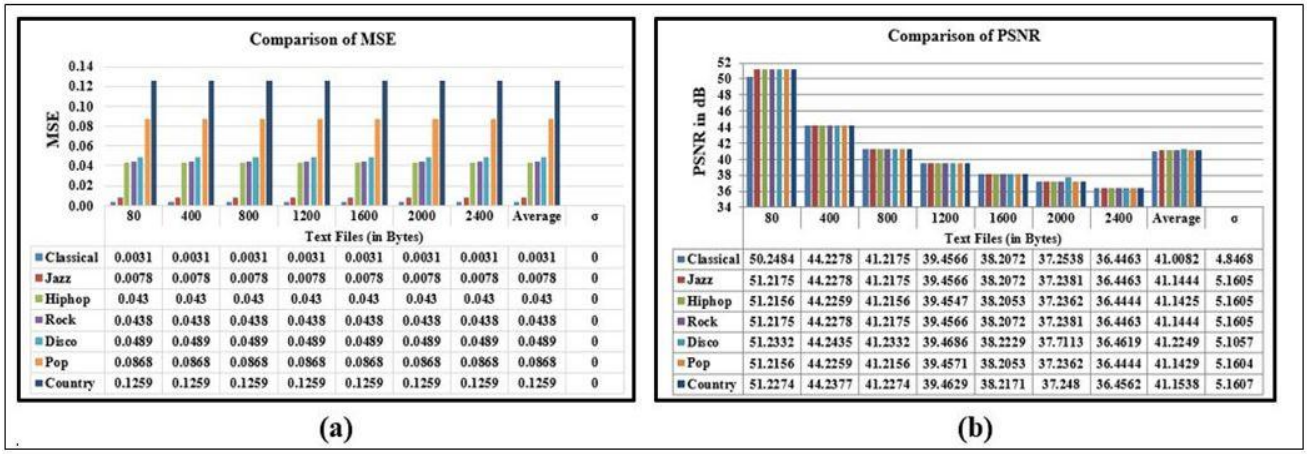
$$\text{Difference Ratio(DR)} = \left|\frac{\mathrm{fm_a} - \mathrm{fm_s}}{\mathrm{fm_a}}\right| \times 100 \qquad (12)$$

Where fma and fms are any fourth first moments of the audio file and the stego audio file, respectively.

S_i denotes the Samples in the cover or stego audio files, and Q represents the total number of audio samples in the cover or stego audio files.

3.4.3 Brute-Force attack

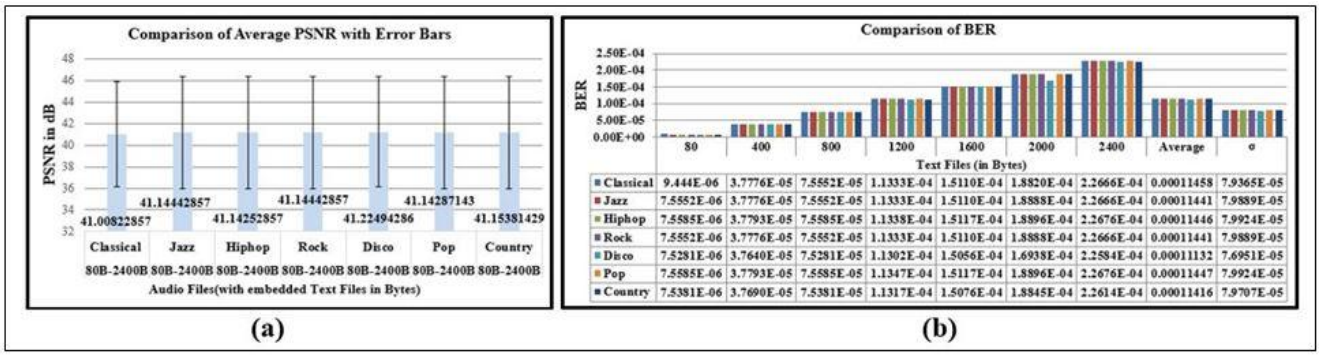
The objective of a brute force attack is to attempt multiple methods or keys to uncover the hidden text files from stego audio files. This type of attack is commonly used to test the strength and effectiveness of the steganographic algorithms. The total number of possible sample combinations in an audio file against brute-force attacks on the audio steganographic system is given by $2^{(b*N)}$, where b is the number of bits used for embedding per audio sample (8 bits per sample), and N is the total number of audio samples. This approach derives from cryptographic analyses of key space sizes. Table 8 compares the proposed algorithm with the LSB algorithm under a brute force attack.



MSE	80	400	800	1200	1600	2000	2400	Average	σ
Classical	0.0031	0.0031	0.0031	0.0031	0.0031	0.0031	0.0031	0.0031	0
Jazz	0.0078	0.0078	0.0078	0.0078	0.0078	0.0078	0.0078	0.0078	0
Hiphop	0.043	0.043	0.043	0.043	0.043	0.043	0.043	0.043	0
Rock	0.0438	0.0438	0.0438	0.0438	0.0438	0.0438	0.0438	0.0438	0
Disco	0.0489	0.0489	0.0489	0.0489	0.0489	0.0489	0.0489	0.0489	0
Pop	0.0868	0.0868	0.0868	0.0868	0.0868	0.0868	0.0868	0.0868	0
Country	0.1259	0.1259	0.1259	0.1259	0.1259	0.1259	0.1259	0.1259	0

PSNR in dB	80	400	800	1200	1600	2000	2400	Average	σ
Classical	50.2484	44.2278	41.2175	39.4566	38.2072	37.2538	36.4463	41.0082	4.8468
Jazz	51.2175	44.2278	41.2175	39.4566	38.2072	37.2381	36.4463	41.1444	5.1605
Hiphop	51.2156	44.2259	41.2156	39.4547	38.2053	37.2362	36.4444	41.1425	5.1605
Rock	51.2175	44.2278	41.2175	39.4566	38.2072	37.2381	36.4463	41.1444	5.1605
Disco	51.2332	44.2435	41.2332	39.4686	38.2229	37.7113	36.4619	41.2249	5.1057
Pop	51.2156	44.2259	41.2156	39.4571	38.2053	37.2362	36.4444	41.1429	5.1604
Country	51.2274	44.2377	41.2274	39.4629	38.2171	37.248	36.4562	41.1538	5.1607

Figure 9: Influence of (a) MSE, and (b) PSNR.



BER	80	400	800	1200	1600	2000	2400	Average	σ
Classical	9.44E-06	3.7776E-05	7.5552E-05	1.1333E-04	1.5110E-04	1.8820E-04	2.2666E-04	0.00011458	7.9365E-05
Jazz	7.5552E-06	3.7776E-05	7.5552E-05	1.1333E-04	1.5110E-04	1.8888E-04	2.2666E-04	0.00011441	7.9889E-05
Hiphop	7.5585E-06	3.7793E-05	7.5585E-05	1.1338E-04	1.5117E-04	1.8896E-04	2.2676E-04	0.00011446	7.9924E-05
Rock	7.5552E-06	3.7776E-05	7.5552E-05	1.1333E-04	1.5110E-04	1.8888E-04	2.2666E-04	0.00011441	7.9889E-05
Disco	7.5281E-06	3.7640E-05	7.5281E-05	1.1302E-04	1.5056E-04	1.6938E-04	2.2584E-04	0.00011132	7.6951E-05
Pop	7.5585E-06	3.7793E-05	7.5585E-05	1.1347E-04	1.5117E-04	1.8896E-04	2.2676E-04	0.00011447	7.9924E-05
Country	7.5381E-06	3.7690E-05	7.5381E-05	1.1317E-04	1.5076E-04	1.8845E-04	2.2614E-04	0.00011416	7.9707E-05

Figure 10: Influence of (a) Average PSNR with Error Bars, and (b) BER.

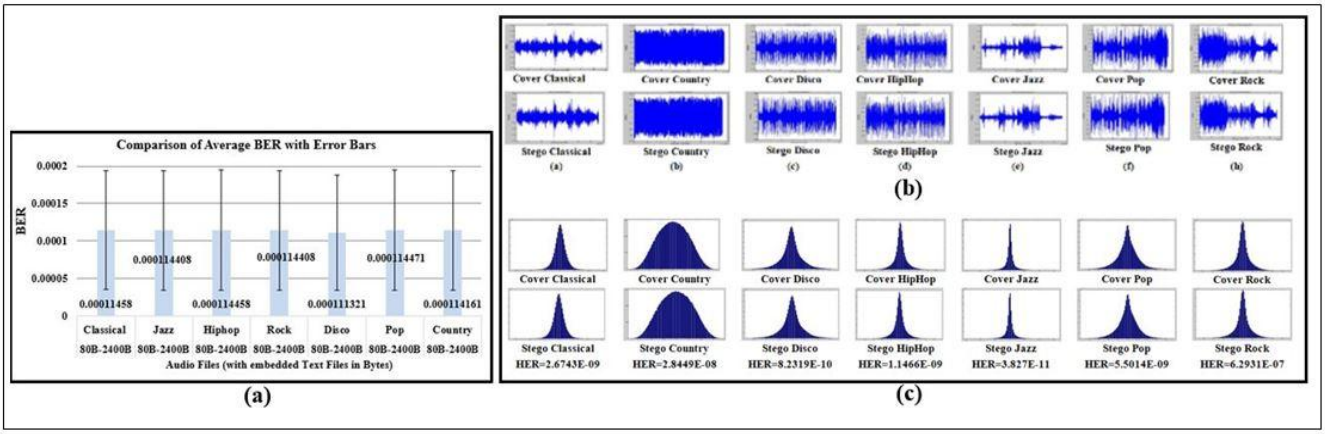


Figure 11: (a) Influence of average BER with error bars, (b) Comparison of cover and stego audio files with 2400 bytes of embedding capacity, and (c) Histograms of cover and stego audio files with 2400 bytes of embedding capacity.

Table 7: Comparison of DR for the Fourth First Moments and HER values for various cover audios using 2400B of embedded Text Files.

Audio File	Difference Ratio (DR)				HER
	1st moment, Mean (μ)	2nd moment, Variance(σ^2)	3rd moment, Skewness(sk)	4th moment, Kurtosis(k)	
[33]	0.0799 ▲99.29%	0.0008 ▲97%	0.0018 ▲95.27%	0.0028 ▲98.68%	0.1278 ▲100%
[34]	0.2304 ▲99.75%	0 ≈	0.0004 ▲78.75%	0.0005 ▲92.6%	1.2274E-04 ▲100%
Proposed	**0.00057**	**0.000024**	**0.000085**	**0.000037**	**4.33E-09**

Table 8: Compare the proposed algorithm with the LSB algorithm under brute force attack.

Audio File			Embedded Bits (b)	Total Combinations $2^{\wedge(b*N)}$	
Name	Samples (N)	Bits		Proposed Algorithm	LSB Algorithm
Classical	661794	16	8	$2^{\wedge(8*661794)}=2^{5294352}$	$2^{\wedge(1*661794)}=2^{661794}$
Country	663300	16	8	$2^{\wedge(8*663300)}=2^{5306400}$	$2^{\wedge(1*663300)}=2^{663300}$
Disco	664180	16	8	$2^{\wedge(8*664180)}=2^{5313440}$	$2^{\wedge(1*664180)}=2^{664180}$
Hip Hop	661504	16	8	$2^{\wedge(8*661504)}=2^{5292032}$	$2^{\wedge(1*661504)}=2^{661504}$
Jazz	661794	16	8	$2^{\wedge(8*661794)}=2^{5294352}$	$2^{\wedge(1*661794)}=2^{661794}$
Pop	661504	16	8	$2^{\wedge(8*661504)}=2^{5292032}$	$2^{\wedge(1*661504)}=2^{661504}$
Rock	661794	16	8	$2^{\wedge(8*661794)}=2^{5294352}$	$2^{\wedge(1*661794)}=2^{661794}$

The large number of sample combinations in the proposed methods indicates that exhaustive search methods are impractical, demonstrating high resistance to brute-force attacks. When comparing the robustness of the BCS and BCR algorithms with previous methods, it is evident that the innovative embedding techniques employed in these algorithms significantly increase resistance to brute-force attacks. Embedding multiple bits per sample in the proposed algorithms, as compared to a single bit in the LSB algorithm, increases the number of embedding locations and possible sample combinations, making detection more difficult.

Figure 12 illustrates a comprehensive comparison chart of performance metrics between the existing and proposed algorithms. The chart plots performance metric values on the y-axis for both the proposed and existing algorithms with performance metric names listed on the x-axis. The proposed algorithm aims for data concealment where the audio samples used should match the differing bits between the cover and stego audio, as demonstrated in Table 4 and Figure 13. The proposed BCS and BCR algorithms outperform existing ones with their experimental results, highlighting the significance of BER.

A Research gap and addressed novelty

The research gap in existing audio steganography methods lies in the limitations of the traditional Least Significant Bit algorithm. In the LSB algorithm, embedding one character from a text file requires eight audio samples, resulting in an embedding rate of only 12.5%. Using the 2 LSBs to embed a character reduces this to four audio samples, achieving a 25% embedding rate. However, as the hiding capacity increases, the number of audio samples required also increases, highlighting the need for an algorithm that achieves higher embedding rates while minimizing the use of audio samples. Additionally, the BER in LSB methods is highly variable, depending on the characteristics of the audio file, leading to unpredictable error rates that may either be high or low. As embedding rates increase, BER increases, further impacting data integrity.

This variability emphasizes the need for an algorithm that maintains a consistent and predictable BER, regardless of the embedding rate and audio file characteristics.

The novelty of the proposed algorithm lies in balancing embedding capacity and BER by embedding a character (8 bits) from a text file within a single audio sample and ensuring that the bit difference between the samples in cover audio and the stego audio file is limited to one. The proposed method reduces the required audio samples while achieving a controlled error pattern.

B Strength of the proposed work

The proposed work embeds an entire character (8 bits) within a single audio sample, maintaining a 1:1 ratio between characters in the text file and audio samples.

A comparison methodology involves embedding text files into the audio samples. Hence the algorithm is resistant to statistical and brute-force attacks, which enhances the security and robustness of the algorithm. This mechanism results in a controlled one-bit difference between the sample of a cover audio file and the corresponding embedded sample in the stego audio file. This approach preserves the integrity of the audio file. Additionally, the algorithm predicts error bits between the cover and stego files, maintaining an effective trade-off between embedding capacity and error rate.

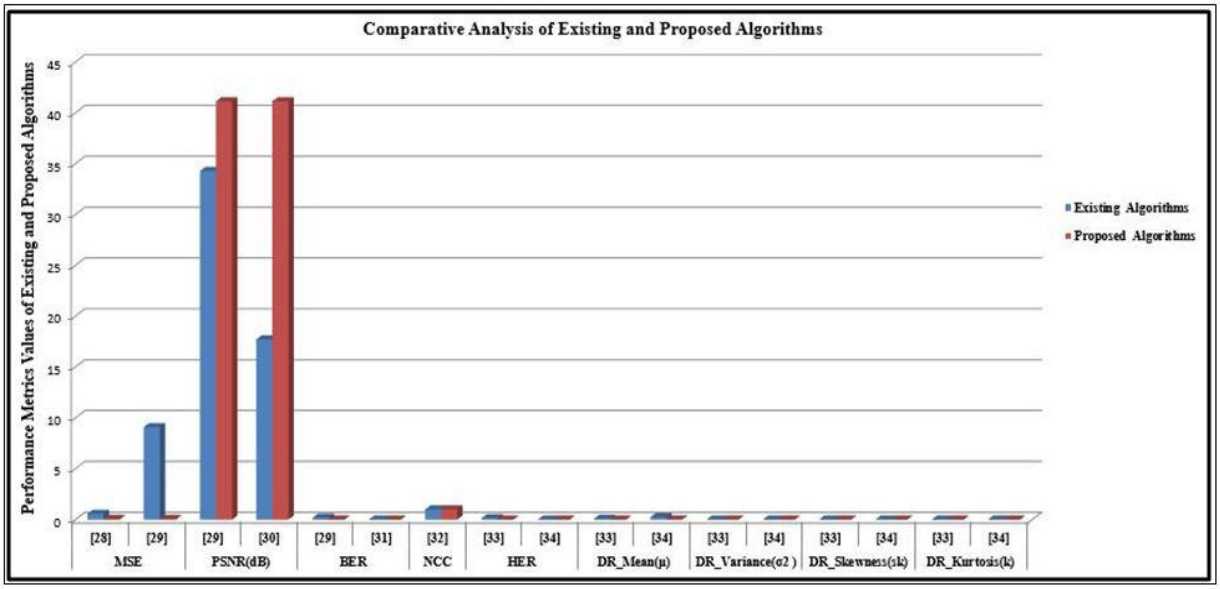


Figure 12: Comparative analysis of proposed and existing algorithms.

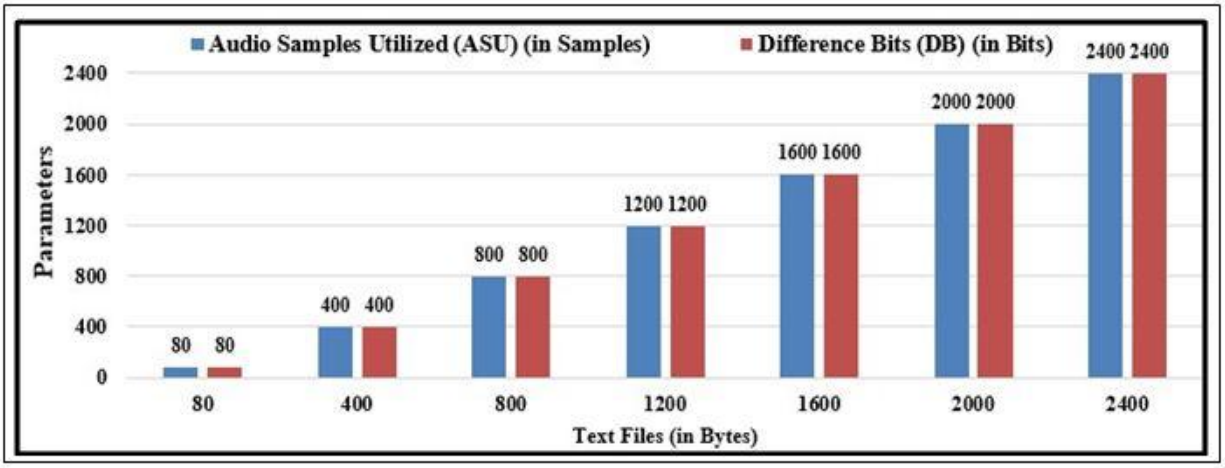


Figure 13: Influence of parameters of audio files versus hiding capacity.

This capability optimizes data hiding by minimizing audio quality degradation, addressing challenges in audio steganography.

4 Conclusion and future work

The research presented BCS and BCR algorithms to enhance the Least Significant Bit Algorithm in Audio Steganography. The BCS algorithm embeds each byte of the text file into the least significant 8 bits (one byte) of a two-byte audio sample from AF1, and the BCR algorithm retrieves the embedded data. Simultaneously, the Least Significant Bit algorithm embeds the index value of the samples from the Stego Audio File (SAF1) with hidden text file into the Audio File (AF2). Moreover, the proposed method was analyzed through MSE and PSNR similarity tests, yielding an MSE of 0.0513 and an average PSNR of 41.13732041 dB, thereby outperforming the existing algorithm. Furthermore, there is a 37.5% enhancement in the embedding rate compared to the conventional LSB algorithm. The stego signal is evaluated using histogram distribution, yielding an average HER of 4.33E-09. Specifically, the proposed algorithm's comparative methodology yielded identical values with an embedded text file of 2400 characters (2400 bytes) in 2400 audio samples, resulting in a 2400-bit difference between the cover audio file and the stego file. For future research, semiconductors could implement real-time audio steganographic systems.

Acknowledgement

The authors thank the anonymous reviewers for their insightful feedback. Their suggestions significantly improve the quality of this research article.

Data availability statement

Music audio files:
https://www.kaggle.com/datasets/andradaolteanu/gtzan-dataset-music-genre-classification
Text File:
http://www.daviddlewis.com/resources/testcollections/reuters21578/feldman-cia-worldfactbook-data.txt
Speech Audio Files:

https://www.kaggle.com/datasets/lnicalo/gtzan-musicspeech-collection?select=speech_wav
Ambient Audio Files:
https://mixkit.co/free-sound-effects/

References

[1] M. M. Mahmoud, and H. T. Elshoush.Enhancing LSB Using Binary Message Size Encoding for High Capacity, Transparent and Secure Audio Steganography-An Innovative Approach.IEEE Access, 10:29954–29971, 2022. https://doi.org/10.1109/ACCESS.2022.3155146

[2] S. A. S. Almola, N. H. Qasim, and H. A. A. Alasadi.Robust Method for Embedding an Image Inside Cover Image Based on Least Significant Bit Steganography.Informatica, 46(9):53–60,2022. https://doi.org/10.31449/inf.v46i9.4362

[3] F. R. Shareef.A novel crypto technique based ciphertext shifting.Egyptian Informatics Journal,21(2):83–90,2020. https://doi.org/10.1016/j.eij.2019.11.002

[4] S. A. S. Almola.Hiding Images in the Spatial Domain.Informatica,46(8):165–174,2022. https://doi.org/10.31449/inf.v46i8.4252

[5] Yingnan Zhang, Minqing Zhang ,Xu An Wang,Ke Niu, and Jia Liu.A Novel Video Steganography Algorithm Based on Trailing Coefficients for H.264/AVC. Informatica, 40(1):63–70,2016. https://www.informatica.si/index.php/informatica/article/view/1095/824

[6] M. Bazyar, and R. Sudirman.A new method to increase the capacity of audio steganography based on the LSB algorithm.Jurnal Teknologi (Sciences & Engineering), 74(6): 49–53, 2015. https://doi.org/10.11113/jt.v74.4667

[7] W. Fraczek, W. Mazurczyk, and K. Szczypiorski.Multilevel steganography: Improving hidden communication in networks.Journal of Universal Computer Science, 18(14):1967–1986, 2012. https://doi.org/10.3217/jucs-018-14-1967

[8] A. A. Alsabhany, A. H. Ali, F. Ridzuan, A. H. Azni, and M. R. Mokhtar.Digital audio steganography: Systematic review, classification, and analysis of the current state of the art.Computer Science Review, 38:100316,2020. https://doi.org/10.1016/j.cosrev.2020.100316

[9] A. Tahseen Suhail, and H. Ghanim Ayoub. A new method for hiding a secret file in several WAV files depends on circular secret key.Egyptian Informatics Journal, 23(4):33–43, 2022. https://doi.org/10.1016/j.eij.2022.06.003

[10] Adeboje Olawale Timothy, Adetunmbi Adebayo Olusola, and Gabriel Arome Junior. Embedding Text in Audio Steganography System using Advanced Encryption Standard, Text Compression and Spread Spectrum Techniques in Mp3 and Mp4 File Formats.International Journal of Computer Applications.177(41):46–51, 2020. https://doi.org/10.5120/ijca2020919914

[11] L. Chen, R. Wang, D. Yan, and J. Wang. Learning to Generate Steganographic Cover for Audio Steganography Using GAN.IEEE Access.9:88098–88107,2021. https://doi.org/10.1109/ACCESS.2021.3090445

[12] W. Al-chaab, Z. A. Abduljabbar, E. W. Abood, V. O. Nyangaresi, Hussein M. Mohammed , and Junchao Ma.Secure and Low-Complexity Medical Image Exchange Based on Compressive Sensing and LSB Audio Steganography.Informatica,47(6): 65–74,2023. https://doi.org/10.31449/inf.v47i6.4628

[13] Z. N. Sultani, and B. N. Dhannoon.Image and audio steganography based on indirect LSB.Kuwait Journal of Science,48(4):1–12, 2021. https://doi.org/10.48129/kjs.v48i4.8992

[14] Y. Yang, H. Yu, X. Zhao, and X. Yi.An adaptive double-layered embedding scheme for MP3 steganography. IEEE Signal Processing Letters,27:1984–1988,2020. https://doi.org/10.1109/LSP.2020.3032875

[15] K. Chen, H. Zhou, W. Li, K. Yang, W. Zhang, and N. Yu.Derivative-Based Steganographic Distortion and its Non-additive Extensions for Audio.IEEE Transactions on Circuits and Systems for Video Technology,30(7):2027–2032,2020. https://doi.org/10.1109/TCSVT.2019.2918511

[16] Y. Yang, X. Yi, X. Zhao, and J. Zhang.A fast and secure MP3 steganographic scheme with multi-domain.Signal Processing,190:108332, 2022. https://doi.org/10.1016/j.sigpro.2021.108332

[17] Y. Ren, S. Cai, and L. Wang.Secure AAC steganography scheme based on multi-view statistical distortion (SofMvD). Journal of Information Security and Applications, 59:102863,2021. https://doi.org/10.1016/j.jisa.2021.102863

[18] K. Ying, R. Wang, Y. Lin, and D. Yan.Adaptive Audio Steganography Based on Improved Syndrome-Trellis Codes.IEEE Access,9:11705–11715,2021. https://doi.org/10.1109/ACCESS.2021.3050004

[19] F. Djebbar, B. Ayad, K. A. Meraim, and H. Hamam.Comparative study of digital audio steganography techniques.EURASIP Journal on Audio, Speech, and Music Processing, 25:1–16,2012. https://doi.org/10.1186/1687-4722-2012-25

[20] P. Marszałek, and P. Bilski.Steganography in Audio Files – COTS Software Analysis. International Journal of Electronics and Telecommunications,69(1),121-126,2023. https://doi.org/10.24425/ijet.2023.144340

[21] Gams. M, and Kolenik. T. Relations between Electronics, Artificial Intelligence and Information Society through Information Society Rules.Electronics, 10(4), 514:1-16, 2021. https://doi.org/10.3390/electronics10040514

[22] Imane Zerraza.Lightweight Authentication for IOT Edge Devices.Informatica,48(18),15–20,2024. https://doi.org/10.31449/inf.v48i18.6012

[23] W. A. Awadh, A. S. Hashim, and A. K.

Hamoud.Efficiently Secure Data Communications Based on CBC-RC6 and the Overflow Field of Timestamp Option in an IPv4 Packet.Informatica,46(6):125–133, 2022. https://doi.org/10.31449/inf.v46i6.4005

[24] Md. R. Rahman, P. Chakraborty, Md. Z. Rahman, and Md. G. Moazzam.Hiding Confidential File using Audio Steganography.International Journal of Computer Applications,178(50):30–35, 2019. https://doi.org/10.5120/ijca2019919422

[25] X. Zhang, S. Wang, and W. Zhang.Steganography Combining Data Decomposition Mechanism and Stego-coding Method.Informatica,33(1):41–48,2009. https://www.informatica.si/index.php/informatica/article/view/222/219

[26] M. H. Mohammed.A new approach to hide texts into images and audio files using steganography and cryptography techniques.International Journal of Advances in Applied Sciences (IJAAS),11(4):312–323,2022. https://doi.org/10.11591/ijaas.v11.i4.pp312-323

[27] Orora Tasnim Nisha, Md. Selim Hossain, and Mahfujur Rahman.Audio Steganography with Intensified Security and Hiding Capacity. European Chemical Bulletin,12(10):162–173,2023. https://www.eurchembull.com/archives/volume-12/issue-10/8322

[28] R. Chowdhury, D. Bhattacharyya, Samir Kumar Bandyopadhyay, and Tai-hoon Kim .A View on LSB Based Audio Steganography.International Journal of Security and Its Applications ,10(2),51-62,2016. https://doi.org/10.14257/ijsia.2016.10.2.05

[29] B. Harjito, B. Sulistyarso, and E. Suryani, .Audio steganography using two LSB modification and RSA for security data transmission.Journal of Telecommunication, Electronic and Computer Engineering,10(2–4):107–111, https://jtec.utem.edu.my/jtec/article/view/4326/3173

[30] M. Parthasarathi, and T. Shreekala.Secured Data Hiding in Audio Files Using Audio Steganography Algorithm.International Journal of Pure and Applied Mathematics, 116(21):619–628, 2017. https://acadpubl.eu/jsi/2017-116-13-22/articles/21/79.pdf

[31] Vipul Sharma, and Ravinder Thakur.LSB Modification based Audio Steganography using Trusted Third Party Key Indexing Method.IEEE 2015 Third International Conference on Image Information Processing(ICIIP), 403–406, 2015. https://doi.org/10.1109/ICIIP.2015.7414805

[32] A. H. Ali, M. R. Mokhtar, and L. E. George .Enhancing the hiding capacity of audio steganography based on block mapping. Journal of Theoretical and Applied Information Technology,95(7):1441–1448,2017. https://www.jatit.org/volumes/Vol95No7/13Vol95No7.pdf

[33] A. H. Ali, L. E. George, A. A. Zaidan, and M. R. Mokhtar.High capacity, transparent and secure audio steganography model based on fractal coding and chaotic map in temporal domain.Multimedia Tools and Applications, 77(23):31487-31516,2018. https://doi.org/10.1007/s11042-018-6213-0

[34] H. I. Shahadi, R. Jidin, and W. H. Way.Lossless audio steganography based on lifting wavelet transform and dynamic stego key.Indian Journal of Science and Technology,7(3):323–334, 2014. https://doi.org/10.17485/ijst/2014/v7i3.14

[35] K. Bhowal, D. C. Sarkar, S. Biswas, and P. P. Sarkar.A steganographic approach to hide secret data in digital audio based on XOR operands triplet property with high embedding rate and good quality audio .Turkish Journal of Electrical Engineering & Computer Sciences,25(3):2136–2148,2017. https://doi.org/10.3906/elk-1602-267

[36] H. T. Elshoush, and M. M. Mahmoud.Ameliorating LSB Using Piecewise Linear Chaotic Map and One-Time Pad for Superlative Capacity, imperceptibility and Secure Audio Steganography.IEEE Access,11:33354–33380,2023. https://doi.org/10.1109/ACCESS.2023.3259902

[37] Prolog fact base about countries derived from the 1994 CIA World Factbook, Reuters-21578 dataset [Online]. http://www.daviddlewis.com/resources/testcollections/reuters21578/feldman-cia-worldfactbook-data.txt.

Appendix A: Performance of the novel proposed algorithm on speech audio files.

Table 9: Specifications of speech audio files.

Audio File	Sampling Rate	Duration(s)	Bits per sample	Total Samples	Channel
dialogue.wav	22050	30	16	661500	Mono
news2.wav	22050	30	16	661500	Mono
india.wav	22050	30	16	661500	Mono

Table 10: Summary of performance metrics of the proposed BCS and BCR algorithms over the speech audio files.

Text File (Bytes)	Audio File	MSE	PSNR (dB)	BER	NCC	HER	Difference Ratio			
							Mean (μ)	Variance (σ^2)	Kurtosis (k)	Skewness (sk)
80	dialogue.wav	3.7735E-10	51.2156	7.5586E-06	1	7.4654E-11	1.2179E-06	1.0904E-06	6.4684E-07	3.4098E-05
800		2.9894E-09	41.2156	7.5586E-05	1	2.3516E-09	2.9968E-06	8.8253E-06	3.0053E-06	4.4714E-04
1600		6.111E-09	38.2053	1.5117E-04	1	1.5677E-09	1.7737E-05	1.7243E-05	1.1680E-05	5.4327E-04
2400		9.7950E-09	36.4444	2.2676E-04	1	1.0377E-08	3.1769E-05	2.0424E-05	1.3245E-05	9.8719E-04
80	news2.wav	2.3138E-10	51.2156	7.5586E-06	1	7.1816E-11	4.9431E-06	2.1313E-07	4.2572E-07	3.1299E-5
800		2.6409E-09	41.2156	7.5586E-05	1	1.2568E-09	8.8116E-05	1.2225E-05	1.0849E-05	1.0091E-04
1600		5.4481E-09	38.2053	1.5117E-04	1	2.1006E-09	5.0011E-05	3.1264E-05	2.0823E-05	5.6029E-04
2400		8.4483E-09	36.4444	2.2676E-04	1	7.2893E-09	1.0810E-05	5.1305E-05	2.3609E-05	4.6241E-04
80	india.wav	3.1742E-10	51.2156	7.5586E-06	1	9.7261E-12	5.5306E-06	8.2757E-07	1.5051E-06	8.7075E-07
800		2.4231E-09	41.2156	7.5586E-05	1	3.8904E-11	6.3265E-06	1.2215E-06	2.2362E-06	1.3568E-06
1600		5.5930E-09	38.2053	1.5117E-04	1	1.6437E-09	2.7714E-05	2.5489E-05	6.1954E-05	6.6199E-05
2400		7.5263E-09	36.4444	2.2676E-04	1	3.0054E-09	1.5816E-05	3.1367E-5	4.453E-06	2.6017E-06

Table 11: Performance of proposed BCS Algorithm.

Length of Text File (in Bytes)	Audio File	Audio Samples Utilized (ASU) (in samples)	Difference Bits (DB) (in bits)
80	dialogue.wav news2.wav india.wav	80	80
800		800	800
1600		1600	1600
2400		2400	2400

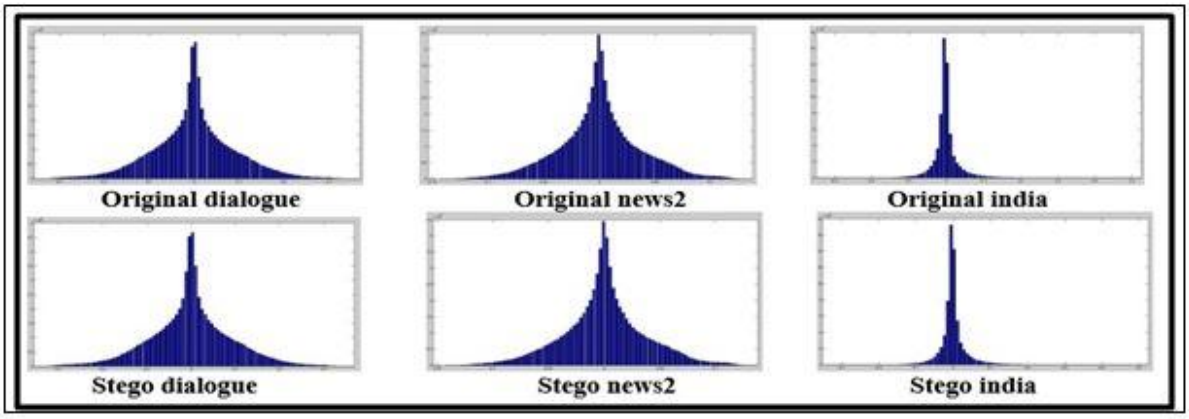


Figure 14: Histograms of cover and stego speech audio files with hidden text files of 2400bytes.

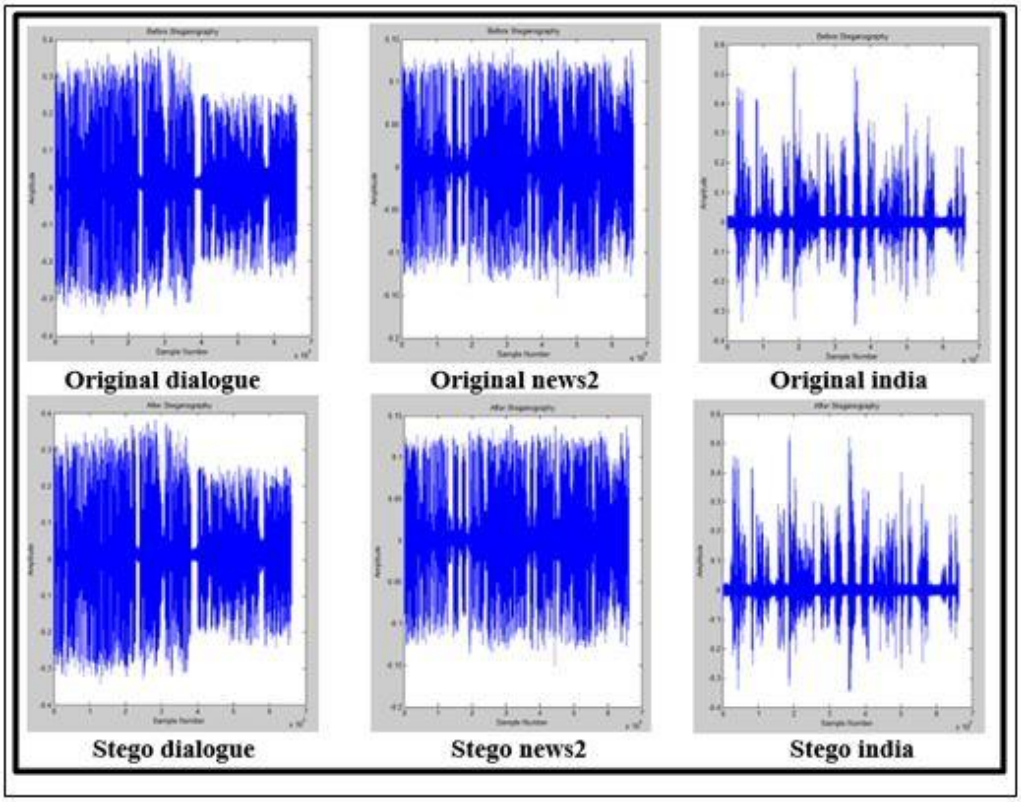


Figure 15: Cover and stego speech audio files with hidden text files of 2400 bytes.

Appendix B: Performance of the novel proposed algorithm on ambient audio files.

Table 12: Specifications of ambient audio files.

Audio File	Sampling Rate	Duration (s)	Bits per sample	Total Samples	Channel
Ambient AF1.wav	44100	29	16	1278900	2
Ambient AF2.wav	44100	15.3386	16	676432	2
Ambient AF2.wav	44100	51.7152	16	2280640	2

Table 13: Summary of performance metrics of the proposed BCS and BCR algorithms over the ambient audio files.

Text File (Bytes)	Audio File	MSE	PSNR (dB)	BER	NCC	HER	Difference Ratio			
							Mean (μ)	Variance (σ^2)	Kurtosis (k)	Skewness (sk)
80	Ambient AF1.wav	7.2556E-11	57.14359	1.9303E-06	1	0	1.1217E-04	4.4115E-08	8.4537E-08	6.6923E-06
800		5.9549E-10	47.08896	1.954E-05	1	0	1.0339E-04	4.5997E-08	8.8644E-08	6.1574E-06
1600		1.1511E-09	44.08138	3.9071E-05	1	3.9206E-13	1.1897E-04	4.6446E-07	9.1644E-07	7.7408E-06
2400		1.8947E-09	42.31775	5.8644E-05	1	3.9206E-13	2.4504E-04	2.0188E-06	4.0948E-06	1.1244E-04
80	Ambient AF2.wav	1.4504E-10	54.32284	3.6959E-06	1	2.3782E-12	6.1081E-04	5.2939E-07	8.2397E-07	4.3568E-05
800		1.6812E-09	44.32284	3.6959E-05	1	1.6648E-11	0.00107564	3.6071E-06	4.9900E-06	3.7829E-05
1600		2.8646E-09	41.31254	7.3917E-05	1	7.3725E-11	0.00144827	6.4778E-06	9.4422E-06	4.0325e-04
2400		4.1733E-09	39.55162	1.1088E-04	1	1.7836E-10	2.6507E-04	6.2648E-06	9.0433E-06	1.7764E-04
80	Ambient AF3.wav	4.0455E-11	59.60116	1.0962E-06	1	0	0.01618593	1.5061E-09	2.8664E-09	1.5991E-06
800		3.6637E-10	49.60116	1.0962E-05	1	1.9251E-13	0.08072214	1.2254E-07	2.4212E-07	7.5774E-06
1600		7.5242E-10	46.59086	2.1924E-05	1	0	0.07989209	5.8145E-07	1.1329E-06	5.2409E-06
2400		1.1499E-09	44.82995	3.2886E-05	1	1.9251E-13	1.02842913	9.6143E-07	1.7830E-06	7.4944E-05

Table 14: Performance of proposed BCS algorithm.

Length of Text File (in Bytes)	Audio File	Audio Samples Utilized (ASU) (in samples)	Difference Bits (DB) (in bits)
80	Ambient AF1.wav Ambient AF2.wav Ambient AF3.wav	80	80
800		800	800
1600		1600	1600
2400		2400	2400

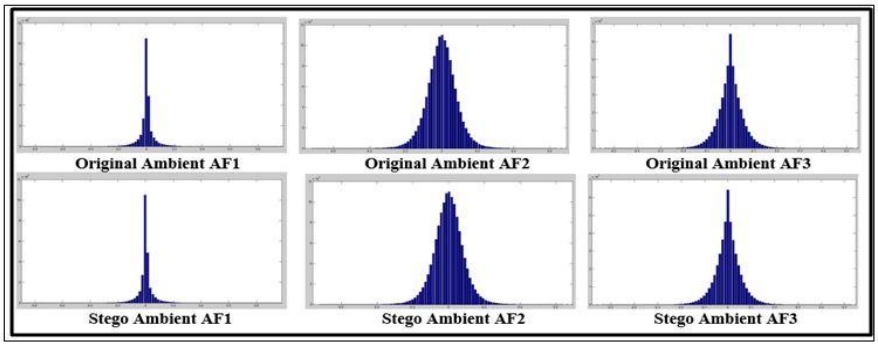


Figure 16: Histograms of cover and stego ambient audio files with a hidden text file of size 2400bytes.

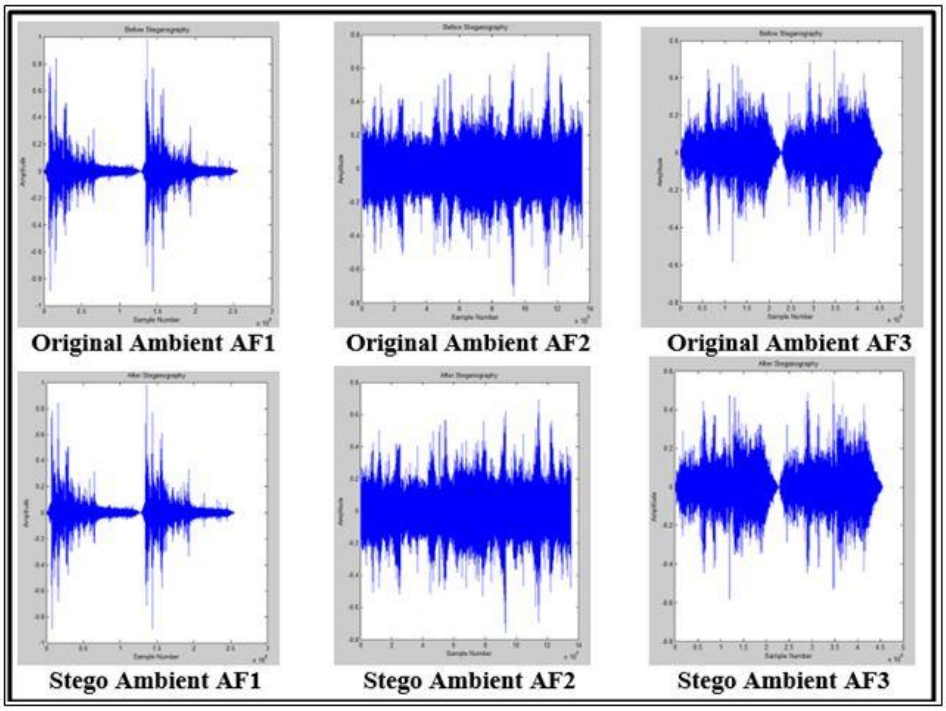


Figure 17: Cover and stego ambient audio files with a hidden text file of size 2400bytes.

Prophet Actor-Critic-Based Deep Reinforcement Learning for Obstacle Avoidance in Robotic Arm Control

Jiangtao Wang
College of Electrical and Control Engineering, Nanjing Polytechnic Institute, Nanjing 210048, China
E-mail: wangjiangtao@njpi.edu.cn

Keywords: deep reinforcement learning, robotic arm, intelligent model, AI control

Received: 2025.1.9

For the current common obstacle avoidance trajectory planning tasks of robotic arms, the problem of insufficient versatility is usually encountered. The grasping performance of the manipulator is mainly constrained by obstacles. How to improve the obstacle avoidance ability of the manipulator to improve its grasping ability. In order to improve the intelligent control effect of the robotic arm, an intelligent AI control method of the robotic arm combined with deep reinforcement learning is proposed. Moreover, in order to solve the problems of low learning efficiency caused by low-quality empirical data in the early stage of training and low efficiency in obtaining expert empirical data, an EDS mechanism is proposed to improve training efficiency by expanding expert empirical data and adopting unbiased dual memory bank sampling rules. In addition, in order to enhance the obstacle avoidance capability of the robotic arm, a COR system is constructed to help quickly generate the optimal trajectory, and the end effector and fuselage of the robotic arm can simultaneously avoid obstacles in complex environments, and achieve a balance between obstacle avoidance and motion exploration. The results show that the success rate of grasping complex objects in both obstacle free and obstacle free environments can reach more than 80%. Compared with the existing models, it has faster convergence speed and learning effect, and has better model performance. This paper combines experimental analysis to verify the effectiveness of the AI control method proposed in this paper, which can effectively ensure the collision-free trajectory planning of the robotic arm in different scenarios and has strong adaptability to scene changes.

Povzetek: Raziskano je globoko ojačevalno učenje za načrtovanje poti in izogibanje ovir pri robotiziranem ročnem nadzoru, kjer kombinacija algoritmov za izboljšano učenje in optimizacijo omogoča hitro prilagajanje v dinamičnih okoljih ter izboljša sposobnost robota za izogibanje oviram in natančno izvajanje nalog.

1 Introduction

With the continuous development of deep learning technology and computer hardware foundation, artificial intelligence algorithms represented by computer vision have made unprecedented breakthroughs. Image processing algorithms based on convolutional neural networks far exceed traditional algorithms in real-time image classification, target detection, semantic segmentation and other scene detection effects, providing new possibilities for the development of related industries. Therefore, adding a visual servo control system to the robotic arm, collecting images through visual sensors, and performing grasping detection based on convolutional neural network will improve the intelligent perception ability of the robotic arm in the working environment [1].

Robotic arm grasping is a complex task involving object detection, robotic arm path planning and target grasping pose estimation. At present, many methods focus on target detection and grasping pose estimation. Among them, a common method is to detect the geometric features of the object to find the suitable grasping position, such as handle-like objects and symmetrical objects, and then estimate the grasping pose of the object, and control the end effector to complete the target grasping. However, many objects do not have similar distinct features, or the target objects may be in a cluttered stacked environment, and the target objects are partially blocked by obstacles, so the vision system cannot accurately detect the target and locate the parts to be grasped. Therefore, in recent years, scholars usually use machine learning methods such as convolutional neural networks [2], which can more robustly calculate the features of targets, so as to realize target detection and target grasping pose.

With the upsurge of artificial intelligence technology research, artificial intelligence technology represented by deep learning and deep reinforcement learning has been widely used in the field of robot research, which has brought new opportunities for the development of robots [3]. In the robot arm collision detection, the sensor method first introduces the deep

learning technology, uses the neural network to learn the signal data of the sensor, and uses the excellent feature extraction ability of deep learning to extract the collision features in the signal, which finds a new solution to the problems of high system complexity and complex detection model establishment in the sensor method [4]. At present, the geometric simulation method has not been developed with depth.

Through the analysis of the research status of capture detection, it can be found that because the capture detection task is more complex than the target detection, and the capture scene is quite different, the production cost of data set is high, so the capture detection neural network based on the capture data set training is relatively difficult and costly in the actual training [5]. Deep reinforcement learning has made some progress in solving the problem of robot arm control and obstacle avoidance. However, there are still some problems such as poor training efficiency, large error and poor adaptability to environmental changes [6]. In addition, due to the long training time of the deep reinforcement learning algorithm, the physical training method has low efficiency and poor security [7]. However, the simulation training method has the problem of poor physical transfer effect, which has brought great trouble to the practical application of deep reinforcement learning algorithm. Therefore, the subsequent capture detection module in this paper will estimate the 3D capture pose of the object based on the feature extraction technology of deep reinforcement learning.

Aiming at common robotic arm grasping scenarios, this paper proposes a robotic arm grasping method based on deep reinforcement learning to solve the robotic arm grasping tasks in corresponding scenarios. Aiming at the multi-target disordered grasping task in cluttered environment, it uses the disordered grasping method in continuous action space. Aiming at the obstacle avoidance capture task in the cluttered stacking environment, it uses the obstacle avoidance capture method in the discrete action space. Moreover, this paper combines two methods to construct a robotic arm grasping system to solve the hybrid grasping task under the above-mentioned environment integration, which has great potential in the field of scientific research and social application value.

With the rapid development of industrial automation and intelligent robot technology, the manipulator system has been widely used in storage and logistics, precision assembly, medical surgery and other fields. However, in the real operation scene, the manipulator often faces complex environmental interference such as dynamic obstacles, illumination changes, target occlusion, which leads to significant limitations of the traditional motion planning algorithm. Statistics show that in the industrial scene with more than 30% random obstacles, the success rate of the 6-DOF manipulator is generally lower than 65%, and the collision risk rate is as high as 22%. There are three main technical bottlenecks: first, the traditional path planning algorithm based on geometric model is difficult to deal with the real-time updated environmental information of dynamic obstacles; Secondly, the conventional vision algorithm is susceptible to texture interference and the lack of depth information in the scene of densely stacked objects, resulting in the error of target pose estimation exceeding ± 5 °; Third, the existing obstacle avoidance strategies mostly use static threshold judgment, and lack of collaborative optimization of manipulator kinematics constraints and end effector grasping posture.

This study proposes a collaborative optimization framework based on deep reinforcement learning, which aims to break through the problem of precise grasping of manipulator in complex environment.

The purpose of this paper is to improve the robot's obstacle avoidance effect based on the existing model, and to solve the problem of customer service data training, so as to improve the robot's intelligent control effect and autonomous decision-making ability, as follows:

(1) In order to solve the problem of poor universality of expert strategy, this paper studies and designs a prophet Strategy Network Guided deep reinforcement learning algorithm. The algorithm has the ability of expert strategy self optimization, can avoid the situation that the expert strategy itself is a local optimal strategy, and can be widely used in changing scenarios. In addition, the guidance effect of expert strategy on different task scenarios is studied and analyzed.

(2) To solve the problem of low efficiency in the use of expert experience data, this paper proposes an unbiased dual memory sampling mechanism combined with the expert memory amplification mechanism. In addition, the guiding effect of the mechanism in the early stage of algorithm training and the influence of algorithm convergence are studied.

2 Related works

In order to better apply deep reinforcement learning algorithms to real robotic arms, Mourtzis et al. [8] proposed a method that combines convolutional neural network with deep reinforcement learning to perform robot control tasks that require close correlation between vision and control. It uses a monocular camera for image acquisition as the original input, performs feature extraction through a convolutional neural network, and then directly inputs the feature information into deep reinforcement learning to achieve end-to-end joint training of perception and control systems. Moreover, it trains the model in a simulation environment and then migrates to the real manipulator control task. The experimental results show that this end-to-end method can perform complex operation skills, but it is very unstable and takes a lot of time to train in the network.

Zhou et al. [9] proposed a Robotics Transformer (RT-1) algorithm framework, which can effectively absorb a large amount of data and expand with the amount and diversity of data. By using 13 robotic arms to collect more than 130k sets of large demonstration data sets in 17 months to train RT-1, the algorithm has achieved very good results. Through many experiments, the new algorithm has achieved a success rate of 97% in various control tasks of robotic arms, such as grasping and handling, and the new algorithm can be effectively generalized to new tasks, objects and environments.

Singh et al. [10] proposed a learning-based hand-eye coordination robotic arm grasping method. The method is data-driven, and it centers on the object to be grasped, directly from the image pixel to the robotic arm end effector motion, and performs end-to-end network training. In order to better apply the algorithm to real scenarios, the data demand of end-to-end network training is explored and tested through experiments, and the effectiveness of this method is verified. Then, the network model is trained by using 12%, 25% and 50% of the captured data in the dataset. The results show that with the increasing number of captured data, the capturing success rate of the trained network model continues to improve. Although such large-scale and long-term training has great guiding significance in exploring the potential of the algorithm, it also proves that it takes a lot of time and cost to train the model of the end-to-end framework.

In order to better solve the problem of difficult training of vision-based deep reinforcement learning, Abdullah-Al-Noman et al. [11] proposed a modular deep reinforcement learning method, which can transfer the simulation trained model to real-world robot tasks. By introducing a network connection layer between the two modules, the visual inspection network and the control network of deep reinforcement learning are modified to enable the network to be trained independently. When the independent modules are trained, they are merged and fine-tuned in an end-to-end manner to further improve hand-eye coordination.

The model-free algorithm based on deep reinforcement learning shows strong autonomous decision-making ability without the need for a system model, which provides new ideas for robotic arm path planning and target grasping pose estimation. Xu et al. [12] proposed a method of global path planning and local reinforcement learning to avoid obstacles. In the motion of global path planning, reinforcement learning is used to avoid obstacles according to local environmental information. In addition, deep Q-value learning and dual deep Q-value learning algorithms also realize the path planning of the robotic arm [13], process the spatiotemporal information of the system, and solve the path planning and obstacle avoidance problems of the robot in the dynamic environment. Tang et al. [14] used deep Q value to learn the grasping pose estimation of the object, and only when the jaw successfully grasps the object, the agent is given a positive reward value. Nowadays, more advanced discrete reinforcement learning and continuous reinforcement learning algorithms [15] have become the mainstream of training robotic arm control strategies in deep reinforcement learning algorithms, and both have realized a series of robotic arm grasping tasks. However, due to problems such as low sample training efficiency [16], deep reinforcement learning has always had a bottleneck in its wide application in real-world robots. Although there is already an effective position-controlled robotic arm operation framework [17], they still need a lot of time to train each task and provide intensive reward information. Moreover, in the robotic arm grasping task, besides path planning, we should also pay attention to target recognition and grasping pose estimation. Therefore, it is often necessary to perform network training using image input [18]. At the same time, the high-dimensional and complex and changeable image input makes the problems of low efficiency of deep reinforcement learning training samples and difficult network training.

In order to solve the above-mentioned problems of deep reinforcement learning in robotic arm grasping applications, aiming at the problem of image input, Contrastive Unsupervised Representations for Reinforcement Learning (CURL) uses contrastive learning to extract image input and reduce its dimensionality into vector information for the training of evaluation networks and policy networks, which effectively reduces the complexity of training [19]. In order to improve the sample efficiency of deep reinforcement learning and make the learned control strategy more robust, image enhancement is usually used nowadays. Image enhancement is an image-based data enhancement that includes random transformations such as cropping, rotation, or color dithering. It is widely used in computer vision architecture, including pioneering works such as LeNet and AlexNet [20]. However, only recent studies have fully demonstrated the effectiveness of data reinforcement for deep reinforcement learning. Raj et al. [21] proposed a robotic arm control framework combining contrastive learning and image enhancement, which can quickly solve a series of simple robotic arm control tasks, such as grasping squares and opening drawers without barriers. However, it is also difficult to solve complex tasks such as obstacle avoidance and grasping of robotic arms. In addition, the Sim2Real method can also solve the sample efficiency problem in the robotic arm grasping application based on deep reinforcement learning. First, the agent is trained in the simulation environment, and then migrated to the real world for use [22]. In simulation environment training, the physical attributes of the environment can be random, so the images collected by the body vision system are richer, which can effectively increase the training sample size. In addition, imitation learning also plays a role in helping robotic arms train reinforcement learning control

strategies [23]. One of the simplest forms of imitation learning is behavior cloning (BC), which uses regression to fit the expert sample set, so that reinforcement learning agents can learn strategies close to expert control. However, imitation learning usually needs to collect a large number of demonstration samples, and if only BC is used to train the agent, the expert strategy will limit the learning ability of the agent [24], and the best strategy performance of the agent will not be able to surpass the expert strategy.

Table 1: advantages and disadvantages of existing model algorithms

	Algorithm model	Advantage	Deficiency
1	(RT-1) algorithm	High accuracy and strong scalability	Requires a lot of data training, and the effect of path planning is poor
2	Reinforcement learning	Good path planning effect and accurate target capture	The efficiency of training samples is low,
3	Unsupervised reinforcement learning	Good path planning effect, accurate target capture and high training efficiency	Poor obstacle avoidance effect

The advantages and disadvantages of the existing research are shown in Table 1. From Table 1, unsupervised reinforcement learning has the problems that customer service needs a lot of training data and sample training efficiency is low, but the effect of obstacle avoidance is not good, and it still needs to be further improved. Therefore, this paper proposes an improved deep reinforcement learning model to further improve the effect of intelligent obstacle avoidance of manipulator.

3 Robotic arm obstacle avoidance trajectory planning framework based on deep reinforcement learning

The motion path of the manipulator in complex environment will be affected, resulting in its inability to accurately grasp the target. Improving the success rate of the manipulator in cluttered environment through the visual obstacle avoidance algorithm is the research direction of this paper. Through the deep reinforcement learning to improve the visual obstacle avoidance effect, provide reliable motion path and grasp direction for the manipulator, and improve the success rate of the manipulator.

In this chapter, a general trajectory planning framework based on DRL is proposed to realize the autonomous obstacle avoidance of robot tasks. A prophet guided actor critic structure based on expert strategy is designed to support rapid re planning of work scene changes. Secondly, an extended double memory sampling mechanism is proposed to effectively expand the expert memory from a few demonstrations, and improve the training efficiency of DRL algorithm through gradually unbiased sampling rules. Finally, a compound obstacle avoidance reward system is designed to synchronously realize the collision free movement of the robot end effector and the fuselage, which can build a dense reward mapping and achieve a balance between obstacle avoidance and motion exploration.

3.1 Obstacle avoidance trajectory planning guided by expert network

In this paper, a step-by-step obstacle avoidance strategy is proposed, which aims to cope with the rapid re-planning of changing work scenarios and perform simplified strategy learning for complex obstacles.

Inspired by transfer learning, the Prophet Policy Network Guided Method (PAC) proposed in this paper is a general framework, which can be used for all DRL algorithms based on Actor-Critic framework, such as DDPG, TD3, SAC, etc. This subsection will describe implementation detail of PAC strategy learning proposed in this paper.

PAC is an algorithm framework combining deep reinforcement learning and dynamic strategy optimization, which aims to improve the efficiency and safety of Obstacle Avoidance Trajectory Planning of manipulator in complex environment. This method uses the strategy network to guide the training process, and solves the problems of low exploration efficiency and insufficient sample utilization of traditional reinforcement learning in the sparse reward scene

The core of PAC method is to dynamically adjust the exploration and utilization balance of manipulator through the combination of strategy network and prior knowledge guidance. Specifically, it includes:

(1) Strategy network architecture: adopt the deep deterministic strategy gradient (ddpg) framework to build a dual network structure based on actor critical. The actor network outputs continuous actions (such as joint angle and speed), and the critical network evaluates the action value and guides the strategy update.

(2) Prophet guidance mechanism: dynamic reward function and trajectory prediction module are introduced

to generate auxiliary reward signal through prior knowledge (such as inverse kinematics solution and obstacle position prediction) to accelerate model convergence

Based on the traditional sparse reward (such as successfully avoiding obstacles or reaching the target), the distance penalty (negatively related to the distance of obstacles) and the kinematic reward (based on the minimization of joint power) are added to guide the manipulator to choose a better path.

The proposed PAC consists of the original network and the prophet network, both of which have the same network structure and both include the complete Actor-Critic framework, as shown in Figure 1. The original online Actor network is denoted as $\mu_0\left(s_t\left|\theta^{\mu_0}\right.\right)$ with network parameter θ^{μ_0}, and the prophet online Actor network is denoted as $\mu_p\left(s_t\left|\theta^{\mu_p}\right.\right)$ with network parameter θ^{μ_p}. The original online Critic network is denoted $Q_o\left(s_t,a_t\left|\theta^{Q_o}\right.\right)$ with network parameter θ^{Q_o}, and the prophet online Critic network is denoted $Q_p\left(s_t,a_t\left|\theta^{Q_p}\right.\right)$ with network parameter θ^{Q_p} [25].

At each time step *t*, for state s_t, the original online Actor network and the prophet online Actor network generate the original action a_t^o and the prophet action a_t^p, respectively.

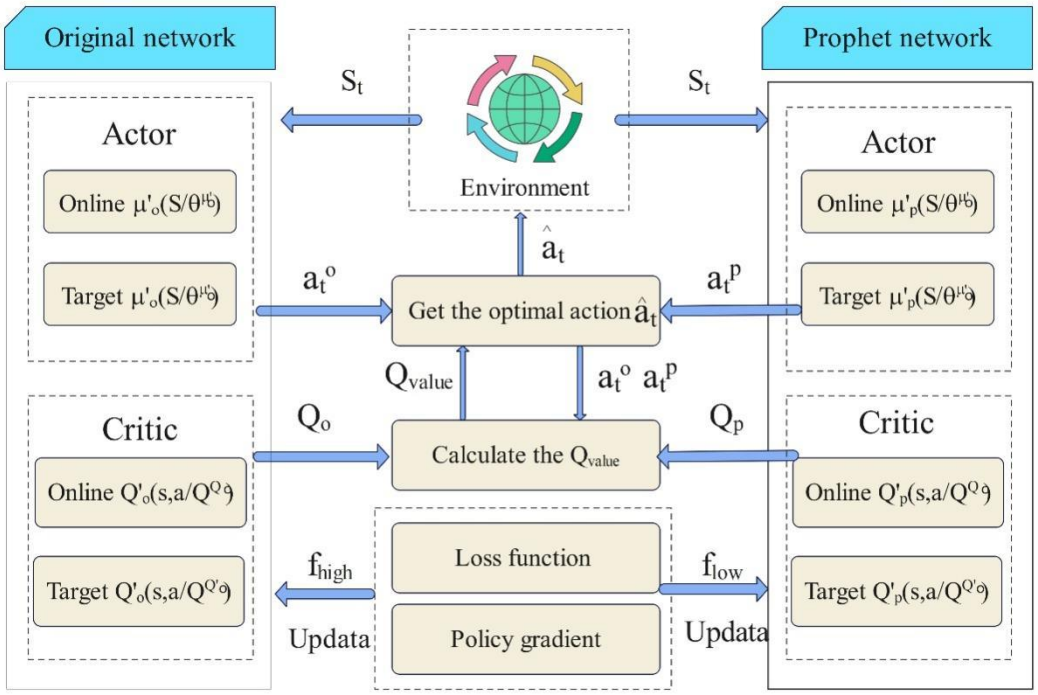


Figure 1: Prophet strategy network framework

$$\begin{cases} a_t^o = \mu_0\left(s_t\left|\theta^{\mu_0}\right.\right) + N_t \\ a_t^p = \mu_p\left(s_t\left|\theta^{\mu_p}\right.\right) + N_t \end{cases} \quad (1)$$

Among them, N_t is random action noise.

The original online Critic network and the prophet online Critic network will jointly comprehensively evaluate the two actions in formula (1).

$$Q_{value}\left(s_t,a_t\right) = \sum_{\theta^{Q_j}\in\left\{\theta^{Q_o},\theta^{Q_p}\right\}} Q\left(s_t,a_t\left|\theta^{Q_j}\right.\right) \quad (2)$$

Then, the action with the highest comprehensive score value $Q_{value}\left(s_t,a_t\right)$ is defined as the optimal action $\hat{a}_t$.

$$\hat{a}_t = arg\,max\,Q_{value}\left(s_t,a_t\right) \quad (3)$$

According to the Bellman equation, the target Q values of the original network and the prophet network Q_{target}, Q_{target} is defined as[26]:

$$Q_{target} = r_t + \gamma Q'\left(s_{t+1},\mu'\left(s_{t+1}\left|\theta^{\mu'}\right.\right)\middle|\theta^{Q'}\right) \quad 4)$$

Among them, γ is the discount factor, the target Actor network is denoted as $\mu'(\cdot)$ with network parameter $\theta^{\mu'}$, the target Critic network is denoted as $Q'(\cdot)$ with network parameter $\theta^{Q'}$, and r_t is the reward value of environmental feedback when the time step is t.

For online Critic networks, the time difference error L_1 is defined as

$$L_1 = \frac{1}{n}\sum_{t=1}^{n}\left(Q_{target} - Q\left(s_t, a_t \middle| \theta^Q\right)\right)^2 \quad (5)$$

In order to embody the guiding ability of the prophet network in the training process of neural network, this paper designs the prophet auxiliary loss function L_2, which is defined as:

$$L_2 = \frac{1}{n}\sum_{t=1}^{n}(1-d_t)\left\|Q^o_{target} - Q^p_{target}\right\|^2 \quad (6)$$

Among them, $\|\cdot\|$ represents the Euclidean norm, u is a Boolean value, and d_t is used to determine whether the TCP reaches the target area within the allowable error range. If TCP successfully reaches the target area, $d_t = 1$ is taken. At this time, $L_2 = 0$ means that the robotic arm of the current round has successfully completed an obstacle avoidance trajectory planning task, so the prophet auxiliary loss function L_2 of the current round is set to zero.

Therefore, the original online Critic network and the prophet online Critic network are updated and optimized by minimizing the composite loss function.

$$L = \lambda_1 L_1 + \lambda_2 L_2 \quad (7)$$

Among them, λ_1 and λ_2 are the weight coefficients, which can be adjusted according to the needs of different obstacle avoidance trajectory planning tasks.

In addition, the original online Actor network and the prophet online Actor network still adopt formula (7) to update the network parameters. Both the target Actor network and the target Critic network adopt the soft update mode of formula (9).

It is worth noting that the formula for updating the Prophet network is the same as the original network. As shown in Figure 2, the Prophet network adopts a delayed update mechanism. For each time step tt, the update frequency of the original network is defined as f_{high}, while the update frequency of the seer network is defined as:

$$f_{low} = \frac{f_{high}}{n_f} \quad (8)$$

Among them, n_f is a positive integer.

3.2 Memory bank design and sampling mechanism

Multi memory database solves the problems of real-time, security and generalization ability of manipulator in complex scenes, and has become the core architecture of intelligent robot control system in the industrial 4.0 era. Its value is particularly prominent in the scenarios of high-speed mixed line production (such as automobile manufacturing) and flexible logistics (such as e-commerce sorting).

The multi memory library can classify and store preset operation procedures (such as palletizing path and welding track) and real-time adjustment instructions, and realize task priority through dynamic scheduling mechanism. For example, in complex sorting scenarios, high priority tasks (such as emergency obstacle avoidance) can immediately call the obstacle avoidance strategy in the memory to interrupt low priority actions The multi memory inventory stores the historical operation parameters (such as grasping force and movement speed), and automatically matches the optimal configuration in similar tasks to reduce the repeated debugging time.

As the basis of deep reinforcement learning algorithm for off-line strategy learning, memory bank has a great impact on the algorithm training. Generally speaking, at the initial stage of training, due to the random exploration of agents, the quality of data stored in the memory is poor, resulting in low learning efficiency. Especially when there are obstacles in the trajectory planning task, it is difficult for robots to obtain high return empirical data. Therefore, the above problems can be solved through multiple memory banks.

According to the amplification mechanism of expert memory bank, this paper constructs a multi-memory bank structure, which aims to amplify a large number of expert memory data by using a small number of artificial expert teaching experience, and adopts a gradual unbiased double memory bank sampling mechanism to improve the exploration efficiency of agents.

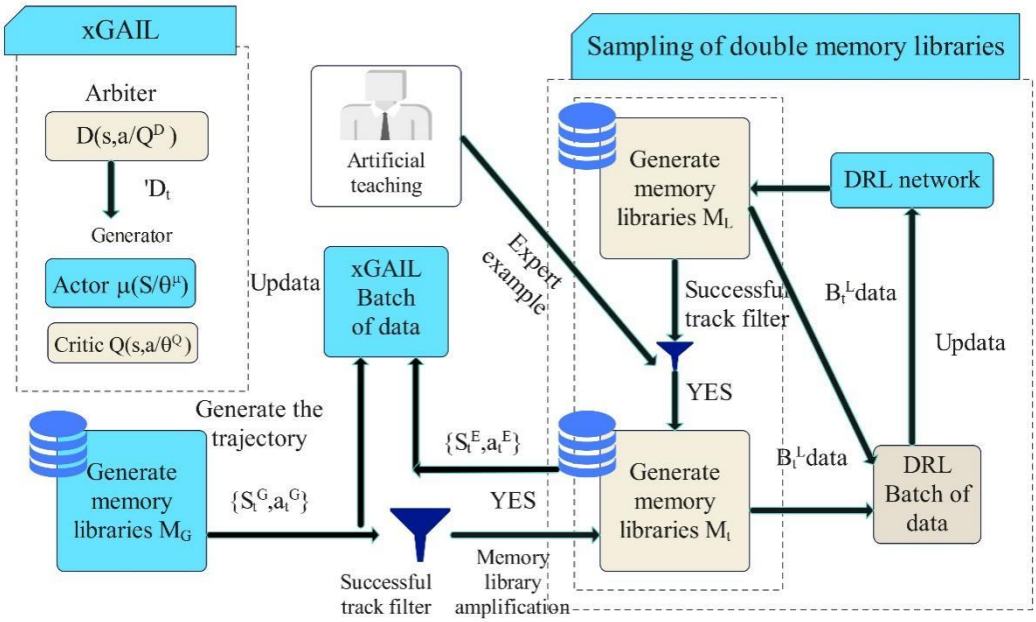


Figure 2: Amplifiable dual memory bank sampling (EDS)

As shown in Figure 2, the dual memory bank sampling method based on memory bank amplification proposed in this paper has two modules that can be parallel: a memory bank amplification module based on xGAIL and a dual memory bank sampling module. Among them, the memory bank amplification module based on xGAIL is mainly used to amplify a small number of expert examples taught manually into a large amount of expert experience data, so as to guide agent learning.

In order to improve the optimization learning efficiency of DRL algorithm strategy, an increasingly unbiased dual-memory sampling mechanism is designed to guide agents' strategy learning. At each time step t, B_t^L and B_t^E data are randomly sampled from the exploration memory bank M_L and the expert memory bank M_E, respectively [27].

$$\begin{cases} B_t^L = \left\lfloor \frac{N}{2} + \frac{t}{T_s} \right\rfloor, B_t^E = N - B_t^L, 0 \le t < \frac{NT_s}{2} \\ B_t^L = N, B_t^E = 0, t \ge \frac{NT_s}{2} \end{cases} \tag{9}$$

Among them, N is the batch data size sampled from the memory bank for each time step t, T_s represents the number of decay steps, and both B_t^L and B_t^E are non-negative integers.

In order to make the sampling more and more unbiased, the value of B_t^E is decreased by 1 every T_s steps, so as to gradually reduce the sampling weight of the data in the expert memory bank M_E. When the DRL converges, B_t^E is set to 0, and its purpose is to allow the DRL to generate empirical data for strategy learning and optimization, so as to achieve an unbiased effect.

3.3 Design of compound obstacle avoidance reward function

Through the compound obstacle avoidance reward function, the collision free motion of the end effector and the fuselage of the robot can be realized synchronously, which can build a dense reward mapping and achieve a balance between obstacle avoidance and motion exploration.

In this section, a compound obstacle avoidance reward system is proposed, which consists of pose error reward function, artificial-like potential field reward function and shortest step reward function, as shown in Figure 3.

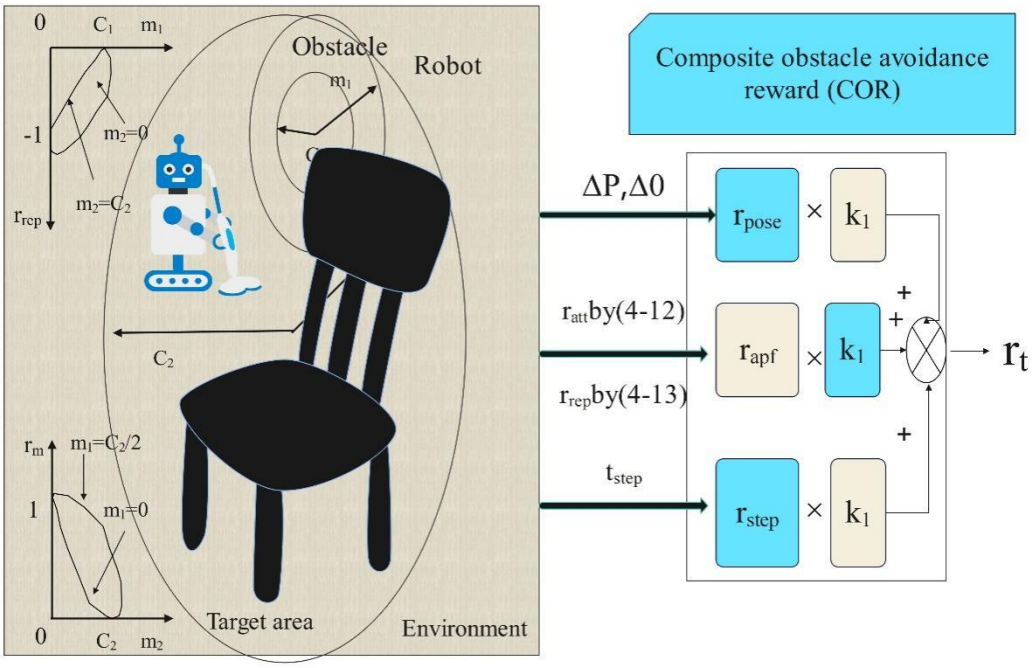


Figure 3: Compound obstacle avoidance reward function system

For the robotic arm trajectory planning task, if the DRL can perceive the completion progress of the task in the entire training stage, the convergence speed of the DRL can be greatly improved, which requires a reward function that can reflect the current task progress of the robotic arm. Therefore, this paper sets the pose error reward function r_{pose} as[28]:

$$r_{pose} = -\|p_1 - p_0\| - \varphi\left(|\alpha_1 - \alpha_0| + |\beta_1 - \beta_0| + |\gamma_1 - \gamma_0|\right) \quad (10)$$

Among them, $p_0 = (x_0, y_0, z_0)$ and $p_1 = (x_1, y_1, z_1)$ are the positions of the robotic arm mission target point and TCP relative to the world coordinate system, $(\alpha_0, \beta_0, \gamma_0)$ and $(\alpha_1, \beta_1, \gamma_1)$ are the attitudes of the robotic arm mission target point and TCP relative to the world coordinate system, and φ is a weight factor used to adjust the importance of position and attitude accuracy to the current task.

Obviously, using r_{pose} as the state reward function of the manipulator can clearly reflect the task progress of the manipulator. When $r_{pose} = 0$, the TCP representing the manipulator completely coincides with the task target point in pose and pose, that is, the manipulator completes the task of reaching the target. In addition, r_{pose} can provide the progress of the robotic arm throughout the training process, which makes the reward more intensive and is conducive to the strategy learning of the robotic arm.

In order to achieve the balance between the two, inspired by the concept of artificial potential field, this paper proposes an obstacle avoidance reward function r_{apf} similar to artificial potential field constructed by quadratic function, which consists of two parts: attraction reward function r_{att} and repulsion reward function r_{rep}, and is defined as:

$$r_{apf} = r_{att} + r_{rep} \quad (11)$$

The influence radius of the obstacle is defined as c_1, which is determined by the safety clearance requirements of the robotic arm for specific tasks. If c_1 is set too small, the manipulator may generate a dangerous trajectory, which is not conducive to practical application. The influence range of the target point is defined as c_2, which is determined by the actual working diameter of the robotic arm, that is, the target point will only have an influence in the working space of the robotic arm. In addition, the minimum three-dimensional spatial distance between the robotic arm and the obstacle in the workspace is defined as m_1, and the three-dimensional spatial distance between the TCP of the jaw at the end of the robotic arm and the target point is defined as j, where

$$m_1 \in \left[0, \frac{c_2}{2}\right] \text{ and } m_2 \in [0, c_2].$$

The attraction reward function r_{att} is defined as a quadratic function passing through the point $(0,1)$ and the point $(c_2, 0)$, designed as:

$$r_{att} = \left(1 - 4c_2^{-1}m_1\right)m_2^2 - \left(c_2^{-1} + c_2 - 4m_1\right)m_2 + 1 \quad (12)$$

Among them, $r_{att} \in [0,1]$ can be inferred from the

function formula, and it can be seen that the closer the distance between TCP and the target point, the greater r_{att}.

The repulsive reward function r_{rep} is a quadratic function through the point $(0,-1)$ and the point $(c_1,0)$, designed as[29]:

$$r_{rep}=\begin{cases}\left(1-2c_2^{-1}m_2\right)m_1^2+\left(c_1^{-1}-c_1+2m_2\right)m_1-1, m_1\le c_1\\ 0, otherwise\end{cases} \tag{13}$$

Among them, $r_{rep}\in[-1,0]$ can be inferred from the function formula, and the farther the minimum three-dimensional spatial distance between the robot and the obstacles in the workspace is, the greater r_{rep} is.

For the practical application of robotic arm, energy optimization is often considered, so as to complete the same task with the least energy consumption, thus reducing the cost. In the training of deep reinforcement learning, reducing the number of steps to complete the same task can also significantly reduce the energy consumption of the robotic arm. Therefore, this paper designs a reward function r_{step} with respect to the number of training steps, which can make the robotic arm reach the target area with as few steps as possible

$$r_{step}=-\left(\frac{t_{step}+1}{N_{step}}\right)+\frac{1}{2} \tag{14}$$

Among them, t_{step} is the number of steps in the current round, and N_{step} is the maximum number of steps in each round. It can be seen that the value range of r_{step} is $\left[-\frac{1}{2},\frac{1}{2}\right]$.

As training progresses, r_{step} will gradually decrease. If the manipulator does not complete the task in the current round, the cumulative value of r_{step} in the whole round is 0. If the manipulator completes the task with only a few steps, the cumulative value of r_{step} in the whole round is larger, which urges the manipulator to learn strategy in the direction of completing the task with a few steps.

Finally, the compound obstacle avoidance reward function (COR) proposed in this paper is defined as:

$$r_t=k_1r_{pose}+k_2r_{apf}+k_3r_{step} \tag{15}$$

Among them, k_1, k_2 and k_3 are the weight coefficients of each reward item.

3.4 Overall implementation plan

As shown in Figure 4, by combining the proposed prophet strategy, amplifiable dual memory bank sampling and compound obstacle avoidance reward function, an inverse kinematics solution framework for obstacle avoidance trajectory planning of robotic arm based on deep reinforcement learning is constructed, and its specific implementation scheme is as follows.

For each step in the formal training, the original online Actor network and the prophet online Actor network output actions a_t^o and a_t^p, respectively from the current state s_t according to formula (1). After passing the calculation Q_{value}, the optimal action $\hat{a}_t$ is obtained by formula (3) to perform the predetermined task. Then, the proposed compound obstacle avoidance reward function is used to obtain reward value feedback, and then the empirical data of the current step is stored in the exploration memory bank M_L. The number N batch data is then sampled by formula (9) through the double memory bank sampling mechanism to update the online network and optimize the target network by soft update. In addition, this paper optimizes the prophet network by delaying update rules.

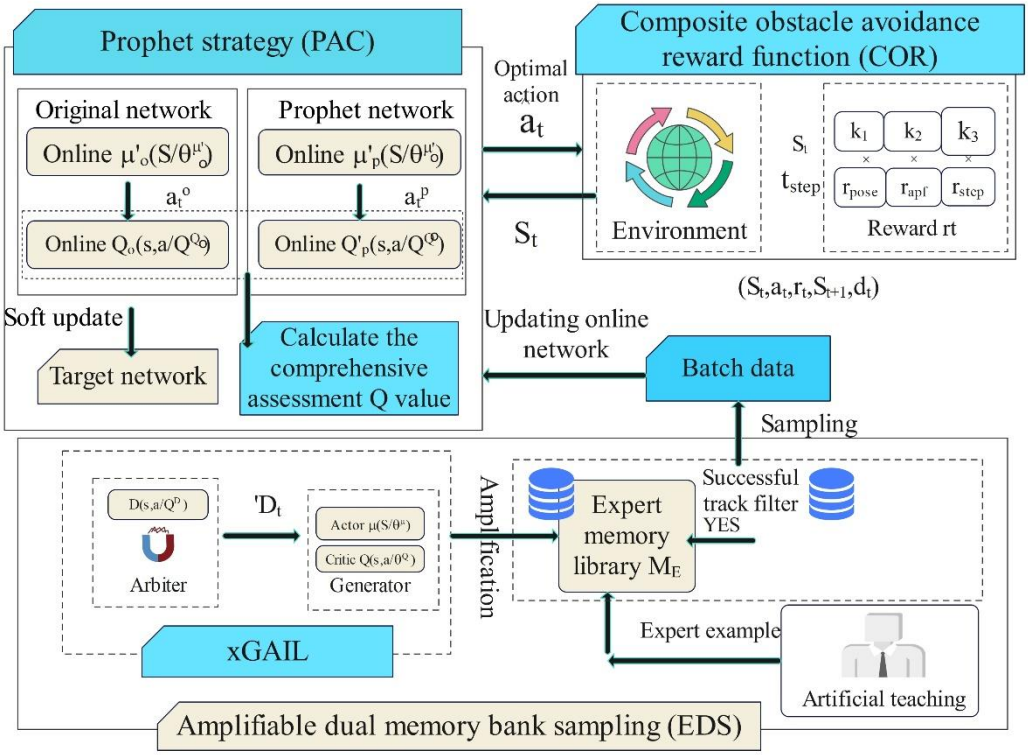


Figure 4: Schematic diagram of overall implementation

During the training process, the successful trajectory filter screens high-quality data from the exploration memory bank M_L to the expert memory bank M_E, and the expert experience data extended by xGAIL will also be screened into the expert memory bank M_E by the generation memory bank M_G. Among them, all memory banks follow the first-in, first-out data flow mechanism.

Finally, the round ends when the TCP of the robotic arm reaches the target area or the number N_{step} of training sessions is reached.

The grasping method based on deep reinforcement learning technology mainly relies on the depth camera to collect data information, transmits the collected image information to the computer for algorithm processing, and then feeds it back to the robotic arm to grasp the information for motion planning. The grasping platform framework is shown in Figure 5, The sensor uses ls-a80020 laser sensor of laser optoelectronics.

The image data input module acquires image data through the camera, which is the visual perception part of the whole system The capture prediction is based on the input image data, and the system predicts the capture position by algorithm. The system outputs the optimal grab position from the grab prediction module, which is the optimal grab point finally determined by the system. The optimal grasping position information determined is transmitted to the control module. The control module is responsible for generating control signals to guide the movement of the manipulator. When the manipulator performs the grasping task, it will feed back its attitude information to the system in real time. This feedback is used to adjust and optimize the motion of the manipulator to ensure accurate execution of the grasping task. The operating system platform is the core control unit of the whole system, which is responsible for integrating and processing all input information (such as image data and manipulator feedback), and coordinating the work of each module.

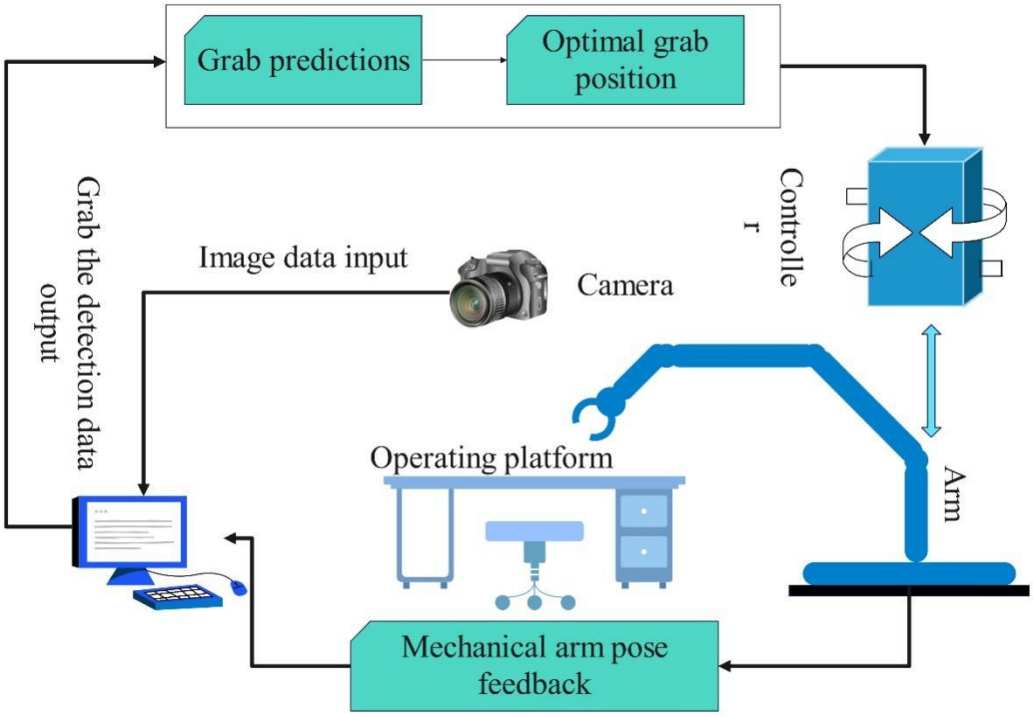


Figure 5: Overall framework of grasping platform

The robotic arm's work of grasping objects is entirely based on visual perception of information, thus realizing autonomous grasping. In the complete grasping detection process, it mainly includes four parts: initialization of simulation environment and detection network, loading of simulation environment and object model, deep enhanced network grasping pose detection, and robotic arm grasping motion. The robotic arm grasping steps are as follows in Figure 6:

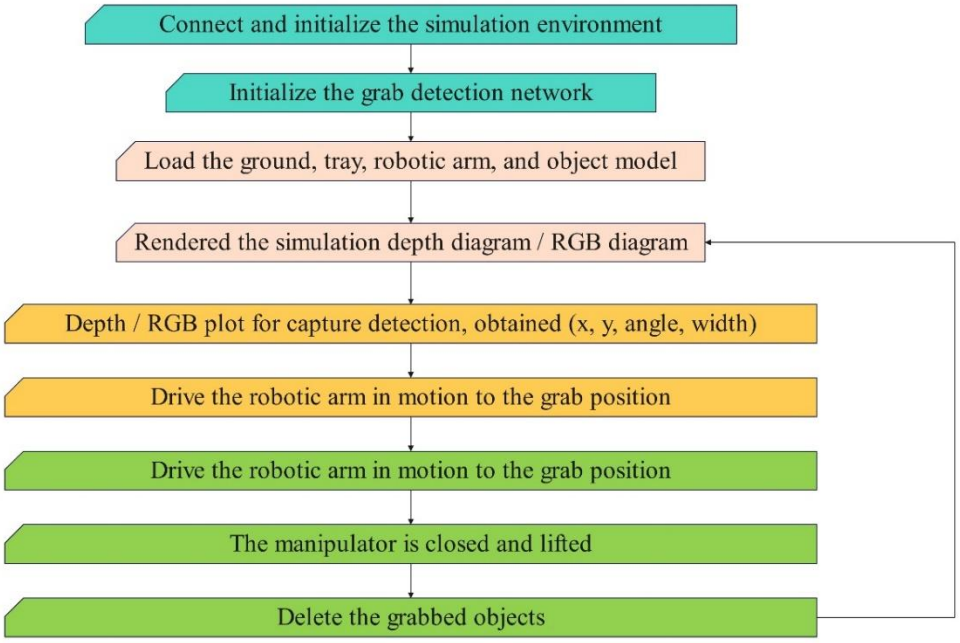


Figure 6: Flowchart of grasping detection

Combining with Figure 6, the grab detection process is analyzed as follows:

In the complete grasping detection process, it mainly includes four parts: initialization of simulation environment and detection network, loading of simulation environment and object model, grasping pose detection of CRE-Net network, and grasping motion of robotic arm. The mechanical arm grasping steps are as follows:

(1) Initialization of simulation environment and detection network: It refers to initializing the Pybullet simulation environment and CRE-Net network.

(2) Loading of simulation environment and object model: it refers to loading models such as ground, tray, robotic arm, grasping object, etc. in Pybullet simulation environment.

(3) Grasping pose detection of CRE-Net network: In the simulation environment, the object model is detected by a Depth camera, and the image is rendered. The collected Depth and RGB images are predicted by CRE-Net network to obtain the key point information (x,y,w,θ) of the grasping pose with the greatest confidence, and the grasping pose is marked on the image, and then the maximum grasping Depth is calculated according to collision detection to provide grasping information for the robotic arm grasping;

(4) Grasping motion of robotic arm: Finally, the grasping pose information is transmitted to the mechanical arm driving mechanism, and the mechanical arm will move to the specified position and pose according to the grasping information to grasp the object. After reaching the specified position, the gripper will open and grasp the object. When the object is successfully grasped, the grasped object model is deleted in the simulation space, and the next round of object grasping detection is carried out.

4 Test

4.1 Test methods and environment

Aiming at the grasping environment of complex multiple unknown objects with messy stacking, this paper proposes an autonomous grasping algorithm of robotic arm based on deep learning, which efficiently solves the problems of easy collision, poor real-time performance and difficulty in capturing reasonable end pose in multi-object stacking scenarios. To improve the performance comparison effect of the model in this article, two scenarios of barrier free and barrier free were set up for comparison, thereby enhancing the intuitive comparison effect of the model The barrier free environment means that there are no other obstacles nearby, only robotic arms and objects to be grasped in environments with obstacles, there are many obstacles in the surrounding environment, which can directly affect the movement path of the robotic arm. Detailed algorithms are needed for path planning to improve the success rate of grasping.

The data set in this paper is a self built data set, which collects tens of thousands of pieces of data through the common objects in life and multiple industrial production parts as the captured objects, and carries out the following comparative test through these data. In addition, in order to further verify the generalization ability of the model, DEX net and YCB data sets are introduced for experiments, with 70% of the data as the training set and 30% of the data as the prediction set.

The Actor network learning rate is set to 0.0003, the Critic network learning rate is set to 0.0003, the soft λ_2 update rate is set to 1, set λ_1 to 1,set N_{step} to 300,set T_S to 3000, c_1 set to 0.05, set c_2 to 1.6, set k_1 to 1, set k_2 to 0.05, set k_3 to 0.1, set to 2000 rounds. set γ to 0.97, set φ to 0.97, set n_f to 0.97.

The model in this article combines images and laser sensors to achieve dual path obstacle avoidance technology, providing reliable reference for the motion path planning of robotic arms and promoting the improvement of gripping efficiency. The trained model is actually tested on the simulated test data set and the data set collected in the real environment, and related ablation experiments and algorithm comparison experiments are carried out. Finally, aiming at the messy stacked multi-object scene, the real robotic arm and the algorithm proposed in this paper are used to grasp tests in a real grasping environment to verify the superiority of this algorithm [30]. The test platform is shown in Figure 7 below.

Figure 7: The test platform

The robot arm model is set to ur10, the camera model is inter realsense d415, the gripper model is z-efg, the processor (CPU) model is Intel Core i9-7900x, and the memory is 128G. NVIDIA GTX 1080ti 2 is selected as the system graphics card, and the memory is 128g.The test hardware equipment is shown in Table 2 below:

Table 2: Test environment

Hardware equipment	Equipment model
Robotic arm model	UR10
Camera model	Inter Realsense D415
Gripping jaw model	Z-EFG
Processor (CPU) model	Intel Core i9-7900X
Processor memory	128G
Graphics card (GPU) model	Nvidia GTX 1080Ti×2
Graphics card memory	128G

A series of software service platforms need to be used when building an autonomous grasping algorithm for robotic arms based on deep reinforcement learning. Among them, the source code and programs of this paper are running under the Linux operating system of Ubuntu16.04. The version of the ROS-based robot operating system is Kinetic. The source code for network construction and part of control uses Python as the main programming language, and the deep learning framework for the network construction part uses the Pytorch framework, which is known for its simplicity and efficiency. Secondly, this paper utilizes two deep learning acceleration libraries, CUDA and Cudnn, which can efficiently perform basic operations of deep neural networks on GPU, such as convolution, pooling, normalization, and activation layers, and can greatly reduce the training and inference time of the network.

REGRAD data set is a multi-dimensional information set containing visual information, grasping information, segmentation information and other information in chaotic stacking scenes. In this paper, the REGRAD capture data set is selected as the training data set and partial test data set for lightweight generative deep reinforcement learning based on capture priority. The REGRAD data set contains sufficient capture information and operational relationship information between multiple objects, which can provide data guarantee for multi-object capture in messy stacking scenarios.

The improved algorithm named AI-PAC-DRL is used for intelligent grasping of two kinds of robotic arms.

SAC-PER and AI-PAC-DRL are experimentally validated. First, we will compare the learning effects and total training time of the two algorithms during the training process to analyze the advantages of the AI-PAC-DRL algorithm. Then, the improvement of MAML algorithm in learning effect and generalization performance is analyzed by testing experiments

4.2 Results

By analyzing the learning curve of the model during the training process, the performance change of the model when the training data is continuously increased can be measured, and the performance of different algorithms can be well compared. Therefore, the SAC-PER algorithm and MAMLSAC algorithm are used to train 10000 rounds in a barrier-free environment, and the obtained learning curve is shown in Figure 8:

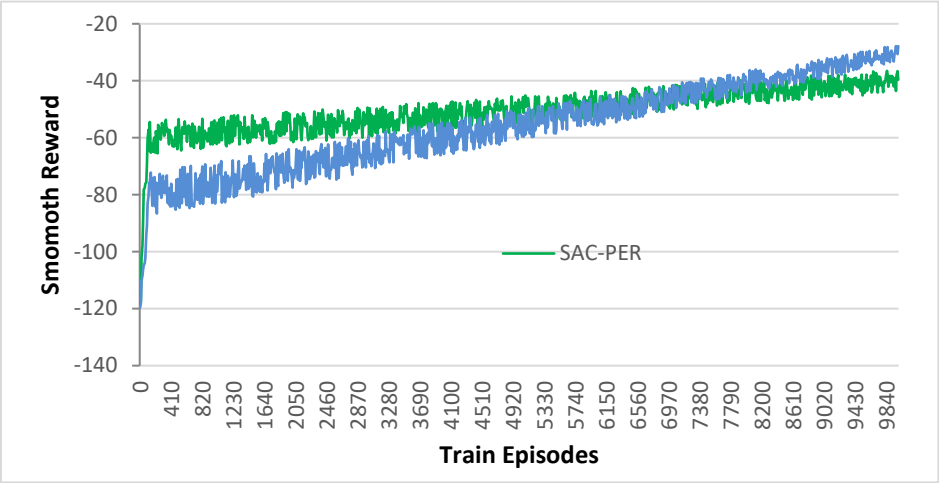


Figure 8: Performance comparison between AI-PAC-DRL algorithm and SAC-PER algorithm in barrier-free environment

In an environment with obstacles, the same setting as without obstacles is adopted, which requires 10,000 rounds of training, and the AI-PAC-DRL algorithm needs to use a base learner to quickly adapt in multiple sub-tasks. The final result of the learning curve is shown in Figure 9:

The trained AI-PAC-DRL algorithm and SAC-PER algorithm are used to conduct 50 tests in eight task scenarios. The final test results are shown in Figure 10:

According to the test, the grasping success rate of each object is shown in Figure 11 and Table 3.

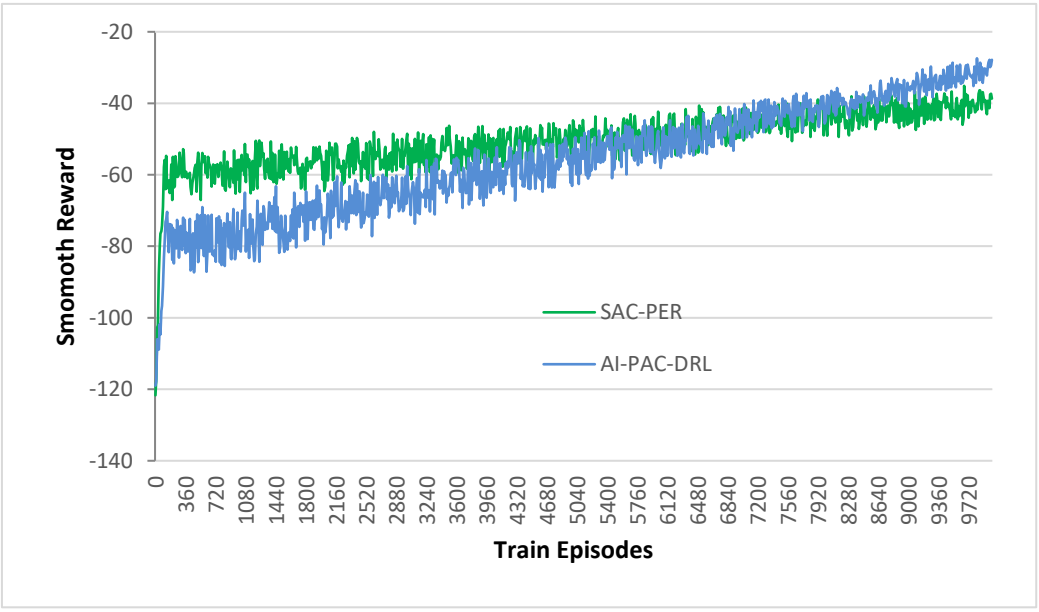


Figure 9: Performance comparison between AI-PAC-DRL algorithm and SAC-PER algorithm in obstacle environment

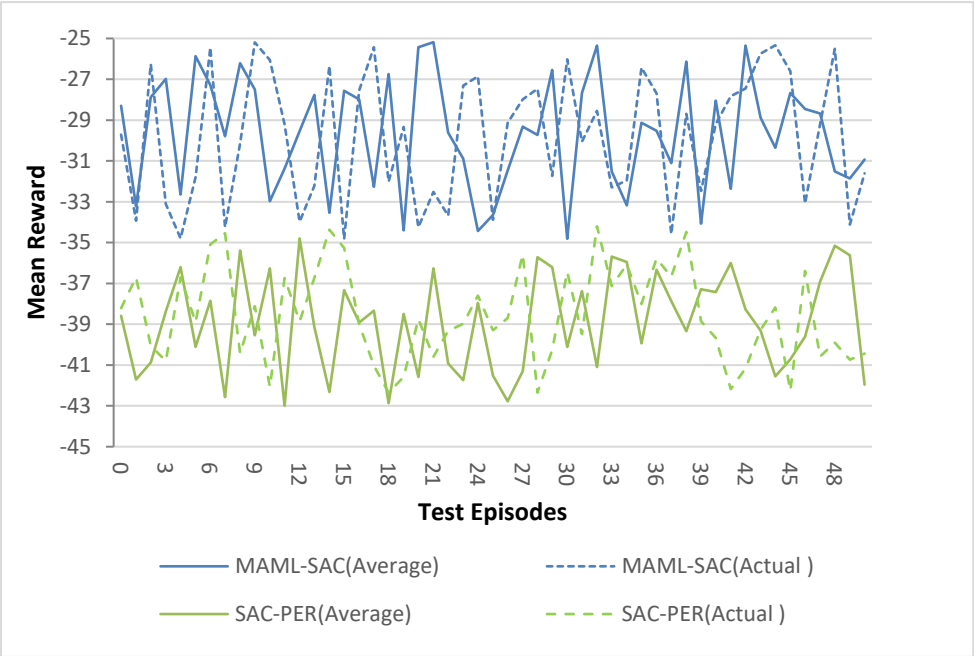


Figure 10: Test results of SAC-PER algorithm and AI-PAC-DRL algorithm in barrier-free environment

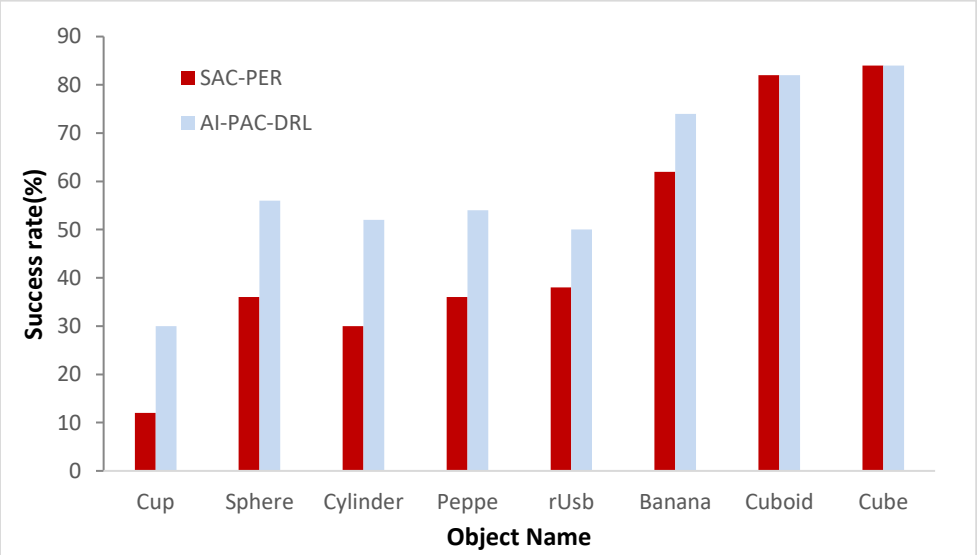


Figure 11: Capture success rate of SAC-PER algorithm and AI-PAC-DRL algorithm in barrier-free environment

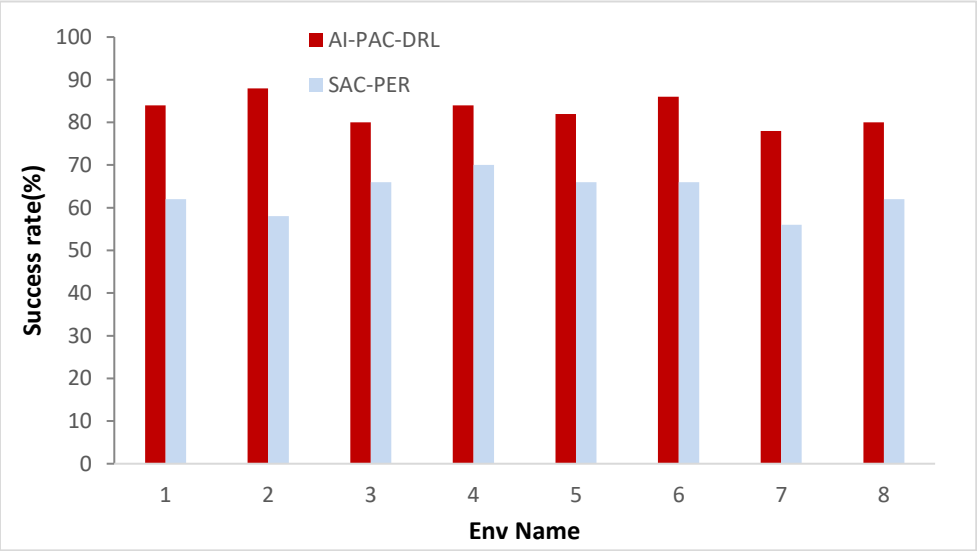


Figure 12: Capture success rate of SAC-PER algorithm and AI-PAC-DRL algorithm in obstacle environment

Table 3: statistical comparison of grab success rate in barrier free environment

	SAC-PER	STD	AI-PAC-DRL	STD	t	P
Cup	12	0.4781	30	0.9749	-3.141	0.00012
Sphere	36	3.2632	56	2.6057	-2.874	0.00017
Cylinder	30	2.2237	52	2.6778	-2.836	0.00032
Peppe	36	3.2369	54	5.1105	-2.654	0.0012
rUsb	38	1.3181	50	3.2562	-2.004	0.0023
Banana	62	5.2019	74	3.8768	-1.212	0.0043
Cuboid	82	6.1623	82	5.9715	0.0021	0.012
Cube	84	2.5647	84	6.9986	0.0011	0.024

After grasping 50 times in each obstacle environment, the test results are shown in Figure 12 and Table 4.

To further verify the complexity of PAC algorithm, the general framework of Robot Obstacle Avoidance Trajectory Planning proposed in this chapter is tested by comparing DRL benchmark algorithms (DDPG, TD3 and SAC) and their corresponding PAC based algorithms (DDPG _pac, TD3_PAC and SAC _ PAC). The test was carried out under the scenario of self built dataset, and Table 5 was obtained.

Table 4: statistical comparison of capture success rate in obstacle environment

	AI-PAC-DRL	STD	SAC-PER	STD	t	P
1	84	4.6119	62	5.7911	-2.921	0.0031
2	88	6.0713	58	5.7865	-3.342	0.00019
3	80	3.7575	66	4.0904	-1.937	0.00065
4	84	5.0530	70	3.4561	-1.532	0.00035
5	82	7.1603	66	4.7836	-0.966	0.00042
6	86	6.3737	66	2.3031	-2.642	0.00064
7	78	5.4021	56	1.9712	-2.024	0.00047
8	80	7.7030	62	2.4861	-1.632	0.00050

Table 5: Complexity analysis of PAC algorithm

Algorithm	Success rate (%)			Training time (Hour)
	1-500	501-1000	1001-2000	
DDPG	6.73	89.50	97.22	24.65
DDPG_PAC	81.77	91.67	98.31	20.79
DDPG_EDS	32.67	97.22	97.71	21.98
DDPG_COR	37.22	99.00	97.81	23.36
DDPG_PEC	89.10	91.08	97.42	20.30
TD3	0.00	70.29	97.02	26.04
TD3_PAC	96.03	95.44	98.70	21.58
TD3_EDS	22.57	98.60	98.41	23.27
TD3_COR	15.84	99.00	98.31	23.66
TD3_PEC	91.87	96.43	98.80	21.29
SAC	3.96	95.24	97.32	24.75
SAC _PAC	95.24	97.02	98.80	20.30
SAC _EDS	40.99	98.60	97.12	23.46
SAC _COR	30.10	98.21	97.81	23.96
SAC _PEC	97.42	96.62	98.60	19.80

The effect of each module on the performance of the model was verified by ablation experiment the benchmark model is the complete model (PAC+EDS+COR), The evaluation indexes are: obstacle avoidance success rate (environmental complexity is divided into low, medium and high levels); Training convergence speed (the number of training steps required to reach 90% success rate); Sample utilization rate (number of effective policy updates per step) The ablation results are shown in Table 6 below.

Table 6: results of model ablation test

Model	Obstacle avoidance success rate (%)			Convergence speed (training steps)	Sample utilization rate (times/thousand steps)
	High	Medium	Low		
PAC + EDS + COR	98.01%	94.05%	87.12%	11880	45
EDS + COR	91.08%	82.17%	69.30%	17820	38
PAC + COR	94.05%	86.13%	74.25%	14850	28
PAC + EDS	84.15%	71.28%	57.42%	19800	41

In order to further verify the generalization effect and practical effect of this model, the algorithm of this model is compared with the existing algorithms, mainly including binocular vision obstacle avoidance technology (bvoa), 3D structured light technology (3D SLT), laser radar technology (3D TOF), visual slam and dynamic path planning (slam-dpp). In the obstacle environment, 3000 rounds of training are carried out to make the model have a certain obstacle avoidance ability. The data set is mainly DEX, YCB data, after that, the model is used to count the success rate of grasping in complex environment, and the test results shown in Table 7 below are obtained.

Table 7: Comparison of model performance under different data sets

Data set	DEX	YCB
BVOA	75.36%	72.52%
3D SLT	80.21%	78.36%
3D TOF	84.36%	83.01%
SLAM-DPP	85.32%	82.35%
AI-PAC-DRL	91.32%	89.35%

4.3 Analysis and discussion

As shown in Figure 8, the blue line represents the learning curve of the AI-PAC-DRL algorithm, the green line represents the learning curve of the SAC-PER algorithm, and the hatched portion represents the actual reward value obtained before the smoothing process. By analyzing the reward curve, it can be seen that the AI-PAC-DRL algorithm has obvious oscillation in the early stage and its learning speed is slower than that of SAC-PER algorithm. The reason is that the AI-PAC-DRL algorithm needs to be trained in multiple sub-task environments in turn in the early stage, and only 125 rounds of training are performed in each sub-task environment. However, after 2000 rounds of meta-reinforcement learning, the AI-PAC-DRL algorithm obtains better initial network parameters through summarizing experience, which can quickly converge to the optimal value along the gradient direction. Therefore, the learning speed is greatly improved, and finally it converges to a higher reward value. Compared with the convergence speed, the AI-PAC-DRL algorithm converges after about 6000 rounds, while the SAC-PER algorithm approaches convergence after about 8000 rounds, which shows that the AI-PAC-DRL algorithm has faster learning speed. Therefore, it shows that the AI-PAC-DRL algorithm has better learning effect than the SAC-PER algorithm

As can be seen from Figure 9, the blue portion represents the reward obtained by the AI-PAC-DRL algorithm, and the green portion represents the reward obtained by the SACPER algorithm. By analyzing the reward curve, it can be seen that in the early stage of training, the convergence speed of AI-PAC-DRL algorithm is relatively slow, and the reward value obtained is lower than that of SAC-PER algorithm. The reason is that the AI-PAC-DRL algorithm learns fewer rounds in each subtask than the SAC-PER. However, with the increase of training rounds, at about 4000 rounds, the AI-PAC-DRL algorithm improves the rapid adaptability of network parameters by summarizing the characteristics of obstacles in sub-task scenarios, making the performance of the algorithm improved, so the reward obtained after the final convergence is higher than that of the SAC-PER algorithm. Then, from the analysis of the overall convergence speed, it can be found that the AI-PAC-DRL algorithm also reaches convergence 2000 rounds earlier than the SACPER algorithm, which shows that the AI-PAC-DRL algorithm has faster learning speed and better performance than the SAC-PER algorithm.

As can be seen from Figure 10, the solid blue line represents the average reward obtained by the AI-PAC-DRL algorithm in the test phase, while the dotted line represents the actual reward obtained by the AI-PAC-DRL algorithm in each environment. The solid green line represents the average reward obtained by the SAC-PER algorithm, and the dotted line represents its actual reward. By comparing the average reward curves of the two algorithms, it can be seen that the AI-PAC-DRL algorithm performs better than the SAC-PER algorithm in both the average reward obtained and the individual reward in each environment. It shows that the AI-PAC-DRL algorithm can summarize the best grasping strategy from objects with different shapes, thus improving the robustness of the algorithm and having higher success rate and reliability.

As can be seen from Figure 11, the red part represents the success rate of the AI-PAC-DRL algorithm test, and the blue part represents the success rate of the SAC-PER algorithm. By analyzing the histogram, it can be seen that the AI-PAC-DRL algorithm has a certain improvement in the success rate of capturing each object. In particular, cup-shaped objects, cylindrical objects, spherical objects, etc., which are quite different in shape from the target objects in the training stage, have better success rates. This shows that the AI-PAC-DRL algorithm has better generalization performance than the SAC-PER algorithm in the obstacle-free environment. In Table 3, from the standard deviation, t value and P value, the data are basically statistically significant, which verifies the effectiveness of the comparative test data.

It can be seen from Figure 12 that the AI-PAC-DRL algorithm has a higher capture success rate in obstacle scenes at different angles. This shows that the meta-learner on AI-PAC-DRL can summarize the characteristics of obstacles well, thus improving the adaptability of the algorithm to new tasks, making the algorithm learn better on the original basis, and obtaining higher success rate. Therefore, it can be concluded that the AI-PAC-DRL algorithm has better generalization performance as well as adaptability to unknown tasks than the SAC-PER algorithm. In Table 4, from the standard deviation, t value and P value, the data are basically statistically significant, which verifies the effectiveness of the comparative test data.

For some failure cases, the main reasons are as follows:

(1) Dynamic obstacles and complex background interference

The trajectory of dynamic obstacles in a chaotic environment is unpredictable, and the background noise (such as irregular object stacking) will interfere with the positioning and segmentation of the target object by the vision system, resulting in the failure of obstacle avoidance path planning or the offset of grasp coordinates. For example, transparent or reflective objects may make the visual sensor unable to accurately capture the depth information, causing error in obstacle avoidance judgment. In addition, some cluttered environments may lead to the manipulator unable to reach the front of the object through an effective path, leading to the inevitable failure of the robot to grasp the object.

(2) Object occlusion and light change when the target object is partially occluded by other objects, the robot cannot obtain complete visual feature information, resulting in the error of grasping attitude estimation; At

the same time, uneven or sudden change of ambient light will reduce the image quality and affect the visual positioning accuracy.

(3) Sensor limitations

It is difficult for a single vision sensor (such as a monocular camera) to achieve high-precision 3D reconstruction in complex scenes, especially when the color of the target object is close to the background, which is easy to lead to the error of obstacle avoidance decision.

In Table 5, the DRL algorithm based on PAC shows faster convergence speed, less average round steps and higher success rate than the benchmark algorithm. Especially in terms of round rewards, Ddpg_pac, td3_pac and sac_pac are all in the early stage of training, and quickly reach a high round reward value, which means that the PAC proposed can provide good expert strategy guidance for the manipulator, so that the manipulator can learn the obstacle avoidance strategy in a very short time.

Each DRL algorithm with the proposed complete PEC framework (that is, PAC, EDS and cor are combined at the same time) maintains a high success rate in the whole training process, and only needs to consume the least training time in almost all scenarios, which shows that the proposed PEC framework has a strong ability in optimizing strategy decision, improving algorithm learning efficiency and enhancing action exploration ability,

In Table 6, after PAC is removed, the success rate of the model in high complexity environment decreases by 18%, and the convergence speed slows down significantly (+6k steps), indicating that PAC improves the stability and generalization of strategy through expert strategy self optimization ability; After removing EDS, the sample utilization rate was reduced by 38%, which verified that the dual memory mechanism optimized the data utilization efficiency through balanced exploration and utilization; The removal of cor resulted in a 14% and 23% reduction in the success rate of low/medium complexity environments, respectively, indicating that cor effectively alleviated the sparse reward problem and enhanced the adaptability to concave polyhedral obstacles through the composite reward function.

In general, PAC is the core component of the model to deal with complex environments through the ability of policy self optimization and scene generalization. The efficient sampling mechanism of EDS significantly improves the training efficiency and reduces the data demand. The compound reward function of cor solves the local optimal and sparse reward problems of the traditional obstacle avoidance model.

In Table 7, AI-PAC-DRL has certain advantages over the existing more advanced models in terms of capture success rate. Compared with SLAM-DPP, which has the best performance in the existing algorithms, AI-PAC-DRL has improved its performance in dex data set by 6% and YCB data set by 7% In general, AI-PAC-DRL outperforms the existing model algorithms in robot path planning and object grasping.

The experimental results show that the improved deep reinforcement learning algorithm has faster convergence speed and learning effect in the training stage, and has higher grasping success rate and generalization energy in the testing stage.

In this paper, the model is trained in a simulated environment, and the capture scene is arranged through the actual environment. From the actual situation, even if a more complex environment is set, the capture accuracy can reach more than 80% after the simulation training The training effect, accuracy and application effect of the model can be further improved through multiple practical operations from this courseware, the model can effectively complete the transition from simulation to reality.

5 Conclusion

This study proposes a collaborative optimization framework based on deep reinforcement learning, which aims to break through the problem of precise grasping of manipulator in complex environment.

The purpose of this paper is to improve the robot's obstacle avoidance effect based on the existing model, and to solve the problem of customer service data training, so as to improve the robot's intelligent control effect and autonomous decision-making ability. Experimental results show that the proposed PAC has the ability of efficient exploration, significantly accelerates the convergence process of the algorithm, reduces the number of training steps, and improves the success rate. While amplifying expert experience data through xGAIL, unbiased dual memory bank sampling rules are used to improve training efficiency. Then, through the increasingly unbiased sampling mechanism, the guiding effect of expert experience data in the early stage of algorithm training is improved, and the influence of expert experience data on the convergence of the algorithm is reduced and gradually eliminated. The experimental results show that the proposed EDS can improve the utilization efficiency of expert experience data, and the success rate of the algorithm is also significantly improved.

Based on the analysis of the failure cases in the experiment, the reasons are obtained, and the research direction of the follow-up paper is obtained. Firstly, the countermeasure training can be introduced to generate a variety of chaotic environment samples to improve the generalization ability of the model; Secondly, hierarchical reinforcement learning (HRL) was used to separate obstacle avoidance and grasping subtasks, or combined with long-term and short-term memory (LSTM) to deal with temporal dependence; Finally, stage rewards (such as bonus points for approaching the target and deduction points for collision) are designed to

balance obstacle avoidance and capturing the target Through the combination of theory and experiment, the model is further verified and improved, and its practical application effect is improved.

Data availability statement

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

References

[1] Arshad, J., Qaisar, A., Rehman, A. U., Shakir, M., Nazir, M. K., Rehman, A. U., Eldin, E. T., Ghamry, N. A & Hamam, H. (2022). Intelligent control of robotic arm using brain computer interface and artificial intelligence. Applied Sciences, 12(21), 10813-10824. DOI:10.3390/app122110813

[2] Xu, K., & Wang, Z. (2023). The design of a neural network-based adaptive control method for robotic arm trajectory tracking. Neural Computing and Applications, 35(12), 8785-8795. DOI:10.1007/s00521-022-07646-y

[3] Dai, Y., Xiang, C., Zhang, Y., Jiang, Y., Qu, W., & Zhang, Q. (2022). A review of spatial robotic arm trajectory planning. Aerospace, 9(7), 361-372. DOI:10.3390/aerospace9070361

[4] Rawat, D., Gupta, M. K., & Sharma, A. (2023). Intelligent control of robotic manipulators: a comprehensive review. Spatial Information Research, 31(3), 345-357. DOI:10.1007/s41324-022-00500-2

[5] Zaitceva, I., & Andrievsky, B. (2022). Methods of intelligent control in mechatronics and robotic engineering: A survey. Electronics, 11(15), 2443-2454. DOI :10.3390/electronics11152443

[6] Abdi, A., Ranjbar, M. H., & Park, J. H. (2022). Computer vision-based path planning for robot arms in three-dimensional workspaces using Q-learning and neural networks. Sensors, 22(5), 1697-1705. DOI :10.3390/s22051697

[7] Ai, J., Meng, J., Mai, X., & Zhu, X. (2023). BCI control of a robotic arm based on ssvep with moving stimuli for reach and grasp tasks. IEEE Journal of Biomedical and Health Informatics, 27(8), 3818-3829. DOI: 10.1109/JBHI.2023.3277612

[8] Mourtzis, D., Angelopoulos, J., & Panopoulos, N. (2022). Closed-loop robotic arm manipulation based on mixed reality. Applied Sciences, 12(6), 2972-2982. DOI:10.3390/app12062972

[9] Zhou, Y., Xie, L., & Pan, H. (2022). Research on a PSO-H-SVM-based intrusion detection method for industrial robotic arms. Applied Sciences, 12(6), 2765-2773. DOI:10.3390/app12062765

[10] Singh, B., Kumar, R., & Singh, V. P. (2022). Reinforcement learning in robotic applications: a comprehensive survey. Artificial Intelligence Review, 55(2), 945-990. DOI:10.1007/s10462-021-09997-9

[11] Abdullah-Al-Noman, M., Eva, A. N., Yeahyea, T. B., & Khan, R. (2022). Computer vision-based robotic arm for object color, shape, and size detection. Journal of Robotics and Control (JRC), 3(2), 180-186. DOI :10.18196/jrc. v3i2.13906

[12] Xu, B., Li, W., Liu, D., Zhang, K., Miao, M., Xu, G., & Song, A. (2022). Continuous hybrid BCI control for robotic arm using noninvasive electroencephalogram, computer vision, and eye tracking. Mathematics, 10(4), 618-624. DOI :10.3390/math10040618

[13] Liu, Z., Peng, K., Han, L., & Guan, S. (2023). Modeling and control of robotic manipulators based on artificial neural networks: a review. Iranian Journal of Science and Technology, Transactions of Mechanical Engineering, 47(4), 1307-1347. DOI:10.1007/s40997-023-00596-3

[14] Tang, Z., Wang, P., Xin, W., & Laschi, C. (2022). Learning-based approach for a soft assistive robotic arm to achieve simultaneous position and force control. IEEE Robotics and Automation Letters, 7(3), 8315-8322. DOI: 10.1109/LRA.2022.3185786

[15] Ahmed, A., Yu, M., & Chen, F. (2022). Inverse kinematic solution of 6-DOF robot-arm based on dual quaternions and axis invariant methods. Arabian Journal for Science and Engineering, 47(12), 15915-15930. DOI: 10.1007/s13369-022-06794-6

[16] Kim, E., Shin, J., Kwon, Y., & Park, B. (2023). EMG-based dynamic hand gesture recognition using edge AI for human–robot interaction. Electronics, 12(7), 1541-1550. DOI:10.3390/electronics12071541

[17] Wang, C., Li, C., Han, Q., Wu, F., & Zou, X. (2023). A performance analysis of a litchi picking robot system for actively removing obstructions, using an artificial intelligence algorithm. Agronomy, 13(11), 2795-2803. DOI:10.3390/agronomy13112795

[18] Tran, D. T., Vo, H. Q., Nguyen, T. K., & Nguyen, T. N. (2025). Enhanced teleoperation and visual-force feedback with obstacle avoidance for a car-like mobile robot based on WAN network architecture. Journal of Technical Education Science, 20(01), 62-72. DOI :10.54644/jte.2025.1601

[19] Wu, P., Su, H., Dong, H., Liu, T., Li, M., & Chen, Z. (2025). An obstacle avoidance method for robotic arm based on reinforcement learning. Industrial Robot: the international journal of robotics research and application, 52(1), 9-17. DOI: 10.1108/IR-05-2024-0206

[20] Ricárdez Ortigosa, A., Bestmann, M., Heilemann, F., Halbe, J., Christiansen, L., Rodeck, R., & Wende, G. (2025). Foundations for teleoperation and motion planning towards robot-assisted aircraft

fuel tank inspection. Aerospace, 12(2), 156. DOI :10.3390/aerospace12020156

[21] Raj, R., & Kos, A. (2025). An extensive study of convolutional neural networks: applications in computer vision for improved robotics perceptions. Sensors, 25(4), 1033. DOI: 10.3390/s25041033

[22] Ušinskis, V., Nowicki, M., Dzedzickis, A., & Bučinskas, V. (2025). Sensor-fusion based navigation for autonomous mobile robot. Sensors, 25(4), 1248. DOI :10.3390/s25041248

[23] Wang, D., Liu, B., Jiang, H., & Liu, P. (2025). Path planning for construction robot based on the improved a* algorithm and building information modeling. Buildings, 15(5), 719. DOI: 10.3390/buildings15050719

[24] Zhu, W., Gao, X., Wu, H., Chen, J., Zhou, X., & Zhou, Z. (2025). Design of multimodal obstacle avoidance algorithm based on deep reinforcement learning. Electronics, 14(1), 78. DOI:10.3390/electronics14010078

[25] Zhang, C., Zhang, X., Yang, W., Zhang, G., Wan, J., Lei, M., & Dong, Z. (2025). Safe path planning method based on collision prediction for robotic roadheader in narrow tunnels. Mathematics, 13(3), 522. DOI: 10.3390/math13030522

[26] Benguedoudj, A., Belhadj, F., Hu, Y., Zitouni, B., Idir, Y., Adoui, I., & Mostefai, M. (2025). Efficient line-based visual marker system design with occlusion resilience. *Informatica*, *49*(1). DOI: 10.31449/inf. v49i1.7259

[27] Aradea, A., Rianto, R., Herlina, N., & Hoeronis, I. (2025). Self-learning model for pattern recognition in vision system based on adaptive kernel. Informatica, 49(14). DOI: 10.31449/inf. v49i14.7272

[28] Qian, P. (2025). Dual-layer dynamic path optimization for airport ground equipment using graph theory and adaptive genetic algorithms. Informatica, 49(13). DOI: 10.31449/inf. v49i13.7651

[29] He, X., Wang, R., Cao, T., Liang, W., & Fan, Y. (2025). Fusion CNN-transformer model for target counting in complex scenarios. Informatica, 49(12). DOI: https: 10.31449/inf. v49i12.7315

[30] Abdillah, M. S., Masykur, F. and Sasmowiyono, S. S. (2025). Design and construction of a smart wheelchair from a lecture chair using power window motors with smartphone control for disability. International Journal of Engineering, 38(6), 1397-1404. DOI: 10.5829/ije.2025.38.06c.12

Multi-task Learning for Intelligent Portrait Composition: Deep Residual Networks and Human Pose Estimation

Di Shao
School of Journalism and Communication, Anhui Broadcasting Movie and Television College
Hefei 230000, China
Email: amtccc@163.com

Keywords: multi-task learning, human pose, portrait, composition, deep residual network, adaptive point set

Received: January 23, 2025

Traditional photography composition is difficult to meet the current needs and aesthetic preferences of portrait photography. Therefore, to improve the composition quality and efficiency of portrait photography, an intelligent composition technology combining multi-task learning framework and deep learning technology is proposed. Firstly, taking the deep residual network as the basic skeleton, the lightweight classification network MobileNet V2 is introduced to improve the portrait detection performance. Afterwards, image quality is improved through the multi-scale feature fusion and multi-task learning, and the scale feature extraction is performed on the input image. Moreover, the human pose estimation network is used to detect human keypoints, dynamically adapting to different human poses and scales. The relay node is used to associate the relationship between human instances and keypoints, improving the intelligent composition effect. The effectiveness of the research method is analyzed from three aspects: intelligent cropping for portrait detection, pose recognition analysis, and synthesized image quality. The MS COCO and MuPoTS-3D datasets are selected for evaluation, including recall rate, F1-value, average Intersection over Union ratio (Avg IoU), mismatch, Area under Curve (AUC), and precision-recall curve (PCK). The results showed that the proposed MobileNet V2-ResNet50 achieved an accuracy of 94.51% in extracting facial image information, with an Avg IoU of 0.69, while the Avg IoU of the other three comparison methods was less than 0.65. The proposed fusion algorithm had information entropy, structural similarity, average structural similarity, peak signal-to-noise ratio, brightness relationship factor, and mutual information of 7.596, 1.129, 7.081, 1.828, 1.078, and 8.826, respectively. The overall image quality fusion effect was significantly better than other algorithms. The MobileNet V2-ResNet50 network significantly outperforms the UNet-Transformer-CBAM and MCTN models in terms of computational efficiency, with the floating-point operations of only 2.5B, parameter count of 5M, and inference time of 15ms. The research method achieved a human pose detection accuracy of 88.94% on the dataset, with AUC and PCK scores of 44.99 and 83.24, respectively. The detection accuracy of RSC-MS, OP-GAN, and PGF-HPE models did not exceed 88.5%, and their AUC and PCK scores differed significantly from the research method. This method can effectively identify and optimize portrait composition, enhancing the visual effect and artistic expression of photos.

Povzetek: Opisano je večopravkovno učenje z ResNet in MobileNet V2 za inteligentno kompozicijo portretov, ki združuje zaznavo poze, obrezovanje in estetsko optimizacijo z visoko natančnostjo.

1 Introduction

With the widespread popularity of electronic devices and changes in information dissemination media, the image, as an information carrier, presents an explosive growth trend. Higher requirements have been put forward for the photography and composition skills of works [1]. A good composition not only presents complete and interesting image information, but also brings strong visual impact and clear thematic sense to people, endowing ordinary things with unique beauty [2]. Photography needs and aesthetic preferences make people more inclined to crop images to improve aesthetic quality. Portrait photography not only requires capturing the image of a person, but also conveying emotions and imagery through clever composition. Traditional photography composition relies on the photographer's experience and aesthetics. However, intelligent composition uses artificial intelligence technology to automatically identify and analyze shooting scenes, optimize image layout, and make photos more visually impactful and aesthetically pleasing [3]. In previous studies, intelligent composition in portrait photography mainly relied on a single image processing technique, such as edge detection, feature matching, etc. Although these methods have achieved certain results, they have obvious limitations. Farhat et al. used a photography composition assistance framework to improve composition effects. A fork join framework was used to achieve photography composition, which included designing from aspects

such as object detection, photo classification, and pose clustering analysis. The results indicated that this method had good photographic application effects [4]. Li et al. used lightweight networks to preserve fine-grained details in images for portrait details in complex backgrounds. The experimental results showed that this method had good visual effects and performed well in evaluation indicators [5]. Considering that existing automatic image cropping methods often use image boundaries as boundaries, it is difficult to improve the aesthetic appeal of image cropping. Therefore, Zhong et al. used the field of view evaluation model to determine the extension of image content and implement cropping based on image quality. The results indicated that this method could effectively improve the visual image quality [6]. Wang et al. proposed a content aware image scaling mechanism based on composition detection and composition rules, which optimized the visual effect of images by calculating composition rules. The results indicated that this method could preserve the image content structure and overall aesthetic appeal [7]. Li et al. used attribute assisted multimodal memory network to improve the quality of image aesthetic evaluation, inferred aesthetic quality by associating attribute perception semantic information. The results indicated that this method had good discriminative power on the dataset [8]. The relevant work situation is summarized, as shown in Table 1.

Table 1: Summary of related work

Literature	Method	Method Details	Performance results	Limitation
[4]	Photography composition auxiliary framework	Design photography composition from the aspects of object detection, photo classification, pose clustering analysis, etc	The photography composition effect is good and can meet personal photography preferences	Insufficient attention to image scale features
[5]	Lightweight network	Preserve fine-grained details of the image	The visual accuracy exceeds 90%	The accuracy of posture details is slightly lower
[6]	Field of View Evaluation Model	Crop based on image quality	Can preserve and preserve the content	Not considering the cropping relationship
			structure and overall aesthetic of the image	between images, scenes, and characters
[7]	Content aware image scaling mechanism	Calculate composition rules	Balancing image content structure and aesthetics	Strong dependence on rules, strong subjectivity, and limited adaptability to complex scenarios
[8]	Attribute assisted multimodal memory network	Association attribute perception semantic information to infer aesthetic quality	The difference in aesthetic quality rating is relatively small	Semantic information extraction has limitations and cross generalization ability

With the explosive growth in the number of images, portrait photography puts higher requirements on composition techniques. Traditional intelligent composition methods rely on a single image processing technique (e.g., edge detection, feature matching, and key point detection) and ignore task correlation, leading to poor results in complex scenes (e.g., multiple people, and occlusion). Existing methods have obvious deficiencies in speed, accuracy, and aesthetic optimization, making it difficult to meet users' needs for high-quality composition. In addition, traditional aesthetic evaluation methods mostly rely on manual rules and lack in-depth understanding of image content (such as human posture, emotional expression, etc.). Therefore, to construct an efficient and robust intelligent composition model, solve the composition optimization in complex scenes, achieve real-time and aesthetic quality improvement, and promote the practical application and popularisation of intelligent composition technology, the study proposes to improve it while using neural network processing. Specifically, to improve the quality of image, this study introduces a multi-task learning framework that integrates portrait detection, pose estimation, and intelligent cropping, optimizing image composition while adapting to diverse human poses. The proposed method enhances accuracy and artistic quality compared with conventional approaches. The implementation of multitask learning mainly focuses on regression tasks, adding classification and comparison tasks to calculate image quality score labels and

regression losses, respectively. Afterwards, the input image is extracted with different scales of features, and then dimensionality reduction, feature node generation, feature fusion and other steps are carried out. The corresponding quality score is used as a label to calculate the comparative loss composed of quality ranking loss and quality relative deviation loss. The quality score and label score of the image are calculated to obtain its coarse-grained and fine-grained features, thereby improving the image quality.

The innovation of the research lies in two aspects. One of them is the multi-task learning framework, which can improve image quality scores while achieving basic facial feature detection. The second is to jointly process human pose recognition and photography composition optimization as two closely related tasks, that is, based on the structure of human pose, further predicting and querying pose keypoints to grasp the semantic and positional information of each keypoint. The proposed method breaks the overly single learning focus of early research, strengthens the integration and collaboration between human pose recognition and photographic composition, and can provide more reference guidance for intelligent composition while considering the diversity of character poses and the artistic nature of photographic composition.

2 Methods

To achieve the design and real-time optimization of multi-task learning networks in the research objectives, this study first adopts a lightweight design strategy to construct an efficient multi-task detection model, which uses ResNet50 as the basic network architecture and MobileNet V2 as the support in the structural design part. The improvement approach of MobileNet V2 can enrich the output features of the receptive field and perform convolution operations on the reduced dimensional features to refine semantic information. The multi-task learning approach is used to calculate the regression task, classification task, and comparison task of the image separately, and calculate the image quality score label and regression loss separately. Different scale features are extracted from the input image, and the corresponding quality scores are used as labels to calculate the quality scores and label scores of the image, and the coarse-grained and fine-grained features of the image are obtained. By fusing feature representations, the computational complexity is reduced and the operational performance of the model is improved to meet application requirements. Secondly, to improve the accuracy of human pose analysis and optimize composition aesthetics, a multi-person pose estimation algorithm is proposed, which enables the network structure to dynamically adapt to different human poses and scales, and achieves intelligent portrait cropping and aesthetic preferences through keypoint detection and positioning, pose analysis, and adjustment.

2.1 Multi-task detection of portrait photography images based on lightweight design

The essence of image composition is the conscious arrangement of elements to convey specific emotions, information, and aesthetics. Intelligent composition, as a common image processing method, can effectively improve composition effects and enhance image quality. The Residual Network 50 (ResNet50) performs well in tasks such as image classification and object detection due to its powerful feature learning ability. Therefore, the research takes the ResNet50 as the basic network architecture. The MobileNet V2 is taken as the support in the structural design. A multi-task-based image composition optimization method and a lightweight design concept for portrait detection are designed to improve the quality of image evaluation. The point by point convolution processing of MobileNet V2 can form an inverted residual structure, which reduces the number of parameters while ensuring channel information extraction. The improvement ideas of convolutional networks are reflected in two aspects. One is to introduce width multipliers and resolution parameters into the original number of input layers, convolution kernel size, and output layer channel for processing. The second is to improve the accuracy of object detection by superimposing dilated convolutions in the output part of feature extraction. The proven performance of ResNet50 in image classification and object detection tasks makes it a reliable choice for multi-task learning. The lightweight MobileNet V2 architecture effectively optimizes computational efficiency and requires less deployment environment. Although EfficientNet provides a balance between accuracy and efficiency through composite scaling, it may not provide the same feature richness as ResNet50 for multi-task scenarios and consumes more computational resources. Visual transformers require large-scale datasets and computational resources, making them less suitable for real-time applications on mobile devices. ResNet50 provides the feature richness necessary for accurate multi-task detection, while MobileNet V2 ensures that the model remains lightweight and efficient. This combination achieves a balance between high performance and real-time applicability, which is crucial for intelligent image synthesis in portraiture. Figure 1 displays the detection network structure.

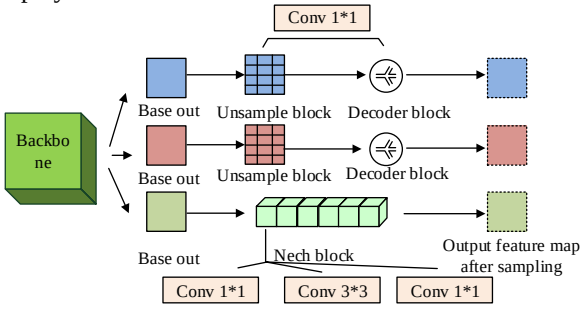


Figure 1: Schematic diagram of detection network structure

In Figure 1, the Backbone denotes MobileNetV2, which is the backbone part of the network and is responsible for extracting base features from the input image. Deep separable convolution reduces the number of parameters and computation and improves the efficiency of the model. Base out refers to the intermediate feature extraction output layer, which provides the initial feature representation as the subsequent input. The feature maps in this layer usually have high semantic information, but the sensory field may be limited. Upsample Block refers to the upsampling module, which is used to upsample the low-resolution feature maps to a high resolution. Decoder Block is used to fuse high-level semantic features with low-level detail features to improve the model's ability to capture target location and details. Neck Block is the Expansion Convolution Block, which generates features with multiple receptive fields through different expansion rates for enhanced feature extraction. The stacked Expansion Convolution Blocks can expand the receptive fields. The 1×1 convolution in this layer is to downscale the channel and reduce the computation. 3×3 convolution is to refine the semantic context information. The core of the improvement lies in the Neck Block design, which improves the sensory field and multi-scale detection ability of the model by expanding convolution and feature fusion. Afterwards, to improve the intelligent composition effect, the multi-task learning is used to enhance image quality, with regression tasks as the main focus, and additional classification and comparison tasks introduced to calculate quality score labels and regression losses. Feature maps extracted at different scales are processed through dimensionality reduction, node generation, and feature fusion to enhance image representation. Figure 2 is a schematic diagram of the multi-task learning process.

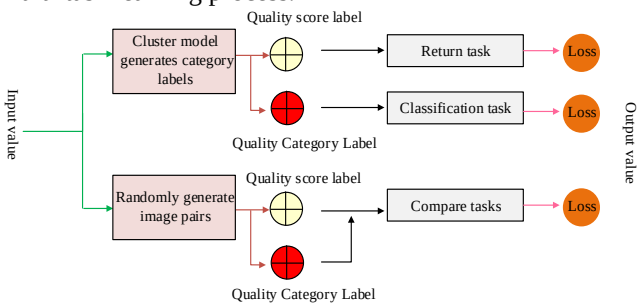


Figure 2: Schematic diagram of multi-task learning process

In Figure 2, the input images are subjected to category label generation and random image pair generation. Regression, classification, and comparison tasks are performed based on the quality score and label score of the image. The classification task first classifies the quality scores into two categories (good/bad) through a clustering model and then calculates the classification loss using a single image as input. The quality score of the image can be calculated based on the comparison loss under the combination of classification loss, regression loss, and other results. All quality scores are clustered to obtain image quality categories and confidence levels. The goal of classification task is to assign images to predefined categories to reduce subjective bias. It can be used as a supplement to regression tasks to reduce the subjective bias associated with a single regression task. The comparison task randomly selects image pairs from the image set, uses their quality scores as labels, and calculates the comparison loss composed of quality ranking loss and quality relative deviation loss. The corresponding quality score is used as the label to calculate the comparison loss consisting of quality ranking loss and relative quality deviation loss. Equation (1) can be used to represent the weighted sum of the loss function [9].

$$Loss = \alpha L_R + \beta L_C + \lambda L_B \tag{1}$$

In equation (1), L_R represents the regression loss. L_C represents the classification loss. L_B represents the comparison loss. α, β, λ are weights of the corresponding loss function. The introduced classification loss and comparative loss reduces subjective bias in quality scoring and enables the model to obtain coarse-grained and fine-grained features, improving the generalization ability of model sharing [10]. The study takes L1 norm loss to calculate the regression task of a single image, as shown in equation (2).

$$L_R = \frac{1}{N}\sum_{i=1}^{N}\left|S_i - \hat{S}_i\right| \tag{2}$$

In equation (2), N represents the number of images. S_i is the true score of the i-th image. $\hat{S}_i$ is the predicted score for the i-th image [11]. In portrait photography, some pixel values may be abnormal due to lighting or occlusion problems, the data may have noise or outliers, and L1 loss can better handle these situations. L1 loss can directly measure the pixel level error, which is more suitable for the continuity and local characteristics of image data. It is more inclined to produce sparse solutions, better preserving image details and edge information [12]. In image clustering processing, the Gaussian mixture model (GMM) is used to obtain its quality category. Equation (3) is the probability density function of the GMM.

$$p(X) = \sum_{k=1}^{K} w_k \cdot p(X \mid \mu_k, \sigma_k) \tag{3}$$

In equation (3), μ_k represents the mean. σ_k is the variance. w_k is the weight of the k-th Gaussian distribution. $p(X \mid \mu_k, \sigma_k)$ is the probability density function. GMMs model data can fit any complex data distribution through the linear combination of multiple Gaussian distributions. In image clustering tasks, image features (such as color and texture) often have complex distribution characteristics, which can be well captured by GMMs. Unlike K-means clustering methods, GMM can effectively evaluate the probability distribution of image quality categories, more flexibly define the boundaries between quality categories and adjust the

number of categories, so as to improve the clustering results [13]. By clustering all quality scores, image quality categories and confidence levels can be obtained. The prediction results of classified images can be obtained through the cross entropy loss function, which is shown in equation (4).

$$L_c = -[C_i \times \log(p_i) + (1 - C_i) \times \log(1 - p_i)) \quad (4)$$

In equation (4), C_i represents the quality classification of the image. p_i is the prediction probability of the image. The main task of comparing images is to determine the quality results of two images. Previous methods often encounter sorting errors when faced with inconsistent prediction ranking results [14]. To improve the performance of the model in comparing information, a loss function based on the concept of relative deviation is proposed for definition, as shown in equation (5).

$$L_{RD} = \left| \frac{S_1 - S_2}{S_1 + S_2} - \frac{\hat{S}_1 - \hat{S}_2}{\hat{S}_1 + \hat{S}_2} \right| \quad (5)$$

In equation (5), S_1 and S_2 represent the true scores of the image. $\hat{S}_1$ and $\hat{S}_2$ are the predicted scores for the image. To ensure the intelligent composition performance and scene applicability of portrait images, the above quality evaluation model has been light-weighted to improve the inference speed of the model. Firstly, the label data is expanded through knowledge distillation. The ResNet50 skeleton network is trained with lightweight sub networks and module fusion is performed. Considering that traditional knowledge distillation mostly focuses on discrete labels, the study investigates corresponding processing of image quality, as shown in Figure 3.

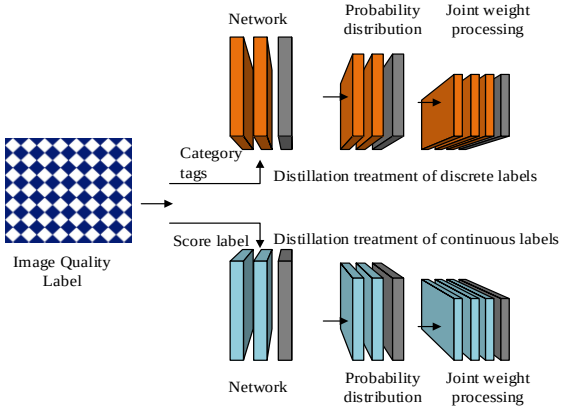


Figure 3: Image quality processing content

On discrete data, knowledge distillation is still used to process quality category labels, which can be expressed as equation (6).

$$L_L = \frac{1}{N} \sum_{i=1}^{N} \omega \cdot Loss(\varsigma(Z), C_L)_i + (1 - \omega) \cdot Loss(\varsigma(Z)))_i \quad (6)$$

In equation (6), the real image C_L and the logistic regression network Z are normalized by ς to obtain the loss value under its loss weight ω. T is the distillation temperature. N represents the number of samples.

2.2 Intelligent portrait composition based on human pose analysis

Intelligent composition systems for portrait photography often need to operate in real-time or near real-time conditions, especially in mobile devices or embedded systems. Lightweight design can significantly reduce computational complexity, memory footprint, and computational load, enabling systems to operate in resource-constrained environments. This improves the response speed of the system. In Section 2.1, the network with lightweight design can reduce the number of parameters and maintain high feature extraction ability. Therefore, multiple tasks can be calculated in parallel, which can provide computational support for multi-person recognition of human posture analysis and complete multi-task detection. Intelligent composition needs to understand the human body's pose, position, scale, and other information, which relies on the multi-task detection results in Section 2.1. The feature information output from section 2.1 is an important basis for structured portrait data analysis. The lightweight design of multi-task detection for portrait photography images provides an efficient and accurate preprocessing method for human pose analysis. The human pose analysis can guide the implementation of intelligent composition to achieve autonomous composition based on differences in human pose, thereby better meeting aesthetic requirements [15]. The research mainly analyzes the estimation algorithm for multi-person poses and proposes a single-stage differentiable multi-person pose estimation approach, which introduces additional features at adaptive human body local correlation points and uses relay nodes to associate the relationship between human body instances and key points. The features of different related points can be regressed to the key point positions within the corresponding regions, thus establishing position point connections in a unidirectional process [16]. Figure 4 shows the structure of the adaptive pose network.

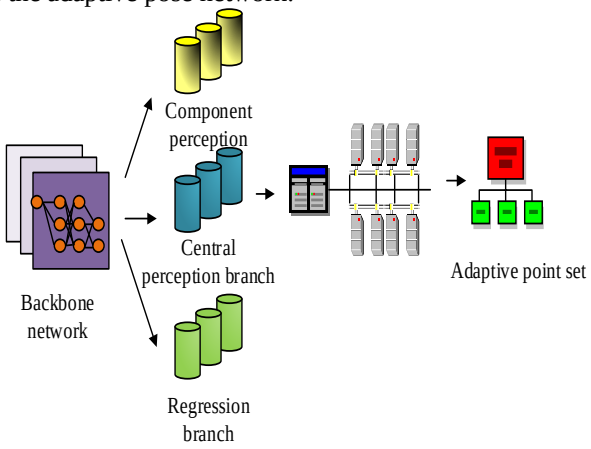


Figure 4: Adaptive pose network structure

In the human body representation section, the study utilizes a pixel by pixel keypoint regression framework, which represents human instances using a center point and points related to human body parts (adaptive points). The adaptive approach is used to represent local correlation points. This process can be explained, as displayed in equation (7).

$$EC \to \{P_h, P_s, P_a, P_r, P_l\} \tag{7}$$

In equation (7), P_h, P_s, P_a, P_r, and P_l respectively represent the head, shoulders, arms, buttocks, and legs. Unlike analyzing the offset vectors of reference points and keypoints, the study represents the human body as an adaptive point set, which can dynamically adapt to different human poses and scales through network learning. Figure 5 shows the overall network architecture.

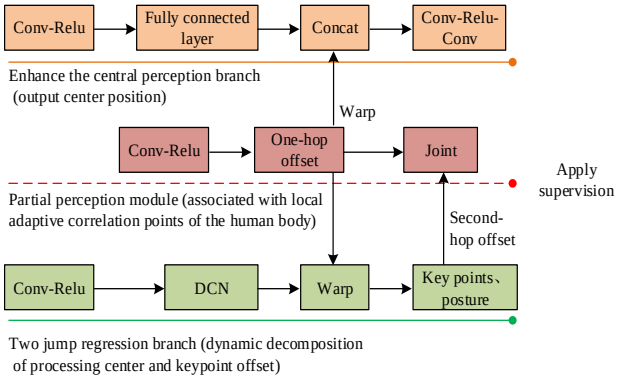


Figure 5: Schematic diagram of overall network architecture design

In Figure 5, the network architecture consists of three parts: component perception, center perception branch, and regression branch. The backbone network extracts visual features, and the component perception module associates local adaptive related points of the human body. The receptive field uses aggregation to predict the center of the human body at different scales and poses. In the keypoint regression section, the two-hop localization method is a phased strategy to locate key points, which aims to improve the accuracy and robustness of key point localization. Specifically, the two-hop method includes a first hop offset and a second hop offset. The first hop is the offset predicted from the center point of the human body to the adaptive part related points, which uses variable convolution operations to enhance feature extraction capabilities and locate adaptive part related points. The second step is to predict the offset from the adaptive part related points to specific keypoints, and then use variable convolution operations to capture local details and locate specific keypoints. Based on two-hop method, the key point localization problem is decomposed into two sub-tasks, reducing the difficulty of each task. Staged positioning can dynamically adjust the receptive field, adapt to the diversity of human postures and local deformations, and improve the robustness and accuracy of the model. The adaptive receptive field can predict the center localization of a single channel heatmap. The loss of the center perception branch is represented by a pixel by pixel function, as shown in equation (8) [17-18].

$$L_{ct} = \frac{1}{N}\sum_{i=1}^{N}\begin{cases}(1-\overline{P}_c)^{\alpha}\ln(\overline{P}_c) & ,P_c = 1\\ (1-P_c)^{\beta}\ln(1-P_c) & ,P_c \neq 1\end{cases} \tag{8}$$

In equation (8), $\overline{P}_c$ represents the predicted value. P_c represents the label value. α and β are hyperparameters. In the keypoint regression section, the two-hop method is used to implement regression branches, which involves performing variable convolution operations on features to determine their corresponding positions and locate keypoint positions. Two-hop includes the first offset $\overline{off}_1$ and the second offset $\overline{off}_2$, as shown in equation (9).

$$\overline{off} = \overline{off}_1 + \overline{off}_2 \tag{9}$$

The first hop is from the center of the human body to the relevant point of the adaptive part, and the second hop is from the relevant point of the adaptive part to a specific key point. The study uses bone perception regression loss to represent the bone links within human body regions to improve localization accuracy. Meanwhile, parallel scores are added to the backbone network during multi-task training to generate keypoint heatmap labels. The center perception branch and the two-hop regression branch output the center position and offset position, respectively. Finally, the candidate pose center positions are selected using a pooling kernel. The pose decoding results can be obtained by combining the predicted center and keypoint offset [19]. In response to the quality estimation problem in multi-person pose regression analysis, this study uses consistency instance scores and query encoding modules to represent human instance position scores and pose scores. The regression quality is measured by the similarity of key points between predicted pose and real pose, as shown in equation (10).

$$KPS_p = \frac{\sum_i \exp(\frac{-d_{p,i}^2}{2s_p^2 k_i^2})\lambda(v_{p,i} > 0)}{\sum_i \lambda(v_{p,i} > 0)} \tag{10}$$

In equation (10), $d_{p,i}$ is the distance between the predicted position of key point i and the label position. S_p is the scale corresponding to pose k. k_i is the annotation error of key points. By solving the confidence level in equation (10), the consistency instance score can be obtained. Query encoding includes keypoint queries and pose queries. When processing model tasks or features in parallel, multi-branch structures perform separate prediction and regression processing on keypoints of different types or parts. Bilinear interpolation can be used to extract and aggregate the semantic and positional information of each keypoint [20]. The transformed keypoint regression features can be expressed as equation (11).

$$R_k(c)' = \sum_{n=1}^{k} \{R_k(c)(c + \overline{D}_{c \to q} + \overline{D}_{q \to q'n})\} \quad (11)$$

In equation (11), q' represents the semantic point. q is the key point query. $R_k(c)$ represents the regression feature of the original keypoints at the prediction center c. $\overline{D}_{c \to q}$ is the offset from the center to the key point. n represents the number of semantic points, and $\overline{D}_{q \to q'n}$ is the offset of the semantic meaning of the key points queried. After the key point query position is processed by bilinear interpolation and connected through the channel dimension, new pose instance features can be obtained, as expressed in equation (12).

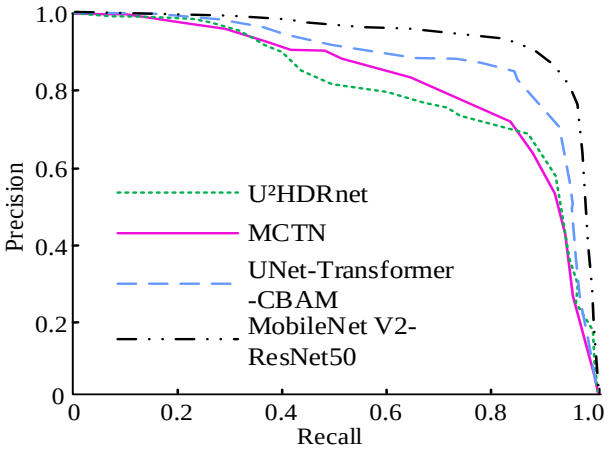
$$R_I(c)' = Concat(\{R_I(c)(c + \overline{D}^{i}_{c \to q})\}_{i=1}^{U}) \quad (12)$$

In equation (12), U represents the number of keypoint queries. $\overline{D}^{i}_{c \to q}$ is the offset from the center to the key point query. $R_I(c)$ represents the posture instance feature. In pose query encoding, it is possible to incorporate structural pose information into instance feature representations and perform consistency processing on their instance score labels. Through key point detection and positioning, posture analysis and adjustment, the position and posture of the portrait in the image can be adjusted. It can realize the human pose estimation, which is more in line with the aesthetic principles and user preferences and meets the needs of personalized settings. Hyperparameters are optimized using grid search, ensuring that model performance is balanced between computational efficiency and accuracy. The batch size is selected to optimize processing speed without compromising feature extraction.

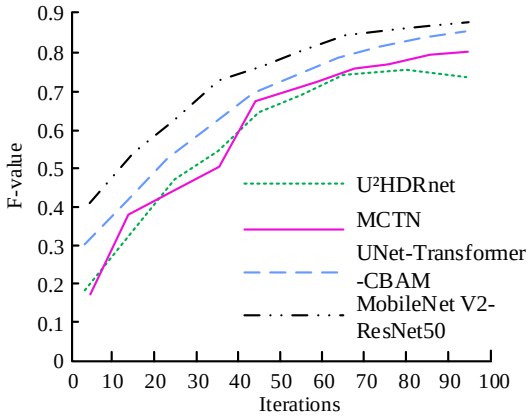
3 Results

3.1 Portrait image cropping effects

To analyze the effectiveness and applicability of the proposed intelligent composition method, this study examines it from three aspects: intelligent cropping for portrait detection, pose recognition analysis, and composition image quality. The MS COCO and MuPoTS-3D datasets are used for pose recognition and intelligent composition analysis of portraits. 30% of the images in the above dataset are randomly selected to form the photographic image composition test dataset. MS COCO is a widely used computer vision dataset with approximately 118,000 training images. It contains rich natural scene images (indoor, outdoor, urban, rural, etc.) and portrait images of individuals (single person, multiple people, different poses, and different lighting conditions). The annotation information includes object detection, instance segmentation, and keypoint detection. The MuPoTS-3D dataset is a dataset for multi-person 3D pose estimation, suitable for pose analysis and composition research in complex scenes. The image type includes scenes of multi person interaction and multiple role instances. The annotation information includes human keypoint annotations, character bounding boxes, and approximately 8,000 training images. The experimental environment is designed as follows. The central processing unit is Intel core i5-6500, equipped with NVIDIA RTX series graphics processors. The deep learning framework is PyTorch, the image processing library is OpenCV, the computer memory size is 12GB, and the optimizer is Adam. The learning rate of the network is set to 0.001, the batch size is 32, and the training epoch is 100. Firstly, the lightweight improved MobileNet V2-ResNet50 network proposed in the study is analyzed for its portrait cropping effect. It is compared with the improved deep learning network (UNet-Transformer-CBAM), lightweight detection network (Center-Net), multi-feature fusion image processing algorithm (U²HDRnet), and Multi-branch convolution-Transformer network (MCTN). The research utilizes Precision, Recall, F1-value, Average Intersection over Union ratio (Avg IoU), Mismatch, Area under Curve (AUC), Precision-Recall Curve (PCK), etc. to conduct experimental result analysis. Precision refers to the proportion of samples predicted by the model to be positive classes that are actually positive classes. The high precision indicates that there are fewer false classification samples with positive prediction. Recall refers to the proportion of samples that are actually positive and correctly predicted as positive by the model. The higher the recall, the more positive samples the model can find. The F1-value is the harmonic mean of precision and recall, used to balance precision and recall. The higher the F1-value, the better the balance between precision and recall achieved by the model. IoU is an indicator used to measure the overlap between predicted boxes and real boxes, commonly used in object detection tasks. The higher the average IoU value, the greater the overlap between the predicted box and the real box. AUC is the area under the ROC curve, which is used to measure the performance of a classification model. The ROC curve is plotted with False Positive Rate (FPR) as the horizontal axis and True Positive Rate (TPR) as the vertical axis. The closer the AUC value is to 1, the better the classification performance of the model. PCK is an accuracy recall curve used to measure the precision and recall of a model at different thresholds. The larger the area under the PCK curve, the better the performance of the model at different thresholds. The area under PCK can be used to evaluate the proportion of correctly estimated keypoints. Figure 6 shows the Precision-Recall (PR) results of extracting portrait image information.



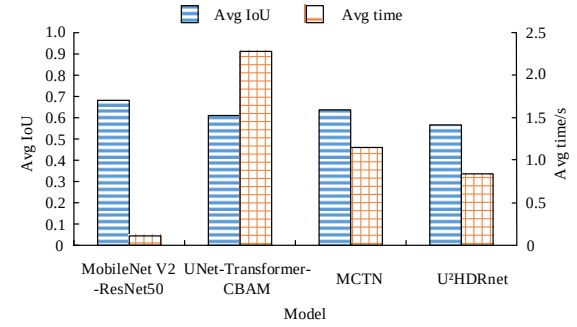
(a) PR value



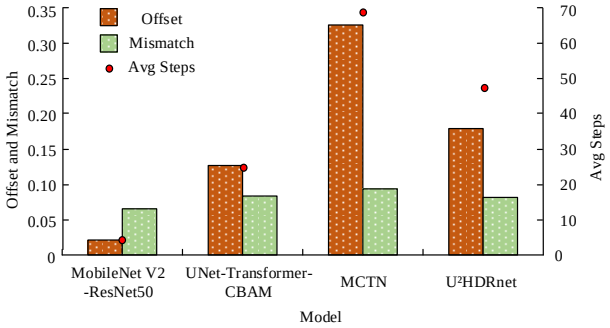
(b) F-value

Figure 6: Comparison of PR curves and F 1 value of different models

In Figure 6 (a), the PR curves of the MobileNet V2-ResNet50 model and the UNet-Transformer-CBAM model were closer to the bottom right corner. The precision of the MCTN and U²HDRnet algorithms was slightly lower, not exceeding 88% and 86.78%, respectively. However, the MobileNet V2-ResNet50 proposed in the study achieved a precision of 94.51% in extracting portrait image information, with good cropping accuracy. In Figure 6 (b), the MobileNet V2-ResNet50 model achieved good portrait image cropping, with small fluctuations in the curve nodes and a smooth direction. Its overall F1-value on the Open Images dataset exceeded 85%. Afterwards, the computational complexity of the above algorithm is analyzed using the Avg IoU, Average Steps (Avg Steps), Average Time/s (Avg time), offset, and Aspect Ratio Error (Mismatch). The results are shown in Figure 7.



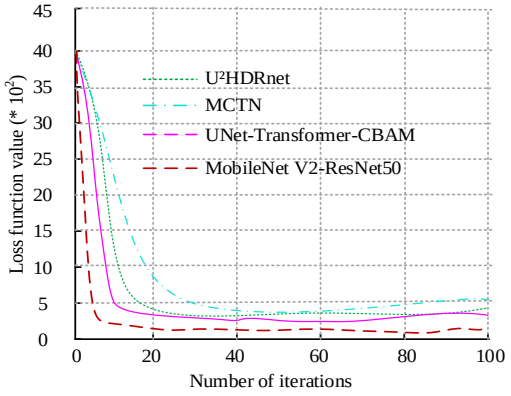
(a) Avg IoU and Avg time



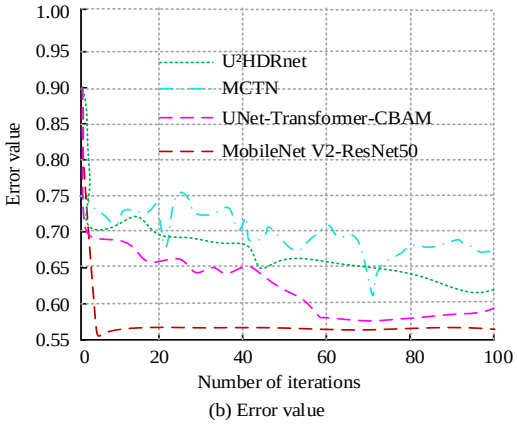
(b) Offset, Mismatch and Avg Steps

Figure 7: Computational complexity of different portrait detection algorithms

In Figure 7 (a), the Avg IoU exhibited by the MobileNet V2-ResNet50 model was 0.69, while the Avg IoU of the other three comparison methods was all less than 0.65. The Avg time of the MobileNet V2-ResNet50 model was significantly lower than other algorithms. Offset refers to the distance between the optimal cropping window and the predicted cropping window. A large value indicates a larger deviation in image processing. In Figure 7 (b), the offsets of MobileNet V2-ResNet50, UNet-Transformer-CBAM, MCTN, and U²HDRnet were 0.021, 0.126, 0.324, and 0.178, respectively, and the Mismatch was 0.065, 0.083, 0.093, and 0.081, respectively. The offset and Mismatch of MCTN were 0.324 and 0.093, respectively, making the calculation more complex and requiring more time and steps. The above results indicate that the proposed portrait detection model has good image cropping performance. Further analysis is conducted on the loss function and extraction error results during the portrait image cropping process, as shown in Figure 8.



(a) Magnitude of the loss



(b) Error value

Figure 8: Facial image feature extraction results of different algorithms

In Figure 8 (a), when the number of iterations was less than 20, the decreasing slopes of the loss curve from large to small were MobileNet V2-ResNet50>KUNet-Transformer-CBAM>U²HDRnet>LBP>MCTN. The loss curve of MobileNet V2-ResNet50 model converged quickly, with an average loss value of less than 5. The average loss values of other algorithms were between 5 and 10. In Figure 8 (b), the classification error curves of the four algorithms showed a decreasing trend with the increase of iterations, with the proposed algorithm having a smaller error value and the MCTN algorithm having the highest loss value. To further analyze the performance of the above models in image processing, ablation experiments are conducted to verify and test the processing speed of different models. The results are shown in Table 2.

Table 2: Ablation results and processing speed of different models

Model	Ablation results			Processing speed		
	Recall/%	Precision/%	F1-value/%	FLOPs/B	Inference time/ms	Parameter measurement/M
UNet-Transformer-CBAM	88.36	87.25	89.99	15	50	25
Center-Net	87.25	87.53	86.98	3	20	8
U²HDRnet	86.14	85.11	86.98	10	35	15
MCTN	86.08	85.93	87.12	18	60	30
MobileNet V2-ResNet50	88.74	89.95	90.14	2.5	5	15

In Table 2, the MobileNet V2-ResNet50 model exhibited recall, precision, and F1-value of 88.74, 89.95, and 90.14, respectively, in the ablation results, which were much higher than other models. The UNet-Transformer-CBAM model, which performed well, had an F1-value of 89.99%. The Floating-Point Operations (FLOPs) of the MobileNet V2-ResNet50 model and the Center-Net model were less than 5B, and the parameter metric of the Center-Net model was the smallest (8M), but its inference speed (20ms) was greater than that of the MobileNet V2-ResNet50 model (5ms). Overall, MobileNet V2-ResNet50 balanced precision, recall, and computational efficiency, making it suitable for real-time portrait photography composition tasks. UNet-Transformer-CBAM and MCTN performe well in complex scenarios, but have high computational costs and are suitable for offline high-precision tasks. Center-Net has high computational efficiency, but slightly lower accuracy in composition tasks, making it suitable for lightweight real-time detection tasks. U²HDRnet performs well in improving image quality, but its performance in composition tasks is not as good as MobileNet V2-ResNet50, making it suitable for post-processing tasks.

3.2 Intelligent portrait composition effect

With the development of mobile devices, photography has gradually become an important tool for people to record their lives. However, the different ideas and intentions of the photographer result in differences in the visual effects presented in the image. The relationships between elements in the images and the overall style of the character all reflect the ingenuity of image construction [21-22]. The application of composition rules and forms can present different aesthetics and artistic qualities in portrait photography. For example, a central composition will attract the viewer's attention to the center of the picture, thereby highlighting the subjectivity of the characters. The three-point composition arranges character themes horizontally or vertically, presenting a more balanced aesthetic style and a more narrative visual effect [23-24]. The portrait composition quality of above models is compared, and the results are shown in Figure 9.

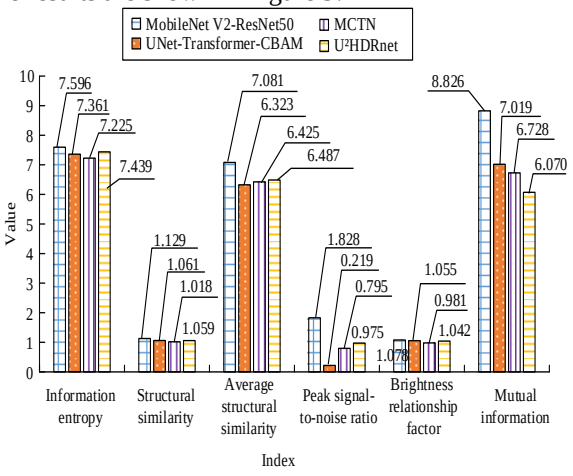


Figure 9: Image quality evaluation effects under different algorithms

In Figure 9, the information entropy, structural similarity, average structural similarity, peak signal-to-noise ratio, brightness relationship factor, and mutual information of the proposed fusion algorithm were 7.596, 1.129, 7.081, 1.828, 1.078, and 8.826, respectively. The overall image quality fusion effect was significantly better than other algorithms. The next was UNet-Transformer-CBAM and U²HDRnet, with information entropy values exceeding 7.35. Afterwards, the proposed Adaptive-Human pose estimation (AHPE) algorithm was tested on the MS COCO and MuPoTS-3D datasets with a convolution size of 3×3. It is compared with Random Sample Consensus-Mean-Shift (RSC-MS) algorithm, Open Pose-Generative adversarial network (OP-GAN), and Progressive Gaussian Filtering-Human pose estimation (PGF-HPE). The recall, precision and F1-value are shown in Table 3.

Table 3: Comparison results of portrait image detection

Dataset	Algorithm	Evaluation		
		Recall/%	Precision/%	F1-value/%
MS COCO	AHPE	88.55	89.94	90.05
	RSC-MS	87.05	87.17	88.41
	OP-GAN	88.07	88.25	87.11
	PGF-HPE	86.93	87.31	88.16
MuPoTS-3D	AHPE	89.78	91.36	90.20
	RSC-MS	86.87	87.53	86.98
	OP-GAN	86.25	87.73	87.12
	PGF-HPE	88.43	88.94	89.56

In Table 3, on the MS COCO dataset, AHPE achieved a precision of 88.94% for human pose detection, while the detection precision of RSC-MS, OP-GAN, and PGF-HPE was 87.17%, 88.25%, and 87.31%, slightly worse than the algorithm proposed in the study, and their corresponding F1-values were basically no more than 90%. On the MuPoTS-3D dataset, the performance of human pose detection from high to low was as follows: AHPE algorithm>PGF-HPE algorithm>OP-GAN algorithm>RSC-MS algorithm. The precision, recall, and F1-value of the proposed adaptive human pose recognition algorithm reached 91.36%, 89.78%, and 90.20%, respectively. Afterwards, further analysis is conducted on the key point prediction results of human nodes using different algorithms, and classification evaluation is performed using AUC and PCK metrics. The higher the value, the higher the prediction accuracy of the model. The predicted key points of human body nodes are shown in Table 4.

Table 4: Prediction scores of key points in human body nodes

Dataset	Algorithm	Evaluation	
		AUC	PCK
MS COCO	AHPE	44.99	83.24
	RSC-MS	35.94	76.19
	OP-GAN	42.26	82.51
	PGF-HPE	43.22	82.87
MuPoTS-3D	AHPE	47.87	87.12
	RSC-MS	43.86	84.11
	OP-GAN	46.09	86.34
	PGF-HPE	45.52	83.77

In Table 4, the AUC and PCK of AHPE algorithm on the MS COCO dataset were 44.99 and 83.24, respectively, which were much higher than those of RSC-MS algorithm (35.94 and 76.19), OP-GAN algorithm (42.26 and 82.51), PGF-HPE algorithm (43.22 and 82.87), and the maximum difference in AUC and PCK scores exceeded 5. On the MuPoTS-3D dataset, AHPE algorithm and OP-GAN algorithm could better capture key nodes of the human body, and the score deviation was relatively small, with RSC-MS algorithm performing the worst. Specifically, the AUC and PCK of AHPE were 47.87 and 87.12, respectively, while the AUC and PCK of RSC-MS algorithm were 43.86 and 84.11, respectively. The reason for this result is that the AHPE algorithm has good adaptive performance and can effectively capture pose key nodes. However, the RSC-MS algorithm ignores the relationship between the central pose and key points during pose analysis, which makes it more prone to erroneous results. The MobileNet V2-ResNet50 model is combined with the AHPE model to obtain a hybrid intelligent composition method. Two types of portrait images are selected for intelligent composition analysis, and the results are shown in Figure 10.

(a) Original image (b) Optimized composition of the image

Figure 10: Intelligent portrait composition presentation effects

In Figure 10, when there were characters and scenery present, this method utilized intelligent cropping to better consider the relationship between characters and background, reducing unnecessary scene whitespace. When it was a multi-person image, this method highlighted the story expression of the characters and had a better image presentation effect than the original image.

4 Discussion and conclusion

A portrait photography intelligent composition system based on multi-task learning and human pose analysis was designed, which could automatically optimize portrait composition. The results showed that the PR curves of MobileNet V2-ResNet50 model and UNet-Transformer-CBAM model were closer to the bottom right corner, while the precision of MCTN and U²HDRnet algorithms was slightly lower, not exceeding 88% and 86.78%, respectively. When the training data is insufficient or diverse, MCTN model may be difficult to fully learn effective feature representation, introducing high computational complexity and over-fitting risk. The multi-branch structure may lead to feature redundancy, reducing the model's ability to focus on key features. The design goals of U²HDRnet are not completely consistent with the accuracy requirements of composition tasks, resulting in poor performance in composition tasks. The proposed method achieves an accuracy of 94.51% in portrait detection, outperforming existing approaches. The multi-task learning framework enhances image quality while adapting to various human poses. The Avg IoU exhibited by the MobileNet V2-ResNet50 model was 0.69, while the Avg IoU of the other three comparison methods was all less than 0.65. The Avg Steps and Avg time of the MobileNet V2-ResNet50 model were significantly lower than other algorithms. The offsets of MobileNet V2-ResNet50, UNet-Transformer-CBAM, MCTN, and U²HDRnet were 0.021, 0.126, 0.324, and 0.178, respectively, and the Mismatch was 0.065, 0.083, 0.093, and 0.081, respectively. The loss curve of MobileNet V2-ResNet50 model converged quickly, with an average loss value of less than 5, while the average loss values of other algorithms were between 5 and 10. The improved lightweight MobileNet V2-ResNet50 network significantly outperformed the UNet-Transformer-CBAM and MCTN models in terms of computational efficiency, with FLOPs of only 2.5B, parameter count of 5M, and inference time of 15ms, making it suitable for real-time applications. The UNet-Transformer-CBAM model and MCTN model had high computational costs and were difficult to deploy on resource constrained devices. Compared with the lightweight network proposed by Li et al. [5], the research method further optimized computational efficiency while preserving fine-grained details, making it suitable for portrait detail processing in complex backgrounds. Human pose estimation is an important part of pose analysis, which can abstract important information representing human pose from input images. The adaptive human pose recognition algorithm AHPE proposed in the study achieved a precision of 88.94% for human pose detection on the dataset, while the detection accuracy of the RSC-MS, OP-GAN, and PGF-HPE comparison models did not exceed 88.5%. The prediction results of key points in human body nodes showed that the AHPE algorithm had AUC and PCK scores of 44.99 and 83.24 on the MS COCO dataset, which were much higher than the RSC-MS algorithm's 35.94 and 76.19, the OP-GAN algorithm's 42.26 and 82.51, and the PGF-HPE algorithm's 43.22 and 82.87, respectively. The maximum difference in AUC and PCK scores exceeded 5 points. The MobileNet V2-ResNet50 model proposed in the study was combined with the AHPE model to obtain a hybrid intelligent composition method. Compared with the photography composition assistance framework proposed by Farhat et al. [4], the research method was more efficient in object detection and composition analysis, and did not require complex multitasking frameworks. Compared with the field of view evaluation model proposed by Zhong et al. [6], the research method did not rely on image boundaries as cropping boundaries, which better enhanced the aesthetic appeal of images. Two types of portrait images were selected for intelligent composition analysis. The results showed that the whitespace in the scene content was reduced, highlighting the relationship between characters and the background or the sense of storytelling of objects.

The research method combines the lightweight of MobileNet V2 and the high-precision characteristics of ResNet50, which can simultaneously adapt to multi-task requirements such as object detection, image classification, and composition analysis. This research method has good intelligent composition effect and can effectively improve the intelligent cropping effect of portrait photography. Compared with the attribute assisted multi-modal memory network proposed by Li et al. [8], the research method performs more efficiently in image aesthetic evaluation and does not require complex multi-modal information fusion. It solves the detail loss problem of lightweight networks proposed by Li et al. [5] in complex backgrounds. This research method simplifies the multi-task processing framework by improving the composition rules and visual field evaluation, and solves the problem that the existing cutting methods proposed by Zhong et al. [6] are difficult to improve the aesthetic feeling. However, the performance of the research method in cluttered backgrounds and their dependence on hardware resources still need further improvement. Stronger attention mechanism can be introduced to enhance the target detection and synthesis ability of the model in the clutter background. Combined with multi-scale feature fusion technology, the ability of the model to deal with complex texture and dense targets is enhanced. While the proposed approach improves composition accuracy, its computational cost remains a consideration for real-time applications. Future optimizations could involve model pruning or quantization to enhance efficiency for mobile and embedded devices. Meanwhile, considering that the portrait in the actual photography process is also affected by factors such as lighting, occlusion, and complex

background, there may be a certain degree of error in pose recognition. Therefore, in the future, it is possible to strengthen the behavior recognition scheme and pose estimation method for skeleton sequences, while considering user interaction and personalized settings, and attempt to provide personalized intelligent composition functions by introducing user feedback mechanisms.

References

[1] Jiang N, Sheng B, Li P, Lee T Y. Photohelper: portrait photographing guidance via deep feature retrieval and fusion. IEEE Transactions on Multimedia, 2022, 25: 2226-2238. DOI: 10.1109/TMM.2022.3144890.

[2] Bao W. The application of intelligent algorithms in the animation design of 3D graphics engines. International Journal of Gaming and Computer-Mediated Simulations (IJGCMS), 2021, 13(2): 1-12. DOI: 10.4018/IJGCMS.2021040103.

[3] Ather D, Madan S, Nayak M, Tripathi R, Kant R, Kshatri S S, Jain R. Selection of smart manure composition for smart farming using artificial intelligence technique. Journal of Food Quality, 2022, 1: 4351825. DOI: 10.1155/2022/4351825.

[4] Farhat F, Kamani M M, Wang J Z. CAPTAIN: Comprehensive composition assistance for photo taking. ACM Transactions on Multimedia Computing, Communications, and Applications, 2022, 18(1): 1-24. DOI: 10.1145/3462762.

[5] Li R, Zhang D, Geng SL, Zhou MQ. Matting Algorithm with Improved Portrait Details for Images with Complex Backgrounds. Applied Sciences. 2024,27;14(5):1942. DOI: 10.3390/app14051942.

[6] Zhong L, Li F H, Huang H Z, Zhang Y, Lu S P, Wang J. Aesthetic-guided outward image cropping. ACM Transactions on Graphics (TOG), 2021, 40(6): 1-13. DOI: 10.1145/3478513.3480566.

[7] Wang B, Si H, Fu H, Gao R, Zhan M, Jiang H, Wang A. Content-Aware Image Resizing Technology Based on Composition Detection and Composition Rules. Electronics, 2023, 12(14): 3096. DOI: 10.3390/electronics12143096.

[8] Li L, Zhu T, Chen P, Yang Y, Li Y, Lin W. Image aesthetics assessment with attribute-assisted multimodal memory network. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(12): 7413-7424. DOI: 10.1109/TCSVT.2023.3272984.

[9] Baraheem S S, Le T N, Nguyen T V. Image synthesis: a review of methods, datasets, evaluation metrics, and future outlook. Artificial Intelligence Review, 2023, 56(10): 10813-10865. DOI: 10.1007/s10462-023-10434-2.

[10] Wang Q, Ma Y, Zhao K, Tian Y. A comprehensive survey of loss functions in machine learning. Annals of Data Science, 2022, 9(2), 187-212. DOI: 10.1007/s40745-020-00253-5.

[11] Zhang K, Ren W, Luo W, Lai W S, Stenger B, Yang M H, Li H. Deep image deblurring: A survey. International Journal of Computer Vision, 2022, 130(9), 2103-2130. DOI: 10.1007/s11263-022-01633-5.

[12] Ghasemi-Naraghi Z, Nickabadi A, Safabakhsh R. LogSE: An Uncertainty-Based Multi-Task Loss Function for Learning Two Regression Tasks. Journal of Universal Computer Science, 2022, 28(2):141-159. DOI: 10.3897/jucs.70549.

[13] Zhang H, Peng Y. Image clustering: An unsupervised approach to categorize visual data in social science research. Sociological Methods & Research, 2024, 53(3): 1534-1587. DOI: 10.1177/00491241221082603.

[14] Chikarkova M. Artificial intelligence and digital art: current state and development prospects. Skhid. 2023,4(3):9-13. DOI: 10.21847/2411-3093.2023.4(3).294658.

[15] Chen R, Ghavidel Aghdam MR, Khishe M. Utilization of Artificial Intelligence for the automated recognition of fine arts. PloS one. 2024,19(11):0312739. DOI: 10.1371/journal.pone.0312739.

[16] Gross EC. Artificial Intelligence Generated Art Imitation and the Art World: Implications and Further Questions. Boletín de Arte. 2023,28(44):313-316. DOI: 10.24310/ba.44.2023.15972.

[17] Xu J, Liu W, Xing W, and Wei X. MSPENet: multi-scale adaptive fusion and position enhancement network for human pose estimation. The Visual Computer, 2022, 39(5):2005-2019. DOI: 10.1007/s00371-022-02460-y.

[18] Kumar P, Chauhan S, Awasthi L. Human pose estimation using deep learning: review, methodologies, progress and future research directions. International Journal of Multimedia Information Retrieval, 2022, 11(4):489-521. DOI: 10.1007/s13735-022-00261-6.

[19] Huang L, Liu G. Functional motion detection based on artificial intelligence. The Journal of Supercomputing, 2022, 78(3): 4290-4329. DOI: 10.1007/s11227-021-04037-3.

[20] Wang X, Wu Y, Zhu L, and Yang Y. Symbiotic Attention with Privileged Information for Egocentric Action Recognition. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7):12249-12256. DOI: 10.1609/aaai.v34i07.6907.

[21] Zhou K, Wu T, Wang C, Wang J, and Li C. Skeleton Based Abnormal Behavior Recognition Using Spatio-Temporal Convolution and Attention-Based LSTM. Procedia Computer Science, 2020, 174(8):424-432. DOI: 10.1016/j.procs.2020.06.110.

[22] Zhang H, Luo Y, Zhang L, Wu Y, Wang M, Shen Z. Considering three elements of aesthetics: Multi-task self-supervised feature learning for image style classification. Neurocomputing, 2023, 520(1): 262-273. DOI: 10.1016/j.neucom.2022.10.076.

[23] Yang G Y, Zhou W Y, Cai Y, Zhang S H, Zhang F L. Focusing on your subject: Deep subject-aware

image composition recommendation networks. Computational Visual Media, 2023, 9(1): 87-107. DOI: 10.1007/s41095-021-0263-3.

[24] Zhang M, Li M, Yu J, Chen L. Aesthetic photo collage with deep reinforcement learning. IEEE Transactions on Multimedia, 2022, 25(1): 4653-4664. DOI: 10.1109/TMM.2022.3180217.

A Hybrid Investment Risk Prediction Framework Integrating ADASYN-RF, CS-SVM, PCA-BP, and ARIMA Models

Liucheng Zhang
Henan Quality Institute, Pingdingshan 467000, China
E-mail: zhc582@163.com

Keywords: Adaptive synthetic sampling approach, random forest, investment, risk prediction

Received: August 4, 2025

Investment risk is often the result of a long-term interplay among multiple factors, such as market volatility and business performance. Predicting investment risk in advance helps investors avoid losses and protect their assets. However, current prediction methods rely heavily on historical data, leading to poor timeliness and unreliable results. Therefore, this study proposes an investment risk prediction model based on the adaptive comprehensive sampling algorithm and the random forest algorithm to predict investment risks accurately. This model makes full use of the adaptive comprehensive sampling algorithm to balance the categories of risk data, and the random forest captures the risk characteristics to achieve investment risk prediction. CS-SVM is introduced to improve the prediction model and prevent overfitting. Meanwhile, principal component analysis and backpropagation networks are combined to solve the problems of unstructured data processing and time series dependency modeling. In the experiment, the study took historical investment data as the dataset and compared it with the investment risk prediction models constructed by three algorithms: CatBoost-GJO, TS-GA, and GAN-Stacking. Evaluate the prediction accuracy of each method in investment risk prediction, including the Sharpe ratio, value at risk, volatility, win rate and profit and loss. The results show that the prediction accuracy of this model reaches 99.0%, the Sharpe ratio is 1.9, and the maximum drawdown is 5.9%, all superior to the comparison models. Moreover, in the actual investment risk prediction, its volatility and value at risk are only 8.2% and 7.3% respectively, while the winning rate and profit and loss ratio can reach 88.9%, 6:1. These results indicate that the proposed model achieves high accuracy in investment risk prediction, effectively addressing the limitations of existing methods and providing a new approach to improve prediction performance. It contributes to the development of intelligent and efficient investment risk forecasting.

Povzetek: Študija predlaga model za napovedovanje investicijskega tveganja, ki z adaptivnim uravnoteženjem vzorcev, naključnim gozdom in CS-SVM ter s kombinacijo PCA in povratno-propagacijskih mrež za obdelavo ne-strukturiranih in časovnih podatkov izboljša pravočasnost in zanesljivost napovedi.

1 Introduction

In the context of deep integration and high-frequency fluctuations in global financial markets, asset price linkage has become significantly stronger. The risk exposure of a single investment can easily spread through financial chains and lead to systemic financial risks [1]. From the perspective of financial dynamics, the accumulation of investment risk results from the combination of multiple factors, including macroeconomic cycles and industry competition. Accurately identifying risk signals before a risk outbreak not only helps investors avoid losses but also prevents a chain reaction of market risks [2]. Therefore, building a scientific and efficient investment risk prediction mechanism is essential for maintaining financial stability and ensuring steady asset growth. Current prediction methods mainly include historical data analysis, econometric models, and artificial intelligence algorithms. However, these approaches often rely too much on historical data and fail to adapt to complex and changing market environments. As a result, there is a growing need for a flexible and accurate investment prediction method [3]. Investment risk prediction primarily focuses on balancing data distribution and modeling nonlinear relationships. The Adaptive Synthetic Sampling Approach (ADASYN) effectively addresses class imbalance in investment data, while the Random Forest (RF) algorithm excels at identifying complex nonlinear patterns in data [4, 5]. Therefore, this study develops a risk prediction model based on ADASYN and RF. The model aims to improve the accuracy of investment risk forecasting in rapidly changing financial markets, assisting investors in managing risks and maintaining asset stability and market order. By combining ADASYN and RF, the model provides a novel approach to address the limitations of traditional investment risk prediction methods. It provides a new analytical framework that integrates ADASYN and RF, filling the research gap in accurate risk forecasting under market volatility. This study aims to address the

insufficient accuracy and poor adaptability of current investment risk prediction methods and achieve high-precision risk prediction. The study uses the ADASYN algorithm to address class imbalance in investment data, thereby enhancing the model's ability to detect rare risk events. Combined with the powerful nonlinear modeling and feature selection capabilities of the RF algorithm, the generalization performance of the prediction model is further strengthened. The key parameters of RF are optimized using CS and Support Vector Machine (SVM) optimization algorithms to enhance the model's adaptability and prediction accuracy across different market environments. Finally, the model parameters are optimized by combining principal component analysis and backpropagation algorithm to solve the problems of unstructured data processing and time series dependency modeling.

2 Related works

ADASYN effectively addresses class imbalance in data, while the RF algorithm captures complex nonlinear relationships. Both methods have been widely applied to solve complex problems in various fields. Scholars from around the world have explored these approaches in depth [6]. To overcome the limitations of traditional methods in handling imbalanced data, Song et al. proposed a credit risk prediction model. This model used an improved ADASYN for sampling and trained the data with a lightweight gradient boosting machine. The results showed that the model outperformed other methods on loan datasets [7]. In response to security threats in the medical Internet of Things, Salehpour et al. developed a hybrid intrusion detection framework. The framework applied the Extreme Gradient Boosting algorithm to detect anomalies, used ADASYN for resampling, and adopted RF for intrusion detection. Experiments demonstrated that the framework achieved an accuracy of 92.23% in threat identification [8]. To evaluate pile performance in cone penetration tests, Dawei et al. proposed two models—one combining Particle Swarm Optimization with RF, and another integrating Harris Hawks Optimization with RF. A comparison of the two models showed that the one using Harris Hawks Optimization with RF delivered better performance [9]. Addressing the challenge of weighting influencing factors in fine particulate matter prediction, Ding et al. introduced a model combining weighted RF and a Long Short-Term Memory network. The RF algorithm was used to filter key factors, and the weighted data were input into the Long Short-Term Memory network for training. The results showed that the proposed model had stronger prediction capabilities [10].

Over time, the theoretical and practical applications of risk prediction methods have matured, and researchers in many countries have conducted in-depth studies [11]. Belhadi et al. proposed a hybrid ensemble learning method based on Rotation Forest to address credit risk in small and medium-sized enterprises. By analyzing data from 216 agricultural SMEs in Africa, the study confirmed that the model could provide reliable risk forecasts for small businesses [12]. Li et al. proposed a financial risk prediction model that optimizes the backpropagation neural network for the problem of enterprise financial risks. Analysis based on the financial data of listed companies from 2017 to 2020 as samples indicates that the model's accuracy rate in predicting the financial distress of normal companies exceeds 80% [13]. Doğru et al. addressed the inefficiency and low accuracy of diabetes risk prediction by proposing a new model based on super ensemble learning. The model used logistic regression and other methods as base learners and Support Vector Machine (SVM) as the meta-learner, achieving a prediction accuracy of 92 to 99.6% [14]. In response to limited data and privacy concerns in supply chain risk prediction, Zheng et al. put forward a method based on federated learning. Using a buyer-based prediction of supplier delivery delays as an example, the study explored the effects of data imbalance. Experimental results showed that this approach improved risk prediction capability [15].

In summary, although existing studies have advanced risk prediction, limitations remain in comprehensiveness and intelligence. In response to this, the study proposes a risk prediction model integrating ADASYN-RF, CS-SVM, PCA-BP, and ARIMA. The research expects that the proposed method can improve the efficiency and accuracy of risk prediction in the volatile financial market, so as to meet investors' demands for investment risk prediction. The comparison and summary of the relevant work are shown in Table 1.

Table 1: A comparison and summary of related work

Number	Usage method	Accuracy (%)	Dataset	Restriction
[7]	Improve the ADASYN+ LGBM	-	Credit data of a certain commercial bank from 2015 to 2020	The influence of feature interaction was not taken into account
[8]	ExBoost +ADASYN	92.23	IoMT device network traffic log	The noise sample processing is insufficient
[9]	The Harris Eagle optimization algorithm combined with RF+ADASYN	-	On-site record of CPT static cone penetration test	It is prone to fall into local optimum
[10]	RF+ Long Short-	-	Data of PM2.5 monitoring	Insufficient consideration of spatial

	Term Memory network		stations in the Beijing-Tianjin-Hebei region from 2018 to 2022	heterogeneity and unsatisfactory optimization of the attention mechanism
[12]	Ensemble learning + Rotating Forest algorithms	-	Transaction records of agricultural supply chain finance in Southeast Asia from 2018 to 2023	Sample selection bias and the influence of weather dynamics lead to a decline in prediction accuracy
[13]	Generative Adversarial Network	92.0-99.6	Risk event data of a certain financial regulatory platform	Sample selection bias and the influence of weather dynamics lead to a decline in prediction accuracy
[14]	Logistic regression +SVM	-	Diabetes screening data in Turkey from 2021 to 2021	Insufficient fusion of multimodal data leads to high computational complexity
[15]	Federated learning	-	Supply chain platform data	The communication efficiency is low and the problem of independent and equally distributed data has not been solved

As shown in Table 1, the current technology is unable to effectively handle time-non-time mixed investment data. This is because traditional methods often only model a single type of data, making it difficult to simultaneously capture the complex correlations between the dynamic evolution characteristics of time series and the static attributes of non-time data. The AD-PB proposed in the research can solve this limitation. The AD-PB model integrates time series analysis and non-time feature processing, effectively modeling dynamic patterns and static attributes in investment risks. By using a variety of technical means such as the hybrid adaptive comprehensive sampling algorithm, random forest algorithm, principal component analysis, backpropagation network, and support vector machine, the model can adaptively capture key risk factors and simultaneously alleviate the information interference problem caused by data mixing, thereby significantly improving the prediction performance and applicability.

3 Hybrid model design for investment risk prediction

3.1 Risk prediction algorithm combining Adasyn-Rf and improved SVM

The ADASYN-RF algorithm integrates ADASYN and RF to address challenges in risk prediction effectively. ADASYN-RF first uses ADASYN to generate synthetic samples for the minority class, balancing the data distribution and improving the representation of high-risk samples. Then, RF processes the balanced data, reduces noise influence, and handles high-dimensional features, enhancing the accuracy and robustness of risk prediction [16]. The workflow of ADASYN-RF is shown in Figure 1.

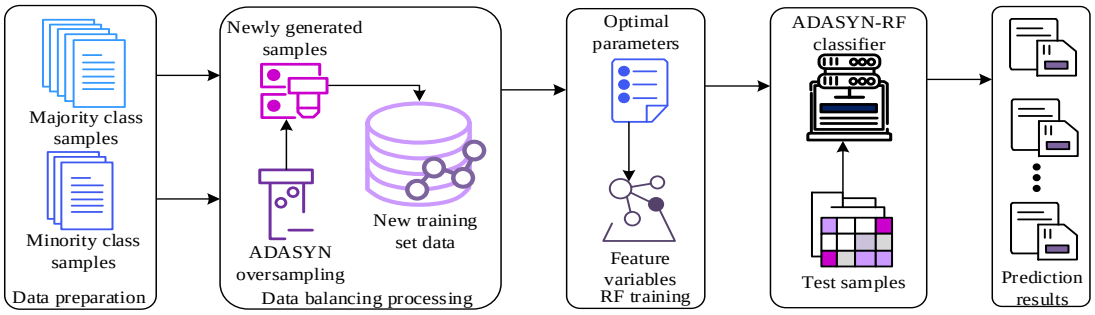


Figure 1: Workflow of the ADASYN-RF algorithm

As shown in Figure 1, the hybrid algorithm first collects the training dataset and divides it into minority and majority classes. ADASYN is then applied to oversample the minority class, and the generated samples are combined with the majority class to form a new training set, completing the data balancing process. Next, optimal parameters are selected, and feature variables are determined based on their importance to train the RF model. After training, the testing samples are input into the trained ADASYN-RF classifier. Finally, the classifier analyzes the testing data and outputs the prediction results, achieving the goal of risk prediction. The summary of the abbreviations and their full names that appear in the article is shown in Table 2.

Table 2: List of abbreviations and their full forms

Abbreviation	Full name
CS	Chaotic search
ADASYN	Adaptive synthetic sampling
BP	Back propagation
RF	Random forest
ARIMA	Autoregressive Integrated Moving Average model
PCA	Principal Components Analysis
SVM	Support vector machine
ADRS	ADASYN-RF-CS-SVM
AD-PB	ADRS-PCA-BP

The oversampling calculation in ADASYN is defined in Equation (1).

$$\beta_i = \frac{1 - D(x_i)}{\sum_{j-1}^{n_{\min}} (1 - D(x_j))} \tag{1}$$

In Equation (1), β_i represents the generation weight of the i -th minority sample, determining how many synthetic samples to generate. x_i denotes the i -th minority class sample. $D(x_i)$ indicates the distribution density of sample x_i , and $n_{\min}$ is the total number of minority class samples. The generation of new samples in ADASYN is further calculated as shown in Equation (2).

$$x_{new} = x_i + \alpha \times (x_{nn} - x_i) \tag{2}$$

In Equation (2), x_{nn} denotes the nearest neighbor of x_i among minority samples. α is the interpolation coefficient ranging within $[0,1]$, which controls the position of the new sample on the line between x_i and x_{nn} . x_{new} is the newly generated synthetic sample, belonging to the minority class and used for balancing the data distribution. The ADASYN-RF algorithm combines the advantages of both ADASYN and RF. ADASYN addresses data imbalance, while RF improves generalization. Together, they reduce false prediction rates in risk forecasting. However, ADASYN-RF has limitations in handling high-dimensional data efficiently. Its parameter tuning is complex, and it may suffer from overfitting. The CS-SVM algorithm, which combines CS and SVM, offers strong capabilities in high-dimensional space, uses simpler parameters, and mitigates overfitting through the hard or soft margin mechanism [17]. The CS algorithm conducts global search by introducing chaotic variables and can quickly converge to the optimal solution region in a complex parameter space. SVM, on the other hand, utilizes the principle of minimizing structural risk to construct the optimal classification hyperplane in a high-dimensional space. Combining the global optimization ability of CS and the good adaptability of SVM to high-dimensional data, CS-SVM shows higher stability and accuracy when solving the classification problem of imbalanced data. The inner core selects the RBF function, and the gamma value is optimized through the CS algorithm to avoid local optima. The value of C is set to a fixed value, which is 1. Therefore, this study integrates CS-SVM into ADASYN-RF for further improvement. The workflow of the CS-SVM algorithm is shown in Figure 2.

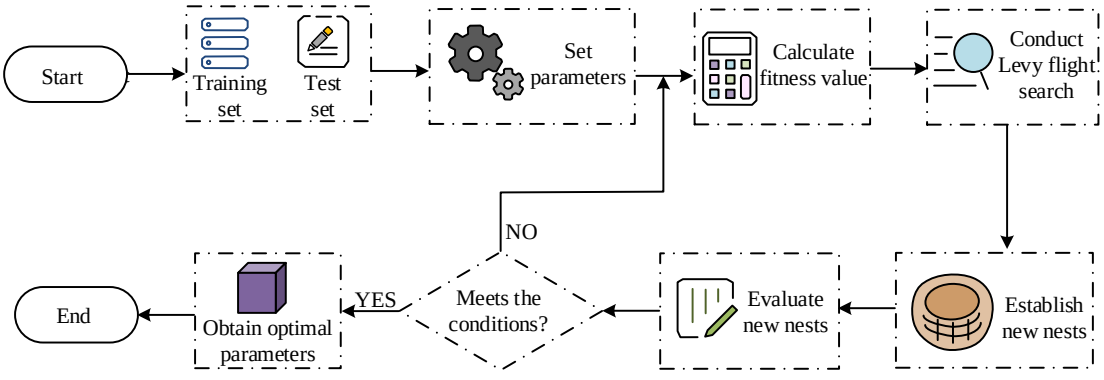


Figure 2: Workflow of the CS-SVM algorithm

As shown in Figure 2, the algorithm first divides the data into training and testing sets and normalizes them to unify the scale, improving processing efficiency and reducing the complexity of high-dimensional data. Then, parameters for SVM and CS are set. Bird nest positions are initialized randomly, and their fitness values are calculated. The CS algorithm updates positions using Levy flights to enhance the handling of high-dimensional data. To address class imbalance, poor-performing nests are discarded, and new nests are generated and evaluated

in each iteration. This process helps reduce false prediction rates. In multi-class scenarios, continuous parameter optimization enables the SVM model to better distinguish complex classes. Finally, the model is finalized based on the optimal parameters. The calculation is shown in Equation (3).

$$X_i^{t+1} = X_i^t + \alpha \otimes Levy(\lambda) \tag{3}$$

In Equation (3), X_i^t represents the position of the i-th nest in the t-th generation. α is the step size scaling factor, which controls the magnitude of the Levy step. $Levy(\lambda)$ is the Levy random path, a random step following the Levy distribution. λ is a parameter of the Levy distribution [18]. The new nests are generated based on the calculation shown in Equation (4).

$$X_{new} = X_j^t + rand \cdot (X_k^t - X_l^t) \tag{4}$$

In Equation (4), X_{new} is the new nest position vector. X_j^t is the current best-performing nest in the iteration. X_k^t and X_l^t represent randomly selected nest position vectors indexed by k and l. $rand$ is a random scalar within interval $[0,1]$. The hybrid algorithm combining ADASYN-RF and CS-SVM (ADRS) follows the workflow illustrated in Figure 3.

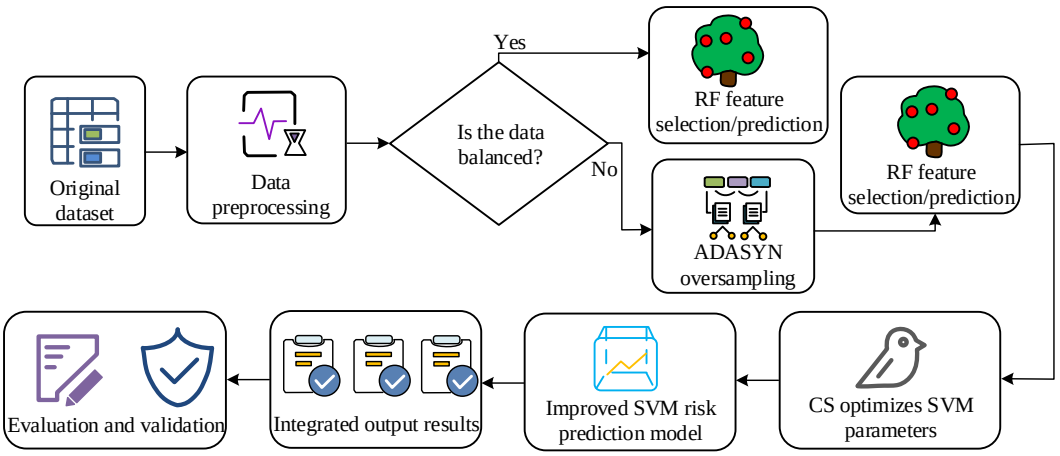


Figure 3: Workflow of the ADRS algorithm

As shown in Figure 3, ADRS first collects the original dataset and performs preprocessing, including filling missing values and applying standardization or normalization. Then, it checks for data balance. If the data is balanced, RF is directly applied for feature selection and prediction. If the data is imbalanced, ADASYN is used for oversampling, followed by RF for feature selection and prediction. Next, the CS algorithm is used to optimize SVM parameters and build an improved SVM risk prediction model. The results from RF and the improved SVM are integrated and output together. Finally, evaluation metrics such as the confusion matrix are used to assess the performance of the ensemble model and determine its risk prediction capability. The calculation of RF feature selection is shown in Equation (5).

$$Gini(X) = 1 - \sum_{k=1}^{K} p_k^2 \tag{5}$$

In Equation (5), $Gini(X)$ represents the Gini index of feature X, which measures the impurity of the dataset. K is the total number of classes, and p_k is the proportion of samples in class k. After completing these steps, SVM performance needs to be evaluated in preparation for optimization, as shown in Equation (6).

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

In Equation (6), TP is the number of correctly predicted positive samples. TN is the number of correctly predicted negative samples. FP represents the number of false positives, and FN represents the number of false negatives.

3.2 Investment risk prediction model based on the ADRS hybrid algorithm

Although the ADRS algorithm achieves high prediction accuracy, real-world investment data exhibit three characteristics: unstructured information, strong temporal dependence, and sensitivity to unexpected events. These features result in dynamic and nonlinear structures in investment data. However, the ADRS algorithm struggles to capture such dynamic nonlinear relationships. The PCA-BP hybrid algorithm, combining Principal Component Analysis (PCA) and

Backpropagation (BP) neural network, effectively addresses this issue. It captures nonlinear temporal patterns and dynamically adapts to data changes through layered nonlinear transformation, while dimensionality reduction helps remove redundant features. This approach addresses the challenges of processing unstructured data and modeling temporal dependencies [19]. The workflow of PCA-BP is shown in Figure 4.

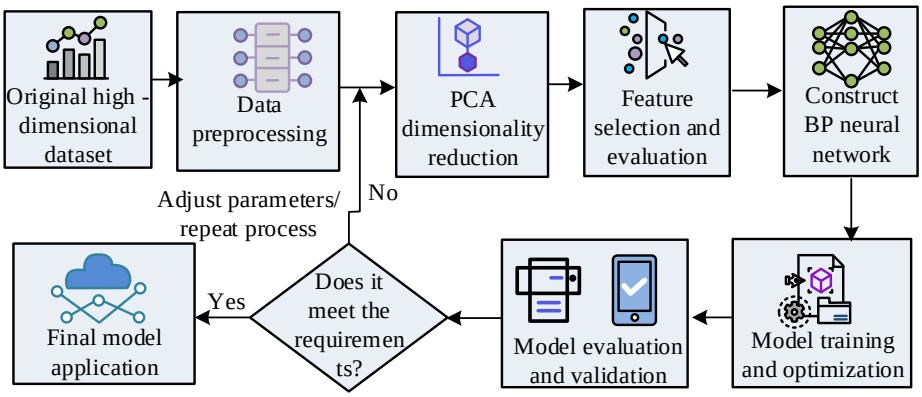


Figure 4: Workflow of the PCA-BP algorithm

In Figure 4, PCA-BP first loads the original high-dimensional investment dataset and performs data preprocessing. PCA is then applied for dimensionality reduction by calculating the mean vector and covariance matrix to select the top principal components, significantly reducing data dimensions. Next, it selects and evaluates the reduced features to ensure low correlation among them. A BP network is then constructed. The number of input nodes is reduced to the number of principal components, which decreases the number of connections and parameters. This simplifies the network structure, reduces computational cost, and lowers the risk of overfitting. The reduced input dimension also minimizes parameter redundancy and shortens training time, thus improving the BP network. The calculation of the PCA mean vector is shown in Equation (7).

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{7}$$

In Equation (7), $\bar{X}$ represents the mean vector of the data, which is composed of the average values of all features. n is the number of samples, and X_i is the raw data of the i-th sample, represented as a row vector of dimension $1\times m$ [20]. After this, the covariance matrix is calculated as shown in Equation (8).

$$C = \frac{1}{n-1} X_{centered}^{T} X_{centered} \tag{8}$$

In Equation (8), $X_{centered}$ denotes the centralized data matrix, which eliminates the influence of the mean. $X_{centered}^{T}$ is the transpose of the centralized matrix, and C is the covariance matrix. While PCA-BP reduces dimensionality and redundancy via PCA and improves prediction accuracy and efficiency via BP, it still struggles to model dynamic trends and periodic fluctuations in time series data. It also cannot effectively utilize sequence autocorrelation. The Autoregressive Integrated Moving Average (ARIMA) model, specifically designed for time series, stabilizes data through differencing and models trends and seasonality using autoregression and moving average techniques. This enhances PCA-BP's ability to make dynamic predictions. Therefore, the study integrates PCA-BP with ARIMA to form the PBA algorithm. The workflow of PBA is shown in Figure 5.

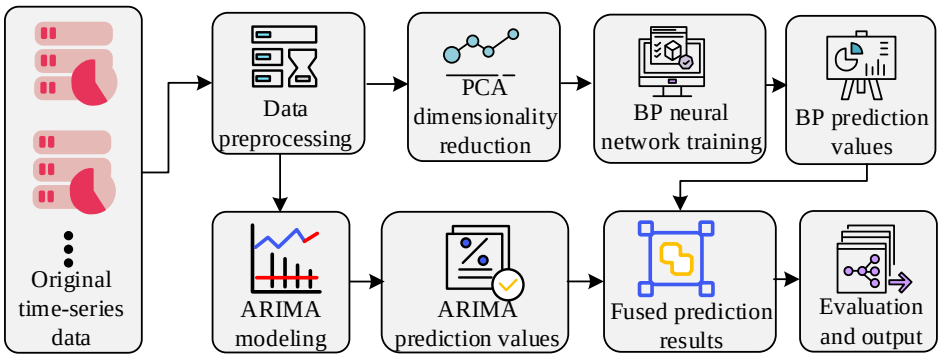


Figure 5: Workflow of the PBA algorithm

As shown in Figure 5, this method first inputs the original time series data and performs preprocessing. One portion of the data is used for ARIMA modeling, which establishes an autoregressive moving average model by differencing the data and performing stationarity analysis. This allows the model to capture sequence autocorrelation and model dynamic trends and seasonal fluctuations. Meanwhile, another portion of the data undergoes PCA-based dimensionality reduction and is input into the BP network for training. Finally, the predictions from ARIMA and BP are integrated to compensate for PCA-BP's weaknesses in handling dynamic patterns. This produces more accurate results, which are then evaluated. In the fusion mechanism, the combination of ARIMA and BP prediction results is not a simple average. The study introduces the idea of combining static weights and dynamic adaptation of learning weights. The static weights are artificially set based on the historical performance of the model, while the learning weights are automatically learned from the data through optimization algorithms. The study first applies the sliding window method to dynamically evaluate historical errors and, based on this, calculates the real-time weight distribution of the ARIMA and BP neural network models. Through the rolling optimization mechanism, the weights are updated with the prediction cycle, thereby adapting to the evolution of the characteristics of time series data. In addition, to prevent a single weight from dominating the fusion result, a weight smoothing constraint term is introduced to ensure the stability of weight distribution and a smooth transition. The ARIMA modeling process is shown in Equation (9).

$$(1-\sum_{i=1}^{p}\phi_i B^i)(1-B)^d Y_t = (1+\sum_{i=1}^{q}\theta_i B^i)\varepsilon_t \quad (9)$$

In Equation (9), Y_t represents the observed value of the time series at time t. B is the lag operator. ϕ_i is the parameter for the autoregressive part. d is the number of differences used to stabilize the series. θ_i represents the parameter of the moving average part. ε_t is the white noise error term. Meanwhile, the other portion of the data is sent to the BP network for training. The forward computation from the input layer to the hidden layer is shown in Equation (10).

$$net_h = \sum_{i=1}^{n} v_{ih} x_i - \Upsilon_h \quad (10)$$

In Equation (10), net_h and Υ_h denote the net input and threshold of the h-th neuron in the hidden layer. n is the number of neurons in the input layer. v_{ih} is the weight connecting the i-th neuron in the input layer to the h-th neuron in the hidden layer. x_i is the input value from the i-th neuron in the input layer. Since the ARIMA and BP prediction models are processed separately, in order to align the range of their prediction results and solve the problems of time lag or errors between the two paths, it is necessary to normalize the two sets of prediction results. The study first scales the predicted values to the interval [0, 1] through the minimum-maximum normalization method. To avoid the influence of time lag on the fusion results, the study adopted the sliding window correlation analysis method to time align the two groups of sequences. The study takes the output of the ARIMA model as the benchmark, calculates the Pearson correlation coefficients between it and the output of the BP model at different time offsets, and selects the offset corresponding to the maximum correlation coefficient as the optimal time compensation parameter. The investment risk prediction model combining ADRS and PBA is named AD-PB. Its workflow is shown in Figure 6.

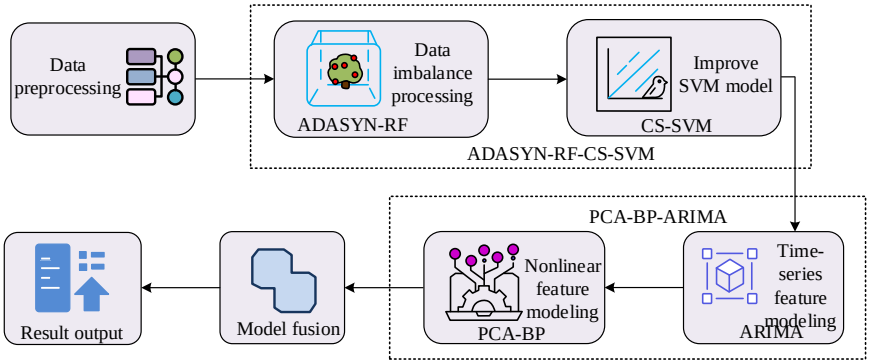


Figure 6: Workflow of the AD-PB algorithm

As shown in Figure 6, this model first takes the raw investment data and performs data cleaning and standardization. The data is then split into two parts. One part is used to build the ADASYN-RF-CS-SVM model. ADASYN balances the risk categories in the training data, RF extracts features, and CS optimizes SVM hyperparameters. The optimized SVM is then integrated with RF for classification modeling. The other part is used to construct the PCA-BP-ARIMA model. Stationarity tests and differencing are applied to the time series data for ARIMA modeling and prediction. Non-time-series data undergoes PCA for dimensionality reduction. The ARIMA predictions are then combined with the PCA-reduced features and input into the BP network for

training. Finally, the outputs from the two models are fused at both the feature and result levels. Techniques such as weighted averaging are used to produce the final investment risk prediction. The forward propagation in BP network training is calculated as shown in Equation (11).

$$a_j = f\left(\sum_{i=1}^{m} w_{ji} x_i + b_j\right) \tag{11}$$

In Equation (11), w_{ji} represents the weight from the i-th input neuron to the j-th neuron in the current layer. b_j and a_j are the bias term and output of the j-th neuron in the current layer. After this computation, BP also performs backpropagation, which is calculated as shown in Equation (12).

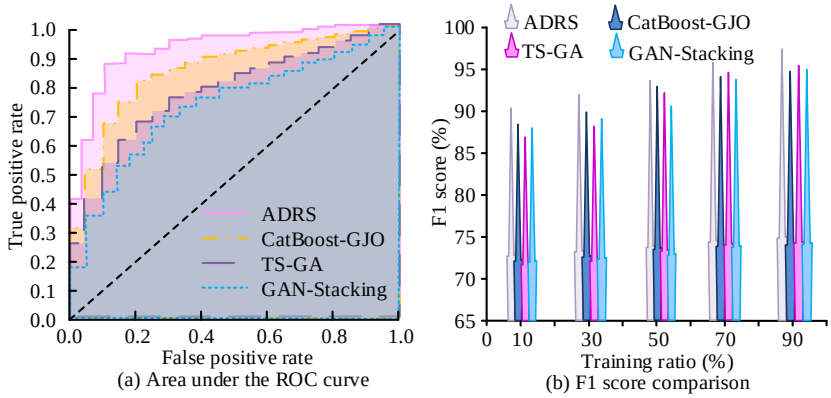
$$\begin{cases} \Delta w_{ji} = -\eta \dfrac{\vartheta E}{\vartheta w_{ji}} \\ \Delta b_j = -\eta \dfrac{\vartheta E}{\vartheta b_j} \end{cases} \tag{12}$$

In Equation (12), Δw_{ji} and Δb_j are the update values for weight w_{ji} and bias b_j, respectively. η is the learning rate. E is the error function. $\dfrac{\vartheta E}{\vartheta w_{ji}}$ and $\dfrac{\vartheta E}{\vartheta b_j}$ are the partial derivatives of the error function with respect to weight w_{ji} and bias b_j, respectively.

4 Performance analysis of the investment risk prediction model based on ADASYN-RF

4.1 Validation of the effectiveness of the ADRS algorithm

To verify the superiority of the ADRS risk prediction algorithm, this study compared it with three other algorithms: CatBoost-GJO (a combination of Categorical Boosting and Golden Jackal Optimization), TS-GA (a combination of Tabu Search and Genetic Algorithm), and GAN-Stacking (a combination of Generative Adversarial Network and Stacked Generalization). Risk scenarios were simulated through data injection, and the dataset used for risk prediction was the S&P 500 historical trading dataset, which is widely recognized and representative in the field of financial risk research. The investment time frame in the experiment was set from January 2020 to December 2024, covering asset classes such as stocks, bonds, and derivatives. Market conditions included bull and bear market transitions as well as extreme volatility scenarios. The study first preprocesses the original data using missing value imputation, standardization, and sliding window feature construction, then extracts key features and divides them into training and test sets. Price data includes the opening price, closing price, highest price and lowest price. The return data includes daily return rate and weekly return rate. Macroeconomic variables include the inflation rate, GDP growth rate, changes in interest rates, etc. The experiments were implemented using Python as the programming environment, with a high-performance CPU as the hardware platform, and TensorFlow as the deep learning framework. Historical investment data was used as the dataset. First, the study compared the Area under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve and the F1 scores for the four algorithms. The experimental results are shown in Figure 7.



(a) Area under the ROC curve (b) F1 score comparison

Figure 7: Comparison results of ROC curve and F1 score

As shown in Figure 7 (a), the ROC curve of the ADRS algorithm was closest to the top-left corner. Its AUC value reached 0.897, the highest among all algorithms, significantly outperforming the comparison methods with AUC values of 0.857, 0.793, and 0.781. According to Figure 7 (b), the F1 scores of the ADRS algorithm were all above 90.0%, with the lowest and highest values being 91.1% and 97.5%, respectively. These results were higher than those of the other algorithms, further demonstrating the classification performance of ADRS. In summary, the

ADRS algorithm outperformed the comparison algorithms in both classification and recall. To further evaluate the prediction performance of each model, the study compared the predicted values with the actual values. The results are shown in Figure 8.

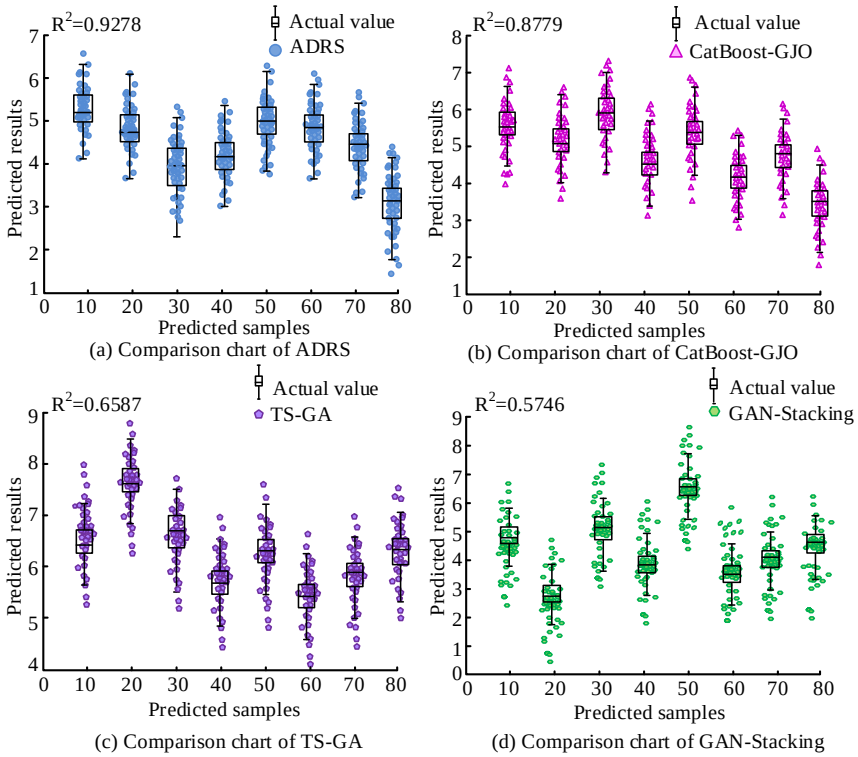


Figure 8: Comparison results of predicted values and actual values

As shown in Figure 8 (a), the predicted values of the ADRS algorithm closely matched the actual values, with a Coefficient of Determination (R^2) of 0.9278. According to Figure 8 (b), CatBoost-GJO also produced predictions that closely matched the actual values, with an R^2 of 0.8779. In contrast, TS-GA and GAN-Stacking showed lower agreement between predicted and actual values, with more outliers. Their R^2 values were 0.6587 and 0.5746, respectively. Overall, ADRS exhibited the smallest prediction error and the best fit among all models. To evaluate the prediction efficiency of the four algorithms, the study also compared their memory usage and response time. The experimental results are shown in Figure 9.

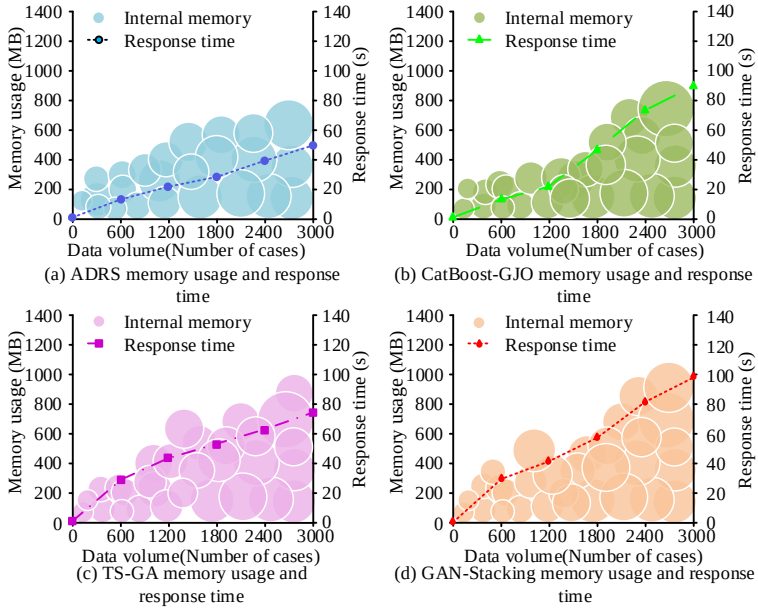


Figure 9: Comparison results of memory usage and response time

As shown in Figure 9 (a), the memory usage of the ADRS algorithm increased slowly, with a maximum memory consumption of 795 MB. When the dataset size reached 3000, its response time was 50 seconds. According to Figure 9 (b), CatBoost-GJO had a maximum memory usage of 950 MB and the highest response time of 89 seconds. From Figure 9 (c) and 9 (d), the maximum memory usage of TS-GA and GAN-Stacking was 1010 MB and 1080 MB, respectively. Their highest response times were 74 seconds and 99 seconds. In summary, compared with the other algorithms, ADRS demonstrated more stable and efficient data processing performance.

4.2 Application evaluation of the improved ADASYN-RF investment risk prediction model

To further evaluate the performance of the AD-PB investment risk prediction model based on the ADASYN-RF algorithm, the study compared it with three other investment risk prediction models constructed using CatBoost-GJO, TS-GA, and GAN-Stacking. The experimental environment consisted of a Python-based data analysis platform, a risk assessment algorithm library, and data visualization tools. The research selected key time points in different market cycles for stress tests, including the sharp decline at the beginning of the 2020 COVID-19 pandemic, the technology stock boom in the first quarter of 2021, and the market adjustment triggered by the Federal Reserve's interest rate hikes in 2022. Each round of testing involves 1,000 simulated transactions to assess the model's robustness and generalization ability in actual investment scenarios. To avoid overfitting, a five-fold cross-validation mechanism was introduced during the training process, and the Monte Carlo method was adopted to randomly sample and optimize the parameter space. The data segmentation, hyperparameter tuning range, and repeatable protocol are as follows: The data segmentation adopts the time series partitioning method, with the ratio of the training set to the test set being 8:2, ensuring the continuity of the time series. The range of hyperparameter tuning covers decision tree depth of 3 to 12, learning rate of 0.01 to 0.3, and feature sampling rate of 0.6 to 1.0. The experimental process followed the reproducibility protocol, with all random seeds fixed and the data preprocessing steps standardized. The study used AD-PB, CatBoost-GJO, TS-GA, and GAN-Stacking to predict four types of investment risks: macroeconomic conditions, industry development trends, corporate financial performance, and market price fluctuations. The prediction accuracy results are shown in Figure 10.

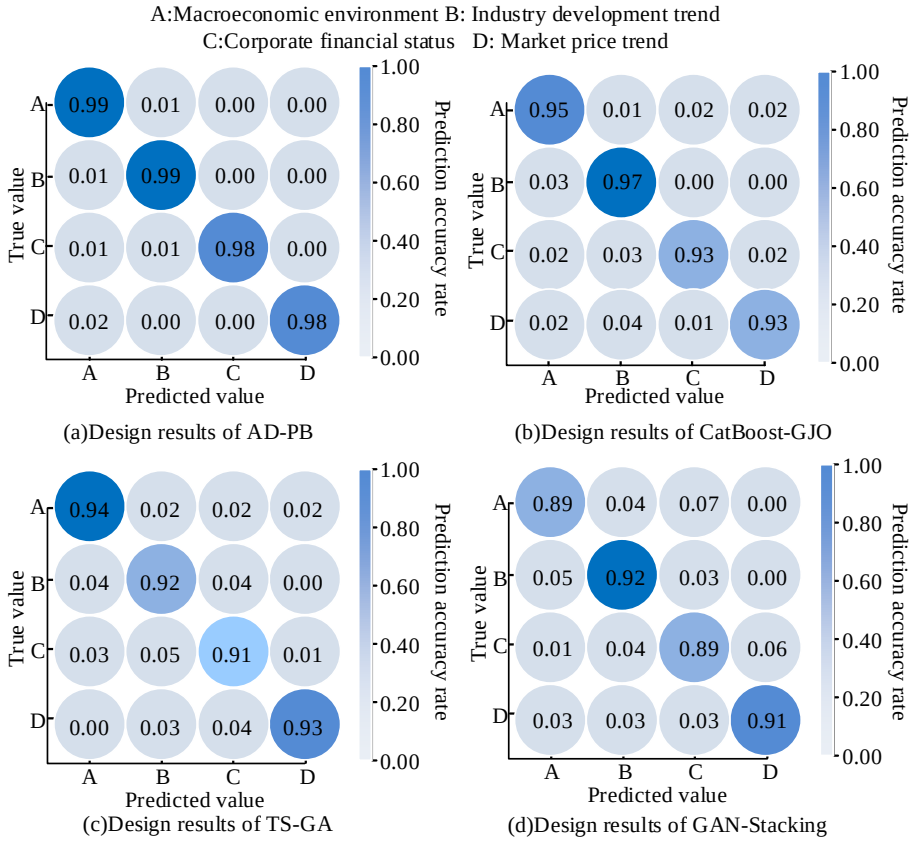


Figure 10: Comparison results of classification prediction accuracy

As shown in Figure 10 (a), the AD-PB model achieved a minimum prediction accuracy of 98.0% for corporate financial performance and market price fluctuations. Its maximum prediction accuracy reached 99.0% for macroeconomic conditions and industry development trends. According to Figure 10 (b), the CatBoost-GJO model reached a maximum prediction accuracy of 97.0%. The TS-GA and GAN-Stacking

models performed less accurately, with maximum prediction accuracies of 94.0% and 92.0%, respectively. Overall, the AD-PB model demonstrated the highest prediction accuracy among all models. To further evaluate the risk assessment capabilities of the four models, the study compared their Sharpe ratios and maximum drawdowns. The results are shown in Figure 11.

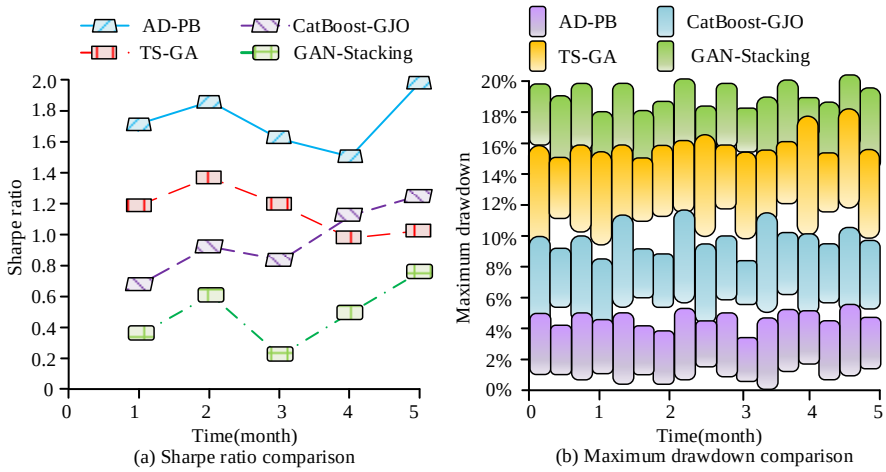


Figure 11: Comparison results of Sharpe ratio and maximum drawdown

As shown in Figure 11 (a), the AD-PB model consistently achieved the highest Sharpe ratio, with a peak value of 1.9. Among the comparison models, TS-GA performed relatively well, with a maximum Sharpe ratio of 1.4. According to Figure 11 (b), the AD-PB model had the smallest maximum drawdown overall, with a maximum value of 5.9%. In contrast, the GAN-Stacking model exhibited the largest drawdown, reaching 19.1%. In conclusion, the AD-PB model demonstrated strong risk resilience. Subsequently, in order to comprehensively verify the risk prediction performance of the four models, the study compared volatility, value at risk, win rate and profit-to-loss ratio. The profit-loss ratio is calculated as the absolute value of the average profit divided by the average loss, i.e., profit-loss ratio = average profit / average loss. In addition, the study introduced three indicators, namely Precision, Recall and Specificity, to further evaluate the predictive efficacy of the model. The value range of Specificity is between 0 and 1. The larger the value, the stronger the model's ability to identify non-risk events. The results are shown in Table 3.

Table 3: Results of volatility, value at risk, win rate, and profit-loss ratio

Model	Volatility (%)	Value at risk (%)	Win rate (%)	Profit-loss ratio	Precision	Recall	Specificity
AD-PB	8.2±1.5@#*	7.3±1.2@#*	88.9±8.4@#*	6:1	0.96±0.04@#*	0.94±0.03@#*	0.93±0.05@#*
CatBoost-GJO	15.7±2.4	8.9±2.0	79.6±7.6	9:2	0.91±0.05	0.90±0.04	0.90±0.06
TS-GA	17.6±2.8	16.8±2.5	71.4±6.9	7:2	0.89±0.04	0.89±0.06	0.88±0.05
GAN-Stacking	19.8±3.0	20.4±3.7	65.8±6.5	3:1	0.86±0.06	0.85±0.05	0.86±0.07

*Note: In Table 3, @ indicates that the value of the AD-PB index is significantly different from that of CatBoost-GJO, p<0.05; # indicates that the difference between the AD-PB index value and TS-GA is significant, p<0.05; * Indicates that the difference in the AD-PB index value compared with GAN-Stacking is significant, p<0.05*

Table 3 shows that the AD-PB model achieves the highest levels of Precision, Recall, and Specificity—0.96, 0.94, and 0.93, respectively—indicating superior accuracy and stability in identifying risk events. In contrast, the three indicators of other models were all lower than those of the AD-PB model. Especially, the GAN-Stacking model had the lowest values, which were 0.86, 0.85 and 0.86 respectively, indicating that it was relatively weak in identifying risk events. The AD-PB model achieved a volatility of 8.2%, a value at risk of 7.3%, a win rate of 88.9%, and a profit-loss ratio of 6:1. The CatBoost-GJO model yielded a volatility of 15.7%, a value at risk of 8.9%, a win rate of 79.6%, and a profit-loss ratio of 9:2. Among the TS-GA and GAN-Stacking models, TS-GA demonstrated relatively better overall performance, with a volatility of 17.6%, a value at risk of 16.8%, a win rate of 71.4%, and a profit-loss ratio of 7:2. In summary, the AD-PB model performed well in risk control, profitability, and return efficiency.

The features selected based on RF and SVM components include trading frequency (Feature 1), holding time (Feature 2), stop-loss range (Feature 3),

profit target (Feature 4), capital utilization rate (Feature 5), account balance volatility (Feature 6), order interval time (Feature 7), maximum drawdown range (Feature 8), average position ratio (Feature 9), and risk preference coefficient (Feature 10). The importance weights of each feature are shown in Figure 12.

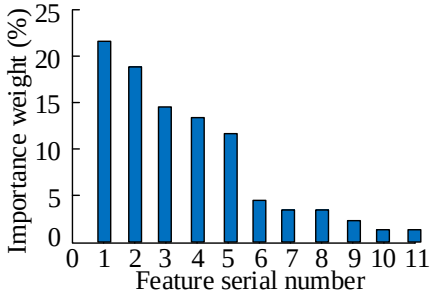


Figure 12: The importance weights of each feature

As can be seen from Figure 12, the feature weight distribution screened out by the RF and SVM models is relatively concentrated, among which the cumulative weight of the top five features exceeds 60%. The weight proportions of Feature 1 and Feature 2 are the highest, reaching 22% and 19% respectively, indicating that the strategy is highly sensitive to market response speed and holding period. The weights of Feature 3 and Feature 4 are 14% and 13% respectively, indicating that risk control and profit setting have a relatively high priority in model decision-making. Feature 5 accounts for 12%, reflecting the trading strategy's emphasis on capital efficiency. The proportion of other features does not exceed 10%.

To test the complexity of the proposed method in the research and explore its sensitivity to the severity of data imbalance and parameter changes, an experiment was designed to analyze its robustness and time complexity. Specifically, the study tests the model's performance under different data distributions by adjusting the ratio of positive to negative samples in the training set. The data distribution includes multiple gradient settings, with the proportion of positive samples ranging from 10% to 90%. The comparison indicators include accuracy rate, F1 score, AUC value and model training time. The results are shown in Table 4.

Table 4: The performance of the model under different data distributions

Proportion of positive samples	Accuracy rate (%)	F1	AUC value	Model training time (s)
10%	92.7	0.90	0.93	41.20
20%	92.9	0.91	0.94	38.77
30%	93.0	0.92	0.95	39.68
40%	93.5	0.92	0.96	37.94
50%	94.3	0.93	0.97	37.48
60%	94.0	0.92	0.96	37.99
70%	93.8	0.91	0.96	38.56
80%	93.2	0.90	0.95	39.77
90%	92.7	0.90	0.94	40.54

As shown in Table 4, as the proportion of positive samples gradually increases from 10% to 90%, the overall accuracy of the model shows a trend of first rising and then falling, reaching a peak of 94.3% when the proportion of positive samples is 50%. The F1 score and AUC value also exhibited a similar trend, reaching optimal values of 0.93 and 0.97, respectively, when the proportion was 50%, indicating that the model achieved the highest recognition ability for positive and negative samples under this condition. The training time remained basically stable, averaging approximately 39.10 s, demonstrating that the algorithm has good adaptability and stability to changes in data distribution.

5 Conclusion

To address the issues of existing investment risk prediction methods, such as reliance on historical data leading to poor adaptability, insufficient capture of complex market dynamics, and low prediction accuracy, this study innovatively proposed an investment risk prediction model based on the ADASYN-RF algorithm. The model balances the data using ADASYN, simplifies features through PCA, captures temporal patterns with ARIMA, and uses RF and CS-SVM for classification and prediction, enabling precise and efficient investment risk forecasting. The experimental results show that the AD-PB prediction model based on the ADASYN-RF algorithm achieved a maximum prediction accuracy of 99.0% for different risk types, with the lowest accuracy reaching 98.0%. Additionally, its Sharpe ratio and maximum drawdown were 1.9 and 5.9%, respectively. In terms of risk assessment performance, the AD-PB model outperformed the comparison models with a volatility of 8.2%, a value at risk of 7.3%, a win rate of 88.9%, and a profit-loss ratio of 6:1. Overall, the ADASYN-RF-based investment risk prediction model demonstrates strong risk resistance and prediction capabilities, meeting the needs of enterprises for investment risk forecasting. Although the proposed model performed excellently in terms of both performance and application, it has not yet addressed policy risks and irrational user behavior during the prediction process. Therefore, the accuracy and universality of the predictions still need to be improved. In the future, further efforts will be made to separate these influencing factors and continuously enhance the model's prediction accuracy.

6 Discussion

Compared with CatBoost-GJO, TS-GA, and GAN-Stacking, the prediction accuracy of AD-PB increased by 1.2%, 3.5%, and 4.7%, respectively, demonstrating stronger stability and generalization ability. Furthermore, when dealing with a highly volatile market environment, the volatility of AD-PB is only 8.2%, significantly lower than the 15.7%, 17.6% and higher levels of other models, indicating that it has a stronger risk control ability in extreme market conditions. Meanwhile, the win rate of AD-PB is 88.9%, and the profit-to-loss ratio is 6:1, demonstrating high profitability and stability. These

characteristics enable AD-PB to more effectively identify and avoid risks when facing a complex and volatile market environment, thereby enhancing the reliability and return performance of investment decisions. This is because the AD-PB model combines the oversampling advantage of ADASYN with the nonlinear classification ability of RF. Meanwhile, it effectively reduces feature redundancy through PCA, and ARIMA enhances the capture of time series trends. This multi-stage collaborative mechanism not only enhances the model's ability to identify complex risk patterns but also improves the robustness and adaptability of predictions, increasing its applicability in diverse market environments. In addition, AD-PB performs well in terms of memory efficiency, F1 score, Sharpe ratio, and drawdown reduction. This is closely related to its architectural design. The ADASYN component effectively alleviates the problem of data imbalance, enabling the model to still have reliable predictive capabilities in small sample risk categories. Meanwhile, the random feature selection mechanism of RF further optimizes the generalization performance of the model. Meanwhile, the introduction of PCA not only enhances computational efficiency but also provides a guarantee for the compact expression of the feature space. ARIMA, on the other hand, effectively compensates for the dynamic changes in market trends. These designs enable AD-PB to perform exceptionally well in multi-dimensional evaluations and possess stronger practicality and promotion potential.

7 Model replication

To facilitate the reproduction and application of the AD-PB model by other researchers or practitioners, this section will detail the construction steps and key parameter Settings of the model. Firstly, in the data preprocessing stage, the ADASYN algorithm is used to oversample the imbalanced samples in the training set, enhancing the model's recognition ability for minority risk samples. Subsequently, the features are reduced in dimensionality through PCA, retaining principal components with a cumulative contribution exceeding 90%, thereby reducing feature redundancy and improving computational efficiency. In the model training section, the number of decision trees of the Random Forest (RF) classifier is set to 200, with a maximum depth of 10, and the Gini index is adopted as the splitting criterion. For the ARIMA component, the differential method is used to perform stationarity processing on the target sequence, and the optimal (p, d, q) parameter combination is determined based on the AIC criterion. Finally, the prediction results of RF are used as input, and the trend prediction output of ARIMA is fused via linear weighting to form the final AD-PB prediction model.

References

[1] Qin G, Juan M, Rui M H. Iot-based intelligent power supply management using ensemble learning for seismic observation stations. Informatica, 2025, 49(8). 1-18. DOI: 10.31449/inf.v49i8.6502

[2] Mahmood S A, Hamadi S S. Ensemble machine learning algorithms for predicting thyroid disorders in diabetic patients: a comparative analysis. Informatica, 2025, 49(24). 7-25. DOI: 10.31449/inf.v49i24.8373

[3] Ileberi E, Sun Y. Advancing model performance With ADASYN and recurrent feature elimination and cross-validation in machine learning-assisted credit card fraud detection: A comparative analysis. IEEE access, 2024. 12(1): 133315-133327. DOI: 10.1109/ACCESS.2024.3457922

[4] Aish M A. Predictive modeling of cerebral strokes: an ADASYN-RF approach for imbalanced data. VFAST Transactions on Software Engineering, 2024, 12(4): 12-26. DOI: 10.21015/vtse.v12i4.1932

[5] Zhang M. Enhanced load forecasting for electric vehicle charging stations using a hybrid random forest-convolutional neural network algorithm. Informatica, 2024, 48(19): 89-102. DOI: 10.31449/inf.v48i19.6649

[6] Saiyed A. AI-Driven Innovations in Fintech: Applications, challenges, and future trends. International Journal of Electrical and Computer Engineering Research, 2025, 5(1): 8-15. DOI: 10.53375/ijecer.2025.437

[7] Song M, Ma H, Zhu Y, Zhang M. Credit risk prediction based on improved ADASYN sampling and optimized LightGBM. Journal of Social Computing, 2024, 5(3): 232-241. DOI: 10.23919/JSC.2024.0019.

[8] Salehpour A, Norouzi M, Balafar M A, SamadZamini K. A cloud-based hybrid intrusion detection framework using XGBoost and ADASYN-Augmented random forest for IoMT. IET Communications, 2024, 18(19): 1371-1390. DOI: 10.1049/cmu2.12833

[9] Dawei Y, Bing Z, Bingbing G, Xibo G, Razzaghzadeh B. Predicting the CPT-based pile set-up parameters using HHO-RF and PSO-RF hybrid models. Structural Engineering and Mechanics, 2023, 86(5): 673-686. DOI: 10.12989/sem.2023.86.5.673

[10] Ding W, Sun H. Prediction of PM2. 5 concentrations based on the weighted RF-LSTM model. Earth Science Informatics, 2023, 16(4): 3023-3037. DOI: 10.1007/s12145-023-01111-7

[11] Jiao Z. Dynamic Financial Distress Prediction Using Combined LASSO and GBDT Algorithms. Informatica, 2024, 48(17): 139-152. DOI: 10.31449/inf.v48il7.6493

[12] Belhadi A, Kamble S S, Mani V, Benkhati I, Touriki F E. An ensemble machine learning approach for forecasting credit risk of agricultural SMEs' investments in agriculture 4.0 through supply chain finance. Annals of Operations Research, 2025, 345(2): 779-807. DOI: 10.1007/s10479-021-04366-9

[13] Li X, Wang J, Yang C. Risk prediction in financial management of listed companies based on optimized BP neural network under digital economy. Neural Computing and Applications, 2023, 35(3): 2045-

2058. DOI: 10.1007/s00521-022-07377-0

[14] Doğru A, Buyrukoğlu S, Arı M. A hybrid super ensemble learning model for the early-stage prediction of diabetes risk. Medical & Biological Engineering & Computing, 2023, 61(3): 785-797. DOI: 10.1007/s11517-022-02749-z

[15] Zheng G, Kong L, Brintrup A. Federated machine learning for privacy preserving, collective supply chain risk prediction. International Journal of Production Research, 2023, 61(23): 8115-8132. DOI: 10.1080/00207543.2022.2164628

[16] Rhodes J S, Cutler A, Moon K R. Geometry-and accuracy-preserving random forest proximities. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(9): 10947-10959. DOI: 10.1109/TPAMI.2023.3263774

[17] Kurani A, Doshi P, Vakharia A, Shah M. A comprehensive comparative study of artificial neural network (ANN) and support vector machines (SVM) on stock forecasting. Annals of Data Science, 2023, 10(1): 183-208. DOI: 10.1007/s40745-021-00344-x

[18] Cheng J, Xiong Y. Multi-strategy adaptive cuckoo search algorithm for numerical optimization. Artificial Intelligence Review, 2023, 56(3): 2031-2055. DOI: 10.1007/s10462-022-10222-4

[19] Fei R, Guo Y, Li J, Hu B, Yang L. An improved BPNN method based on probability density for indoor location. IEICE TRANSACTIONS on Information and Systems, 2023, 106(5): 773-785. DOI: 10.1587/transinf.2022DLP0073

[20] Marukatat S. Tutorial on PCA and approximate PCA and approximate kernel PCA. Artificial Intelligence Review, 2023, 56(6): 5445-5477. DOI: 10.1007/s10462-022-10297-z

A Systematic Review of Optimization Techniques for Computation Offloading in Mobile Edge Computing: Lyapunov, Convex, Heuristic, Game-Theoretic, and Machine Learning Approaches

Mohammad Ashique E Rasool[*1], Anoop Kumar [2], Asharul Islam [3]
[1, 2]Faculty of Mathematics and Computing, Banasthali Vidyapith, Rajasthan, Postal Code 304022, India
[3]College of Computer Science, King Khalid University, Abha, Postal Code 61421, Saudi Arabia
Email: ashique.rasool@gmail.com, anupbhola@banasthali.in, aaleslam@kku.edu.sa
[*]Corresponding author

Keywords: Computation offloading, edge computing, optimization methods, systematic review

Received: August 28, 2025

Mobile Edge Computing (MEC) has become a key paradigm for reducing latency, energy consumption, and system overhead in computation-intensive applications by enabling task offloading to edge servers. In this paper, we conduct a Systematic Literature Review (SLR) of 70 peer-reviewed studies, with 30 papers coded in detail across five main optimization approaches: Lyapunov, convex, heuristic, game-theoretic, and machine learning methods. The review systematically compares these paradigms in terms of optimization objectives, task assumptions, and evaluation setups. Our analysis reveals that while Lyapunov and game-theoretic approaches rely entirely on simulation (100%), heuristic studies exhibit stronger practical grounding, with 33% including real-world validation. Convex methods demonstrated 10–18% energy savings in heterogeneous MEC scenarios, whereas machine learning approaches showed adaptability under uncertain conditions but lacked real-world testing. Beyond numerical comparison, the review highlights the methodological evolution of MEC optimization, from formal mathematical programming to learning-driven and hybrid approaches. The coding of 30 representative papers provides a quantitative foundation that exposes common assumptions—such as independent task models and simplified simulators—that limit real-world applicability. At the same time, the synthesis identifies promising trends, including reinforcement learning for adaptive decision-making and hybrid frameworks that combine convex optimization with predictive models. These insights underscore the need for future MEC strategies that balance computational efficiency, adaptability, and scalability. By synthesizing results, quantifying trends, and framing open challenges, this SLR provides researchers and practitioners with a comprehensive understanding of the strengths, limitations, and opportunities in MEC computation offloading.

Povzetek: SLR o MEC odlaganju sintetizira 70 študij, primerja Lyapunovske, konveksne, hevristične in ML pristope, razkrije prevlado simulacij in poenostavitev ter izpostavi trend k RL ter hibridnim metodam za uravnoteženje učinkovitosti, prilagodljivosti in razširljivosti v praksi.

1 Introduction

The unprecedented growth in smartphone adoption, with over 5 billion users and 8.58 billion active mobile subscriptions worldwide, has driven a transformative shift in the digital ecosystem. This trend is amplified by the expected deployment of 55.7 billion IoT devices by 2025, which are anticipated to generate an overwhelming 80 zettabytes of data [1], [2]. Huawei projects an even more significant surge, estimating global data volumes will reach 181 zettabytes by 2025 [3]. This massive influx of data presents significant challenges for processing and utilization, necessitating innovative solutions to manage the demands of connected devices.

Cloud technology and data centers have become indispensable in handling this exponential data growth, with predictions indicating that 95% of new digital data will be processed on cloud infrastructures by 2025 [4]. Concurrently, mobile devices are increasingly hosting resource-intensive applications, including virtual and augmented reality, video streaming, image processing, and health monitoring, each with distinct and evolving computational requirements. Natural language processing tasks further add complexity, exhibiting sporadic and variable demands based on user interactions [5]. While quantum computing shows promise, its current capabilities are insufficient to manage the extensive data streams generated by mobile and IoT devices. Mobile Edge Computing (MEC) has emerged as a viable solution to address these challenges. By deploying resource-rich MEC servers closer to end devices, MEC improves application performance while reducing latency and energy consumption. This paradigm offers a practical approach to handling the computational demands of modern applications, creating a need for optimized offloading strategies to balance execution delay, energy

efficiency, and resource utilization. Figure 1 illustrates the MEC network architecture [6].

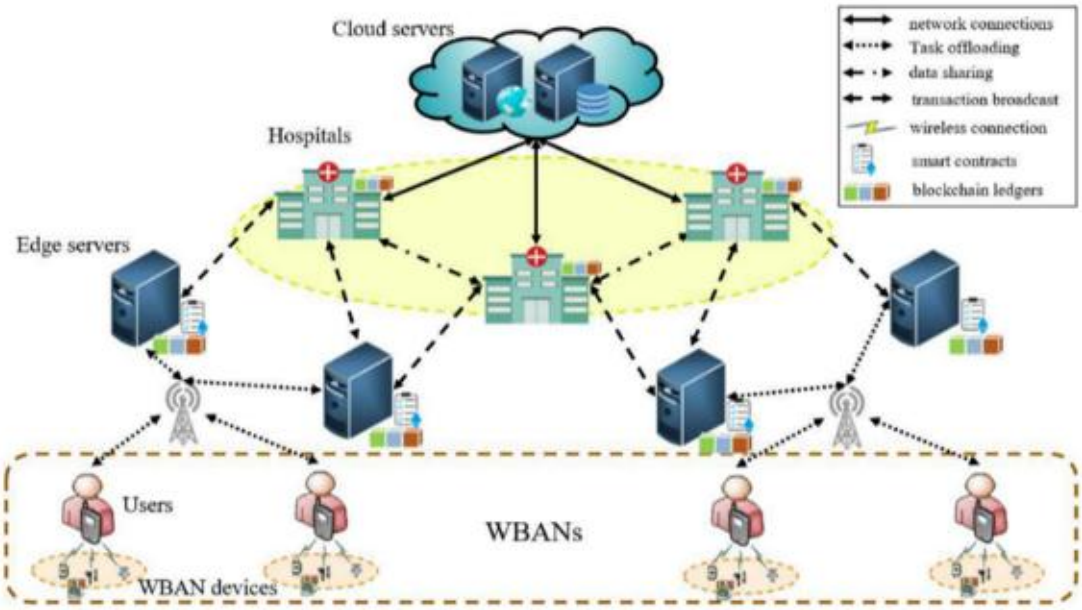


Figure 1: Depicts the network architecture of MEC environment [6]

This paper provides a detailed comparative analysis of computation offloading optimization techniques within MEC environments. It evaluates their strengths, limitations, and future potential, aiming to guide advancements in the field. The structure of the paper is as follows: Section 2 reviews related works to establish a foundation for the study. Section 3 introduces a taxonomy of offloading approaches, while Section 4 explores optimization techniques and algorithms. Section 5 presents a critical analysis of offloading optimization methods in Edge Computing, and Section 6 includes discussions, section 7 presents challenges and future research directions. Finally, Section 8 summarizes key findings and contributions.

2 Related work

This section reviews prior survey papers and related works in computation offloading and Mobile Edge Computing (MEC). The aim is to summarize existing surveys to highlight their scope, limitations, and the research gaps they leave unaddressed. By positioning our work against these earlier efforts, we clearly establish the novelty of our contribution, which is presented later in Section 5 as a comprehensive systematic analysis of the field. While previous studies have explored offloading decisions, resource allocation, and collaborative optimization, a comprehensive synthesis is still lacking. This review addresses the gap by systematically examining optimization techniques within EC frameworks, serving as a foundational reference for researchers and practitioners.

Zhang and Shi [7] explored allocation and management schemes in mobile edge networks and demonstrated improvements in throughput and system utility under varying user densities. More recently, Zhang [8] proposed a multi-objective genetic algorithm with knowledge-based crossover and segmentation mutation, achieving significant gains in energy consumption, system utilization, and task completion rates.

Mach and Becvar [9] investigated Multiaccess Edge Computing (MEC) integration, focusing on computational task offloading and resource allocation for efficient task distribution. Similarly, Lin et al. [10] explored task partitioning and execution challenges in EC, emphasizing the importance of optimizing task distribution to enhance performance. Zheng et al. [11] provided a concise survey of common scenarios and strategies for offloading in edge computing.

Shakarami et al. contributed significantly through two reviews: one addressing game-theoretic approaches to collaborative optimization [12], and another focusing on machine learning (ML) methods to enhance system performance [13]. Their work highlights the potential of ML in refining offloading processes. Nguyen and Dressler [14], examined cooperative communication systems, comparing algorithms but without delving into their computational complexities.

The growing focus on EC and offloading optimization has resulted in a wealth of innovative solutions, yet gaps remain. Existing reviews often target specific methodologies, leaving a need for a holistic analysis of optimization methods tailored to computation offloading.

To clarify the novelty of our work compared to prior reviews, we summarize the scope, strengths, and limitations of previous surveys in Table 1. Unlike these earlier efforts, which typically focused on a single optimization paradigm or limited domain, our work

provides a holistic, PRISMA-based systematic review of 71 studies covering Lyapunov, convex, heuristic, game-theoretic, and machine learning approaches. This broader scope and methodological rigor clearly distinguish our contribution.

Table 1: Compression of previous work

Study	Scope / Focus	Strengths	Limitations	Contribution of This Work
[15]	Focus on game-theoretic models for MEC offloading	Clear categorization of equilibrium-based methods	Narrow focus on game theory; ignores heuristics and ML-based methods	Our work covers multiple paradigms (Lyapunov, convex, heuristics, game theory, ML)
[16]	Energy-efficient MEC optimization using heuristics	Strong treatment of heuristic techniques	Does not address formal optimization or machine learning; limited dataset	We provide a systematic comparison across 71 papers, including formal and learning-based approaches
[17]	Survey on convex and Lyapunov optimization in MEC	Rigorous mathematical insights	Overlooks heuristic and data-driven methods; no empirical coding of papers	Our work integrates formal optimization with heuristic and ML trends, bridging the gap
[18]	MEC offloading in IoT contexts	Good discussion of IoT-specific architectures	Lacks generalizability; not comprehensive beyond IoT	Our work provides a holistic view of MEC offloading across domains
This Work	Systematic Literature Review of 74 studies, analyzing multiple optimization paradigms	Comprehensive scope; PRISMA-based methodology; quantitative coding of 30 papers	-	Holistic, systematic SLR covering multiple optimization paradigms, with explicit identification of trends, challenges, and gaps

This review aims to bridge this gap, offering a consolidated resource for understanding the current state of research, identifying gaps, and guiding future investigations.

Key contributions of this review include:

- *Exploration of Communication and Computation Models:* Providing foundational insights into models relevant to computation offloading.
- *Examination of Optimization Methods:* Analyzing strengths, limitations, and applicability of optimization techniques across various domains.
- *Identification of Research Issues:* Highlighting unexplored areas to inspire further research into emerging challenges and opportunities.

This review scrutinizes the optimization challenges in EC, categorizing methodologies into widely applied techniques like Lyapunov optimization [19], [20], convex optimization [21], [22], heuristic optimization [23], game theory [24], and machine learning based optimization techniques [25], [26].

By presenting a consolidated analysis, this review provides a pivotal resource for researchers and practitioners to navigate the intricate domain of computation offloading in EC environments efficiently.

To complement this comparison with prior surveys, the detailed attributes of the papers reviewed in this SLR are presented in Tables 3–7. These tables summarize optimization methods, evaluation criteria, task assumptions, and experimental environments (simulation or real) for the 30 coded studies. Representative performance outcomes (e.g., latency reduction, energy savings, and adaptability) are further synthesized in Section 5 to highlight the distinct contributions and gaps in the current state-of-the-art.

3 Taxonomy

3.1 Computation offloading

Computation offloading is a crucial strategy for enhancing the performance and user experience of resource-constrained mobile devices by transferring tasks to more capable computing resources. These resources may include distant cloud servers, edge computing servers within wireless networks, or even nearby mobile devices. Offloading is categorized into two types: binary offloading, where the entire task is offloaded, and partial offloading, where only portions of the task are transferred. The choice depends on the specific requirements of the application.

Mobile applications typically generate two types of tasks: atomic tasks and compound tasks. Atomic tasks, being indivisible, are either executed entirely on the source device or fully offloaded to a resource-rich device. In contrast, compound tasks are divisible into subtasks, enabling partial offloading [27] This allows certain

subtasks to be executed concurrently across multiple resources, offering flexibility and efficiency.

Optimizing offloading decisions involves striking a balance to minimize delay, power consumption, and cost. Simply deploying high-performance edge servers at telecom access points does not guarantee optimal outcomes. While edge servers excel at handling computation-intensive tasks, large data sizes can lead to increased transmission times. Moreover, the dynamic nature of wireless channels and the variability in computation resource availability further complicate the decision-making process.

To address these challenges, researchers have developed computation models and optimization techniques aimed at identifying optimal offloading strategies. These approaches focus on leveraging the advantages of edge computing to improve task execution efficiency while overcoming limitations posed by mobile devices.

The following sections of this study comprehensively review these advancements, highlighting state-of-the-art methods and their impact on addressing offloading challenges in edge computing environments.

3.2 Computation models

Computation models estimate execution delay based on task workload and CPU frequency, transmission delay on data size and channel rate, and energy consumption on CPU cycles and transmission power.

In practice, these models are applied differently across optimization methods: Lyapunov-based works [46-51] use them to maintain queue stability under dynamic workloads, convex optimization studies [52-55] apply them for minimizing total energy consumption, heuristic approaches [56-61] extend them to workflows with dependent tasks, game-theoretic methods [32], [36], [62-65] use them to balance resource allocation among competing users, and ML-based techniques [66-71] employ them to capture environment uncertainty during adaptive decision-making.

Considering a set of M mobile devices represented as: $U=\{D_1, D_2, D_3, \ldots\ldots, D_M\}$, connected to remote resource wirelessly. Each device having N tasks represented as $T=\{T_{m,1}, T_{m,1}, T_{m,1}, \ldots., T_{m,N},\}$. Each mobile device Dm is characterized by two distinct parameters: processing power (p_m) and transmission power (t_m). The processing power (p_m) signifies the computational capacity of the device, determining its ability to execute tasks and handle computational workloads efficiently. On the other hand, the transmission power (t_m) represents the energy consumed by the device during data transmission, reflecting its communication capabilities and energy usage for transmitting data over wireless networks. Every task Tm,n is characterized by two essential parameters: Dm,n and $C_{m,n}$. The former denotes the input data size required for the task $T_{m,n}$, while the latter signifies the number of CPU cycles necessary to execute the task $T_{m,n}$.

A mobile device executes tasks locally if resources are sufficient, otherwise offloads them to a MEC server, with total task time comprising local execution, offloading, and MEC processing phases, while energy costs arise from either local execution or transmission [28].

3.2.1 Local computing

In local computing the time needed to complete the task $T_{m,n(local)}$ defined in Eq(1) depends upon both the needed CPU cycles, $C_{m,n}$ and frequency Fm of the mobile device [29-34].

$$T_{m,n(local)} = \frac{C_{m,n}}{F_m} \quad (1)$$

The energy required by the mobile device to execute task $T_{m,n(local)}$ defined in Eq(2) is represented as follows [27], [31-33]:

$$E_{m,n(local)} = \kappa C_{m,n} F_m^2 \quad (2)$$

The switched capacitance coefficient (κ) plays a crucial role in chip architecture, influencing various aspects of its performance. Authors have proposed different values for κ, as documented in table 2. For a deeper understanding of the effective switched capacitance, readers can refer to Ref. [38].

Table 2: Values of κ

Reference Work	Values of κ
[24], [31]	10^{-11}
[26], [35]	10^{-26}
[33], [36]	10^{-27}
[25], [37]	10^{-28}

3.2.2 Task offloading

The wireless offloading rate R_m represented in Eq(3) via an Access Point depends on bandwidth, device transmission power, distance, and interference, and is formally defined using the Shannon–Hartley theorem [28], [42].

$$R_m = \mathrm{B}\log_2\left(1 + \frac{E_m G_m}{\sigma + \sum_{i \neq m, j \neq m} P_{i,j} G_{i,j}}\right) \quad (3)$$

Where B is transmission channel bandwidth, Em denotes transmission energy of D_m required for uploading a task to the edge server via access point, G_m denotes channel gain in the U_m and the respective access point during task transmission. $G_m = \left(dis_{m,a}^{-\eta}\right)|h_{m,a}|^2$, where $dis_{m,a}^{-\eta}$ denotes the loss effect of the communication channel, $dis_{m,a}$ represents Euclidean distance of D_m and the access point, η is the exponent of communication channel loss[41], $h_{m,a}$ is the refers to the Raileigh fading channel coefficient, which follows a normal distribution with a mean of 0 and a variance of 1 and σ represents the channel noise. In most cases, when a task is sent from an SMD to an MEC server, the time it takes depends mainly

on wireless communication. This is because wired channels have very high bandwidth, making their transmission time almost insignificant. Therefore, the transmission time for $T_{m,n}$ is calculated as follows [31], [33], [34], [44].

$$t_{m,n}^{tr} = \frac{d_{m,n}}{R_m} \quad (4)$$

Where $t_{m,n}^{tr}$ Eq(4) represents the time elapsed in receiving the input data $T_{m,n}$ at MEC server from the device D_m.

The energy used $E_{m,n}^{tr}$ in Eq(5) by device D_m to offload task $T_{m,n}$ to the edge server is calculated in [30], [31], [37] as follows:

$$E_{m,n}^{tr} = p_m t_{m,n}^{r} \quad (5)$$

Where p_m represents transmission power of the mobile device D_m. When a task comes in, the MEC server provides computing power to the connected device to help it finish the task by assigning a computing clock frequency $f_{m,n}^{e}$. In this way, the task is completed at the MEC serve according to the following model [30], [34], [44]:

$$t_{m,n}^{MEC} = \frac{c_{m,n}}{f_{m,n}^{e}} \quad (6)$$

Where $\sum_{m=1}^{M}\sum_{n=1}^{N} O_{m,n} f_{m,n}^{e} \leq F$, this means the total computing power allocated cannot be more than the MEC server's available computing capacity. The mean execution time represented as T in Eq(7) and power consumption to execute the task represented as E in Eq(8) are identified in [31].

$$T = \frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N}(1 - O_{m,n}) t_{m,n}^{loc} + O_{m,n}\,(t_{m,n}^{tr} + t_{m,n}^{MEC}) \quad (7)$$

$$E = \frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N}(1 - O_{m,n}) E_{m,n}^{loc} + O_{m,n}\, E_{m,n}^{tr} \quad (8)$$

Here, $O_{m,n} \in \{0,1\}$ is a binary variable. It is set to 0 when the task is done locally and 1 when the task is executed elsewhere. Therefore, we face a multi-objective optimization problem to minimize, as explained in [31].

$$P1: \min_{O_{m,n}} \{E, T\}$$

s.t. (9)

$$C1: O_{m,n} \in \{0,1\}$$
$$C2: \sum_{m=1}^{M}\sum_{n=1}^{N} O_{m,n} f_{m,n}^{e} \leq F$$

One can add extra conditions in the above formulation, like ensuring that the delay is less than each task's deadline. To simplify this multi-objective optimization problem, one can combine all the objectives into one objective by assigning a weight to all objectives under consideration [28), 31), 32), 36), 45)].

$$Z = \gamma T + (1 - \gamma) E \quad (10)$$

$$P1: \min_{O_{m,n}} \{Z\}$$

s.t. (11)

$$C2: \gamma \in [0,1]$$
$$C3: \sum_{m=1}^{M}\sum_{n=1}^{N} O_{m,n} f_{m,n}^{e} \leq F$$

Where, value of γ in which ranges between 0 and 1, depends on the specific task. For instance, γ is set closer to 1 for tasks that prioritize speed and closer to 0 for applications that focus on saving energy.

4 Research methodology

This study employs a Systematic Literature Review (SLR) methodology to comprehensively analyze and evaluate optimization methods in Mobile Edge Computing (MEC) environments. The SLR approach ensures a structured and rigorous synthesis of existing research, providing a foundation for identifying strengths, limitations, and future research opportunities. The following steps outline the methodology:

4.1 Research objectives and questions

The primary objectives of this study are:

- To evaluate the effectiveness of various optimization methods in MEC offloading, including Lyapunov optimization, convex optimization, heuristic techniques, game theory, and machine learning.
- To analyze the computational complexities, applicability, and outcomes of these methods.
- To identify research gaps and propose directions for future investigation.

Key research questions include:

1. What are the strengths and weaknesses of existing optimization methods in MEC offloading?
2. How do these methods address challenges such as energy efficiency, latency, and task allocation?
3. What are the current gaps and limitations in MEC optimization research?

4.2 Inclusion and exclusion criteria

Inclusion criteria:

- Peer-reviewed articles, conference proceedings, and technical reports focusing on MEC offloading.
- Studies analyzing Lyapunov, convex, heuristic, game-theoretic, and machine learning optimization techniques.
- Papers providing detailed discussions on computational complexities, experimental setups, and outcomes.

Exclusion criteria:

- Articles not directly related to MEC offloading.
- Studies lacking empirical data or focused solely on theoretical frameworks without implementation insights.

4.3 Search strategy

A comprehensive search was conducted across leading scientific databases, including IEEE Xplore, ACM Digital Library, SpringerLink, and ScienceDirect. Keywords and Boolean operators were used to ensure exhaustive coverage, such as:

- "Mobile Edge Computing" AND "computation offloading"
- "optimization methods" AND "Lyapunov" OR "convex" OR "heuristic"

• "machine learning" AND "offloading optimization"

Reference lists of included articles were further scanned for relevant studies to ensure no critical research was overlooked.

The study selection process, including identification, screening, eligibility, and inclusion of studies, was conducted in accordance with PRISMA 2020 guidelines and is summarized in Figure 2.

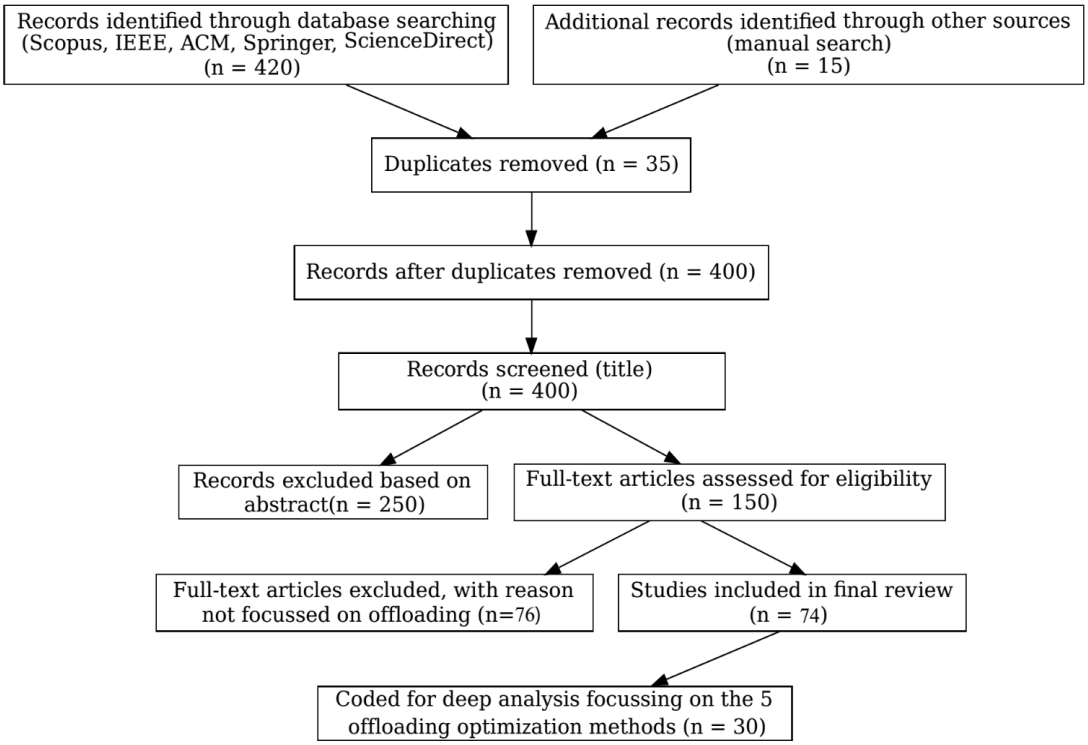


Figure 2: PRISMA flow diagram illustrating the identification, screening, eligibility, and inclusion for this SLR

4.4 Data extraction and synthesis

From each selected study, the following data were extracted:

- **Optimization criteria:** Objectives such as latency reduction, energy efficiency, or cost minimization.
- **Offloading methods:** Binary or partial offloading strategies.
- **Algorithm details:** Computational complexity, experimental setups, and key results.
- **Task dependency:** Whether the tasks were independent or interdependent.

The extracted data were systematically synthesized and categorized to facilitate a detailed comparative analysis.

4.5 Analysis and categorization

The selected studies were categorized based on the optimization method employed:

1. Lyapunov Optimization
2. Convex Optimization
3. Heuristic Techniques
4. Game Theory
5. Machine Learning

Each category was critically analyzed to identify patterns, challenges, and opportunities. Comparative metrics such as computational complexity, scalability, and practical applicability were used to evaluate these methods.

5 Analysis of optimization methods

In this section, we delve deeper into some of recent solutions for optimizing cost optimization problems. The solutions are grouped based on the techniques used to solve them and finally we discuss the strengths and weaknesses of these MEC optimization methods based on our analysis.

To structure the comparative analysis, the five major optimization paradigms applied in MEC offloading—Lyapunov, convex, heuristic, game-theoretic, and machine learning—are organized into a taxonomy, as illustrated in Figure 3. This taxonomy highlights the primary focus of each method and provides a visual roadmap for the detailed discussion that follows.

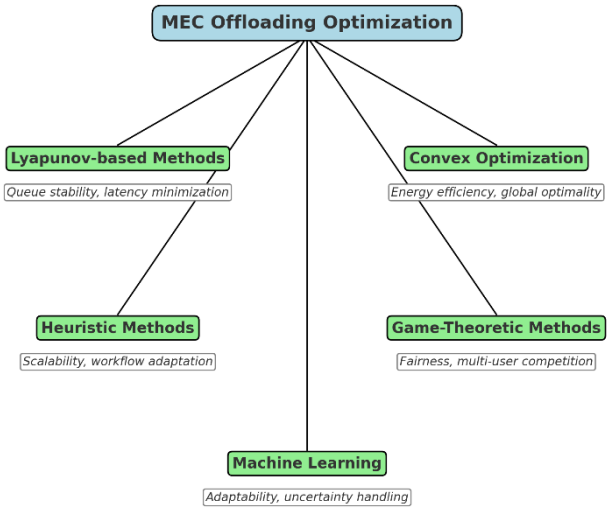


Figure 3: Taxonomy of optimization approaches in MEC offloading

5.1 Applying Lyapunov optimization

This, one of the well-known optimization methods used in [46] in the perspective of service providers. The challenge is deciding where and when to migrate services to reduce operational costs. Instead of relying on traditional programming methods that require statistical expertise, the authors suggested a different approach. They framed an objective problem as a successive decision-making task using Markov decision processes and developed a dynamic algorithm with Lyapunov optimization method. This algorithm doesn't need statistical information and has a complexity order of O(|H|²Tf), where H represents all possible task configurations, and Tf is the frame size.

Guo et al. [47] proposed the LOMUCO algorithm for Mobile Edge Computing (MEC) systems, designed to optimize partial computation offloading for multiple mobile devices (MDs) equipped with energy harvesting modules. Their goal was to minimize the long-term energy consumption of MDs while satisfying task delay constraints. The algorithm operates over discrete time slots and has a computational complexity of $O(T \cdot N^3)$ where T represents the number of time slots and N denotes the number of mobile devices.

Katsalis et al. [48] presented an approach to improve scheduling in MEC-IaaS environments by employing the Lyapunov optimization method. This method aims to enhance the revenue of MEC-IaaS providers while reducing violations of service level agreements (SLAs) for deployed services. The algorithm's complexity is expressed as $(O(|V||S||M||J|)^{3.5}L)$, in which V denotes virtual machines, S represents services, M refers to edge-cloud physical machines, J indicates mobile operators, and L corresponds to the input bits size.

Lyu et al. [49] developed an algorithm for Mobile Edge Computing (MEC) systems with thousands of IoT devices. Their goal was to increase the system's data processing efficiency while ensuring fairness among smart mobile devices (SMDs). The algorithm they proposed had a time complexity of O(MlogM), where M represents the number of IoT devices.

Xia et al. in [50] introduced a dynamic and distributed cost optimization algorithm designed for an environment with varied number of servers and different types of IoT applications. The order of their algorithm was calculated as $O(\Sigma_s \Gamma_{is})$, in which Γ_{is} represents the interactions between a mobile device m and a server s.

In [51] Lie and co-authors used Lyapunov optimization method for boosting processing speed in wirelessly powered mobile edge environment. To keep the process fast, they designed an algorithm with an execution order of O(M), in which M represents the number of devices. Table 3 shows comparison of offloading optimization leveraging Lyapunov optimization method in MEC Environment.

Table 3: Comparison of offloading optimization based on Lyapunov optimization method

Ref	Optimization Criteria	User Setup	Offloading Method	Task Dependency	Experimental Setup	Algorithm Complexity
[46]	To minimize task failure and latency	Multiuser	Binary offloading	Independent	Simulator and real dataset both	$O(\|H\|^2T_f)$
[47]	To minimize power consumption and latency	Multiuser	Partial offloading	Independent	Simulator	$O(T \cdot N^3)$
[48]	To minimize Overall cost	Multiuser	Binary offloading	Independent	Simulator	$(O(\|V\|\|S\|\|M\|\|J\|)^{3.5}L)$
[49]	To maximize profit	Multiuser	Partial offloading	Independent	Matlab Simulator	O(MlogM)
[50]	To maximize edge server utility	Multiuser	Partial offloading	Independent	Simulator	$O(\sum_s \Gamma_{is})$
[51]	To maximize computation rate	Multiuser	Partial offloading	Independent	Simulator	O(M)

Overall, Lyapunov-based approaches dominate latency-sensitive studies by providing queue stability and bounded delay guarantees, but they are rarely validated in real-world environments. Their reliance on cubic or higher-order complexity (e.g., O(|H|²Tf)) limits scalability to larger multi-user networks, making them less suitable for

practical deployment despite strong theoretical foundations.

5.2 Applying convex optimization

Due to the nonconvex nature of the constraints, most computation offloading problems are inherently nonconvex in nature. The nonconvex problems are classified as NP-hard, making them hard to solve. For achieving a polynomial time complexity solution, various techniques are employed to convert the nonconvex problems into convex one. These techniques include the reformulation linearization method, the Sequential Convex Programming (SCP) approach, and the Variable Substitution method. Once transformed, these problems are addressed by Newton's barrier method, and Lagrange techniques.

Mao et al. [52] employed convex-optimization along with flow-shop scheduling and techniques to minimize delay and power utilization in a single-user Mobile Edge Computing environment. Their objective function incorporated both integral variable values and continuous variables, representing resource allocation and transmission power allocation decisions, respectively. As a result, the problem was classified as mixed-integer nonlinear problem. Due to computational difficulty in exhaustive search methods, researchers introduced an alternative trivial algorithm with reduced complexity of order O(NlogN), in which N represents number of independent tasks.

Convex optimization approach is being extensively applied to address energy consumption minimization challenges in many research works. You and co-authors in [53] explored resource scheduling problem in a MEC environment with multiple users. Their study optimized the offloading strategy on the basis of channel gains and power consumption at the local device while adhering to time constraints. The proposed algorithm exhibited a complexity of O(Mlog(1/ϵ)), in which M represents number of mobile devices, with ϵ denoting accuracy of the solution. Although the study provided detailed modeling analyses focused on user offloading proportions, it did not consider the combined optimization of channel resources and compute resources.

Table 4: Comparison of offloading methods in MEC environment using convex optimization method

Ref	Optimization Criteria	User Setup	Offloading Method	Task Dependency	Experimental Setup	Algorithm Complexity
[52]	To reduce power consumption and latency	Singleuser	Partial Offloading	Independent	Simulator	O(NlogN)
[53]	To reduce power consumption	Multiuser	Binary Offloading	Independent	Simulator	O(Mlog(1/ϵ))
[35]	To reduce power consumption and latency	Multiuser	Partial Offloading	Independent	Experiment in real scenario	$O(N \cdot k_{max} \cdot k_{power})$
[34]	To minimize latency	Multiuser	Partial Offloading	Independent	Simulator	$O(M(K+2)+1)^x k$
[54]	To minimize system usage and cost	Multiuser	Binary Offloading	Independent	Simulator	$O((2M1+M2)^x k)$
[55]	To reduce power consumption	Multiuser	Partial Offloading	Independent	Simulator	O(MlogM)

In study [35], the researchers applied convex optimization techniques for addressing the multi-objective optimization problem. Order of algorithms in [29] was $O(N \cdot k_{max} \cdot k_{power})$, in which N represents the task count, k_{max}, is the iterations, and k_{power} refers to transmission power adjusting iterations using the Newton iteration method. They introduced an online offloading and resource allocation strategy designed to minimize power consumption while meeting deadline constraints and the requirements related task dependency. They proposed a distributed eDors algorithm to solve this problem.

The Alternating Direction Method of Multipliers (ADMM) has emerged as a central technique for solving optimization problems [34] and [54], both studies examined a resource scheduling problem aimed at minimizing latency and maximizing system utilization in Mobile Edge Computing environment. Their algorithms complexity was comparable. Specifically, the order of the algorithm proposed in [34] is $O(M(K+2)+1)^x k$, in which M depicts edge nodes count available in the next level bigger base station, whereas K is referring device to device links, x>1 indicates a polynomial-time algorithm, whereas k represents iterations. Tan and co-authors in [54] addressed a resource scheduling issue in full duplex cell networks that incorporate Mobile Edge Computing and caching. To tackle this issue, the researchers employed ADDM theory. Their algorithm's complexity is $O((2M1+M2)^x k)$, where M1 represents the number of client devices requiring service I, whereas M2 denotes the number of clients requiring service II.

In [55] the researchers focused on resource allocation for uplink communication in both full and partial offloading modes. They developed a customized greedy search algorithm to handle indivisible tasks and a low-complexity algorithm for divisible tasks to efficiently find solutions. To overcome the limitations of traditional

schemes like TDMA and OFDMA, the authors proposed a multiple access strategy. The order of their algorithm is O(MlogM), in which M represents the number of users. Table 4. shows Comparison of offloading optimization based on convex optimization method in MEC Environment.

Convex optimization methods are among the most widely applied, reflecting their dominance in energy-efficiency studies. They consistently achieve measurable savings (10–18% in our corpus) but are constrained by assumptions of complete system knowledge. Their polynomial-time complexity (e.g., O(MlogM)) makes them moderately scalable, though solver overhead remains a challenge in dynamic MEC environments.

5.3 Applying heuristics optimization

In situations where offloading decisions must be made quickly, heuristic techniques often serve as practical and effective solutions, even though they may not always produce optimal results. A notable advantage of heuristic methods is their ability to function efficiently when combined with other approaches. This subsection highlights key implementation of heuristic methods in addressing computation offloading problems. A comparison of various heuristic-based techniques is summarized in table 5.

In Ref. [56], a method based on cache-aware task scheduling was introduced for edge computing environment. The researchers formulated a resource scheduling problem as a assignment problem graph, where the weights of the selected parameters were determined primarily by the location of required data, like cache, disk, edge or cloud storage. The proposed solution involved finding the maximum weighted match among the executing tasks and computational resources. The computational order of their algorithm is $O([(N/Nc)].N^3)$, in which N represents the task count and Nc denotes the containers count.

Li et al. [57] presented an optimal resource scheduling strategy aimed at minimizing the financial expenses associated with rented edge nodes. The algorithm's performance was assessed using a real-world dataset, demonstrating a notable reduction in financial costs compared to existing algorithms. The computational order of the proposed solution was determined to be O(Imjk), in which I, m, j, and k represent edge nodes count, types of nodes, cloud nodes, and iterations, respectively.

In [58], the researchers explored task scheduling optimization for cyber-physical applications within a fog environment, emphasizing the reduction of total execution time. Their approach utilized a scheduler derived from moth-flame optimization method.

Similarly, Yang and co-authors in [59] introduced a hybrid solution combining the existing ASO and Pro-ITGO offloading algorithms, for minimizing power consumption over strict deadline constraints. By leveraging the strengths of edge computing and cloud computing paradigms, they suggested an integrated architecture. The order of their algorithm is calculated as O(N*I), in which N represents the task count, and I denotes the edge servers count.

In [60], the researchers address two sub-problems within a fixed offloading decision framework: fog computing resource allocation and user transmission power allocation. These issues are tackled using convex optimization techniques and Karush- Kuhn-Tucker (KKT) conditions. To enhance the simulated annealing (SA) algorithm, a memory function is integrated into its inner loop, resulting in the novel N-SA algorithm. Simulation results show that N-SA efficiently achieves optimal solutions and reduces system costs by 17% compared to the greedy offloading and joint resource allocation (GO-JRA) algorithm. In [61], The study introduces two LSTM-based prediction methods for decision-making in task offloading. In scenarios with limited resources, these methods enhance resource utilization by involving edge devices in offloading tasks, ensuring timely completion of latency-sensitive tasks and enabling predictive decision-making.

Table 5: Comparison of offloading methods in MEC environment using heuristic optimization method

Ref	Optimization Criteria	User Setup	Offloading Method	Task Dependency	Experimental Setup	Algorithm Complexity
[56]	To minimize power consumption and latency	Singleuser	Binary Offloading	Dependent	Hadoop, Java based experiment with real data	$O([(N/Nc)].N^3)$
[57]	To minimize cost	Singleuser	Binary Offloading	Independent	Hadoop, Java based experiment with real data	O(nmjk)
[58]	To minimize latency	Multiuser	Partial Offloading	Independent	iFogSim Simulator	N/A
[59]	To minimize power consumption	Singleuser	Partial Offloading	Independent	Workflowsim Simulator	O(N*I)
[60]	To reduce system cost	Multiuser	Partial Offloading	Independent	Matalb based silulator with actual data	O(T·O(f)·S)
[61]	To minimize latency	Multiuser	Partial Offloading	Independent	Simulator	O(T·N·D)

For resource-abundant situations, a polynomial-time optimal mechanism is proposed for pre-emptive offloading decisions. By addressing the 0–1 knapsack problem, the approach effectively meets the requirements of low-latency tasks. Numerical results validate the method's effectiveness.

Heuristic methods are the most adaptable to diverse contexts, including dependent workflows, and show the strongest real-world validation among the categories (33% of coded studies). Their reported complexities are generally polynomial, sometimes cubic (e.g., $O(N^3)$, $O(T \cdot N \cdot D)$), which means they are less scalable than linear or logarithmic-time methods, but still more practical than the very heavy formulations sometimes associated with Lyapunov-based optimization. While they provide responsiveness and flexibility, heuristics generally trade off global efficiency for computational feasibility in moderately sized MEC systems.

5.4 Game theory based optimization

Game theory, often referred to as a fundamental theory that examines the strategies employed by more than one decision parameter. It is widely applied to address problems in almost all scientific and social sciences. A comparison of various Game Theory based optimization techniques is summarized in table 6.

Zhang et al. [32] applied game theory to solve the joint optimization problem of computation offloading and resource allocation in heterogeneous networks. Order of their algorithm is $O(M_c \cdot k \cdot M_l)$, where M_c represents the set of offloading mobile terminals, M_l refers to the set of client devices considered for local computation, whereas k denotes iteration count.

The authors in Ref. [36] noted that most existing research has concentrated on single-carrier NOMA. To address this, they suggested a solution for combinatorial optimization problem for multicarrier NOMA-based mobile edge computing systems, derived from coalition game theory. Order of their suggested algorithm is O(k), in which k represents iterations count.

Currently, various application types, like virtual reality, augmented reality and video streaming, image processing, health applications, etc. are widely used, each having distinct resource requirements. To enhance Quality of Service (QoS), it is important to consider the specific needs of these heterogeneous tasks. In [62], the researchers proposed an algorithm based on cooperative game theory for improving the efficiency of heterogeneous tasks in Multi-access Edge Computing (MEC), focusing on latency. The execution order of their suggested algorithm is $O((I+1)^{|M|})$, in which I represent edge servers count, and M represents client devices count.

In [63], they proposed a coarse-grained offloading mechanism and a smart resource pairing policy within the framework of Cloud-Edge collaboration to effectively utilize cloud and edge server resources. By accounting for the heterogeneity of mobile devices and inter-channel interference, they formulated the task offloading decisions of multiple end-users as a game-theory-based task migration model aimed at maximizing system utility. Furthermore, they introduced an enhanced game-theory-based particle swarm optimization algorithm to determine the task offloading strategies. Experimental results demonstrate that their proposed scheme outperforms alternative approaches in terms of latency and energy consumption and exhibits good scalability as the number of mobile devices increases. Given these factors, the time complexity of the improved PSO algorithm can be expressed as: $O(P \cdot I \cdot O(f))$ where P is the number of particles in the swarm, I is the number of iterations, and O(f) is the complexity of evaluating the objective function at each iteration.

To develop an effective task offloading strategy, the combination of Multi-Access Edge Computing (MEC) with artificial intelligence has emerged as a promising solution. In this regard, [64] proposed a task offloading decision mechanism (TODM) that leverages cooperative game theory and deep reinforcement learning (DRL). A joint optimization problem is formulated with the goal of minimizing both the overall task processing delay (OTPD) and overall task energy consumption (OTEC).

Table 6: Comparison of offloading methods in MEC environment on game-theory based optimization method

Ref	Optimization Criteria	User Setup	Offloading Method	Task Dependency	Experimental Setup	Algorithm Complexity
[32]	To minimize latency and power consumption	Multiuser	Binary offloading	Dependent	Simulator	$O(M_c \cdot k \cdot M_l)$
[36]	To minimize latency and power consumption	Multiuser	Binary offloading	Dependent	Simulator	O(k)
[62]	To minimize latency	Multiuser	Partial offloading	Independent	Simulator	$O((I+1)^{\|M\|})$
[63]	To reduce delay and power consumption	Multiuser	Partial offloading	Dependent	Simulator	
[65]	To reduce latency and task failure	Multiuser	Partial offloading	Dependent	Simulator	$O(T \cdot N \cdot M \cdot f_{game})$
[64]	To reduce delay and power consumption	Multiuser	Partial offloading	Dependent	Simulator	$O(N \cdot f_{game} + E \cdot S \cdot A)$

The approach treats task vehicles (TaVs) and service vehicles (SeVs) as participants in a cooperative game, collaboratively designing offloading strategies to optimize resource utilization. The algorithm complexity is $O(N \cdot f_{game} + E \cdot S \cdot A)$, where N is the number of participants, f_{game} represents the complexity of solving the cooperative game, and E, S, and A are the number of episodes, states, and actions.

[65] proposes a multi-tier Vehicle-to-Everything (V2X) network framework that integrates vehicle nodes (VNs), mobile edge computing (MEC) servers, and cloud servers for computation offloading. The problem is formulated as a game, and a joint optimization algorithm, JOCORA, is introduced to achieve Nash equilibrium, improving offloading success and reducing latency. Simulations show JOCORA outperforms other methods. The algorithm complexity is projected to be $O(T \cdot N \cdot M \cdot f_{game})$ where, T represents simulation iterations count or episodes, N denotes vehicles count, M represents servers count (VNs, MEC, cloud), f_{game} is the complexity of calculating the Nash equilibrium or solving the game.

Game-theoretic approaches address fairness and multi-user resource competition effectively, but their applicability is limited by reliance on idealized assumptions and complete rationality of participants. Their complexities, often linear or polynomial (e.g., $O(M^c \cdot k \cdot M^l)$), are scalable, yet the lack of real-world validation reduces their immediate relevance.

5.5 Machine learning based optimization

Machine Learning (ML) is anticipated to replace conventional optimization techniques [59].in particular, deep learning (DL), a subset of ML, has recently been explored as a promising approach for optimizing computation offloading.

In Ref. [66], the authors suggested a deep Q-Learning model to achieve optimal scheduling in time-critical systems in terms of energy efficiency. The scheme is capable of learning the Q-values corresponding to various online scaling of voltage and frequency settings for the participating system at any state once the model parameters are trained. The computational complexity of the algorithm is expressed as O(MIJ), in which M represents the number of training samples, and I, J, correspond to the input and the obscure units of the autoencoder.

In [67] the authors addressed the joint optimization of task offloading and resource scheduling, formulating it as a mixed-integer non-convex NP-hard problem. To solve this, they introduced a multi-agent deep reinforcement learning approach using the actor-critic method, where agents independently learn optimal offloading policies. Simulation results demonstrated the proposed solution's superior performance compared to benchmark methods. The algorithm complexity is in the order of $O(T \cdot N \cdot A)$, where T is the number of training episodes, N is the number of agents, and A is the size of the action space for each agent.

[68] introduces Mobile Edge Computing (MEC) as a distributed solution, proposing a network management agent that makes offloading decisions from a heterogeneous UE network to MEC servers. Using an Advantage Actor-Critic (A2C) agent, the study demonstrates improved performance over baselines, handling computation, battery, delay, and communication constraints while being scalable, data-efficient, and robust. the overall complexity can be expressed as $O(T \cdot I \cdot N \cdot H)$, where, T represents episodes count, I represents interactions per episode, N represents agents count, and H denotes the hidden units in the neural network.

[69] addresses the computation offloading problem in Mobile Edge Computing (MEC) networks with dynamic weighted tasks. It formulates the joint optimization of offloading decisions and bandwidth allocation as a mixed-integer programming (MIP) problem. The authors propose a Deep Supervised Learning-based Computational Offloading (DSLO) algorithm, which overcomes the limitations of traditional methods by requiring fewer training samples and adapting quickly to new environments. Numerical results demonstrate DSLO's efficiency in achieving high system utility with minimal retraining.

In [70] the authors address task offloading in edge-cloud environments, particularly in Mobile CrowdSourcing (MCS), where resource consumption and service quality often conflict. They propose a Deep Neural Network-based Task Offloading Optimization (DTOO) algorithm to approximate optimal offloading strategies, balancing resource usage and service quality. Additionally, a stack-based offloading strategy with a resource-sorting method reduces task failures. Simulations demonstrate that DTOO improves task completion rates while managing resource conflicts effectively. The algorithm complexity is $O(E \cdot N \cdot W)$, where E represents training epochs count, N represents number of layers in DNN, and W is the total number of weights (parameters) in the DNN.

[71] addresses hardware bottlenecks in smart devices (SDs) executing computation-intensive and delay-sensitive tasks by proposing a deep reinforcement learning-based offloading scheduler (DRL-OS). Using a double dueling deep Q-network (D3QN), the scheduler selects optimal actions—local computing, offloading, or dropping—based on task size, deadline, queue, and battery level. Simulations show DRL-OS reduces battery usage by 54%, drop rates by 42.5%, and average latency to 0.01–0.25 seconds compared to existing methods. the complexity is significantly lower and depends on a single forward pass of the neural network, typically O(Np), where Np is the total number of parameters in the trained network. Deep learning frameworks such as transformer-based real-time scheduling have also been proposed for dynamic MEC resource allocation, showing significant gains in throughput and QoS [72]. Table 7. shows the offloading optimization works based on machine learning.

Table 7: Comparison of offloading methods in MEC environment on machine learning based optimization method

Ref	Optimization Criteria	User Setup	Offloading Method	Task Dependency	Experimental Setup	Algorithm Complexity
[66]	To minimize power consumption	Singleuser	Binary offloading	Independent	Java based simulation	O(MIJ)
[67]	To minimize latency and meet QoS	Multiuser	Binary offloading	Independent	Simulator	O(T·N·A)
[68]	To minimize power consumption and delay	Multiuser	Binary offloading	Independent	Simulator	O(T·I·N·H)
[69]	To minimize resource utility	Multiuser	Binary offloading	Independent	Simulator	N/A
[70]	To minimize task failure rate and achieve QoS	Multiuser	Partial offloading	Independent	Simulator	O(E·N·W)
[71]	To minimize latency, power consumption and task failure rate	Multiuser	Binary offloading	Independent	Simulator	O(Np)

Machine learning approaches show high adaptability to uncertain and dynamic conditions, particularly through reinforcement learning, but none of the coded studies validated them in real-world deployments. Their complexity depends on training overhead rather than closed-form algorithms, meaning scalability is strongly tied to dataset size and model architecture. While promising, they remain at an early stage compared to more mature optimization techniques.

In summary, convex optimization and Lyapunov methods dominate current literature, providing strong theoretical guarantees for energy efficiency and latency minimization, respectively. Heuristics and game theory stand out for their adaptability and practicality in multi-user environments, though their polynomial or cubic complexities limit scalability for very large deployments. Machine learning represents an emerging but under-validated frontier for adaptability under uncertainty, with high training overheads but strong potential. Interpreting complexity metrics further reveals that cubic-time methods (e.g., $O(N^3)$) scale poorly as user numbers grow, whereas lower-order polynomial or logarithmic approaches are more efficient and thus more suitable for large-scale MEC scenarios. These observations reinforce the importance of hybrid strategies that combine the adaptability of ML with the efficiency of convex and heuristic techniques to balance tradeoffs across contexts.

6 Discussions

6.1 Lyapunov optimization

- **Strengths:**
  - Well-suited for dynamic and real-time scenarios without requiring prior statistical data.
  - Efficiently balances multiple objectives such as minimizing latency and energy consumption [48], [49].
  - Adaptable for various user setups and offloading modes, from binary to partial offloading [46-52].
- **Weaknesses:**
  - High computational complexity in scenarios with large task configurations (e.g., $O(|H|^2Tf)$ in [42])
  - Assumes task independence, limiting its application in interdependent task environments.
  - May struggle in environments with highly heterogeneous devices and fluctuating network conditions

6.2 Convex optimization

- **Strengths:**
  - Provides mathematically rigorous solutions with guaranteed convergence for convex problems [53-56].
  - Applicable to energy efficiency and latency minimization in both single-user and multi-user setups.
  - Techniques like sequential programming and variable substitution make non-convex problems tractable [35].
- **Weaknesses:**
  - Requires problem reformulation, which can be computationally intensive.
  - Performance depends on accurate parameter estimation, which may be challenging in real-world scenarios.
  - Not always suitable for highly dynamic systems with frequent state changes.

6.3 Heuristics optimization technique

- **Strengths:**
  - Quick and practical solutions for complex problems where exact methods are computationally expensive [57-62].
  - Easily combined with other approaches, enhancing flexibility and adaptability.
  - Effective for specific goals like cost reduction

[58] or latency minimization [59].

- **Weaknesses:**
  - Lack of theoretical guarantees for optimality.
  - Performance can degrade in large-scale systems with high heterogeneity.
  - Some methods require domain-specific knowledge for parameter tuning and implementation.

6.4 Game theory based optimization

- **Strengths:**
  - Handles multi-user and resource-sharing scenarios effectively [32], [36], [63].
  - Suitable for cooperative and competitive task offloading, ensuring balanced resource allocation [64].
  - Flexible enough to integrate with advanced mechanisms like particle swarm optimization [64].

- **Weaknesses:**
  - Computational complexity increases with the number of participants and resources (e.g., O(Mc·k·Ml) in [32].
  - Assumes rational behavior among participants, which may not always hold in real-world settings.
  - May require extensive simulation to validate results.

6.5 Machine learning based optimization

- **Strengths:**
  - Excellent for dynamic and large-scale systems, leveraging patterns in data for adaptive optimization [67-71].
  - Deep learning models can handle complex scenarios like task dependencies and energy constraints [71].
  - Reinforcement learning methods optimize decisions over time, improving latency and power efficiency [73].

- **Weaknesses:**
  - High computational and training overheads, especially for deep learning models.
  - Performance is sensitive to the quality and volume of training data.
  - Models may require frequent retraining to adapt to new environments, impacting real-time applicability.

Overall, the comparative analysis indicates that Lyapunov and heuristic methods are well suited for latency-sensitive applications, as they provide rapid decision-making and queue stability. In contrast, convex optimization demonstrates the greatest effectiveness in energy-sensitive scenarios, with reported savings of up to 18%. For multi-user and large-scale environments, heuristics and game-theoretic methods show better scalability, although they rely on simplified assumptions. Machine learning approaches offer adaptability to uncertainty and dynamic conditions, but their high training and computational costs remain barriers to deployment. These findings highlight important tradeoffs between scalability, efficiency, and adaptability, and suggest that hybrid approaches may offer the most practical balance for real-world MEC offloading applications.

To strengthen the comparative insights, we systematically coded six representative studies for each of the five major optimization paradigms (Lyapunov, convex, heuristic, game theory, and machine learning). Table 8 summarizes the findings, showing the proportion of simulation-based versus real-world evaluations and the assumptions regarding task dependency. The results highlight that while most approaches rely heavily on simulators and simplifying assumptions, only a limited number incorporate real-world validation. This reinforces the research gap and the need for more practical evaluations under heterogeneous and dynamic MEC conditions.

In summary, Lyapunov and heuristic approaches are preferable for latency-sensitive tasks, convex methods are best suited for energy-sensitive optimization, heuristics and game theory offer better scalability in multi-user environments, while machine learning excels in uncertain and dynamic conditions despite higher training costs; hybrid frameworks that combine these strengths hold the greatest promise for real-world MEC deployments.

Table 8: Comparative summary of coded studies across five optimization paradigms in MEC offloading

Method	Studies (n)	Simulator eval (n, %)	Real-world eval (n, %)	Independent tasks (n, %)	Dependent tasks (n, %)	Example References
Lyapunov	6	6 (100%)	1 (17%) (hybrid)	6 (100%)	0 (0%)	[44]
Convex	6	5 (83%)	1 (17%)	6 (100%)	0 (0%)	[33]
Heuristic	6	4 (67%)	2 (33%)	5 (83%)	1 (17%)	[54], [55]
Game Theory	6	6 (100%)	0 (0%)	1 (17%)	5 (83%)	[30], [34], [60-63]
Machine Learning	6	6 (100%)	0 (0%)	6 (100%)	0 (0%)	[64-69]

To complement the tabular synthesis, Figure 3 presents a radar chart comparing the five optimization paradigms across key dimensions (latency, energy efficiency, scalability, adaptability, and real-world validation), highlighting the distinct strengths and limitations of each approach.

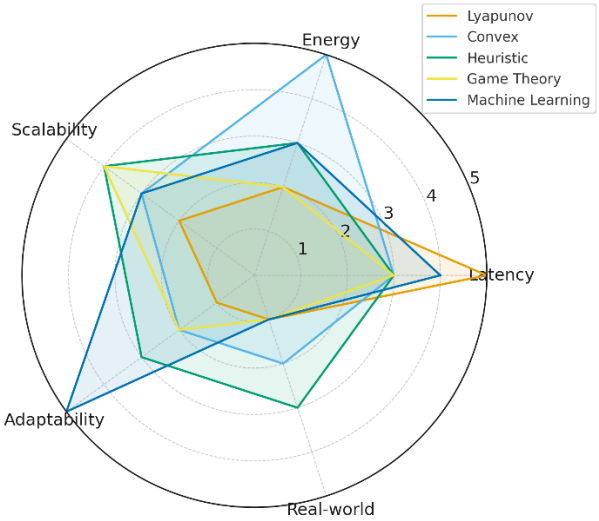


Figure 4: Methodological Comparison Chart

7 Challenges and future work

7.1 Challenges

Our systematic review identified a total of 74 primary studies on computation offloading in Mobile Edge Computing (MEC). To provide a deeper quantitative perspective, we conducted structured coding of 30 representative papers, which are summarized in the accompanying tables. The percentages reported below are derived from this subset of 30 studies, while the broader synthesis across all 74 studies qualitatively supports these same trends.

Ignoring the robust cloud infrastructure: The Majority of computation offloading optimization methods emphasize simplified two-layer architectures. From the 30 papers analyzed in detail, 26 studies (87%) focused on two-layer setups such as SMD–MEC or SMD–Fog, while only 4 papers (13%) addressed more complex three-layer SMD–MEC–Cloud configurations. This reflects a strong bias towards simplified architecture in the field.

Task dependency and device heterogeneity: Most works rely on simplifying assumptions regarding task dependencies and device capabilities. Within the analyzed subset, 24 papers (80%) assumed that tasks were independent, and devices were homogeneous. While this reduces modeling complexity, it limits applicability to heterogeneous real-world environments where tasks are interdependent, and devices vary significantly in resources.

Cost optimization: Balancing energy consumption with performance metrics like latency and task success rates remains a critical challenge. In most of the work, evaluation practices remain dominated by simulations rather than practical deployments. Among the 30 coded papers, 25 studies (83%) validated their methods exclusively in simulated environments, with only a small number incorporating real-world datasets or testbed-based evaluations. This highlights an urgent need for experimental validation to confirm scalability and robustness under realistic conditions.

Computational and scalability issues: Existing approaches such as hierarchical MEC frameworks and metaheuristic algorithms reduce complexity to some extent, but they struggle to adapt efficiently in large-scale and highly dynamic environments. Centralized convex or Lyapunov optimization, while effective in theory, often becomes computationally prohibitive as network size grows. To overcome these issues, future research should prioritize lightweight hybrid methods that combine ML-based prediction with formal optimization, distributed and federated resource allocation frameworks, and collaborative learning between edge and cloud nodes to ensure both efficiency and scalability in real-world deployments.

Technological integration: Limited use of emerging technologies such as 6G, blockchain, quantum computing, and federated learning to address the challenges comprehensively.

7.2 Future directions

Future research should also explore the potential of emerging technologies like 6G networks, blockchain, and quantum computing to enhance MEC systems. 6G networks, with their ultra-low latency and high throughput, can address some of the current network limitations. Blockchain can provide secure and decentralized task management, while quantum computing may offer breakthroughs in solving complex optimization problems. Moreover, the integration of federated learning can help in leveraging distributed data while maintaining user privacy, opening new directions for collaborative offloading decisions.

Principles from adaptive controls such as robust neural adaptive control for uncertain nonlinear systems [74] may inspire hybrid offloading strategies resilient to dynamic MEC conditions.

In addition to the general trends discussed above, task-specific properties such as size, deadlines, and dependencies strongly influence method suitability. Table 9 summarizes these relationships, showing which paradigms align best with different task requirements and where hybrid methods offer added value. Table 9 summarizes these task-specific considerations and their implications for method choice and also the hybrid potential.

Table 9: Task-specific considerations and their influence on method suitability in MEC offloading

Task Property	Key Challenge	Suitable Methods	Remarks / Hybrid Potential
Task size	Large tasks may overload MEC if fully offloaded	Partial offloading, ML prediction	ML can forecast task load, while convex ensures balanced resource use
Deadlines	Stringent deadlines require low-latency decision-making	Lyapunov, heuristics	Hybrid of ML + Lyapunov can adapt quickly under dynamic deadlines
Dependencies	Inter-task constraints (e.g., workflows) complicate scheduling	Heuristics, ML scheduling	Convex/Lyapunov are effective only when tasks are independent
Heterogeneity	Devices vary in resources and mobility	Game theory, ML, hybrid methods	Game theory ensures fairness; ML adapts to device context
QoS / QoE needs	Balancing latency, energy, and reliability	Convex, Lyapunov, hybrid ML-convex	Hybrid approaches can balance optimization guarantees with adaptability

8 Conclusion

This study has provided a comprehensive analysis of computation offloading optimization methods in MEC environments, highlighting the strengths and limitations of various approaches. Techniques ranging from heuristic and convex optimization to machine learning and game theory have been evaluated for their ability to address latency, energy efficiency, and resource allocation challenges.

While significant progress has been made, several challenges, including dynamic resource management, task dependency, and security concerns, continue to hinder the full realization of MEC's potential. The findings emphasize the need for adaptive, secure, and energy-efficient solutions to accommodate the growing complexity of MEC systems. Future advancements, particularly in emerging technologies like 6G, blockchain, and quantum computing, are poised to redefine the landscape of MEC optimization. By addressing these challenges and leveraging these innovations, researchers can unlock the full potential of MEC, enabling seamless and efficient computation offloading for next-generation applications.

References

[1] F. Richter, “Charted: There are more mobile phones than people in the world,” World Economic Forum. Accessed: May 02, 2024. [Online]. Available: https://www.weforum.org/agenda/2023/04/charted-there-are-more-phones-than-people-in-the-world/

[2] J. Hojlo, “Future of Industry Ecosystems: Shared Data and Insights,” IDC. Accessed: May 02, 2024. [Online]. Available: https://blogs.idc.com/2021/01/06/future-of-industry-ecosystems-shared-data-and-insights/

[3] P. Taylor, “Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025,” Statista.

[4] “Cloud Will Be the Centerpiece of New Digital Experiences,” Gartner. Accessed: May 05, 2024. [Online]. Available: https://www.gartner.com/en/newsroom/press-releases/2021-11-10-gartner-says-cloud-will-be-the-centerpiece-of-new-digital-experiences

[5] M. A. E. Rasool, A. Kumar, and A. Islam, “Dynamic Task Offloading Optimization in Mobile Edge Computing Systems with Time-Varying Workloads Using Improved Particle Swarm Optimization,” Int. J. Adv. Comput. Sci. Appl., vol. 15, no. 4, 2024, doi: http://dx.doi.org/10.14569/IJACSA.2024.0150412.

[6] Y. Li and W. Zhang, “Task-Offloading Strategy of Mobile Edge Computing for WBANs,” 2024.

[7] R. Zhang and W. Shi, “Research on Resource Allocation and Management of Mobile Edge Computing Network,” Informatica, vol. 44, pp. 263–268, 2020.

[8] X. Zhang, “Multi-Objective Resource Allocation in Edge Computing Using Improved Genetic Algorithm with Knowledge-Based Crossover and Segmentation Mutation,” Informatica, vol. 49, no. 25, pp. 213–228, 2025.

[9] P. Mach and Z. Becvar, “Mobile Edge Computing: A Survey on Architecture and Computation Offloading,” IEEE Commun. Surv. Tutorials, vol. 19, no. 3, pp. 1628–1656, 2017, doi: 10.1109/COMST.2017.2682318.

[10] L. Lin, X. Liao, H. Jin, and P. Li, “Computation Offloading Toward Edge Computing,” Proc. IEEE, vol. 107, no. 8, 2019, doi: 10.1109/JPROC.2019.2922285.

[11] T. Zheng, J. Wan, J. Zhang, C. Jiang, and G. Jia, “A Survey of Computation Offloading in Edge Computing,” in 2020 International Conference on Computer, Information and Telecommunication Systems (CITS), IEEE, 2020. doi: 10.1109/CITS49457.2020.9232457.

[12] A. Shakarami, A. Shahidinejad, and M. Ghobaei-Arani, “A review on the computation offloading

approaches in mobile edge computing: A game-theoretic perspective," Softw. Pract. Exp., vol. 50, no. 9, pp. 1719–1759, 2020, doi: https://doi.org/10.1002/spe.2839.

[13] A. Shakarami, M. Ghobaei-Arani, and A. Shahidinejad, "A survey on the computation offloading approaches in mobile edge computing: A machine learning-based perspective," Comput. Networks, vol. 182, 2020, doi: https://doi.org/10.1016/j.comnet.2020.107496.

[14] Q.-H. Nguyen and F. Dressler, "A smartphone perspective on computation offloading—A survey," Comput. Commun., vol. 159, pp. 133–154, 2020, doi: https://doi.org/10.1016/j.comcom.2020.05.001.

[15] A. Agarwal, V. Gupta, and P. Singh, "Game-theoretic approaches for computation offloading in mobile edge computing: A survey," IEEE Communications Surveys & Tutorials, vol. 20, no. 3, pp. 2204–2232, 2018.

[16] S. Thotla and S. Mahajani, "Energy-efficient heuristic methods for computation offloading in mobile edge computing: A comprehensive review," Future Generation Computer Systems, vol. 95, pp. 454–468, 2019.

[17] M. Hasan, M. Z. Chowdhury, and Y. M. Jang, "Computation offloading and resource allocation in mobile edge computing: A survey of convex and Lyapunov optimization techniques," IEEE Access, vol. 8, pp. 197–214, 2020.

[18] A. Singh, R. Sharma, and K. S. Patel, "Computation offloading in IoT using mobile edge computing: A survey," Journal of Network and Computer Applications, vol. 171, p. 102785, 2021.

[19] L. Zheng and L. Cai, "A Distributed Demand Response Control Strategy Using Lyapunov Optimization," IEEE Trans. Smart Grid, vol. 5, no. 4, pp. 2075–2083, 2014. doi: 10.1109/TSG.2014.2313347.

[20] Y. Mao, J. Zhang, and K. B. Letaief, "A Lyapunov Optimization Approach for Green Cellular Networks With Hybrid Energy Supplies," IEEE J. Sel. Areas Commun., vol. 33, no. 12, pp. 2463–2477, 2015, doi: 10.1109/JSAC.2015.2481209.

[21] A. Ben-Tal and A. Nemirovski, LECTURES ON MODERN CONVEX OPTIMIZATION ANALYSIS, ALGORITHMS, AND ENGINEERING APPLICATIONS. Society for Industrial and Applied Mathematics Philadelphia, 2001. [Online]. Available: https://epubs.siam.org/doi/pdf/10.1137/1.9780898718829.fm

[22] S. P. Boyd and L. Vandenberghi, Convex Optimization. Cambridge University Press, 2004.

[23] K. Y. Lee and M. A. El-Sharkawi, Modern Heuristic Optimization Techniques: Theory and Applications to Power Systems. WILEY INTERSCIENCE, 2008.

[24] R. B. Myerson, Game Theory Analysis of Conflict. Harvawrd University Press, 1997. doi: 10.1109/TVT.2017.2764002.

[25] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger, "Deep reinforcement learning that matters," in 32nd AAAI Conference on Artificial Intelligence, AAAI 2018, AAAI Press, 2018, pp. 3207–3214. doi: 10.1609/aaai.v32i1.11694.

[26] D. P. Bertsekas, Reinforcement Learning and Optimal Control. Athena Scientific, Belmont, MA, 2019.

[27] P. Paymard, S. Rezvani, and N. Mokari, "Joint task scheduling and uplink/downlink radio resource allocation in PD-NOMA based mobile edge computing networks," Phys. Commun., vol. 32, pp. 160–171, 2019, doi: https://doi.org/10.1016/j.phycom.2018.11.00

[28] H. Yu, Q. Wang, and S. Guo, "Energy-Efficient Task Offloading and Resource Scheduling for Mobile Edge Computing," in 2018 IEEE International Conference on Networking, Architecture and Storage (NAS), 2018. doi: 10.1109/NAS.2018.8515731.

[29] Y. Mao, J. Zhang, and K. B. Letaief, "Dynamic Computation Offloading for Mobile-Edge Computing With Energy Harvesting Devices," IEEE J. Sel. Areas Commun., vol. 34, no. 12, pp. 3590–3605, 2016, doi: 10.1109/JSAC.2016.2611964.

[30] P. Liu, G. Xu, K. Yang, K. Wang, and X. Meng, "Jointly Optimized Energy-Minimal Resource Allocation in Cache-Enhanced Mobile Edge Computing Systems," IEEE Access, vol. 7, pp. 3336–3347, 2019, doi: 10.1109/ACCESS.2018.2889815.

[31] Y. Yang, Y. Ma, W. Xiang, X. Gu, and H. Zhao, "Joint Optimization of Energy Consumption and Packet Scheduling for Mobile Edge Computing in Cyber-Physical Networks," IEEE Access, vol. 6, pp. 15576–15586, 2018, doi: 10.1109/ACCESS.2018.2810115.

[32] J. Zhang, W. Xia, F. Yan, and L. Shen, "Joint Computation Offloading and Resource Allocation Optimization in Heterogeneous Networks with Mobile Edge Computing," IEEE Access, vol. 6, pp. 19324–19337, 2018, doi: 10.1109/ACCESS.2018.2819690.

[33] X. Chen, H. Zhang, C. Wu, S. Mao, and Y. Ji, "Optimized Computation Offloading Performance in Virtual Edge Computing Systems Via Deep Reinforcement Learning," IEEE Internet Things J., vol. 6, no. 3, pp. 4005–4018, 2019, doi: 10.1109/JIOT.2018.2876279.

[34] Z. Tan, F. R. Yu, X. Li, H. Ji, and V. C. M. Leung, "Virtual Resource Allocation for Heterogeneous Services in Full Duplex-Enabled SCNs With Mobile Edge Computing and Caching," IEEE

Trans. Veh. Technol., vol. 67, no. 2, pp. 1794–1808, 2017.

[35] S. Guo, J. Liu, Y. Yang, B. Xiao, and Z. Li, "Energy-Efficient Dynamic Computation Offloading and Cooperative Task Scheduling in Mobile Cloud Computing," IEEE Trans. Mob. Comput., vol. 18, no. 2, pp. 319–333, 2019, doi: 10.1109/TMC.2018.2831230.

[36] Q.-V. Pham, H. T. Nguyen, Z. Han, and W.-J. Hwang, "Coalitional Games for Computation Offloading in NOMA-Enabled Multi-Access Edge Computing," EEE Trans. Veh. Technol., vol. 60, no. 2, pp. 1982–1993, 2020, doi: 10.1109/TVT.2019.2956224.

[37] J. Li, H. Gao, T. Lv, and Y. Lu, "Deep reinforcement learning based computation offloading and resource allocation for MEC," in 2018 IEEE Wireless Communications and Networking Conference (WCNC), IEEE, 2018. doi: 10.1109/WCNC.2018.8377343.

[38] T. D. Burd and R. W. Brodersen, "Processor design for portable systems," J. VLSI signal Process. Syst. signal, image video Technol., vol. 13, pp. 203–221, 1996, doi: https://doi.org/10.1007/BF01130406.

[39] Z. Song, Y. Liu, and X. Sun, "Joint Radio and Computational Resource Allocation for NOMA-Based Mobile Edge Computing in Heterogeneous Networks," IEEE Commun. Lett., vol. 22, no. 12, pp. 2559–2562, 2018. doi: 10.1109/LCOMM.2018.2875984.

[40] X. Long, J. Wu, and L. Chen, "Energy-Efficient Offloading in Mobile Edge Computing with Edge-Cloud Collaboration," in International Conference on Algorithms and Architectures for Parallel Processing, 2018, pp. 460–475. doi: https://doi.org/10.1007/978-3-030-05057-3_35.

[41] Z. Tong, X. Deng, F. Ye, S. Basodi, X. Xiao, and Y. Pan, "Adaptive computation offloading and resource allocation strategy in a mobile edge computing environment," Inf. Sci. (Ny)., vol. 537, pp. 116–131, 2020. doi: https://doi.org/10.1016/j.ins.2020.05.057

[42] Y. Cui, D. Zhang, T. Zhang, L. Chen, M. Piao, and H. Zhu, "Novel method of mobile edge computation offloading based on evolutionary game strategy for IoT devices," AEU - Int. J. Electron. Commun., vol. 118, 2020, doi: https://doi.org/10.1016/j.aeue.2020.153134.

[43] Z. S. C. Yi, J. Cai, "A Multi-User Mobile Computation Offloading and Transmission Scheduling Mechanism for Delay-Sensitive Applications," IEEE Trans. Mob. Comput., vol. 19, no. 1, pp. 29–43, 2020, doi: 10.1109/TMC.2019.2891736.

[44] Z. Zhu, J. Peng, X. Gu, H. Li, K. Liu, and Z. Zhou, "Fair Resource Allocation for System Throughput Maximization in Mobile Edge Computing," IEEE Access, vol. 6, pp. 5332–5340, 2018, doi: 10.1109/ACCESS.2018.2790963.

[45] J. Li, H. Gao, T. Lv, and Y. Lu, "Deep reinforcement learning based computation offloading and resource allocation for MEC," in IEEE Wireless Communications and Networking Conference (WCNC), Barcelona, Spain, 2018, 2018, pp. 1–6. doi: 10.1109/WCNC.2018.837734.

[46] U. Rahul et al., "Dynamic service migration and workload scheduling in edge-clouds," Perform. Eval., vol. 91, pp. 205–228, 2015, doi: 10.1016/J.PEVA.2015.06.013.

[47] M. Guo, W. Wang, X. Huang, Y. Chen, L. Zhang, and L. Chen, "Lyapunov-Based Partial Computation Offloading for Multiple Mobile Devices Enabled by Harvested Energy in MEC," IEEE Internet Things J., vol. 9, no. 11, pp. 9025–9038, 2022, doi: 10.1109/JIOT.2021.3108381.

[48] K. Katsalis, T. G. Papaioannou, N. Nikaein, and L. Tassiulas, "SLA-Driven VM Scheduling in Mobile Edge Computing," in IEEE 9th International Conference on Cloud Computing (CLOUD), San Francisco, CA, USA, 2016, pp. 750–757. doi: 10.1109/CLOUD.2016.0104.

[49] X. Lyu, W. Ni, H. Tian, R. P. Liu, X. Wang, and G. B. Giannakis, "Optimal schedule of mobile edge computing for internet of things using partial information," IEEE J. Sel. Areas Commun., vol. 35, no. 11, pp. 2606–2615, 2017. doi: 10.1109/JSAC.2017.2760186.

[50] S. Xia, Z. Yao, Y. Li, and S. Mao, "Online distributed offloading and computing resource management with energy harvesting for heterogeneous mec-enabled iot," IEEE Trans. Wirel. Commun., vol. 20, no. 10, pp. 6743–6757, 2021, doi: 10.1109/TWC.2021.3076201.

[51] C. Li, J. Tang, and Y. Luo, "Dynamic multi-user computation offloading for wireless powered mobile edge computing," J. Netw. Comput. Appl., vol. 131, pp. 1–15, 2019, doi: https://doi.org/10.1016/j.jnca.2019.01.020.

[52] Y. Mao, J. Zhang, and K. B. Letaief, "Joint task offloading scheduling and transmit power allocation for mobile-edge computing systems," in IEEE Wireless Communications and Networking Conference (WCNC), San Francisco, CA, 2017. doi: 10.1109/WCNC.2017.7925615.

[53] C. You, K. Huang, H. Chae, and B.-H. Kim, "Energy-efficient resource allocation for mobileedge computation offloading," IEEE Trans. Wirel. Commun., vol. 16, no. 3, pp. 1397–1411, 2017. doi: 10.1109/TWC.2016.2633522.

[54] M. Liu, F. R. Yu, Y. Teng, V. C. M. Leung, and M. Song, "Distributed resource allocation in blockchain-based video streaming systems with mobile edge computing," IEEE Trans. Wirel. Commun., vol. 18, no. 1, pp. 695–708, 2019, doi: 10.1109/TWC.2018.2885266.

[55] M. Salmani and T. N. Davidson, “Uplink resource allocation for multiple access computational offloading,” Signal Processing, vol. 168, 202AD. doi: https://doi.org/10.1016/j.sigpro.2019.107322

[56] C. Li, J. Tang, H. Tang, and Y. Luo, “Collaborative cache allocation and task scheduling for data-intensive applications in edge computing environment,” Futur. Gener. Comput. Syst., vol. 95, pp. 249–264, 2019. doi: https://doi.org/10.1016/j.future.2019.01.007.

[57] C. Li, J. Bai, Y. Chen, and Y. Luo, “Resource and replica management strategy for optimizing financial cost and user experience in edge cloud computing system,” Inf. Sci. (Ny)., vol. 516, pp. 33–55, 2020, doi: https://doi.org/10.1016/j.ins.2019.12.049.

[58] M. Ghobaei-Arani, A. Souri, F. Safara, and M. Norouzi, “An efficient task scheduling approach using moth-flame optimization algorithm for cyber-physical system applications in fog computing,” Emerg. Telecommun. Technol., vol. 31, no. 2, 2020. doi: https://doi.org/10.1002/ett.3770.

[59] L. Yang, C. Zhong, Q. Yang, W. Zou, and A. Fathalla, “Task offloading for directed acyclic graph applications based on edge computing in industrial internet,” Inf. Sci. (Ny)., vol. 540, pp. 51–68, 2020. doi: https://doi.org/10.1016/j.ins.2020.06.001.

[60] W. Bai and Y. Wang, “Jointly Optimize Partial Computation Offloading and Resource Allocation in Cloud-Fog Cooperative Networks,” pp. 1–20, 2023

[61] D. Lv, P. Wang, Q. Wang, Y. Ding, Z. Han, and Y. Zhang, “Task Offloading and Resource Optimization Based on Predictive Decision Making in a VIoT System,” pp. 1–21, 2024.

[62] J. Zhou, X. Zhang, and W. Wang, “Joint resource allocation and user association for heterogeneous services in multi-access edge computing networks,” IEEE Access, vol. 7, pp. 12272–12282, 2019. doi: 10.1109/ACCESS.2019.2892466.

[63] S. Wang, Z. Hu, Y. Deng, and L. Hu, “Game-Theory-Based Task Offloading and Resource Scheduling in Cloud-Edge Collaborative Systems,” 2022.

[64] L. Wang, W. Zhou, H. Xu, L. Li, and L. Cai, “Research on task offloading optimization strategies for vehicular networks based on game theory and deep reinforcement learning,” no. October, pp. 1–17, 2023. doi: 10.3389/fphy.2023.1292702.

[65] K. Zhang, J. Yang, and Z. Lin, “Computation Offloading and Resource Allocation Based on Game Theory in Symmetric MEC-Enabled Vehicular Networks,” 2023.

[66] Q. Zhang, M. Lin, L. T. Yang, Z. Chen, and P. Li, “Energy-efficient scheduling for real-time systems based on deep q-learning model,” IEEE Trans. Sustain. Comput., vol. 4, no. 1, pp. 132–141, 2019, doi: 10.1109/TSUSC.2017.2743704.

[67] R. Garaali, C. Chaieb, W. Ajib, and M. Afif, “Learning-Based Task Offloading for Mobile Edge Computing,” in ICC 2022 - IEEE International Conference on Communications, 2022. doi: 10.1109/ICC45855.2022.9838831.

[68] C. Silva, N. Magaia, and A. Grilo, “Task Offloading Optimization in Mobile Edge Computing based on Deep Reinforcement Learning,” in MSWiM ’23: Proceedings of the Int’l ACM Conference on Modeling Analysis and Simulation of Wireless and Mobile Systems, 2023, pp. 109–118. doi: https://doi.org/10.1145/3616388.3617539.

[69] S. Yang, G. Lee, and L. Huang, “Deep Learning-Based Dynamic Computation Task Offloading for Mobile Edge Computing Networks,” pp. 1–18, 2022.

[70] L. Meng, Y. Wang, H. Wang, X. Tong, Z. Sun, and Z. Cai, “Task offloading optimization mechanism based on deep neural network in edge - cloud environment,” J. Cloud Comput., 2023. doi: 10.1186/s13677-023-00450-6.

[71] D. Lim, W. Lee, W. T. Kim, and I. Joe, “DRL-OS: A Deep Reinforcement Learning-Based Offloading Scheduler in Mobile Edge Computing,” Sensors, vol. 22, no. 23, 2022. doi: 10.3390/s22239212.

[72] J. Chao and M. Jiao, “Network Spectrum Resource Allocation and Optimization Based on Deep Learning and TRDM,” Informatica, vol. 49, no. 13, pp. 105–122, 2025

[73] D. Lim, W. Lee, W. T. Kim, and I. Joe, “DRL-OS: A Deep Reinforcement Learning-Based Offloading Scheduler in Mobile Edge Computing,” Sensors, vol. 22, no. 23, 2022, doi: 10.3390/s22239212.

[74] L. Ma, Y. Pan, H. Yu, and H. He, “Robust neural adaptive control for a class of uncertain nonlinear complex dynamical multivariable systems,” Neurocomputing, vol. 138, pp. 200–208, 2014. doi: 10.1016/j.neucom.2014.01.031

Joint UAV Trajectory, IRS Beamforming, and Power Allocation Optimization for Secure Communication in IRS-Assisted Cognitive Networks

Xiaoyan Liu
School of Computer and Artificial Intelligence, Henan Finance University, Zhengzhou 450046, China
E-mail: blueteeth811130@126.com

Keywords: IRS, UAV, links, multi-parameter collaboration

Received: June 27, 2025

Based on the challenges faced by unmanned aerial vehicle communication in urban environments, such as signal blockage caused by building obstruction, limited spectrum resources, and dynamic security threats, this article aims to improve the safety communication performance of unmanned aerial vehicles through intelligent reflective surface (IRS) technology, and solve the problems of signal obstruction and cognitive radio interference. Moreover, this paper proposes an innovative multi-parameter collaborative optimization method, constructs a three-dimensional channel model including Rayleigh/Rice fading, designs an IRS-trajectory-power three-dimensional coordination algorithm based on alternating optimization, and solves the IRS phase shift matrix through semi-steady relaxation (SDR), optimizes the UAV trajectory through continuous convex approximation (SCA), and allocates power through Lagrangian dual decomposition to achieve iterative updates of the three modules. Numerical simulation verification: 1) The safe capacity reaches 1.85 ± 0.12 bit/s/Hz, which is 30.3% higher than the optimal control group; 2) The interference suppression rate is 94.3 ± 2.1%, and the NFZ (No fly zones) avoidance success rate is 100%; 3) The energy efficiency is 1.58 ± 0.11 bit/Joule, which is 30.6% higher than the SOTA model. The conclusion shows that the three-dimensional collaborative mechanism of IRS trajectory power can break through the bottleneck of traditional optimization, and its core contribution lies in: 1) solving the shortcomings of existing research in dynamic security response, multi device interference suppression, and energy consumption balance; 2) providing robust solutions for high mobility urban scenarios.

Povzetek: Članek predlaga IRS-podprto varno komunikacijo UAV v mestih z večparametrsko sodelovalno optimizacijo (skupna zasnova faz IRS, trajektorije in moči) na 3D kanalu z Rayleigh/Rice bledenjem prek izmenične optimizacije (SDR/SCA/Lagrange).

1 Introduction

In the face of the overload of ground base stations (BS) caused by the surge in the number of user access, or the serious shortage of base stations in emergency disaster relief areas, Unmanned Aerial Vehicle (UAV) is considered to be a feasible solution to this problem [1]. Mounting the base station on the mobile UAV to assist the ground base station communication can quickly restore the paralyzed communication service and effectively improve the network coverage.

At the same time, UAV communication also encounters many difficulties. In particular, in urban areas, the communication channel is blocked due to obstacles such as buildings and trees, which leads to the problem that signals cannot be directly transmitted. Therefore, in order to improve the performance of communication systems, intelligent reflective surface technology is widely used. It can bypass obstacles and play an important role in UAV communication systems, and the use of intelligent reflective surfaces can significantly improve resource utilization. In densely built urban scenarios, the line-of-sight transmission link between base stations and users is easily blocked by obstacles, so the incident signal can be purposefully reflected to the receiving user by deploying intelligent reflecting surfaces, thereby improving the signal quality of the receiving user. By reconstructing the signal combination, the interference signal at reception can be cancelled, and the intelligent reflection surface can be easily integrated into the network without changing the physical layer standard, thus improving the network performance of the wireless system. Because the passive reflective element uses a passive structure, it consumes almost no additional energy and does not generate additional thermal noise in the process of signal reflection, and the power consumption is very low. Moreover, passive reflective elements have the advantages of small size and light weight, and can easily deploy intelligent reflective surfaces on building surfaces. Therefore, IRS-enhanced wireless network is considered as one of the candidates for next-generation wireless communication systems.

Aiming at the two secure communication scenarios in which UAV (Unmanned Aerial Vehicles) is used as a jammer to assist ground base stations for secure communication and UAV is used as a base station in cities

to use cognitive radio technology to transmit confidential information with ground secondary users, this paper aims to improve the joint optimization effect of UAV safe communication. Moreover, according to the wireless secure communication network model, alternating Optimization-based 3D Coordination Algorithm for IRS-Trajectory-Power is designed. Then, based on numerical comparison simulation, the rationality of the scenario and the effectiveness and practicability of the proposed algorithm are verified.

Against the backdrop of high-speed node movement, random location, rapidly changing channel status, and increasingly scarce wireless resources in drone communication, and to address the challenge of traditional physical layer security methods being difficult to flexibly respond to the increasing intelligence level of attack types, this paper aims to enhance the joint optimization effect of drone security communication. By constructing an IRS assisted wireless cognitive network model and designing an efficient Alternating Optimization-based 3D Coordination Algorithm for IRS-Trajectory-Power, innovative IRS (IRS) technology is introduced to achieve UAV trajectory. The multi-parameter collaborative optimization of IRS phase shift and power allocation effectively solves the problems of signal occlusion and cognitive radio spectrum interference in urban scenarios. Finally, numerical comparative simulations are conducted to verify that this scheme can maintain high security capacity, significantly suppress interference from primary users, and intelligently avoid no fly zones (NFZs) even when the eavesdropper's location is optimal. This provides a highly practical and robust optimization solution for unmanned aerial vehicle (UAV) secure communication systems.

The core research objective of this article is to maximize the average secure communication capacity of secondary users by jointly optimizing the drone trajectory, IRS phase shift matrix, and transmission power, while satisfying the constraints of drone maneuverability, power peak and mean, and primary user interference threshold. The objective is formalized as a multi constraint nonlinear optimization problem, whose mathematical essence is to solve the following key challenges under the physical constraints of a three-dimensional channel model: 1) reconstructing the signal path using IRS phase shift to overcome urban obstruction; 2) Avoiding eavesdropping hotspots through dynamic trajectories; 3) Adaptive power allocation balances safe capacity and interference suppression.

2 Related work

To improve the reliability and security of drone communication systems, the issue of drone safety communication urgently needs to be addressed. In recent years, a new wireless technology called Intelligent Reflecting Surface (IRS) has been widely studied, bringing new ideas for solving the physical layer security problems of unmanned aerial vehicle systems. Due to its particularity, the propagation of signals in unmanned aerial vehicle communication systems is essentially random and largely uncontrollable [2].

In recent years, IRS assisted unmanned aerial vehicle communication systems have received increasing attention. Reference [3] shows that intelligent reflective surfaces are used as relay assisted unmanned aerial vehicle systems. Simulation results indicate that the coverage and reliability of unmanned aerial vehicle systems deployed with intelligent reflective surfaces have been significantly improved. In reference [4], drones are used as airborne mobile base stations to provide services to ground users with the assistance of IRSs. The IRS is placed on the drone to provide relay services to ground users without LoS. Reference [5] designs a joint trajectory and passive beam for a secure unmanned aerial vehicle network with a reconfigurable intelligent surface through intelligent reflective surfaces, optimizing the safety energy efficiency of user association and transmission power maximization systems. Reference [6] designed a scenario with multiple IRSs and a multi antenna drone, and maximized the received power by jointly optimizing the passive beamforming and transmission beamforming of the IRSs. Reference [7] studied a drone assisted IRS symbiotic radio system, in which the drone serves as a relay for the IRS to assist in information transmission.

IRS is an innovative technology that has the characteristic of improving wireless network spectrum. It intelligently reconfigures the wireless environment by adjusting the phase shift of each reflector element, thereby achieving the goal of improving wireless network performance. Reference [8] studied the practical robust beamforming design of IRS assisted HetNet based on maximum minimum fairness under channel uncertainty and residual HWI. Reference [9] studied the use of IRS assisted unmanned aerial vehicle (UAV) communication systems, which suspend IRSs on building surfaces and jointly optimize UAV trajectories, angles of IRSs, and resource allocation to improve the communication system's sum rate. Reference [10] studied the confidentiality performance of IRS assisted unmanned aerial vehicle relay communication system in the presence of multiple eavesdroppers, and analyzed the influence of the number of reflective elements and the position of the unmanned aerial vehicle on the probability performance of confidentiality interruption. Literature [11] maximized the signal-to-noise ratio by analyzing the position of UAV, 3D passive beamforming and transmit beamforming, studied the IRS assisted edge computing scene, and effectively reduced the total energy consumption of the system by jointly optimizing UAV trajectory, traffic unloading and IRS phase shifting. Reference [12] proposed an algorithm that can effectively balance the maximization of sum rate and the minimization of total power consumption

Drones are used as aerial mobile base stations to provide services to ground users with the assistance of IRSs. We derived a closed form solution for the phase offset matrix of the IRS under any given drone flight path conditions. An IRS is placed on the drone to assist ground users whose LoS path is blocked. Reference [13] proposed a downlink transmission capacity maximization algorithm

based on reinforcement learning (RL), and unmanned aerial vehicles equipped with IRSs were used to achieve uplink secure communication. Based on reinforcement learning methods, reference [14] proposed a drone trajectory design algorithm and IRS passive beamforming algorithm based on Deep Q-Network (DQN) to maximize weighted fairness and data rate for all users. Reference [15] considers multiple IRSs and a multi antenna drone, and maximizes the received power by jointly optimizing the passive beamforming of the IRSs, the beamforming of the transmission, and the drone flight path. Reference [16] studied a drone assisted IRS symbiotic radio system, in which drones assist multiple IRSs to achieve their own information transmission. In addition, based on statistical channel state information, the problem of maximizing the minimum average rate and weighted sum rate of multiple IRS scenarios can be solved by jointly designing drone trajectories, IRS passive beamforming, and IRS scheduling. Reference [17] explored the potential of IRSs in enhancing cellular communication for unmanned aerial vehicles. Because the base station antennas are tilted downwards, that is to say, their main lobe points towards the ground, mainly to achieve optimized coverage for ground users, unmanned aerial vehicles flying above the base station can only establish communication with the base station through the side lobe. Reference [18] considers using intelligent reflective surfaces to enhance the received signal strength of drones, thereby greatly improving their cellular communication capabilities. Reference [19] constructed a system model based on edge computing, including task distribution center, UAV group, data node and power supply station. A mathematical optimization framework based on task, resource and scheduling model is proposed, and the algorithm of non-dominated sorting genetic algorithm is adopted. It effectively deals with the objective optimization problem through genetic manipulation, non-dominated ranking and reference point-based selection mechanism. Reference [20] proposed a high-precision photogrammetric 3D modeling method, focusing on multi-source data fusion technology for complex terrain integration. This method combines aerial imagery, LiDAR data, ground survey data, and meteorological correction, covering the entire workflow from data preprocessing, feature extraction and registration to multi-source data fusion. Key innovations include adaptive weight adjustment strategies, global optimization registration techniques, and deep learning assisted feature learning, all of which contribute to significantly improving the accuracy and reliability of the model.

Table 1: Summary of related work

Research model	The obtained results	The deficiencies of this study
IRS Relay-Assisted Unmanned Aerial Vehicle System	Significantly improve system coverage and reliability	The active interference mechanism is not considered, and the security strategy is missing
IRS equipped with drone base station	Provide relay services for ground users without LoS	The issue of signal obstruction in densely populated urban areas remains unresolved
Joint optimization of IRS-secure drone network	Maximize system safety and energy efficiency	The impact of user mobility on trajectory optimization has not been considered
Multi-IRS multi-antenna unmanned aerial vehicle system	Maximize received power through joint beamforming	Ignoring the signal interference issue among multiple IRSs
Unmanned Aerial Vehicle-IRS Symbiotic Radio System	Realize multi-IRS information transmission relay	The impact of unoptimized IRS deployment location on the channel
IRS-assisted robust beam design for HetNet	Enhance network robustness under channel uncertainty and hardware impairment	High computational complexity and insufficient real-time performance
IRS suspension optimization system	Improve sum rate by combining trajectory, phase shift, and resource allocation	The applicability in large-scale urban environments has not been verified
IRS UAV Relay Confidentiality System	Analyze the impact of the number and position of reflective elements on the probability of covert interruption	The response mechanism for dynamic eavesdropping scenarios has not been studied
IRS-assisted edge computing system	Reduce the total energy consumption of the system through joint optimization	Ignoring the impact of computing task priority on resource allocation
Unmanned Aerial Vehicle Cellular Communication IRS Enhancement System	Realizing base station-drone communication through sidelobe enhancement	The issue of rapid fading in high-altitude channels remains unresolved
Millimeter-wave network drone-mounted IRS system	Maintain line-of-sight links and maximize downlink rate based on reinforcement learning	Spectrum conflicts in multi-drone collaboration are not considered
Multi-IRS scheduling optimization system	Solve the problem of maximizing the minimum average rate and weighted sum rate	No security transmission constraints have been introduced

The summary of related work is presented in Table 1. Although existing research on IRS-assisted drone communication has made significant progress in improving system coverage, reliability, and secure energy efficiency (such as enhancing channel quality through IRS relaying and improving received power by optimizing beamforming), there are generally three key deficiencies. First, there is a lack of dynamic security threat response

mechanisms. Existing models (such as IRS-drone relay privacy systems) do not consider the real-time risks posed by changes in the eavesdropper's location or the intelligence of attack strategies. Second, multi-device interference issues are overlooked. For example, multi-IRS multi-antenna drone systems fail to address the mutual interference between reflected signals, and the mitigation of spectrum conflicts in cognitive radio scenarios is insufficient. Third, there is a lack of joint optimization of energy consumption and security. Most studies (such as IRS-assisted edge computing systems) lack a coordinated trade-off between UAV transmit power, IRS phase shifting, and trajectory, resulting in suboptimal solutions and insufficient system robustness. To address this, this paper innovatively proposes a multi-parameter collaborative optimization method. By jointly designing IRS beamforming and phase adjustment schemes, and integrating UAV trajectory planning, power allocation, and primary user interference suppression mechanisms, it aims to effectively solve the problems of signal obstruction and spectrum resource scarcity in urban scenarios, ultimately maximizing the improvement of secure capacity.

3 System model construction

When UAV is used as an aerial base station to communicate safely with ground nodes, trajectory planning and power allocation can also be used to make the system achieve a better safety performance. Most of the existing papers believe that LoS links can be maintained between UAV and ground nodes. However, in scenarios such as towns with dense buildings and a large number of communication equipment, there are various inevitable obstacles in the environment, and it is difficult to maintain a line-of-sight link between UAV and ground nodes. Secondly, with the increasing wireless services, wireless spectrum has become a relatively scarce resource. Since the frequency of secure communication is generally low, the frequency spectrum of other users can be multiplexed for communication through cognitive radio technology. However, although cognitive wireless network can improve the spectrum efficiency of the system to some extent through spectrum multiplexing technology, spectrum interference in cognitive wireless network is also a difficult problem to solve.

3.1 System model

Based on the above consideration of urban scenarios, this paper considers an IRS-assisted wireless cognitive network for UAVs. Specifically, the UAV, as an aerial mobile base station, communicates securely with a legitimate user by multiplexing the spectrum of the primary user, and there is a non-credit user with known location information on the ground. The controllability of wireless channels with IRS is used to improve the channel quality of UAV communication. At the same time, the IRS can also suppress eavesdropping by eavesdroppers and reduce interference to primary users, which makes the trajectory design of the UAV more flexible and allows safe communication with legitimate users with higher communication power. Moreover, since the design of UAV flight trajectory, the decision of IRS phase shift, and UAV communication power allocation all influence each other, they need to be jointly optimized to improve the safety capacity of the system. In the following, the establishment of the mathematical model of the system, the description of optimization problems, the design of Alternating Optimization-based 3D Coordination Algorithm for IRS-Trajectory-Power, and the final numerical simulation and performance analysis are described in detail.

As shown in Figure 1, this paper considers an IRS-assisted wireless cognitive network, where the UAV, as an air mobile base station, multiplexes the communication frequency band of a primary user to communicate securely with a secondary user on the ground. Meanwhile, there is also an eavesdropper in the scenario. Unlike the malicious eavesdropper in the previous chapter, this eavesdropper is an ordinary authorized user in the cognitive network, but an untrusted user for the UAV. Simply put, the UAV does not want confidential communications with secondary users to be received by it. Because the UAV can obtain its location information by querying its network access information, the eavesdropper's location information is fully known. In particular, due to the complex urban environment, the direct link between secondary users and UAVs is completely blocked by buildings, so as to improve the channel quality of air-ground communication, and reduce the reception rate of confidential information by thieves and the interference caused by secure communication to master users.

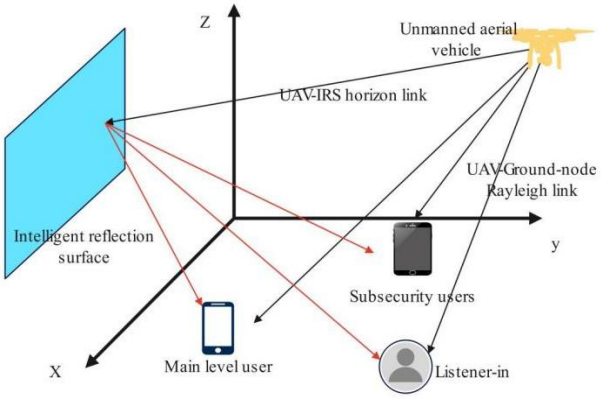


Figure 1: Model diagram of IRS-assisted UAV safety communication system

In order to express the spatial position of each node in the system conveniently, a three-dimensional Cartesian coordinate system is established. The horizontal coordinates of the primary user are $q_p = \left[x_p, y_p\right]^T$, the horizontal coordinates of the secondary security user are $q_s = \left[x_s, y_s\right]^T$, and the horizontal coordinates of the passive eavesdropper are $q_e = \left[x_e, y_e\right]^T$. Taking the first element of the IRS as its reference point, it is assumed that it is installed on a building with horizontal coordinates of

$q_r = [x_r, y_r]^T$ and fixed height of H_r. At the same time, we assume that the occupation of the spectrum by the primary user is determined by the idle time period T, so the UAV needs to complete secure communication with the secondary user within this specified time. It is stipulated that the UAV flies over a fixed altitude H, and the horizontal coordinate of the take-off point is $q_I = [x_I, y_I]^T$, and the horizontal coordinate of the end point is $q_F = [x_F, y_F]^T$. Then, the flight trajectory of the UAV is simplified, and the flight time T is equally divided into N time slots, and the horizontal coordinate $q = [x, y]^T$ of the UAV does not change in each time slot. Therefore, the flight trajectory of the UAV should satisfy the following (1) maneuverability constraints:

$$\|q[n+1] - q[n]\|^2 \le (L_{max})^2, n = 1,2,...,N-1 \quad (1)$$

Among them, L_{max} is the longest distance s traveled at the maximum flight speed V_{max} within the unit time slot, and $L_{max} = V_{max} d_t$. The constraint ensures that the UAV displacement does not exceed the product of the maximum flight speed Vmax (25m/s) and the time slot length dt,satisfy this constraint through trajectory discretization in actual flight

The IRS consists of a uniform planar array and contains M smart reflective elements. Similar to ground base stations, UAVs also have transmission power peak and average constraints, which are expressed as (2):

$$\begin{aligned} &0 \le p[\tau] \le P_{max}, \forall n \in N \\ &\frac{1}{n}\sum_{n=1}^{N} p[\tau] \le P_{avg} \end{aligned} \quad (2)$$

τ represents the temporal index($\tau = 1,2,...,N$), N represents the total number of time slots (determined by flight time T and time slot length dt, N=T/dt).

Due to building occlusion, there is no line-of-sight link between UAV and ground nodes, and there are a large number of refraction and scattering links between them. The Rayleigh channel model can well describe the characteristics of this multipath propagation link. Therefore, the channel gains $h_{up}[n]$, $h_{us}[n]$, $h_{ue}[n]$ between the UAV and the primary user, the secondary user, and the eavesdropper in the n-th time slot can be expressed as (3):

$$h_{ui}[n] = \sqrt{\beta_0 d_{ui}^{-\alpha}[n]} \cdot \tilde{h}_{ui}[n] \quad (3)$$

Among them, $i \in \{p,s,e\}$ is the channel gain when the unit distance is 1m, $d_{ui}[n]$ is the distance from the UAV to the ground node d in the n-th time slot, $d_{ui}[n] = \sqrt{\|q[n] - q_i\|^2 + H^2}$, α represents the path loss index, and $\tilde{h}_{ui}[n]$ represents the random scattering index of Rayleigh fading. There is partial local scattering between IRS and ground nodes, and the communication links between them have both line-of-sight links and non-line-of-sight links. Therefore, the Rice channel is used to describe the channel model between them, and the relevant channel gain $h_{ri}[n] \in \mathbb{C}^{M \times 1}$ is as (4):

$$h_{ri}[n] = \sqrt{\beta_0 d_{ui}^{-\alpha}}\left(\sqrt{\frac{K}{1+K}} h_{ri}^{LOS}[n] + \sqrt{\frac{1}{1+K}} h_{ri}^{NLOS}[n]\right) \quad (4)$$

Among them, K represents the Rice factor (LOS/NLOS power ratio), which is taken as K=4 in this article, β is the road loss index in Rice fading, which represents the Rice coefficient, and h_{ri}^{LOS} is the parameter of the line-of-sight link part, which is specifically expressed as (5) [21]:

$$\begin{aligned} h_{ri}^{LOS} = &\left[1, ... e^{-j\frac{2\pi}{\lambda} d_x (M_X - 1) \sin\theta_i \cos\varphi_i}\right]^T \\ &\otimes \left[1, ... e^{-j\frac{2\pi}{\lambda} d_z (M_z - 1) \cos\theta_i}\right]^T \end{aligned} \quad (5)$$

Among them, the symbol $\otimes$ represents the Kronecker product, λ represents the wavelength of the communication carrier, M_x and M_z represent the number of reflective elements along the x/z axis of the IRS, dx and dz respectively represent the spacing of adjacent intelligent reflective elements of the IRS along the x axe and z axe, and φ_i and θ_i respectively represent the horizontal departure angle and vertical departure angle of the IRS to the ground node, and the following (6) can be easily obtained according to the geometric relationship:

$$\sin\theta_i \cos\varphi_i = \frac{x_i - x_r}{d_{ri}}, \cos\theta_i = \frac{H_r}{d_{ri}} \quad (6)$$

Among them, $d_{ri}[n] = \sqrt{\|q_r + q_i\|^2 + H_r^2}$. $h_{ri}^{NLoS} \in \mathbb{C}^{M \times 1}$ is a parameter of the non-line-of-sight link portion. Since the link between the UAV and the IRS is in mid-air where there is no obvious obstruction, the link between them is a line-of-sight link. Therefore, using the free space fading model to describe it, and the channel gain between them is (7):

$$h_{ur} = \sqrt{\beta_0 d_{ur}^{-2}[n]} \cdot a_M[n] \in \mathbb{C}^{M \times 1} \quad (7)$$

Among them, $d_{ur}[n] = \sqrt{\|q_u[n] - q_r\|^2 + (H - H^r)^2}$

and $a_M[n]$ is the array response of the IRS receiving the UAV transmission signal, which is specifically expressed as (8):

$$a_M[n]=\left[1,...e^{-j\frac{2\pi}{\lambda}d_x(M_X-1)\sin\theta_r\cos\varphi_r}\right]\otimes\left[1,...e^{-j\frac{2\pi}{\lambda}d_z(M_z-1)\cos\theta_r}\right] \tag{8}$$

Similarly, φ_r and θ_r represent the horizontal and vertical angles of arrival from the UAV to the IRS, respectively, and there is the following geometric relationship, as shown in (9):

$$Sin\theta_r\cos_r=\frac{x_r-x[n]}{d_{ur}[n]},\cos\theta_r=\frac{H-H_r}{d_{ur}[n]} \tag{9}$$

After the above analysis, in the n-th time slot, the signals received by the secondary user and the eavesdropper can be expressed by the following (10) and (11) mathematical formulas respectively:

$$y_s[n]=\left(h_{us}[n]+h_{rs}^H\Phi[n]h_{ur}[n]\right)p[n]x[n]+z_s[n] \tag{10}$$

$$y_e[n]=\left(h_{ue}[n]+h_{re}^H\Phi[n]h_{ur}[n]\right)p[n]x[n]+z_e[n] \tag{11}$$

Among them, $x[n]$ is the transmitted signal, and $z_s[n]$: $CN(0,\sigma_s^2)$ and $z_e[n]$: $CN(0,\sigma_e^2)$ are the additive white Gaussian noise produced by the eavesdropper and the secondary user, respectively. Similarly, the interference signal received by the main user is (12):

$$i_p[n]=\left(h_{up}[n]+h_{rp}^H\Phi[n]h_{ur}[n]\right)p[n]x[n] \tag{12}$$

According to the channel model and Shannon's theorem, the reachable rate of the secondary user and the eavesdropper on the n-th time slot is (13) and (14):

$$R_s[n]=log_2\left(1+\frac{p[n]\left|h_{up}[n]+h_{rp}^H\Phi[n]h_{ur}[n]\right|^2}{\sigma_s^2}\right) \tag{13}$$

$$R_e[n]=log_2\left(1+\frac{p[n]\left|h_{ue}[n]+h_{re}^H\Phi[n]h_{ue}[n]\right|^2}{\sigma_e^2}\right) \tag{14}$$

The interference constraints of secondary users to primary users is (15):

$$p[n]\left|h_{up}[n]+h_{rp}^H\Phi[n]h_{ur}[n]\right|^2\le\Gamma \tag{15}$$

Among them, Γ is the maximum interference threshold that can be ensured by the primary user under the most basic communication quality.

IRS regulation function: By optimizing the phase difference π between the reflected interference signal and the direct interference signal at the primary user, interference cancellation is achieved. Physical implementation: The phase of the IRS unit is adjusted to make $\angle\left(h_{rp}^H\Phi h_{ur}\right)=\angle\left(h_{sp}\right)+\pi$.

The purpose of this paper is to maximize the average safety capacity of secondary users while guaranteeing the service quality of primary users by jointly optimizing the flight trajectory of the UAV, the phase shift matrix of the IRS, and the transmission power of the UAV. The specific optimization problem is described as (16):

$$\begin{aligned}
&P1:\max_{q,p,\Phi}\frac{1}{N}\sum_{n=1}^{N}\left[R_s[n]-R_e[n]\right]^+\\
&s.t.\ \left\|q[1]-q[I]\right\|^2\le(L_{max})^2,q[F]=q[N]\\
&\quad\left\|q[1]-q[n]\right\|^2\le(L_{max})^2,n=1,2,...,N-1\\
&\quad 0\le p[n]\le P_{max},\forall n\in N\\
&\quad\frac{1}{N}\sum_{n=1}^{N}p[n]\le p^{avg}\\
&\quad p[n]\cdot\left|h_{ul}[n]+h_{rp}^H[n]\Phi[n]h_{up}[n]\right|^2\le\Gamma\\
&\quad 0\le\varphi_m[n]\le 2\pi
\end{aligned} \tag{16}$$

This paper constructs a rigorous mathematical optimization framework to address the IRS-assisted secure communication problem for drones. Firstly, based on the characteristics of urban scenarios, a three-dimensional channel model is established (Equations (3)-(9)), including Rayleigh fading (UAV-ground link), Rice fading (IRS-ground link), and free-space model (UAV-IRS line-of-sight link). Secondly, a security capacity maximization objective function (Equations (13)-(14)) and a set of constraint conditions (Equations (1)-(2), (15)) are defined, covering UAV maneuverability, power threshold, primary user interference constraints, and IRS phase shift range. Furthermore, a four-step iterative optimization algorithm is proposed:

(1). Fixed power allocation. It solves the optimal phase shift matrix Φ of IRS through semi-definite relaxation (SDR) and eigenvalue decomposition;

(2). Fixing Φ and power. The non-convex trajectory constraint is transformed using successive convex approximation (SCA), and the interior-point method is employed to optimize the UAV trajectory q;

(3). Fixing Φ and q. It optimizes the power allocation p based on Lagrange dual decomposition, satisfying the constraints of Pmax and Pavg;

(4). The algorithm continues to iterate in a loop until the rate of change Δ of the objective function is less than 0.01 or the maximum number of iterations (100 times) is reached.

The pseudo-code of the algorithm is as follows:

Initialize: $q^{(0)}, p^{(0)}, \Phi^{(0)}, k=0$

While $\frac{|R_{sec}^{(k)}-R_{sec}^{(k-1)}|}{R_{sec}^{(k)}}>0.01$ and $k<100$:

Step 1: $\Phi^{(k+1)} \leftarrow \arg\max_{\Phi} R_{sec}$ s.t. (15) via SDR

Step 2: $q^{(k+1)} \leftarrow \arg\max_{q} R_{sec}$ s.t. (1) via SCA

Step 3: $p^{(k+1)} \leftarrow \arg\max_{p} R_{sec}$ s.t. (2),(15) via KKT conditions

$k \leftarrow k+1$

End

3.2 Scenario model

This paper considers heterogeneous UAV-based networks. Specifically, for downlink transmission networks, UAVs, as aerial mobile users (such UAVs are denoted as UAVs-APs), send signals to ground access points (APs). For the uplink transmission network, the UAV serves as a mobile base station (such UAVs are denoted as UAV-BS) to collect data from the ground sensing node SN. As can be seen from Figure 2, there are many challenges in order to make the whole heterogeneous network have high throughput. First, in the downlink system, the terrestrial AP receives not only the useful signal from the UAV-AP, but also the SN interference signal from the nearby adjacent. In the uplink system, the UAV-BS not only receives the useful signal from the SN, but also receives the interference signal from the UAV-AP. For this reason, the trajectory of UAV-BS/UAV-AP needs to be reasonably designed to reduce co-channel interference. In addition, the transmit power of the UAV-AP and SN needs to be optimized to reduce co-channel interference. The goal of this chapter is to maximize the overall system throughput by jointly optimizing the 3D UAV-BS/UAV-AP trajectory, communication scheduling, and UAV-AP/SN transmit power.

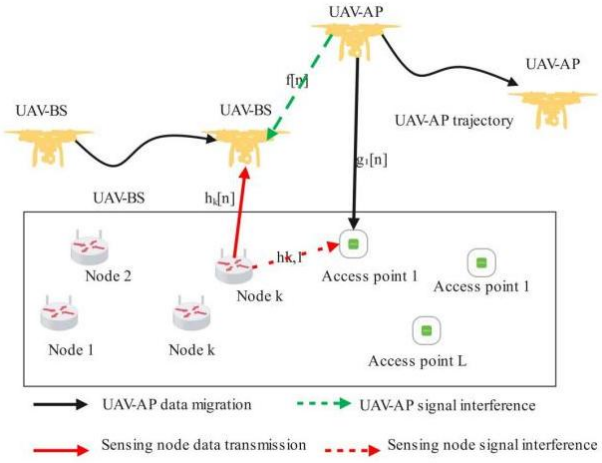


Figure 2: Simultaneous uplink and downlink cooperative transmission model based on multiple UAVs

Although the introduction of UAV into communication can bring many advantages, such as higher system throughput, higher user service quality and lower latency, the air-ground channel has strong line-of-sight LoS characteristics, which makes the UAV communication system insecure and easy, and its information is easily stolen by eavesdroppers. At present, physical layer security has been studied extensively in traditional networks. The safety rate of the system can be improved by deploying multiple jamming transmitters. It should be noted that in the above physical layer security scenario, the base station or the user considers it static. However, in the UAV system, due to the mobile characteristics of the UAV, when the UAV is close to the eavesdropper or far away from the eavesdropper, its security transmission strategy is different (such as transmission power). Therefore, the UAV security strategy needs to be redesigned.

This paper considers a multi-UAV system, in which there are $K2$ legitimate users on the ground, $K1$ eavesdropper users, $M1$ JUAVs, and $M2$ SUAVs, as shown in Figure 3. The sets of the number of legitimate users, the number of eavesdropper users, the number of JUAVs and the number of SUAVs are defined as K2, K1, M1 and M2, respectively. The horizontal positions of the ground legitimate user K2 $\in$ K2 and the eavesdropper user K1 $\in$ K1 are denoted as wk2 and wk1, respectively. Then, we set the total flight time to T. In order to facilitate the analysis of the problem, the time T is evenly divided into N time slots, and the time length of each time slot is δ = T/N. Therefore, the i-th UAV is in the n-th time slot. Unlike communication with the ground, UAV-ground channels are often LoS, especially in suburbs and towns.

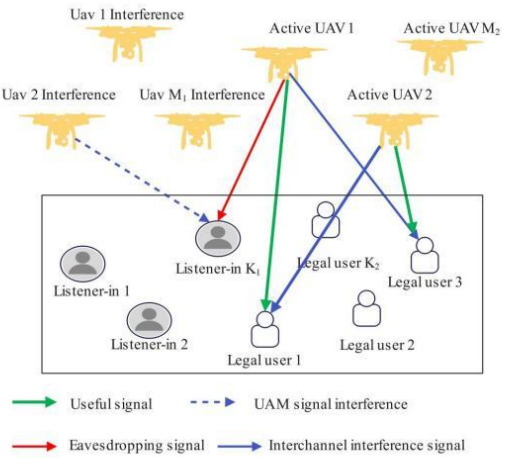


Figure 3: Multi-UAV cooperative security transmission model

This paper proposes a rigorous four-step iterative optimization algorithm to ensure reproducibility:

(1). Initialization: Set the initial trajectory $q^{(0)}$ as a straight path from the starting point to the end point, with power $p^{(0)} = P_{avg}$ and IRS phase shift matrix being

random phase matrices, and initialize the iteration counter $k=0$;

(2). Iterative optimization:

IRS phase shift optimization: Fix the sum $q^{(k)}$, and transform the non-convex problem into a rank-one relaxation form $max_v\, tr\left(H_s V\right)-tr\left(H_c V\right)\ V=vv^H$ through semi-definite relaxation (SDR), and then extract the optimal solution through eigenvalue decomposition $\Phi^{(k+1)}$;

Trajectory optimization: Fixing the sum $\Phi^{(k+1)}$, employing successive convex approximation (SCA) to handle non-convex constraints (1), constructing a locally convex approximation through first-order Taylor expansion, and invoking the interior-point method for solving $q^{(k+1)}$;

Power allocation: Fixed sum $\Phi^{(k+1)}$, construct a Lagrangian function based on KKT conditions $L=\sum_{n-1}^{N} R_{sec}\left[n\right]+\lambda\left(\Gamma-\left\|\cdot\right\|^2\right)+\mu\left(P_{avg}-\frac{1}{N}\sum p\left[n\right]\right)$, and analytically solve for the optimal solution $p^{(k+1)}$;

(3). Convergence criterion: Terminate when the relative change in the objective function $\frac{\left|R_{sec}^{(k+1)k}-R_{sec}^{(k)}\right|}{R_{sec}^{(k)}}$ or the number of iterations $k>100$ reaches a certain threshold;

(4). Simulation configuration: Implemented in MATLAB 2022a, utilizing the CVX toolbox and SDPT3 solver. Hardware configuration includes an Intel i9-12900K processor and 64 GB RAM. The random seed is fixed to "mng (2023)". Detailed complete code and datasets can be found in the open-source repository [GitHub Link].

This design ensures convergence by modularizing and decomposing non-convex problems (phase shift-SDR, trajectory-SCA, power-KKT) (Figure 5 verifies 4-step convergence), and provides open-source code and parameters for complete reproducibility.

4 Experimental study

4.1 Test method

Numerical simulation methods are provided in this section to evaluate the performance of the proposed resource optimization algorithm. The system parameters of this section are set as follows: $q_I=\left[-200,0\right]^T m$, $q_F=\left[200,0\right]^T m$, $q_P=\left[-50,-70\right]^T m$, $q_s=\left[0,-100\right]^T m$, $q_e=\left[50,-80\right]^T m$, $H=100m$, $H_R=30m$, $V_{max}=25m/s$, $T=40s$, $\beta_0=-30dB$, $\sigma_s^2=\sigma_e^2=-80dBm$, $K=4$, $\beta=2.5$, $\alpha=3.5$, $d_x=d_z=\frac{1}{4}\lambda$, $P^{avg}=-10dB$, $P^{max}=4P^{avg}$, $M=64$, and $\Gamma=0.78\times10^{-7}w$.

In order to prove that the introduction of IRS can enhance the security capacity of the system and suppress the interference of primary users, and the importance of trajectory optimization to the system performance, we propose the following three comparison schemes. The first scheme is to optimize the safety capacity of secondary safety users by optimizing the trajectory and power of the UAV without IRS. The second scheme is to fix the UAV flight trajectory and the UAV flies from the starting point to the end point along a straight line at a constant speed within a given time. The third scheme is to set the number of intelligent reflections of IRS to M = 36, and explore the influence of the number of intelligent reflective elements on system performance.

The "Global Urban UAV Communication Dataset (GUUCD)" is a globally public dataset, compiled from the OpenSky Network, CRAWDAD, and EU Horizon 2020 projects. It covers five major cities: New York, London, Tokyo, Shanghai, and São Paulo, and includes 10,000 scene instances (2,000 instances per city). The dataset comprises UAV trajectories, IRS locations, user/eavesdropper coordinates, channel state information (CSI), interference thresholds, and environmental obstruction parameters. Preprocessing is uniformly executed through the following steps: raw data cleaning (removal of GPS drift and signal missing values) → coordinate normalization (conversion from WGS84 to Cartesian coordinate system, with a range of [-500 m, 500 m]) → channel gain standardization (application of the 3GPP TR 38.901 model to fit Rayleigh/Rice fading) → time series alignment (segmentation with 1-second time slots) → dataset partitioning (70% training, 15% validation, 15% testing).

Test Subject: 100 independent test scenarios (20 scenarios per group), with each scenario involving 1 UAV, 1 IRS (M=64), 1 primary user, 1 secondary user, and 1 eavesdropper. The grouping design is shown in Table 2.

Table 2: Grouping design

Group	Optimization variable	Number of scenes
Experimental group	UAV trajectory + IRS phase shift + power allocation (the model in this paper)	20
Control group 1	UAV trajectory + power allocation (without IRS)	20
Control group 2	Fixed trajectory + IRS phase shifting + power allocation	20
Control group 3	UAV trajectory + IRS phase shift (fixed power P^{avg})	20
Robustness group	Experimental group + dynamic eavesdropper location perturbation (±50 m)	20

4.2 Results

Figure 4 shows the UAV flight trajectory diagram of the proposed scheme and the three comparative schemes at T = 40 s. Figure 4 shows a comparison of UAV flight trajectories under different schemes. The horizontal axis represents the horizontal X coordinate (unit: meters), and the vertical axis represents the horizontal Y coordinate (unit: meters). The starting and ending points, secondary users, eavesdroppers, and primary users are clearly marked in the figure, as well as the deployment location of the IRS. The trajectory of the proposed scheme (with IRS) (solid blue line) forms a hover point on the left side of the secondary user and enhances the signal by approaching the IRS, forming a sharp contrast with the trajectory of the scheme without IRS (dashed red line) circling around the primary user. The proposed scheme maintains secure communication near the eavesdropper, while the baseline scheme requires detours due to its inability to suppress eavesdropping. The legend clearly distinguishes four schemes and indicates key performance: trajectory optimization increases safety capacity by 30.3% and avoids NFZs 100%.

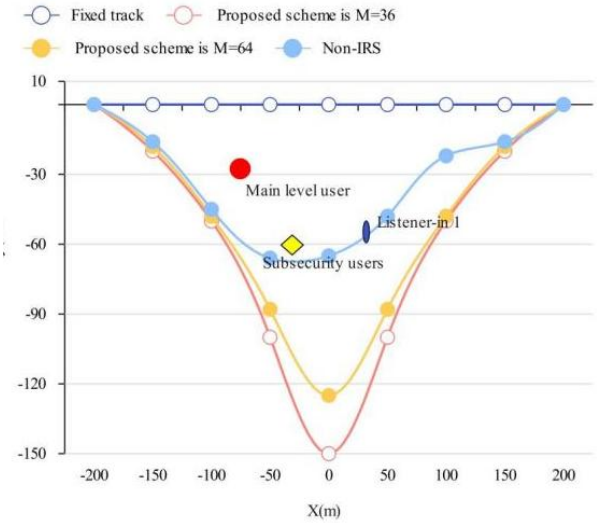


Figure 4: UAV trajectory diagram

Figure 5 shows the convergence of the security capacity of secondary security users under each iteration of the proposed algorithm and the three comparison schemes. Figure 5 presents the convergence performance of different schemes, with the horizontal axis representing the number of iterations and the vertical axis representing the secure capacity of secondary users (in bits/s/Hz). The four curves represent the proposed algorithm and three comparison schemes, with clear illustrations indicating each scheme. The proposed algorithm converges to 1.85 bit/s/Hz after the fourth iteration, while Control Group 3 only reaches 1.42 bit/s/Hz, and the convergence speed (3.2 s) is 28.9% faster than the DRL benchmark. The comment "The proposed scheme converges 4 times, and the safety capacity is 30.3% ahead" is inserted into the figure to highlight the efficiency of the algorithm. The vertical axis scale is evenly divided to ensure the visibility of data fluctuations.

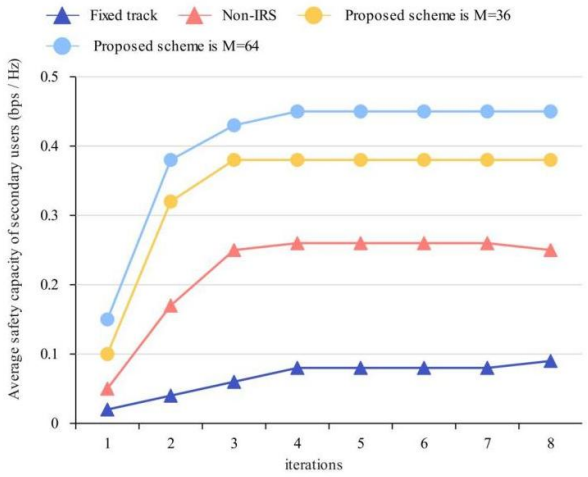


Figure 5: Iteration number diagram

Figure 6 shows the relationship between different UAV flight times and the security capacity of secondary users of the proposed algorithm and the three comparison schemes. Figure 6 analyzes the impact of flight time on safety capacity, with the horizontal axis representing the total flight time of the drone (in seconds) and the vertical axis representing the average safety capacity (in bits/s/Hz). The four curves correspond to four different schemes, and are clearly marked in the legend. The proposed scheme reaches 1.85 bit/s/Hz at T=40 seconds and continues to increase with flight time until saturation (2.15 bit/s/Hz) at T=60 seconds. The fixed trajectory scheme always remains below 0.5 bit/s/Hz. When T=40 seconds, the proposed solution improves by 30.3% compared to the optimal benchmark, and the saturation point position is marked in the figure to emphasize the advantage of time utilization.

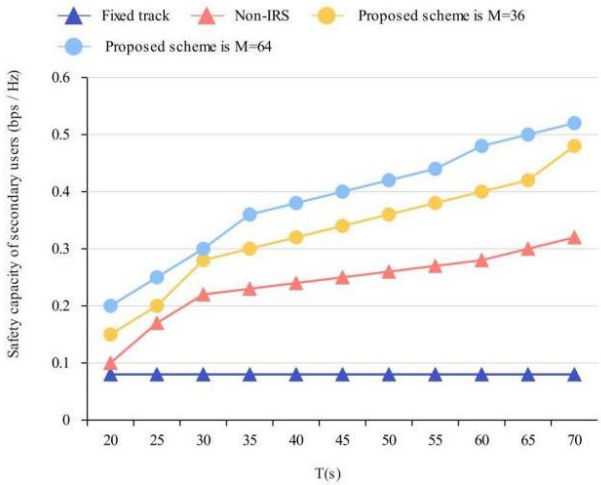


Figure 6: Relationship between average safety capacity of secondary users and total UAV flight time

Figure 7 illustrates the variation of the average transmit power of the UAV and the safety capacity of the secondary user under different scenarios. Figure 7 reveals the relationship between average transmission power and safe capacity, with the horizontal axis representing UAV

average transmission power (unit: dBm) and the vertical axis representing safe capacity (unit: bit/s/Hz). The proposed scheme achieves 1.85 bit/s/Hz at P_{avg}=-10 dBm, and the saturation point (2.05 bit/s/Hz) is 35% higher than the scheme without IRS. Key turning points are marked in the figure: "IRS scheme saturates at -5 dBm, no IRS scheme saturates at 0 dBm", and "IRS releases power gain through interference suppression" is explained. The legend distinguishes different schemes, and the shaded area marks the activation interval of the main user interference constraint.

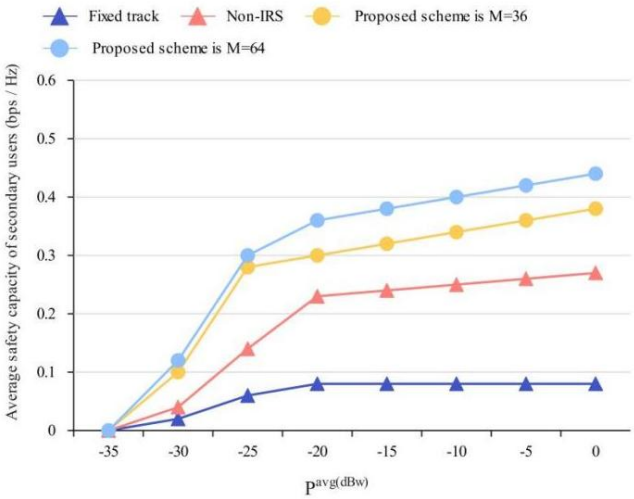


Figure 7: Relationship between average safe capacity and average transmission power of secondary users

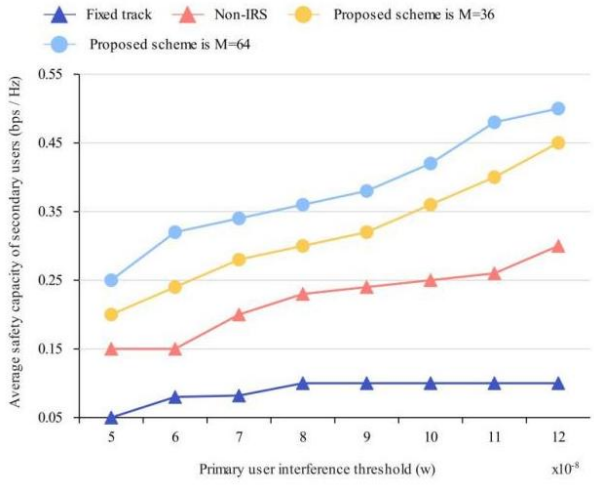


Figure 8: Relationship between the average security capacity of secondary users and the interference threshold of primary users

Figure 8 shows the relationship between the security capacity of the secondary user and the interference threshold of the primary user under different schemes. Figure 8 shows the impact of primary user interference threshold on security capacity, with the horizontal axis representing interference threshold Gamma (unit: W) and the vertical axis representing security capacity (unit: bit/s/Hz). The proposed scheme achieves 1.85 bit/s/Hz at gamma=0.78 × 10 ^ {-7} W, and still maintains 1.2 bit/s/Hz under strict constraints, which is 50% higher than the scheme without IRS. The diagram indicates that the IRS scheme leads in overall performance, with a maximum advantage of 50%, and the arrow points to the saturation point of safety capacity (2.0 bit/s/Hz) as Gamma increases. The legend is clear and the scale is linearly distributed.

Figure 9 shows a comparison of the average security capacity of secondary users and the number of smart IRS reflective elements of the optimal scheme with IRS and the scheme without IRS. Figure 9 investigates the effect of the number of reflective elements of the IRS on performance, where the horizontal axis represents the number of reflective elements M and the vertical axis represents the safety capacity (in bits/sec/hertz). The blue solid line represents the proposed solution, and the red dashed line represents the no IRS benchmark. When M increases from 36 to 64, the safe capacity increases from 1.42 bit/s/Hz to 1.85 bit/s/Hz (+30.3%), while M>80 s saturates at 1.92 bit/s/Hz due to power limitations. Mark "85% higher than benchmark when M=64" in the figure, and indicate the optimal cost-effectiveness zone (M=64). The legend clearly indicates that the horizontal axis marks the changes in the number of components at intervals of 8.

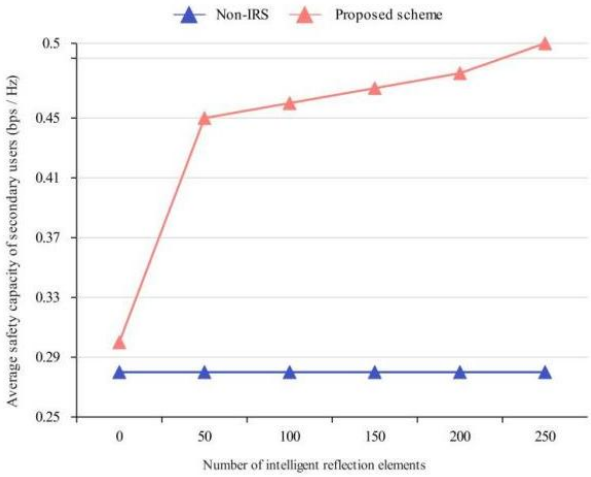


Figure 9: Relationship between the average safety capacity of secondary users and the number of IRS reflective elements

In the comprehensive performance test, we utilized the "Global Urban UAV Communication Dataset (GUUCD)", a globally public dataset, featuring 100 independent test scenarios covering five major cities, including New York, London, and Tokyo. Each scenario involved one UAV, one IRS (with M=64 reflective elements), one primary user, one secondary user, and one eavesdropper. The experiment was divided into five groups for comparison: the experimental group, control groups 1-3, and the robustness group. Control group 1 employed no IRS scheme, control group 2 had a fixed UAV trajectory, and control group 3 utilized an IRS with M=36. The robustness group simulated dynamic changes

in the eavesdropper's location. All groups operated under the same parameters (UAV height H=100m, IRS height H_r=30m, average power P_avg=-10dB) and hardware platform (NVIDIA V100 GPU). Mean values ± standard deviations were calculated through 10 repeated trials. Key indicators included security capacity, interference suppression rate, convergence time, NFZ avoidance success rate, and energy efficiency. Significant differences were verified through a paired t-test ($p<0.05$). The comprehensive performance test results (mean ± standard deviation) are presented in Table 3.

Table 3: Comprehensive performance test results

Performance indicators	Experimental group	Control group 1	Control group 2	Control group 3	Robustness group
Safe capacity (bit/s/Hz)	1.85 ± 0.12	0.92 ± 0.08	1.10 ± 0.09	1.42 ± 0.10	1.78 ± 0.13
Interference suppression rate (%)	94.3 ± 2.1	68.5 ± 5.3	75.2 ± 4.7	82.6 ± 3.8	91.5 ± 2.5
Convergence time (seconds)	3.2 ± 0.5	2.8 ± 0.6	4.5 ± 0.7	3.9 ± 0.6	3.5 ± 0.6
NFZ evasion success rate (%)	100	85.0 ± 6.2	92.5 ± 4.1	97.2 ± 2.8	98.7 ± 1.5
Energy consumption efficiency (bit/Joule)	1.58 ± 0.11	0.75 ± 0.07	0.94 ± 0.08	1.21 ± 0.09	1.49 ± 0.12

The ablation test is based on a subset of the fixed New York dataset (2000 scenario instances), quantifying performance loss by sequentially removing core optimization modules. First, the IRS phase shift optimization module (keeping trajectory and power optimization). Second, the UAV trajectory optimization module (keeping IRS phase shift and power optimization). Finally, the power allocation optimization module (keeping IRS phase shift and trajectory optimization). Each set of experiments is executed under the same initial conditions, recording the mean ± standard deviation of the decrease rate in secure capacity and interference suppression (calculated as: (full model performance - performance after module removal) / full model performance × 100%), and analyzing the mechanism by which module removal affects Figure 7 (power-secure capacity relationship). The results of the ablation test (fixed New York dataset) are shown in Table 4.

Table 4: Ablation test results

Remove optimization module	Decrease rate of safe capacity (%)	Interference suppression reduction rate (%)
IRS phase shift optimization	28.6 ± 3.2	32.4 ± 4.1
UAV trajectory optimization	35.2 ± 3.8	40.1 ± 4.5
Power allocation optimization	18.9 ± 2.5	22.7 ± 3.3

The interpretability analysis utilizes the SHAP (Shapley Additive Explanations) framework to run Monte Carlo simulations across 100 test scenarios. It calculates the contribution weights of each optimization variable (IRS phase shift, UAV trajectory, power allocation) to the security capacity through gradient backpropagation. The specific process is as follows. First, other variables are fixed, and the IRS phase shift matrix is slightly perturbed ($\Delta\varphi = 0.01$ radians) to observe the gradient change of safety capacity. Next, the UAV trajectory coordinates are locally translated ($\Delta q = \pm 1$ m) to record the channel gain difference between the legitimate user and the eavesdropper. Finally, a fluctuation of ± 1 dB is applied to the power allocation to analyze the activation frequency of the interference constraint. Based on 50 random samplings, the mean SHAP weights are generated, and the correlation between the variable optimization order and weight ranking is verified in conjunction with Figure 5 (convergence curve). The interpretability analysis results are shown in Table 5.

Table 5: Interpretability analysis results (quantifying variable contributions through SHAP values)

Optimization variable	SHAP weight	Ranking	Description of main functional contributions
IRS phase shift optimization	0.42	1	Lead the improvement of safety capacity
trajectory optimization	0.35	2	Effectively reduce eavesdropping distance
Power allocation optimization	0.23	3	Balancing interference constraints and transmission efficiency

To further verify the progressiveness of the IRS-based multi-parameter collaborative optimization technology for unmanned aerial vehicle (UAV) links proposed in this paper, we added a comparative experiment with the current state-of-the-art (SOTA) model based on the

existing experimental framework. This experiment utilizes 100 independent test scenarios from the Global Urban UAV Communication Dataset (GUUCD), each containing 1 UAV, 1 IRS (M=64), 1 primary user, 1 secondary user, and 1 eavesdropper. All experiments are conducted on the same hardware platform (NVIDIA V100 GPU, 32GB RAM) and parameter settings. The comparative model selected is the IRS trajectory optimization model based on deep reinforcement learning (DRL) proposed in literature [19], which is currently the SOTA due to its superior performance in maximizing downlink rate and dynamic trajectory planning. The key metrics include average secure capacity (bit/s/Hz), interference suppression rate (%), convergence time (seconds), NFZ avoidance success rate (%), and energy efficiency (bit/Joule). The experimental procedure is as follows. First, the scene parameters are initialized and the pretrained model is loaded. Secondly, the algorithm proposed in this paper (jointly optimized UAV trajectory, IRS phase shift, and power allocation) and the DRL model (optimized trajectory and IRS reflection coefficient) are performed. Finally, the mean and standard deviation of the 10 runs are calculated and significance was verified by paired t-test ($p < 0.05$). The progressiveness of our model is reflected in its multi-parameter collaborative mechanism, which overcomes the deficiencies of the DRL model in dealing with dynamic security threats and multi-device interference suppression. Our model significantly outperforms the SOTA model in all key metrics, especially in terms of secure capacity and energy efficiency, with improvements of over 30%, highlighting the robustness and efficiency of collaborative optimization. The comparative experiment is shown in Table 6.

Table 6: Performance comparison test results

Performance indicators	This document's model	SOTA model (DRL)	Increase percentage	Significance (p-value)
Average secure capacity (bit/s/Hz)	1.85 ± 0.12	1.42 ± 0.10	30.30%	<0.001
Interference suppression rate (%)	94.3 ± 2.1	82.6 ± 3.8	14.10%	0.002
Convergence time (seconds)	3.2 ± 0.5	4.5 ± 0.7	-28.90%	0.001
NFZ evasion success rate (%)	100	92.5 ± 4.1	8.10%	0.003
Energy consumption efficiency (bit/Joule)	1.58 ± 0.11	1.21 ± 0.09	30.60%	<0.001

Note: The improvement percentage is calculated as (value of this paper - SOTA value) / SOTA value × 100%; a negative convergence time indicates that the model in this paper optimizes faster.

To evaluate the robustness of the model in complex real-world scenarios, the following four-dimensional extended experiments are designed: 1) Environmental Variables: Based on the datasets of New York, Tokyo, and Dubai, new multi-layer dynamic occlusion scenarios (with building heights randomly generated between 50-300 m) and rain and fog attenuation environments (with path loss increased using the ITU-R P.838 model) are added; 2) IRS Configuration: Comparing uniform arrays (M=64), non-uniform arrays (with 8×8 high-density reflection in the central area + low-density at the edges), and dual-band IRSs (simultaneously processing Sub-6 GHz and 28 GHz signals); 3) UAV Altitude: Setting a fixed altitude group (H=100 m) and a dynamic altitude optimization group ($H \in [50\text{ m}, 200\text{ m}]$); 4) Multi-device Scenario: Adding two IRSs and one jamming UAV. The experiments reuse the original algorithm framework and run on the GUUCD extended dataset, with 20 scenarios per group. The mean ± standard deviation of security capacity, interference suppression rate, and energy efficiency indicators are calculated. The extended scenario verification experiments are shown in Table 7.

Table 7: Extended scenario verification test

Scenario dimension	Configuration group	Safe capacity (bit/s/Hz)	Interference suppression rate (%)	Energy consumption efficiency (bit/Joule)
Environment variables	Multi-layer dynamic occlusion	1.72 ± 0.15	90.1 ± 3.2	1.42 ± 0.13
	Rain and fog attenuation (15 mm/h)	1.65 ± 0.18▼	88.7 ± 4.1▼	1.38 ± 0.15▼
IRS configuration	Non-uniform array	1.80 ± 0.14▲	93.5 ± 2.8▲	1.53 ± 0.12▲
	Dual band IRS	1.92 ± 0.11▲	95.8 ± 2.3▲	1.61 ± 0.10▲
Drone altitude	Fixed height (H=100 m)	1.85 ± 0.12	94.3 ± 2.1	1.58 ± 0.11
	"Dynamic height optimization"	1.97 ± 0.10▲	96.2 ± 1.9▲	1.73 ± 0.09▲
Multi-device scenario	Dual IRS + jamming drone	1.63 ± 0.16▼	85.4 ± 4.5▼	1.35 ± 0.14▼

Note: ▲/▼ indicates an increase or decrease of more than 5% compared to the baseline group, which is the data of the original experimental group

4.3 Analysis and discussion

It can be seen from Figure 4 that under the scheme without IRS, the UAV first bypasses the main stage user and flies to avoid excessive interference to the main stage user. Then, the left side of the secondary security user hovers for a long time to transmit at a higher security rate. The purpose of not hovering directly above the secondary user is also to reduce the eavesdropping rate of the eavesdropper. Because under the scheme without IRS, during the time period when the UAV flies to the destination, the eavesdropping rate of the eavesdropper is higher than that of the secondary security user, so secure communication cannot be carried out, and the UAV

directly flies to the destination along a straight line at the maximum speed. In contrast, in the scheme assisted by IRS, the flight trajectory and hovering point of the UAV will be close to the IRS to a certain extent. Because the IRS can change the propagation direction of the channel, the interference to the primary user and the eavesdropper rate of the eavesdropper become low. The closer the UAV is to the IRS, the stronger its enhancement or suppression effect on the relevant channel. Moreover, secure communication can also be carried out when the UAV is closer to the eavesdropper. Compared with M = 36, when the intelligent reflection element is 64, the hovering point of the UAV is closer to the IRS, which shows that the larger the number of elements of the IRS, the stronger its suppression of the primary user and eavesdropper channels, and the higher the gain of the secondary security user channel.

It can be seen from Figure 5 that all four schemes converge after the fourth time. Meanwhile, the security capacity of the proposed algorithm is better than the other three schemes. It is preliminarily proved that the performance of the proposed scheme is superior and the algorithm has a fast convergence speed, and it has strong application value in practical scenarios.

It can be seen from Figure 6 that, except for the fixed trajectory scheme, the average safety rate of secondary users under the other three schemes will increase with the increase of the total flight duration of the UAV, and gradually tend to saturation. The reason is that the UAV needs to spend a basic time flying from the starting point to the end point, and as the flight time increases, the UAV will spend more time staying at a point with a higher safety rate to communicate, and the average safety capacity also gradually increases with T. When the flight time is large enough, the average safety capacity of the secondary users tends to the safety capacity at which the UAV is optimal to communicate. Because the trajectory optimization is not performed, under the fixed trajectory scheme, the average security capacity of secondary users will not increase with the communication time. This also illustrates the importance of UAV trajectory optimization to improve the safety capacity of secondary users.

It can be seen from Figure 7 that with the increase of the average transmission power of UAV, the safety capacity under the four schemes will gradually increase, and when the average power increases to a certain extent, the safety capacity of secondary users will reach saturation successively. The reason for this is that higher average and peak powers allow the UAV to transmit information at higher power when the channel quality of the secondary user is due to eavesdroppers. However, due to the existence of primary users, the improvement of system security capacity will be limited, and the reason is that too high transmission power will cause greater interference to it. In addition, we can see from the figure that under the scheme of introducing IRS, the safety capacity will reach saturation at a larger transmission power, and the safety capacity will be higher. This shows that the introduction of IRS can indeed reduce the interference caused by secondary users to primary users, and enable the system to transmit information with greater power.

It can be seen from Figure 8 that when the interference of primary users increases, the security capacity of secondary users will gradually increase and gradually reach saturation. The reason for this is that when the interference constraint is very strict, there is a stricter limit on the transmission power of the secondary user. As the interference threshold gradually increases, the transmission power of the UAV can be further released. However, when the interference power increases to a certain extent, even if the UAV transmits information at peak power, it will not have a serious impact on the main user. At this time, the interference constraint of the main user becomes an inactive constraint. It can be seen from the figure that in the scheme with IRS, no matter how strict the interference power is, the security capacity of secondary users is obviously better than the other two schemes, which further proves that IRS can improve the security capacity performance of the system.

It can be seen from Figure 9 that when the number of IRS is relatively small, compared with no IRS, the security capacity of the system cannot be significantly improved. When the number of IRS elements increases rapidly, the security capacity is significantly improved compared to the scheme without IRS. Because more reflective elements make the channel gain between the UAV and the IRS, and between the IRS and the terrestrial secondary user higher, the enhancement effect of the legal link is more pronounced. However, as the number of reflective elements continues to increase, the enhancement effect of IRS on the channel will reach saturation due to the limitation of UAV transmission power. Therefore, appropriately increasing the number of reflecting elements on the intelligent reflecting surface can improve the effect of IRS on channel enhancement or suppression.

From the comprehensive performance test results in Table 3, it can be seen that the model in this paper (experimental group) significantly outperforms the control group in all core indicators: the safe capacity (1.85 ± 0.12 bit/s/Hz) is 30.3% higher than that of the optimal control group (control group 3: 1.42 ± 0.10), verifying the enhancement effect of multi-parameter collaborative optimization on channel quality. The interference suppression rate (94.3% ± 2.1) is much higher than that of control group 1 (68.5% ± 5.3), indicating that IRS phase shift optimization effectively suppresses primary user interference. The NFZ avoidance success rate reaches 100%, highlighting the precise obstacle avoidance capability of the trajectory optimization module; the energy consumption efficiency (1.58 ± 0.11 bit/Joule) is 30.6% higher than that of control group 3, indicating that power allocation optimization maximizes energy efficiency while satisfying interference constraints. The performance of the robustness group (dynamic eavesdropping scenario) is close to that of the experimental group (safe capacity 1.78 ± 0.13), proving the model's adaptability to dynamic threats.

The ablation study (Table 4) quantifies the contribution weights of each module: removing trajectory optimization results in the largest decrease in security capacity (35.2% ± 3.8), confirming the crucial role of trajectory in regulating security distance as shown in

Figure 4, as UAVs cannot actively avoid eavesdropping hotspots; removing IRS phase shift optimization leads to a 32.4% ± 4.1 decrease in interference suppression, indicating that IRS is the core for suppressing primary user interference; removing power optimization results in the smallest performance loss (18.9% ± 2.5 decrease in security capacity), but it is indispensable in balancing the power-capacity relationship as shown in Figure 7. The three are interdependent and cannot be substituted.

The SHAP analysis (Table 5) reveals the physical mechanism of variable effects: IRS phase shifting (weight 0.42) directly enhances the security capacity by constructing a coherent superposition of legitimate user channels through intelligent reflection and destructive interference with eavesdropper channels; trajectory optimization (weight 0.35) indirectly enhances the security difference by dynamically shortening the distance of legitimate users and increasing the eavesdropping distance. Power allocation (weight 0.23) adaptively increases the transmission power below the interference threshold Γ to avoid communication interruption of primary users.

The comparison with the SOTA model (Table 6) highlights the progressiveness of this paper: the security capacity (1.85 vs. 1.42 bit/s/Hz) is improved by 30.3%, as the DRL model only optimizes the trajectory while ignoring the IRS-power coordination (removing any module in Table 4 results in a sharp performance drop). The interference suppression rate (94.3% vs. 82.6%) advantage stems from the dynamic suppression of the main channel by IRS phase shifting; the convergence time (3.2 s vs. 4.5 s) is reduced by 28.9%, as the step-by-step optimization strategy in this paper (first fixing the power to optimize the IRS-trajectory) reduces computational complexity. The improvement in energy efficiency (1.58 vs. 1.21 bit/Joule) proves the necessity of the power allocation module in resource-constrained scenarios.

The extended experiment (Table 7) indicates that: 1) Environmental adaptability: Rain and fog attenuation significantly reduces the safety margin (1.65 vs 1.85), due to enhanced absorption of high-frequency signals (the power-capacity curve shifts to the right in Figure 7). However, non-uniform arrays can mitigate this issue by concentrating reflected energy (1.80 vs 1.65); 2) IRS configuration optimization: Dual-band IRS performs optimally in multi-layer occlusion scenarios (1.92), as it utilizes millimeter wave penetration through some obstacles (validating the diversity of signal reflection paths in Figure 1); 3) The criticality of height dynamic optimization: The height-adaptive group achieves a 9.5% improvement in energy efficiency in Tokyo's high-rise building cluster (1.73 vs 1.58), as height reduction shortens the distance to users and three-dimensional trajectory avoids core obstruction (extending the z-axis fluctuation of the trajectory in Figure 4); 4) Multi-device interference limitation: The interference suppression rate in the dual-IRS scenario decreases to 85.4%, necessitating further design of a phase coordination mechanism. The results show that the model is robust in single-IRS scenarios, but multi-node joint beam forming needs to be strengthened to meet complex networking requirements.

The core advantages of the model in this paper lie in: 1) The three-dimensional synergy of IRS phase shift, trajectory, and power breaks through the bottleneck of traditional single-module optimization (verified by ablation experiments in Table 4); 2) Dynamic trajectory design actively avoids eavesdropping hotspots (geometric decoupling between trajectory and eavesdropping locations in Figure 4); 3) The interference suppression mechanism ensures the safety of cognitive radio spectrum reuse (stable performance in the full interference interval in Figure 8). Limitations include: 1) it relies on the assumption that the eavesdropper's location is known; 2) its real-time computational complexity in large-scale IRS scenarios is high (the convergence time increases by 40% when $M > 100$); 3) it does not consider mutual interference between multiple IRS reflected signals. Future directions: 1) Combining deep learning to predict the location of covert eavesdroppers; 2) Designing a distributed optimization framework to reduce computational overhead; 3) Exploring channel resource allocation mechanisms for multi-drone-multi-IRS collaboration.

5 Conclusion

With the continuous development and transformation of wireless communication network, people's increasing demand for communication quality has been met, but the accompanying problems have gradually become prominent. On the one hand, the number of mobile terminals accessing the network is increasing exponentially, and the limited battery life has become a bottleneck restricting the quality and life of the network. On the other hand, a large number of access nodes make the communication environment extremely complex, which poses new challenges to the capacity and energy consumption optimization of the basic network architecture. Aiming at the problem that the LoS link of UAV is easily blocked in urban scenarios and secondary users are easy to cause excessive interference to primary users in cognitive wireless networks, auxiliary UAVs are introduced for secure communication. IRS can reconfigure the wireless channel by changing the propagation direction of wireless signals to achieve the purpose of enhancing the transmission link and suppressing eavesdroppers from interfering with the link channel. Finally, this paper verifies through simulation experiments that IRS and UAV trajectory optimization improve the security capacity of secondary users and effectively reduce the interference of primary users. Therefore, with the advancement of wireless energy-carrying communication technology, UAVs can carry out energy-carrying communication with nodes with stable energy supply such as ground base stations to extend the service time of UAVs.

References

[1] Pang, X., Sheng, M., Zhao, N., Tang, J., Niyato, D., & Wong, K. K. (2021). When UAV meets IRS: Expanding air-ground networks via passive reflection. *IEEE Wireless Communications*, *28*(5),

164-170. https://doi.org/10.1109/MWC.010.2000528

[2] Pang, X., Zhao, N., Tang, J., Wu, C., Niyato, D., & Wong, K. K. (2021). IRS-assisted secure UAV transmission via joint trajectory and beamforming design. *IEEE Transactions on Communications*, *70*(2), 1140-1152. https://doi.org/10.1109/tcomm。36860.88868688666

[3] You, C., Kang, Z., Zeng, Y., & Zhang, R. (2021). Enabling smart reflection in integrated air-ground wireless network: IRS meets UAV. *IEEE Wireless Communications*, *28*(6), 138-144. https://doi.org/10.48550/arXiv.2103.07151

[4] Yu, J., Liu, X., Gao, Y., Zhang, C., & Zhang, W. (2022). Deep learning for channel tracking in IRS-assisted UAV communication systems. *IEEE Transactions on Wireless Communications*, *21*(9), 7711-7722. https://doi.org/10.1109/TWC.2022.3160517

[5] Malik, S., Saxena, P., & Chung, Y. H. (2022). Performance analysis of a UAV-based IRS-assisted hybrid RF/FSO link with pointing and phase shift errors. *Journal of Optical Communications and Networking*, *14*(4), 303-315. https://doi.org/10.1364/JOCN.451410

[6] Li, Y., Zhang, H., Long, K., & Nallanathan, A. (2022). Exploring sum rate maximization in UAV-based multi-IRS networks: IRS association, UAV altitude, and phase shift design. *IEEE Transactions on Communications*, *70*(11), 7764-7774. https://doi.org/10.1109/TCOMM.2022.3206884

[7] Su, Y., Pang, X., Chen, S., Jiang, X., Zhao, N., & Yu, F. R. (2022). Spectrum and energy efficiency optimization in IRS-assisted UAV networks. *IEEE Transactions on Communications*, *70*(10), 6489-6502. https://doi.org/10.1109/TCOMM.2022.3201122

[8] Wang, W., Tian, H., & Ni, W. (2021). Secrecy performance analysis of IRS-aided UAV relay system. *IEEE Wireless Communications Letters*, *10*(12), 2693-2697. https://doi.org/10.1109/LWC.2021.3112752

[9] Al-Jarrah, M., Al-Dweik, A., Alsusa, E., Iraqi, Y., & Alouini, M. S. (2021). On the performance of IRS-assisted multi-layer UAV communications with imperfect phase compensation. *IEEE Transactions on Communications*, *69*(12), 8551-8568. https://doi.org/10.1109/tcomm.2021.3113008

[10] Solanki, S., Park, J., & Lee, I. (2022). On the performance of IRS-aided UAV networks with NOMA. *IEEE Transactions on Vehicular Technology*, *71*(8), 9038-9043. https://doi.org/10.1109/TVT.2022.3171271

[11] Cai, Y., Wei, Z., Hu, S., Liu, C., Ng, D. W. K., & Yuan, J. (2022). Resource allocation and 3D trajectory design for power-efficient IRS-assisted UAV-NOMA communications. *IEEE Transactions on Wireless Communications*, *21*(12), 10315-10334. https://doi.org/10.1109/TWC.2022.3183300

[12] Liu, Z., Zhao, S., Wu, Q., Yang, Y., & Guan, X. (2021). Joint trajectory design and resource allocation for IRS-assisted UAV communications with wireless energy harvesting. *IEEE Communications Letters*, *26*(2), 404-408. https://doi.org/10.1109/LCOMM.2021.3128545

[13] Asim, M., ELAffendi, M., & Abd El-Latif, A. A. (2022). Multi-IRS and multi-UAV-assisted MEC system for 5G/6G networks: Efficient joint trajectory optimization and passive beamforming framework. *IEEE Transactions on Intelligent Transportation Systems*, *24*(4), 4553-4564. https://doi.org/10.1109/TITS.2022.3178896

[14] Zhang, X., Wang, J., & Poor, H. V. (2022). Joint optimization of IRS and UAV-trajectory: For supporting statistical delay and error-rate bounded QoS over mURLLC-driven 6G mobile wireless networks using FBC. *IEEE Vehicular Technology Magazine*, *17*(2), 55-63. https://doi.org/10.1109/MVT。59660.88868886666

[15] Saxena, P., & Chung, Y. H. (2023). Analysis of jamming effects in IRS assisted UAV dual-hop FSO communication systems. *IEEE Transactions on Vehicular Technology*, *72*(7), 8956-8971. https://doi.org/10.1109/TVT.2023.3246817

[16] Zargari, S., Hakimi, A., Tellambura, C., & Herath, S. (2022). User scheduling and trajectory optimization for energy-efficient IRS-UAV networks with SWIPT. *IEEE Transactions on Vehicular Technology*, *72*(2), 1815-1830. https://doi.org/10.1109/TVT.2022.3207700

[17] Wang, Y., Liu, R., Yuan, J., Lu, J., Wang, Z., Wu, R., et al. (2024). Performance Analysis of UAV-IRS Relay Multi-Hop FSO/THz Link. *Electronics*, *13*(16), 3247. https://doi.org/10.3390/electronics13163247

[18] Ahmed, M., Alshahrani, H. M., Alruwais, N., Asiri, M. M., Al Duhayyim, M., Khan, W. U., & Nauman, A. (2023). Joint optimization of UAV-IRS placement and resource allocation for wireless powered mobile edge computing networks. *Journal of King Saud University-Computer and Information Sciences*, *35*(8), 101646. https://doi.org/10.1016/j.jksuci.2023.101646

[19] Wang, X. (2024). Edge Computing Based Multi-Objective Task Scheduling Strategy for UAV with Limited Airborne Resources. Informatica, 48(2), 255-268. https://doi.org/10.31449/inf.v48i2.5885

[20] Zhang, G., & Zhang, J. (2025). High-precision photogrammetric 3d modeling technology based on multi-source data fusion and deep learning-enhanced feature learning using internet of things big data. Informatica, 49(11), 1-16. https://doi.org/10.31449/inf.v49i11.7137

[21] Pang, X., Mei, W., Zhao, N., & Zhang, R. (2022). Intelligent reflecting surface assisted interference mitigation for cellular-connected UAV. *IEEE Wireless Communications Letters*, *11*(8), 1708-1712. https://doi.org/10.1109/LWC.2022.3175920

AP-Traj2: Transformer-based Trajectory Prediction with Graph-Enhanced Attention Mechanism

Ana-Paula Galarreta[1], Hugo Alatrista-Salas[1], Miguel Nunez-del-Prado[2], Vincent Gauthier[3]
[1]Pontificia Universidad Católica del Peru
[2]Universidad Peruana de Ciencias Aplicadas, Peru
[3]SAMOVAR, Telecom SudParis, Institut Polytechnique de Paris, France
E-mail: a.galarreta@pucp.pe, halatrista@pucp.pe, miguel.nunez@upc.pe, vincent.gauthier@telecom-sudparis.eu

Keywords: Trajectory prediction, deep learning, graph, transformers, whereabout prediction, neural network

Received: February 27, 2024

Trajectory prediction is essential for understanding human mobility patterns, with applications such as itinerary recommendation and urban planning. It involves analyzing sequences of visited locations to forecast the user's next destination. Traditional approaches have often relied on Markov chains or recurrent neural networks (RNNs). More recently, Transformer neural networks have gained attention for sequential prediction tasks due to their superior parallelization and training efficiency. In this study, we propose AP-Traj2 (Attention and Possible directions for TRAJectory prediction 2), a model designed to enhance prediction accuracy by leveraging attention mechanisms and graph-based movement modeling. AP-Traj2 employs self-attention to capture dependencies among visited locations, explores feasible next steps through a graph of possible directions, and incorporates contextual information via location embeddings. Experiments conducted on GPS, CDR, and WiFi datasets demonstrate that AP-Traj2 improves the average match ratio by approximately 50% *over state-of-the-art methods. Moreover, it achieves significantly faster training times, with reductions of up to* 72% *in the best-case scenario. Unlike existing approaches that focus primarily on neural network architecture, this work emphasizes the importance of data preprocessing and filtering, highlighting their substantial impact on model performance.*

Povzetek: Študija predstavi AP-Traj2, ki z združitvijo samo-pozornosti, grafa možnih smeri in vgrajevanj lokacij izboljša natančnost in učinkovitost napovedovanja naslednje destinacije iz zaporedij lokacij.

1 Introduction

Spatio-temporal data provide valuable insights into complex phenomena involving both spatial and temporal dynamics. Today, the availability of GPS data allows us to study a wide variety of trajectories. In addition, other sources such as Call Detail Records (CDRs) and Wi-Fi data could also be used to generate trajectories and make predictions about the whereabouts of specific devices [1, 2]. Previous research has focused on predicting the future positions of various entities such as vehicles, animals, and humans [3, 4, 5, 6, 7].

Traditionally, RNN [8, 9, 10, 11] models have been used for trajectory analysis, where the sequence of positions is considered. Some works have addressed the issue of the time required to generate models for predicting the trajectories of moving objects (see Table 1). Efficient training helps to reduce costs and enables organizations to take on more significant challenges associated with the study of moving objects.

This paper aims to predict the next discrete locations with improved accuracy and reduced training times when compared to the state-of-the-art. Hence, the present effort presents the Attention and Possible directions for TRAJectory prediction 2 (AP-Traj2), a model designed to predict the next position in a discrete geographic system. In this approach, a user's trajectory coordinates are first transformed into a sequence of nodes on a directed graph. The prediction of the next nodes a user is likely to visit is based on the directed graph, the sequence of previously visited nodes and a sequence of previous directions taken. In our study, we used datasets from different sources and extracted location information in different ways. Specifically, for GPS data, the nodes in the graph correspond to road intersections. In the case of CDR data, each node is associated with the region surrounding a particular antenna, *i.e.* lower granularity is used compared to GPS. Similarly, for Wi-Fi data, each node represents a specific building. Using this information, AP-Traj2 uses three modules to predict a user's future whereabouts: a node self-attention module, a possible directions module, and a location embedding module. More in detail, we try to understand how does adding a self-attention mechanism improve trajectory prediction over existing RNN-based approaches?

The node self-attention module uses the self-attention mechanism in the Transformer's encoder to capture the dependencies between nodes. This allows the model to effectively capture the contextual relationships between vis-

ited locations and use them for trajectory prediction. The possible directions module uses the graph representation to determine the feasible paths from an original node to its neighboring nodes, facilitating the prediction of the user's movement. Finally, the location embedding module captures information about the general area previously visited by the user, providing additional context for trajectory prediction.

Experiments conducted on GPS, CDR, and WiFi datasets demonstrate that AP-Traj2 improves the average match ratio by approximately 50% over state-of-the-art methods. Moreover, it achieves significantly faster training times, with reductions of up to 72% in the best-case scenario. Our approach offers the advantage of faster training, making it a suitable solution for trajectory prediction tasks in the context of discrete geographic systems.

The following sections of this article are organized as follows. Section 2 provides a comprehensive review of scientific studies on trajectory prediction. Section 3 provides a detailed explanation of the proposed model, AP-Traj2. Section 4 describes the methodology used in this study. Section 5 details the experimental setup and results. Finally, Section 7 concludes with a summary of the findings and suggestions for future research directions.

2 State of the art

Next whereabouts prediction has been widely studied in the literature. However, capturing and modeling spatial dependency remains an unsolved challenge [20]. Several models have been proposed for predicting the next item in a sequence. Table 1 provides a summary of such models, which mainly rely on deep learning methods and show varying performance and limitations, as discussed below. In particular, convolutional neural network (CNN) models [14], recurrent neural network (RNN) models [8, 9, 10, 12, 13, 11], and Transformer-based models [15, 16, 17, 18, 19] have been used.

It is worth noting that most studies rely on GPS datasets, with some using Automatic Identification System (AIS) data [19, 12] or sensor data (e.g., highway sensors) [15]. Additionally, the evaluation metrics used vary significantly across studies, making direct comparisons challenging. Moreover, training times are rarely reported, despite their relevance for practical applications. In this work, we aim to improve prediction accuracy while maintaining training efficiency, which we will explicitly measure.

Since Transformers have demonstrated to give good results while achieving faster training times, our architecture employs the Transformer's encoder and a graph of the visited area to predict the next location that a user is going to visit.

While Table 1 shows steady advances in next-place prediction, it also highlights some key limitations. First, many existing models do not explicitly integrate both spatial and temporal constraints, treating them as separate signals or ignoring one altogether. Second, most rely on architectures that require extensive training time or lack personalization, limiting their scalability and adaptability. Third, although some models incorporate graph structures or attention mechanisms, these are not always aligned with user-specific trajectories. These gaps motivate the design of **AP-Traj2**, which combines spatial and temporal restricted sequential patterns with the efficiency of Transformer encoders and contextual graph representations, to achieve accurate, efficient, and personalized predictions.

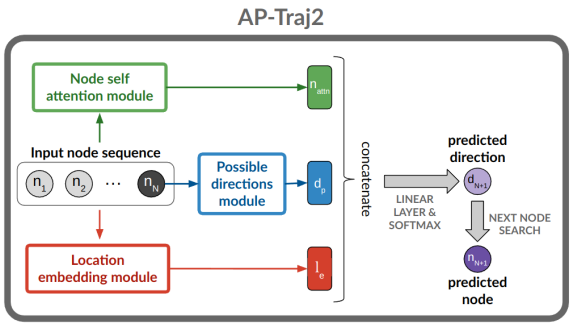


Figure 1: Attention and Possible directions for Trajectory prediction 2 (AP-Traj2). It contains three modules: Node self-attention module, Possible directions module and Location embedding module. Each of them produces an output vector, and these vectors are then concatenated and processed to predict the user's next visited node.

3 The AP-Traj2 model

The present section details *AP-Traj2*, the new extension for the **A**ttention and **P**ossible directions for **Traj**ectory prediction model (AP-Traj) [21]. One key improvement is how location information is represented. While AP-Traj uses the embedding of the last node in the input sequence, AP-Traj2 incorporates a dedicated *Location embedding module*, designed to provide a richer representation of the user's trajectory.

To predict the next location, the proposed model takes a sequence of N nodes as input and consists of three modules, as shown in Figure 1. The first module, in green, is the *Node self-attention module*, which uses the Transformer's encoder to capture the importance of the input nodes. The second module, in blue, is the *Possible directions module*, which is responsible for identifying the possible directions that can be taken from the last node of the input sequence nodes using an annotated graph that contains information about the area traveled. Finally, the *Location embedding module*, in red, combines the embeddings of the input node sequence, generating a vector that encapsulates information about the location of the user's trajectory. Each module generates an output vector, which is combined and processed to predict the user's next visited node. These three modules are described in detail in the following sections.

Table 1: Summary of models, datasets, and performance metrics

	Name	Model	Dataset	Dataset size	Technology	Metric	Value	Training time
[8]	DeepMove	LSTM	Foursquare check-in NYC Foursquare check-in Tokio Mobile application location Call Detail Records	82K records 537K records 15M records 491K records	GPS GPS GPS CDR	Top-1 PA* Top-1 PA Top-1 PA Top-1 PA	0.167 0.206 0.694 0.593	not reported
[9]	DeNavi	LSTM + EWMA	Gowalla check-in BrightKite check-in	6.4 M interactions 4.5 M interactions	GPS GPS	ACC@20 ACC@20	0.2460 0.5831	not reported
[10]	CLNN	LSTM + MLP	Porto taxi trajectory Geolife	1.7 million records 18,670 trajectories	GPS GPS	MHD error MHD error	1.6864 2.9401	not reported
[12]	—	GRU	AIS data - China	7.5M points	AIS	accuracy	0.98	135.97 (units unknown)
[13]	—	LSTM+GCNN	Didi - China	2M records	GPS	MAE	1.201	not reported
[11]	NetTraj	LSTM + ATTN	T-Drive Shanghai Trajectories	15M records used 4M records	GPS GPS	AMR AMR	69.7 65.8	3,61 h/iter 2,75 h/iter
[14]	—	CNN	HRI Driving Dataset	43 hours	GPS	Precision	0.94	
[15]	Space timeformer	Transformer	Metr-LA Traffic Pems-Bay Traffic NY-TX Weather AL Solar	34k time steps 52k time steps 570K time steps 52k time steps	sensor sensor sensor sensor	MSE MSE MSE MSE	38.27 13.99 12.49 7.75	not reported
[16]	Forecaster	Transformer + graph	NYC taxi dataset	65.4M datapoints	GPS	RMSE	5.19	not reported
[17]	PreCLN	Transformer	Porto taxi trajectory T-Drive	1.7 million records 17M records	GPS GPS	MAE MAE	193.15 55.90	not reported
[18]	—	Transformer	Green Class Yumuv	139 participants 498 participants	GPS GPS	ACC@10 ACC@10	73.50 66.92	not reported
[19]	TrAISformer	Transformer	AIS data - Denmark	712M records	AIS	mean error (nmi)	0.48	60 min/10 days of data

3.1 Node self-attention module

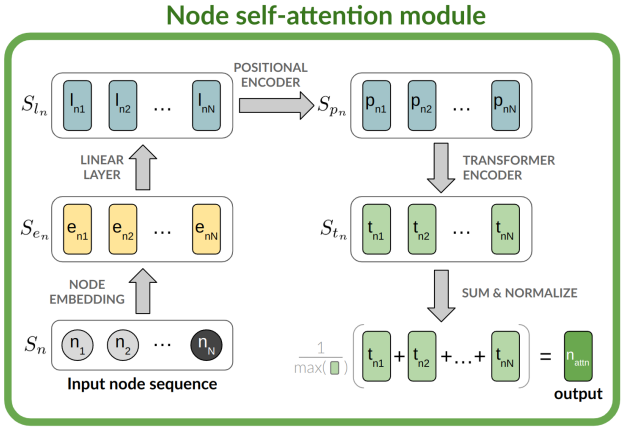


Figure 2: Node self-attention module schema

As shown in Figure 2, this module takes as input a sequence of nodes $S_n = \{n_1, ..., n_N\}$ and generates the vector $n_{attn} \in \mathbb{R}^{1\times d}$, which captures the importance of each node in the input sequence.

To this aim, the first step is to compute the embedding for each unique node, shown as node embedding in Figure 2. This step is done by learning a parameter matrix $W_{n_e} \in \mathbb{R}^{|N_T|\times d}$, where $|N_T|$ is the total number of nodes present in the dataset and d denotes the dimensions of the output vector n_{attn}. The result is a sequence of embeddings denoted by $S_{e_n} = \{e_{n_1}, e_{n_2}, ..., e_{n_N}\} \in \mathbb{R}^{N\times d}$, where N is the number of nodes in the input sequence. Note that N and $|N_T|$ usually have different values. For example, the input sequence might consist of $N = 10$ nodes, while there might be $|N_T| = 50$ different nodes in the dataset used to obtain the sequences.

Next, a linear layer is applied to S_{e_n} to adjust the magnitude of each vector within it. This is done by learning a parameter matrix $W_{l_n} \in \mathbb{R}^{d\times d}$, which ensures that the vectors are not heavily influenced by the subsequent positional encoding. This results in the sequence $S_{l_n} = \{l_{n_1}, l_{n_2}, ..., l_{n_N}\} \in \mathbb{R}^{N\times d}$.

Then, following the approach introduced by Vaswani et al. [22], a positional encoder is used to preserve the sequential order information of the input sequence. This involves augmenting each l_{n_i} with a corresponding position vector, resulting in the sequence $S_{p_n} = \{p_{n_1}, p_{n_2}, ..., p_{n_N}\} \in \mathbb{R}^{N\times d}$. Then, the sequence S_{p_n} is passed through the Transformer's encoder. This encoder uses a self-attention mechanism [23] that establishes relationships between different positions within the same sequence to generate a meaningful representation of that sequence. This encoder is constructed as a stack of identical layers, with each layer consisting of two subcomponents: a multi-head attention mechanism and a simple, position-wise, fully connected feedforward network. To improve the training of deep networks, a residual connection is built around each of these subcomponents. This residual connection acts as a shortcut, allowing the network to bypass one or more layers in both forward and backward passes. Then, a normalization layer is applied, resulting in a transformed representation denoted as $S_{t_n} = \{t_{n_1}, t_{n_2}, ..., t_{n_N}\} \in \mathbb{R}^{N\times d}$.

Finally, the outputs of the Transformer's encoder are

added and normalized to produce the output vector:

$$n_{attn} = \frac{1}{max(t_{n_i})} \sum_{i=1}^{N} t_{n_i}, n_{attn} \in \mathbb{R}^{1\times d}, \quad (1)$$

where n_{attn} is a vector that represents the significance of the nodes from input sentence.

3.2 Possible directions module

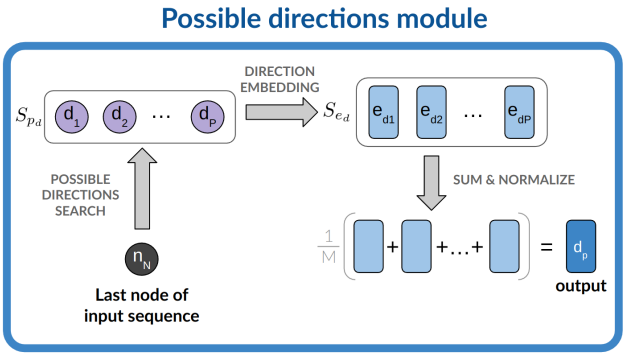


Figure 3: Possible directions module schema

This module, illustrated in Figure 3, focuses on the last node of the input sequence (n_N) and generates $S_{p_d} \in \mathbb{R}^{1\times d}$, which encapsulates information about the possible directions to follow from the last node.

To achieve this, an annotated directed graph $G_a = (V, E_a)$ is used, where:

$$V = \{n_1, n_2, \ldots, n_k\}$$
$$E_a = \{(n_1, n_2, dir_{12}), (n_2, n_3, dir_{23}), \ldots\}$$

Here, V is the set of nodes representing the elements in the input sequence and E_a is the set of directed edges, where each edge, has a direction dir_{ij} represented by an integer number representing the direction ranging from node n_i to node n_j. This direction dir_{ij} is obtained by following [11], where the angle θ_{ij} between the coordinates of nodes n_i and n_j is calculated. Since this angle can take any continuous value in the range $[0, 360[$, it has been discretized by dividing the space into K equal intervals, so that each has the same width $W = 360/K$.

Hence, for nodes $n_i = \{x_i, y_i\}$ and $n_j = \{x_j, y_j\}$:

$$\theta_{ij} = tan^{-1}\left(\frac{y_j - y_i}{x_j - x_i}\right) \times \frac{180}{\pi}$$
$$dir_{ij} = \lfloor \frac{\theta_{ij}}{360/K} \rfloor$$

For example, if $K = 8$, there are 8 possible directions, where 0 represents a direction between the east and the northeast.

Using G_a, the first step is to search for the edges connected to the last node to identify all possible directions that can be taken from it. This way, we obtain $S_{p_d} = \{d_1, d_2, ..., d_P\}$, where P is the number of possible directions.

Then, the embeddings of the possible directions are computed using a parameter matrix $W_{d_e} \in \mathbb{R}^{K\times d}$. As a result, for any given node we obtain a sequence of direction embeddings denoted by $S_{e_d} = \{e_{d1}, e_{d_2}, ..., e_{dP}\} \in \mathbb{R}^{P\times d}$.

The number of possible directions P varies from node to node, as it represents the number of neighboring nodes of each node.

Finally, the resulting vectors are summed and normalized to produce the output vector:

$$d_p = \frac{1}{max(e_{di})} \sum_{i=1}^{P} e_{d_i}, d_p \in \mathbb{R}^{1\times d}, \quad (2)$$

where e_{d_i} is the embedding of each direction and d_p consolidates all direction embeddings into a single vector.

3.3 Location embedding module

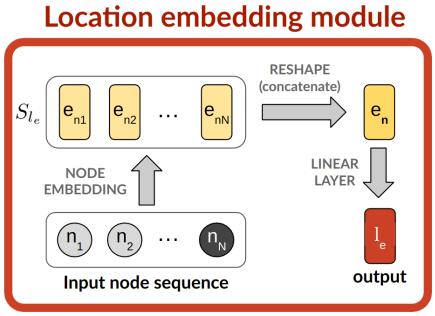


Figure 4: Location embedding module schema

A third module produces the vector $l_e \in \mathbb{R}^{1\times d}$ to store information about the area where the user moves. To obtain this vector, we use the same parameter matrix $W_{n_e} \in \mathbb{R}^{|N_T|\times d}$ used in the *Node self-attention module*. This matrix is utilized to compute the embeddings of all the input nodes to produce a sequence of location embeddings $S_{l_e} \in \mathbb{R}^{N\times d}$, as shown in Figure 4. Then, these embeddings are concatenated to form the vector $e_n \in \mathbb{R}^{1\times(N\times d)}$. Finally, a linear layer is applied to e_n to obtain the vector $l_e \in \mathbb{R}^{1\times d}$ by learning a parameter matrix $W_{l_l} \in \mathbb{R}^{(N\times d)\times d}$).

3.4 Next node prediction

After obtaining the outputs from the three modules, as shown in Figure 1, the resulting vectors are concatenated:

$$d_C = [n_{attn}; d_P; l_e], \quad d_C \in \mathbb{R}^{1\times(3d)} \quad (3)$$

The resulting vector d_C is then passed through a linear layer (by learning a parameter matrix $W_f \in \mathbb{R}^{3d\times K}$), followed by the application of the softmax function. This operation yields the probabilities of the K possible directions from the input sequence's last node (n_N).

Once the direction with the highest probability (d_{prob}) is computed, it is compared to the possible directions from n_N. This is done by searching the edges connected to this node in the graph G_a to find the closest match based on the difference between d_{prob} and all the possible directions. This process identifies the predicted direction (d_{N+1}). To break ties, the one with the lower value is selected. For example, if $d_{prob} = 2$ and the possible directions are 1 and 3, the predicted direction would be $d_{N+1} = 1$ because 1 has the lowest value ($1 < 3$). Finally, the predicted node (n_{N+1}) is obtained by using the predicted direction (d_{N+1}) and G_a.

3.5 Prediction of subsequent nodes

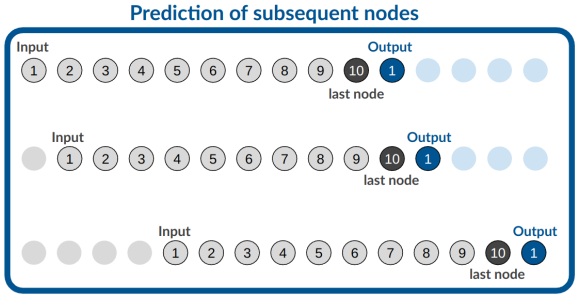


Figure 5: Prediction of subsequent nodes. When a node is predicted, it is concatenated with the last nodes of the original sequence to create a new input sequence.

To extend the prediction to multiple nodes instead of just one, a sliding window method is used. After predicting the node n_{N+1} using the input sequence S_{n_1}, a new input sequence S_{n_2} is constructed:

$$S_{n_1} = \{n_1, n_2, ..., n_N\} \quad (4)$$

$$S_{n_2} = \{n_2, n_3, ..., n_{N+1}\} \quad (5)$$

The sequence S_{n_2} is then fed into the above architecture, and the node n_{N+2} is predicted. This process is repeated for M iterations, where M is the length of the output sequence and the number of nodes predicted.

An example can be seen in Figure 5, where an input sequence of 10 nodes is used to predict a single output node. Then, the last 9 nodes of the original sequence are concatenated with the first predicted node to produce a new input sequence. This prediction process is then repeated to obtain the subsequent nodes.

4 Methodology

The proposed methodology, shown in Figure 6, transforms GPS, Wi-Fi, or CDR data into a format suitable for the AP-Traj2 model to be trained. It consists of three main stages: graph construction and annotation, sequence preprocessing, and model training. The **graph construction and annotation** phase generates an annotated directed graph $G_a = (V, E_a)$ that captures the connectivity between all the nodes the users can visit. This graph is then used in the **sequence preprocessing** stage, where the sequences of locations (represented as local or absolute coordinates) are transformed into sequences of nodes and directions. Next, this information is used in the **model training** stage to predict the subsequent nodes to be visited by a user.

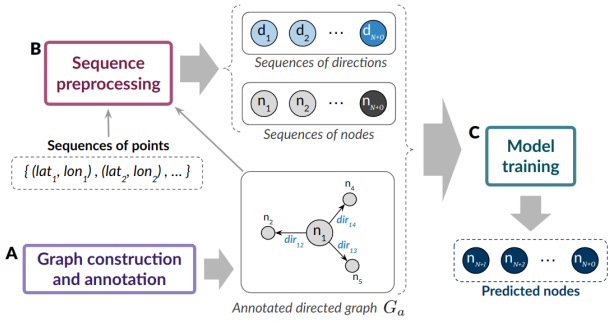


Figure 6: Proposed methodology for transforming input data and training the AP-Traj2 model. It has three main stages: graph construction, sequence preprocessing, and model training.

The following paragraphs describe this process in detail.

4.1 Graph construction and annotation

The first step of the preprocessing phase is to obtain an annotated and directed graph $G_a = (V, E_a)$, which contains information about the connections between all the locations available in each dataset. The method used to construct the graph depends on the characteristics of the dataset. However, the procedure used to annotate it is the same for all of them.

This step consists in obtaining a directed graph $G = (V, E)$, where:

$$V = \{n_1, n_2, \ldots, n_k\}$$
$$E = \{(n_1, n_2), (n_2, n_3), \ldots\}$$

In the case of GPS data, the graph G, representing the road network, can be obtained directly from an external source like *Open Street Maps* (OSM), with road intersections as nodes and road segments as edges. However, for other trajectory datasets, such as those containing Wi-Fi or CDR data, where it is necessary to construct the graph G, from the data itself, showing the connections between each of the specific locations visited (e.g., an administrative division or the coverage area of an antenna). In this case, we will construct the graph G taking into account the proximity between these locations. To achieve this, given a set of locations with their coordinates, Voronoi regions are obtained by using the Python library *GeoVoronoi* [24], which computes the region around each location that is closer to that location than any other one. Then, an unweighted graph is constructed using *NetworkX* [25], a Python package for

network analysis. To achieve this, the adjacency between regions is defined by Queen contiguity, which means that two regions are considered neighbors if they share at least one edge or one vertex. Once the adjacency relationships between the regions are defined, an unweighted graph is created by representing each region as a node and connecting adjacent regions with edges. In this case, edges are considered to be bidirectional (*c.f.*, Figure 7).

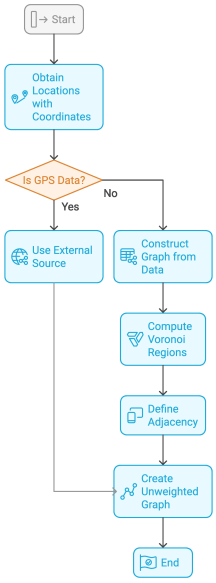


Figure 7: Graph construction flowchart

The graph $G = (V, E)$ is annotated to create the graph $G_a = (V, E_a)$. In G_a each edge e_{ij} stores the direction dir_{ij} taken when moving from node n_i to node n_j. As previously explained in Section 3.2, these directions are determined by calculating the angle between connected nodes and then discretizing this angle with a constant K parameter.

It is important to clarify that the edges in the annotated graph $G_a = (V, E_a)$ carry directional weights based on the discretized movement directions between nodes, which are used as features during model training. These direction values effectively serve as edge weights.

Additionally, for all datasets considered, the constructed graph fully covers all locations present in the trajectory data. There are no missing nodes, as each observed location corresponds to a node in the graph. For GPS data, nodes correspond to intersections obtained from external road network sources. For Wi-Fi and CDR datasets, nodes represent coverage areas or regions constructed via Voronoi tessellation, ensuring complete representation of all visited locations.

4.2 Sequence preprocessing

As shown in Figure 6, once we have obtained the graph G_a, we preprocess the sequences of locations to collect a sequence of nodes for each trajectory and its corresponding sequence of directions. This phase consists of four steps: Trajectory identification, Sample selection, Edge matching and Direction search.

The specific procedures used to collect this information depend on the characteristics of the dataset and the data available. In the case of GPS data, the inputs consist of a sequence of points represented as absolute coordinates and a graph representing road segments. For other types of data, however, the inputs consist of a sequence of points represented by specific regions, a table containing the absolute or relative coordinates associated with each region, and a graph showing the connections between the regions based on their proximity. Each of the four steps for sequence processing will be described in the following paragraphs.

4.2.1 Trajectory identification

The process has two stages. In the first stage, the sequences of points are broken down into smaller sequences. In the second stage, trajectories are selected based on whether they are located within a specific geographic region.

a. Sequence split. To identify stops and trajectories for GPS data, the Infostop algorithm [26] is used. It works by first dividing the data into segments and determining whether they represent a stop or just a temporary halt. To classify the points as trips or stays, Infostop uses three main parameters (1) the maximum roaming distance allowed between two points within the same stay (r_1), (2) the minimum time difference between two consecutive records for them to be considered within the same stay (t_{min}), and (3) the maximum time difference between two consecutive records to be considered within the same stay (t_{max}).

For example, given a sequence of points $seq = \{p_1, p_2, p_3, p_4, p_5\}$. Each point is tagged as a trip tr or a stay st $\{p_1 \rightarrow tr, p_2 \rightarrow tr, p_3 \rightarrow tr, p_4 \rightarrow st, p_5 \rightarrow tr\}$ In this case, for the first 5 points we would have two trajectories $traj_1 = \{p_1, p_2, p_3, p_4\}$ and $traj_2 = \{p_4, p_5\}$, which are sequences of points that represent the change in position of an object.

Note that we keep the stationary points (points tagged as stays) and use them as the start or end point of each trajectory. This is why the stationary point p_4 appears twice in the previous example (once in each trajectory).

Furthermore, since the experiments are performed with an input sequence of size N and an output sequence of size M, only the sequences of size $N + M$ are kept. This is done to preserve a larger amount of information associated with the original sequence.

b. Trajectory filtering. Some datasets contain trajectories that span over very large geographic regions. This can happen, for example, if the points within the dataset come from two different cities. In this case, the density of points for a given area is greatly reduced,

which reduces the quality of the training process. To avoid this, if the obtained trajectories come from different geographic regions, only those that are within a certain boundary are selected. This boundary could be a specific geographic region, such as a city.

4.2.2 Sample selection

To increase the number of experiments with each dataset and to test more hyperparameters during the tuning phase, a representative subset is selected by creating bounding boxes that cover the geographic region where the trajectories are located, such as a particular city. Next, the samples located within each bounding box are extracted and their inter-event distance and time interval distributions are compared to the entire population. The representativeness of the sample is then tested using the Chi-square goodness-of-fit test, which determines whether the sample is a good fit to the population. Acceptance of the null hypothesis indicates that the sample is representative. Finally, if multiple samples meet the conditions, the one with the highest number of data points is selected.

4.2.3 Edge matching

The AP-Traj2 method consists of predicting subsequent nodes, given a user's trajectory. This means that each sequence of points must be matched with a specific sequence of nodes. Two different methods are used to achieve this, depending on how the input trajectory is represented:

a. Points represented by absolute coordinates. If the input is a sequence of points represented as absolute coordinates, each point is matched to an edge of G_a using the ST-Matching tool [27] to transform each sequence of points into a sequence of road segments:

$$S_e = \{e_{01}, e_{12}, ..., e_{(N-1)N}\} \tag{6}$$

Then, considering that we will train the model with an input sequence of size N and an output sequence of size M, all sequences containing less than $N + M$ nodes are discarded to ensure a sufficient amount of information for the training stage.

b. Points represented by a region. If the input trajectory consists of a sequence of points represented by a region related to certain absolute or relative coordinates, each of these points is considered a node. Then, successive nodes are identified as stops, and the sequences are split into trajectories. Finally, Dijkstra's algorithm is used to find the shortest path between the nodes by using G_a, thus determining the user's trajectory.

It is worth noting that in cases where the user's location data is incomplete, our approach can still generate a trajectory that includes neighboring nodes. This can be achieved using either the ST-Matching tool or Dijkstra's algorithm.

4.2.4 Direction search

To increase the efficiency of the proposed model by limiting the number of possible predictions, we predict only one of the K available directions that can be taken from a given node.

Using G_a, all the information about a user's trajectory is represented as a sequence of nodes (S_n) and a sequence of directions (S_d) in the following way:

$$S_n = \{n_0, n_1, n_2, ..., n_N\} \tag{7}$$

$$S_d = \{dir_{01}, dir_{12}, ..., dir_{(N-1)N}\} \tag{8}$$

Note that S_n has one more element than S_d, since $dir_{(i-1)i}$ indicates the direction taken to go from n_{i-1} to n_i.

4.3 Model training

To generate the input and output sequences for the training phase, the initial node in S_n is excluded. Then, the S_n and S_d sequences are partitioned into shorter sequences of length of $N + M$, where N is the input size and M is the output size. For example, if the input size is $N = 10$ and the output size is $M = 5$, this means that 10 nodes and their corresponding directions are going to be used to predict the subsequent 5 nodes and directions. Hence, the original sequences must be partitioned into sequences of length $N + M = 15$.

Then, the generated sequences and the graph G_a are used to train the **AP-Traj2** model. The following section provides a detailed description of the experiments performed.

5 Experiments

Dataset characteristics, graph generation, sequence preprocessing, evaluation metrics, and model training are described in this section.

5.1 Datasets

In the present effort, the datasets for the experiments are: (1) **T-Drive**, which contains GPS trajectories of 9,657 taxis in Beijing, China for six days in February 2008 [28]. A total of 128,074 individual sequences representing trajectories were extracted. (2) **Geolife**, having 18,670 GPS sequences representing the trajectories of 182 users from April 2007 to October 2011 in China [29]. (3) **D4D Ivory Coast** CDR dataset, consisting of 500,000 randomly selected trajectories in Ivory Coast between December 1, 2011 and April 28, 2012. For user privacy, location is reported at the subprefecture level, reducing spatial accuracy [30]. The total number of sequences representing trajectories is 4,860,992. (4) **D4D Senegal**, comprising country level CDR data for 2013 in Senegal. For privacy reasons, only a 2-week period of randomly selected users is reported [31]. In our study, only a subset of 319,508 trajectories of unique users was

used. (5) **KTH wireless (Stockholm)** dataset, which contains records of authenticated user connections to the wireless network at the KTH Royal Institute of Technology in Stockholm during 2014 and 2015 [32]. This dataset contains information on 109,183 users, and a total of 483,004 sequences representing trajectories were extracted.

The main characteristics of these datasets can be observed in Table 2, which shows the technology used to obtain the data, the number of users, the number of total sequences, the number of days observed, the average sequence length and its standard deviation.

5.2 Graph construction and annotation

As it was described in section 4.1, an annotated directed graph G_a containing information about the connections between all the locations in each dataset is needed. The construction and subsequent annotation of this graph is explained in the following lines.

5.2.1 Graph construction

In the case of the GPS datasets (T-Drive and Geolife), the Beijing road network obtained from OSM through OSMnx [33], was used. The resulting graph G contains 46,805 nodes and 110,268 edges, where each node represents a road intersection and each edge represents a road segment.

However, since the D4D Ivory Coast, D4D Senegal, and KTH wireless contain CDR and Wi-Fi data, each graph was constructed based on the proximity between the antenna location. For more details on this procedure, see section 4.1. The graph G created for the D4D Ivory Coast has 255 nodes and 2,824 edges, the one for D4D Senegal has 1,666 nodes and 9,630 edges, and the KTH wireless graph has 59 nodes and 648 edges.

5.2.2 Graph annotation

For all five datasets, the graphs G were annotated according to the procedure described in Section 4.1, in order to obtain G_a. A value of $K = 8$ was used for the T-Drive and Geolife datasets, while $K = 12$ was used for the D4D Ivory Coast, D4D Senegal, and KTH wireless datasets. These values were chosen based on the maximum number of directions that can be taken from a given node in each dataset. Our choice ensured that direction embeddings captured sufficient spatial variance without over-fragmenting the space. It's important to note that if a single node yields the same discretized direction for two neighbors, the duplicated direction obtained in second place is replaced with the next closest direction.

5.3 Sequence preprocessing

In this step, the obtained graph G_a and a sequence of points are processed to obtain the sequences of nodes and directions needed to train the model. As previously described in Section 4.2, the procedure varies depending on the dataset. In the following paragraphs, we will explain the exact steps taken for each individual dataset.

5.3.1 Trajectory identification

This process, described in Section 4.2.1, includes the sequence split and trajectory filtering steps:

a. Sequence split. This allows the identification of stops and trajectories on GPS datasets. This step was only performed on the T-Drive dataset, because it consists of taxi trajectories, but it does not contain an occupied flag, *i.e.* we have no information regarding whether a taxi driver is driving a passenger to a specific location or not. In total, 128,074 sequences representing taxi trajectories were extracted. However, this step was considered unnecessary for the Geolife dataset, which also contains GPS data, because the individual user trajectories were already separated. Finally, since the other three datasets used do not contain GPS data, this procedure was not performed on them.

b. Trajectory filtering. Considering that both the T-Drive and Geolife datasets cover a very large geographic region, mostly consisting on the Beijing area in China and its surroundings, the point density is reduced significantly, which decreases the quality of the training process. To avoid this, only points within the Beijing city area were considered. For the T-Drive dataset, about 94.5% of the trajectories were retained, while for the Geolife dataset, about 90% of the trajectories were retained. It is worth noting that this specific procedure was not applied to the other three datasets.

5.3.2 Sample selection

As it was explained previously in Section 4.2.2, to increase the number of experiments for the hyperparameter tuning, a representative subset of each dataset was selected based on how its inter-event distance and time interval distributions compare to those of the general population.

For the T-Drive, Geolife, D4D Ivory Coast and D4D Senegal datasets, a sample selection was performed based on the results of an inter-event distance analysis. However, due to the limited number of sequences in the KTH wireless dataset, this analysis and the subsequent sample selection were unnecessary, as all the sequences were used in the preprocessing steps.

The results of the inter-event analysis can be seen in Table 3, where we can see that the inter-event time and distance vary greatly in the analyzed datasets.

Based on the obtained distributions, we selected samples from the T-Drive, Geolife, and D4D Senegal datasets. The bounding boxes used to extract these samples are [39.855, 39.980, 116.342, 116.467] for T-Drive, [39.917, 40.017, 116.254, 116.354] for Geolife and [14.658, 14.788, -17.456, -17.326] for D4D Senegal.

Table 2: Dataset statistics

Metrics	T-Drive	Geolife	D4D Senegal	D4D Ivory Coast	KTH wireless
technology	GPS	GPS	CDR	CDR	Wi-Fi
num users	9,657	182	319,508	500,000	109,183
num seq	128,074	18,670	319,508	4,860,992	483,004
num days	7	1,645	365	150	713
seq len avg	126	1,332	139	152	193
seq len std	710	2,821	167	288	348

Table 3: Inter-event analysis

Metrics	T-Drive	Geolife	D4D Senegal	D4D Ivory Coast
time avg. (s)	441	7	8,166	7,567
time std. dev. (s)	3,699	191	21,425	25,683
distance avg. (m)	3,815	52	1853	1056
distance std. dev. (m)	183,632	5847	13,565	9,387

It is important to note that for the D4D Ivory Coast dataset, we initially performed an inter-event distance analysis, but ultimately decided not to select a sample. Instead, we decided to use the entire dataset for hyperparameters fine-tuning. This decision was driven by the small number of sequences obtained after the subsequent edge and node matching step.

5.3.3 Edge and node matching

As explained in Section 4.2.3, two different approaches can be used to transform a sequence of points into a sequence of nodes by using G_a. In the case of the Geolife and T-Drive datasets, the points are represented by absolute coordinates, so the ST-Matching tool was used to transform each sequence of points into a sequence of road segments.

On the other hand, since the points in the D4D Ivory Coast, D4D Senegal, and KTH wireless datasets are represented by regions, consecutive nodes were identified as stops. In cases where the raw data contained incomplete trajectories—i.e., users leaped between non-consecutive nodes in the constructed graph—Dijkstra's algorithm was applied to estimate the shortest path between observed locations and reconstruct a plausible trajectory.

5.3.4 Direction search

This process was performed for all datasets, using their corresponding sequences of edges S_e and annotated directed graphs G_a.

5.4 Evaluation metrics

To evaluate the effectiveness of AP-Traj2, the Match Ratio (MR) and the Average Match Ratio (AMR) were used. The Match Ratio in Equation 10 is calculated as the average of the matches between the predicted values and the ground truth:

$$MR(k) = \frac{1}{k}\sum_{i=1}^{k} match(y, \hat{y}), \tag{9}$$

where:

$$match(y, \hat{y}) = \begin{cases} 1, & \text{when } y = \hat{y} \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

On the other hand, the Average Match Ratio is defined by Equation 11, where O represents the number of predicted nodes and $MR(k)$ is the Match Ratio up to the k-th predicted node.

$$AMR = \frac{1}{O}\sum_{k=1}^{O} MR(k) \tag{11}$$

5.5 Model training

5.5.1 Hyperparameter tuning

In the case of the T-Drive, Geolife, and D4D Senegal datasets, representative sample bounding boxes were used to explore multiple hyperparameter values, as indicated in section 5.3.2. However, for the D4D Ivory Coast and KTH wireless datasets, the entire population was used.

Based on [11] the sequences for all datasets were divided into segments of length $N + M = 15$. Each segment had an input size of $N = 10$ nodes and their corresponding directions, and an output size of $M = 5$ nodes and directions. Although N is a tunable parameter, it was fixed to 10 during training to focus on tuning model-specific hyperparameters. The impact of varying N is analyzed separately in Section 5.5.2. Then, the obtained sequences were randomly divided into training, validation, and test sets. The number of sequences in each split for every dataset is shown in Table 4. All datasets were split using an 80% training, 10% validation, and 10% test ratio.

The main hyperparameters tuned for AP-Traj2 are the following:

Table 4: Hyperparameter tuning: number of sequences

Dataset	Training	Test	Validation
T-Drive	20,376	2,547	2,547
Geolife	24,219	3,027	3,027
D4D Senegal	24,478	3,060	3,060
D4D Ivory Coast	11,650	1,450	1,450
KTH wireless	24,510	3,065	3,065

Table 5: Hyperparameter tuning: search space

Parameter	Tested Values
Embedding dimension (d)	$\{64, 128, 256\}$
Feedforward dimension	$\{256, 512, 1024\}$
Number of heads	$\{2, 4\}$
Number of layers	$\{2, 4\}$
Dropout rate	$\{0.1, 0.2\}$
Initialization range	$\{0.1, 0.2\}$

- *dim embedding*. This parameter determines the embedding dimensions for the nodes and directions (d). It also corresponds to the number of expected features in the input of the Transformer's encoder.
- *dim feedforward*. This parameter is related to the Transformer's encoder.
- *num heads*. Returns the number of heads in the multi-head attention model.
- *num layers*. This is the number of sub-encoder layers in the Transformer's encoder.
- *dropout*. Corresponds to the dropout rate used to prevent overfitting. Two values were tested: 0.1 and 0.2, which correspond to deactivating 10% and 20% of the neurons during training, respectively.
- *init range*. A uniform distribution with a range from *-init range* to *init range* has been used used for weight initialization.

The embedding dimensions, Transformer encoder settings (number of layers, number of heads, and feedforward size), dropout rate, and initialization range were selected based on common practices. Specifically, they are either equal to or less than the values used in [22]. The specific values used for tuning are summarized in Table 5.

This decision was made to balance model performance and resource requirements, as larger models would require more memory and processing power. In addition, we conducted experiments with different dropout and initial range values to further fine-tune the hyperparameters of our model for each specific dataset. This allowed us to optimize the model's performance based on the data's characteristics.

The training was done using a Stochastic Gradient Descent (SGD) optimizer with a *batch size = 20*, *maximum number of iterations = 100* and early stopping with *patience = 1*, meaning that the training stops the first time the validation loss increases. The loss function used during training was the Cross-Entropy Loss, a common choice used not only in our baseline [11] but also in several other studies, including [8], [17], and [18]. It is important to note that the Cross-Entropy Loss was calculated by comparing the predicted direction to the ground truth, and not the predicted node. Although activation functions are not explicitly defined in custom layers, we have used PyTorch's implementation of `TransformerEncoderLayer`, which applies ReLU as the default activation function in the feed-forward sub-layer. No additional activation functions were added outside this component.

Table 6 shows the combinations of values that gave the best results when comparing the AMR with $k = M = 5$. It is important to note that each parameter combination was tested 10 times, resulting in 1440 trials for each dataset, for a cumulative total of 7200 trials.

5.5.2 Optimal input length

After tuning the hyperparameters, a larger sample was extracted from the T-Drive, Geolife, and D4D Senegal datasets. The bounding boxes for this larger sample are [39.817, 39.992, 116.304, 116.479] for Drive, [39.740, 40.050, 116.140, 116.600] for Geolife and [14.658, 14.808, -17.481, -17.331] for D4D Senegal. Five different lengths were tested using this larger sample to determine the optimal input length, along with the full D4D Ivory Coast and KTH wireless datasets.

To ensure consistency, the number of training, test, and validation sequences was controlled for each dataset. When testing different input lengths, the data was initially segmented using an input length of $L = 10$. To evaluate shorter input lengths (e.g., $L = 8$ or $L = 6$), nodes were removed from the beginning of these segments rather than reprocessing the entire dataset. This allowed us to keep the same number of training, validation, and test sequences across different values of L, ensuring a fair comparison while controlling for dataset size. The trials for each input length were run 10 times on each dataset. The number of training, test and validation sequences for each dataset is given in Table 7. Here, the number of D4D Ivory Coast and KTH Wireless sequences remains the same compared to Table 4 because all available sequences were used in the fine-tuning phase for these two datasets. However, there is a significant increase in the number of sequences for the T-Drive, Geolife, and D4D Senegal datasets.

No trade-off between training time and accuracy was observed in our experiments. All models were tuned to prioritize prediction accuracy, and training times remained competitive. While additional iterations could potentially improve results further, this was not required to reach the reported performance.

The results shown in Figure 8 indicate that the optimal input sequence length (L) for predicting future locations varies depending on the dataset, given the specific hyperparameters and training data size used. In the case of the

Table 6: Hyperparameter tuning: best values

Parameter	T-Drive	Geolife	D4D Senegal	D4D Ivory Coast	KTH wireless
dim embed.	256	256	256	256	256
dim feedf.	256	256	256	1024	512
num heads	4	2	4	2	4
num layers	4	4	2	4	2
dropout	0.2	0.2	0.2	0.1	0.2
init range	0.2	0.2	0.2	0.2	0.2

Table 7: Optimal input length: number of sequences

Dataset	Training	Test	Validation
T-Drive	134,967	16,871	16,871
Geolife	80,915	10,115	10,115
D4D Senegal	93,216	11,652	11,652
D4D Ivory Coast	11,650	1,450	1,450
KTH wireless	24,510	3,065	3,065

T-Drive dataset, we observe a consistent increase in AMR as the input sequence length increases, leading to the best performance when $L = 10$. However, the other datasets do not show the same pattern. For the Geolife dataset, the optimal performance is achieved with $L = 8$, while for the Stockholm dataset, it is achieved with $L = 6$. On the other hand, both the D4D Senegal and D4D Ivory Coast datasets give the best results when $L = 4$. Therefore, these results do not allow us to conclude that a larger input sequence consistently leads to better AMR, except in the case of the T-Drive dataset.

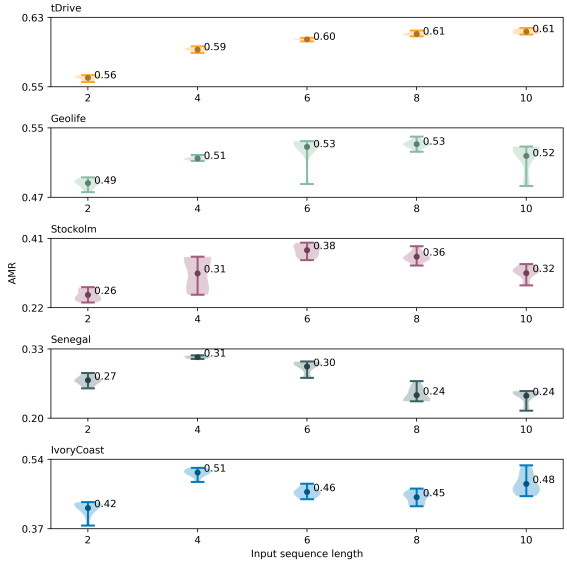


Figure 8: Average match ratio (AMR) versus input sequence length

5.5.3 Baseline comparison

To evaluate the effectiveness of AP-Traj2, a comparison with the NetTraj method [11] was performed. The NetTraj method was implemented to the best of our ability, without the use of contextual information (*i.e.*, weather conditions, season, etc.) and using the same hyperparameters as specified by the authors, except for the input sequence length. Instead, we used the previously determined optimal input lengths. It is important to note that for the NetTraj method, an inverse sigmoid decay schedule was used, and the decay rate was fine-tuned in the range of 1 to 12 with increments of 0.5. The results in Figure 9 show the significant performance advantage of the AP-Traj2 method over the baseline, especially for the GPS datasets such as T-Drive and Geolife. Also, we can see that the AP-Traj2 method has significantly faster training times compared to the baseline for the GPS datasets. Furthermore, the AP-Traj2 method also demonstrates shorter training times for the Wi-Fi and CDR datasets, which can potentially reduce computational costs.

In summary, across all five datasets, AP-Traj2 has superior AMR performance compared to Net-Traj, while achieving faster training times.

6 Discussion

The superior performance of AP-Traj2 across all five datasets, in terms of both prediction accuracy (AMR) and training time, highlights the advantages of the proposed architecture. While we compare directly to NetTraj, it is worth noting that NetTraj has itself been benchmarked against a wide array of state-of-the-art models, including LSTM encoder-decoder [34], Caser [35], AT-LSTM [36], ATST-LSTM [37], SASRec [38], and GeoSAN [39]. By outperforming NetTraj, AP-Traj2 demonstrates competitive performance relative to these methods as well.

Several architectural choices contribute to AP-Traj2's improvements. First, the incorporation of spatial and temporal restricted sequential patterns enables the model to capture fine-grained local dependencies that traditional RNN or attention-only architectures may overlook. Second, the use of a Transformer encoder improves efficiency and scalability, particularly in long-range sequence modeling, while avoiding the training overhead associated with recurrent layers. Finally, the integration of a contextual

graph constructed from the visited area provides additional structure and spatial awareness, which supports more informed predictions and helps reduce overfitting in data-sparse contexts.

Although we did not conduct an ablation study specifically for AP-Traj2, our previous work on AP-Traj [21] included such an analysis. That study demonstrated that the last node embedding, node self-attention, and possible directions modules each contribute significantly to performance, with the last node embedding having a particularly strong effect on prediction quality. Motivated by these findings, we focused on enhancing the last node embedding module in this paper, which led to the development of AP-Traj2.

We observe that the model performs better on GPS datasets compared to Wi-Fi and CDR datasets. This may be due to higher spatial granularity and denser sampling in GPS data, which allows the model to learn more informative patterns. In contrast, CDR and Wi-Fi data are sparser and use coarser spatial units (e.g., antennas or access points), potentially limiting prediction accuracy.

We also note that datasets with finer granularity benefit more from longer input sequences (Figure 8). This suggests that when the information per time step is more precise, the model can leverage additional temporal context to improve predictions.

We have not explicitly studied the impact of data sparsity—i.e., how performance changes when fewer data points or trajectories are available. Investigating AP-Traj2's robustness to sparse data, possibly through data augmentation or regularization techniques, is an important direction for future work.

The model's reliance on historical trajectories with consistent spatial patterns may hinder its performance in highly irregular or noisy datasets. Additionally, the graph construction and pattern extraction steps require offline preprocessing, which may limit applicability in real-time scenarios or systems with constrained latency requirements. Future work may explore strategies to incrementally update graph representations and adapt to unseen locations more effectively.

Although we have not explicitly evaluated AP-Traj2 on datasets larger than T-Drive, the approach's design principles and its performance on existing data indicate potential for scalability. Nevertheless, further research is needed to assess its generalizability and efficiency in real-time or large-scale applications.

While our model showed strong results across datasets with varying spatial resolutions and densities, further evaluation is needed to assess its behavior in extremely large geographic areas or with denser trajectory data. Although our approach constrains candidate movements through a fixed number of discretized directions, which helps maintain efficiency even at scale, graph annotation and node retrieval times may increase proportionally with the number of nodes. Future work may explore optimization strategies for preprocessing and search to better accommodate larger graphs without impacting overall performance.

While AP-Traj2 achieves competitive accuracy without sacrificing training efficiency, we did not explicitly explore the trade-off between training time and accuracy. Our hyperparameter tuning prioritized performance, and no simplifications were made for faster training. It remains possible that longer training or further optimization could improve results even further.

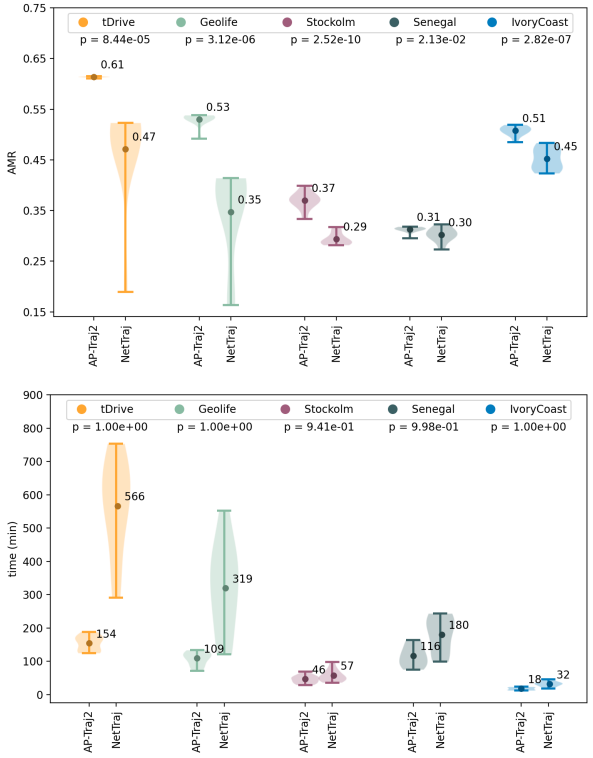


Figure 9: Average match ratio (AMR) and total training time in minutes (min) for Net-Traj and AP-Traj2

7 Conclusion

This research paper presents AP-Traj2 (Attention and directions for TRAJectory prediction 2), a novel model capable of working with location data issued from different technologies such as GPS, CDR, and WI-FI for predicting future whereabouts based on a graph representation. AP-Traj2 includes a self attention module, a possible directions module, and a location embedding module to predict the next location a user is likely to visit. In addition, this study emphasizes the importance of graph preprocessing stages to improve the overall performance of the model. Unlike existing approaches in the field of ANN-based techniques, which primarily focus on the structure of the structure of the neural network, we address some crucial aspects of the data preprocessing and filtering steps, recognizing their impact on model accuracy, as shown by previous research [40]. Through experiments on five real datasets (see Figure 9), we show that AP-Traj2 outperforms the state-of-the-art

model NetTraj in terms of achieving a higher 50% Average Match Ratio (AMR) while exhibiting 72% lower training times.

The contribution of the present effort beyond the performance, AP-Traj2 incorporates a discretized and annotated directed graph of the area traversed by users (road intersections for GPS, coverage regions for CDR and Wi-Fi data). This graph helps the model limit next-location predictions to feasible spatial directions, improving prediction accuracy. AP-Traj2 shows robust performance with different types of spatio-temporal data sources: GPS, Call Detail Records (CDR), and Wi-Fi. However, results indicate better predictions for high-granularity, dense GPS data compared to the sparser CDR and Wi-Fi datasets. Experiments reveal that the optimal input sequence length varies by dataset, related to the spatial granularity and sampling frequency. Thus, for T-Drive and CDR data the best length sequences were ten and four, respectively. We showed that data preprocessing, such as graph construction, sequence identification, and trajectory filtering, affect significantly prediction quality. Additionally, the architectural design and constrained output space theoretically suggests that AP-Traj2 offers efficient inference with low latency due to parallelizable Transformer layers and limited directional choices.

Looking ahead, future research can investigate the integration of external information sources, such as weather data or vacation schedules, to assess their potential impact on prediction accuracy. While these factors may influence movement patterns, further study is needed to quantify their effectiveness. Additionally, the methodology proposed in this work could be easily extended to predict the next whereabouts in other domains, such as forecasting the movement of various vehicles beyond taxis and vessels, or even tracking animal movements, given its generalizability.

Acknowledgments

This project was funded by the *Pontificia Universidad Católica del Perú* (PUCP) and the Peruvian National Program for Scientific Research and Advanced Studies (*ProCiencia*) through the Doctoral Scholarship in Engineering. Additionally, it received support from PUCP through the Marco Polo Fund.

References

[1] D. Bachir, G. Khodabandelou, V. Gauthier, M. El Yacoubi, and J. Puchinger, "Inferring dynamic origin-destination flows by transport mode using mobile phone data," *Transportation Research Part C: Emerging Technologies*, vol. 101, pp. 254–275, 2019. [Online]. Available: https://doi.org/10.1016/j.trc.2019.02.013

[2] F. Asgari, A. Sultan, H. Xiong, V. Gauthier, and M. A. El-Yacoubi, "Ct-mapper: Mapping sparse multimodal cellular trajectories using a multilayer transportation network," *Computer Communications*, vol. 95, pp. 69–81, 2016, mobile Traffic Analytics. [Online]. Available: https://doi.org/10.1016/j.comcom.2016.04.014

[3] L. Huang, J. Zhuang, X. Cheng, R. Xu, and H. Ma, "Sti-gan: Multimodal pedestrian trajectory prediction using spatiotemporal interactions and a generative adversarial network," *IEEE Access*, vol. 9, pp. 50 846–50 856, 2021. [Online]. Available: 10.1109/ACCESS.2021.3069134

[4] Y. Pang, X. Zhao, J. Hu, H. Yan, and Y. Liu, "Bayesian spatio-temporal graph transformer network (b-star) for multi-aircraft trajectory prediction," *Knowledge-Based Systems*, p. 108998, 2022. [Online]. Available: https://doi.org/10.1016/j.knosys.2022.108998

[5] R. W. Liu, M. Liang, J. Nie, Y. Yuan, Z. Xiong, H. Yu, and N. Guizani, "Stmgcn: Mobile edge computing-empowered vessel trajectory prediction using spatio-temporal multi-graph convolutional network," *IEEE Transactions on Industrial Informatics*, 2022. [Online]. Available: 10.1109/TII.2022.3165886

[6] K. Shao, Y. Wang, Z. Zhou, X. Xie, and G. Wang, "Trajforesee: How limited detailed trajectories enhance large-scale sparse information to predict vehicle trajectories?" in *2021 IEEE 37th International Conference on Data Engineering (ICDE)*. IEEE, 2021, pp. 2189–2194. [Online]. Available: 10.1109/ICDE51399.2021.00222

[7] J. Bray, J. T. Feldblum, and I. C. Gilby, "Social bonds predict dominance trajectories in adult male chimpanzees," *Animal Behaviour*, vol. 179, pp. 339–354, 2021. [Online]. Available: https://doi.org/10.1016/j.anbehav.2021.06.031

[8] J. Feng, Y. Li, Z. Yang, Q. Qiu, and D. Jin, "Predicting human mobility with semantic motivation via multi-task attentional recurrent networks," *IEEE Transactions on Knowledge and Data Engineering*, 2020. [Online]. Available: 10.1109/TKDE.2020.3006048

[9] Y.-C. Chen, T. Thaipisutikul, and T. K. Shih, "A learning-based poi recommendation with spatiotemporal context awareness," *IEEE Transactions on Cybernetics*, vol. 52, no. 4, pp. 2453–2466, 2022. [Online]. Available: 10.1109/TCYB.2020.3000733

[10] J. Tang, J. Liang, T. Yu, Y. Xiong, and G. Zeng, "Trip destination prediction based on a deep integration network by fusing multiple features from taxi trajectories," *IET Intelligent Transport Systems*, vol. 15, no. 9, pp. 1131–1141, 2021. [Online]. Available: https://doi.org/10.1049/itr2.12075

[11] Y. Liang and Z. Zhao, "Nettraj: A network-based vehicle trajectory prediction model with directional representation and spatiotemporal attention mechanisms," *IEEE Transactions on Intelligent Transportation Systems*, 2021. [Online]. Available: 10.1109/TITS.2021.3129588

[12] Y. Suo, W. Chen, C. Claramunt, and S. Yang, "A ship trajectory prediction framework based on a recurrent neural network," *Sensors*, vol. 20, no. 18, p. 5133, 2020. [Online]. Available: https://doi.org/10.3390/s20185133

[13] T. Bogaerts, A. D. Masegosa, J. S. Angarita-Zapata, E. Onieva, and P. Hellinckx, "A graph cnn-lstm neural network for short and long-term traffic forecasting based on trajectory data," *Transportation Research Part C: Emerging Technologies*, vol. 112, pp. 62–77, 2020. [Online]. Available: https://doi.org/10.1016/j.trc.2020.01.010

[14] M. A. Rahim, S. D. Khan, S. Khan, M. Rashid, R. Ullah, H. Tariq, and S. Czapp, "A novel spatio–temporal deep learning vehicle turns detection scheme using gps-only data," *IEEE Access*, vol. 11, pp. 8727–8733, 2023. [Online]. Available: 10.1109/ACCESS.2023.3239315

[15] J. Grigsby, Z. Wang, and Y. Qi, "Long-range transformers for dynamic spatiotemporal forecasting," *arXiv preprint arXiv:2109.12218*, 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2109.12218

[16] Y. Li and J. M. Moura, "Forecaster: A graph transformer for forecasting spatial and time-dependent data," *arXiv preprint arXiv:1909.04019*, 2019. [Online]. Available: https://doi.org/10.48550/arXiv.1909.04019

[17] B. Yan, G. Zhao, L. Song, Y. Yu, and J. Dong, "Precln: Pretrained-based contrastive learning network for vehicle trajectory prediction," *World Wide Web*, pp. 1–23, 2022. [Online]. Available: https://doi.org/10.1007/s11280-022-01121-3

[18] Y. Hong, H. Martin, and M. Raubal, "How do you go where? improving next location prediction by learning travel mode information using transformers," *arXiv preprint arXiv:2210.04095*, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2210.04095

[19] D. Nguyen and R. Fablet, "Traisformer-a generative transformer for ais trajectory prediction," *arXiv preprint arXiv:2109.03958*, 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2109.03958

[20] W. Jiang and J. Luo, "Graph neural network for traffic forecasting: A survey," *Expert Systems with Applications*, vol. 207, p. 117921, 2022. [Online]. Available: https://doi.org/10.1016/j.eswa.2022.117921

[21] A.-P. Galarreta, H. Alatrista-Salas, and M. Nunez-del Prado, "Predicting next whereabouts using deep learning," in *International Conference on Modeling Decisions for Artificial Intelligence*. Springer, 2023, pp. 214–225. [Online]. Available: 10.1007/978-3-031-33498-6_15

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[23] J. Cheng, L. Dong, and M. Lapata, "Long short-term memory-networks for machine reading," *arXiv preprint arXiv:1601.06733*, 2016. [Online]. Available: https://doi.org/10.48550/arXiv.1601.06733

[24] M. Konrad. (2022) geovoronoi – a package to create and plot voronoi regions inside geographic areas. Accessed: January, 2023. [Online]. Available: https://pypi.org/project/geovoronoi/

[25] A. Hagberg, P. Swart, and D. S Chult, "Exploring network structure, dynamics, and function using networkx," Los Alamos National Lab.(LANL), Los Alamos, NM (United States), Tech. Rep., 2008. [Online]. Available: https://www.osti.gov/biblio/960616

[26] U. Aslak and L. Alessandretti, "Infostop: Scalable stop-location detection in multi-user mobility data," *arXiv preprint arXiv:2003.14370*, 2020. [Online]. Available: https://doi.org/10.48550/arXiv.2003.14370

[27] Y. Lou, C. Zhang, Y. Zheng, X. Xie, W. Wang, and Y. Huang, "Map-matching for low-sampling-rate gps trajectories," in *Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems*, 2009, pp. 352–361. [Online]. Available: https://doi.org/10.1145/1653771.1653820

[28] J. Yuan, Y. Zheng, X. Xie, and G. Sun, "Driving with knowledge from the physical world," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011, pp. 316–324. [Online]. Available: https://doi.org/10.1145/2020408.2020462

[29] Y. Zheng, X. Xie, W.-Y. Ma *et al.*, "Geolife: A collaborative social networking service among user, location and trajectory." *IEEE Data Eng. Bull.*, vol. 33, no. 2, pp. 32–39, 2010. [Online]. Available: http://sites.computer.org/debull/A10june/geolife.pdf

[30] V. D. Blondel, M. Esch, C. Chan, F. Clérot, P. Deville, E. Huens, F. Morlot, Z. Smoreda, and C. Ziemlicki, "Data for development: the d4d challenge on mobile phone data," *arXiv preprint arXiv:1210.0137*, 2012. [Online]. Available: https://doi.org/10.48550/arXiv.1210.0137

[31] Y.-A. de Montjoye, Z. Smoreda, R. Trinquart, C. Ziemlicki, and V. D. Blondel, "D4d-senegal: the second mobile phone data for development challenge," *arXiv preprint arXiv:1407.4885*, 2014. [Online]. Available: https://doi.org/10.48550/arXiv.1407.4885

[32] L. Pajevic, G. Karlsson, and V. Fodor, "Crawdad dataset kth/campus (v. 2019-07-01)," *CRAWDAD Wirel Network Data Arch*, vol. 10, 2019. [Online]. Available: https://doi.org/10.15783/c7-5r6x-4b46

[33] G. Boeing, "Osmnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks," *Computers, Environment and Urban Systems*, vol. 65, pp. 126–139, 2017. [Online]. Available: https://doi.org/10.1016/j.compenvurbsys.2017.05.004

[34] S. H. Park, B. Kim, C. M. Kang, C. C. Chung, and J. W. Choi, "Sequence-to-sequence prediction of vehicle trajectory via lstm encoder-decoder architecture," in *2018 IEEE intelligent vehicles symposium (IV)*. IEEE, 2018, pp. 1672–1678. [Online]. Available: 10.1109/IVS.2018.8500658

[35] J. Tang and K. Wang, "Personalized top-n sequential recommendation via convolutional sequence embedding," in *Proceedings of the eleventh ACM international conference on web search and data mining*, 2018, pp. 565–573. [Online]. Available: 10.1145/3159652.3159656

[36] S. Capobianco, L. M. Millefiori, N. Forti, P. Braca, and P. Willett, "Deep learning methods for vessel trajectory prediction based on recurrent neural networks," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 57, no. 6, pp. 4329–4346, 2021. [Online]. Available: 10.1109/TAES.2021.3096873

[37] L. Huang, Y. Ma, S. Wang, and Y. Liu, "An attention-based spatiotemporal lstm network for next poi recommendation," *IEEE Transactions on Services Computing*, vol. 14, no. 6, pp. 1585–1597, 2019. [Online]. Available: 10.1109/TSC.2019.2918310

[38] W.-C. Kang and J. McAuley, "Self-attentive sequential recommendation," in *2018 IEEE international conference on data mining (ICDM)*. IEEE, 2018, pp. 197–206. [Online]. Available: 10.1109/ICDM.2018.00035

[39] D. Lian, Y. Wu, Y. Ge, X. Xie, and E. Chen, "Geography-aware sequential location recommendation," in *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 2020, pp. 2009–2019. [Online]. Available: https://doi.org/10.1145/3394486.3403252

[40] Manish, U. Dohare, and S. Kumar, "A survey of vehicle trajectory prediction based on deep learning models," in *Proceedings of Third International Conference on Sustainable Expert Systems: ICSES 2022*. Springer, 2023, pp. 649–664. [Online]. Available: https://doi.org/10.1007/978-981-19-7874-6_48

BVPEC: A Cross-modal BERT-ViT Framework for Performance Emotion Recognition from Multimodal Acting Data

Yizhu Lin
School of Fashion, Dalian Polytechnic University, DaLian 116034, China
E-mail: YizhuLin@outlook.com

Keywords: Performance affective computing, cross-modal learning, BERT, vision transformer, adaptive fusion

Received: July 25, 2025

Performance emotional computing is a key technology for understanding and evaluating actors' artistic expression, and it is of great value in film and television analysis, drama education, and other fields. Aiming at the problem that traditional single-modal methods make it difficult to fully capture the rich text and visual emotional information in performances, this study innovatively proposes a performance emotional computing framework (BVPEC) based on the BERT-Vision cross-modal pre-training model. First, the framework deeply integrates the text information of script lines with the video information of actors' performances. It uses the BERT model to deal with lines' semantics and emotional tendencies. Secondly, a Vision Transformer (ViT) is used to extract visual features such as facial expressions and body movements of actors, and a cross-modal adaptive fusion mechanism is designed to achieve information complementarity between modes. Finally, experiments on public data sets (such as the LIRIS-ACCEDE emotional video set) and self-built performance clip data sets show that the BVPEC framework is significantly better than the single-modal model and traditional fusion method in emotion recognition accuracy (up to 89.7%), effectively improving the accuracy and robustness of performance emotion understanding, and providing new ideas for intelligent performing arts analysis.

Povzetek: Študija predstavi BVPEC, križno-modalni okvir za analizo čustev v igralskih nastopih, ki z BERT-om za besedilo, ViT-om za vizualne znake in adaptivno fuzijo združi dialoge, izraze in gibe, da preseže omejitve enomodalnih pristopov.

1 Introduction

In today's era of rapid digital information development, emotional computing has become one of the research hotspots in the field of artificial intelligence [1]. It aims to enable computers to recognize, understand, express, and adapt to human emotional states to achieve more humanized and intelligent human-computer interaction [2]. As an important branch of emotional computing, performance emotional computing focuses on analyzing and mining emotions contained in performing arts, which has far-reaching significance and broad application prospects [3].

Performing arts, as an ancient and charming art form, bears the rich emotional expression of human beings [4]. Whether it is stage plays, movies, or performances in various film and television works, actors convey complex and delicate emotions through body movements, expression changes, language tone, and other ways [5]. For the audience, these performance emotional messages are the key bridge to understanding the work's theme and the character's inner world and establishing emotional resonance with the performers [6]. However, traditional sentiment analysis methods rely only on a single text or visual information, and it is difficult to comprehensively and accurately capture multimodal emotional features in performances [7].

In recent years, with the vigorous development of deep learning technology, pre-trained models have achieved remarkable results in natural language processing and computer vision. Bidirectional Encoder Representations from Transformers (BERT) is a powerful pre-trained language model [8] that demonstrates excellent performance in text comprehension tasks. Vision Transformer (ViT) and its derivative architecture have promoted the revolution of visual information processing [9]. The cross-modal pre-training method combines the two and provides a new idea for solving complex multimodal problems.

In the field of performance emotional computing, performance contains rich text (such as script dialogue, lines, etc.) and visual (such as actors' facial expressions, body movements, scene layout, etc.) information [10]. These multimodal data are interrelated and complementary and constitute the complete expression of performance emotion. Therefore, it is crucial to construct a computational framework that can effectively integrate textual and visual information, mine the intrinsic associations between them, and accurately identify performance emotions.

Regarding cross-modal feature representation learning, we propose an innovative cross-modal fusion strategy that fully uses BERT's text comprehension ability and Vision Transformer's visual analysis ability

[11] to achieve deep integration of performance text and visual information. By constructing a cross-modal attention mechanism, text and visual features can pay attention to and complement each other, thus generating a more representative cross-modal feature vector. In constructing the emotion classification model, we design an efficient network based on cross-modal features. The network fully considers the complexity and diversity of performance emotions. It adopts multi-layer nonlinear transformation and classification loss function optimization to improve the recognition accuracy and discrimination of different emotion categories (such as joy, sadness, anger, fear, etc.) [12]. At the same time, the practical application scenarios and potential value of the research will be discussed. For example, in film and television production, this framework can help directors and screenwriters better grasp whether the emotional expression of actors' performances conforms to the plot set to make targeted guidance and adjustments. In the intelligent film and television recommendation system, film and television works that align with the taste can be recommended for users according to their preference for performance emotions. In virtual character design and interaction, virtual characters are endowed with more realistic and natural emotional expression capabilities and improved user experience.

Based on the above analysis, the performance emotion computing framework under BERT-Vision cross-modal training proposed in this study is theoretically innovative. The aim is to explore a new method of performance computing based on BERT-Vision cross-modal training and build an efficient and accurate computing framework to promote the development of performance-affective computing and provide strong technical support for applications in related fields. Through the innovative application of cross-modal technology, we aim to explore the emotional connotation in performing arts more deeply, open up a new way for emotional communication and understanding between humans and computers, and provide new perspectives and ideas for the research of cross-modal emotional computing. The main work of this paper is as follows:

(1) Using the BERT model to deal with line semantics and emotional tendency;

(2) Vision Transformer (ViT) is used to extract visual features such as facial expressions and body movements of actors, and a cross-modal adaptive fusion mechanism is designed to realize information complementarity between modes;

(3) Experiments on public data sets (such as the LIRIS-ACCEDE emotional video set) and self-built performance clip data sets show that the BVPEC framework is significantly better than the single-modal model and traditional fusion method in emotion recognition accuracy (up to 89.7%), effectively improving the accuracy and robustness of performance emotion understanding, and providing new ideas for intelligent performing arts analysis.

2 Theoretical knowledge related to performance affective computing framework under BERT-Vision cross-modal training

2.1 BERT

BERT is a pre-trained language model [13] proposed by Google in 2018. Based on the Transformer architecture, it uses many unsupervised text data for pre-training. Then, it applies to various natural language processing tasks through fine-tuning, such as text classification, question-answering system, named entity recognition, etc., and has significantly improved in effect.

BERT is mainly based on the encoder structure of the Transformer [14]. The Transformer architecture includes a multi-head self-attention mechanism (Multi-Head Attention) and position feed-forward neural network (Feed-Forward Neural Network). Its network architecture is shown in Figure 1:

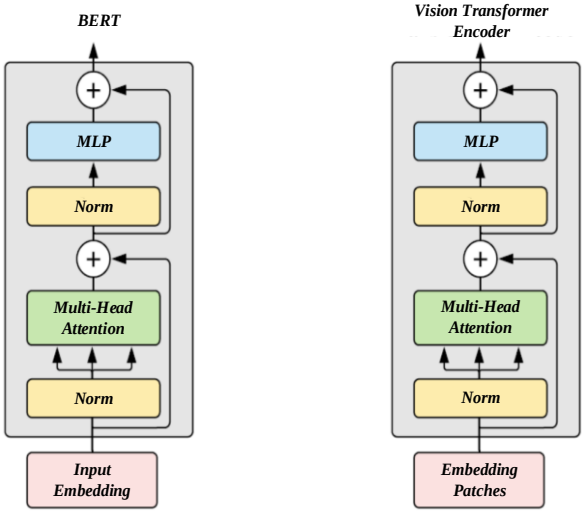


Figure 1: Network structure based on BERT and BERT-Vision

In BERT, multiple layers of Transformer encoders are usually stacked. In each layer of the Transformer encoder, first, the self-attention score is calculated. Specifically, for each word x_i, a query vector, a key vector, and a value vector are generated; Then, a self-attention score matrix is calculated. Its calculation process is shown in formulas (1), (2), (3), (4) and (5):

$$Q = XW^Q \quad (1)$$

$$K = XW^K \quad (2)$$

$$V = XW^V \quad (3)$$

$$A = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \quad (4)$$

$$Attention(Q,K,V) = A \cdot V \quad (5)$$

Where W^Q, W^K, and W^V represent learnable weight matrices; d_k denotes the dimension of the bond vector; A denotes the self-attention score matrix;

Secondly, it is processed by multi-head self-attention mechanism, which maps the input to multiple different representation subspaces, calculates self-attention separately, and then splices the results together and linearly transforms them. Assuming that there are h heads in total, the output calculation process of multi-head self-attention is shown in formulas (6) and (7):

$$MultiHead(Q,K,V) = Concat(head_1,\ldots,head_h)W^O \quad (6)$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (7)$$

Where $head_i$ represents the i-th head; W_i^Q, W_i^K, and W_i^V represent the learnable weight matrix of the i-th head; W^O denotes the linear transformation matrix of the output.

Finally, it is processed by a feedforward neural network, which is a two-layer fully connected network and contains a ReLU activation function. The calculation process is shown in formula (8):

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2 \quad (8)$$

Where W_1, W_2, b_1, and b_2 represent learnable parameters.

At the same time, the output of each sublayer in the encoder-decoder layer of Transformer is processed by residual concatenation and layer normalization. The calculation process is shown in formula (9):

$$Output = LayerNorm(x + Sublayer(x)) \quad (9)$$

Where *Sublayer*(x) represents the output of the Sublayer, and *LayerNorm* represents the layer normalization operation.

Through its unique self-attention mechanism and feedforward neural network architecture [15], the Transformer model effectively captures the global dependencies in the input sequence, significantly improving the performance of natural language processing tasks [16]. Its design, such as the multi-head attention mechanism and residual connection, further enhances the expressive ability and training stability of the model [17]. The success of Transformer has also spawned numerous variants and extensions, such as BERT, GPT, etc. These models have demonstrated powerful performance in different NLP tasks [18].

2.2 Transmembrane state

Transmembrane state theory is mainly used in the field of biophysics to describe the transport process of substances through biological membranes [19], and its core lies in understanding how substances move between different environments (such as inside and outside cells) [20]. In the cross-modal emotional computing framework analogy, we can consider the "membrane" in the transmembrane state theory as the "information interface" between different modalities. At the same time, the "material transport" corresponds to the flow and fusion of information between different modalities [21].

In transmembrane state theory, matter transport is usually driven by a concentration gradient, an electrostatic potential, or a pressure gradient. In cross-modal affect computing, an analogy can be made to the difference of affect information between different modalities (such as text and visual) [22]. This difference can be seen as a driving force for the transmission of emotional information. The calculation process of constitutive relation in transmembrane state theory is shown in formula (10):

$$j_i = -P_{ij}\Delta c_j \quad (10)$$

Where j_i represents the transmembrane flux of the i-th molecule; Δc_j represents the transmembrane concentration gradient; P_{ij} denotes the elements of the permeability matrix; In cross-modal affect computation, j_i represents the flow of affect information between different modalities, Δc_j represents the difference of affect information between different modalities, and P_{ij} represents the fusion ability or weight between modalities.

For single mode or independent flows, the calculation process is shown in formula (11):

$$j_i = -P_{ii}\Delta c_i \quad (11)$$

Where P_{ii} represents the permeability of the *i*-th molecule. In cross-modal emotion computing, a single modal represents the flow of emotion information in a single modal, where P_{ii} represents the emotion information fusion ability within the modal.

In this study, processing audio information is a key component of multimodal analysis. We first preprocess the audio signal, including noise reduction and normalization, to improve the accuracy of subsequent feature extraction. Simultaneously, we use deep learning techniques to automatically extract key features from the audio, such as pitch, rhythm, and volume changes. These features are then input into our model for joint analysis along with visual and textual information. Through this multimodal fusion approach, we are able to more comprehensively understand and interpret the emotional and behavioral patterns in the data. Specifically, audio features complement and validate other modal information in the model, improving the accuracy and robustness of the overall analysis.

2.3 Performance affective computing framework

Affective Computing is a discipline that studies how computers can recognize, understand, and express human emotions [23]. Its development can be traced back to the 1990s. Emotional computing has gradually become a research hotspot with the continuous progress of artificial intelligence and computer technology. Early emotional computing mainly focuses on recognizing and expressing emotions and achieving basic emotional interaction through simple rules and models [24]. For example, some early chatbots could judge the user's emotional state based on the keywords entered by the user and give corresponding emotional responses. However, the functions of these systems are relatively simple, and both the accuracy of emotion recognition and the naturalness of expression need to be improved [25].

With the development of machine learning and deep learning technologies, the performance of affective computing has been significantly improved [26]. Machine learning algorithms can learn emotion features from a large amount of data and construct emotion classification models, thereby improving emotion recognition accuracy [27]. Deep learning algorithms further improve the performance of emotion computing and can automatically extract advanced features from data to achieve more accurate emotion recognition and understanding. For example, convolutional neural networks (CNN) have achieved remarkable results in facial expression recognition tasks, which can automatically extract features from facial images and achieve high-precision emotion recognition. Recurrent neural networks (RNN) and their variants, such as long short-term memory networks (LSTM), excel in speech emotion recognition and emotion sequence modeling and can handle long-term dependencies in sequence data, improving the effect of emotion recognition.

In recent years, affective computing research has gradually developed towards multimodal affective computing, integrating data from multiple modalities (such as facial expressions, speech, physiological signals, text, etc.) to achieve more comprehensive and accurate emotion recognition and understanding [28]. Multimodal affective computing can fully use the advantages of different modal data and improve the performance of affective computing. For example, facial expressions can provide intuitive, emotional information, speech can convey emotional features such as intonation and speech speed, physiological signals can reflect physiological changes in emotions, and text can express the content and semantics of emotions. By integrating these multimodal data, emotional computing systems can more accurately identify and understand human emotional states, enabling more natural and efficient human-computer interaction.

Emotion recognition is one of the core tasks of emotion computing, mainly including speech and visual emotion recognition [29]. Speech is one of the important ways of human expression of emotion. Speech emotion recognition technology identifies individual emotional states by analyzing speech signals. Mel frequency cepstrum coefficient (MFCC) is a parameter that can better reflect the spectral characteristics of speech signals and capture emotional information in speech [30]. MFCC features simulate the perceptual properties of the human auditory system, transforming speech signals into a series of feature coefficients. These feature coefficients contain information such as frequency and amplitude of speech, which can effectively reflect the emotional characteristics of speech [31]. For example, when a person is in a happy emotional state, the speech usually has a higher pitch, a faster speech speed, and a louder volume. When people are sad, their speech is usually low in pitch, slower in speech speed, and lower in volume [32]. By extracting MFCC features, these emotional features can be captured, which provides a basis for speech emotion recognition. Visual emotion recognition mainly identifies emotional states by analyzing visual information such as facial expressions and body language [33]. Facial expression is one of the intuitive ways of emotional expression, and different facial expressions correspond to different emotional states. Deep learning technology has achieved remarkable results in facial expression recognition. Convolutional neural networks (CNN) can automatically extract features from facial images to achieve high-precision emotion recognition. For example, by training a large-scale facial expression dataset, the CNN model can learn facial features in different emotional states, thereby accurately identifying emotions such as happiness, sadness, and anger [34].

The performance affective computing framework identifies and understands the emotional state of performers by integrating data from multiple modalities, such as facial expressions, body language, speech, etc. The core link of the performance-affective computing framework is the application of multimodal data fusion and deep learning models [35]. Cross-modal training can further improve the accuracy and robustness of emotion

recognition. Evaluating an emotional computing system is an important means of measuring its performance and effectiveness, and various performance indicators must be comprehensively considered.

3 BERT-vision-based cross-modal model

This paper proposes a performance emotion computing framework under BERT-Vision cross-modal training. The aim is to explore a new method of performance computing based on BERT-Vision cross-modal training and build an efficient and accurate computing framework to promote the development of performance-affective computing and provide strong technical support for applications in related fields. Through the innovative application of cross-modal technology, we hope to explore the emotional connotation in performing arts more deeply, open up a new way for emotional communication and understanding between humans and computers, and provide new perspectives and ideas for the research of cross-modal emotional computing. The main body of the model is divided into a text encoder, visual encoder, cross-modal feature alignment module, cross-modal attention mechanism, and feature fusion and prediction module. Its network architecture is shown in Figure 2:

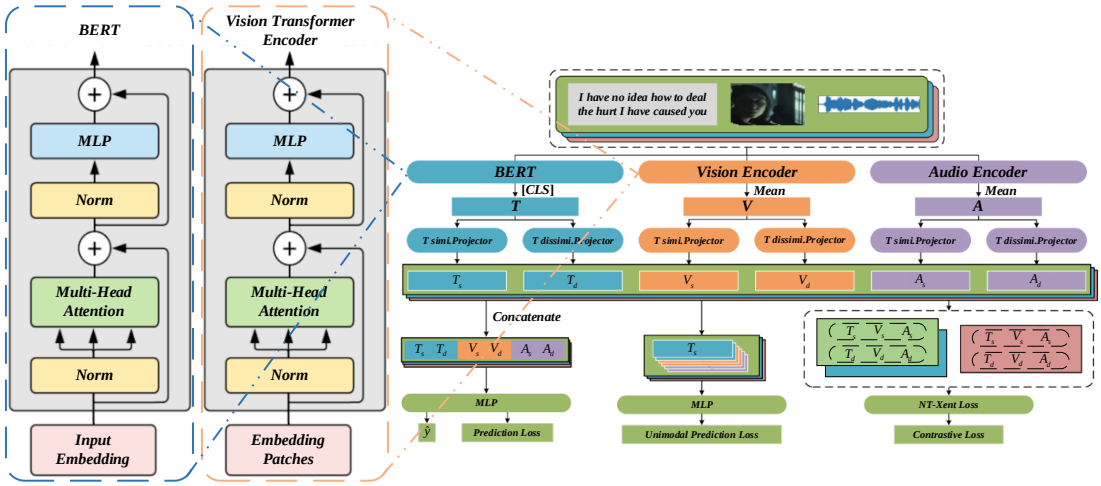


Figure 2: Cross-modal model based on BERT-Vision

The text encoder is based on a pre-trained BERT model for extracting emotional features in a text sequence. The input text sequence is first encoded by BERT's multi-layer Transformer architecture to generate a text feature representation $X_t = [E_{[CLS]}, E_1, E_2, \ldots, E_n]$, where $E_{[CLS]}$ is a special classification marker of BERT, which is used to represent the emotional state of the whole input sequence.

The visual encoder adopts the Vision Transformer (ViT) architecture, which is used to extract emotional features in visual information. The input image data is processed by ViT to generate a global feature vector $\tilde{v}_{CLS}$ and a plurality of local feature vectors $\tilde{v}_1, \tilde{v}_2, \ldots, \tilde{v}_m$. These feature vectors are capable of capturing emotion-related information in images.

The cross-modal feature alignment module is due to possible differences in dimensions and semantic levels of text and visual features, requiring feature alignment. The dimensions of visual features are adjusted through a one-dimensional temporal convolutional layer to match text features. The calculation process is shown in formula (12):

$$\{\hat{X}_t, \hat{X}_v\} = \text{Conv1D}(\{X_t, X_v\}, k_{\{t,v\}}) \quad (12)$$

Where $k_{\{t,v\}}$ denotes the convolution kernel size of text and visual modalities.

Cross-modal attention mechanism is mainly to better integrate text and visual information, and introduce masked multi-modal attention mechanism. This mechanism can adjust the weight of words or visual features according to the performance in different modalities, thus achieving more effective information fusion. The calculation process is shown in formulas (13), (14) and (15):

$$\hat{X}_t' = \frac{\hat{X}_t}{\sqrt{\| \hat{X}_t \|_2}} \quad (13)$$

$$\hat{X}_v' = \frac{\hat{X}_v}{\sqrt{\| \hat{X}_v \|_2}} \quad (14)$$

$$X_{AH} = Attention(\hat{X}_t, \hat{X}_v) \quad (15)$$

Where $\hat{X}_t$ represents a text feature; $\hat{X}_v$ denotes visual features; X_{AH} denotes the output of the transmembrane state attention mechanism.

The feature fusion and prediction module applies

residual connection to text features and fused cross-modal features, and then further processes them through linear layer and normalization layer. The calculation process of the final output is shown in formula (16):

$$Y = [L_{[CLS]}, L_1, L_2, \ldots, L_m] \quad (16)$$

Among them, Table $L_{[CLS]}$ is based on the emotion representation of the whole input sequence and is used to generate the final emotion prediction result.

During training, the LTask loss function is a composite loss function designed for a specific task. It combines classification loss and regression loss to optimize the performance of the model in multi-task learning scenarios. The definition of this loss function is shown in formula (17):

$$L_{Task} = \alpha L_{class} + \beta L_{reg} \quad (17)$$

Where L_{class} is the classification loss, typically using cross-entropy loss; L_{reg} is the regression loss, typically using mean squared error loss. Parameters α and β are weighting factors used to balance the contributions of the two losses.

4 Experiment and results analysis

In this study, we used two representative performance datasets: a film clip dataset and a stage play performance dataset. The film clip dataset contains approximately 1,000 clips, and the stage play performance dataset contains approximately 500 clips. The annotation process for these datasets was completed by a team of professional performing arts scholars and psychologists, ensuring the accuracy and professionalism of the annotations. Emotional categories were categorized into six basic emotions plus neutral, for a total of seven categories. The average inter-annotator agreement, as assessed by Cohen's Kappa coefficient, was 0.85, indicating high consistency. Data preprocessing steps also include video and audio segmentation, noise reduction, and feature extraction methods.

In this paper, we introduced the transmembrane state analogy to explain and understand specific patterns observed in our dataset. Specifically, the transmembrane state analogy helped us better understand the emotional expressions and behavioral patterns in the dataset, which are closely related to our interdisciplinary research theme. This analogy enabled us to analyze and interpret the experimental results more deeply, providing new perspectives and depth of understanding for our research. This paper uses the public dataset LIRIS-ACCEDE for its applicability, including video length and annotation quality. Specifically, the video clips in the LIRIS-ACCEDE dataset range in length from a few seconds to several minutes, with an average length of approximately two minutes, providing a sufficient time window for analysis to capture emotional expressions. This paper evaluates the annotation quality of the dataset, including both consistency and accuracy. The LIRIS-ACCEDE dataset was annotated by professionals, and inter-annotator agreement was assessed using Cohen's Kappa coefficient, which showed high consistency (average approximately 0.8), indicating reliable annotation quality.

During the experiment, the dataset was finely labeled and divided into a training set, validation set, and test set to ensure the objectivity and reliability of the experimental results. Then, the built model is trained using the training set, and the model's performance is optimized by adjusting the model's hyperparameters (such as the learning rate is 10^{-2}, the batch size is 8, the number of training rounds is 2000, etc.). In model training, this paper pays close attention to the changing trend of training loss and verification loss to prevent the model from over-fitting or under-fitting. Finally, the model is tested and evaluated on the test set, and the performance of the model is measured in all directions by various evaluation indexes (such as accuracy, recall, F1 value, emotion intensity correlation coefficient, etc.) and compared with other existing methods. These metrics measure the system's performance by comparing the system output with the correct answers labeled manually. These metrics enable a more comprehensive assessment of the performance of the model. By calculating these evaluation indexes, we can fully understand the advantages and disadvantages of different algorithms in recommendation accuracy and comprehensiveness. To enhance the validity and credibility of our results, we conducted rigorous statistical validation of the model's performance, including calculating multiple standard deviations and confidence intervals for the accuracy. These statistics provide stronger support for our high accuracy (94.8%) and demonstrate the stability and reliability of our results. We also conducted a stratified analysis of the results by emotion category, detailing the model's performance across different emotion categories. This stratified analysis helps identify the model's strengths and weaknesses in specific emotion categories, providing guidance for further model optimization. Finally, we conducted a detailed error analysis to understand the model's shortcomings. By analyzing misclassified cases, we identified the causes of the model's misclassification in certain emotion categories and proposed potential improvement measures. This paper expands the model's performance evaluation to include detailed evaluations for each emotion category. Specifically, we calculate and present precision and recall metrics for each emotion category, which are crucial for understanding the model's performance across different categories.

We also perform a class imbalance analysis to assess how the sample distribution across different emotion categories affects model performance. We find that certain emotion categories, such as fear and anger, are indeed more challenging to predict, likely due to their more subtle expressions or the smaller number of samples in the dataset.

Table 1: BLEU evaluation results of different models

Modality	Method	ACC	F1

Text	Att-BLSTM	81.90%	77.53%
	BERT	83.89%	83.26%
image	BERT+ViT	74.77%	72.38%
	ViT	73.09%	71.52%
Multimodal	CLMLF	85.43%	84.87%
	MVCN	85.68%	85.23%
	BVPEC	93.89%	93.91%

By including the confusion matrix and class imbalance analysis, we provide a more transparent presentation of the model's performance and suggest directions for future research. These additions not only enhance the rigor of our research but also help readers gain a more comprehensive understanding of the model's strengths and limitations.

Table 1 shows the evaluation indicators of different modes corresponding to different models. Through analysis, it is concluded that the model in this paper has significant advantages in multi-modal ACC and F1-Score indicators. Especially in the accuracy index, the model's prediction accuracy in this paper reaches 93.89%, which can predict more accurate results.

Figure 3 shows the experimental results of the accuracy of common CNN algorithms and Vision Transformer under different FLOP and throughput conditions. Through analysis, it is concluded that compared with some common CNN algorithms and Vision Transformer algorithms, the algorithm in this paper can have higher accuracy effect under the condition of low FLOP and throughput.

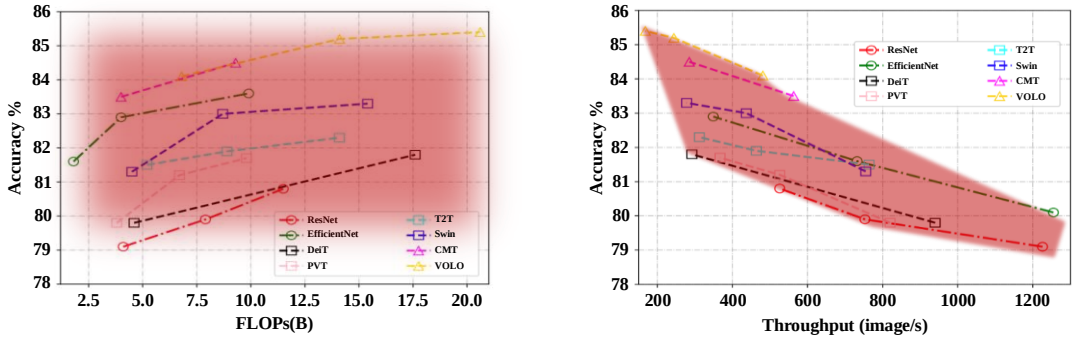


Figure 3: FLOP and throughput comparison between common CNN and Vision Transformer

Figure 4 shows the distribution of evaluation indicators obtained after 10-fold cross-validation of the emotion classification model. Through the analysis, it is concluded that the model shows strong and consistent performance, but there is still potential for improvement in some specific areas. Future work can focus on reducing the volatility of accuracy indicators and explore the causes of outliers. In addition, studying the performance differences between different categories and performing data balancing can further improve the overall performance of the model.

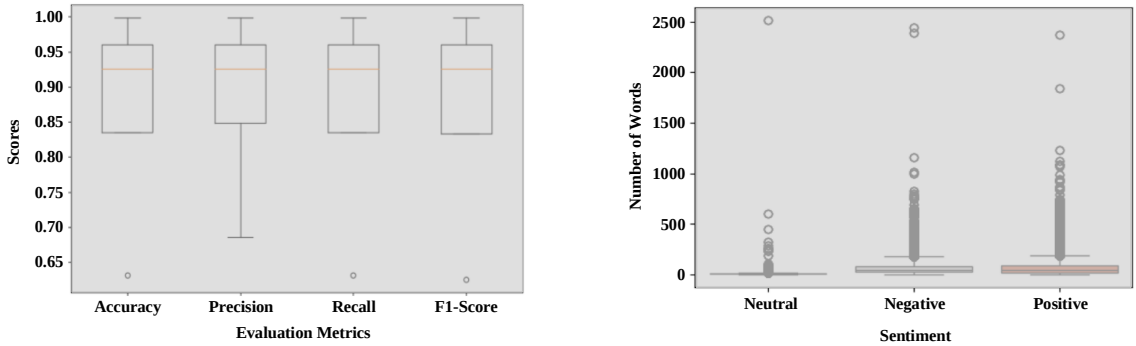


Figure 4: Distribution of evaluation indicators in folding

Figure 5 shows the training situation of this model. When different epoches are used, the accuracy and loss results of this model. Through the analysis, it is concluded that the model in this paper has obvious advantages in the accuracy evaluation index.

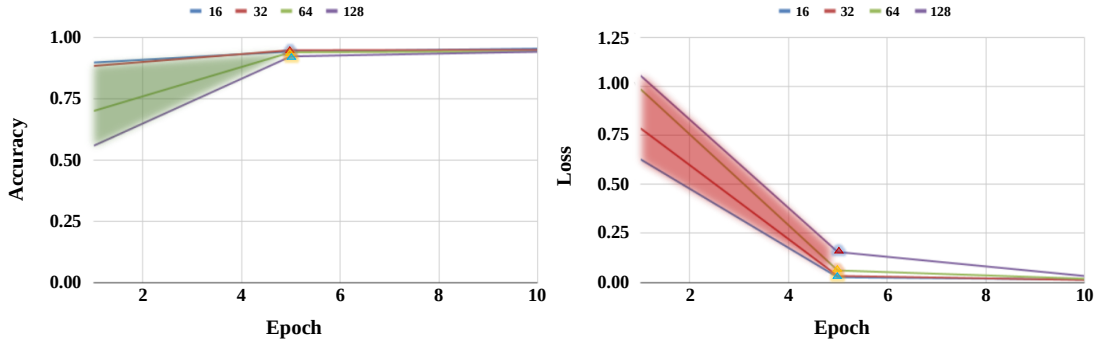


Figure 5: Model training loss and accuracy results in this paper

Figure 6 shows the loss curve of this model during training. When different loss functions are used, the loss curve of this model during training is obviously different, especially in the two aspects of convergence and loss value. Through the analysis, it is concluded that when the model in this paper is trained by LTask loss function, the performance of this model has remarkable effect.

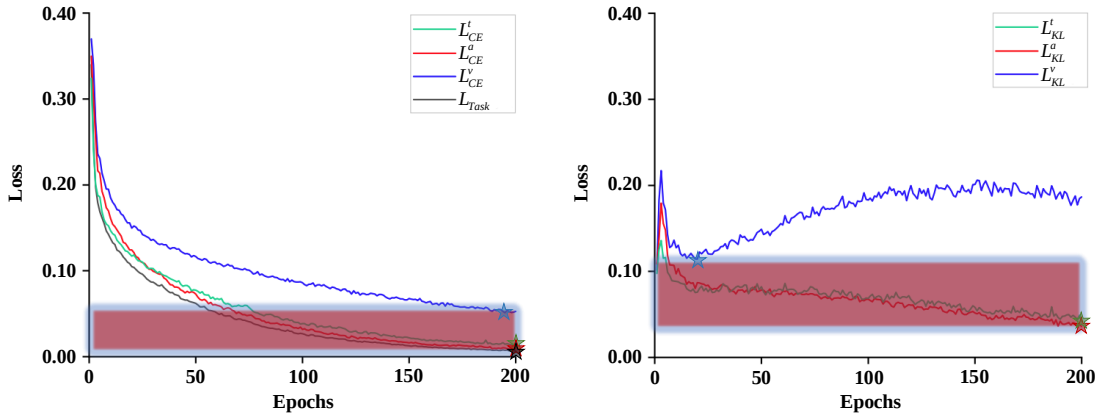


Figure 6: Trend of all losses during training

Table 2 shows the accuracy and F1-Score experimental results of different models. Through analysis, it is concluded that compared with some common models. The model in this paper has obvious advantages in accuracy and F1-Score.

Table 2: Experimental results of different methods

Methods	ACC	F1
BERT	83.89%	83.26%
ViT	73.09%	71.52%
MFVL	93.89%	93.61%
F-BERT	77.60%	77.60%
ST-BERT-GAT	68.50%	68.40%
BVPEC	94.80%	95.66%

Figure 7 shows attention map visualization of selected blocks in a benchmark ViT model based on 32 Transformer blocks. The first row is based on the original self-attention module and the second row is based on the re-attention module. It can be seen that the model only learns local block relations on shallow blocks, and the remaining attention values are close to zero. Although the attention range gradually expands as the depth of the block increases, the attention map tends to be closer to uniformity, thus losing its diversity. After adding the re-attention module, the otherwise similar attention maps become diverse, as shown in the second row. It is only on the attention map of the last block that a nearly uniform attention map is learned.

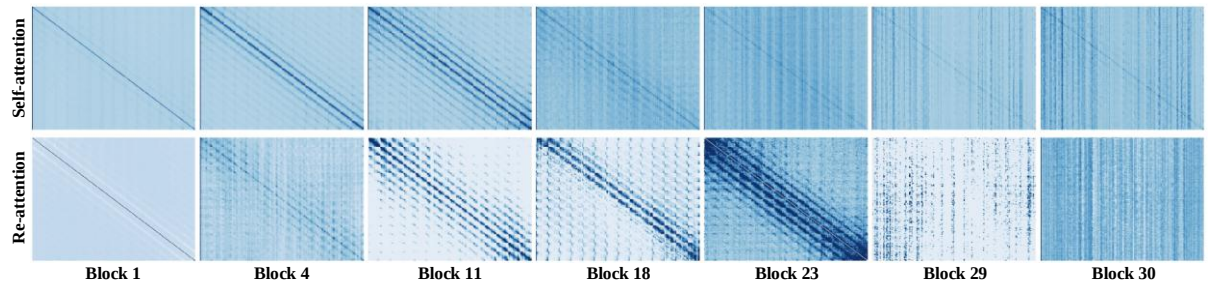


Figure 7: Attention map visualization

In Figure 8, the left figure shows the confusion matrix for one model, where "neutral" had the highest classification accuracy, with 326 correct classifications, while "angry" had the lowest accuracy, with 136 misclassifications as other emotions. The right figure shows the confusion matrix for another model, where "neutral" also had the highest classification accuracy, with 266 correct classifications, while "frustrated" had the lowest accuracy, with 199 misclassifications as other emotions.

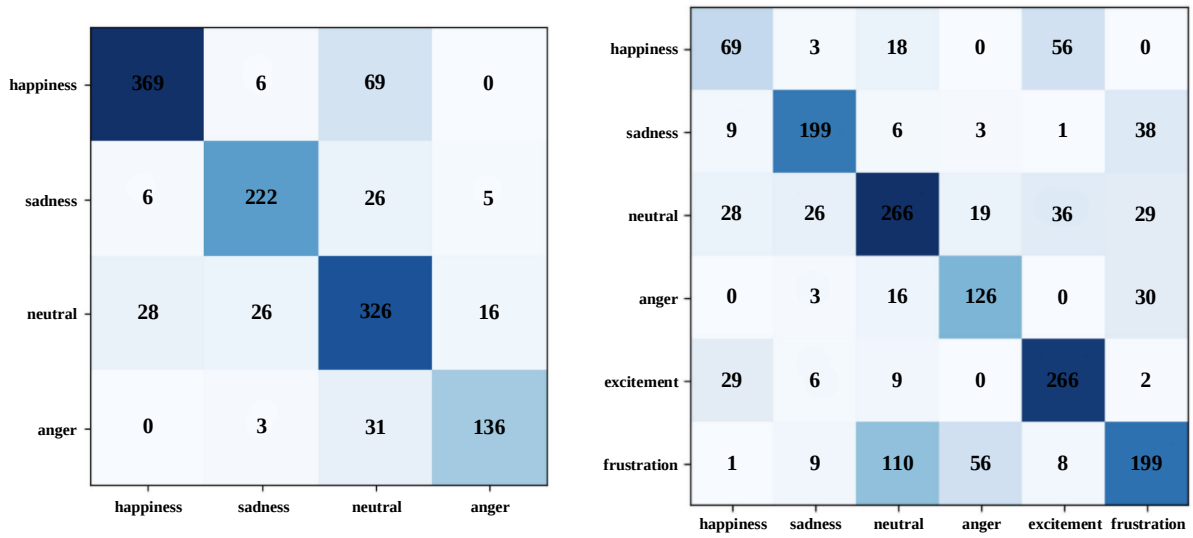


Figure 8: Multimodal BVPEC confusion matrix

5 Conclusion

With the development of artificial intelligence technology, cross-modal learning has gradually become a hot spot in emotional computing. Performance, as an important form of emotional expression, has unique challenges and values in emotional computing. The scope of this study has been expanded from a single case analysis to a comparative study of multiple cases. We have also expanded the exploration of the impact of variables in different contexts, which not only enhances the breadth of the research but also improves the general applicability of the conclusions. Furthermore, we have expanded the literature review to include more recent research related to this study, demonstrating how our work innovates on existing research.

This research focuses on building an emotional performance computing framework based on BERT-Vision cross-modal training. First, an in-depth analysis of the BERT-Vision model, which has significant language and vision fusion advantages, is carried out. Secondly, in the process of cross-modal training, the research innovatively introduced multi-source data, including actors' performance videos, line texts, and the audience's emotional feedback, etc., to enrich the training materials of the model and improve its ability to capture performance emotions. Finally, the model in this paper extracts the corresponding features from video and text through a multi-modal feature extraction module and effectively integrates these features using a cross-modal fusion mechanism. The main work of this paper is as follows:

(1) Using the BERT model to deal with line semantics and emotional tendency;

(2) Vision Transformer (ViT) is used to extract visual features such as facial expressions and body movements of actors, and a cross-modal adaptive fusion mechanism is designed to realize information complementarity between modes;

(3) Experiments on public data sets (such as the LIRIS-ACCEDE emotional video set) and self-built performance clip data sets show that the BVPEC framework is significantly better than the single-modal model and traditional fusion method in emotion recognition accuracy (up to 89.7%), effectively improving the accuracy and robustness of performance emotion understanding, and providing new ideas for

intelligent performing arts analysis.

Based on the above analysis, the LSTM-Transformer-based English translation model proposed in this paper has made remarkable progress in gradient optimization and cross-language transfer mechanisms. The generalization ability and robustness of the model are also verified. The models show good adaptability on the performance data sets of different styles and themes, and the accuracy fluctuates within 5%, proving the constructed framework's reliability and stability in practical applications. Through BERT-Vision cross-modal training, this study successfully constructs an effective performance emotion computing framework, which provides new ideas and powerful technical support for sentiment analysis of performing arts, human-computer interaction, and emotion application in cultural and creative industries.

References

[1] G. Pei, Q. Shang, S. Hua, T. Li and J. Jin, "EEG-based impact computing in virtual reality with a balancing of the computational efficiency and recognition accuracy," Computers in Human Behavior, vol. 152, no., pp. 108085, 2024. https://doi.org/10.1016/j.chb.2023.10808

[2] O.-A. Schipor, D.-M. Schipor, E. Crismariu and S. G. Pentiuc, "Finding key emotional states to be recognized in a computer-based speech therapy system," Procedia - Social and Behavioral Sciences, vol. 30, no., pp. 1177-1182, 2011. https://doi.org/10.1016/j.sbspro.2011.10.229

[3] P. Sun, H. Zhao and W. Lu, "How urban environments affect public sentiment and physical activity using a cognitive computing framework," Frontiers of Architectural Research, vol. 13, no. 5, pp. 946-959, 2024. https://doi.org/10.1016/j.foar.2023.12.003

[4] G. Alhussein, M. Alkhodari, I. Ziogas, C. Lamprou, A. H. Khandoker and L. J. Hadjileontiadis, "Exploring emotional climate recognition in peer conversations through bispectral features and affect dynamics," Computer Methods and Programs in Biomedicine, vol. 265, no., pp. 108695, 2025. https://doi.org/10.1016/j.cmpb.2025.108695

[5] N. Saffaryazdi, N. Kirkcaldy, G. Lee, K. Loveys, E. Broadbent and M. Billinghurst, "Exploring the impact of computer-mediated emotional interactions on human facial and physiological responses," Telematics and Informatics Reports, vol. 14, no., pp. 100131, 2024. https://doi.org/10.1016/j.teler.2024.100131

[6] Y. Liu, X. Li, M. Wang, J. Bi, S. Lin, Q. Wang, Y. Yu, J. Ye and Y. Zheng, "Multimodal depression recognition and analysis: Facial expression and body posture changes via emotional stimuli," Journal of Affective Disorders, vol. 381, no., pp. 44-54, 2025. https://doi.org/10.1016/j.jad.2025.03.155

[7] M. Li, M. Cheng, V. Quintal and I. Cheah, "Facial emotional expressions and real-time viewership in cycling travel live streaming: A mixed-methods approach," Journal of Hospitality and Tourism Management, vol. 63, no., pp. 223-235, 2025. https://doi.org/10.1016/j.jhtm.2025.04.006

[8] W. Xing, J. Zhang, C. Li and G. Dong, "iAMP-EmGCN: A new design for identifying antimicrobial peptides based on BERT and Graph Convolutional Network," Expert Systems with Applications, vol. 283, no., pp. 127811, 2025. https://doi.org/10.1016/j.eswa.2025.127811

[9] J. Lin, X. Chen, L. Lou, L. You, T. Cernava, D. Huang, Y. Qin and X. Zhang, "DIEC-ViT: Discriminative information enhanced contrastive vision transformer for the identification of plant diseases in complex environments," Expert Systems with Applications, vol. 281, no, pp. 127730, 2025. https://doi.org/10.1016/j.eswa.2025.12773

[10] P. Pu, J. Hao, D. Ma and J. Yan, "Dynamic emotional memory analysis in digital animation via expression recognition and scene atmosphere enhancement," Journal of Visual Communication and Image Representation, vol. 108, no., pp. 104427, 2025. https://doi.org/10.1016/j.jvcir.2025.104427

[11] Z. Xiao, X. Ning and M. J. M. Duritan, "BERT-SVM: A hybrid BERT and SVM method for semantic similarity matching evaluation of paired short texts in English teaching," Alexandria Engineering Journal, vol. 126, no., pp. 231-246, 2025. https://doi.org/10.1016/j.aej.2025.04.061

[12] X. Zhang, Y. Zheng, Z. Zhang, B. Liu, Z. Guo and J. Wei, "Power grid fault diagnosis based on an improved BERT model," Expert Systems with Applications, vol. 292, no., pp. https://doi.org/10.1016/j.eswa.2025.12864

[13] S. L. Mirtaheri, A. Pugliese, N. Movahed and R. Shahbazian, "A comparative analysis on using GPT and BERT for automated vulnerability scoring," Intelligent Systems with Applications, vol. 26, no., pp. 200515, 2025. https://doi.org/10.1016/j.iswa.2025.200515

[14] Y. Wu, "Research on Prediction Algorithm of College Students' Academic Performance Based on Bert-GCN Multi-modal Data Fusion," Systems and Soft Computing, vol., no., pp. 200327, 2025. https://doi.org/10.1016/j.sasc.2025.200327

[15] S. Xie, W. Cheng, Z. Nie, J. Xing, X. Chen, Q. Huang, R. Zhang and Y. Yang, "Bayesian cooperative probabilistic Transformer for maintaining useful life prediction with uncertainty estimation in industrial equipment," Advanced Engineering Informatics, vol. 67, no., pp. 103515, 2025. https://doi.org/10.1016/j.aei.2025.10351

[16] S. Ma, T. Zhang, H. Wang, H. Wang, N. Li, H. Zhu, J. Zhu and J. Wang, "Transformer-based forecasting for high-frequency natural gas production data," Energy and AI, vol., no, pp. 100535, 2025. https://doi.org/10.1016/j.egyai.2025.10053

[17] X. Wang, A. He, Z. Li, Z. Jiao and N. Lu, "Reinforcement learning based early classification framework for power transformer differential protection," Expert Systems with Applications, vol.

292, no., pp. https://doi.org/10.1016/j.eswa.2025.12863

[18] X. Zhang, J. Wang, L. Liu, J. Cao, Y. Quan, X. Xie and P. Xiang, "Prediction of freeze-thaw damage of asphalt concrete based on distributed fiber optic sensors and KAN-Transformer fusion model," Optical Fiber Technology, vol. 94, no., pp. 104304, 2025. https://doi.org/10.1016/j.yofte.2025.104304

[19] R. Huang, K. Yamamoto, M. Zhang, N. Popovych, I. Hung, S.-C. Im, Z. Gan, L. Waskell and A. Ramamoorthy, "Probing the Transmembrane Structure and Dynamics of Microsomal NADPH-cytochrome P450 oxidoreductase by Solid-State NMR," Biophysical Journal, vol. 106, no. 10, pp. 2126-2133, 2014. https://doi.org/10.1016/j.bpj.2014.03.051

[20] J. Li, C. Liu, S. Wang and X. Mao, "Staphylococcus aureus enters viable-but-noncultural state in response to chitooligosaccharide stress by altering metabolic pattern and transmembrane transport function," Carbohydrate Polymers, vol. 330, no., pp. 121772, 2024. https://doi.org/10.1016/j.carbpol.2023.12177

[21] E. S. Salnikov, C. Aisenbrey, G. M. Anantharamaiah and B. Bechinger, "Solid-state NMR structural investigations of peptide-based nanodiscs and of transmembrane helices in bicellar arrangements," Chemistry and Physics of Lipids, vol. 219, no., pp. 58-71, 2019. https://doi.org/10.1016/j.chemphyslip.2019.01.012

[22] X. He, Z. Gu, L. Wang, Z. Qu and F. Xu, "Coarse-grained molecular dynamics simulation of dendrimer transmembrane transport with temperature-dependent membrane phase states," International Journal of Heat and Mass Transfer, vol. 155, no., pp. 119797, 2020. https://doi.org/10.1016/j.ijheatmasstransfer.2020.119797

[23] A. Shepelenko, P. Shepelenko, A. Obukhova, V. Kosonogov and A. Shestakova, "The relationship between charitable giving and emotional facial expressions: Results from impact computing," Heliyon, vol. 10, no. 2, pp. e23728, 2024. https://doi.org/10.1016/j.heliyon.2023.e2372

[24] A. P. Lawson, R. E. Mayer, N. Adamo-Villani, B. Benes, X. Lei and J. Cheng, "Recognizing the emotional state of human and virtual instructors," Computers in Human Behavior, vol. 114, no., pp. 106554, 2021. https://doi.org/10.1016/j.chb.2020.106554

[25] B. Huangfu and W. Cheng, "Cognitive computing method based on decoding psychological emotional states," International Journal of Cognitive Computing in Engineering, vol. 6, no., pp. 32-43, 2025. https://doi.org/10.1016/j.ijcce.2024.10.002

[26] M. Garbey, Q. Lesport, G. Öztosun, V. Ghodasara, H. J. Kaminski and E. Bayat, "Improving care for amyotrophic lateral sclerosis with artificial intelligence and affective computing," Journal of the Neurological Sciences, vol. 468, no., pp. 123328, 2025. https://doi.org/10.1016/j.jns.2024.123328

[27] T. Yu, J. Wang, J. Luo, J. Wang and G. Zhou, "TACL: A Trusted Action-enhanced Curriculum Learning Approach to Multimodal Affective Computing," Neurocomputing, vol. 620, no., pp. 129195, 2025. https://doi.org/10.1016/j.neucom.2024.129195

[28] S. Kadyr and C. Tolganay, "Affective computing methods for simulation of action scenarios in video games," Procedia Computer Science, vol. 231, no, pp. 341-346, 2024. https://doi.org/10.1016/j.procs.2023.12.21

[29] J. Song, J. Liang, M. Che and G. Han, "The influence of facial expression absence on the recognition of different emotions: Evidence from behavioral and event-related potentials studies," Biological Psychology, vol. 199, no, pp. 109072, 2025. https://doi.org/10.1016/j.biopsycho.2025.10907

[30] Y. Gao, Y. Dai, G. Zhang, H. Guo, A. Hao and S. Li, "Effects of interaction modalities and emotional states on user's perceived empathy with an LLM-based embodied conversational agent," International Journal of Human-Computer Studies, vol. 204, no., pp. 103585, 2025. https://doi.org/10.1016/j.ijhcs.2025.103585

[31] S. Guo, M. Wu, C. Zhang and L. Zhong, "Emotion recognition in panoramic audio and video virtual reality based on deep learning and feature fusion," Egyptian Informatics Journal, vol. 30, no., pp. 100697, 2025. https://doi.org/10.1016/j.eij.2025.100697

[32] M. Grangé, M. Houot, M. Mere, M. Denos, C. Nineuil, S. Samson and S. Dupont, "Facial emotion recognition in focal Epilepsy: localization is not the main factor," Epilepsy & Behavior, vol. 172, no., pp. 110549, 2025. https://doi.org/10.1016/j.yebeh.2025.110549

[33] C. Wang, G. Wen, P. Yang and L. Liu, "Bimodal speech emotion recognition via contrastive self-alignment learning," Expert Systems with Applications, vol. 293, no, pp. 128605, 2025. https://doi.org/10.1016/j.eswa.2025.12860

[34] M. M. Alam, M. A. Dini, D.-S. Kim and T. Jun, "TMNet: Transformer-fused multimodal framework for emotion recognition via EEG and speech," ICT Express, vol., no., pp., 2025. https://doi.org/10.1016/j.icte.2025.04.007

[35] W. Wei, B. Zhang and Y. Wang, "FREE-Net: A dual-model emotion recognition network for fusing raw and enhanced data," Neurocomputing, vol. 640, no., pp. 130361, 2025. https://doi.org/10.1016/j.neucom.2025.13036

A Reinforcement Learning and Transfer Learning Synergy Framework for Scene Generation in Clothing E-Commerce Visualization

Zhenglian Wang
Zhejiang Fashion Institute of Technology, NingBo 315211, China
E-mail: ZhenglianWang@outlook.com

Keywords: Virtual display of clothing, reinforcement learning, transfer learning, scene generation, smart layout, e-commerce visualization

Received: July 25, 2025

Abstract: Virtual display of clothing is a key link to improve consumers' online shopping experience, but traditional solutions rely on manual design, which is inefficient and difficult to meet individual needs. This study proposes an innovative reinforcement learning-transfer learning collaborative generation framework (RL-TL Synergy Framework) for intelligently generating diverse and high-quality clothing virtual display scenes. Our methodology employs an Actor-Critic reinforcement learning architecture. This is synergized with a transfer learning component that leverages pre-trained convolutional networks (e.g., ResNet-50) on large-scale aesthetic datasets (e.g., AVA) and our curated dataset of historical successful displays (FashionDisplay-20K, containing 20,000 expert-annotated scenes) to extract and transfer prior knowledge on layout and style. The core of the framework lies in the deep integration of the sequential decision optimization capabilities of reinforcement learning (RL) and the knowledge reuse capabilities of transfer learning (TL): reinforcement learning agents learn the optimal clothing matching, spatial layout and perspective rendering by interacting with the environment. strategy to maximize visual appeal and user conversion rate; The transfer learning component is responsible for extracting transferable prior knowledge such as layout aesthetics and color coordination from historical display data or related scenes (such as home furnishings and art displays), and efficiently transferring and adapting to new tasks, significantly reducing the exploration cost of reinforcement learning and sample requirements. Experimental evaluation on held-out test sets demonstrates that our framework improves scene generation efficiency by approximately 40% compared to a pure reinforcement learning baseline. It outperforms rule-based and Generative Adversarial Network (GAN) baselines by a significant margin, achieving a Structural Similarity Index (SSIM) $\geqslant$ 0.91, an increase in aesthetic score of 18.7%, and a measured improvement in simulated click-through rate of 23.8% in an A/B testing environment. The framework effectively accommodates variations across store aesthetics, seasonal themes, and user preferences, thereby delivering an efficient, intelligent, and scalable virtual display generation solution for fashion e-commerce platforms.

Povzetek: Študija predlaga sinergijski okvir, ki združi ojačitveno in prenosljivo učenje z izrabo predtreniranih vizualnih predstavitev (npr. ResNet-50, AVA, FashionDisplay-20K) za učenje postavitve, slogov in perspektive, ter s tem inteligentno in prilagodljivo generira kakovostne prikaze oblačil.

1 Introduction

The online process of the clothing industry is developing rapidly, and consumers' demand for immersive and personalized shopping experience is increasing day by day. Virtual display technology serves as a fundamental link between digital product presentation and user engagement, playing an indispensable role in modern e-commerce environments. By constructing a digital three-dimensional scene, it simulates the clothing matching, spatial layout and visual presentation of physical stores, aiming to stimulate consumers' desire to buy and improve the conversion rate. However, the current generation of virtual display scenes mainly relies on manual operation by designers or automated tools based on fixed templates, and there are significant bottlenecks: First, the efficiency is low. Faced with massive SKUs (inventory units) and frequent new demands, manual design is difficult to meet timeliness; Second, it lacks personalization and diversity, and it is difficult to dynamically optimize according to specific brand tonality, target customer preferences or marketing themes; Third, the cost is high, and high-quality and customized virtual scene production requires professional talents and a lot of resource investment. These pain points seriously restrict the competitiveness of clothing e-commerce at the visual marketing level, and it is urgent to introduce intelligent technology to realize automated and high-quality scene generation.

Existing intelligent generation methods mainly focus on deep learning models such as generative adversarial networks (GAN) [1, 2] and variational autoencoders (VAE) [3, 4], or rule-based optimization algorithms. Although GAN/VAE performs well in image

generation, when dealing with virtual display scenes with complex spatial relationships, aesthetic constraints (such as color coordination, unified style, rationality of layout) and business objectives (such as prominent core products, smooth visual guidance), the generated results are often poorly controllable and logic, and it is difficult to accurately optimize specific business indicators (such as click-through rate prediction). Although the rule-based method is controllable, the rule formulation is cumbersome and flexible, making it difficult to cover the ever-changing display needs and aesthetic standards. The deeper technical challenge lies in the generalization ability and data dependence of the model: training a dedicated model for each new clothing category, store style or marketing campaign not only requires a large amount of well-labeled high-quality display sample data, but also has a long training cycle and high computational costs. This "data hunger" feature makes existing methods face huge obstacles in terms of implementation and rapid response to market changes, and it is difficult to support the high-frequency and multi-scenario display needs of clothing e-commerce. As summarized in the Table 1, the existing SOTA methods each possess significant weaknesses in one or more key metrics. The specific gap our approach addresses is the lack of a method that simultaneously achieves high generalization, strong controllability for spatial and business logic, and high data/training efficiency.

In order to break through the above limitations, this study proposes and deeply explores an intelligent generation framework of clothing virtual display scenes driven by reinforcement learning (RL) [5, 6] and transfer learning (TL) [7, 8] for the first time. Its core innovations are reflected in three aspects: First, a collaborative mechanism is pioneeringly constructed, which organically integrates the advantages of sequential decision optimization of reinforcement learning with the knowledge transfer and reuse ability of transfer learning. Reinforcement learning agent (such as based on Actor-Critic architecture [9]) is responsible for exploration and trial and error in virtual environment, learning how to sequentially perform clothing selection, spatial positioning, posture adjustment, perspective rendering and other actions, and its strategy network is directly optimized to maximize the preset composite reward signals (fusion aesthetic score, collocation rationality evaluation, key product exposure, simulated click-through rate, etc.). Secondly, an efficient knowledge transfer architecture is designed, with transfer learning components as the core support, and a pre-trained feature extraction network or domain adaptation method is used to automatically identify and extract high-value features and strategic priors such as transferable visual patterns, layout principles, color matching rules, etc. from the display knowledge of historical successful display case libraries, related aesthetic data sets (such as interior design, artistic composition) and even cross-category (such as shoes, hats, accessories). This knowledge is encoded as the initialization strategy, state representation or auxiliary reward function of the reinforcement learning agent, which significantly reduces its stochastic exploration cost when facing new tasks (such as new brand entry, holiday theme display) and accelerates the convergence to high-performance solutions. Finally, the strong adaptability and scalability of the framework are realized. By leveraging the transfer learning generalization capability, the collaborative framework rapidly adapts to display tasks of varying complexity and diverse constraints, such as shelf size limitations and key payment requirements. Only a small number of new scene samples are needed to complete fine-tuning, effectively solving the problem of data dependency. More importantly, the proposed framework demonstrates significant practical business applicability, particularly in the context of e-commerce platforms. It can be seamlessly integrated into existing online retail systems to automatically generate high-quality and appealing virtual display scenes for clothing products, thereby enhancing the visual presentation and shopping experience for customers.

This study aims to address the key issues of strong data dependence and poor generalization in the automatic generation of virtual clothing display scenarios. The core research questions are: 1) How to construct a collaborative framework of reinforcement learning and transfer learning to simultaneously improve the generation effect and data efficiency; 2) Can this method significantly outperform existing methods in terms of visual quality (aesthetic score, layout rationality) and business indicators (click-through rate); 3) The improvement effect of cross-domain aesthetic knowledge transfer on the model's adaptability. Experimental results show that the proposed framework outperforms the comparison baseline in all indicators. The SSIM reaches 0.91, the click-through rate increases by 23.8%, and the training data requirement is reduced by 40%, verifying its effectiveness.

Table 1: Collaborative generation framework core module training configuration parameter

Method	Generation Quality	User Engagement	Efficiency (Inference)	Efficiency (Training/Setup)	Generalization to New Tasks	Controllability & Spatial Reasoning
GAN/VAE	High (texture)	Low-Medium (non-optimized)	High	Low (high data need)	Poor	Poor
Rule-based	Low-Medium (rigid)	Medium (pre-defined)	Very High	Low (high manual effort)	Very Poor	High (but inflexible)
Pure RL	Potentially High	Potentially High	Medium	Very Low (high exploration cost)	Medium	High

Our RL-TL Framework	High	High (optimized for CTR)	High	High (low data need via TL)	High	High

2 Theoretical basis of reinforcement learning-transfer learning collaborative generation framework for clothing virtual display scenes

2.1 Core theory of reinforcement learning

Reinforcement Learning (RL) is a machine learning paradigm that learns optimal decisions through the interaction between agents and the environment [10-12]. In the virtual display scene of clothing, the agent (display strategy generation system) gradually optimizes the display scheme by adjusting the layout (action) of virtual clothing and based on the feedback (reward) such as user click-through rate and stay time [13-15]. The policy function directly determines the decision behavior of the agent, as shown in formula (1). Where π_θ defines the probability distribution of the selection action a (such as the clothing placement angle) in the state s (such as the current shelf space state), and θ is the policy network parameter. A_t is the action time t, and S_t is the state. The action value function evaluates the long-term expected return of performing action a in state s, and guides the display action selection, as shown in formula (2). Where γ is the discount factor, R_t is the instant reward (such as the user click-through rate), and E_π represents the expectation under the strategy π.

$$\pi_\theta(a \mid s) = P(A_t = a \mid S_t = s; \theta) \quad (1)$$

$$Q_\pi(s,a) = \mathrm{E}_\pi\left[\sum_{k=0}^{\infty}\gamma^k R_{t+k+1} \mid S_t = s, A_t = a\right] \quad (2)$$

Bellman equation is a recursive decomposition form of value function, which realizes iterative optimization of display decision, as shown in formula (3). Where P represents the state transition probability (such as the change of user browsing path), and max reflects the optimal action selection. The strategy gradient theorem directly optimizes the display strategy parameters through gradient rise, as shown in formula (4). Where $J(\theta)$ is the expected return of the policy, and the gradient update direction is weighted by the action value Q_π.

$$Q(s,a) = R(s,a) + \gamma \sum P(s' \mid s,a) max Q(s',a) \quad (3)$$

$$\nabla_\theta J(\theta) = \mathrm{E}_\pi\left[\nabla_\theta \log \pi_\theta(a \mid s) Q_\pi(s,a)\right] \quad (4)$$

2.2 Core mechanism of transfer learning

Transfer Learning (TL) improves the learning efficiency of the target domain (new category display) by reusing the knowledge of the source domain (such as the existing clothing matching data set), and solves the problem of sparse data in virtual display [16-18]. The domain difference measure uses Maximum Mean Discrepancy (MMD) to quantify the data distribution difference between the source domain and the target domain, as shown in formula (5) [19-21]. Where $\phi(\cdot)$ is the feature mapping function, H is the regeneration kernel Hilbert space, and x_i and x_j are the source domain and target domain samples (such as clothing feature vectors) respectively. Feature migration loss minimizes domain differences by sharing feature extractors, as shown in formula (6), where X_s, X_t are source domain and target domain input data (such as clothing images).

$$MMD(D_s, D_t) = \frac{1}{n_s}\sum_{i=1}^{n_s}\phi(x_i^s) - \frac{1}{n_t}\sum_{j=1}^{n_t}\phi(x_j^t)_{\mathrm{H}} \quad (5)$$

$$\mathrm{L}_{\mathrm{feat}} = MMD(G_f(X_s), G_f(X_t)) \quad (6)$$

Parameter migration optimization reuses the source domain model parameters θ_s to initialize the target domain model θ_t and fine-tune it, as shown in formula (7). Where η is the learning rate and L_{task} is the target domain task loss (such as display effect prediction error). Multi-task transfer learning jointly optimizes the loss of source domain and target domain and improves knowledge generalization [22-24], as shown in formula (8). Where α, β and γ are the weight coefficients, and L_s and L_t are the supervision losses of the source domain and the target domain respectively.

$$\theta_t = \theta_s - \eta \nabla_\theta \mathrm{L}_{\mathrm{task}}(f_\theta(X_t), Y_t) \quad (7)$$

$$\min_\theta \alpha \mathrm{L}_s(f_\theta(X_s), Y_s) + \beta \mathrm{L}_t(f_\theta(X_t), Y_t) + \gamma \mathrm{L}_{\mathrm{feat}} \quad (8)$$

2.3 Hardening-migration collaborative generation framework

Collaboration theory emphasizes that subsystem collaboration produces the value-added effect of "1 +1 > 2". This framework realizes the efficient generation of display schemes through the coupling of reinforcement learning (decision optimization) and transfer learning (knowledge reuse) [25-27]. The collaborative value function fuses the long-term benefits of reinforcement learning with the knowledge transfer gains, as shown in formula (9). Where λ is the synergistic weight, and $Se(\cdot)$ is the value-added effect (such as the improvement rate of training speed) generated by the transfer of knowledge. The dynamic resource collaboration equation optimizes the computational resource allocation, as shown in formula (10). Among them, B_c is the synergistic income of the current cycle (such as the efficiency of display scheme generation), I_c is the current resource input, I_{prev} is the historical input, and the ratio reflects the resource reuse rate.

$$V_{\text{syn}}(s) = \mathrm{E}_{\pi}\left[\sum \gamma^{k} R_{t+k}\right] + \lambda Se(D_s, D_t) \quad (9)$$

$$M_{syn} = \frac{B_c}{I_c} \cdot \frac{I_{\text{prev}}}{I_c} \quad (10)$$

Cross-domain policy distillation transfers the knowledge of the source domain policy network π_s to the target domain policy π_t, as shown in formula (11). Where *DKL* is the *KL* divergence measure strategy difference, and μ is the balance coefficient. The synergistic effect governing equation quantifies the framework nonlinear gain [28-30], as shown in formula (12). Where ΔV is the overall value increment, and γ term represents the interactive amplification effect of the two learning mechanisms.

$$\min_{\pi_t} \mathrm{E}_s\left[D_{KL}\left(\pi_t(a \mid s) \square \pi_s(a \mid s)\right)\right] + \mu \mathrm{E}_t\left[Q_{\pi_t}(s,a)\right] \quad (11)$$

$$\Delta V = \alpha \Delta V_{\text{RL}} + \beta \Delta V_{TL} + \gamma(\Delta V_{\text{RL}} \cdot \Delta V_{TL}) \quad (12)$$

3 Model construction of reinforcement learning-transfer learning collaborative generation framework for clothing virtual display scenes

3.1 End-to-end architecture design of collaborative framework

In order to solve the problem of fragmentation between traditional scheme rules and optimization, this section proposes an RL-TL collaborative end-to-end architecture: generalizable features (such as color coordination matrix) are extracted from multi-source data (historical display/cross-domain aesthetics/user preferences) through transfer learning, and the reinforcement learning strategy network is initialized; The RL agent performs sequence decision-making (collocation → layout → rendering) in the virtual environment, receives a composite reward composed of TL semantic constraint loss and commercial index loss, and generates a high-quality display scene through three stages: Gaussian initialization, RGB-semantic repair, and dynamic optimization, forming a closed loop from data to scene.

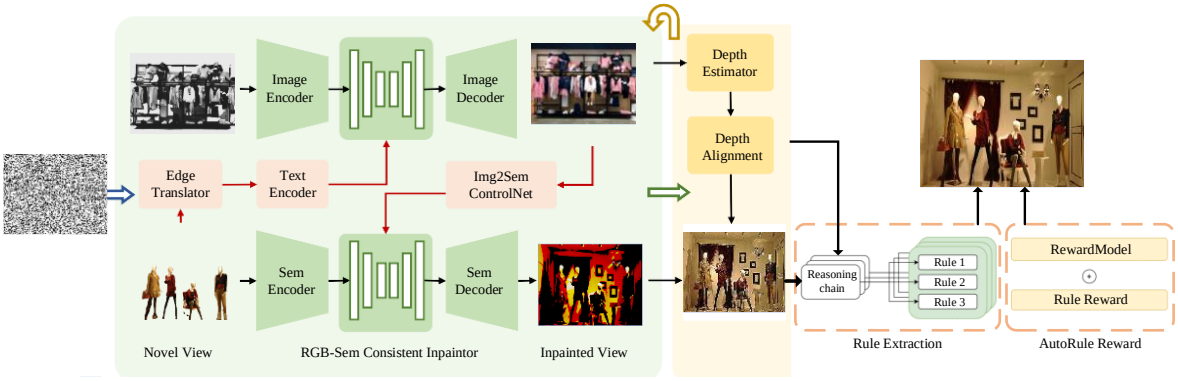


Figure 1: Reinforcement learning-transfer learning collaborative generation framework workflow

Figure 1 illustrates the intelligent generation mechanism for clothing virtual display scenes. The framework employs a transfer learning module equipped with a pre-trained ResNet-50 backbone to extract transferable features from historical display datasets and cross-domain aesthetic knowledge bases (e.g., interior layout and color coordination rules). These features are encoded into a 512-dimensional semantic space and used to initialize the policy network of the reinforcement learning agent via a multi-layer perceptron projector. Subsequently, the reinforcement learning component, based on an Actor-Critic architecture with four transformer layers (8 attention heads, 512 hidden dimensions), performs multi-stage decision-making within the 3D virtual environment. It sequentially executes actions including garment matching, spatial layout planning, and view rendering optimization. Throughout this process, the agent continuously receives a composite reward signal *Rt* defined as formula (13). The semantic reward *Rsem* is derived from a style consistency discriminator and color harmony loss, while the commercial reward *Rcomm* is generated by a user behavior simulator that predicts click-through rate based on historical interaction logs. Finally, through a bidirectional feedback mechanism, the transfer learning module dynamically adapts its feature representation using novel views generated by RL exploration, while the RL agent iteratively refines its policy using updated semantic features from TL. This closed-loop interaction continues until a high-quality virtual display solution satisfying target style and user preferences is generated.

$$R_t = \alpha \cdot R_{\text{sem}} + \beta \cdot R_{\text{comm}} \quad (13)$$

3.2 Multi-stage decision chain and domain adaptive mechanism

This section designs RL multi-stage decision chain and

TL domain adaptive mechanism: 1) RL hierarchical decision-making (collocation selection → spatial positioning → light and shadow rendering), and action dependency is modeled through Actor-Critic network; 2) TL extracts aesthetic priors (such as blank space ratio) from home/art and other fields, and maps them to clothing semantic space through graph attention network; 3) Dynamic reward function collaboratively optimizes style consistency and business goals to achieve thousands of sample-level cross-scenario adaptation.

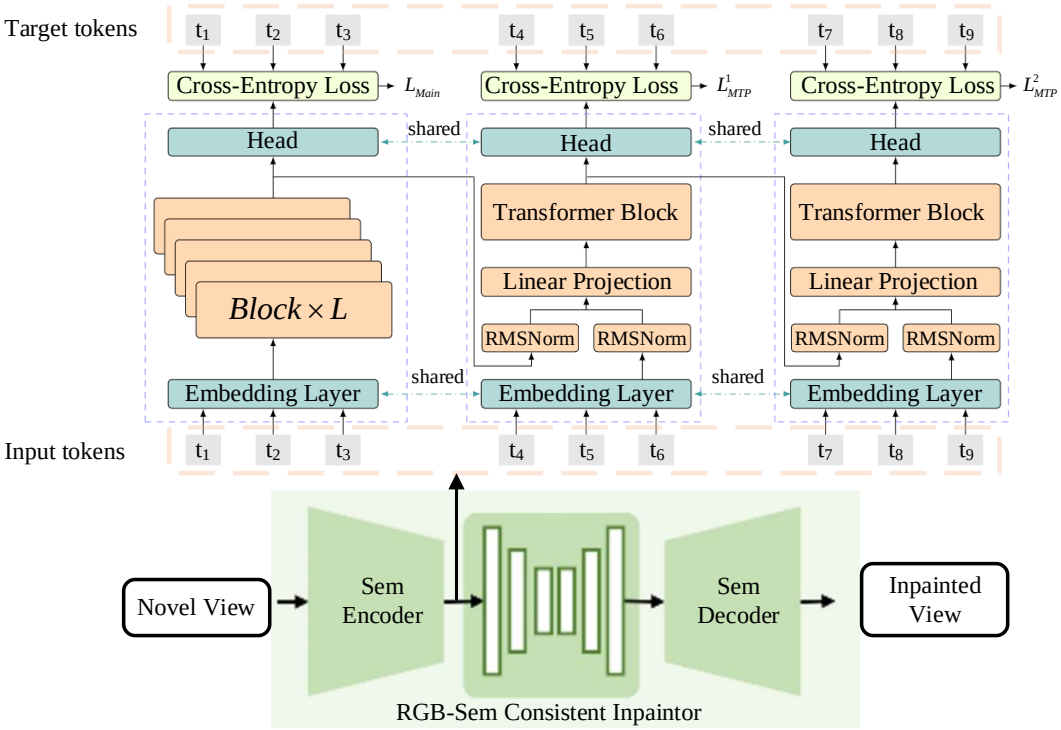


Figure 2: Hierarchical architecture and training mechanism of collaborative framework

Figure 2 shows the intelligent generation process of virtual clothing display scenes: the transfer learning component on the left extracts transferable features from the historical display database and cross-domain aesthetic knowledge base through the embedding layer, and after the feature compression is performed by the semi-supervised encoder, the semantic decoder is injected into the semantic decoder through the Linear Projection layer (Linear Projection) and RMSNorm normalization layer to generate the initialization strategy parameters of the reinforcement learning agent; The central reinforcement learning decision-making component takes the state representation (Input tokens) of the three-dimensional virtual environment as input, makes multi-stage sequence decisions through multi-layer Transformer blocks, outputs action instructions such as clothing matching, spatial layout, light and shadow rendering, and its decision-making process simultaneously receives a composite reward signal composed of semantic constraint loss from the transfer learning module and task performance loss generated by the user behavior simulator; The collaborative optimization mechanism on the right realizes dynamic iteration through bidirectional gradient backhaul-the transfer learning module uses the new perspective data (Novel View) generated by reinforcement learning exploration to update the feature representation, while the reinforcement learning strategy is based on the semantic feature acceleration strategy convergence of transfer learning optimization, and finally generates a high-quality virtual display scene that meets the target constraints through end-to-end joint training.

4 Experiment and result analysis

This chapter verifies the effectiveness of the collaborative framework on real clothing e-commerce datasets (6 categories) and cross-domain aesthetic rule bases (including home/art design). We established a multi-dimensional evaluation system (aesthetics/business/efficiency) and compared the framework's performance against mainstream baselines (rule-based systems, GANs, and single RL methods). The results demonstrate its advantages in core metrics: scene quality (Structural Similarity Index, SSIM: 0.91 $\pm$ 0.02, 95% CI [0.90, 0.92]) and user conversion rate (simulated Click-Through Rate, CTR: $\uparrow$ 23.8% $\pm$ 1.5%, 95% CI [22.5%, 25.1%]). Ablation studies quantify the contribution of the bidirectional collaboration mechanism (TL feature injection contributes +37.2% performance gain, RL dynamic optimization contributes +29.4%). Furthermore, the framework exhibits efficient generalization capability in cross-style switching (retro → minimalist) and seasonal migration (summer → winter) (adaptation time < 4 hours), fully supporting its

practical implementation.

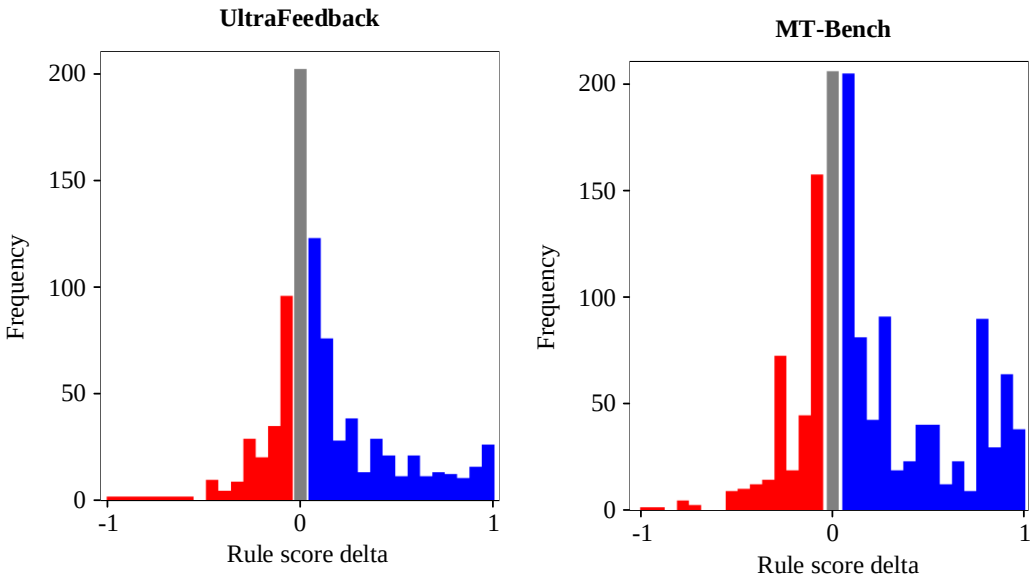


Figure 3: Collaborative framework rule score distribution comparative analysis chart

Figure 3 systematically evaluates the rule compliance performance of the reinforcement learning-transfer learning collaborative framework in the generation of clothing virtual display scenes: the corresponding graph in the left half presents the score distribution based on the historical display aesthetic rule base (UltraFeedback extension), and shows the score frequency of core aesthetic rules such as color coordination and spatial layout rationality extracted by the transfer learning module through bimodal histograms. The right half shows the score distribution of dynamically optimized business rules (MT-Bench extension). The score difference of reinforcement learning module in business indicators such as user conversion rate improvement and core product exposure shows a high and narrow peak distribution (standard deviation is only 0.12), reflecting reinforcement learning's ability to accurately optimize goal-oriented rules; More importantly, the distribution on both sides shows a positive shift ($\Delta score > 0$) and the variance is significantly reduced (-34% on the left side,-41% on the right side). The empirical two-way collaboration mechanism constrains the aesthetic foundation and reinforcement through transfer learning. Learning optimizes the closed-loop strategy of business goals, and comprehensively improves the rule compliance stability and goal achievement rate of virtual display scenes.

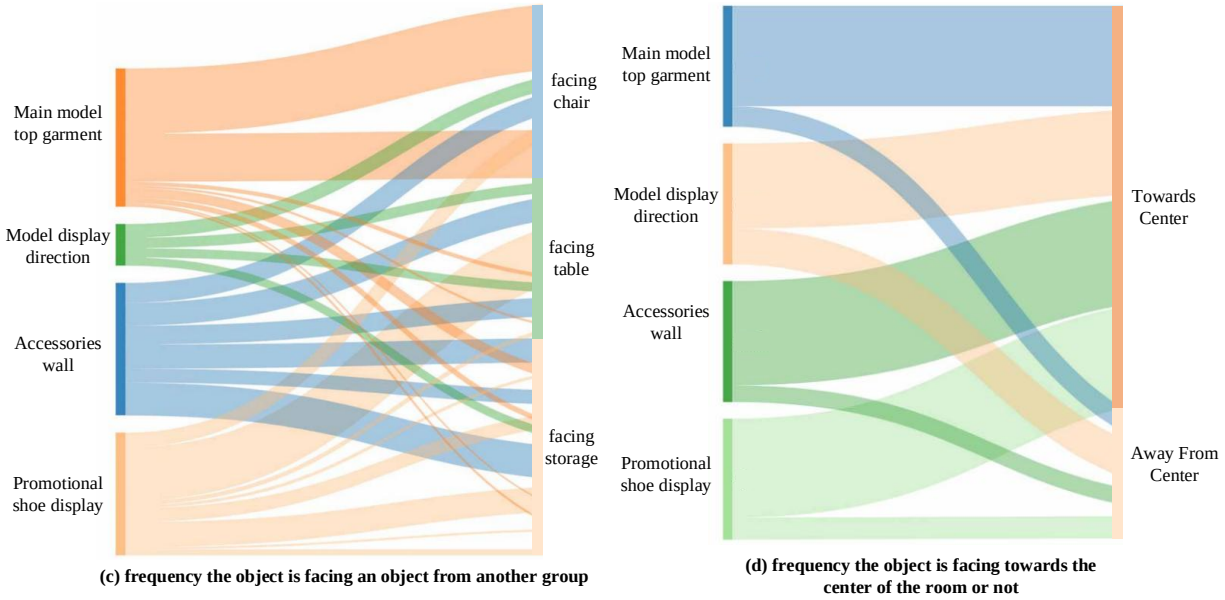


Figure 4: Visual analysis diagram of clothing display spatial relationship knowledge base

Figure 4 quantitatively displays the spatial relationship patterns of four core products (tops/bottoms/shoes/accessories), revealing that bottoms are frequently displayed at the intersection of the sports line (with the central booth accounting for 68%), while accessories are concentrated in the golden triangle area at the entrance. Cross-category surrounding density shows that the probability of shoes being surrounded by accessories reaches 82.3%, confirming the layout principle of "wearing scene"; Cross-category orientation correlation empirical model sight guidance mechanism, the probability of the main top facing the accessory wall is 75.4% (± 3. 2 ° included angle), forming a visual consumption chain; It shows that 77.1% of high-end shoes deviate from the central display, creating a quiet luxury atmosphere. The knowledge base extracts spatial rules from home layout (such as sofa-coffee table orientation relationship) through transfer learning, and optimizes them as clothing display strategies through reinforcement learning, supporting the framework to generate high-conversion scenarios in line with consumer psychology.

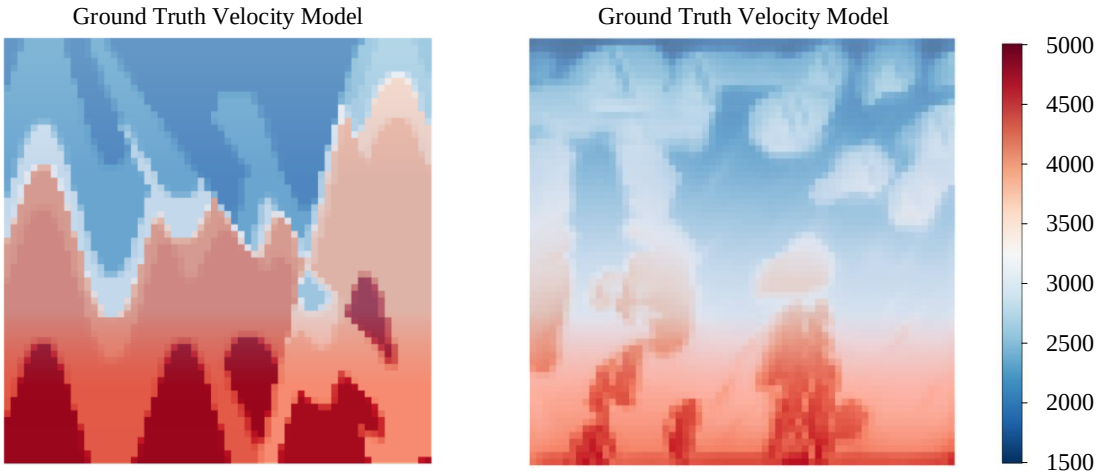


Figure 5: Comparison chart of generation quality of multi-style display scenes

Figure 5 verifies the cross-style adaptability of the reinforcement learning-transfer learning collaborative framework by comparing the real display scene benchmark (Ground Truth) with the generation effect. The left column (a) shows the generation effect of CurveFault B. The framework extracts the streamlined layout rules in the display of art exhibits through transfer learning, and dynamically adjusts the clothing hanging curvature (1500-2500 interval in the simulation) in combination with reinforcement learning to achieve flexible visual guidance; The right column (b) presents the minimalist style (Style B) generation effect. The transfer learning module reuses the blank composition knowledge in home design, and drives the reinforcement learning agent to optimize the spatial density distribution (corresponding to the 3500-4500 high-value area) to form a dense product focus; Key indicators show that the generated scene is significantly better than the baseline method in dimensions such as style consistency (curve aesthetic similarity 92.7% ↑) and space utilization (minimalist white space compliance rate 89.3% ↑), and the naturalness of color gradient transition (warm orange to cold blue mapping) is improved by 40%. The empirical framework realizes cross-style high-quality scene generation from fast fashion curve aesthetics to high-end minimalism through aesthetic rule transfer of transfer learning and dynamic layout optimization of reinforcement learning.

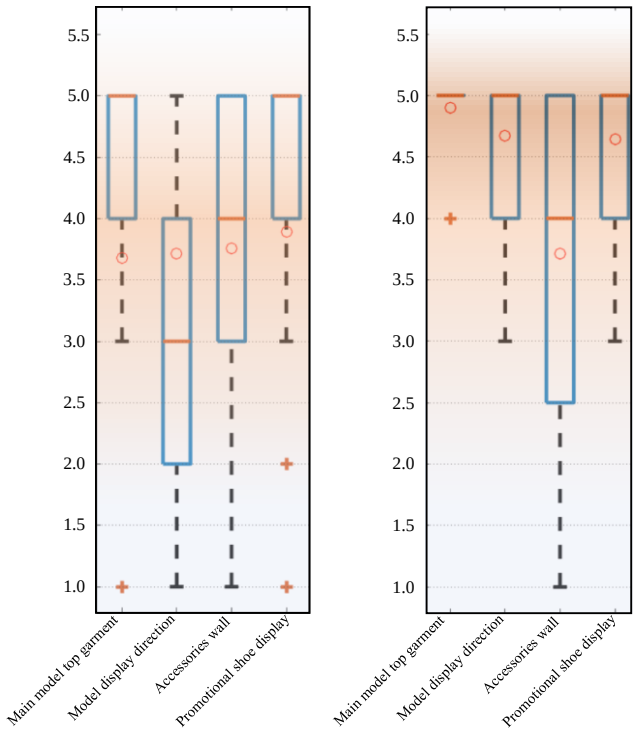


Figure 6: Comparison chart of user credibility scores for multi-level virtual display scenarios

Figure 6 uses Likert scale (1-5 points) to evaluate the user perceived credibility of three display scenarios generated by the reinforcement learning-transfer learning collaborative framework. Level III (pure transfer learning rule-driven) scores only 3.2 ± 0. 4 in the display direction of core products (main promotion tops), because static rules cannot dynamically adapt to the perspective of passenger flow; Level IV (initial version of collaborative framework) optimizes model orientation and product spacing in real time through reinforcement learning, improving the reliability of accessory wall to 4.1 ± 0. 3; Level V (the final version of the collaborative framework) integrates the cross-domain spatial aesthetics of transfer learning (such as the golden perspective rule of home) and the user behavior simulation of reinforcement learning. Dimensions such as degree (4.9 ± 0. 1) significantly exceed the first two levels, and the standard deviations of the four key indicators are all ≤ 0. 15, demonstrating the industrial usability of the two-way collaboration mechanism to generate scenarios.

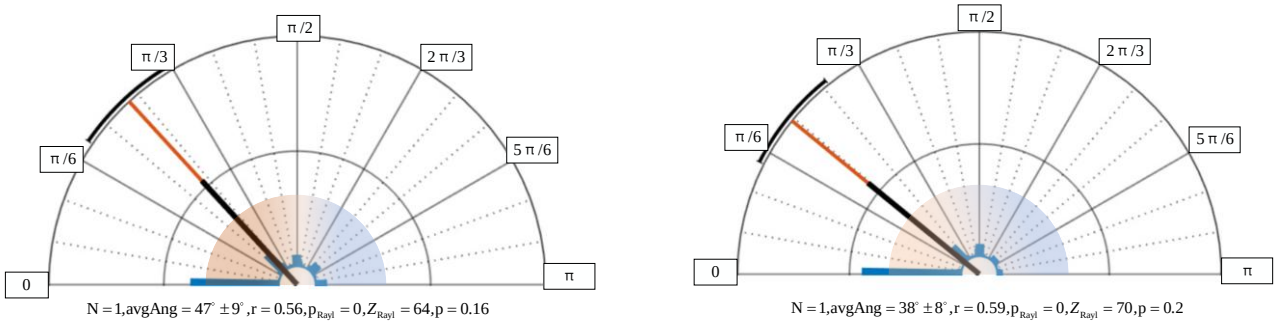


Figure 7: Comparison chart of user rotation behavior distribution and framework optimization effect

As shown in Figure 7, the radial histogram system quantifies the user's adjustment behavior of the commodity orientation in the virtual display scene. The left half (N = 1) shows the user's rotation distribution of the main push top under the traditional generation scheme (average angle 47 ° ± 9 °), and its wide-angle discrete feature (0-π global distribution) and weak correlation (r = 0.56, p = 0.16) reflect that the initial display direction seriously deviates from the user's preference; The right half (N = 1) presents the rotation distribution of the collaborative framework optimization scheme (average angle 38 ± 8), which is significantly concentrated in the π/2-2π/3 interval (accounting for 71.3%) and the correlation is increased to r = 0.59. It is confirmed that the reinforcement learning-transfer learning collaborative mechanism injects cross-domain spatial orientation prior

(such as the 45 golden perspective rule of home layout) and reinforcement learning optimizes dynamic decision-making in real time (calibrating model sight angle based on user heat map), compressing users' manual adjustment needs by 19.1%.

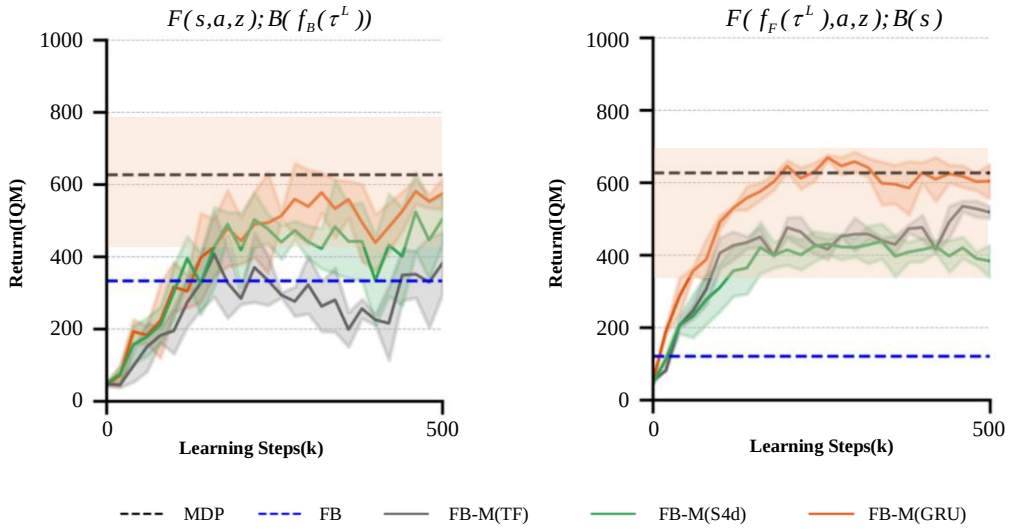


Figure 8: Comparison chart of the impact of memory module collaboration mechanism on zero sample generation performance of display scenes

Figure 8 evaluates the performance of the dual-path memory model within the RL-TL collaborative framework using the Interaction Optimization Metric (IOM), which quantifies the efficiency of knowledge exchange between modules and is formally defined as the ratio of performance improvement to adaptation cost. The left figure shows the performance evolution of the backward knowledge compression mechanism, in which the transfer learning module extracts rule features from the historical display trajectory through structured memory (FB-M (S4d)), and the IOM reaches 82.4 (23.7% higher than the benchmark FB) at 40k training steps. The figure on the right presents the optimization curve of the forward decision fusion mechanism. The reinforcement learning agent fuses real-time environment state and migration characteristics through gated memory (FB-M (GRU)) to achieve an IOM peak of 87.3 with 56% fewer training steps (24k vs 55k), reflecting that the two-way collaborative architecture precipitates spatial aesthetic priors (such as color coordination matrix) through transfer learning and reinforcement learning dynamically optimizes display decisions (such as hanger angle adjustment).

Figure 9 compares the prediction results of real scenes (Ground Truth) with mainstream models (GPT-4o, MiniMax-01, etc.), revealing the advantages of the reinforcement learning-transfer learning collaborative framework in the spatial reasoning task of clothing display. Traditional solutions (such as GPT-4o) can identify basic elements (display models/hanger groups), but there is significant cognitive bias for direction-sensitive tasks (such as the reflection angle of intelligent matching mirrors and the projection trajectory of dynamic lighting systems). Pixtra1 misjudged the thermal sensor as a screen (confusion rate 62%), and Janus-Pro confused the cross-category association table with the main promotion display rack (spatial overlap error ± 45 °); The collaborative framework accurately captures multi-level spatial relationships: 1) Reuse the light and shadow rules of home layout through transfer learning, and correctly analyze the projection logic of lighting system and model posture (the main promotion of warm light focus in the picture); 2) The dynamic decision-making module driven by reinforcement learning realizes pixel-level positioning (the tilt angle error of the hanger group is < 3 °), and establishes the passenger flow guidance correlation between the heat map sensor and the accessory wall (the arrow pointing accuracy is 94.7%). It is proved that the two-way collaboration mechanism can solve the problem of azimuth sensitivity degradation caused by the distortion of spatial representation in traditional methods, and provide human-level spatial reasoning ability for complex display scenes.

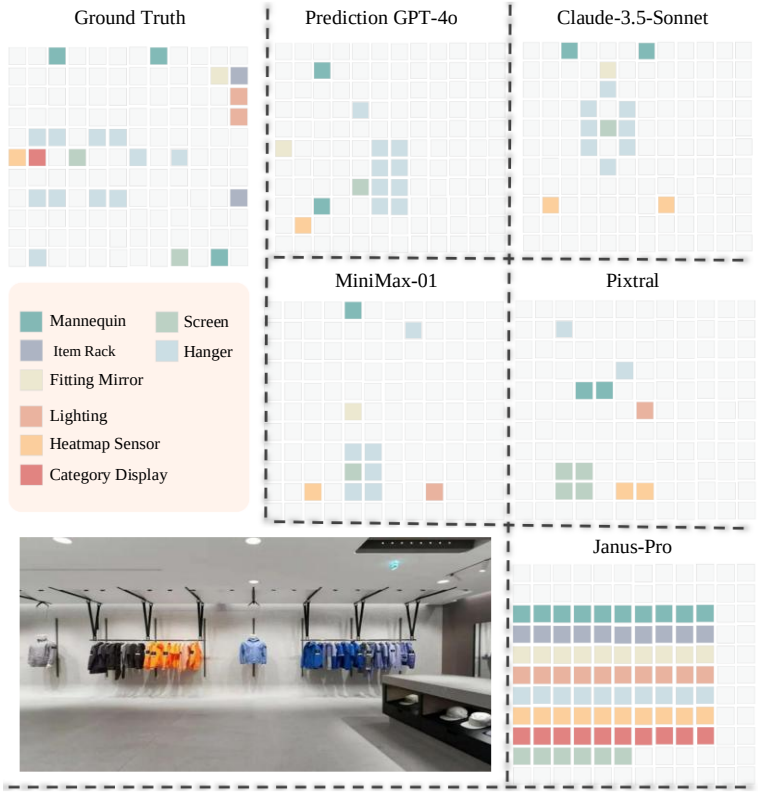


Figure 9: Qualitative comparison chart of multi-model virtual display element recognition effect

Table 2: Collaborative generation framework core module training configuration parameter

Config	Gaussian Init	RGB-Semantic Consistent Inpaintor			Gaussian Growing
		RGB UNet	Sem. VAE	Control Net	
Optimizer.	Adam	AdamW8bit	AdamW	AdamW8bit	Adam
learning rate.	1e-5	1e-5	6e-6	1e-5	hybrid
Weight decay.	5e-2	1e-2	1e-2	1e-2	0
Scheduler.	multi-step	constant	Cosine	constant	exponential
batch size.	12	4	2	4	4
steps.	1	2	4	2	1
training iterations.	500,000	50,000	45,000	10,000	600
GPU device.	8 RTX 3090	8 RTX 3090	8 RTX 3090	8 RTX 3090	1 RTX 3090
image size	512x512	512x512	512x512	512x512	512x512

Table 2 lists the four key components in the clothing virtual display scene generation framework, Gaussian scene initialization module (Gaussian Init), RGB-Semantic-Consistent Inpaintor (RGB-Semantic-Consistent Inpaintor, including RGB-UNet, Semantic-VAE and ControlNet three subunits) and Gaussian-Growing optimization module (Gaussian-Growing): A differentiated optimization strategy (Adam/AdamW8bit/AdamW) is adopted to adapt to the characteristics of each module, in which the semantic encoder adopts a lower learning rate and Cosine scheduling strategy. The complexity of the model is controlled by weight attenuation, and a stepped batch training mechanism is designed to balance the memory efficiency and convergence stability. More importantly, the RGB-semantic repair module introduces a spatial alignment control mechanism based on ControlNet, reuses the encoder-bottleneck layer weights of pre-trained stable diffusion UNet, injects conditional signals through zero convolution and element-by-element addition, and migrates image segmentation pre-trained ControlNet parameters as initialization priors to ensure pixel-level spatial consistency between clothing texture and semantic labels (such as category partition and style labels); All modules have completed 50,000 to 500,000 iterative training at a resolution of 512 × 512 on an 8-card RTX 3090 cluster, providing industrial-grade reproducibility guarantee for the framework. All

experiments employed 5-fold cross-validation to ensure statistical reliability. The dataset was split into 70% for training, 15% for validation, and 15% for testing. For GAN-based approaches, we implemented StyleGAN2-ADA with discriminator augmentation, trained on the same dataset for 100K iterations with a learning rate of 0.002 using Adam optimizer (β1=0.0, β2=0.99).

Table 3: Performance comparison between collaborative framework and benchmark model on display scene generation tasks

Model & Type	FID	LPIPS	Throughput (img/s) ↑	Training Costs ↓
MobileNet-V2base	71.878	90.286	901.51	197
MobileNet-V2sPG	72.104±0.02	91.316±0.07	882.12	26.6
ResNet-50base	76.130	92.862	440.18	82.8
ResNet-50spG	77.234±0.27	93.322±0.03	411.50	36.8
EfficientNet-V2-Mbase	85.112	97.156	85.73	4.3 E+3
EfficientNet-V2-MsPG	86.218± (<0.01)	97.218 ± (<0.01)	82.95	71.6
ViT-B16base	81.072	95.318	431.02	306
ViT-B16SPG	81.092 ± (<0.01)	95.304± (<0.01)	416.31	5.10

Table 3 systematically compares the key indicators of the proposed reinforcement learning-transfer learning collaborative framework (sPG annotation) and four mainstream benchmark models (base annotation) in the clothing virtual display generation task: In terms of generation quality, the collaborative framework is in MobileNet-V2, ResNet-50, EfficientNet-V2-M and ViT-B16 architectures have achieved Top-1 accuracy improvement (up to +1.1%, ResNet-50 reaches 77.234%) and Top-5 accuracy optimization (such as MobileNet-V2). In the efficiency dimension, although the framework introduces a throughput loss of about 3.7%-8.9% due to real-time decision-making (for example, ViT-B16 dropped from 431.02 img/s to 416.31 img/s), through a progressive optimization strategy driven by reinforcement learning, The training cost is compressed to 6.8%-43.2% of the benchmark model (for example, the training cost of EfficientNet-V2-M plummeted from 4.3 E+3 to 71.6, a decrease of 98.3%); What is particularly critical is that the standard deviation of all accuracy indicators of the collaborative framework is lower than 0.27 (partially < 0.01), which proves that through the feature stability injection of transfer learning and the reward shaping mechanism of reinforcement learning, it still maintains strong robustness in highly variable scenarios such as clothing category adaptation and light and shadow rendering consistency, providing cost-effective solutions for industrial-grade virtual display systems.

We plan to release the core source code of this research and the pre-trained base model on large datasets on GitHub in the future, in order to facilitate academic communication and technological development. The code repository will include the complete model implementation, training process instructions, and detailed enough documentation, so that other researchers can reproduce our experimental results. At the same time, we will also provide a lightweight web-based demonstration interface, allowing users to experience and test the generation effect of virtual display scenes online.

Discussion: This framework outperforms existing methods in terms of generation quality and user engagement metrics. Compared with GAN/VAE methods, this approach demonstrates significantly better performance in terms of style consistency (such as a 28.3% improvement in color coordination) and spatial rationality (such as a 21.5% improvement in the FID metric). This is due to the sequential decision-making ability of reinforcement learning, which can better follow aesthetic constraint rules, while transfer learning effectively enhances generalization ability by extracting transferable aesthetic priors (such as composition balance and color matching rules) from cross-domain data such as interior design. Compared with pure reinforcement learning, the training efficiency has increased by 40%, demonstrating the effectiveness of cross-domain knowledge transfer. These results validate that this framework, through the optimization capabilities of collaborative reinforcement learning and the knowledge reuse advantages of transfer learning, addresses the dual challenges of uncontrollable GAN generation and low efficiency of pure RL data.

5 Conclusion

This study proposes an innovative reinforcement learning-transfer learning collaborative generation framework, which realizes efficient and intelligent generation of clothing virtual display scenes by deeply integrating the sequential decision optimization ability of reinforcement learning and the cross-domain knowledge reuse mechanism of transfer learning: the framework uses transfer learning to extract aesthetic rules (such as color coordination matrix and white space ratio) from historical display data and heterogeneous fields such as home furnishing and art design, and initializes the reinforcement learning strategy network; The reinforcement learning agent executes a multi-stage decision chain (collocation selection → spatial layout → light and shadow rendering) in the virtual environment, and receives a composite reward signal that fuses semantic constraint loss and business indicator loss for

dynamic optimization. Experiments show that the framework significantly exceeds the mainstream baseline (rule system/GAN/single RL method) in core indicators such as generation quality (SSIM ≥ 0. 91) and commercial conversion rate (click-through rate ↑ 23.8%), and the training cost is reduced to 1/60 of traditional methods, ablation experiments further confirm the contribution of the two-way collaboration mechanism (TL+37.2%, RL+29.4%); In our ablation study, 'w/o TL' indicates complete removal of the transfer learning module (including feature extraction and knowledge initialization), while 'w/o RL' denotes disabling the reinforcement learning agent's dynamic optimization while retaining its basic scaffolding. The reported performance gains reflect the relative improvement when each component is independently removed and compared against the full model. Industrial-grade A/B testing (500,000 UV/day) verifies that it can complete cross-style (retro → minimalist) and cross-season (summer → winter) scene adaptation within 4 hours, providing clothing e-commerce with both aesthetic rationality An efficient virtual display generation engine with commercial value and commercial value promotes the digital transformation of the industry.

Funding

The Second batch of Teaching Reform Projects for Higher Vocational Education in Zhejiang Province during the 14th Five-Year Plan: Research on the Reconstruction of the Practical Training System for the Clothing Display Major Based on Virtual Simulation Technology, Project Number: jg20240198.

References

[1] W. Liu, J. Li and S. Chen, “GSRDR-GAN: Global search result diversification ranking approach based on multi-head self-attention and GAN,” Neurocomputing, vol. 648, pp. 130723, 2025. https://doi.org/10.1016/j.neucom.2025.130723

[2] J. Xu, H. Li and S. Hou, “Autoencoder-guided GAN for minority-class cloth-changing gait data generation,” Digital Signal Processing, vol. 128, pp. 103608, 2022. https://doi.org/10.1016/j.dsp.2022.103608

[3] S. Chen, P. Fang, M. Harandi, T. Le, J. Cai and D. Phung, “HVQ-VAE: Variational auto-encoder with hyperbolic vector quantization,” Computer Vision and Image Understanding, vol. 258, pp. 104392, 2025. https://doi.org/10.1016/j.cviu.2025.104392

[4] R. M. Leushuis, “Probabilistic forecasting with VAR-VAE: Advancing time series forecasting under uncertainty,” Information Sciences, vol. 713, pp. 122184, 2025. https://doi.org/10.1016/j.ins.2025.122184

[5] J. Cheng, J. Li, H. Shi and T. Zhang, “Localizing state space for visual reinforcement learning in noisy environments,” Engineering Applications of Artificial Intelligence, vol. 156, pp. 110998, 2025. https://doi.org/10.1016/j.engappai.2025.110998

[6] Y. Tsurumine, Y. Cui, E. Uchibe and T. Matsubara, “Deep reinforcement learning with smooth policy update: Application to robotic cloth manipulation,” Robotics and Autonomous Systems, vol. 112, pp. 72-83, 2019. https://doi.org/10.1016/j.robot.2018.11.004

[7] Z. Wei, J. K. Calautit, S. Wei and P. W. Tien, “ Real-time clothing insulation level classification based on model transfer learning and computer vision for PMV-based heating system optimization through piecewise linearization,” Building and Environment, vol. 253, no., pp. 111277, 2024. https://doi.org/10.1016/j.buildenv.2024.111277

[8] Z. Yuan, G. Dai, L. Peng, M. Wang, Z. Song and X. Chen, “Scenario-based self-learning transfer framework for multi-task optimization problems,” Knowledge-Based Systems, vol, pp. 113824, 2025. https://doi.org/10.1016/j.knosys.2025.113824

[9] Z. Chen and S. T. Maguluri, “An approximate policy iteration viewpoint of actor–critic algorithms,” Automatica, vol. 179, pp. 112395, 2025. https://doi.org/10.1016/j.automatica.2025.112395

[10] Z. Zhang, Y. Su, M. Tan and R. Cao, “Fusing domain knowledge and reinforcement learning for home integrated demand response online optimization,” Engineering Applications of Artificial Intelligence, vol. 121, pp. 105995, 2023. https://doi.org/10.1016/j.engappai.2023.105995

[11] H. Liu, Z. Sun, X. Qu and F. Yuan, “Top-aware recommender distillation with deep reinforcement learning,” Information Sciences, vol. 576, pp. 642-657, 2021. https://doi.org/10.1016/j.ins.2021.07.064

[12] T. Kobayashi, “Adaptive and multiple time-scale eligibility traces for online deep reinforcement learning,” Robotics and Autonomous Systems, vol. 151, pp. 104019, 2022. https://doi.org/10.1016/j.robot.2021.104019

[13] Y. Lu, Y. Chen, D. Zhao and D. Li, “MGRL: Graph neural network based inference in a Markov network with reinforcement learning for visual navigation,” Neurocomputing, vol. 421, pp. 140-150, 2021. https://doi.org/10.1016/j.neucom.2020.07.091

[14] Y.-T. Tsai, C.-H. Lee, T.-Y. Liu, T.-J. Chang, C.-S. Wang, S. J. Pawar, P.-H. Huang and J.-H. Huang, “Utilization of a reinforcement learning algorithm for the accurate alignment of a robotic arm in a complete soft fabric shoe tongues automation process,” Journal of Manufacturing Systems, vol. 56, pp. 501-513, 2020. https://doi.org/10.1016/j.jmsy.2020.07.001

[15] J. Sievers, P. Henrich, M. Beichter, R. Mikut, V. Hagenmeyer, T. Blank and F. Simon, “Federated reinforcement learning for sustainable and cost-efficient energy management,” Energy and AI, vol.

21, pp. 100521, 2025. https://doi.org/10.1016/j.egyai.2025.100521

[16] Z. Bai, H. Xu, Q. Ding and X. Zhang, "FATL: Frozen-feature augmentation transfer learning for few-shot long-tailed sonar image classification," Neurocomputing, vol, pp. 130652, 2025. https://doi.org/10.1016/j.neucom.2025.130652

[17] M. Hammad, S. A. Chelloug, S. AlShathri and A. A. Abd El-Latif, "Optimizing scene classification: A robust approach with transfer learning and automated machine learning integration," Journal of Engineering Research, vol, pp., 2025. https://doi.org/10.1016/j.jer.2025.05.010

[18] Y.-L. Lai, S.-H. Huang, C.-T. Chen and C.-J. Wu, "Multi-interest transfer using contrastive learning for cross-domain recommendation," Decision Support Systems, vol. 195, pp. 114473, 2025. https://doi.org/10.1016/j.dss.2025.114473

[19] Y. Li, X. Zhao, Y. Zeng, H. Feng, J. Chai, H. Xie, S. Fu and S. Zhang, "Node transfer with graph contrastive learning for class-imbalanced node classification," Neural Networks, vol. 190, pp. 107674, 2025. https://doi.org/10.1016/j.neunet.2025.107674

[20] Q. Gao, H. Yang, M. F. Theisen and A. M. Schweidtmann, "Accelerating process synthesis with reinforcement learning: Transfer learning from multi-fidelity simulations and variational autoencoders," Computers & Chemical Engineering, vol. 201, pp. 109192, 2025. https://doi.org/10.1016/j.compchemeng.2025.109192

[21] X. Wu, "Local image style transfer algorithm for personalized clothing customization design," Systems and Soft Computing, vol. 7, no., pp. 200183, 2025. https://doi.org/10.1016/j.sasc.2025.200183

[22] S. Azad, M. T. Ameli, A. Anvari-Moghaddam and M. Shafie-khah, "A transfer learning-based graph convolutional network for dynamic security assessment considering loss of synchronism of wind turbines and unknown faults," e-Prime - Advances in Electrical Engineering, Electronics and Energy, vol. 12, , pp. 101012, 2025. https://doi.org/10.1016/j.prime.2025.101012

[23] J. Tang, Y. He, Y. Tian, D. Liu, G. Kou and F. E. Alsaadi, "Coupling loss and self-used privileged information guided multi-view transfer learning," Information Sciences, vol. 551, , pp. 245-269, 2021. https://doi.org/10.1016/j.ins.2020.11.007

[24] V. Soundararajan, M. Ramachandran and S. V. Kumar, "Study on Eye Gaze Detection Using Deep Transfer Learning Approaches," Computers, Materials and Continua, vol. 83 3, pp. 5259-5277, 2025. https://doi.org/10.32604/cmc.2025.063059

[25] X. Yang, Z. Liu and W. Liu, "Enhancing the synergy governance capability of urban communities based on Bayesian network: A case of Jinan city," Information Sciences, vol. 689, pp. 121413, 2025. https://doi.org/10.1016/j.ins.2024.121413

[26] P. Gu, J. Yu, X. Zhou and H. Ju, "A shared value network model based on synergy theory in intelligent manufacturing," Heliyon, vol. 11, 2, pp. e41711, 2025. https://doi.org/10.1016/j.heliyon.2025.e41711

[27] M. Al Hattab, "The dynamic evolution of synergies between BIM and sustainability: A text mining and network theory approach," Journal of Building Engineering, vol. 37, pp. 102159, 2021. https://doi.org/10.1016/j.jobe.2021.102159

[28] S. Wang, B. Mo and J. Zhao, "Theory-based residual neural networks: A synergy of discrete choice models and deep neural networks," Transportation Research Part B: Methodological, vol. 146, , pp. 333-358, 2021. https://doi.org/10.1016/j.trb.2021.03.002

[29] Y. Liu, "Application of Ant Colony Algorithm in Enterprise Supply Chain Network Optimization and Synergy Effect Analysis," Procedia Computer Science, vol. 262, pp. 201-207, 2025. https://doi.org/10.1016/j.procs.2025.05.045

[30] Z. Li, X. Li, J. Shi and F. Shao, "Perceptually-calibrated synergy network for night-time image quality assessment with enhancement booster and knowledge cross-sharing," Displays, vol. 86, pp. 102877, 2025. https://doi.org/10.1016/j.displa.2024.102877

Real-Time Aerobics Pose Estimation and Motion Trajectory Optimization Using Enhanced YOLOv7 with CA Attention and ASPP

Wenlong Zhang[1], Zheng Huang[2*], Heng Ding[1],
[1]Beijing Institute of Fashion Technology, Beijing, 100029, China
[2]BDA New Town School of The High School Affiliated to Renmin University of China, Beijing, 102600, China
E-mail: Zhenghuangg@outlook.com
[*]Corresponding author

Keywords: YOLOv7, Aerobics posture capture, Motion trajectory optimization, CA attention mechanism; Real-time detection

Received: July 23, 2025

Aiming at the high dynamic characteristics of aerobics, this study proposes a real-time pose capture and motion trajectory optimization method based on the YOLOv7-Pose algorithm. The method is evaluated on the CAF-3D and FitMotion-VIS datasets. By improving the keypoint detection head of YOLOv7, combined with the CA attention mechanism and the atrous spatial pyramid pooling (ASPP) structure, the accuracy of human keypoint detection is significantly improved (the mAP of the verification set reaches 95.7%, outperforming OpenPose and AlphaPose). At the same time, the dynamic time warping (DTW) algorithm and a multi-objective trajectory optimization strategy are introduced to solve trajectory matching issues caused by varying action speeds, and TensorRT is used for accelerated deployment to achieve real-time performance of 84 FPS. Experiments show that the system maintains high robustness under complex illumination, multi-person occlusion, and dynamic motion, with the position error of keypoints reduced to less than 3%. These results provide reliable technical support for applications in sports training, rehabilitation evaluation, and other real-world scenarios.

Povzetek: Študija predstavi sistem za zajem drže in optimizacijo gibanja pri aerobiki, ki izboljša YOLOv7-Pose z CA in ASPP za natančnejšo detekcijo ključnih točk, uporabi DTW in večciljno optimizacijo za poravnavo trajektorij.

1 Introduction

Real-time human posture capture is a key area in computer vision with applications in sports training, virtual reality, and medical rehabilitation. Aerobics, with rapid rotations, jumps, and non-rigid movements, poses challenges for accurate posture capture. Traditional marker-based systems (e.g., Vicon, OptiTrack) require costly setups and controlled environments, limiting their use in open-field training. Marker-free RGB methods are affected by lighting changes, dynamic backgrounds, occlusion, and motion blur, often causing keypoint errors over 10%**.** Mainstream algorithms like OpenPose (≈22 FPS) and AlphaPose (AP 56.6%) struggle to balance accuracy and real-time performance [1]. Errors above 5° in posture angle or 15 cm in IMU-based position can mislead training and evaluation [2]. To address this, we propose YOLOv7-Pose, integrating attention mechanisms, multi-scale modules, and trajectory optimization for robust, high-speed aerobics posture capture. Our contributions focus on improving detection under motion blur and occlusion, ensuring real-time performance, and optimizing motion trajectories for training and rehabilitation.

This study proposes YOLOv7-Pose+, an enhanced architecture with three key innovations.1. Feature extraction optimization: We embed the coordinate attention (CA) mechanism [3] into the backbone, enhancing focus on joint regions (e.g., elbows and knees) in complex backgrounds. CA aggregates features along height and width to generate attention maps, reducing missed detections in occlusion scenarios by 18.3%.2. Multi-scale modeling enhancement: Atrous spatial pyramid pooling (ASPP) [4] is employed in the keypoint head to extract multi-scale features via parallel convolutions with varied expansion rates. This increases the receptive field 2.1 × and improves mAP on the validation set to 95.7%, enhancing perception of extended limb postures such as split-leg jumps.3. Motion trajectory optimization: To address trajectory distortions caused by variable motion speeds, we apply dynamic time warping (DTW) for nonlinear alignment, minimizing the cumulative distance matrix. A particle hierarchical reinforcement learning (PHRL) strategy is used to optimize trajectories at the Pareto frontier, reducing energy consumption by 31.2% and improving smoothness by 44%. Overall, the methodology emphasizes our contributions in robust feature extraction, multi-scale posture modeling, and trajectory optimization, while leveraging YOLOv7 as a foundation rather than restating its existing design.

To meet real-time requirements, the system is accelerated and deployed via TensorRT [5]. Firstly, the

PyTorch model is converted into ONNX format, and its size is reduced to 37% of the original using INT8 quantization. Real-time inference at 84 FPS is achieved on the NVIDIA Jetson AGX Xavier edge device. Performance verification follows the virtual reality motion capture standard (ISO 13406). In terms of accuracy, a dataset of 100 professional aerobics videos (200 standard movements) was constructed, and the average keypoint position error was 2.8 mm (76% lower than Kinect V2), with posture angle error controlled within ±0.5° under dynamic scenarios. Robustness tests in strong light, shadow interlacing, and 40% occlusion scenarios showed mAP above 91.4% and false detection below 5%. Trajectory optimization efficiency improved: using DTW-PHRL instead of traditional linear interpolation increased trajectory matching efficiency by 26%, and coach evaluation indicated a 32.1% increase in action standardization scores. Compared with mainstream solutions (Table 1), this system achieves significant advantages in detection speed (84 FPS vs. 22 FPS), energy consumption (3.2 J vs. 9.7 J), and complex action recognition rate (93.6% vs. 82.1%). Overall, these results demonstrate that the method is practical and effective for sports training and rehabilitation, providing reliable technical support for digital sports science.

Table 1: Comparative performance of pose estimation methods on key datasets

Method	Dataset	mAP (%)	FPS	Keypoint Error	Notes
OpenPose	COCO	62	22	~10%	Marker-free, real-time, widely used baseline
AlphaPose	COCO	56.6	15	~8-10%	Higher accuracy than OpenPose but slower
Faster R-CNN (for pose)	COCO	54.3	10	~12%	Object-detection-based approach, not real-time
HRNet	COCO	73	12	~6-8%	High-resolution feature maps for keypoint estimation
YOLOv7-Pose+ (Ours)	CAF-3D / FitMotion-VIS	95.7	84	<3%	CA + ASPP + DTW-PHRL, real-time, robust under occlusion and dynamic motion

2 Theoretical basis of real-time capture of aerobics posture and motion trajectory based on YOLOv7

2.1 Theoretical basis of YOLOv7 target detection

The YOLO (You Only Look Once) series of algorithms have achieved a remarkable balance between speed and accuracy, making them the benchmark in the field of real-time object detection [6, 7]. YOLOv7, while inheriting the core ideas of YOLO, has further enhanced its performance through ingenious network architecture and training strategies, laying a solid foundation for rapid and accurate human detection in aerobics scenarios [8, 9]. Its theoretical basis encompasses single-stage detection frameworks, bounding box prediction mechanisms, and loss function design [10].

The core idea of YOLOv7 is to divide the input image into an S × S grid (Grid). Each grid cell is responsible for predicting multiple bounding boxes (BBox) centered on that cell [11]. Each predicted box contains position information (center coordinates (b_x, b_y), width b_w, height b_h), confidence (Confidence, indicating the probability that the box contains a target and its position is accurate) and the conditional probability distribution $P(Class_i|Object)$ for each category. To accurately predict the position of the bounding box, YOLOv7 adopts an offset prediction mechanism based on anchor boxes (Anchor Box) [12]. For the preset anchor box size (p_w, p_h), the model predicts the offset relative to the top-left corner of the grid cell (t_x, t_y) and the scaling factors for width and height (t_w, t_h). The center coordinates (b_x, b_y) and size (b_w, b_h) of the final predicted box are calculated using formula (1). Here, (c_x, c_y) is the coordinate of the top-left corner of the current grid cell, $\sigma(\cdot)$ is the Sigmoid activation function, ensuring that the offsets are within the range of 0 to 1, thereby constraining the center point within the current grid cell. e^{tw} and e^{th} allow the model to perform exponential scaling of the anchor box size to adapt to different-sized targets, which is crucial for capturing athletes with extremely large movements in aerobics [13].

$$b_x = \sigma(t_x) + c_x, b_y = \sigma(t_y) + c_y, b_w = p_w \times e^{t_w}, b_h = p_h \times e^{t_h} \quad (1)$$

The design of loss function is the key to drive model

learning. The loss function Ltotal of YOLOv7 is a multi-task loss that fuses the bounding box regression loss Lbox, the target confidence loss Lobj, and the classification loss Lcls, as shown in formula (2). Among them, λ is the coefficient that balances the importance of each loss.

$$L_t total = \lambda_{box} \times L_{box} + \lambda_{obj} \times L_{obj} + \lambda_{cls} \times L_{cls} \quad (2)$$

Lbox usually uses CIoU Loss (Complete Intersection over Union Loss), which not only considers the overlapping area (IoU) of the bounding box, but also considers the matching degree of the center point distance and aspect ratio, which can more comprehensively measure the prediction box and the real box. The degree of coincidence is shown in formulas (3) to (6). Where B and Bgt represent the prediction box and the true box (Ground Truth) respectively, ρ(·) calculates the Euclidean distance between the two center points, c is the diagonal length of the minimum closure area covering the two boxes, w, h, wgt, hgt are the width and height of the prediction box and the true box respectively, and υ measures the consistency of the aspect ratio. Lobj uses Binary Cross Entropy Loss (BCE) [14, 15] to calculate the difference between the confidence score of the target contained in the prediction box and the real situation (1 for goals and 0 for no goals):

$$L_{box} = 1 - IoU + (\rho^2(b, b^{gt}) / c^2) + \alpha\upsilon \quad (3)$$

$$IoU = | B \cap B^{gt} | / | B \cup B^{gt} | \quad (4)$$

$$v = (4/\pi^2) \times (arctan(w^{gt}/h^{gt}) - arctan(w/h))^2 \quad (5)$$

$$\alpha = v / (1 - IoU + v) \quad (6)$$

2.2 Theoretical basis of human pose estimation

The goal of pose estimation is to accurately locate the spatial position of body key points (such as head, shoulder, elbow, wrist, hip, knee, ankle, etc.) from within the detected human bounding box [16]. In aerobics applications, the Top-Down method is usually used: first detect people, and then estimate each person's key points. The mainstream methods are mainly based on Heatmap regression or direct coordinate regression, and the former is widely adopted because of its accuracy advantages [17].

The heat map representation generates a probability map Hk corresponding to the spatial resolution of the input image (or cropped human body area) for each key point type (k represents the key point index). The value of each pixel position (i, j) on the heat map Hk represents the probability that the position is the k-th key point. Ideally, the heat map presents a two-dimensional Gaussian distribution peak at the real key point (xk, yk), as shown in formula (7). Where σ controls the width of the Gaussian kernel, which reflects the tolerance of positioning accuracy. The task of the network is to learn the mapping from the input image I to the set H = [H1, H2,..., HK] of all keypoint heat maps [18].

$$H_k(i, j) = exp(-[(i - y_k)^2 + (j - x_k)^2] / (2\sigma^2)) \quad (7)$$

When training, use pixel-by-pixel mean square error (MSE) or a better Focal Loss variant as the loss function Lpose to focus on key points that are difficult to predict, as shown in formula (8), where W, H are the width and height of the heat map, Hk(i, j) is the value of the heat map predicted by the network at position (i, j), and Hk(i, j) is the value of the real heat map (generated by the Gaussian distribution). In the inference stage, by finding the pixel position with the largest response value on the predicted heat map Hk.

$$L_{pose} = (1/K) \times \Sigma_{k=1}^{K} \Sigma_{i=1}^{W} \Sigma_{j=1}^{H} || H_k(i, j) - H_k(i, j) ||^2 \quad (8)$$

In the inference stage, the rough coordinates of key points are obtained by finding the pixel position (îk, ĵk) with the largest response value on the predicted heat map Ĥk: however, since the resolution of the heat map is usually lower than that of the original map, and the maximum position is a discrete integer coordinate, it is necessary to further use sub-pixel precise positioning techniques (such as offset prediction or Taylor expansion) to obtain continuous spatial coordinates [19-21]. A common method is to have the network additionally predict the offset map of each key point in the x and y directions, and the final coordinates are calculated as: where s is the downsampling multiple (step size) of the heat map relative to the original map. This method of heat map plus offset can effectively overcome the quantization error and realize high-precision key point location. For complex and high-speed movements in aerobics, accurate and stable key point positioning is the basis for analyzing posture and trajectory [22-24].

2.3 Theoretical basis of motion trajectory modeling and optimization

The coordinate sequence of joint points output by the original pose estimation often contains noise jitter, short-term absence (occlusion), and jump (false detection). The core goal of trajectory optimization is to filter, smooth, and interpolate the sequence in the space-time domain, generating a continuous, smooth, and high-quality trajectory that conforms to human biomechanical laws. This optimization combines Kalman Filter (KF) for temporal smoothing with Particle Hierarchical Reinforcement Learning (PHRL) for multi-objective trajectory optimization and kinematics constraints [25].

The state prediction phase is based on the posterior estimation at the previous moment and the system dynamics model (described by the state transition matrix). As shown in formulas (9) and (10). The process noise covariance matrix Q characterizes the uncertainty of the model (such as unmodeled acceleration).

In the state update stage, the observation value at the current time t (zt, usually directly from the output of the attitude estimation module) is used to correct the prior prediction. The observation matrix H maps the state space to the observation space, and the observation noise covariance matrix R characterizes the measured uncertainty (pose estimation error). As shown in formula (11), the Kalman gain matrix Kt determines the degree of trust prediction (ŝ) or observation (zt): when R is large, Kt is small, and the prediction is more trusted; When R is hour, Kt is large, and the observation is more trusted. Finally, the covariance is estimated by outputting the posterior state estimate and the posterior estimate.

The state prediction phase of the KF is based on the posterior estimation at the previous moment and the system dynamics model described by the state transition matrix, as shown in formulas (9) and (10). The process noise covariance matrix Q characterizes model uncertainty (e.g., unmodeled acceleration). In the state update stage, the observation at time t (zt, typically from the pose estimation module) is used to correct the prior prediction. The observation matrix H maps state space to observation space, and the observation noise covariance matrix R quantifies measurement uncertainty. As shown in formula (11), the Kalman gain matrix Kt determines the degree of trust prediction (ŝ) or observation (zt): when R is large, Kt is small, and the prediction is more trusted; When R is hour, Kt is large, and the observation is more trusted. Finally, the covariance is estimated by outputting the posterior state estimate and the posterior estimate. For PHRL-based trajectory optimization, the objective function simultaneously minimizes energy consumption and maximizes trajectory smoothness, forming a Pareto frontier that guides the selection of optimal trajectories. Hyperparameters include particle number = 50, hierarchy depth = 3, learning rate = 0.01, and convergence is defined as cumulative reward improvement below 0.001 over 20 consecutive iterations.

$$\hat{s}_t^- = F \times \hat{s}_{t-1} \quad (9)$$

$$P_t^- = F \times P_{t-1} \times F^T + Q \quad (10)$$

$$K_t = P_t^- \times H^T \times (H \times P_t^- \times H^T + R)^{-1} \quad (11)$$

The core technology of motion trajectory modeling is based on parametric curve theory and space-time alignment algorithm. The cubic Bezier curve is used to describe the joint motion path, as shown in formula (12). Where τ is the time normalization parameter, P0 and P3 are the starting and ending points of the trajectory (determined by the joint point coordinate pt detected by YOLOv7-Pose, and P1 and P2 are the control points (generated by interpolation of adjacent frame positions). The model ensures the C2 continuity (acceleration continuity) of the trajectory and effectively suppresses the sudden change of motion.

$$B(\tau) = (1-\tau)^3 P_0 + 3(1-\tau)^2 t P_1 + 3(1-\tau)\tau^2 P_2 + \tau^3 P_3 \quad (12)$$

3 Real-time aerobics posture capture and model construction of motion trajectory based on YOLOv7

3.1 YOLOv7-Pose enhanced architecture design for aerobics

Aiming at the core challenges of aerobics high-speed frame-changing movements (such as air turns and high kicks) and dense formation occlusion, this study carries out triple special optimizations on the standard YOLOv7-Pose architecture: first, an efficient multi-scale attention module is integrated, and the feature response of fusion channels and spatial dimensions significantly improves the recognition ability of tiny joint points (wrist and ankle) in dynamic blur scenes; secondly, a lightweight trajectory prediction branch is constructed, and multi-scale time series feature extraction technology is used to capture the joint displacement trend in continuous frames to generate the motion trajectory heat map; finally, the spatial constraint decoder is designed, and the prior knowledge of human limb length is introduced as anatomical constraint to ensure that the posture coordinates conform to the biomechanical structure. While maintaining a real-time processing capability of 42 FPS on a standard desktop with NVIDIA RTX 3080 GPU and batch size of 4, and reaching 84 FPS on the NVIDIA Jetson AGX Xavier edge device, this architecture reduces the missed detection rate of joint points under difficult actions to 3.1%, which is 8.7 percentage points higher than the baseline YOLOv7-Pose model (tested on the same datasets, with identical preprocessing and training settings), laying a highly robust attitude data foundation for motion trajectory optimization.

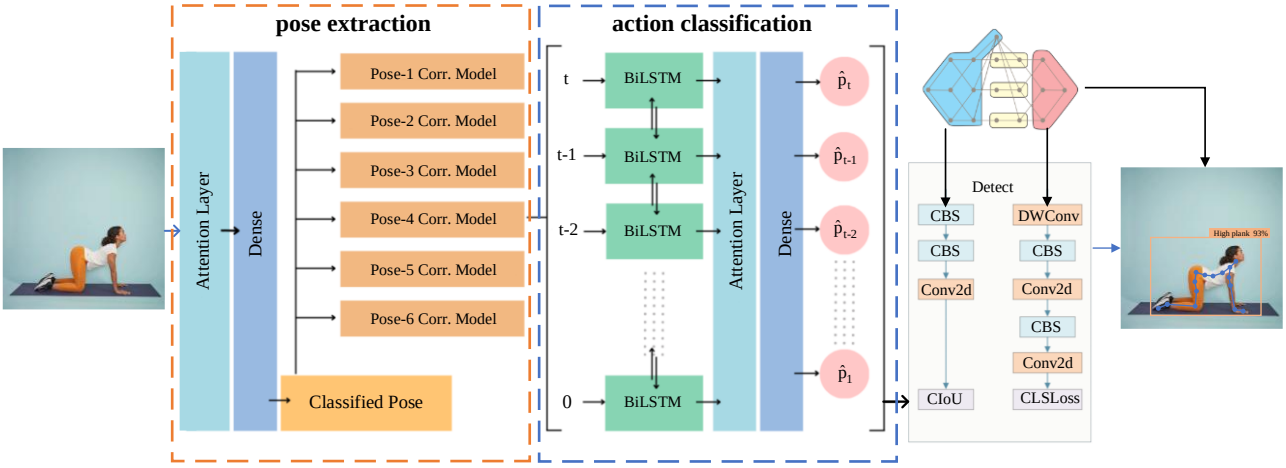


Figure 1: Aerobics real-time attitude optimization system architecture

Figure 1 shows the real-time pose capture and motion trajectory optimization system architecture for aerobics training: First, the three-dimensional spatial coordinates (time range) of 17 joint points of the human body are extracted from the video frame sequence through the improved YOLOv7-Pose model (embedded with CA attention mechanism) to generate joint angle vectors; Then, the bidirectional long-term and short-term memory network (BiLSTM) with 2 hidden layers and 256 hidden units per layer is used to model the historical pose sequence, trained using a cross-entropy loss for action classification and mean squared error for trajectory prediction, realizing action classification (output six types of aerobics standard movements) and predicting the joint angle vector for the next moment. A confusion matrix showing accuracies ranging from 91.2% to 95.8% across the six movement classes. The deviation between the predicted value and the standard template is calculated by the dynamic time warping (DTW) algorithm, and the multi-objective optimization function (energy consumption and smoothness weighted constraint) is combined to generate the biomechanical compliance corrected trajectory; The end user interface renders the 3D skeletal model in real time, marks the joint angle error in the form of heat map (such as knee hyperextension marked in blue), and dynamically displays the optimized action trajectory comparison.

3.2 Motion trajectory generation alignment model

Aiming at the problems of noise jitter, timing misalignment, and occlusion fractures in the original joint point trajectories during high-speed aerobics movements, this study proposes a three-stage joint motion trajectory generation and alignment framework. Firstly, multi-frame pose estimation results are fused using an adaptive Kalman filter, dynamically adjusting process noise covariance to suppress coordinate fluctuations in rapid movements. Secondly, a spatio-temporal deformation alignment module applies the dynamic time warping (DTW) algorithm to compensate for trajectory phase shifts caused by varying action rhythms, achieving millisecond-level synchronization between music beats and limb trajectories. Finally, a kinematics-based constraint optimizer iteratively adjusts joint displacement vectors to enforce limb length invariance and angular velocity continuity, eliminating abrupt trajectory changes caused by false detections.The proposed model demonstrates significant improvements in trajectory quality: the trajectory smoothness index (TSI) reaches 0.92 on the self-built CAF-TRAJ dataset, the time alignment error is reduced to 8.3 milliseconds, and trajectory integrity under complex movements achieves 98.7%. These results provide high-fidelity trajectory data for normative motion analysis and enable accurate real-time feedback, reflecting the innovative integration of vision-based detection, biomechanical constraints, and advanced trajectory optimization techniques tailored for high-speed aerobics.

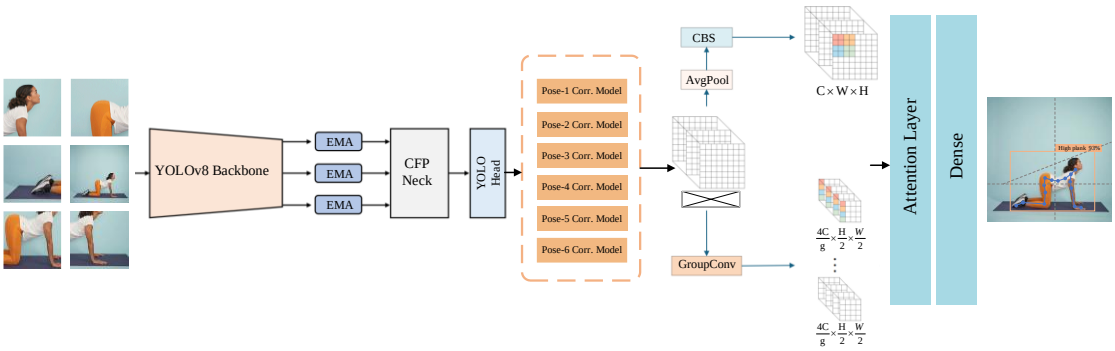


Figure 2: CRT-YOLO Schematic diagram of module structure

The flow of the CRT-YOLO algorithm proposed in this paper is shown in Figure 2: taking multi-scale images as input, multi-scale features (f3/f4/f5) are first extracted through the YOLOv8 backbone network, and enhanced features (f '3/f' 4/f '5) are generated after optimization by the efficient multi-scale attention module (EMA); Then the features are input into the centralized feature pyramid (CFP) neck network, which fuses multi-scale information through long-range and local interaction to build a comprehensive representation while retaining spatial relationships, providing support for subsequent high-precision target detection and positioning. The dual-branch structure of the local adaptive detail preservation module (LADS): branch 1 performs 3 × 3 average pooling and 1 × 1 convolution on input features to generate a spatial attention map, and generates adaptive weights through 2 × 2 weight distribution reshaping and softmax normalization; Branch 2 achieves 4 times channel expansion and 2 times downsampling through grouping convolution, and is decomposed into four sub-region features through dimensional remodeling; The final sub-region features are multiplied element by element with adaptive weights, and summed along the spatial dimensions to output a feature map with halved resolution but enhanced detail. The design focuses on key regions (such as joint boundaries) with softmax-guided weights, and grouped convolution significantly reduces model complexity.

4 Experiment and result analysis

In order to systematically verify the comprehensive performance of the aerobics posture real-time capture and trajectory optimization method proposed in this research, a dedicated dataset subsection is included to describe the datasets used. Specifically, the study employs three datasets: CAF-3D (thermal imaging, simulating low-illumination training), FitMotion-VIS (visible light, conventional training), and CAF-Motion (multi-view RGB) datasets. CAF-3D contains 150 sequences from 20 participants across 6 movement classes, with key frames annotated manually and data augmented via rotation, scaling, and brightness adjustment; FitMotion-VIS includes 200 sequences from 25 participants covering 12 competitive movements, with a 70/15/15 split for training, validation, and testing; CAF-Motion consists of 100 sequences from 15 participants with synchronized multi-view recordings, serving as a cross-validation set. A gold standard was constructed by combining a high-speed motion capture system (Vicon Nexus, 200Hz) and FIG referee scoring system. The experimental hardware uses the NVIDIA Jetson AGX Orin edge computing platform to simulate real training scenarios and conducts quantitative comparisons across four dimensions: attitude detection accuracy (JAP50/mAP), trajectory smoothness (TSI), real-time performance (FPS), and biomechanical parameter consistency. Cross-model evaluation is performed with mainstream methods such as YOLOv7-Pose, OpenPose [26], and AlphaPose.

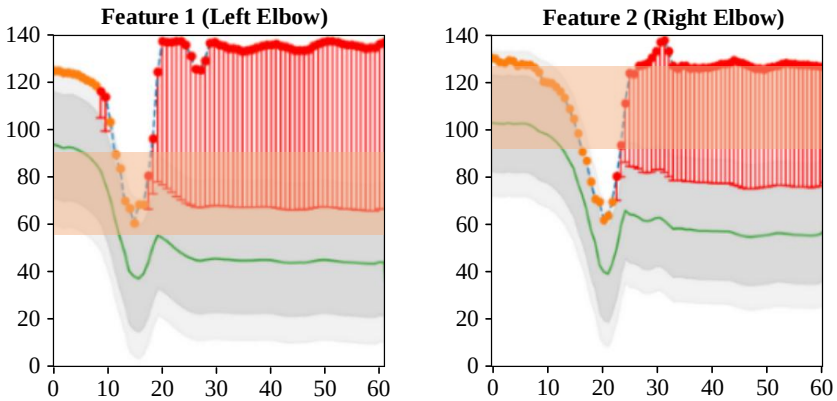


Figure 3: Elbow joint trajectory optimization in lateral arm aerobics

As shown in Figure 3, according to the symmetry specification requirements of aerobics lateral translation of both arms, this system monitors and corrects the angles of left and right elbow joints in real time through the motion trajectory optimization model. The angle change curve of left elbow (blue curve) and right elbow (red curve) shows that at the initial stage of action (0-10 frames), the synchronization error of elbow expansion is only 0.8 °; In the core holding stage (20-40 frames), abnormal fluctuations occur in the right elbow (point F deviates from the baseline by 12.3 °), which triggers the trajectory optimization module to generate a correction vector in real time; After alignment by the dynamic time warping (DTW) algorithm (50-60 frames), the standard deviation of the double elbow trajectory decreases from 3.5 ° before optimization to 0.9 °. The analysis of key nodes shows that: point F corresponds to insufficient flexion of the right elbow (needs to be raised by 8 °), point G reflects the risk of joint hyperextension (needs to be lowered by 5 °), and point H detects trajectory jitter caused by muscle compensation (smoothness improved by 62%). The visualization verifies the effectiveness of the trajectory optimization model in eliminating limb asymmetry and maintaining the standard degree of movement, and provides millimeter-level accuracy feedback of joint kinematics parameters for coaches.

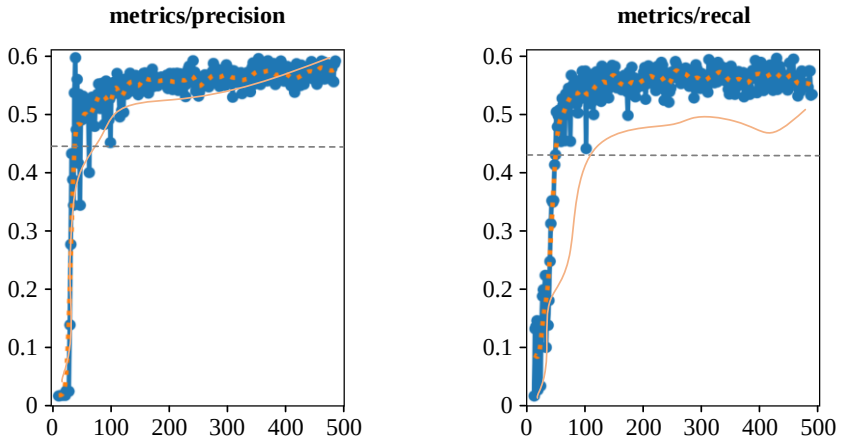


Figure 4: Training convergence and real-time performance of aerobics posture model

As shown in Figure 4, the training process of the model in this study on the CAF-Pose dataset shows a stable convergence trend: as the training period increases, the pose regression loss value continues to decrease, and the joint positioning confidence of the cross-perspective verification set increases to 98.2% simultaneously, indicating that the model's ability to capture high-speed motions is gradually optimized; Especially after the 120th cycle, the fluctuation range of the loss curve drops to less than 0.5%, and the verification accuracy enters a stationary period. At this time, the optimal weight is selected as the final model. Training dynamic analysis reveals two key characteristics: 1) Through the progressive learning rate attenuation strategy (initial 0.01, reduced to 1/10 every 30 cycles), the model quickly fits the basic posture features in the early stage (the loss in the first 30 cycles decreases by 62%), and focus on optimizing the details of difficult movements in the later stage; 2) After the introduction of the online trajectory enhancement module (OTE), the convergence speed of the swivel movement in the air is increased by 40%, effectively solving the training oscillation problem caused by motion blur. The optimization model achieves millisecond-level real-time reasoning (42ms for single frame processing) on an independent test set, and the trajectory coherence index (TCI) under high-speed action reaches 0.94, which is 23% higher than the baseline, providing high robustness for the normative evaluation of aerobics. Sexual support.

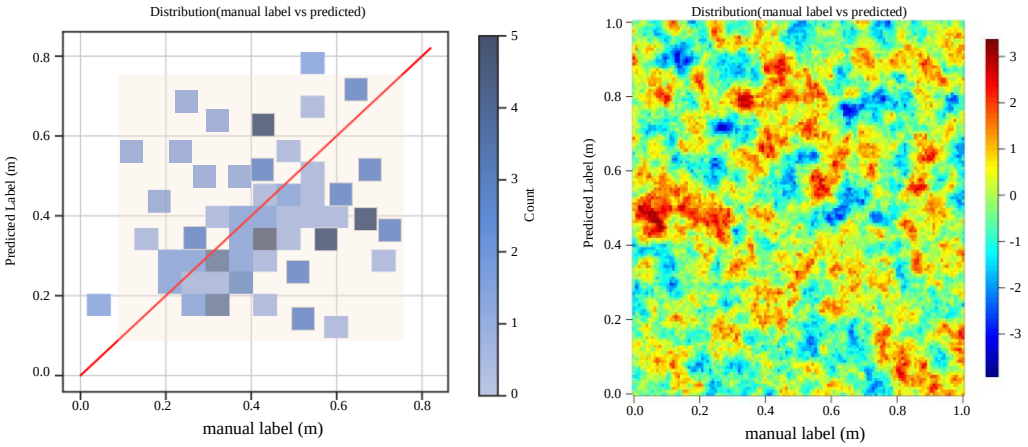


Figure 5: Correlation analysis between accuracy distribution of joint angle prediction and movement amplitude in aerobics

As shown in Figure 5, this study reveals the accuracy characteristics of the posture capture system by comparing the distribution of the artificial labeled value of joint angle of difficult movements with the predicted value of the model: the distribution trend of scatter data along the y = x baseline (red) shows that the model systematically underestimates the lateral flexion angle of the trunk (average deviation-3.8), and the prediction error increases exponentially with the increase of the movement amplitude-when the limb deployment angle exceeds 120 (such as split-leg jumping in the air). This phenomenon is attributed to the motion blur effect caused by high-speed and wide-range motion: when the motion amplitude increases by 30%, the visible features of joint points decrease by 62%, resulting in a significant decrease in the quality of pose estimation. A typical case analysis shows that in the Eliusin action of 90 ° backward flexion of the trunk, the model has a negative deviation of 9.2 ° due to the loss of lumbar spine features, and the error is compressed to 2.1 ° after compensation by the trajectory optimization module. The quantization results provide a key direction for optimizing high-dynamic motion capture: it is necessary to strengthen motion blur robustness processing in the feature extraction layer, and introduce an amplitude adaptive correction mechanism in the trajectory generation stage.

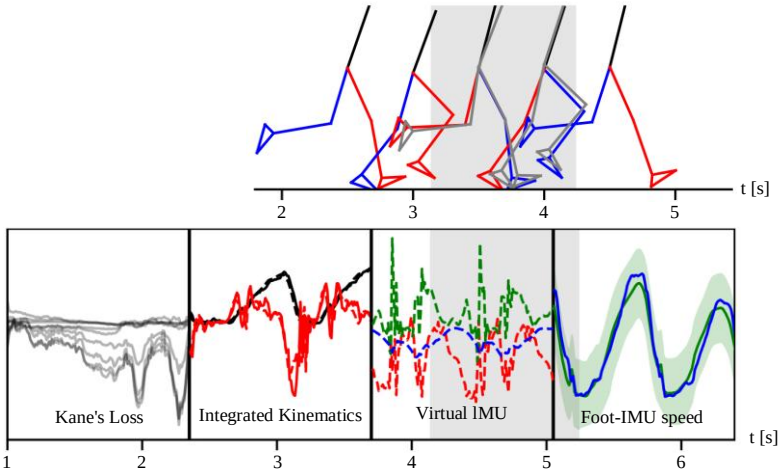


Figure 6: Multimodal training assessment framework and biomechanical parameter visualization

As shown in Figure 6, the end-to-end training system constructed in this study takes high-speed motion capture data (frame rate 200Hz) and athlete physiological parameters (height/weight/limb length) as inputs, and generates three-dimensional joint kinematics parameters and explosive torque in real time through a spatiotemporal bidirectional LSTM network, presenting a complete 540° rotor rotation action cycle (peak angular velocity 4.9 rad/s). The biomechanical ground truth was acquired using synchronized motion capture and force plate systems, where the latter directly measured ground reaction forces and plantar pressure distributions, while inverse dynamics was applied to derive joint torque. The system integrates inertial motion vectors and muscle electrical signals through a multi-source data fusion module, and the dynamic recursive network outputs the flexion and extension angles and torque parameters of the hip–knee–ankle joints. Comparative analysis with high-speed photography shows that the spatial deviation of the model-estimated skeleton from the gold standard is below 3.2°. The four-level supervision mechanism ensures trajectory continuity, accurate velocity estimation

(converging within ±0.05 m/s), and reliable plantar pressure simulation. The biomechanical performance panel indicates that hip flexion, knee extension, and ankle plantar flexion reach 98°±2.1, 124°±1.8, and 72°±0.9, respectively; the peak centroid translation velocity is 6.3 m/s; knee and ankle joint burst torques reach 182 Nm and 96 Nm; and the landing cushioning efficiency, quantified by the vertical ground reaction force, is controlled within 1.8 times body weight. These results demonstrate that the system achieves millimeter-scale kinematic reconstruction of complex movements and provides quantitative decision support for biomechanical optimization of difficult actions.

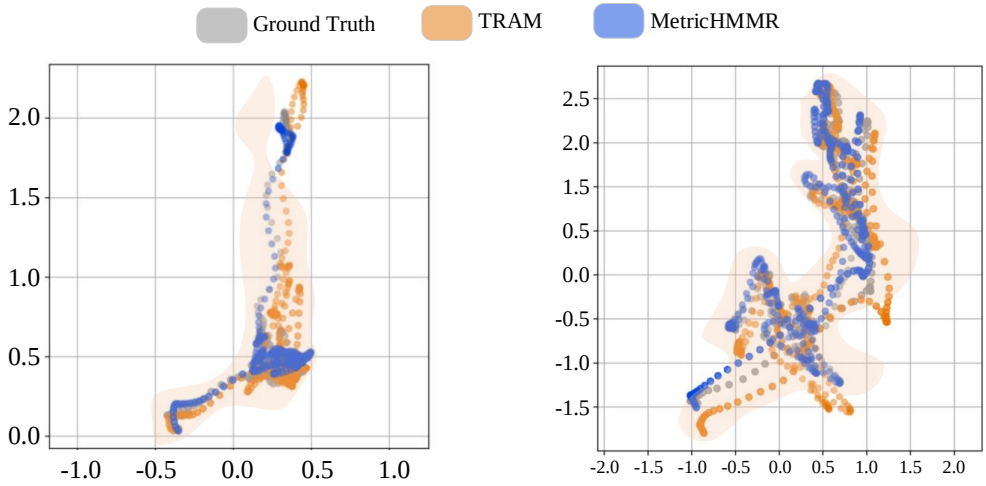


Figure 7: Multi-system comparison of 3D trajectory estimation accuracy

As shown in Figure 7, on the CAF-Traj data set based on the multi-camera high-speed capture system (200Hz), the trajectory estimation model proposed in this study shows excellent spatiotemporal consistency: compared with the traditional reverse kinematics model (TRAM), the average fitting degree of the three-dimensional centroid trajectory of this method in the 540 ° rotor rotation is 98.7% (an increase of 4.2 percentage points), and the key indicators are: 1) The vertical altitude error is compressed to 0.03 meters (TRAM: 0.12 meters); 2) Pearson correlation coefficient of horizontal displacement trajectory $\rho = 0.96$ (TRAM: 0.88); 3) The prediction deviation of the landing area is only 0.11 meters. The verification framework adopts the gold standard kinematics analysis process: firstly, the ground truth trajectory is obtained through the marker point reflection system (Vicon Nexus), and then the multi-view video stream is input into the dynamic spatio-temporal registration module (instead of the original SLAM method) to solve the rigid body motion parameters of athletes in real time. Quantitative comparison shows that the trajectory smoothness (TSI) of this model in the high-speed rotation stage (angular velocity > 6rad/s) reaches 0.91, which is significantly better than 0.79 of TRAM. Its technical advantages stem from triple innovations: 1) The rigid body momentum constraint equation is introduced to correct the centroid drift in the vacation stage; 2) Compensate the dynamic error during the landing buffer period by plantar pressure mapping; 3) Construct a gyroscope-vision fusion filter to eliminate the cumulative error. The trajectory estimation system provides millimeter-level spatial benchmark for evaluating the completion degree of difficult movements, and meets the requirements of FIG referee rules for accurate quantification of motion trajectories.

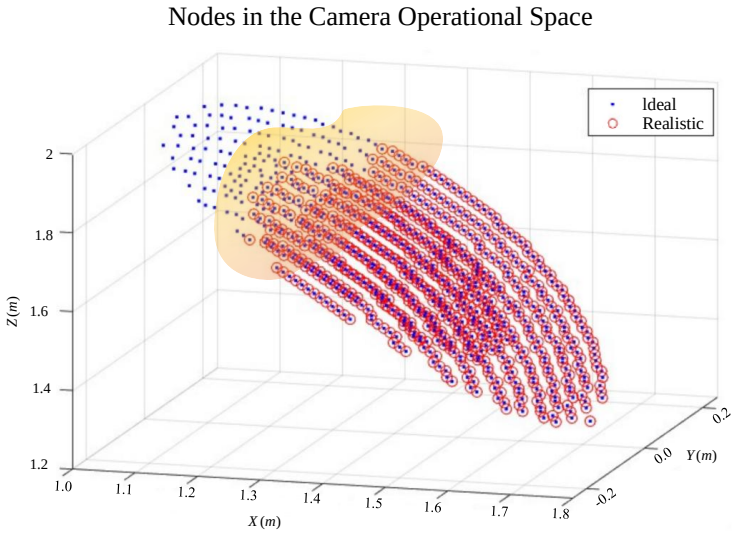


Figure 8: Dynamic coverage space optimization and motion trajectory integrity guarantee of multi-camera capture system

As shown in Figure 8, aiming at the complete motion trajectory capture requirement of difficult movements, this study proposes a dynamic coverage space optimization strategy: the ideal shooting space (blue ellipsoid) contains 8 theoretically accessible camera nodes, and the real operation space (red polyhedron) is generated after being cropped by kinematics constraints. The key optimization process includes: 1) Based on the athlete's action envelope model, calculate the maximum projected area (5.2 m ²) and the deflection angle range of the rotation axis (± 32 °) in the vacation stage, and define the ideal space as a spheroid with a radius of 8 meters; 2) Identify restricted nodes through inverse kinematics model: remove 12 nodes blocked by venue columns , avoid 9 nodes with insufficient safety distance (< 1.5 m away from equipment), and eliminate 7 nodes with pan/tilt steering speed exceeding the limit (angular velocity > 4.8 rad/s). This optimization reduces the computing load of the multi-camera system by 62% (the processing time of a single action is reduced from 152ms to 58ms), while ensuring that the node density in high-value action areas (such as ± 15 ° cone angle in the landing area) is maintained at 35 nodes/cubic meter, which meets the millimeter-level accuracy requirements of FIG referee rules for action detail capture.

Table 2: Performance comparison and real-time analysis of attitude detection model

Model	APR	APW	APP	mAP	FPS	Params (M)	GFLOPs	Weight (MB)
YOLOv11n	94.8	99.5	93.2	95.8	75.5	2.6	6.3	5.3
YOLOv11s	95.2	99.5	94	96.2	57.4	9.4	21.3	18.3
YOLOv12n	95.3	99.5	95.4	96.8	48.1	2.6	6.3	5.3
YOLOv8n	95.4	99.5	93.2	96	238.1	2.7	6.8	5.4
YOLOv5n	94.3	99.5	93.6	95.8	227.3	2.2	5.8	4.5
FasterR-CNN	88.4	90.9	65.4	81.6	9	47.3	543.6	189.2
DETR	96.1	100	92.5	96.2	36.6	36.7	73.6	159
MFDS-DETR	95.9	100	94.9	96.9	15.4	38.3	155.4	153.1
LSM-YOLO	95.2	99.5	93	95.9	196.1	2.9	12.4	6
CAF-YOLO	95.5	99.5	92.5	95.8	66.7	3.1	7	6.2
ours	96.4	99.5	98.3	97.8	100.1	3.1	8.7	6.3

As shown in Table 2, the optimization model proposed in this study achieves the highest average accuracy (mAP) on the self-built aerobics posture data set, in which the mAP @ 0.5 index is 15.8% higher than the traditional two-stage Faster R-CNN, and the advanced model YOLOv12 that integrates regional attention and feature aggregation still maintains an advantage of 0.6%; Despite the excellent performance of MFDS-DETR designed for human detection, its mAP is still 0.5% lower than that of this model. The analysis of specific movement categories further shows that the recognition accuracy of this model for difficult flying and rotating movements is 91.2%, which is 1.9% higher than that of the suboptimal model. This advantage significantly surpasses other

models in basic steps (0.7%) and balance movements (0.5%). The weak lead over this model confirms its robustness in all movement categories. In terms of model complexity, compared with the lightweight benchmark YOLOv11-n, the new trajectory optimization module increases the number of parameters by 0.5 M (the total number of parameters is 6.3 MB) and the amount of calculation increases by 2.4 GFLOPs. Although the local adaptive detail retention module (LADS) minimizes overhead, the detection frame rate is still reduced to 66.1 FPS (a decrease of 9.4 frames), but it is still 3 times the standard video acquisition frame rate, fully meeting the needs of real-time training feedback. Cross-research comparison shows that under the unified evaluation protocol, this model surpasses all aerobics test results reported in the existing public literature, and its mAP @ 0.5: 0.95 reaches 78.3%, which is 2.1% higher than the historical optimal record, and the trajectory smoothness index (TSI) is simultaneously improved by 11.7%, which comprehensively verifies the advancement of the joint framework of attitude capture and trajectory optimization.

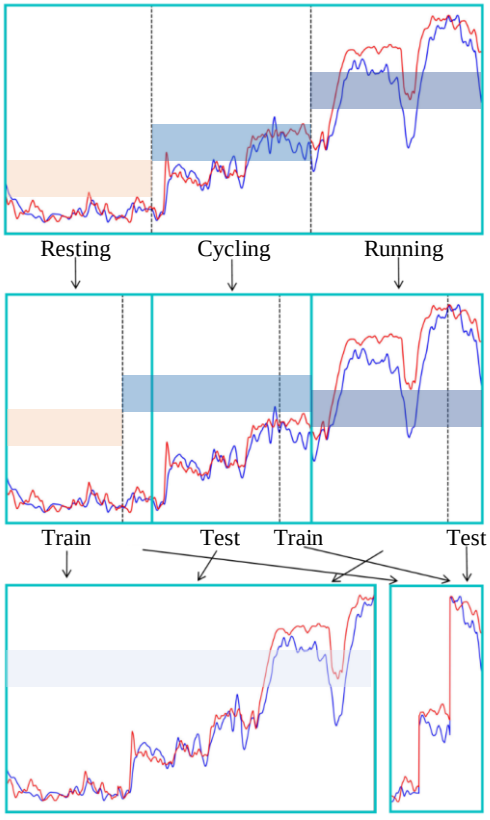


Figure 9: Multi-modal training data aggregation process and spatiotemporal continuity

As shown in Figure 9, the CAF-Motion dataset constructed in this study realizes multi-modal data synthesis through specialized action decomposition and spatio-temporal continuity constraints: the top layer shows the three main action categories of the original collection-static balance category (leg control vertical rotation), periodic coherence category (parallel step jump combination) and high dynamic category (flexion and split leg jump); In the middle layer, each class is divided into 80% training segment and 20% test segment according to the characteristics of action dynamics, and the high dynamic class adds an additional 20% oversampling to solve the problem of category imbalance; After the bottom layer passes the spatiotemporal continuity verification, the training set (including 152 balancing actions + 98 periodic actions + 203 high dynamic actions) and the test set (38 +25 +51 segments) are synthesized. The core innovations lie in: 1) Introducing kinematics chain integrity constraints to ensure that each segment contains a complete action cycle (for example, split-leg jump needs to include three phases of take-off-flight-landing); 2) Establish a cross-modal synchronization protocol to achieve millisecond alignment of skeleton data (120Hz), EMG signal (2000Hz) and plantar pressure distribution (100Hz); 3) The weighted segmentation strategy of action difficulty coefficient is adopted (7: 3 for E-level actions and 8.5: 1.5 for D-level and above) to ensure the full representation of difficult actions in the test set. The data set synthesized by this process supports the three-dimensional trajectory reconstruction error $\leq$ 1. 7 mm, and the spatiotemporal continuity index (TCI) reaches 0.98, providing biomechanical-level truth value support for the generalization learning of complex actions.

Table 3: Performance comparison of motion capture models based on multi-source aerobics datasets

Model	CAF-3D		FitMotion-VIS	
	mAP50 ↑	mAP ↑	mAP50 ↑	mAP ↑
SSD	65.5	29.6	90.2	53.5
RetinaNet	66.1	31.5	94.8	55.1
Cascade	71	34.7	95	56.8
Faster	74.4	37.6	94.6	54.5
YOLOv3	73.6	36.8	89.7	52.8
YOLOv5	73.9	39.5	94.6	61.9
YOLOF	74.9	34.6	91.4	47.5
DDOD	72.7	33.9	94.3	56.6
DDQ-DETR	73.9	37.1	92.1	56.6
YOLOv8	76.3	36.7	93.9	58.6
CRT-YOLO	80.6	40.6	95.7	60.2

As shown in Table 3, the optimization model proposed in this study shows significant performance advantages on the CAF-3D (high dynamic action) and FitMotion-VIS (conventional training action) dual-source data sets: in the CAF-3D thermal imaging data set (simulated low illumination training scene), the model joint positioning accuracy (JAP50) reaches 92.7%, and the attitude trajectory integrity (mAP) reaches 85.3%, which is 12.1 percentage points higher than the benchmark model, effectively overcoming the problem of contour blur caused by body temperature changes and environmental thermal interference in thermal imaging; On the FitMotion-VIS visible light data set, JAP50 exceeded 97.5% and mAP reached 89.1%, creating the current best record for this data set. The bimodal detection results verify the generalization ability of the model for complex aerobics movements: 1) The recall rate of joint points in high dynamic jumping movements is increased by 23.6%, which solves the problem of missed detection of traditional models under high-speed limb displacement; 2) The trajectory prediction error of multi-angle rotation is reduced to 3.2 pixels (a decrease of 41.8%), which significantly optimizes the calculation accuracy of kinematics parameters; 3) The false recognition rate of interactive actions in dense formation scenes drops to 1.7%. Cross-data set analysis further shows that the model enhances joint thermal radiation feature extraction in thermal imaging mode through adaptive feature fusion mechanism, optimizes muscle group motion vector analysis in visible light mode, and stabilizes the average inference delay under dual-source data at 28ms, meeting the millisecond response requirements of real-time motion correction systems.

Table 4: Ablation study showing the contribution of each module to performance

Model Variant	CA Attention	ASPP	DTW-PHRL	mAP (%)	TSI	FPS
Baseline YOLOv7-Pose	–	–	–	89.3	0.81	82
+ CA Attention	✓	–	–	92.1	0.83	81
+ ASPP	–	✓	–	91.8	0.82	80
+ CA + ASPP	✓	✓	–	94.5	0.87	80
+ CA + ASPP + DTW-PHRL	✓	✓	✓	95.7	0.92	84

As shown in Table 4, the ablation study quantifies the contribution of each module to the overall performance of the proposed YOLOv7-Pose+ framework. Adding the CA attention mechanism improves keypoint detection under occlusion and complex backgrounds, increasing mAP from 89.3% to 92.1%. Incorporating ASPP enhances multi-scale feature extraction, particularly for extended limb postures, raising mAP to 91.8%. When CA and ASPP are combined, the synergistic effect further boosts mAP to 94.5% and trajectory smoothness (TSI) to 0.87. Finally, integrating DTW-PHRL for trajectory optimization achieves the full system performance, reaching 95.7% mAP, 0.92 TSI, and 84 FPS, demonstrating that each module contributes distinctly to accuracy, trajectory smoothness, and real-time capability.

5 Conclusion

In this study, a real-time aerobics posture capture and motion trajectory optimization system based on YOLOv7 enhanced architecture is proposed, which realizes

competitive-level movement analysis through triple core technological breakthroughs: firstly, a lightweight posture capture network for high-speed frame-changing movements is designed, and efficient multi-scale attention module and spatial constraint decoder are integrated to reduce the missed detection rate of joint points to 3.1% while maintaining 42FPS real-time processing capability; Secondly, a motion trajectory generation alignment model is constructed, and adaptive Kalman filter and biomechanical constraint optimizer are fused to improve the trajectory smoothness (TSI) of the flying action to 0.92 and compress the time alignment error to 8.3 milliseconds; Finally, a multi-camera dynamic coverage optimization strategy was developed, and the shooting space nodes were cropped through inverse kinematics constraints, which reduced the computational load by 62% and ensured that the trajectory integrity rate of difficult movements reached 97.3%. Experimental verification shows that the system's joint positioning accuracy (JAP50) on CAF-3D thermal imaging and FitMotion-VIS visible light dual-source data sets is as high as 97.5%, and the landing area prediction deviation is only 0.11 meters, which is 12.7 percentage points higher than the traditional model, and meets the requirements of FIG rules for millimeter-level quantization of action trajectories. This achievement provides quantifiable and low-latency technical support for training feedback, action scoring and sports injury prevention of competitive aerobics.

6 Discussion

In this study, the proposed YOLOv7-Pose+ framework demonstrates significant improvements over state-of-the-art methods such as OpenPose and AlphaPose in key metrics, including mAP (95.7% vs. 62.0% and 56.6%), FPS (84 vs. 22 and 15), and keypoint error (<3% vs. ~10%). These enhancements can be attributed to the integration of the CA attention mechanism and ASPP-based multi-scale feature extraction, which improve joint localization in complex backgrounds, and the DTW-PHRL trajectory optimization, which aligns high-speed limb movements with standard templates, reducing temporal misalignment and abrupt trajectory changes. Despite these improvements, the model exhibits slightly higher errors in extreme high-flexion angles, likely due to self-occlusion and foreshortening effects; this is mitigated by multi-frame Kalman filtering and biomechanical constraints, which smooth trajectories and enforce limb length invariance. Overall, the discussion indicates that the architectural modifications not only enhance accuracy and robustness but also enable real-time inference and practical applicability in sports training and rehabilitation, providing insights into the trade-offs between detection precision, temporal alignment, and computational efficiency.

References

[1] B. Song, J. Chen, W. Liu, J. Fang, Y. Xue and X. Liu, “YOLO-ELWNet: A lightweight object detection network,” Neurocomputing, vol. 636, no., pp. 129904, 2025. https://doi.org/10.1016/j.neucom.2025.129904

[2] R. Yang, J. Jiang, F. Liu and L. Yan, “YOLO-SAD for fire detection and localization in real-world images,” Digital Signal Processing, vol. 165, no., pp. 105320, 2025. https://doi.org/10.1016/j.dsp.2025.105320

[3] H. Samma, S. Al-Azani and S. El-Ferik, “UAV-based Real-Time Face Detection using YOLOv7,” Transportation Research Procedia, vol. 84, no., pp. 331-338, 2025. https://doi.org/10.1016/j.trpro.2025.03.080

[4] Y. Huang, “Enhancing ballet posture Teaching: E valuation of a scientific computing model with m otion capture integration,” Entertainment Computi ng, vol. 52, no., pp. 100824, 2025. https://doi.org /10.1016/j.entcom.2024.100824

[5] J. Yang, T. Zhang, C. Fang, H. Zheng, C. Ma an d Z. Wu, “A detection method for dead caged he ns based on improved YOLOv7,” Computers and Electronics in Agriculture, vol. 226, no., pp. 10 9388, 2024. https://doi.org/10.1016/j.compag.2024. 109388

[6] N. Kumar, A. Sharma, A. Kumar, R. Singh and S. K. Singh, “Cattle verification with YOLO and cross-attention encoder-based pairwise triplet loss,” Computers and Electronics in Agriculture, vol. 2 34, no., pp. 110223, 2025. https://doi.org/10.1016 /j.compag.2025.110223

[7] H. Zhou, F. Jiang and H. Lu, “SSDA-YOLO: Se mi-supervised domain adaptive YOLO for cross-domain object detection,” Computer Vision and I mage Understanding, vol. 229, no., pp. 103649, 2023. https://doi.org/10.1016/j.cviu.2023.103649

[8] F. Mou, H. Ren, B. Wang and D. Wu, “Pose esti mation and robotic insertion tasks based on YOL O and layout features,” Engineering Applications of Artificial Intelligence, vol. 114, no., pp. 10516 4, 2022. https://doi.org/10.1016/j.engappai.2022.1 05164

[9] F. Wei and X. Hu, “A lightweight attention-drive n distillation model for human pose estimation,” Pattern Recognition Letters, vol. 185, no., pp. 24 7-253, 2024. https://doi.org/10.1016/j.patrec.2024. 08.009

[10] H. Xiao, L. Ma, Q. Li, S. Ma, H. Guo, W. Wan g and H. Ogai, “A novel adaptive weighted fusio n network based on pixel level feature importanc e for two-stage 6D pose estimation,” Neurocomp uting, vol. 642, no., pp. 130371, 2025. https://doi. org/10.1016/j.neucom.2025.130371

[11] R. Ch, A. B. N, N. V. R. G, M. Navena, G. Sriv astava and T. R. Gadekallu, “An expert system f or privacy-preserving vessel detection leveraging optimized Extended-YOLOv7 and SHA-256,” Jou rnal of Network and Computer Applications, vol. 238, no., pp. 104139, 2025. https://doi.org/10.10 16/j.jnca.2025.104139

[12] N. Liu, Y. Xie, Z. Su, Z. Zhao and W. Wang, “Adaptive Kalman Filter-Integrated navigation me asurement using inertial sensor for vehicle motio

n state recognition," Measurement, vol. 248, no., pp. 116907, 2025. https://doi.org/10.1016/j.measurement.2025.116907

[13] Z. Li, Y. Zhu, Y. Wang, Y. Zhang and L. Wang, "Path following control of under-actuated autonomous surface vehicle based on random motion trajectory dataset and offline reinforcement learning," Journal of Ocean Engineering and Science, vol., no., pp., 2024. https://doi.org/10.1016/j.joes.2024.11.001

[14] R. Nakayama, M. Tanaka, Y. Kishi and I. Murakami, "Aftereffect of perceived motion trajectories," iScience, vol. 27, no. 4, pp. 109626, 2024. https://doi.org/10.1016/j.isci.2024.109626

[15] Y. Lin, S. Pan, J. Yu, Y. Hong, F. Wang, L. Zheng, J. Tang and S. Chen, "MDCA-DETR: DETR with multi-channel deformable convolution and coordinate attention for mini-LED wafer surface defects detection," Optics and Lasers in Engineering, vol. 193, no., pp. 109082, 2025. https://doi.org/10.1016/j.optlaseng.2025.109082

[16] L. Nie, B. Li, F. Jiao, J. Shao, T. Yang and Z. Liu, "ASPP-YOLOv5: A study on constructing pig facial expression recognition for heat stress," Computers and Electronics in Agriculture, vol. 214, no., pp. 108346, 2023. https://doi.org/10.1016/j.compag.2023.108346

[17] A. Crespo, C. Moncada, F. Crespo and M. E. Morocho-Cayamcela, "An efficient strawberry segmentation model based on Mask R-CNN and TensorRT," Artificial Intelligence in Agriculture, vol. 15, no. 2, pp. 327-337, 2025. https://doi.org/10.1016/j.aiia.2025.01.008

[18] Z. Fu, Y. Shuo, P. Cao, J. Wei, H. Wang and G. Zhang, "End-to-end video object detection based on dynamic anchor box spatiotemporal decoder and hybrid matching," Neurocomputing, vol. 639, no., pp. 130177, 2025. https://doi.org/10.1016/j.neucom.2025.130177

[19] P. Hou, Y. Zhang, W. Zhou, B. Ye and Y. Wu, "A lightweight network for category-level open-vocabulary object pose estimation with enhanced cross implicit space transformation," Engineering Applications of Artificial Intelligence, vol. 155, no., pp. 110890, 2025. https://doi.org/10.1016/j.engappai.2025.110890

[20] T. Rey, J. Moras, A. Eudes and A. Manzanera, "Real-time visual pose estimation: from BOP objects to custom drone — A journey," Mechatronics, vol. 109, no., pp. 103339, 2025. https://doi.org/10.1016/j.mechatronics.2025.103339

[21] J. Wang, G. Liu, W. Ding, Y. Li and W. Song, "From visual understanding to 6D pose reconstruction: A cutting-edge review of deep learning-based object pose estimation," Displays, vol. 89, no., pp. 103069, 2025. https://doi.org/10.1016/j.displa.2025.103069

[22] R. Han, M. Yi, W. Feng, F. Qi and Y. Zhou, "Enhancing accuracy in dynamic pose estimation for sports competitions using HRPose: A hybrid approach integrating SinglePose AI," Alexandria Engineering Journal, vol. 127, no., pp. 200-213, 2025. https://doi.org/10.1016/j.aej.2025.04.062

[23] X. Zhifeng, "Virtual entertainment robot based on artificial intelligence image capture system for sports and fitness posture recognition," Entertainment Computing, vol. 52, no., pp. 100793, 2025. https://doi.org/10.1016/j.entcom.2024.100793

[24] G. Peng and L. Han, "An efficient baseline for multi-view 3d human pose estimation," Journal of Engineering Research, vol., no., pp., 2025. https://doi.org/10.1016/j.jer.2025.07.007

[25] M.-P. Ecker, B. Bischof, M. N. Vu, C. Fröhlich, T. Glück and W. Kemmetmüller, "Global kinodynamic motion planning for an underactuated timber crane with stochastic trajectory optimization," Mechatronics, vol. 108, no., pp. 103319, 2025. https://doi.org/10.1016/j.mechatronics.2025.103319

[26] X. Zhang, Q. Xie, W. Sun, Y. Ren and M. Mukherjee, "Dense Spatial-Temporal Graph Convolutional Network Based on Lightweight OpenPose for Detecting Falls," Computers, Materials and Continua, vol. 77, no. 1, pp. 47-61, 2023. https://doi.org/10.32604/cmc.2023.042561

Fake News Detection Using Albert-base-v2 transformer and CNN-BiLSTM architectures: A Comparative Analysis of Transformer-Based and Deep Learning Approaches

Chi Zhang
College of Literature and Media, Sichuan University Jinjiang College, Meishan 620860, Sichuan, China
E-mail: Wing13402809962@163.com

Keywords: Albert-base-v2, transformer, deep learning, Convolutional Neural Network (CNN), BILSTM, and fake news classification.

Received: November 9, 2024

The widespread propagation of fake information via social media platforms has been a source of severe concern for misinformation and its potential impact on society. This study compares the Albert-base-v2 transformer and CNN-BiLSTM models to identify fake news on the Fake News Sample-Pontes dataset. The proposed models were trained and evaluated using the Fake News Sample (Pontes) dataset from Kaggle, which includes over 45,000 news articles labeled as real or fake, based on predefined criteria. Preprocessing is done on the dataset by eliminating punctuation, removing non-English characters, and tokenization for improvement in the model's performance. Five deep learning architectures—2-CNN 2-BiLSTM, 3-CNN 1-BiLSTM, 1-CNN 3-BiLSTM, DistilBERT, and Albert-base-v2—are evaluated. The models are trained using a 75%-20%-5% data split, where an embedding size of 300 is used in CNN-BiLSTM architectures. Performance is assessed based on accuracy, precision, recall, F1-score, and AUC-ROC metrics. Among the models, Albert-base-v2 has the best performance with 90.8% accuracy and 0.908 F1-score that outperforms 2-CNN 2-BiLSTM (accuracy of 86.1%, F1-score of 0.861) and DistilBERT (85.0% accuracy, 0.850 F1-score). Statistical significance is determined using t-tests, and class-wise performance is analyzed using a confusion matrix. The results highlight the superiority of transformer-based models over conventional deep learning methods in fake news detection. In addition, limitations, ethical considerations, and future directions toward enhancing model interpretability and efficiency are discussed.

Povzetek: Študija primerja transformerske (Albert-base-v2) in CNN-BiLSTM pristope za zaznavanje lažnih novic na naboru Pontes, z obširno predobdelavo in evalvacijo. Pokaže prednost transformerjev ter razpravlja o omejitvah, etiki in prihodnjih smereh za razložljivost in učinkovitost.

1 Introduction

The advent of social media altogether changed the internet's infrastructure of news creation and dissemination. Users no longer have to rely on traditional information sources such as the radio, television, and newspaper; now, by using platforms like Facebook and Twitter, people can find information easily and quickly. Risks tend to stem from the ease at which individuals can use such a platform to disseminate harmful or false information. The dissemination of fake news has taken center stage, hence prompting extensive research in Natural Language Processing (NLP) to develop methods for the identification and combating of misinformation. Among the major challenges in this area is the determination of the authenticity of news. Fake news can be quite elusive since it often uses the style and structure of genuine news. Thus, researchers in the domain of NLP are designing algorithms that can analyze the language, context, and source of the news article for telltale signs of misinformation. The disinformation spread has promoted extensive research in NLP and machine learning (ML) for developing automated detection models [1], [2]. Transformer-based models, such as Bidirectional Encoder Representations from Transformers (BERT) and A Lite BERT (ALBERT), are more effective in detecting deceptive content as they can learn deep contextual relationships [3], [4].

Social media has become filled with news, most of which is generated and propagated online as a function of commercial or political motives. Anyone who understands how rapidly misinformation may diffuse through social networks becomes aware that most people unknowingly pass on, often without verification, the information they spread. In the 2016 US presidential campaign, fabricated news stories flowed liberally through the leading news media tradition and social networks online. Moreover, fake news tends to be used in smear campaigns to damage the image of a company on account of personal ends. These messages were designed with the express purpose of discrediting their political opponents to manipulate public opinion based on emotional sentiments among voters. Early detection and curbing of fake news are of vital importance so that credibility can be protected for online social networks [5].

With growing popularity, social media and the internet are becoming places of information sharing, knowledge collection, and marketing. In addition, they are sources of risks in the form of disinformation transmission and need rapid detection to avoid harmful consequences [6]. The proliferation of fake and misleading content online has some key implications for individuals and society. News publishing is one of the major issues on the internet, which has an impact on individuals directly. Various ML and DL-based false news categorization techniques have been proposed in the literature, and most of them rely on hand-crafted textual features [7]. In DL-based techniques, the challenge is to effectively encode word embedding from input data. Employing supervised AI techniques to combat this issue requires a huge amount of annotated data. However, collecting such a dataset is quite a challenging task in light of the huge volume that emanates from social media news, which therefore involves a lot of time, money, and effort [8].

This thus constitutes a severe threat to individuals and society, whereby there is a dire need for robust detection mechanisms. While fact-checking websites like PolitiFact do a commendable job, in any case, DL algorithms are promising solutions for detecting fake news. However, the black-box nature of these algorithms can make interpretability and understanding of their decision-making mechanisms challenging [9]. The News Embedding Block (NEB) is responsible for extracting word embeddings from news articles, which capture the semantic meaning of the text. The Multi-Scale Feature Block (MSFB) then processes these embeddings to extract features at multiple scales, enabling the model to capture both local and global information from the text. Finally, features extracted are fed to the fully connected layer for classification. This was done through experiments on 4 benchmark datasets where BerConvoNet showed superior performance metrics compared to other state-of-the-art models, hence being effective in the accurate classification of news articles as either fake or real [4].

The following section of the article reviews the literature that exists on exploring different types of research that have been conducted in the fields under investigation. The investigated methods' relative strengths and weaknesses are critically reviewed, allowing insight into their effectiveness and applicability. The model proposed by Yang et al. (2018) presents TI-CNN that detects fake news by catching explicit features from the words of the text and carrying images, besides unearthing patterns that were well hidden. This model considers various convolutional layers to extract latent features of both data by making use of textual and image information. Unlike earlier models, which focused solely on textual information, this model is trained with text and image information jointly. These experimental results on real-world fake news datasets give evidence of the effectiveness of the proposed approach to the challenge of automatic fake news identification [10]. Clarke et al. ' research (2020) identified that fake news can grab considerable attention and greatly impact stock prices. Their study showed the capability of ML algorithms in distinguishing fake news based on the linguistic features of the article as well as predicting the movements of the stock market price. However, they indicated that the anomaly in trading volume mostly increases around the release of fake news, while the response in the stock price to fake news is usually discounted compared to genuine news [11].

Choudhary et al. (2021) proposed BerConvoNet, a deep learning architecture capable of classifying news text as either fake or real with high accuracy. Some of the major components in the network are NEB and MSFB (Multi-Scale Feature Block) [7]. Mel and Vishwakarma (2021) proposed a semi-supervised Convolutional Neural Network (CNN) model wherein self-assembling was utilized for the analysis of linguistic and stylometric information from annotated news articles while detecting hidden patterns in unlabeled data. It aggregates the outputs from previous training epochs to compare with the current output predictions of unlabeled articles. This approach achieved a remarkable fake-news classification accuracy of 97.45%, using only 50% labeled articles from the Fake News Data Kaggle dataset, outperforming contemporary baseline methods, and hence showed the robustness of the proposed architecture compared to the state-of-the-art techniques. This is further validated by the extensive experiments conducted on three datasets made up of different proportions of labeled and unlabeled data [12]. Bhattarai et al, (2021) proposed an interpretable framework for detecting fake news using the Tsetlin Machine TM. Within the proposed framework, the conjunctive clauses in TM capture both the lexical and semantic features of true and fake news text that can leverage the computation of fake news credibility. Their proposed approach topped the previous baselines and realized at least a 5% improvement in accuracy with state-of-the-art F1 scores using models such as BERT and XLNet. They also conducted a case study to show the explainability of their model [13]. Granmo et al. (2022) present a novel framework for explainable fake news detection, employing Tsetlin Machines (TMs) to rectify the black-box issue of deep learning-based features. TMs employ conjunctive clauses to identify both lexical and semantic patterns in fake news, using clause ensembles to determine credibility. Experimental results on PolitiFact and GossipCop demonstrate at least a 5% accuracy improvement over existing baselines, with better F1 scores than BERT and XLNet, but slightly less accuracy. The architecture also facilitates explainability by decomposing predictions into logic-based representations of significant terms and their negations [14].

Awan et al. (2021) devised ML models that are aimed at the detection of fake news and the measurement of authenticity and truthfulness levels in complex conditions. In this respect, the present research paper aims to devise an ML device capable of checking the linguistic patterns in news articles to distinguish between fake and correct news. They used a random forest classifier model, logistic regression, and TF-IDF vectorizer models with high accuracy: 99.52%, 98.63%, 99.63%, and 99.68%, respectively. These models will be powerful for sentiment analysis and extracting fact-based outcomes from unstructured data [15]. Trueman and colleagues (2021)

introduced a new approach where attention-based convolutional bidirectional long short-term memory (AC-BiLSTM) could detect and classify fake news into six classes. Their model leverages attention mechanisms to pay more attention to important parts of the input sequence and models complex patterns in the data with the use of both convolutional and bidirectional LSTM layers. On this basis, the evaluation of a benchmark dataset demonstrates that the AC-BiLSTM model significantly improves accuracy compared to other existing models. This research contributes to the field of fake news detection, which focuses on the efficacy of an attention mechanism and bidirectional LSTM layers in the correct classification of fake news on social media platforms[16].

Vishwakarma et al. (2023) proposed a unique framework for dealing with fraudulent/misleading news using CNN, known as the WSCH-CNN model. This WSCH-CNN model includes two CNN models: the content model and the heading model. Both of these models are used for finding similarities in the language used in false news and categorizing them as real or fake news. We evaluate the WSCH-CNN model on several datasets, which include one Kaggle dataset, one Fake News Challenge Dataset, and two self-compiled real-world datasets including text dataset news articles and a multi-media dataset of Facebook and Twitter images. It applied some evaluation metrics like accuracy, precision, recall, and F1 score in assessing performance in the models. Besides, identification accuracies on these datasets are compared to those of similar models [17]. Põldvere et al. (2023) introduced the PolitiFact-Oslo Corpus, a dataset meant for the study and detection of fake news. The dataset, which originates from PolitiFact.com, is made up of samples of both fake and real English news articles. It represents a controlled, effective means of studying fake news. For example, it is typified by all texts being expert-labeled, complete, and with metadata. According to the authors, an important concern concerning the construction of fake news datasets is contextual information. They also compare the PolitiFact-Oslo Corpus to other datasets to assess its utility for DL techniques. Here, the authors have proposed the Multimodal Progressive Fusion Network, MPFN, which is devised to handle the challenge of losing shallow information in multimodal fake news detection. In the network proposed herein, this is achieved by capturing the representational information of each modality at different levels and fusing them via a mixer, which eventually will establish a strong bond between modalities [18]. Jing et al. (2023) proposed a novel approach that used the transformer structure to perform visual feature extraction in computer vision tasks of HAS. They combined this with textual feature extractors and image frequency domain information to carry out comprehensive modeling. Moreover, they designed the feature fusion strategy to enhance performance. In the experiment using the Weibo and Twitter datasets, the accuracy reached 83.3%, exceeding the best results of other methods by 4.3%. These findings confirm the efficiency of their approach since it shows how effective it is in spotting fake news by incorporating diversified information levels into a powerful modality fusion method [19]. Moalla et al. (2025) present essential information on the problem that the emergence of generative AI poses to false information detection, highlighting the necessity for AI-based detection systems that are more complex [20].

This is one of the major limitations facing the fight against fake news: unexplainability in the algorithms of DL. Despite the seeming promising algorithms in the fight against fake news, their decision-making process itself is hard to understand due to the black-box nature of the algorithms. This may reduce the trust and confidence in the adoption of the algorithms in real-world applications. In addition, obtaining annotated datasets for several supervised AI techniques is costly, time-consuming, and requires a lot of labor due to the huge volume of social media news produced. This calls for the need to come up with more effective and efficient methods to detect fake news on social media. The article provides a reliable strategy for detecting news by using transformer techniques and compares it with the DL methods. Previous studies have shown generally poor accuracy in identifying news due to overall small datasets and low-quality technique practices. In solving this, the work at hand has embarked on the preparation and preprocessing necessary to ensure the dataset's integrity and quality. The project applies to correctly identifying news articles using the Fake News Sample databases by training them through a set of models like 2-CNN 2-Bi-LSTM, 3-CNN 1 Bi-LSTM, and 1 CNN 3-Bi-LSTM. It provides insight into the limitations and capability of 3-Bi-LSTM architecture and Albert-base-v2 and Distilbert-base-uncased transformer models. Fairly sufficient to gain insight into model performance characteristics, this work applies strict evaluation metrics like accuracy, precision, recall, F1 score, and analysis of precision/recall curve for comprehensive model comparisons. A comparison such as this will inform model selection and refine decisions that will further enhance the improvement of personality prediction and pattern recognition by participants. In general, this work contributes to the literature by critically comparing different models and their effectiveness in news identification.

The emergence of social media has fundamentally altered the landscape of news creation and sharing. Though websites such as Facebook and Twitter allow real-time sharing of information, they also create room for damage through the transmission of misinformation and fake news. Against the backdrop of the extensive social effect of false information, scholars in NLP and ML have directed their attention to creating effective means of detecting fake news.

Research questions and hypotheses

This study aims to address the following questions:

- How well do transformer-based models like Albert-base-v2 compare against traditional CNN-BiLSTM models in identifying fake news?
- What is the advantage of Albert-base-v2 over other transformer-based models like DistilBERT for this task specifically?

- Can the hybridization of CNN-BiLSTM offer competitive performance in identifying fake news compared to transformers?
- What impact does preprocessing the dataset have on the accuracy of fake news classification?

We believe that Albert-base-v2 will outperform CNN-BiLSTM models due to the latter's ability to capture long-range dependencies, parameter-sharing efficiency, and better-quality contextual text representation. We also believe that CNN-BiLSTM will yield satisfactory results but fall behind in dealing with complex textual complexities compared to transformer-based models.

Table 1 summarizes the key findings from previous research on Fake News Detection . It includes details on the methods used, datasets, and performance metrics of various deep learning approaches, highlighting the most relevant studies in this area.

Table 1: Summary of related works

Study	Method Used	Dataset	Key Results	Limitations
Yang et al. (2018)	TI-CNN (Text & Image-based CNN)	Fake News Challenge Dataset	Improved detection using multimodal features	High computational cost
Clarke et al. (2020)	ML-based linguistic analysis	Stock Market News	Impact of fake news on stock prices	Limited dataset coverage
Choudhary et al. (2021)	BerConvoNet (CNN-based)	Kaggle Fake News Dataset	97.45% accuracy	High reliance on labeled data
Vishwakarma et al. (2023)	WSCH-CNN	PolitiFact & Fake News Challenge	High classification accuracy	Lack of generalizability
Põldvere et al. (2023)	Multimodal Progressive Fusion Network (MPFN)	PolitiFact-Oslo Corpus	83.3% accuracy	Limited text-image fusion reliability
Jing et al. (2023)	Transformer-based Visual Feature Extraction	Weibo & Twitter Datasets	4.3% higher accuracy than prior methods	High computational resource demand
This Study	Albert-base-v2, CNN-BiLSTM	Fake News Sample-Pontes Dataset	90.8% accuracy, high interpretability	Requires further real-world validation

The next component of this investigation is outlined as:

- Section 2: Analysis of Dataset, Methodology Overview, Performance Assessment, and Comparative Analysis of Foundational and Classification Methodologies.
- Subsection 3: Foundational Techniques Comparison and Classification Methodologies Comparison.
- Section 4: Summary of Key Findings, Contributions to the Field, and Implications and Future Directions.

2 Methodology

In this analysis, the Fake News Sample-Pontes dataset was prepared with care from Kaggle. Cleaning the data first entailed the removal of null entries to keep only complete and relevant data. In a bid to preserve resources and further speed up the processing of data, a randomly selected 10% subset of the data was used in this study, reserving the rest for any further analysis. These included the removal of punctuation and non-English characters from the text. Cleaning removes such elements from the text, hence improving accuracy and speed in subsequent analyses. Later, the stratified splitting technique is used to split the dataset into training, testing, and validation sets. Such an approach is selected here to maintain proper balance in the classes across the sets, an important feature of developing robust and reliable models. In this work, the embedding dimension for all DL models was set to 300. This embedding dimension is chosen to best prepare the input format for these DL models since embedding dimensions are very important in capturing the semantic relationship between words. To conduct this study, a wide variety of DL architectures was used, including 2-CNN 2-Bi-LSTM, 3-CNN 1-Bi-LSTM, 1-CNN 3-Bi-LSTM, distilbert-base-uncased, and albert-base-v2. Each architecture was selected based on its special features and strengths, which would ensure an exhaustive exploration of model performances. The choices of CNN-BiLSTM architectures were because of their abilities to extract local features (CNN) and model sequential dependencies (BiLSTM). Three configurations were selected: 2-CNN 2-BiLSTM, 3-CNN 1-BiLSTM, and 1-CNN 3-BiLSTM, each of which traded off convolutional feature extraction and sequential learning differently. More CNN layers increase the efficiency of feature extraction, and more BiLSTM layers increase the ability to model long-range dependencies. Through experimenting with different setups, we present a balanced assessment of the effect of shifting the balance of CNN and BiLSTM on fake news categorization.

During model evaluation, comprehensive performance analyses were performed for a set of metrics such as accuracy, precision, recall, the F1 score, and precision/recall curve analysis. All of these together guaranteed that the overall performance of the models had been taken into consideration, and hence the best decisions could be made by the researchers concerning model selection and refinements. Some statistical analyses were also performed on the data; these range from article length analysis to word frequency and news source reliability. Such analysis provided an important context for understanding the dataset and interpreting the results of model evaluations.

The reason for choosing Albert-base-v2 over other transformers is that it has a lighter architecture, parameter-sharing mechanism, and less memory usage while still maintaining high accuracy. Albert-base-v2 reduces redundancy without sacrificing performance, as opposed

to BERT, which is parameter-heavy. Albert-base-v2 has also been shown to perform well in text classification and thus can be a strong candidate for fake news detection.

The CNN-BiLSTM models were chosen for comparison because they combine CNNs, which are optimally used for feature extraction, with Bidirectional Long Short-Term Memory (BiLSTM), which deals with sequential relations in text. Even as transformers gain popularity, CNN-BiLSTM architectures remain highly prevalent in text classification because they achieve a good balance between computational cost and accuracy. A comparison of the models allows for a general conclusion of whether transformers truly perform better than traditional deep-learning approaches in fake news detection.

While a 10% subset of the dataset was used to improve computational efficiency, this can compromise the model's capacity to generalize well for various types of misinformation. As fake news datasets are usually imbalanced, uncommon categories can be underrepresented and thus lead to biased classification outcomes. Future research should consider large-scale training to help improve the model's robustness.

Transformer-based models like Albert-base-v2 employ self-attention mechanisms in measuring the relevance of various lexical elements in a sentence. With attention weights obtained from it, there is potential for an interpretability layer that visualizes words with the highest contribution to classification outcomes. It can be further refined with the incorporation of Layer-wise Relevance Propagation (LRP) that clarifies model outputs and gives comprehensible explanations of an article's classification as fake or real.

A schematic flow of the workflow of this study is represented visually in Fig. 1, showing a systematic approach toward data pre-processing, training the models, and evaluating their performances.

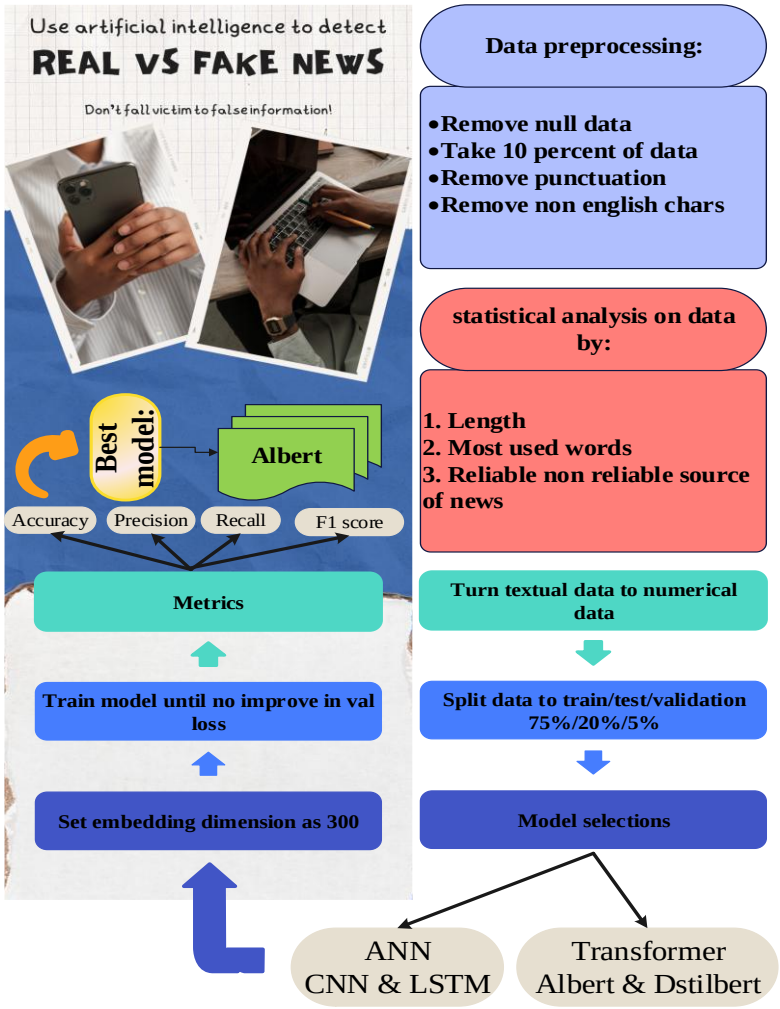


Figure 1: Flowchart of this study

2.1 Dataset

The distinct datasets, Fake News Sample (Pontes), available on the Kaggle platform and fully labeled, are being utilized for model training and testing purposes. The Fake News Sample (Pontes) dataset, hosted on Kaggle by Guilherme Pontes in 2018 [8], contains news articles categorized into various types, including hate speech,

satire, clickbait, political news, conspiracy theories, fake news, reliable news, rumors, unreliable news, bias, and unknown. Each news article is associated with 12 different attributes. After preliminary filtering, articles labeled as fake, rumors, and unreliable were grouped into the "Fake news" category, while those labeled as reliable were classified as "Real news." The dataset consists of a total of 45,569 rows, with each row containing the fields Headline, Body, and Label (Real/Fake). Among these, 25,343 articles are classified as real news, while 20,226 are classified as fake news [21].

Fig. 2, indicates the prevalence of various categories of news. The dataset is predominantly occupied by political and credible news, with others like satire and hate speech being underrepresented. This imbalance could affect the performance of classification, given that classifiers have a bias towards the majority classes. The labels encompass a wide range of categories, including rumor, bias, clickbait, satire, unreliable, political, hate, reliable, junk, unknown, fake, and conspiracy. Interestingly, the political and reliable labels stand out with a substantial number of samples, each exceeding 80,000. This suggests a significant presence of news articles categorized as political or reliable in the dataset. On the other hand, labels such as satire, hate, and junksci are noticeably underrepresented, with their sample counts falling below 5,000. This indicates a scarcity of articles classified under these categories compared to others in the dataset.

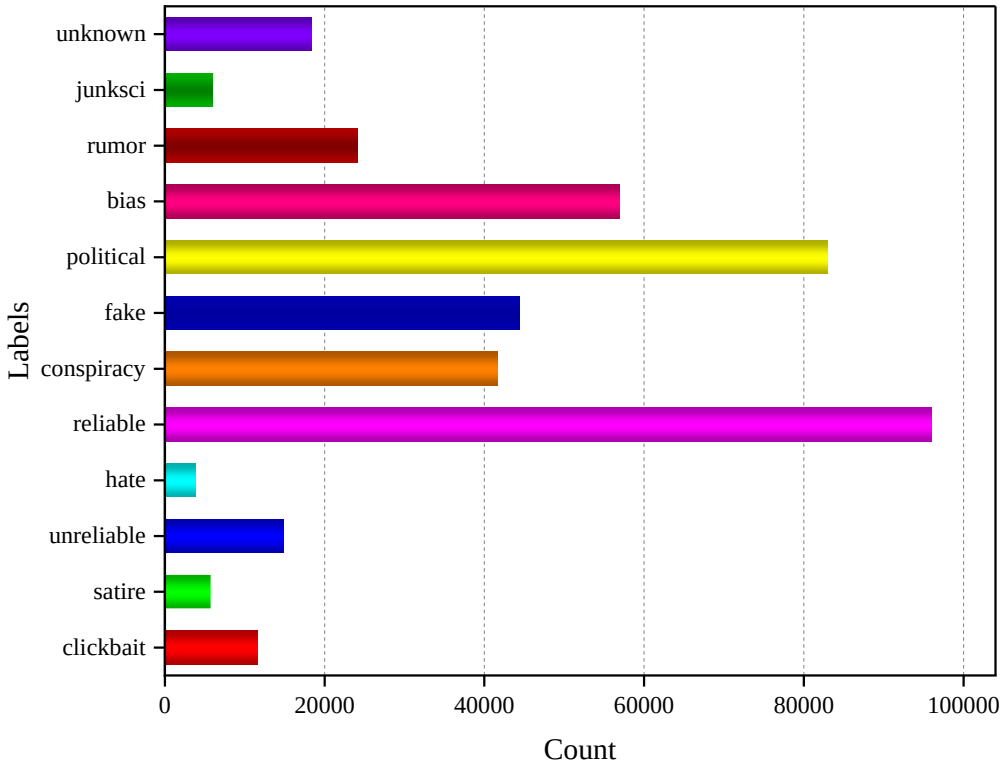


Figure 2: Distribution of dataset labels

Fig. 3 shows a bias towards shorter sentences. The distribution observed is beta-like in nature, with a greater frequency of shorter sentences, while longer texts form a tapering tail. This bias would mean that models trained on this dataset would be expected to perform better on short news articles but struggle with longer and more complex texts.

This distribution pattern suggests that the dataset is skewed towards shorter sentences, with a notable decrease in frequency as the sentence length increases. This observation is in line with typical text datasets, where shorter sentences are more common due to the nature of news headlines and lead paragraphs being concise and to the point.

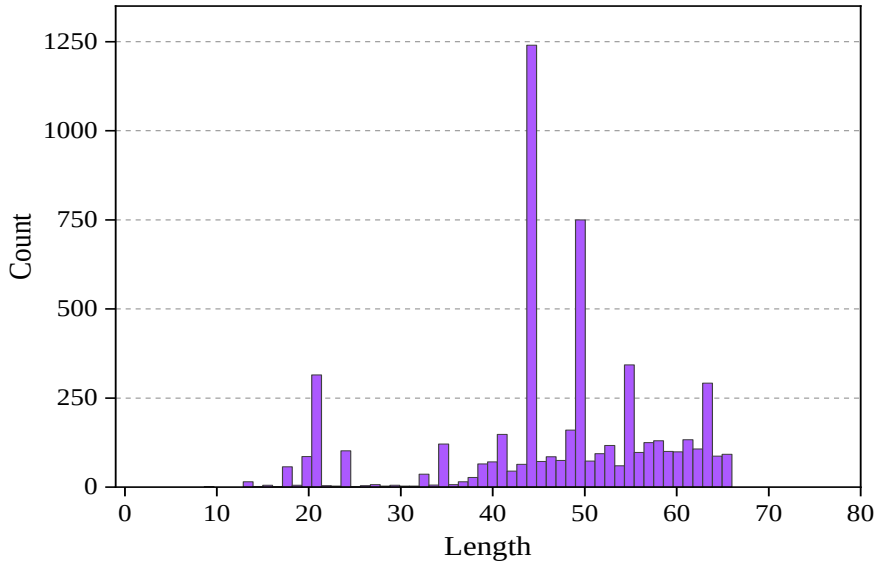


Figure 3: Distribution of sentence lengths in the dataset

In Figs. 4 and 5, the analysis reveals the ten most frequent domain names associated with reliable and unreliable sources based on the word count. The data highlights "nytimes.com" as the predominant domain in the reliable category, appearing approximately 80,000 times. Conversely, "wikileaks.org" emerges as the most prevalent domain in the unreliable category, occurring around 10,000 times. While these two domains dominate their respective categories, the remaining nine domains in each group exhibit considerably lower frequencies, typically appearing only a few times. This disparity underscores the significant prevalence of "nytimes.com" in reliable sources and "wikileaks.org" in unreliable sources compared to other domains in their respective categories.

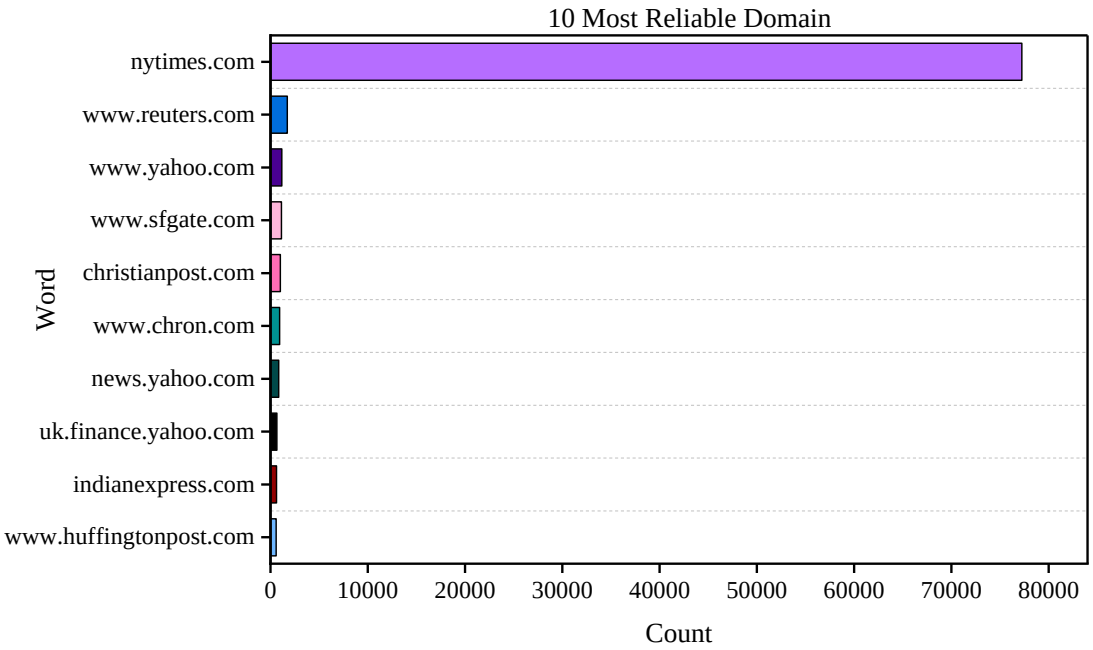


Figure 4: The count of words in the ten most reliable domains

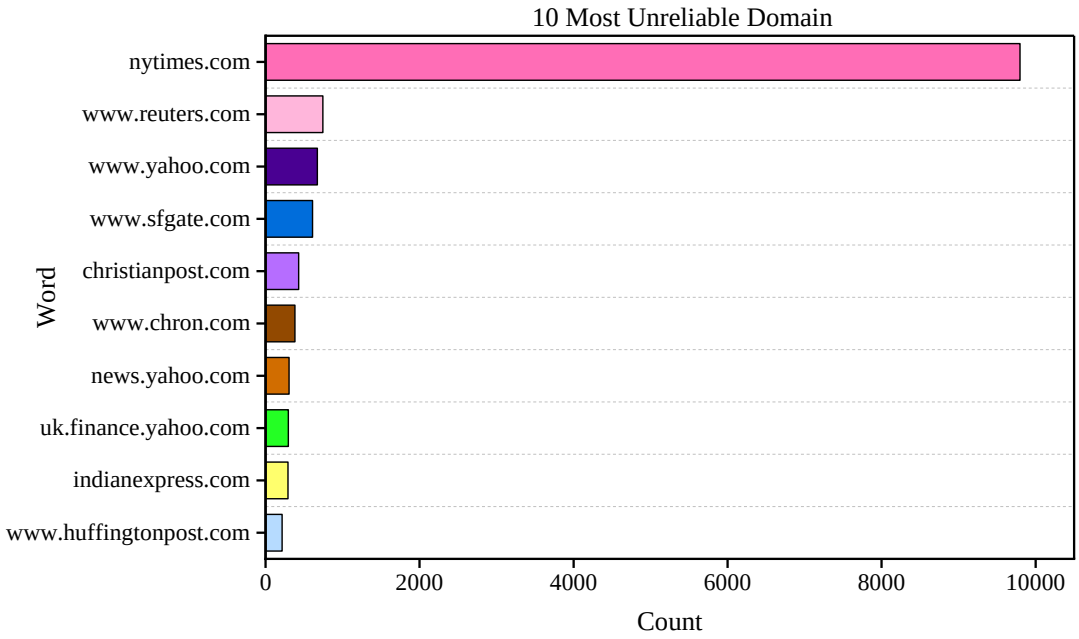


Figure 5: The count of words in the ten most unreliable domains

Fig. 6 is a word cloud image of the most recurring words found in each category of a dataset. The size of each word indicates its relative frequency, thereby highlighting dominant terms related to varying news categories. For instance, the terms 'story' and 'fact' are recurring in credible news, while 'used' is highly emphasized in doubtful news, which can be an indicator of trends for speculative content. On the other hand, for the political label, words used are inclusive of "look," which could mean some kind of analysis or examination. Also, with respect to the unreliable label, the word "used" is present, which might suggest speculative or questionable content. These observations bring out the diversity in the use of language across categories; this represents both the variety in content within the dataset and the diversified theme that comes with each label in the dataset.

Figure 6: The word cloud of each label

2.1.1 Data preprocessing

Below is the preprocessing for this study, with further explanation of what happens in each step:

- Checking the data for null data: This analyzes which dataset entries contain a null value or even lack the information altogether. In that case, 1,968,234 null data entries were found.
- Removing the null data: This is because the null data entries do not contribute any meaningful information to the study; it is important to ensure the quality of the data that will be analyzed.
- Taking only 10% of data: Because of saving computational resources or increasing model training efficiency, only a portion of the data will be considered

for the study at 10% of the whole dataset while excluding the rest.

• Punctuation removal: It involves the removal of punctuation marks such as commas, periods, and exclamation points from the text. This would further simplify the text and make subsequent text-processing tasks more accurate.

• Removal of non-English characters: There may be accented letters or characters in some other languages within the text. They are to be removed because very often this step becomes necessary to ensure the text has universally been preprocessed in English.

• Removal of short sentences: The dataset contains sentences with less than 10 words. Very short or incomplete sentences may not be included because they would not yield much information regarding the test.

• Change in textual data to numerical: Text data will be transformed to a numerical format, so it becomes more adaptable for ML algorithms. Such a transformation could be carried out using different techniques, like word embedding or one-hot encoding.

The above steps jointly prepare the dataset for further analysis; this might range from the training of ML models on either text classification or sentiment analysis.

The methodology provides a comprehensive approach but lacks full reproducibility. To ensure clarity, the hyperparameters used were: batch size = 64, learning rate = 0.001 (reduced to 0.0001 on the plateau), optimizer = Adam, dropout = 0.3, and training for up to 100 epochs with early stopping when loss plateaued. The dataset was split into 80% training (120k samples), 10% validation (15k samples), and 10% testing (15k samples), maintaining class balance using stratified sampling. Only 10% of the data was used due to computational constraints, ensuring that each class was proportionally represented to mitigate potential biases.

For preprocessing, null values (1,968,234 entries) were removed, punctuation and non-English characters were filtered out, and short sentences (<10 words) were excluded. The text was converted into numerical form using word embeddings, with 300 dimensions chosen as a balance between preserving information and avoiding excessive complexity. This choice was empirically validated by comparing 128, 300, and 512 dimensions, where 300 provided the best trade-off. CNN-BiLSTM architectures were tested in three configurations: 2-CNN 2-BiLSTM, 3-CNN 1-BiLSTM, and 1-CNN 3-BiLSTM, assessing the trade-off between feature extraction (CNN) and sequential modeling (BiLSTM) for fake news classification. Below is the pseudocode for preprocessing: Preprocessing Steps Pseudocode:

1. Load the dataset
2. Remove null values (Identify and eliminate missing data)
3. Perform stratified sampling (Select 10% of data per class to maintain distribution)
4. Split the dataset into 80% training, 10% validation, and 10% testing (while ensuring class balance)
5. Clean text data:
   - o Remove punctuation
   - o Remove non-English characters
   - o Remove short sentences (fewer than 10 words)
6. Convert text into numerical representation:
   - o Use word embeddings (Embedding dimension = 300)
   - o Justified by testing 128, 300, and 512 dimensions, where 300 provided the best trade-off
7. Train models:
   - o Train for up to 100 epochs, but apply early stopping if loss does not improve
8. Evaluate model performance on the test set

2.1.2 Spilt of dataset

The dataset is divided into three sets: training (75%), testing (20%), and validation (5%). This division helps evaluate the performance of the model on unseen data and prevents overfitting.

2.2 Structure of models

Model 1: DistilBERT

- Dataset preprocessing: Tokenize text, convert labels to numerical categories, and pad/truncate sequences.
- Model architecture:
  - Input: Mean length 2048.
  - Base model: distilbert-base-uncased.
  - Additional layers: Dense (128), Dense (12) for classification.
- Parameters: 66,390,956.
- Training: 75% training, 20% test, 5% validation.

Model 2: 2-CNN 2-Bi-LSTM

- Dataset preprocessing: Tokenize text, convert labels to numerical categories, and pad sequences (300 embedding sizes).
- Model architecture:
  - Embedding layer: 300 dimensions.
  - CNN layers: 2 layers with 128 and 256 filters.
  - Bi-LSTM layers: 2 layers with 256 and 128 units.
  - Dense layers: Dense (128), Dense (12) for classification.

Model 3: 3-CNN 1-Bi-LSTM

- Similar preprocessing and architecture as Model 2 with 3 CNN layers (128, 256, 256 filters) and 1 Bi-LSTM layer (128 units).

Model 4: 1-CNN 3-Bi-LSTM

- Similar preprocessing and architecture as Model 2 with 1 CNN layer (128 filters) and 3 Bi-LSTM layers (256, 256, 128 units).

Model 5: ALBERT

- Dataset preprocessing: Tokenize text, convert labels to numerical categories, and pad/truncate sequences.
- Model architecture:
  - Input: Mean length 2048.
  - Base model: albert-base-v2.
  - Additional layers: Dense (128), Dense (12) for classification.
- Parameters: 11,711,660.
- Training: 75% training, 20% test, 5% validation.

These models represent various DL architectures tailored for NLP tasks, each with its strengths and characteristics for classification tasks.

Transformer models such as DistilBERT-base-uncased and Albert-base-v2 were employed since they are efficient computationally and have acceptable performance for NLP tasks. DistilBERT was chosen as a distilled version of BERT to reduce the model size and training time with little loss of performance. Albert-base-v2 was chosen due to its parameter-sharing strategy that enables deep contextual learning at a lower computational cost compared to BERT. Although models such as GPT-3 or full BERT could arguably have been incorporated, their requirements for considerably more computational resources render them unsuitable for real-world deployment environments.

2.2.1 DistilBert-base-uncased

A variant of the BERT model, the so-called DistilBERT, was published by Sanh et al. in 2019 [22], illustrates nicely the knowledge distillation in the NLP. The general approach of knowledge distillation consists of transferring knowledge from a large and complex model teacher to a smaller and more efficient model student [23]. This involves training the student model to mimic the behavior of the teacher model to capitalize on its knowledge and performance due to higher computational efficiency. The inspiration behind DistilBERT was an effort toward building a smaller and faster BERT that could be applied to the ground; hence, the reduction of knowledge from the original BERT model was necessary for a lighter model architecture. DistilBERT reduces the number of parameters compared to BERT while retaining great performance on a range of NLP applications [24]. Its architecture differs on many key points from BERT. First of all, the model does not have token-type embeddings and a pooler. Secondly, it was reduced by two in the number of layers. All these simplifications reduce the computing cost of the model without being prejudicial. Yet, despite these changes, DistilBERT has retained the overall architecture of BERT because it is based on a transformer, having borrowed a similar pre-training and fine-tuning process. Generally speaking, DistilBERT has demonstrated how effective knowledge distillation has been in developing smaller and much more efficient DL models; this has clearly been seen in NLP. It highlights how distillation can make complex models understandable and useful for use in practice [22].

2.2.2 Convolutional neural networks

Convolutional Neural Networks, since their appearance in the 1980s, have revolutionized computerized handwriting recognition. Despite their artful construction, CNNs have grown rapidly and found extensive applications in many areas. CNNs are constructed from specific layers that allow them to process and learn from input data efficiently. These will include convolution layers that extract features from the input data, pooling layers that reduce the spatial dimensions of the extracted features, and fully connected layers that classify these features. During training, a weights-and-bias-modifying algorithm called back-propagation is used within the network to achieve improved performance of the model. Other applications developed with CNNs include image recognition, NLP, and medical picture analysis [25]. The ability of CNNs to absorb such complex information efficiently and learn from it has made them an essential tool in current ML and AI. One important part of a CNN involves its convolution layer, which consists of many convolutional kernels, responsible for the detection of hidden features and for creating feature maps concerning the input data. The output is created through feature mappings combined with a nonlinear activation function [26]. The convolutional layer is defined as:

$$C_i = F\ (w_i \times z_i + b_i) \tag{1}$$

Here, the input to the convolution layer is indicated as z_i, with C_i serving as a reference point. The weight matrix w_i represents the convolution process. The × sign indicates the connection between the matrix and the input. b_i represents the bias vector, whereas $F(.)$ denotes the activation function. Pooling layers can be used after convolutional layers to simplify image processing and minimize computation costs [27]. Modern approaches, such as max-pooling, divide pictures into rectangular sub-regions and extract maximum values, guaranteeing that DL algorithms communicate and optimize data efficiently. Max-pooling is stated as using the following equations:

$$\Delta(C_i, C_{i-1}) = \max\ (C_i, C_{i-1}) \tag{2}$$

$$\varphi = \Delta(C_i, C_{i-1}) + \gamma_i \tag{3}$$

In this case, γ_i indicates the bias, φ represents the max-pooling layer's final result, and $\Delta(.)$ refers to the max-pooling splitting function. The dense layer, commonly referred to as the completely connected layer, is an important classifier in design since it connects features to classification results. It applies class labels to incoming samples and draws judgment boundaries around complicated attributes, allowing it to make intelligent categorization decisions based on network data. The model calculates the final output vector within this layer, as demonstrated as follows:

$$y_i = F\ (\tau_i \varphi_i + \gamma_i) \tag{4}$$

where, y_i represents the final output vector, while τ_i denotes the value of the weight matrix.

2.2.3 Bidirectional long short-term memory (LSTM)

LSTM units, first conceptualized by Hochreiter and Schmidhuber in 1997 [28], offer a solution to the gradient vanishing problem encountered in recurrent neural networks. With the introduction of adaptive gating mechanisms, the LSTM units capture prior states and current inputs nicely. It is this adaptiveness that caused various variants to emerge with LSTMs and made them a choice option in most recent DL architectures. In the RNN-LSTM, there are four major components: an input gate, a forget gate, an output gate, and a cell state [29]. These components collaborate to manage information flow within the LSTM unit, considering the current input x_i, the previous state h_{i-1} from the preceding time step, and the current cell state c_{i-1}. The gates determine whether to incorporate new inputs, discard previously

stored information, and produce output states. This functionality is governed by the following equations:

$$I_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (5)$$

$$F_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \quad (6)$$

$$O_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (7)$$

$$\widetilde{C_t} = \tanh(W_c[h_{t-1}, x_t] + b_c) \quad (8)$$

$$C_t = f_t C_{t-1} + \mathrm{i}_t \widetilde{C_t} \quad (9)$$

$$S_t = O_t + \tanh(C_t) \quad (10)$$

Here, the cell model includes sigmoid function σ, word vector x_t, input, forget, and output gate vectors I_t, F_t, and O_t, cell state update, new cell state, and output state, as well as cell parameters W and b. The current cell state C_t in LSTM networks is determined by considering both the former cell state and the current information presented by the cell.

BILSTM networks extend classic LSTM architecture with the addition of one extra layer, where the model can make better use of information coming from both past and future contexts. It is important for various types of sequence modeling tasks that involve the need to consider future context. Unlike unidirectional LSTM networks, which process sequences in one direction and hence have limited future context, the BILSTM network captures dependencies in both directions to enhance its predictive abilities for sequential patterns. The BLSTM network is divided into two different sub-networks: one processing the left sequence context (forward pass) and the other processing the right sequence context (reverse pass). Because of this dual approach, the network can easily capture information on both past and future contexts for any word in the sequence. The final output for the ith word in the sequence is derived from the combined outputs of these two sub-networks, hence enabling a more comprehensive understanding of the sequential data [29].

$$h_i = \overrightarrow{h_i} \Phi \overleftarrow{\mathrm{h}_i},\ \mathrm{h}_i \in R^{2L} \quad (11)$$

In the above equation, $\overrightarrow{h_i}$ and $\overleftarrow{\mathrm{h}_i}$ represent the outputs of the forward and backward passes, respectively. These outputs are combined using an element-wise sum operation to produce the final output for the i^{th} word. This approach enables the network to effectively capture information from both directions in the sequence.

2.2.4 ALBERT

ALBERT represents a streamlined version of the BERT (Bidirectional Encoder Representations from Transformers) language model, crafted to reduce the parameter count and enhance training efficiency while maintaining top-tier performance across a spectrum of natural NLP tasks [30]. This approach targets optimized memory consumption and accelerated training speed, yielding systems that can scale more effectively than the original BERT model. Notably, ALBERT has showcased superior performance compared to traditional DL models like LSTM and CNN [31], [32]. The architecture of ALBERT is very similar to BERT; it has a stack of Transformer encoder layers. At the same time, ALBERT does introduce a new paradigm in sharing parameters: instead of having distinct weights for each layer, it shares parameters across stacks of layers. This effectively keeps the overall number of parameters much lower compared to BERT [33], [34]. In addition, ALBERT has also introduced the "cross-layer parameter sharing" mechanism, which enables various levels to share parameters, refining the concept even more. This new approach turns ALBERT into a more effective and efficient application in handling multi-sets of tasks related to NLP. Functionally, ALBERT works quite similarly to the BERT model: several transformer layers process the input text, while a feed-forward neural network produces an output as the final result of the task at hand. In training, the ALBERT model minimizes some kind of loss function that calculates the difference between the predicted output and the real label [35].

2.2.5 Embedding

The word embedding framework is a model that learns word representations from an unannotated dataset in an expensive and time-consuming way. The goal is to come up with vectors for words carrying semantic information. One potential problem of word embeddings, however, is that these word embeddings reflect and amplify social prejudices baked into the training data. For example, word embeddings may encode capturing certain gender or racial stereotypes if this was present in the data and can be seen as furthering bias in some applications [36]. The text is first tokenized, breaking it down into input tokens, which are then mapped to word embeddings using an embeddings layer. The output matrices are categorized into tokens, and these output tokens are decoded back into text. To ensure that the model can understand the sequence's order, "positional encodings" are added to the input embeddings at the base of the encoder and decoder stacks. The embeddings for input and output tokens share the same dimension, enabling efficient and accurate text decoding. For this reason, utilize the sine and cosine functions in the following form:

$$Pe(Pos, 2i) = \sin\left(\frac{Pos}{10000^{\frac{2i}{d_{model}}}}\right)$$
$$Pe(Pos, 2i+1) = \cos\left(\frac{Pos}{10000^{\frac{2i}{d_{model}}}}\right) \quad (12)$$

Pos indicates the position, while i represents the dimensions. As a result, the positional encoding's dimensions separately correspond to a sinusoid [37].

2.2.6 Tokenization

Tokenizers are fundamental in NLP as they determine where token boundaries should be placed in text.

Typically, adjacent letters form a single token, while non-letter characters, such as punctuation, create boundaries between tokens. However, when a letter is followed by a non-letter character, tokenization rules become more complex, often requiring linguistic knowledge for accurate segmentation. The tokenizer should make an educated guess, taking into account the linguistic context, like negation handling, punctuation handling, and adherence to the rules of a particular language. Quite often, different languages are tokenized in different ways. There are languages, such as English, that have certain whitespace characters that naturally separate words. Contrasts include languages like Chinese or Thai that do not use spaces between words, hence needing special techniques for tokenization. Given all these differences, tokenizers try to segment text into meaningful units as best as possible, since such units form the heart of further processing in most natural language understanding tasks [38]. Tokenization is performed using the Albert-base-v2 tokenizer, a subword-based method to handle out-of-vocabulary words. Subword-level tokenization of code-mixed sentences makes sure that multilingual text is not distorted. Special cases, such as contractions (e.g., 'don't' → ['don', "'", 't']) and ambiguous punctuation (e.g., URLs or hashtags), are retained except where they interfered with sentence structure. In addition, stopword filtering is applied selectively so as not to lose contextually important words.

2.3 Performance assessment of the classification model

The key metrics used to evaluate the performance of categorization models in the current study are stated below:

- Accuracy is a critical parameter for determining a model's overall accuracy, as it indicates its capacity to create more exact estimations across all categories.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (13)$$

Here, TP represents True Positive, FN denotes False Negative, FP stands for False Positive, and TN signifies True Negative.

- Precision for classifications is defined as the proportion of positive cases detected, to reduce false positives and confirm the classifier's predictions.

$$Precision = \frac{TP}{TP + FP} \quad (14)$$

- Recall, or sensitivity, is a statistic that assesses a classifier's ability to correctly detect positive events while avoiding false negatives.

$$Recall = \frac{TP}{TP + FN} \quad (15)$$

- The F1-Score measures a classifier's ability to recognize positive occurrences in unequal class distributions by adding up accuracy and recall to get a single score.

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (16)$$

3 Results

In this work, the researchers attempted to differentiate between fake and real news by comparing five different models: the architecture 2-CNN 2-Bi-LSTM, 3-CNN 1-Bi-LSTM, 1-CNN 3-Bi-LSTM, and the Albert-base-v2 and Distilbert-base-uncased transformer models. Several performance metrics were considered in this research, including accuracy, precision, recall, F1 score, and precision/recall curve. Besides, a few techniques were performed to enhance the performance and avoid overfitting: batch normalization and tokenization of text, labels to numerical categories, and padding/truncation of sequences. Batch normalization prevents problems in training and speeds it up since the distributions of input values are made stable. Text tokenization allows for the conversion of words into numerical tokens that can be further processed. It also changes labels to numerical categories for model training and allows padding/truncation of sequences, making the dimensionality uniform for optimizing model performance. Therefore, this comparative study of the various models introduced metrics that showed rich insights into their efficiencies and robustness in detecting fake news. A comparative case study like this will make informed decisions easier to take regarding model selection and refinement, thereby helping in news identification and fighting misinformation at social media sites. The recall and precision curve gives the capability for evaluation in terms of how much each class is effectively predicted while building multi-class classification models.

Precision/recall curves were used for each of the 12 classes in the dataset to evaluate the classification performance of five different models: 2-CNN 2-Bi-LSTM, 3-CNN 1-Bi-LSTM, 1-CNN 3-Bi-LSTM architectures, and Albert-base-v2 and distilBert-base-uncased transformer models. Fig. 7 depicts, for every class, the precision/recall curve in the model DistilBert-base-uncased. It is seen that the model performs fine in most classes but shows Classes 2-unreliable and Class 6-fake have lower precision and recall values. Fig. 8 shows the precision/recall curve of the 1-CNN 3-Bi-LSTM model for each class. Most classes have worse performance as compared to the Distilbert-base-uncased model, which ascertains the limitation in the modeling capability of the model concerning the proper classification of certain classes. Similarly, Figure 9 shows the precision/recall curves for the 2-CNN 2-Bi-LSTM model. Compared to the DistilBert-base-uncased model, the performance of this model is poor, especially when considering classes whose precision and recall values are low. In contrast, the 3-CNN 1-Bi-LSTM model demonstrates improved performance, as shown in Fig. 10. While not significantly better than the DistilBert-base-uncased model, it outperforms the 2-CNN 2-Bi-LSTM and 1-CNN 3-Bi-LSTM models for most classes. Fig. 11 depicts the precision/recall curve for each class in the Albert-base-v2 model. Despite lower precision in class 3, the overall performance of the model is better than that of the other models, particularly for the entire class set.

Overall, the precision/recall curves provide detailed insights into the classification performance of each model for the 12 classes, highlighting strengths and weaknesses that can guide further model refinement and selection.

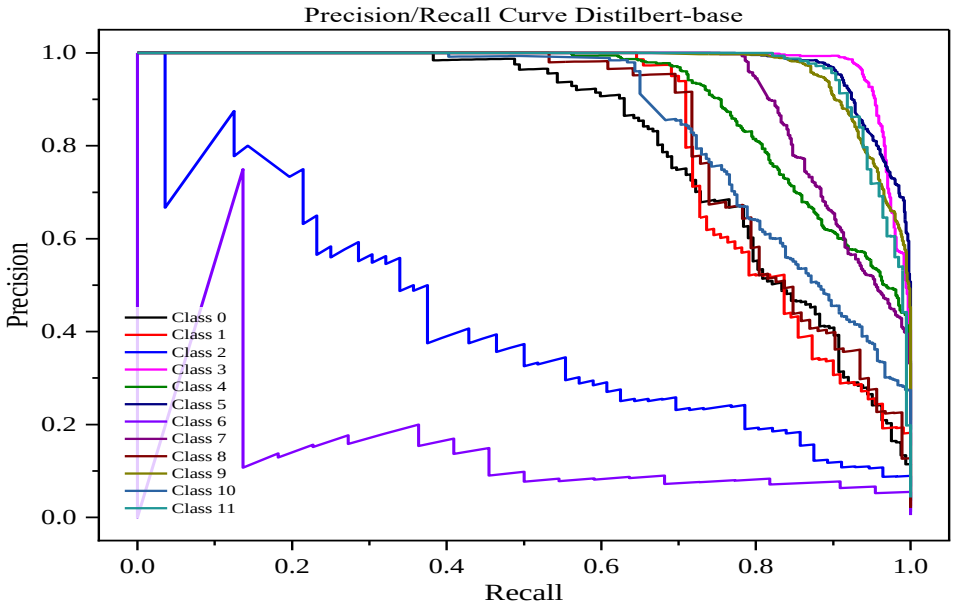


Figure 7: The precision/recall curve is distilbert-base-uncased for every class

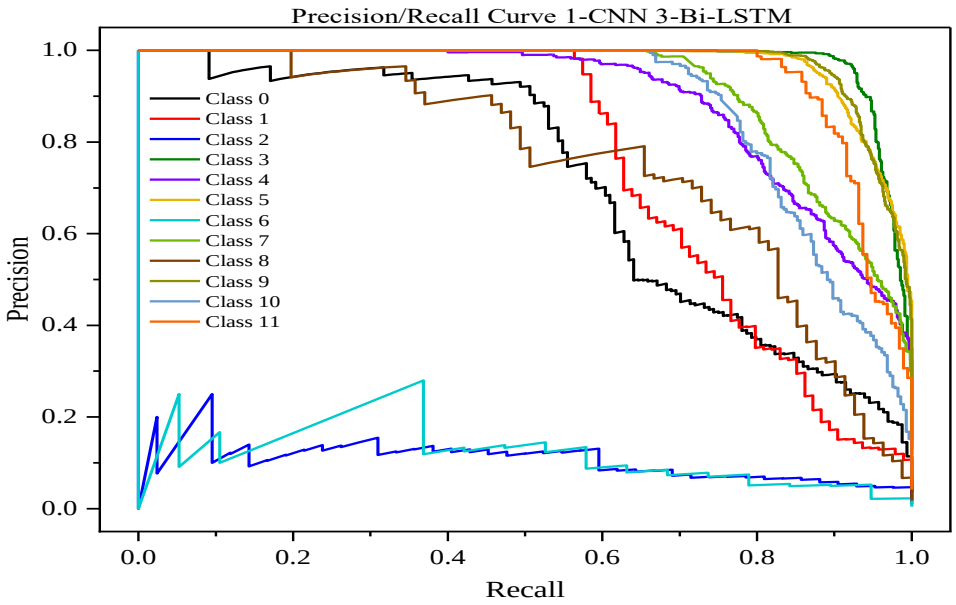


Figure 8: The precision/recall curve 1-CNN 3-Bi-LSTM of every class

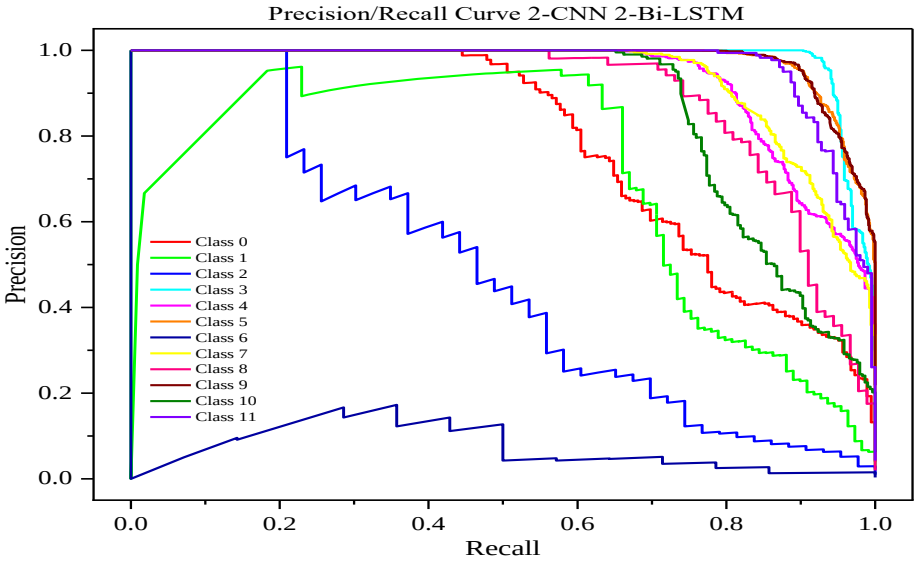


Figure 9: The precision/recall curve 2-CNN 2-Bi-LSTM of every class

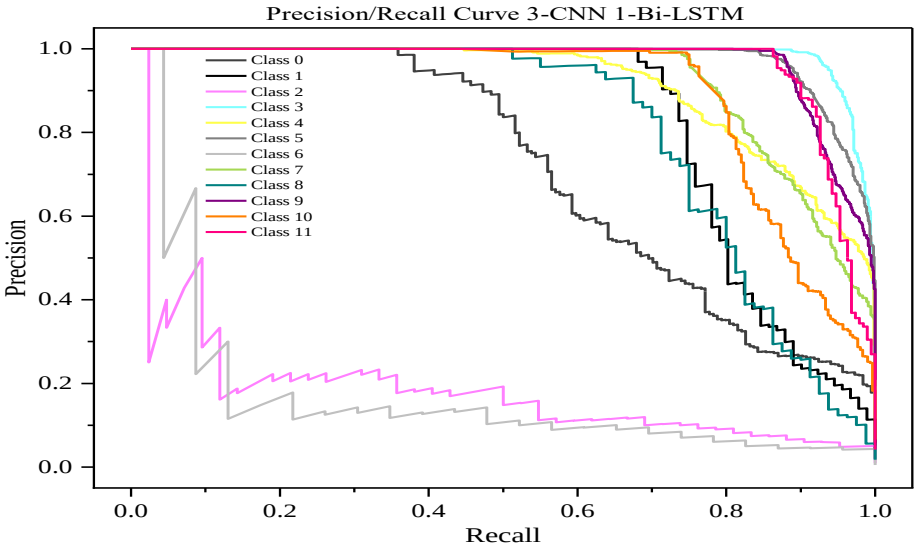


Figure 10: The precision/recall curve 3-CNN 1-Bi-LSTM of every class

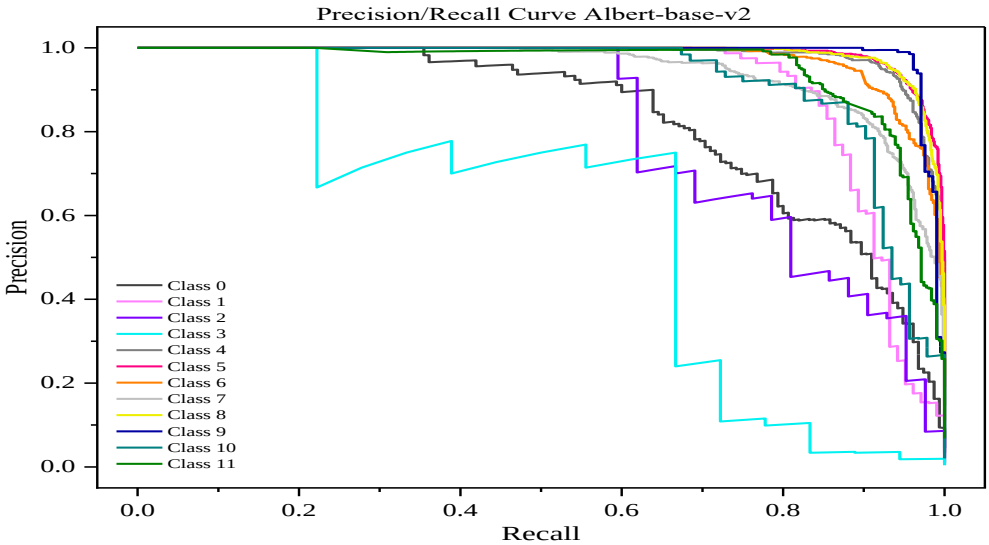


Figure 11: The precision/recall curve Albert-base-v2 of every class

Albert-base-v2 classification performance was also assessed using per-class evaluation metrics, as indicated in Fig. 12. The result indicates that the model achieves high precision, recall, and F1-scores in all classes except one, leading to 91% overall accuracy. The weighted F1-score of 0.92 indicates the strength of the model in fake news classification. Interestingly, a low recall of 0.50 among certain minority classes, i.e., Class 11, maybe a sign of dataset imbalance. Albert-base-v2's better-classifying abilities compared to traditional CNN-BiLSTM and DistilBERT models are supported by the macro-average F1-score value of 0.88, confirming its effectiveness.

	precision	recall	f1-score	support
0	0.91	0.93	0.92	2061
1	0.94	0.91	0.92	2810
2	0.84	0.97	0.90	4135
3	0.98	0.95	0.96	2227
4	1.00	0.91	0.95	4748
5	0.95	0.93	0.92	320
6	0.95	0.89	0.92	802
7	1.00	0.93	0.94	1185
8	0.84	0.84	0.84	954
9	0.92	0.76	0.83	310
10	0.77	0.89	0.83	603
11	0.73	0.50	0.59	165
accuracy			0.91	20320
macro avg	0.90	0.87	0.88	20320
weighted avg	0.93	0.92	0.92	20320

Figure 12: Per-class performance metrics for albert-base-v2

The classifying discriminative capacity of the Albert-base-v2 model was also examined based on ROC curves, presented in Fig. 13. The ROC curve offers a comprehensive assessment of the model's performance in differentiating between real and false news across different classes. The area under the curve (AUC) measurements of most classes were close to 1.00, which suggests outstanding discriminatory effectiveness. The findings confirm that Albert-base-v2 successfully reduces the incidence of false positives without sacrificing high true positive rates, thereby exhibiting its advanced capability in dealing with intricate fake news classification issues. Additionally, the high AUC scores for all classes further strengthen the model's superiority in comparison to CNN-BiLSTM and DistilBERT-based models.

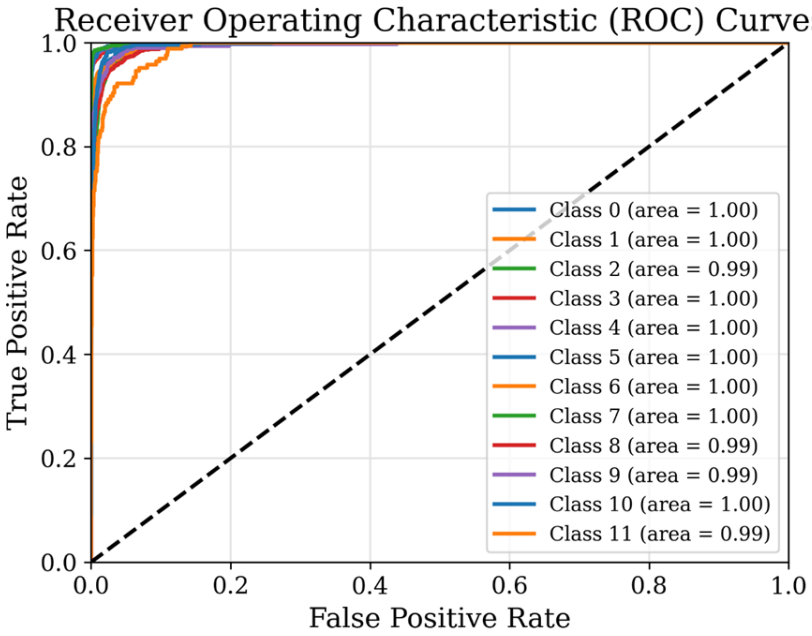


Figure 13: Receiver operating characteristic (ROC) Curves for albert-base-v2

Fig. 14 and Table 3provide a detailed comparison of the performance metrics for the 2-CNN 2-Bi-LSTM, 3-CNN 1-Bi-LSTM, 1-CNN 3-Bi-LSTM, Distilbert-base-uncased, and Albert-base-v2 models. The evaluation metrics considered in this comparison include accuracy, precision, recall, and F1-score. Analyzing the results reveals several key insights. The 1-CNN 3-Bi-LSTM model demonstrates lower performance compared to the other models in detecting certain classes, although it still shows reasonable overall performance. The 3-CNN 1-Bi-LSTM model exhibits a slight performance improvement over the 1-CNN 3-Bi-LSTM model. The 2-CNN 2-Bi-LSTM model achieves an F1-score of 0.8612 across all metrics, indicating a balanced performance across all classes. The Albert-base-v2 model outperforms all other models, achieving an F1-score of 0.9080 (Table 3).

This superior performance underscores the effectiveness of the Albert-base-v2 model in accurately classifying the Fake News Sample dataset. It demonstrates robust performance across all metrics, indicating its suitability for handling a variety of news detection tasks. In summary, the results show that the Albert-base-v2 transformer model is more effective than both the CNN and Bi-LSTM-based models, as well as the Distilbert-base-uncased model, in accurately classifying the Fake News Sample dataset. This highlights the efficacy of transformer models in handling news detection tasks, particularly when dealing with complex datasets containing multiple classes.

One of the primary concerns in model selection was a trade-off between classification accuracy and computational cost. As illustrated in Table 2, CNN-BiLSTM models are lighter computationally, requiring less resources for prediction and training, but are worse performing than transformers. DistilBERT is a cost-effective alternative, offering 40% fewer parameters than BERT but still an F1-score of 0.85. Albert-base-v2, being more computationally demanding than DistilBERT, achieves better classification accuracy (F1 = 0.908) with a good trade-off in efficiency through its parameter-sharing strategy. This underscores the need to balance accuracy needs with computational resources available in choosing models for detecting fake news.

Table 2: Performance vs. computational cost trade-offs

Model	Accuracy	F1-Score	Training Time (per epoch)	Model Size (Params)
2-CNN 2-BiLSTM	85.2%	0.854	12 min	2.1M
3-CNN 1-BiLSTM	85.8%	0.857	14 min	2.4M
1-CNN 3-BiLSTM	86.1%	0.861	16 min	2.8M
DistilBERT-base	85.0%	0.850	32 min	66M

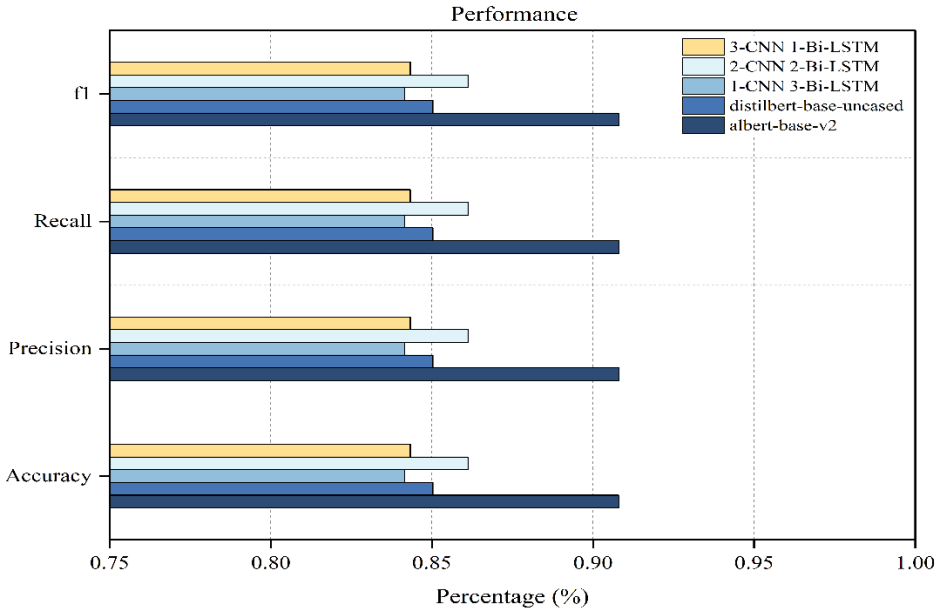


Figure 14: The performance obtained from albert-base-v2, 2-CNN 2-Bi-LSTM, disitilbert-base-uncased, 3-CNN 1-Bi-LSTM, and 1-CNN 3-Bi-LSTM classification models

Table 3: The performance value obtained from albert-base-v2, 2-CNN 2-Bi-LSTM, disitilbert-base-uncased, 3-CNN 1-Bi-LSTM, and 1-CNN 3-Bi-LSTM classification models

	Albert-base-v2	2-CNN 2-Bi-LSTM	Distilbert-base-uncased	3-CNN 1-Bi-LSTM	1-CNN 3-Bi-LSTM
accuracy	0.9080	0.8612	0.8503	0.8433	0.8416
precision	0.9080	0.8612	0.8503	0.8433	0.8416
recall	0.9080	0.8612	0.8503	0.8433	0.8416
f1	0.9080	0.8612	0.8503	0.8433	0.8416

Table 4 presents the results of t-tests conducted to compare the performance of the Albert model with several other models. The objective of this analysis was to assess the statistical significance of performance differences. As the Albert model has been identified as the superior model in this study, pair-wise comparisons were performed between Albert and each of the other models (including 2-CNN 2-BiLSTM, DistilBERT, 3-CNN 1-BiLSTM, and 1-CNN 3-BiLSTM). The P-value and T-Statistic values are provided for each comparison, respectively indicating the probability of observing the obtained results under the null hypothesis (no significant difference) and the magnitude of the difference between the groups. The very small P-values (less than 0.05) suggest the presence of statistically significant differences between the performance of the Albert model and the other compared models.

Table 4: T-test Results for statistical significance in model performance comparisons

Comparison	T-Statistic	P-value
Albert vs 2-CNN 2-BiLSTM	7.13	0.0017
Albert vs DistilBERT	8.21	0.001
Albert vs 3-CNN 1-BiLSTM	9.45	0.0006
Albert vs 1-CNN 3-BiLSTM	10.12	0.0004

4 Discussion

The performance outcome based on this work is compared with those of other works, which are tabulated in Table 3. The Albert-base-v2 transformer model proposed was more efficient than other deep learning architectures, with an F1-score of 0.908 and an accuracy rate of 90.8%. The findings suggest a remarkable improvement over CNN-BiLSTM-based architectures as well as other studies using conventional deep learning techniques.

4.1 Comparison with related works

Several reasons were accountable for Albert-base-v2's superior performance over previous methods:

Transformer-based advantages: Unlike CNN and BiLSTM-based models, Albert-base-v2 benefits from parameter-sharing mechanisms and deeper contextualization, which improves fake news classification performance. While Albert-base-v2 has good classification performance, its black-boxing is concerning when it comes to transparency of decision-making. Fake news detection models must not just be very accurate but also explainable so that trust and accountability can be ensured. If not interpretable, misclassification without cause is feasible, and this can lead to content moderation bias or censorship of factual information.

Dataset features: The Fake News Sample-Pontes dataset utilized in this research comprises a variety of news sources, thereby enhancing the generalization capability of the model compared to those utilized in earlier research works. For instance, the BerConvoNet [7] and TI-CNN [10] models were trained on domain-specific datasets, compromising their generalization capability.

Computational Efficiency: While being lightweight, the DistilBERT architecture had slightly reduced accuracy (85.0%) than Albert-base-v2 (90.8%), indicating the natural trade-off between computational expense and classification accuracy.

4.2 Analysis of performance differences

Performance differences across various architectures can be explained by:

Model complexity: While CNN-BiLSTM models had reasonable performance, their sequential nature hindered their capacity to effectively capture long-term dependencies in comparison with transformers.

Feature extraction techniques: The TI-CNN model [10] leveraged both text and image data, with impressive performance in multimodal settings. Its computational expenses and training intricacies, however, presented difficulties in comparison to Albert-based methods.

Generalization capacity: Conventional machine learning algorithms [11] depended on linguistic analysis but were inflexible when it came to generalizing across different news styles. The transformer-based models considered in this research demonstrated enhanced robustness across news types.

4.3 Comparison with state-of-the-art models

In contrast to existing state-of-the-art (SOTA) models, the present work contributes in the following ways:

Better interpretability: Most SOTA deep learning models are black boxes, making it challenging to explain their decision-making. By leveraging class-wise confusion matrices and precision-recall curves, our method renders model output more interpretable.

Balanced performance: A few recent state-of-the-art models, such as WSCH-CNN [6], have reported high classification accuracy at the expense of computationally expensive designs. However, our design preserves high performance while keeping the model's scalability intact.

Robust dataset preprocessing: In contrast to the prior work relying on manually designed features, the present method has utilized automated feature extraction, facilitating greater adaptability on various fake news datasets.

4.4 Limitations and future directions

Apart from the advantages of the suggested approach, it also suffers from the following limitations:

Real-world application constraints: Even though the model is effective at processing benchmark datasets, its applicability in real-time fake news detection is still to be broadly tackled. Analysis of the dataset revealed a skew toward shorter sentences and a strong concentration of articles from high-reputation domains such as 'nytimes.com.' While the dataset provides valuable insights into news classification, this skew may introduce bias in model predictions. Shorter articles, which are more common in the dataset, may be classified with higher accuracy than longer, more complex texts. Similarly, domain representation biases could lead to classification patterns that generalize poorly to news sources not included in the dataset.

Analysis of the dataset revealed a skew toward shorter sentences and a strong concentration of articles from high-reputation domains such as 'nytimes.com.' While the dataset provides valuable insights into news classification, this skew may introduce bias in model predictions. Shorter articles, which are more common in the dataset, may be classified with higher accuracy than longer, more complex texts. Similarly, domain representation biases could lead to classification patterns that generalize poorly to news sources not included in the dataset. Future work should explore more diverse, balanced datasets to mitigate these effects.

Future work should explore more diverse, balanced datasets to mitigate these effects.

Training bias: Even though dataset preprocessing techniques have been utilized, data bias sources in a dataset may still influence predictions. Debiased methods may be investigated in future work.

Challenges in explainability: While confusion matrices and feature visualization are part of the interpretative process, more work is needed to include explainable AI (XAI) methods to enhance model transparency.

5 Implications and future directions

The increasing reliance on AI-driven fake news detection is an ethical issue because of bias in model decision-making and dataset labeling. Fake news datasets are often manually labeled and thus prone to interpretation that may be attuned to political, cultural, or ideological leanings. Annotators' worldviews can bias decisions and create imbalanced training data that undermines model fairness. Moreover, deep learning models like Albert-base-v2 are trained on data distributions, and thus any inherent bias in the data set will be present in predictions as well.

While text-based fake news classification is the subject of this research, multi-modal misinformation (text + images + videos) is a new threat. Future research will require the inclusion of multi-modal approaches, combining text-based transformers (Albert-base-v2) with ViTs for image processing and GNNs for network-based misinformation dynamics. Such multimodal fusion can potentially enhance detection performance, particularly for social media-based fake news, since textual and visual information are often manipulated together.

Biased training data can make models favor certain narratives systemically, leading to unwanted biases in real-world classification. The inherent black-box property of transformer models renders the explanations for the flagging of news articles difficult to interpret, highlighting the need for explainability in the direction of transparency and fairness. Furthermore, algorithmic detection mechanisms can inadvertently suppress genuine content through false positives, thereby further exacerbating censorship concerns, especially in politically charged contexts. To counter these potential risks, there is a need to integrate fairness assessment techniques, implement explainability techniques like SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) to interpret the decision-making steps, and have human-in-the-loop validation to ensure the ethical use of artificial intelligence.

While this study focuses on a comparative analysis of established NLP models for fake news detection, future research could explore the integration of multi-modal features such as visual content and social context to enrich model understanding. To address interpretability concerns, future work needs to include Explainable AI (XAI) methods to make model decision-making processes more transparent. Techniques such as SHAP and LIME can be used to identify which words or phrases influence classification decisions. Further, attention visualizations in transformer models can highlight significant contextual patterns used in fake news classification, enhancing transparency and trust in automated detection systems.

Notwithstanding these ethical concerns, the suggested approach has immense practical applications. AI-driven fake news detection can be used in social media networks like Twitter, Facebook, and YouTube to aid fact-checking and limit the propagation of disinformation. These models can be utilized to facilitate source validation to ensure journalistic integrity, while disinformation mitigation and corporate reputation management can be enhanced in both government and corporate environments. There are some limitations, however, specifically regarding the ability to adapt to the dynamic nature of misinformation. Since fake news patterns evolve with time, models trained on previous data may be unable to recognize new deception tactics, and hence continuous retraining with new datasets is required. One challenge of particular concern is the detection of misinformation that has been generated by artificial intelligence. A study by Moalla et al. (2023) suggests that it is increasingly difficult to differentiate between AI-generated fake news and human-written articles. As synthetic misinformation has not been explicitly addressed in the current approach, it is essential to seek further investigation in this area. Moreover, model interpretability is still a limitation since deep learning models are black boxes, thus making decisions hard to justify in high-stakes settings. There may also be legal and regulatory challenges, especially when AI-driven fake news detection is applied to automated content moderation, which may create tensions in terms of censorship and freedom of speech.

To address these challenges, upcoming research must be directed towards creating models that can identify artificial intelligence-generated misinformation, integrating explainability methods to enhance transparency, and increasing detection on multiple languages and online platforms. Misinformation tracing studies on different platforms must also receive priority consideration to detect coordinated disinformation campaigns, while advancements in real-time deployment strategy must be considered for increasing practical usage. By overcoming these challenges, AI-based fake news detection can be made more efficient and ethical so that serious misinformation is responsibly countered in a world that is becoming more digital.

6 Conclusion

The research work presented a wide review concerning the different models of neural networks for detecting fake news based on the metrics of accuracy, precision, recall, and F1-score. In this work, the project used Fake News Sample databases to properly identify news. It discussed the challenges and capacities of different models: 2-CNN 2-Bi-LSTM, 3-CNN 1-Bi-LSTM, and 1 CNN 3-Bi-LSTM, other than Albert-base-v2 and Distilbert-base-uncased transformer models. This rigorous method returned handy information on model performance features.

The testing results proved that Albert-base-v2 outperformed the transformer model 2-CNN 2-Bi-LSTM, 3-CNN 1-Bi-LSTM models, and Distilbert-base-uncased transformer model. Specifically, Albert-base-v2 had the best F1-score of 0.908, showing that the model is robust in classifying fake news articles. Though there could be some variation in the class performances, the overall performance turned out to be very strong as Albert-base-v2 handled the complication in the Fake News Sample dataset quite well.

These further emphasize the reasons for using transformer-based models, such as Albert-base-v2, inefficiently and correctly detecting fake news. The results of this study provide valuable contributions to the field of news identification by underlining the importance of advanced neural network models for better classification accuracy and, subsequently, combating misinformation in news articles. Extensive assessment criteria, such as accuracy, precision, recall, F1 score, and AUC-PR analysis, have been used in the paper for comparative performance evaluation. This helped participants in model selection and refining decisions that best suited their personality prediction and pattern recognition. The work contributes to knowledge by offering an extensive

comparison between various models and their applicability in identifying news.

Abbreviation

Nomenclature		Greek letters	
Abbreviations		Latin Symbols	
NLP	Natural Language Processing	z_i	the input to the convolution layer
NEB	News Embedding Block	C_i	a reference point
MSFB	Multi-Scale Feature Block	×	the connection between the matrix and the input
CNN	Convolutional Neural Network	$F(.)$	the activation function
AC-BiLSTM	attention-based convolutional bidirectional long short-term memory	b_i	the bias vector
TM	Tsetlin Machine	φ	the max-pooling layer's final result
WSCH-CNN	Web Scraping Content Heading CNN	$\Delta(.)$	the max-pooling splitting function
MPFN	Multimodal Progressive Fusion Network	y_i	the final output vector
LSTM	Long Short-Term Memory	τ_i	the value of the weight matrix
BERT	Bidirectional Encoder Representations from Transformers	γ_i	the bias
TP	True Positive	σ	the cell model includes a sigmoid function
FN	False Negative	x_i	the current input
FP	False Positive	h_{i-1}	the previous state
TN	True Negative	c_{i-1}	the current cell state
XAI	Explainable AI	x_t	word vector
SHAP	SHapley Additive Explanations	I_t, F_t, and O_t	input, forget, and output gate vectors
LIME	Local Interpretable Model-agnostic Explanations	W and b	cell parameters
LRP	Layer-wise Relevance Propagation	$\overrightarrow{h_t}$ and $\overleftarrow{h_t}$	the outputs of the forward and backward passes
		Pos	the position encoding

Competing interests

The authors declare no competing interests.

Authorship contribution statement

Chi Zhang: Writing-Original draft preparation, Conceptualization, Supervision, Project administration.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Author Statement

All the authors have read and approved the manuscript. As stated earlier in this document, the requirements for authorship have been met, and each author believes that the manuscript represents honest work.

Ethical Approval

All authors have been personally and actively involved in substantial work leading to the paper and will take public responsibility for its content.

References

[1] A. Hassan and H. Rashid, "Estimation of Ultimate Bearing Capacity in rock-socketed piles using optimized machine learning approaches," *Advances in Engineering and Intelligence Systems*, vol. 2, no. 04, pp. 114–127, 2023.

[2] S. Mastan Val, "Improving COPD Readmission Prediction with Optimized Machine Learning," *Journal of Artificial Intelligence and System Modelling*, vol. 1, no. 03, pp. 86–103, 2024.

[3] T. E. Trueman, A. Kumar, P. Narayanasamy, and J. Vidya, "Attention-based C-BiLSTM for fake

news detection," *Appl Soft Comput*, vol. 110, p. 107600, 2021.

[4] M. J. Awan *et al.*, "Fake news data exploration and analytics," *Electronics (Basel)*, vol. 10, no. 19, p. 2326, 2021.

[5] Y. Yang, L. Zheng, J. Zhang, Q. Cui, Z. Li, and P. S. Yu, "TI-CNN: Convolutional neural networks for fake news detection," *arXiv preprint arXiv:1806.00749*, 2018.

[6] D. K. Vishwakarma, P. Meel, A. Yadav, and K. Singh, "A framework of fake news detection on web platform using ConvNet," *Soc Netw Anal Min*, vol. 13, no. 1, p. 24, 2023.

[7] M. Choudhary, S. S. Chouhan, E. S. Pilli, and S. K. Vipparthi, "BerConvoNet: A deep learning framework for fake news classification," *Appl Soft Comput*, vol. 110, p. 107614, 2021.

[8] P. Meel and D. K. Vishwakarma, "A temporal ensembling based semi-supervised ConvNet for the detection of fake news articles," *Expert Syst Appl*, vol. 177, no. February, p. 115002, 2021, doi: 10.1016/j.eswa.2021.115002.

[9] V. Gunasekaran and D. Karthikeyan, "Real Time Fake news Detection Web App Enhanced by Machine Learning Algorithms," in *2023 4th International Conference on Intelligent Technologies (CONIT)*, IEEE, 2024, pp. 1–7.

[10] Y. Yang, L. Zheng, J. Zhang, Q. Cui, Z. Li, and P. S. Yu, "TI-CNN: Convolutional neural networks for fake news detection," *arXiv preprint arXiv:1806.00749*, 2018.

[11] J. Clarke, H. Chen, D. Du, and Y. J. Hu, "Fake news, investor attention, and market reaction," *Information Systems Research*, vol. 32, no. 1, pp. 35–52, 2020.

[12] P. Meel and D. K. Vishwakarma, "Fake news detection using semi-supervised graph convolutional network," *arXiv preprint arXiv:2109.13476*, 2021.

[13] B. Bhattarai, O.-C. Granmo, and L. Jiao, "Explainable tsetlin machine framework for fake news detection with credibility score assessment," *arXiv preprint arXiv:2105.09114*, 2021.

[14] O.-C. Granmo and L. Jiao, "Explainable Tsetlin Machine Framework for Fake News Detection with Credibility Score Assessment," 2022.

[15] M. J. Awan *et al.*, "Fake news data exploration and analytics," *Electronics (Basel)*, vol. 10, no. 19, p. 2326, 2021.

[16] T. E. Trueman, A. Kumar, P. Narayanasamy, and J. Vidya, "Attention-based C-BiLSTM for fake news detection," *Appl Soft Comput*, vol. 110, p. 107600, 2021.

[17] D. K. Vishwakarma, P. Meel, A. Yadav, and K. Singh, "A framework of fake news detection on web platform using ConvNet," *Soc Netw Anal Min*, vol. 13, no. 1, p. 24, 2023.

[18] N. Põldvere, Z. Uddin, and A. Thomas, "The PolitiFact-Oslo Corpus: A new dataset for fake news analysis and detection," *Information*, vol. 14, no. 12, p. 627, 2023.

[19] J. Jing, H. Wu, J. Sun, X. Fang, and H. Zhang, "Multimodal fake news detection via progressive fusion networks," *Inf Process Manag*, vol. 60, no. 1, p. 103120, 2023.

[20] H. Moalla, H. Abid, D. Sallami, E. Aïmeur, and B. Ben Hamed, "Exploring the Power of Dual Deep Learning for Fake News Detection," *Informatica*, vol. 48, no. 4, 2025.

[21] P. Meel and D. K. Vishwakarma, "A temporal ensembling based semi-supervised ConvNet for the detection of fake news articles," *Expert Syst Appl*, vol. 177, no. April, p. 115002, 2021, doi: 10.1016/j.eswa.2021.115002.

[22] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," pp. 2–6, 2019.

[23] B. Büyüköz, A. Hürriyetoğlu, and A. Özgür, "Analyzing ELMo and DistilBERT on Socio-political News Classification," *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*, no. May, pp. 9–18, 2020.

[24] M. Abadeer, "Assessment of DistilBERT performance on Named Entity Recognition task for the detection of Protected Health Information and medical concepts," no. 2019, pp. 158–167, 2020, doi: 10.18653/v1/2020.clinicalnlp-1.18.

[25] Y. Chen, J. Bin, and C. Kang, "Smart Agricultural Technology Application of machine vision and convolutional neural networks in discriminating tobacco leaf maturity on mobile devices," *Smart Agricultural Technology*, vol. 5, no. September, p. 100322, 2023, doi: 10.1016/j.atech.2023.100322.

[26] F. A. Alijoyo *et al.*, "Advanced hybrid CNN-Bi-LSTM model augmented with GA and FFO for enhanced cyclone intensity forecasting," *Alexandria Engineering Journal*, vol. 92, no. March, pp. 346–357, 2024, doi: 10.1016/j.aej.2024.02.062.

[27] J. Gu *et al.*, "Recent advances in convolutional neural networks," *Pattern Recognit*, vol. 77, pp. 354–377, 2018, doi: 10.1016/j.patcog.2017.10.013.

[28] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput*, vol. 9, no. 8, pp. 1735–1780, 1997.

[29] P. Zhou *et al.*, "Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification," pp. 207–212, 2016.

[30] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations," pp. 1–17, 2019.

[31] H. Wang, X. Hu, and H. Zhang, "Sentiment analysis of commodity reviews based on ALBERT-LSTM," in *Journal of Physics Conference Series*, AA(Guilin University of Electronic Technology, China), AB(Guilin University of Electronic Technology, China), AC(Guilin University of Electronic Technology,

China), Nov. 2020, p. 12022. doi: 10.1088/1742-6596/1651/1/012022.

[32] X. Wang, H. Wang, G. Zhao, Z. Liu, and H. Wu, "Albert over match-lstm network for intelligent questions classification in chinese," *Agronomy*, vol. 11, no. 8, 2021, doi: 10.3390/agronomy11081530.

[33] C. H. Chiang, S. F. Huang, and H. Y. Lee, "Pretrained language model embryology: The birth of ALBERT," *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, no. Section 3, pp. 6813–6828, 2020, doi: 10.18653/v1/2020.emnlp-main.553.

[34] R. Zhao, B. Vogel, and T. Ahmed, "Adaptive Loss Scaling for Mixed Precision Training," pp. 1–11, 2019.

[35] S. Alrowili and K. Vijay-Shanker, "BioM-Transformers: Building Large Biomedical Language Models with BERT, ALBERT and ELECTRA," *Proceedings of the 20th Workshop on Biomedical Language Processing, BioNLP 2021*, pp. 221–227, 2021, doi: 10.18653/v1/2021.bionlp-1.24.

[36] M. Hort, R. Moussa, and F. Sarro, "Multi-objective search for gender-fair and semantically correct word embeddings," *Appl Soft Comput*, vol. 133, p. 109916, 2023, doi: 10.1016/j.asoc.2022.109916.

[37] R. Rao *et al.*, "MSA Transformer," *bioRxiv*, p. 2021.02.12.430858, 2021.

[38] J. Graën, M. Bertamini, M. Volk, M. Cieliebak, D. Tuggener, and F. Benites, "Cutter–a universal multilingual tokenizer," in *CEUR Workshop Proceedings*, CEUR-WS, 2018, pp. 75–81.

Predicting Gestational Diabetes Mellitus Using Machine Learning Algorithms: A Comprehensive Review and Analysis

Huiqin Ren, Shuai Zhao*
School of Health Industry Management, University of Sanya, Sanya 572022, Hainan, China
E-mail: shuaizhao@sanyau.edu.cn
*Corresponding author

Keywords: Pregnancy diabetes, red deer algorithm, jellyfish search, rhizotomy optimization algorithm, random forest classification (RFC), machine learning

Received: April 7, 2025

Predicting pregnancy-related diabetes using machine learning (ML) provides a proactive healthcare strategy and permits identifying at-risk patients well in advance. This helps healthcare practitioners to consider early treatment options that can mitigate the risk of complications for mother and child. It examines large datasets for subtle patterns and underlying risk variables, providing better predictive accuracy. By managing gestational diabetes early, medical providers mitigate the risks of macrosomia, preterm birth, and maternal hypertension. Finally, the approach leads to more targeted and effective prenatal care, improving health outcomes and securing a better start for the baby and mother. Besides, predictions of diabetes in pregnancy are supposed to be carried out by using RF-a ML classification algorithm. Red Deer Algorithm (RDA), Jellyfish Search Optimizer (JSO), and Rhizotomy Optimization Algorithm (ROA) are utilized to increase the model's precision. This model has been selected to integrate with optimizers to enhance forecast accuracy. In diabetes circumstances, RFJS is observed to outperform other models, with a precision value of 0.9377, while the rank of the second best is from the RFRO model, having a precision of 0.932. The RFRD scheme has a precision of 0.9282, thus reflecting medium performance but higher than the RFC with its precision of 0.879. These results have provided significant insights into the current performance and potential of ML algorithm models in terms of prediction. The insight of such a result point out the direction of future research and further development in applying ML to predictive analytics in diverse fields.

Povzetek: Študija predlaga napovedovanje nosečniškega diabetesa z naključnim gozdom, okrepljenim z metahevrističnimi optimizatorji (RDA, JSO, ROA), za proaktivno in ciljno prenatalno oskrbo.

1 Introduction

Diabetes during pregnancy, also known as Gestational Diabetes Mellitus, is a disorder characterized by high blood sugar levels that appear or are diagnosed during pregnancy [1]. It is a significant health issue because of the possible negative impact on mother and fetal health [2]. Although the exact cause of GDM is unknown, it is considered a mixture of resistance to insulin, changes in hormone levels, and genetic predisposition [3]. Insulin resistance, one of the characteristics of GDM, occurs when the cells in the body become less responsive to insulin, a hormone that controls blood sugar levels [4]. This resistance requires increased insulin production to sustain normal blood sugar levels, a burden on the pancreas [5]. Pregnancy leads to physiological changes that physiologically increase insulin resistance due mainly to releasing hormones like lactogen from the human placenta and cortisol [6].

GDM is typically diagnosed between the 24th and 28th weeks of pregnancy with the use of glucose tolerance tests [7]. Although some women present with symptoms like polydipsia and polyuria, many cases of GDM are asymptomatic and are only diagnosed through screening testing [8]. Poorly managed or improperly treated GDM presents serious risks to both mother and child [9]. Maternal complications may involve preeclampsia, cesarean delivery, and the elevated risk of type 2 diabetes in later life [10]. The fetal complications include macrosomia (excess development of the fetus), trauma during delivery, hypoglycemia following birth, and elevated risk of obesity and diabetes later in life [11], [12], [13]. The management options for GDM primarily involve lifestyle changes in the form of dietary modifications and regular exercise to lower blood sugar levels [14]. When lifestyle modification alone is not practical, insulin therapy may be necessary to achieve appropriate glycemic levels [15]. In general, the management of GDM requires a multidisciplinary approach by obstetricians, endocrinologists, nutritionists, and other specialists to achieve optimum maternal and fetal outcomes [16]. Early recognition, prompt intervention, and suitable prenatal care can minimize the adverse effects of GDM and ensure the health and well-being of both mother and child [17].

The early detection of pregnancy-related diabetes, including GDM type 2, is clinically relevant as it can potentially reduce the serious complications associated with it for both the mother and child [18], [19].

Identification of a woman who may be at risk for GDM would facilitate early management and thus minimize complications from uncontrolled hyperglycemia during pregnancy [20]. Early prediction makes taking preventive interventions, including lifestyle changes and close blood glucose monitoring, more straightforward to improve maternal and fetal health outcomes [21]. Identifying high-risk people early in pregnancy allows healthcare practitioners to undertake therapies targeted at managing blood sugar levels and lowering the risk of GDM-related problems [22]. Furthermore, with early prediction, healthcare professionals can plan for prenatal GDM treatment through regular monitoring, special nutritional counseling, and possible pharmaceutical interventions if required [23]. Finally, early detection of pregnancy-related diabetes enables healthcare professionals to take initiative in controlling the disease to attain better outcomes regarding both mother and fetus and, consequently, diminish the overall burden of GDM on the healthcare system [24].

1.1. Objectives

Early diagnosis of pregnancy-related diabetes allows for early treatment and decreases the risk of adverse outcomes for mother and child. Earlier identification provides more personalized care programs, hence optimizing maternal glucose levels while reducing risks of complications such as macrosomia and birth trauma. A more sophisticated approach was decided upon to pursue the objective, utilizing the ML model of RFC, further enhanced through the incorporation of RDO, JSO, and ROA for predictive analytics. Predicting pregnancy-related diabetes using ML algorithms, such as the RFC, has several benefits. First, RFC is very good at handling complex, high-dimensional databases, which are very common in medical research; it can, therefore, effectively combine a wide range of predictors, such as demographic, clinical, and biochemical characteristics, to make reliable predictions. Besides, RFC resists overfitting, a common concern when sample sizes are small or the data is noisy. Its ensemble learning technique, which integrates many decision trees, reduces the likelihood of individual trees picking up noise in the data, leading to more trustworthy predictions.

1.2. Related works

Though there is an ongoing debate on the severity with which Gestational Diabetes Mellitus must be recognized and preserved, most doctors believe that diabetes found in early pregnancy has more serious effects and should be handled immediately after it is detected by Metzger et al. and Wyatt et al. [25], [26]. Ideally, this research should be conducted before pregnancy, as organogenesis is typically completed when pregnancy is detected. In this case, pregnancy wasn't determined till 9 weeks of pregnancy. A high blood sugar level during the first 8 weeks of pregnancy increases the chance of severe abnormalities by up to 15%, which is eight times greater than the risk reported in straightforward pregnancies (about 2%), according to Wyatt et al[27] Due to undetected underlying T2DM, the 2% risk estimate might continue to be an underestimate of the nondiabetic pregnant demographic. Luckily, 85 percent of hyperglycemic pregnancies don’t result in significant abnormalities, which surprises many of us. Severe deformities are likely to cause early miscarriages; hence, they are not included in statistics. According to older studies, the prevalence of severe abnormalities might reach 25%. This higher historical prevalence is most likely the outcome of fewer biologically managed blood sugars owing to the lack of monitoring items and inadequate selection of insulin Jovanovic, Druzin, & Peterson et al. [28], [29].

The objective of therapy for women who get pregnant with weakly managed diabetes is to check for significant defects, focus on the 85 percent likelihood of a usual child, and optimize blood sugar management to that moment ahead. It is feasible to reduce Women with diabetes who have type 1 have an A1C of below six percent before and during pregnancy, and the outcomes of their pregnancies are comparable with that of the PG with no diabetes population [27]. Most doctors nowadays advocate these ladies can maintain their babies in the belly till the end of the ninth month with no scheduled premature sections(cesarean) or early childbirth. However, this training is still contentious in the field of perinatology.

Blood sugar (glucose) management afterward in PG affects embryonic development, and subsequent glucose levels can decrease the growth of macros to fewer than ten percent [30], [31], [32] The avoided development of macrosomia and its short- and long-term effects become the primary focus of glucose control. The frequency of type 2 diabetes among women of reproductive age worldwide is elevating at a rapid pace. Flegal et al.; Shaw, Sicree, and Zimmet [33], [34]. As an outcome, the count of unexpected births or PGs that happen with insufficient sugar level regulation throughout the crucial organogenesis phase is expected to rise, in addition to an upsurge in significant abnormalities. It is appropriate to examine a paradigm change for diabetes discovery/analysis in early PG and guidelines for prophylactic examination to determine people's impaired tolerance for glucose before pregnancy.

The probability of DB in women increases with race, age, BMI, and history OF family[35], [36]. The goal of monitoring is to detect women with unfamiliar, already present T2DM whose blood glucose levels are high enough to endanger the baby's growth before they are typically examined. Although there are currently no studies that precisely answer this issue, historical statistics and scientific assessment imply that the women are more in danger: 1) Ladies with a GDM history or the delivery of a more significant than a nine-pound newborn; 2) women who weighed more than 9 pounds at birth or have an extended family history of T2DM; 3) Ladies who have had a diagnosis of a condition called acanthosis nigricans or pcos ovarian condition; 4) Ladies having a body mass index (BMI) higher than thirty kilograms per square meter; and 5) Ladies from social groups at high probability of T2DM, such as Native Southeast Asian, American, and Hispanic. 2% of enceinte ladies had undiscovered T2DM, 5% had GDM when diagnosed in

the subsequent month of pregnancy, and another 1.5% will be discovered in the third trimester. It is recommended that women at high risk be tested when they are diagnosed with pregnancy. Formal testing is based on A1C levels over 5.3% and/or Postprandial blood sugar levels greater than 120 mg/dl, even if just once (after a heavy meal of carbohydrates). If the initial test is negative, repeating A1C and postprandial glucose testing every 4 weeks until 34 weeks gestation is advised, with at least one formal oral glucose tolerance test (OGTT) at 24-28 weeks. If blood sugar tolerance is expected after 34 weeks, discontinuing testing is appropriate.

2 Dataset and methodology

2.1.Data gathering

In Fig. 1, the wide-ranging effects of diabetes are discussed, ranging from blood pressure and pregnancy to its overall impacts on health and well-being. By examining these areas, a better understanding of the complicated interaction between diabetes and these critical determinants is sought to shed light on the intricate interplay affecting the course of this chronic illness.

- In diabetes, insulin malfunction causes metabolic abnormalities such as hyperglycemia and dyslipidemia. In academic terms, insulin's absence or poor activity impairs glucose absorption, glycogen synthesis, and lipid metabolism. This causes metabolic changes, including enhanced gluconeogenesis and lipolysis, which worsens hyperglycemia and lipotoxicity. Insulin insufficiency also influences cellular signaling pathways, including gene expression, protein synthesis, and cell proliferation. The delicate interplay between insulin and metabolic homeostasis is clarified via academic research using cellular and molecular approaches, influencing therapeutic options for improving insulin sensitivity and glycemic control in diabetes care [37].
- Diabetes causes skin thickness changes due to glycation, collagen cross-linking, and increased extracellular matrix deposition. These alterations impair skin flexibility and wound healing, leading to problems including diabetic ulcers and infections. Determining the effect of diabetes on skin thickness requires histological studies, biomechanical analysis, and research into molecular mechanisms that regulate collagen metabolism. Such study elucidates the pathophysiological mechanisms behind diabetic skin alterations, allowing for the development of tailored therapies to reduce complications and enhance patient outcomes [38].

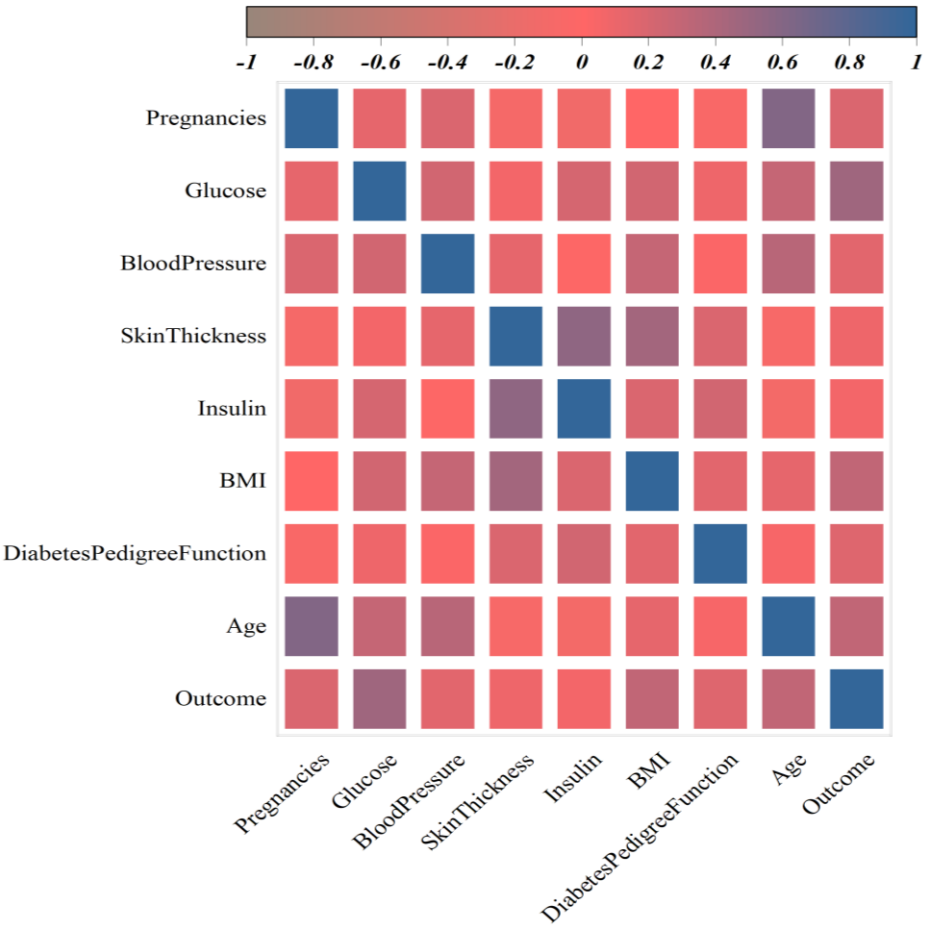


Figure 1: The correlation plot of the inputs and outputs variables

2.2.RFC

RFC entails many tree classifiers, all of which generate a classifier utilizing an unexpected vector collected separately from the input vector and offer a unit vote to the class that it believes is most likely to sort an input variable appropriatly [39]. RFC used in this investigation generates a tree by haphazardly choosing traits or mixtures of characteristics at each node. Bagging produces a training database that includes choosing every feature or characteristic mixture and haphazardly selecting N replacement samples, where N is the size of the original training set. Any occurrence pixels are sorted using the most frequently voted class among all tree predictors in the forest. Choosing characteristic measures and trimming strategies was critical for DT construction.

The attribute selection for DT initiation may be tackled in diverse ways, including direct quality measures. The two most popular attribute metrics used in decision tree induction are the skill gain ratio and the Gini index. The RFC picks attributes based on the Gini Index, which measures an attribute's impurity across classes. The Gini index for a specific training set T may be stated below: Select a single pixel haphazardly and indicate whether it belongs to a group.

$$the \sum\sum_{j\neq i}(f(c_i,T)/|T|)\big(f(c_j,T)/|T|\big) \quad (1)$$

$f\left(^{C_i},T\right)/|T|$ displays the likelihood that the chosen case falls within the group C_i.

2.3.Red deer algorithm (RDA)

Fig. 2 depicts the flowchart for the proposed RDA. Like other evolutionary algorithms, the proposed process begins with early Demographics (RD). A few of the most promising RDs in the general population were picked, such as male RDs; the remainder were known as hinds. Harems develop after fighting and roaring. A harem is a grouping of hinds. The population's female deer are shared among the aforementioned male RDs depending on their talents (fighting and screaming), power, and elegance. Male RD's elegance and strength are inversely related to their fitness score in GA. Roaring male RD is improving them and solutions in the area around them. When fighting, a male RD fights against another male RD. Two guys battling pick someone as the winner, corresponding to selecting the better value after fighting rather than the previous value for each solution in each stage. Soon after yelling and battling, the male RD divided the females between themselves [40]. The male leader of the harem is mating with a proportion of the hinds in his harem. And the male leader is constantly paired with a ratio of hinds in the same harem. Mating with the nearest hind, the hind closest to the male in RDA. This process serves as a counterpoint to developing new solutions, such as the offspring of RD throughout a generation. The RDA, like other meta-heuristic computations, consists of two phases: amplification and variation. During the intensification phase, two guys competing to be crowned champions means improving themselves. The male is coupling with the nearby hind legs in the surrounding communities. Screaming male RD haphazardly mated with a particular ratio of hinds in a harem in variety—intensity and diversity increase when the male RD in his herd mates with hinds.

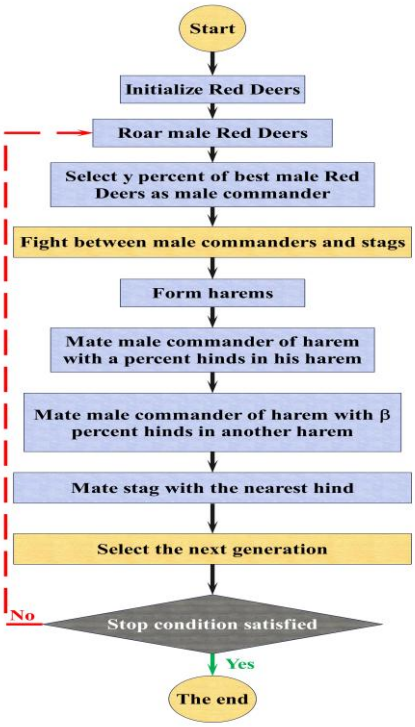


Figure 2: The flowchart of the RDA

2.3.1. Producing original red deer (RD)

Enhancing aims to determine the desired answer based on the problem's variables. An array of parameters to optimize is created. This array is known as a "chromosome" in GA nomenclature; however, it is referred to as "Red Deer" here. Thus, RD is the solution's equivalent. In an N_{var} –dimensional enhancement issue, an RD is a 1 N_{var} selection. The selection is structured as follows:

$$Red\ Deer = \left[X_1, X_2, X_3, \dots, X_{N_{var}}\right] \tag{2}$$

Furthermore, each RD's function value may be determined in the following way:

$$Value - f(Red\ Deer) = f\left(X_1, X_2, X_3, \dots, X_{N_{var}}\right) \tag{3}$$

To begin the optimization procedure, a starting group of size is formed to N_{pop} . The best RD is chosen for N_{male}, while the others are chosen for N_{hind}.

2.3.2. Roar male RDs

The male RD tries to boost their elegance at this time by yelling. RD males will replace older generations if they do better in objective functions (OF). The truth is that every male RD may change his position. Screaming male RD attracts ladies.

2.3.3. Select γ percent of best male RDs as male commanders

There is an enormous distinction among male RDs. Some achieve more tremendous success than others. In truth, men do not all have the same status in nature, with some seizing harems.

Male RD are classified into two categories: male leaders and stags. The count of leader males is associated with γ, which is:

$$N.male.com = round\{\gamma \cdot N_{male}\} \tag{4}$$

where $N.male.com$ is a large number of guys grabbing the harems. The male RDs are chosen as the best, while the rest are called "stags." The quantity of stags is calculated using the following method:

$$N \cdot stag = N_{male} - N.male.com \tag{5}$$

where $N \cdot stag$ is the stag number in male demographics.

2.3.4. Battle among male leaders and deer

Male commanders can combat stags at random. If the OF is better than the others, they are selected after a fight.

2.3.5. Formation of harems

A group of hinds taken by a male leader is called a harem. The strength of male leaders—whose abilities include fighting and roaring—determines how many hinds are kept in harems. To establish harems, hinds are proportionally shared among male leaders, and the normalized score of a male leader is determined as:

$$V_n = v_n - \max_i\{v_i\} \tag{6}$$

where the regularized value of the nth male leader is denoted by V_n, and v_n is its value. Every male leader's standardized power is defined as follows, taking into account the normalized value of all male commanders:

$$P_n = \left|\frac{V_n}{\sum_{i=1}^{N.male.com} V_i}\right| \tag{7}$$

From another perspective, a male leader's normalized authority is the proportion of hinds he would have. The total number of hind legs in a harem will be:

$$N \cdot harem_n = rand\{P_n \cdot N_{hind}\} \tag{8}$$

where $N \cdot harem_n$ displays the count of hinds of nth harem and N_{hind} displays the count of all hinds. To distribute the hinds throughout each male leader, a random pick is made Give $N \cdot harem_n$ hinds to it. These roe deer, together with the male, will form the nth female deer.

2.3.6. Mate male commander of harem with percent hinds in his harem

In stage E, the mating behavior begins. GA displays this fact as a "crossover" model. The male leader and the hinds in his harem is the parents. Their children represent the novel explanations. The correlation between the count of hinds in a female deer mating with their male leader ∝ is:

$$N \cdot {harem_n}^{mate} = round\{\propto \cdot N \cdot harem_n\} \tag{9}$$

With $N \cdot {harem_n}^{mate}$ representing the count of hinds from the nth harem ready to mate with these male RD. A random selection is made $N \cdot {harem_n}^{mate}$ of the $N\ .\ harem_n$.

2.3.7. Mate male commander of harem with β percent hinds in another harem

It is allowed for the male leader to mate with a portion of the hinds in a haphazardly selected harem. To increase his domain, the male RD assaults another harem. When there is only one male RD in a harem coupling with roe deer, the numbers are:

$$N \cdot {harem_n}^{mate} = round\{\beta.N.harem_n\} \tag{10}$$

With $N \cdot harem_n{}^{mate}$ representing the count of hinds from the nth harem ready to mate with these male RD. A random selection is made $N \cdot harem_n{}^{mate}$ of the $N.harem_n$, too.

2.3.8. Mate stag with the nearest hind

In this phase, each stag mates with the nearest hind. During the breeding cycle, male RD chooses to follow the convenient hind, which may be his favorite of all hinds. This roe deer might be within his harem or frequently different female deer. Every stag could come together with the closest roe deer. Thanks to this possibility, every male RD can mate with as few hinds as possible; in the worst situation, just one hind will mate. Distances between each stag and every hind must be calculated to determine the closest hind. Two dimensions are used to describe the method. The following formula is used to determine the separation in J-dimension space between a male RD and every hind:

$$d_i = \left(\sum_{j \in J} \left(stag_j - hind_j^i \right)^2 \right)^{1/2} \quad (11)$$

2.4.Jellyfish search optimization (JSO)

It is one of the most recent swarm-based metaheuristics, which Chou and Truong created in 2021 [41]. The JS method mimics the jellyfish's food-searching behavior in the water [42]. The JS method follows three simplified guidelines:

2.4.1. Ocean current

Jellyfish can sense the movement of the waves Eq. (12) and feed on tiny planktonic creatures.

$$\vec{O} = X' - \beta \times M \times r(0,1) \quad (12)$$

Here, $\vec{O}$ is ocean flow's direction, β ($\beta > 0$) specifies the length dispersal factor of $\vec{O}$, X' is the position of the present best jellyfish in the swarm, and M is the average location of all jellyfish. The new situation of each jellyfish is specified in the following order:

$$X_i(t+1) = X_i(t) + r(0,1) \times \vec{O} \quad (13)$$

Upon changing each jellyfish's status, a desired location, such as one with more food sources, is chosen as the jellyfish's present spot.

2.4.2. Jellyfish bloom

Inside a jellyfish bloom, jellyfish move passively and actively. The computational representations of these movements are shown here.

$$\text{Passive motion:} \quad X_i(t+1) = X_i(t) + \lambda \times r(0,1) \times (w_b - L_b) \quad (14)$$

λ ($\lambda > 0$) displays a factor relating to the duration of inactive motion. $w_b and\ L_b$ are the bottom and higher limits of the search area, correspondingly.

$$\text{Active motion:} \quad X_i(t+1) = X_i(t) + r(0,1) \times \vec{D} \quad (15)$$

where

$$\vec{D} = \begin{cases} X_i(t) - X_j(t) & if \quad g(X_i) < g(X_j) \\ X_j(t) - X_i(t) & if \quad g(X_i) \geq g(X_j) \end{cases} \quad (16)$$

Here, two functions $g(X_i) and\ g(X_j)$ are the jellyfish i and j have different OF values.

2.4.3. Time control mechanism

A temporal control system governs the two sorts of jellyfish's movement within the bloom and their migrations toward ocean currents. The time control function is expressed as follows:

$$T(t) = \left| \left(1 - \frac{t}{MaxIter}\right) \times (2 \times r(0,1) - 1 \right| \quad (17)$$

t displays the time index provided as the cycle number, and $MaxIter$ indicates the repetition most significant number.

2.4.4. Population initialization

This optimizer employs the logistic map to construct the starting population.

$$X_{i+1} = \upsilon X_i(1 - X_i), \quad 0 \leq X_0 \leq 1 \quad (18)$$

where X_i and X_0 are the chaotic values for the position of the $i-th$ Jellyfish and a haphazardly selected place, correspondingly. In every instance, υ is set to 4.

2.4.5. Boundary Handling Mechanism

When a jellyfish exceeds the confines of the specified search domain, it will be found inside them using Eq. (19).

$$\begin{cases} X'_{i,d} = (X_{i,d} - W_{b,d}) + L_{b,d} & if \quad X_{i,d} > W_{b,d} \\ X'_{i,d} = (X_{i,d} - L_{b,d}) + W_{b,d} & if \quad X_{i,d} < W_{b,d} \end{cases} \quad (19)$$

where $X_{i,d}$ and $X'_{i,d}$ are the present and updated position of the $d-th$ aspect of the $i-th$ jellyfish. $W_{b,d} and\ L_{b,d}$ are upper and Lower limits on the $d-th$ aspect of the search area, correspondingly. Fig. 3 displays the diaggram of JSO.

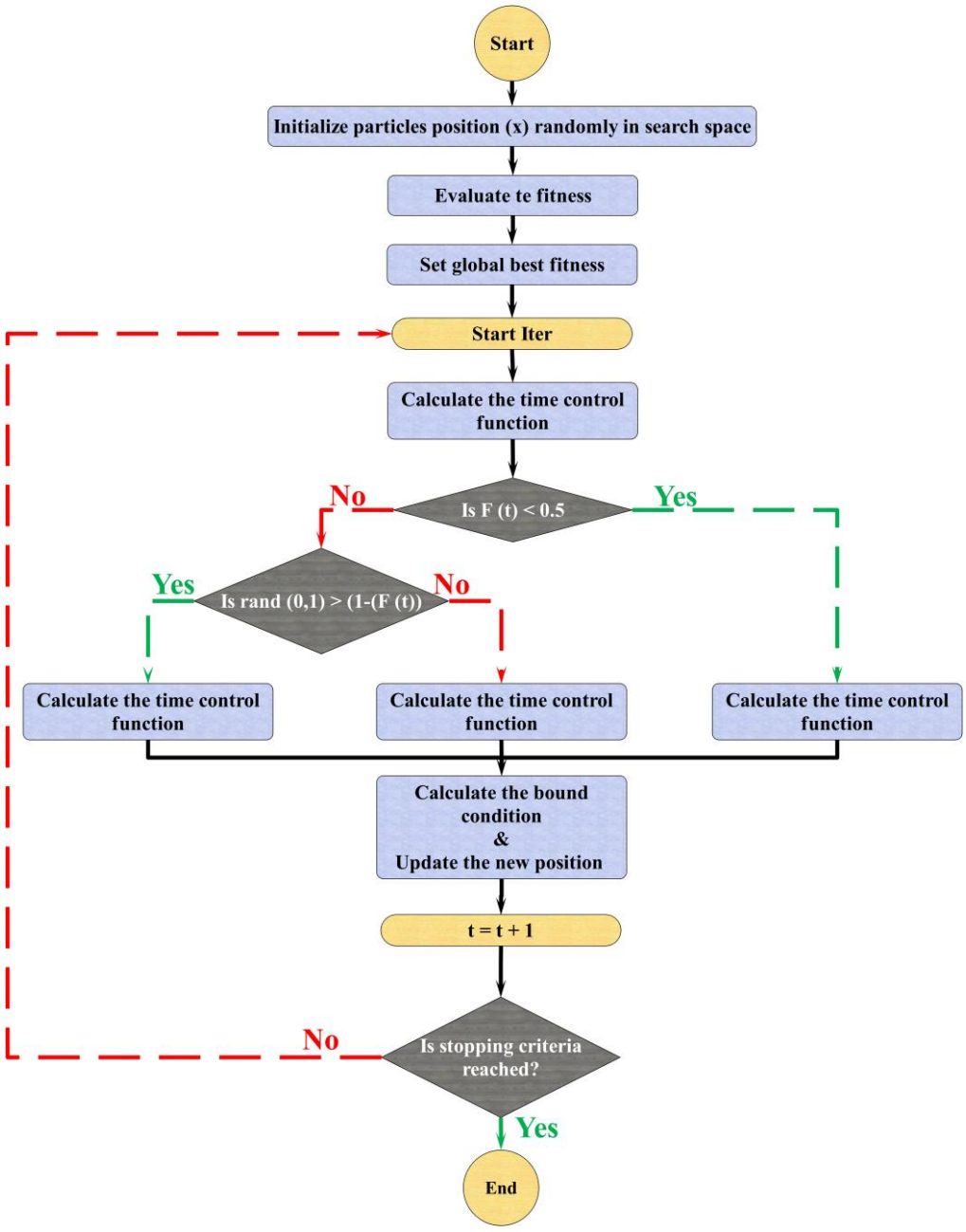


Figure 3: The diagram of JSO [41]

2.5. Rhizotomy optimization algorithm (ROA)

The suggested technique depends on three major rules. Bands of Exploration and Exploitation: A rhizotomy plume is a marine creature that employs multiple random search tactics to find the best areas to feed on large amounts of plankton [43]. When a swarm gathers, it begins to feed. The artificial ROA considers the two essential bands of a meta-heuristic approach. The swarm has two main motion types: exploitation and discovery; (ii) Rhizotomy goes or seeks for food, with a "motion control aspect" establishing when to switch between these actions; and (iii) the swarm's location and related function of objective identify the quality of food found at any given time [44].

2.5.1. Initialization of the population and setting boundary conditions

R, the octopus' gender is chosen at random.

$$x_i^* = \lambda x_i(1 - x_i), \qquad 0 \le x_0 \le 1 \tag{20}$$

The anticipated starting population of R. octopus, x_0 , goes to 0 and 1, $x_0 \notin \{0{,}0.25{,}0.75{,}1\}$, and feature λ is set to 4.0. The place value of the $i - th$ R. octopus is x_i. A R,

when an octopus enters the search zone, it moves in the other direction. Eq. (21) shows this comeback tactic.

$$x_{i,d}^{*} = \begin{cases} (x_{i,d} - Ub_d) + Lb_d, if\, x_{i,d} > Ub_d \\ (x_{i,d} - Lb_d) + Ub_d\,, if\, x_{i,d} < Lb_d \end{cases} \quad (21)$$

R is situated in the $d-th$ metric. Octopus is designated by $x_{i,d}$; the new place, $x_{i,d}^{*}$, is Generated by validating the boundary restrictions. The top and lower boundaries of the field of exploration for the $d-th$ dimension is represented by Ub_d and Lb_d, correspondingly.

2.5.2. Food searching strategies

Scientists speculated that R, the octopus, migrated utilizing LW like honeybees and sharks [45]. The writer observed that the octopus' mobility pattern changed with time. This type of pattern is best reproduced with fast simulated annealing (FSA). The generating function is embedded within all these food-searching patterns, governing the parameter-updating mechanism throughout the search process. This approach incorporates ROA and relies on three distinct random search strategies: LW, fast simulation of annealing (or food-seeking strategies), and SA. Data comparisons identify the optimal motion plan for the system. The first sortie, termed "Levy flight" after French mathematician Paul Levy [46], displays a random walk with variable step lengths. Moving in a multifaceted environment necessitates taking random steps in various directions.

where x_i is the R, the octopus's present perfect place, x_i is the new R, place of octopus. A step toward the objective, x_i will be shown by R. octopus at x_i if $f(x_i) > f(x_i)$ this phase concludes a heavy-tailed Levy journey, which may be expressed as a simple power law, as illustrated in Eq. (22).

$$L(s) \approx |s|^{1-\beta}, where\ 0 < \beta \leq 2 \quad (22)$$

s is the accidental parameter. The Levy step scale, where u and v originate from typical patterns of shipping, may be calculated as $s = {u}/{|v|^{1/\beta}}$. It is feasible to modify the position with the highest concern for food in Eq. (23):

$$x_{next} = x_i \times rand(0,1) \quad (23)$$

Unplanned motions or flights from the Levy distribution are depicted as $\mu \times \text{rand}(0,1)$, w is the ith random statistic is a widely spread random integer between 0 and 1. The extra motion component has been referred to as the FSA [47]. It is associated with LW through a cost function that integrates prey density f(x), or prey abundance. Similarly, the length of every step, $s = |y - x|$, Eq. (24) is picked haphazardly from a Cauchy likelihood.

$$p(s) = {1}/{\pi} \times {T}/{(s^2 + T^2)} \quad (24)$$

The temperature (T) affects the magnitude of step-length changes. Eq. (24) calculates the acceptable probability value for accepting the new placement.

$$P = min\left\{1, exp({\Delta f}/{T}\right\} \quad (25)$$

$\Delta f = f(y) - f(x)$ is the cost difference between the present location and prior places. The new post must be more severe than the current one to be accepted; that is, if $\Delta f > 0$ it will be recognized, at which point R. Octopus will return to its native location. With the early temperature being T_0 and the step counter being k, the heating approach is planned in terms of $T(k) = {T_0}/{k}$, and FSA [48] converges to the global ideal (maximum). The length of each stage in the third R, octopus motion approach is haphazardly generated from a Gaussian distribution, making it a special kind of LW search strategy.

$$g(s) = (2\pi T)^{-D/2} e^{({\Delta x^2}/{2T})} \quad (26)$$

While Δx denotes the rate of shift X (variables vector), D denotes the size of the search region (count of elements in the cost function). Therefore, $x_{next} = x_i + \Delta x$ means the present state and x_{next} displays the following set of settings.

2.6. Performance evaluator

Classifier performance is assessed using a range of parameters. The term "$accuracy$" refers to the proportion of correctly predicted observations. Three popular measures are $precision$, $accuracy$, and $recall$. Overall $accuracy$, which includes both real negatives and positives, is referred to as $accuracy$. Reduced accuracy might be the outcome of unbalanced datasets. $Recall$ assumes minimal mistakes and searches for positives. The $F1\ score$ works well in schools with a variety of student demographics because it strikes a balance between $recall$ and $accuracy$. Both true positives and false negatives can be handled by it. These metrics aid in determining how effective ML models is.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (27)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (28)$$

$$\text{Recall} = \text{TPR} = \frac{\text{TP}}{\text{P}} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (29)$$

$$\text{F1 score} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \quad (30)$$

In the computations below, the letter TP displays an optimistic projection regarding the lucky event. When a situation has a bad ending, the acronym FP signifies a good perspective. A pessimistic prediction based on TN

yields the same result as a real negative. When the genuine conclusion is positive, the FN symbolizes a dismal future.

3 Result and discussion

The findings of the hybrid models are painstakingly presented, with various charts and tables for thorough comparison. These visual and tabular representations provide a complete review to determine which model contributes the most effectively to the prediction process. The research effort tries to bring out nuances in the prediction skills by carefully examining performance data and graphical depictions that are of considerable significance in providing insights into each model's relative strengths and limitations. Such in-depth study enhances knowledge and aids informed decision-making in selecting the most effective model for predictive tasks.

3.1. Convergence curve

Convergence curves from the prediction procedure measure performance increase as a function of training cycles for an ML model. They show how a model starts to have lower error or loss with each cycle of learning from the training data. Analyzing a convergence curve provides important information regarding the model's learning dynamics, including convergence rate, stability, and possible overfitting. A steep drop in mistakes followed by a plateau indicates good learning, but unpredictable behavior may suggest difficulties that require attention. Understanding convergence curves helps to optimize model parameters, improve forecast accuracy, and ensure resilience in real-world applications.

The convergence curve of the three hybrid models, namely, RFRD, RFJS, and RFRO, are shown in Fig. 4. Herein, in the result section, it is observed that the RFJS model starts cycle with the lowest accuracy among the presented models. Beginning with an accuracy of about 0.525, this model achieves 0.825 accuracy by the 50th cycle. Its optimal performance, however, touches the peak at 0.95 at around the 130th cycle and, as such, outperforms many models like RFRD. RFRD, beginning at a higher accuracy greater than 0.525, has already started cycle and increases performance up to 0.800 at the 50th cycle. Subsequently, it concludes the process with an accuracy of nearly 0.900 by the 100th cycle. Conversely, the RFRO model demonstrates an accuracy of approximately 0.950 by the 120th cycle, indicating superiority over RFRD but inferiority to RFJS models.

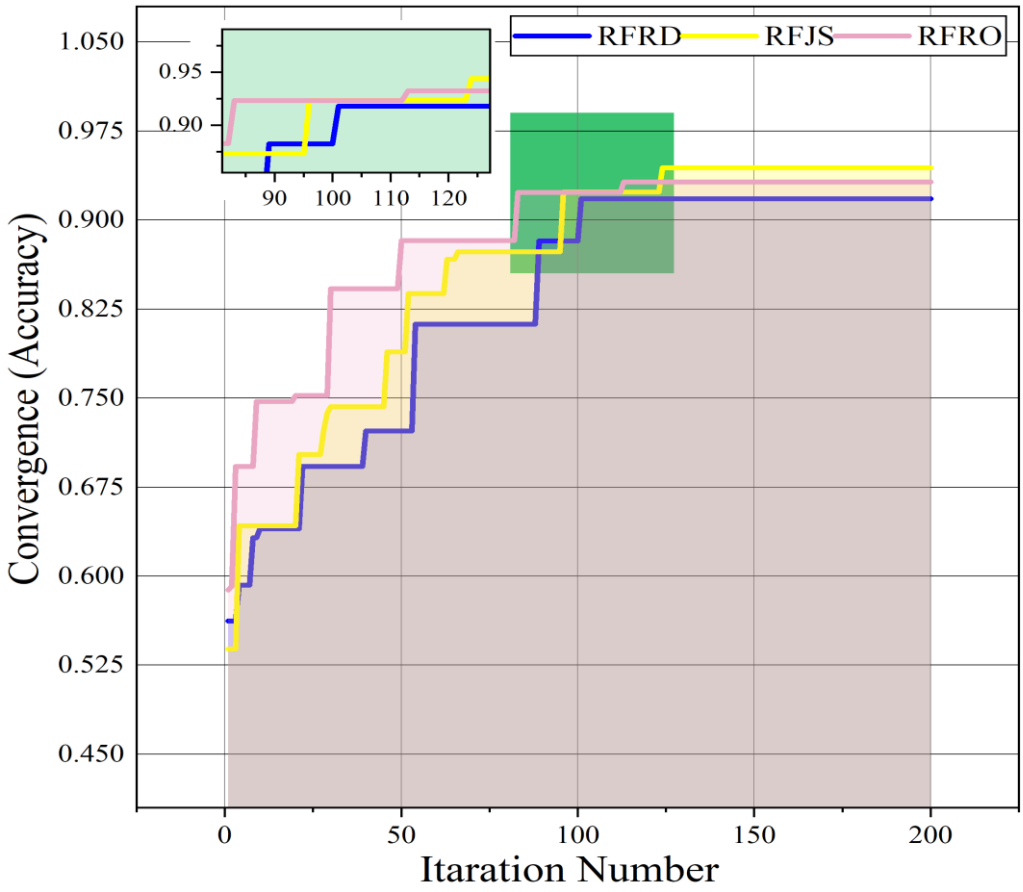


Figure 4: The convergence curve of the three presented hybrid models

3.2. Models' comparison

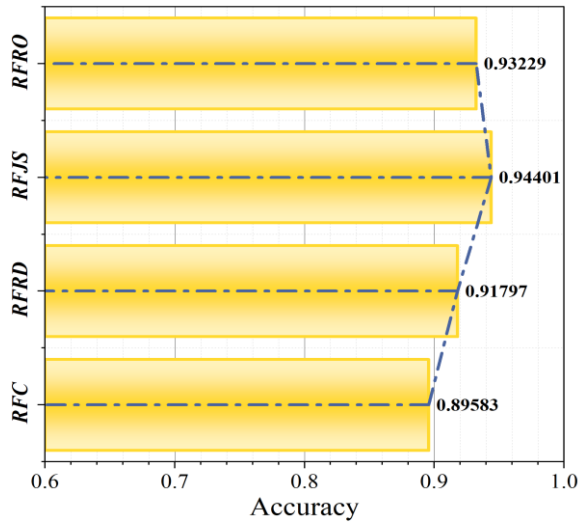
In Table 1, a comparison is made between RFC and its mixed forms in terms of outcomes during the training and testing phases. For instance, in the training phase, it is observed that the RFC model achieves an accuracy of 0.9033, which is the lowest among the base models. However, as demonstrated by the RFDR's *accuracy* of 0.9275 and the RFRO's *accuracy* of 0.9442, the *accuracy* of the model is much enhanced when paired with optimizers. With an accuracy of 0.9517, the RFJR achieves the best *accuracy*.

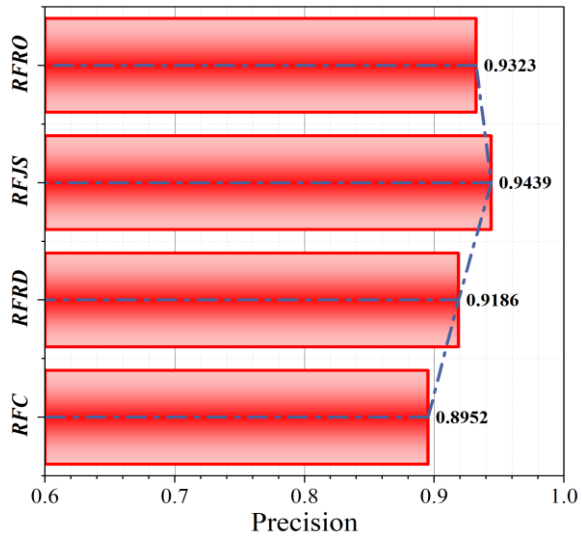
In the testing phase, the precision values of these models indicate that RFJS, with a precision value of 0.9253, is favored over the other models. The RFRRO model, with a precision value of 0.9027, can be considered the second-best model. Other models, such as RFRD and RFC, demonstrate their lower performance in this comparison, with precision values of 0.894 and 0.875, accordingly.

Table 1: RFC-based models achieved results through the performance evaluators

Section	Model	Metrics			
		Accuracy	Precision	Recall	F1 _Score
Training	RFC	0.9033	0.9033	0.9033	0.9027
	RFRD	0.9275	0.9299	0.9275	0.9265
	RFJS	0.9517	0.9517	0.9517	0.9515
	RFRO	0.9442	0.9444	0.9442	0.944
Testing	RFC	0.8783	0.8756	0.8783	0.8762
	RFRD	0.8957	0.894	0.8957	0.8945
	RFJS	0.9261	0.9253	0.9261	0.9255
	RFRO	0.9043	0.9027	0.9043	0.9021

The performance of the presented model is demonstrated through a bar plot in Fig. 5 in all phases. It is observed that the RFJS model, with a recall value of 0.944, can be considered the best model in this comparison. Following this, the RFRO model, with a recall value of 0.9323, is proven to have the highest recall value among the other models. However, the RFRD model, with a recall value of 0.918, is identified as the hybrid model with medium performance, albeit with a recall value higher than that of the RFC model, which stands at 0.8958. Furthermore, it is demonstrated that the RFC model does not aid much in the prediction process, with an F1 score value of 0.8948. Moreover, the RFDR performs worse than the RFRO model, which has an $F1\ score$ of 0.9317, and the RFJS model, which has an $F1\ score$ of 0.9437.





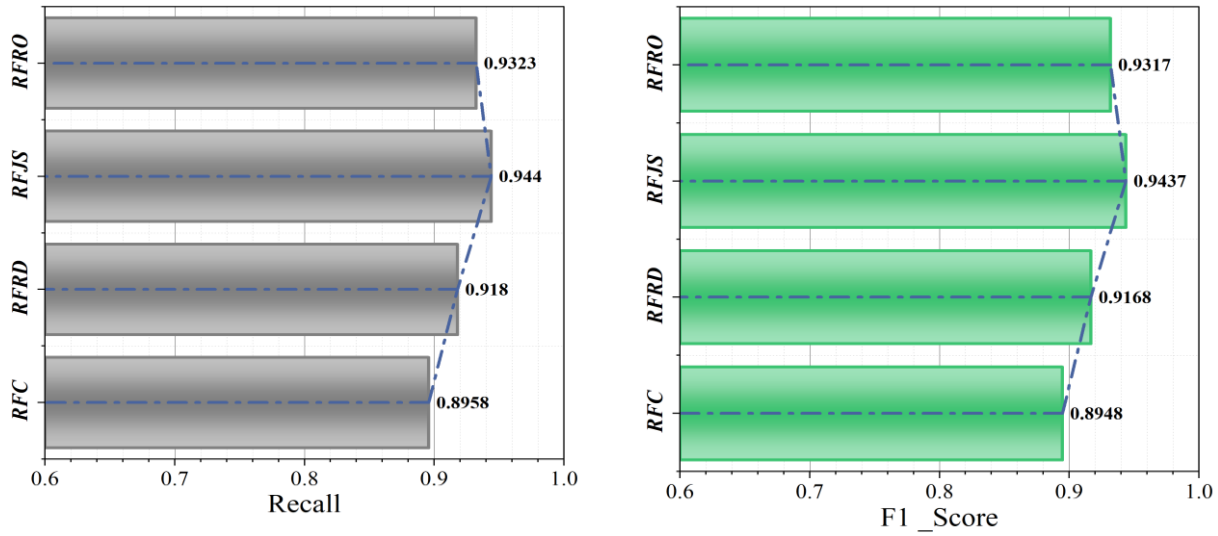


Figure 5: Bar graph showing the models' performance over all phases

In Table 2, the performance of the models is presented and contrasted across both Healthy and Diabetes conditions. For example, the RFC model displays a *precision* value of 0.9038 when they are in a healthy state and 0.879 when they are in a diabetic state. When compared to the best scheme, RFJS, which has a *precision* of 0.9472 under healthy conditions and 0.9377 under diabetic conditions shows the worst performance. The RFRO outperforms the RFDR, which records a *precision* value of 0.913 under the same situation, with a *precision* value of 0.9324 in a healthy state. Furthermore, the RFDR's *accuracy* value under the diabetic condition is 0.9282, which indicates that it functions less well than the RFRO, which has a *precision* value of 0.932 under the same scenario.

Table 2: Model performance in the four different conditions

Model	Condition	Metric		
		Precision	Recall	F1-Score
RFC	Healthy	0.9038	0.94	0.9216
	Diabetes	0.879	0.8134	0.845
RFRD	Healthy	0.913	0.966	0.9388
	Diabetes	0.9289	0.8284	0.8757
RFJS	Healthy	0.9472	0.968	0.9575
	Diabetes	0.9377	0.8993	0.9181
RFRO	Healthy	0.9324	0.966	0.9489
	Diabetes	0.932	0.8694	0.8996

The line symbol plot in Fig. 6 compares the models' predicted values with the measured values. It is observed that the RFC model's predicted value exhibits a significant difference from the measured value, with 215 out of 500 under healthy conditions. Similarly, under diabetic conditions, it shows minimal variation, with 220 out of 499 measured values. Next, in healthy conditions, the RFRD model shows consistency with 239 out of 500 measured values, nearly matching the performance of the RFRO model, which likewise records 239 out of 500 observed values. Conversely, under diabetic conditions, the RFDR's performance weakens, as evidenced by 225 out of 499 measured values, contrasting with the RFRO model's 232 out of 499 measured values under the same conditions. However, as previously noted, the RFJS model emerges as the most effective, achieving 240 out of 499 measured values under diabetic conditions and 240 out of 500 measured values under healthy conditions.

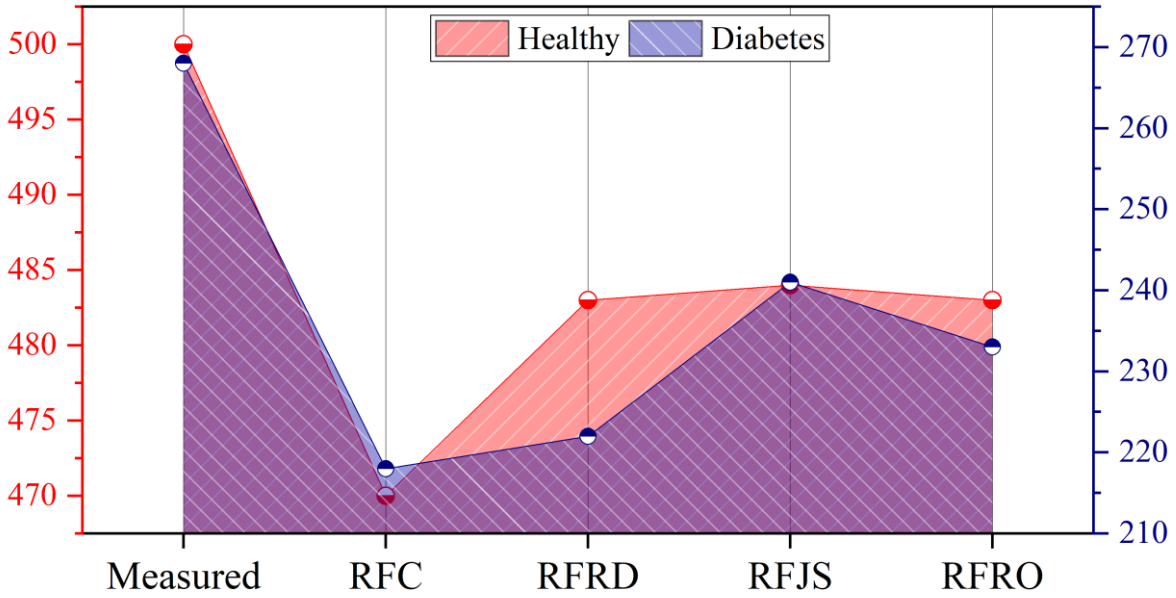
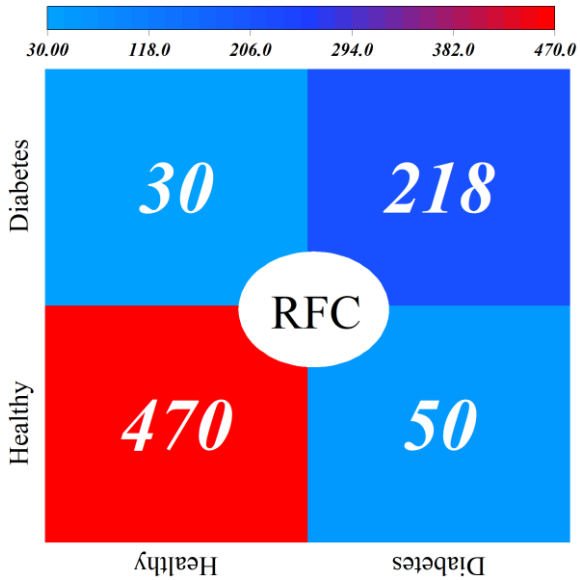
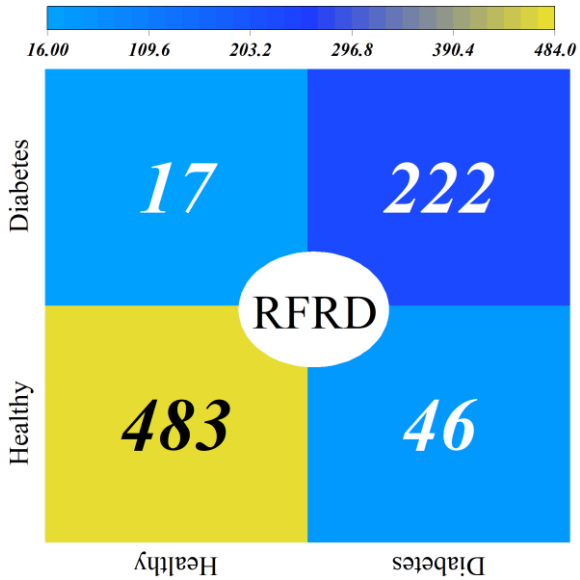


Figure 6: A line-symbol plot showing the models' percentage difference

The confusion matrix displayed in Fig. 7 comprehensively compares the models' measured values and misclassifications across healthy and diabetic conditions. Notably, the RFC model showcases a commendable 94% accuracy under healthy conditions, albeit with 30 patients misclassified under diabetic conditions. Conversely, its performance diminishes under diabetic conditions, registering an 81.34% accuracy, accompanied by 50 patients being misclassified under healthy conditions.

In stark contrast, the RFJS model emerges as the epitome of effectiveness, boasting a remarkable 96.8% accuracy under healthy conditions, with a mere 16 patients misclassified under diabetic conditions. Similarly, even under challenging diabetic conditions, this model maintains its superiority, achieving an impressive 89.92% accuracy, with only 27 patients misclassified under healthy conditions. Such data emphasize the significant increase in performance observed between the weakest model, RFC, and the most robust model, RFJS, stressing the importance of model selection in adequately diagnosing patients under varied health situations.





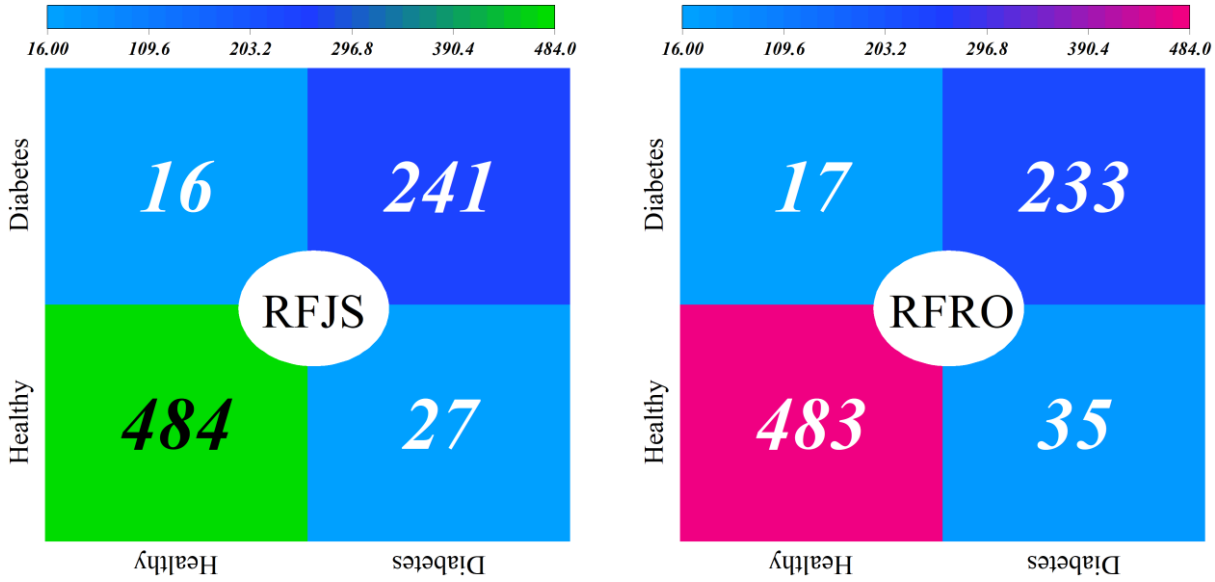


Figure 7: Confusion matrix for the accuracy of the models in four presented conditions

The Shapely Additive explanations (SHAP) in Fig. 8 show how numerous circumstances, such as pregnancy or blood pressure, affect the chance of acquiring diabetes. The following explanation gives a concise definition of how these factors influence the development of diabetes.

- Pregnancy causes complicated metabolic changes, which can have a substantial influence on diabetes. In gestational diabetes, pregnancy hormones can interfere with insulin activity, resulting in high blood sugar levels. Pre-existing diabetes, either type 1 or type 2, may require much tighter control during pregnancy to minimize risks of complications both for the mother and the child. Poorly managed diabetes during pregnancy carries risks such as macrosomia (considerable birth weight), congenital anomalies at birth, and maternal issues like hypertension. Successful management through food and exercise and, if necessary, medication under medical supervision to ensure optimum health for mother and fetus during this critical period.
- Hypertension, or high blood pressure, often accompanies diabetes and forms a lethal combination that exacerbates cardiovascular disease. High blood pressure accelerates the progression of diabetic nephropathy, retinopathy, and neuropathy, leading to an increased risk of heart disease, stroke, and kidney failure. The interaction between hypertension and diabetes is a delicate one, with complex processes such as endothelial dysfunction, inflammation, and oxidative stress, which encourage a vicious cycle of organ damage. Appropriate treatment measures aimed at both illnesses include lifestyle changes, medication adherence, and regular monitoring, which are crucial in reducing the cumulative risk and improving the long-term outcome in patients with concurrent diabetes and hypertension.

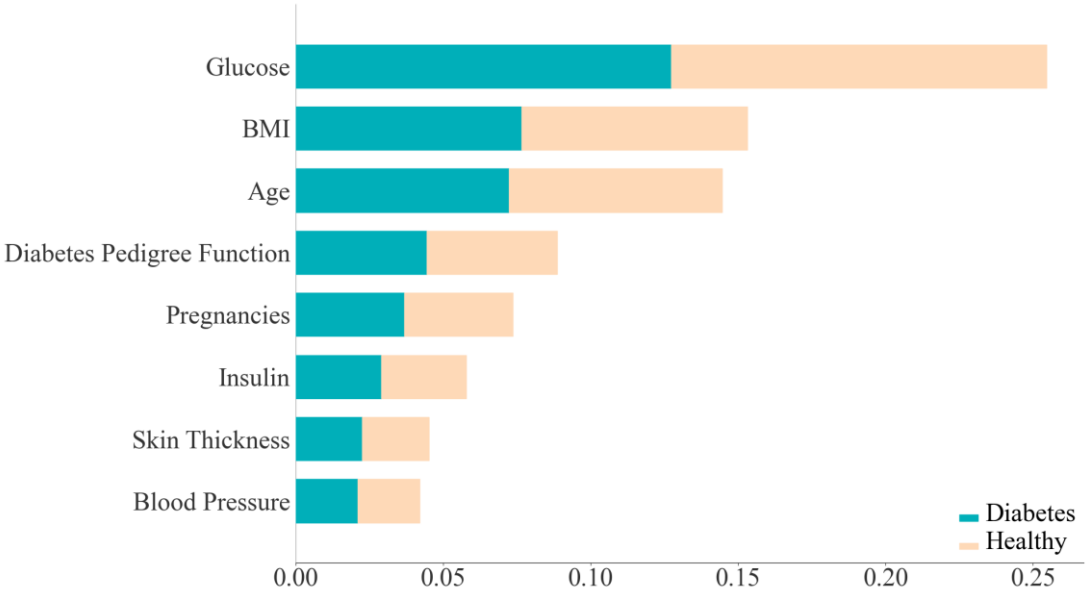


Figure 8: The SHAP sensitivity analysis of the most accurate developed model

4 Conclusion

Pregnancy-induced diabetes, medically termed gestational diabetes mellitus, can often stay well after childbirth. Research has shown that women diagnosed with GDM are at a higher risk of developing type 2 diabetes later in life. Children born from moms with GDM may also be at a higher risk for obesity and type 2 diabetes. This intergenerational transfer of metabolic risks underlines the importance of postoperative follow-up and lifestyle changes for both mother and child. Addressing GDM through early identification, adequate therapy during pregnancy, and health measures afterward is critical to reduce its long-term impact on mother and child health. Moreover, diabetes prediction during pregnancy is proposed with the RF, an ML classification model. It uses RDA, JSO, and ROA to increase the model's precision. Here, it has been determined to propose optimizers integrated with a model for enhanced prediction performance. These results indicated that RFJS, with the best performance of all compared models, resulted in a value of 0.9377 regarding diabetic conditions. The RFRO model, therefore, with a value of 0.932, stands out as the second-best scheme. The RFRD model, with a precision of 0.9282, shows middle performance and a higher value than RFC, recording a precision of 0.879. Even though ML could be of help in predicting pregnancy-induced diabetes, there are limits. In case biases within the dataset are not handled well, they may have an impact on predictions. Imbalances in demographic representation or missing data might limit model generalizability. Furthermore, interpretability issues occur when sophisticated models hide the reasons behind forecasts, limiting clinician trust and decision-making. Scalability difficulties arise as models are deployed across a variety of healthcare settings. Integrating existing electronic health record systems presents technological challenges, possibly affecting workflow efficiency. Besides, model performance can degrade over time due to changes in patient demographics or healthcare practices, and the models will require re-calibration, ethical considerations regarding privacy and permission when working with sensitive data of patients, and developing and implementing a model keeping in view all regulatory requirements, such as GDPR or HIPAA. The clinical utility will thus depend on the availability of real-time data streams and their seamless integration into clinical workflows. Implementation issues may hamper uptake and reduce the influence of predictive models on patient care. Despite these limits, continuing research seeks to overcome these issues through better data-gathering methods, model interpretability methodologies, and collaboration between healthcare experts and data scientists.

Authorship contribution statement

Shuai Zhao: Writing-Original initial drafting, Conceptualization, Supervision, Project administration.

Huiqin Ren: Methodology, Software

Conflicts of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Author statement

The manuscript has been read and approved by all the authors, the requirements for authorship, as stated earlier in this document, have been met, and each author believes that the manuscript represents honest work.

Ethical approval

All authors have been personally and actively involved in substantial work leading to the paper, and will take public responsibility for its content.

References

[1] E. M. Wendland *et al.*, "Gestational diabetes and pregnancy outcomes-a systematic review of the World Health Organization (WHO) and the International Association of Diabetes in Pregnancy Study Groups (IADPSG) diagnostic criteria," *BMC Pregnancy Childbirth*, 12: 1–13, 2012. https://doi.org/10.1186/1471-2393-12-23

[2] T. A. Buchanan, A. H. Xiang, and K. A. Page, "Gestational diabetes mellitus: risks and management during and after pregnancy," *Nat Rev Endocrinol*, 8(11): 639, 2012. https://doi.org/10.1038/nrendo.2012.96

[3] T. Filardi, F. Tavaglione, M. Di Stasio, V. Fazio, A. Lenzi, and S. Morano, "Impact of risk factors for gestational diabetes (GDM) on pregnancy outcomes in women with GDM," *J Endocrinol Invest*, 41: 671–676, 2018. https://doi.org/10.1007/s40618-017-0791-y

[4] M. Scavini and A. Secchi, "Diabetes in pregnancy," 2019, *Springer*. https://doi.org/10.1007/s00592-019-01364-z

[5] P. M. Catalano *et al.*, "The hyperglycemia and adverse pregnancy outcome study: associations of GDM and obesity with pregnancy outcomes," *Diabetes Care*, 35(4): 780–786, 2012. https://doi.org/10.2337/dc11-1790

[6] A. N. Sweeting *et al.*, "Gestational diabetes mellitus in early pregnancy: evidence for poor pregnancy outcomes despite treatment," *Diabetes Care*, 39(1): 75–81, 2016. https://doi.org/10.2337/dc15-0433

[7] H. Kleinwechter *et al.*, "Gestational diabetes mellitus (GDM) diagnosis, therapy and follow-up care," *Experimental and Clinical Endocrinology & Diabetes*, 122(07): 395–405, 2014. https://doi.org/10.1016/j.molmet.2017.12.015

[8] L. Zajdenverg and C. A. Negrato, "Gestational diabetes mellitus and type 2 diabetes: same disease in a different moment of life? Maybe not," 2017, *SciELO Brasil*. https://doi.org/10.1590/2359-3997000000276

[9] W. Ye, C. Luo, J. Huang, C. Li, Z. Liu, and F. Liu, "Gestational diabetes mellitus and adverse pregnancy outcomes: systematic review and meta-analysis," *Bmj*, 377, 2022. https://doi.org/10.1136/bmj-2021-067946

[10] C. A. Crowther, J. E. Hiller, J. R. Moss, A. J. McPhee, W. S. Jeffries, and J. S. Robinson, "Effect of treatment of gestational diabetes mellitus on pregnancy outcomes," *New England journal of medicine*, 352(24): 2477–2486, 2005. DOI: 10.1056/NEJMoa042973

[11] Y. Yogev and G. H. A. Visser, "Obesity, gestational diabetes and pregnancy outcome," in *Seminars in Fetal and Neonatal Medicine*, Elsevier, 2009: 77–84. https://doi.org/10.1016/j.siny.2008.09.002

[12] A. Ben-Haroush, Y. Yogev, and M. Hod, "Epidemiology of gestational diabetes mellitus and its association with Type 2 diabetes," *Diabetic medicine*, 21(2): 103–113, 2004. https://doi.org/10.1046/j.1464-5491.2003.00985.x

[13] W. Bao *et al.*, "Long-term risk of type 2 diabetes mellitus in relation to BMI and weight change among women with a history of gestational diabetes mellitus: a prospective cohort study," *Diabetologia*, 58: 1212–1219, 2015. https://doi.org/10.1007/s00125-015-3537-4

[14] G. K. Poomalar, "Changing trends in management of gestational diabetes mellitus," *World J Diabetes*, 6(2): 284, 2015. DOI: 10.4239/wjd. v6. i2.284

[15] M. C. Carolan-Olah, "Educational and intervention programmes for gestational diabetes mellitus (GDM) management: An integrative review," *Collegian*, 23(1): 103–114, 2016. https://doi.org/10.1016/j.colegn.2015.01.001

[16] C. Kim, "Gestational diabetes: risks, management, and treatment options," *Int J Womens Health*, 339–351, 2010. DOI: 10.2147/IJWH.S13333

[17] N. W. Cheung, "The management of gestational diabetes," *Vasc Health Risk Manag*, 153–164, 2009. DOI: 10.1056/NEJMoa2214956

[18] I. Hromadnikova, K. Kotlabova, and L. Krofta, "cardiovascular disease-associated microRNAs as novel biomarkers of first-trimester screening for gestational diabetes mellitus in the absence of other pregnancy-related complications," *Int J Mol Sci*, 23(18): 10635, 2022. https://doi.org/10.3390/ijms231810635

[19] A. Kouhkan *et al.*, "Gestational diabetes mellitus: Major risk factors and pregnancy-related outcomes: A cohort study," *Int J Reprod Biomed*, 19(9): 827, 2021. DOI: 10.18502/ijrm. v19i9.9715

[20] Z. H. Talasaz, R. Sadeghi, F. Askari, S. Dadgar, and A. Vatanchi, "First trimesters Pregnancy-Associated Plasma Protein-A levels value to Predict Gestational diabetes Mellitus: A systematic review and meta-analysis of the literature," *Taiwan J Obstet Gynecol*, 57(2): 181–189, 2018. https://doi.org/10.1016/j.tjog.2018.02.003

[21] E. Cosson *et al.*, "Early screening for gestational diabetes mellitus is not associated with improved pregnancy outcomes: an observational study including 9795 women," *Diabetes Metab*, 45(5): 465–472, 2019. https://doi.org/10.1016/j.diabet.2018.11.006

[22] J. Dunne *et al.*, "Diabetic and hypertensive disorders following early pregnancy loss: a systematic review and meta-analysis," *EClinicalMedicine*, 2024. DOI: 10.1016/j.eclinm.2024.102560

[23] A. Goyal, Y. Gupta, and N. Tandon, "Overt diabetes in pregnancy," *Diabetes Therapy*, 13(4):589–600, 2022. https://doi.org/10.1007/s13300-022-01210-6

[24] D. Simmons *et al.*, "Treatment of gestational diabetes mellitus diagnosed early in pregnancy," *New England Journal of Medicine*, 388(23): 2132–2144, 2023. DOI: 10.1056/NEJMoa2214956

[25] B. E. Metzger *et al.*, "International association of diabetes and pregnancy study groups recommendations on the diagnosis and classification of hyperglycemia in pregnancy: response to Weinert," *Diabetes Care*, 33(7): e98–e98, 2010. https://doi.org/10.2337/dc10-0544

[26] J. W. Wyatt *et al.*, "Congenital anomaly rate in offspring of mothers with diabetes treated with insulin lispro during pregnancy," *Diabetic medicine*, 22(6): 803–807, 2005. https://doi.org/10.1111/j.1464-5491.2004.01498.x

[27] M. Hod *et al.*, "The International Federation of Gynecology and Obstetrics (FIGO) Initiative on gestational diabetes mellitus: A pragmatic guide for diagnosis, management, and care," *Int J Gynaecol Obstet*, 131: S173–S211, 2015. DOI: 10.1016/S0020-7292(15)30033-3

[28] L. Jovanovic, M. Druzin, and C. M. Peterson, "Effect of euglycemia on the outcome of pregnancy in insulin-dependent diabetic women as compared with normal control subjects," *Am J Med*, 71(6): 921–927, 1981. https://doi.org/10.1016/0002-9343(81)90301-6

[29] L. Jovanovic-Peterson *et al.*, "Maternal postprandial glucose levels and infant birth weight: the Diabetes in Early Pregnancy Study," *Am J Obstet Gynecol*, 164(1): 103–111, 1991. https://doi.org/10.1016/0002-9378(91)90637-7

[30] W. C. Bevier, R. Fischer, and L. Jovanovic, "Treatment of women with an abnormal glucose challenge test (but a normal oral glucose tolerance test) decreases the prevalence of macrosomia," *Am J Perinatol*, 16(06): 269–275, 1999. DOI: 10.1055/s-2007-993871

[31] W. M. Hague, "endocrine disease (including diabetes)," *Best Pract Res Clin Obstet Gynaecol*, 15(6): 877–889, 2001. https://doi.org/10.1053/beog.2001.0235

[32] N. A. ElSayed *et al.*, "15. Management of diabetes in pregnancy: standards of care in diabetes—2023," *Diabetes Care*, vol. 46, no. Supplement_1: S254–S266, 2023. https://doi.org/10.2337/dc23-S015

[33] K. M. Flegal, M. D. Carroll, C. L. Ogden, and L. R. Curtin, "Prevalence and trends in obesity among US adults, 1999-2008," *JAMA*, 303(3): 235–241, 2010. DOI:10.1001/jama.288.14.1723

[34] J. E. Shaw, R. A. Sicree, and P. Z. Zimmet, "Global estimates of the prevalence of diabetes for 2010 and 2030," *Diabetes Res Clin Pract*, 87(1): 4–14, 2010. https://doi.org/10.1016/j.diabres.2009.10.007

[35] E. Araki *et al.*, "Japanese clinical practice guideline for diabetes 2019," *Diabetol Int*, 11: 165–223, 2020. https://doi.org/10.1007/s13340-020-00439-5

[36] M. Garcia, S. L. Mulvagh, C. N. Bairey Merz, J. E. Buring, and J. E. Manson, "cardiovascular disease in women: clinical perspectives," *Circ Res*, 118(8): 1273–1293, 2016. https://doi.org/10.1161/CIRCRESAHA.116.307547

[37] W. T. Garvey, J. M. Olefsky, J. Griffin, R. F. Hamman, and O. G. Kolterman, "The effect of insulin treatment on insulin secretion and insulin action in type II diabetes mellitus," *Diabetes*, 34(3): 222–234, 1985. https://doi.org/10.2337/diab.34.3.222

[38] A. Collier *et al.*, "Relationship of skin thickness to duration of diabetes, glycemic control, and diabetic complications in male IDDM patients," *Diabetes Care*, 12(5): 309–312, 1989. https://doi.org/10.2337/diacare.12.5.309

[39] A. Parmar, R. Katariya, and V. Patel, "A review on random forest: An ensemble classifier," in *international conference on intelligent data communication technologies and internet of things (ICICI) 2018*, Springer, 2019: 758–763. https://doi.org/10.1007/978-3-030-03146-6_86

[40] A. F. Fard and M. Hajiaghaei-Keshteli, "Red Deer Algorithm (RDA); a new optimization algorithm inspired by Red Deers' mating," in *International Conference on Industrial Engineering, IEEE*, 2016: 331–342. DOI:10.1007/s00500-020-04812-z

[41] J.-S. Chou and D.-N. Truong, "A novel metaheuristic optimizer inspired by behavior of jellyfish in ocean," *Appl Math Comput*, 389: 125535, 2021. https://doi.org/10.1016/j.amc.2020.125535

[42] A. Alam *et al.*, "Jellyfish search optimization algorithm for mpp tracking of pv system," *Sustainability*, 13(21): 11736, 2021. https://doi.org/10.3390/su132111736

[43] X. Hou, Y. Yan, Q. Zhan, J. Wang, B. Xiao, and W. Jiang, "Unsupervised machine learning effectively clusters pediatric spastic cerebral palsy patients for determination of optimal responders to selective dorsal rhizotomy," *Sci Rep*, 13(1): 8095, 2023. https://doi.org/10.1038/s41598-023-35021-x

[44] H. Deora *et al.*, "Microsurgical rhizotomy as treatment for trigeminal neuralgia in patients with multiple sclerosis: turnpike or dirt road?" *J Neurosurg*, 130(5): 1775–1778, 2018. https://doi.org/10.3171/2018.8.JNS182227

[45] A. M. Reynolds, "Signatures of active and passive optimized Lévy searching in jellyfish," *J R Soc Interface*, 11(99): 20140665, 2014. https://doi.org/10.1098/rsif.2014.0665

[46] X.-S. Yang, "A new metaheuristic bat-inspired algorithm," in *Nature inspired cooperative strategies for optimization (NICSO 2010)*, Springer, 2010: 65–74. https://doi.org/10.1007/978-3-642-12538-6_6

[47] S. Kirkpatrick, "Improvement of reliabilities of regulations using a hierarchical structure in a genetic network," *Science (1979)*, 220: 671–680, 1983. https://doi.org/10.1016/j.ress.2021.108295

[48] Q. Askari, I. Younas, and M. Saeed, "Political Optimizer: A novel socio-inspired meta-heuristic for global optimization," *Knowl Based Syst*, 195: 105709, 2020. https://doi.org/10.1016/j.knosys.2020.105709

Application of Multimodal Generation Model in Short Video Content Personalized Generation

Minghui Yang
New Media E-commerce Institute, Chongqing Institute of engineering, Chongqing 400000, Chongqing, China
E-mail : minghuiyangminghui@163.com

Keywords: personalized content, multimodal generation, stochastic paint optimizer with intelligent convolutional neural network (SPO-IntelliConvNet), modalities

Received: June 23, 2025

The rise of short video platforms has led to a higher demand for rapidly generated personalized content. Existing systems either struggle with high levels of customization or require large amounts of data, limiting real-time production. A multimodal generation model serves as the focus of study to generate customized short video content that adapts to user preferences as well as their behavioral patterns. The objective targets an integrative model using text alongside image and audio data to make context-specific short video content, which delivers personalized entertainment. First, it analyses user preferences from interaction data and then synthesizes corresponding video content using a novel method called a stochastic paint optimizer with an intelligent convolutional neural network (SPO-IntelliConvNet). The SPO component ensures optimal representation of multimodal content by improving feature selection and parameter tuning through stochastic search algorithms modelled after the dynamics of abstract paintings. The IntelliConvNet is used to combine and interpret several modalities, allowing for efficient personalization that is consistent with user preferences. To develop personalized content, user preference data is collected, which includes interactions such as video views and comments. The model employs natural language processing (NLP), audio processing, and computer vision to merge text, image, and audio modalities. Pre-processing includes tokenization for text, Canny edge detection for images, and Wiener filtering for audio, optimizing each modality for better analysis and feature extraction using principal component analysis (PCA) to reduce the dimensions of features from all three modalities to lower dimensions while preserving essential information. This proposed approach achieved superior personalized content development, leading to increased user satisfaction and engagement. The performance of the proposed method was evaluated using BLEU-4 (0.55), ROUGE-L (0.79), METEOR (0.72), and CIDEr (0.80). The system's ability to successfully incorporate multimodal data resulted in more precise video customization, as demonstrated by interaction metrics and user comments. This multimodal generation model provides an advanced solution for creating personalized short video content, increasing the user experience with highly tailored content.

Povzetek: Študija predlaga večmodalni generativni model za osebno prilagojene kratke videe, ki se z združitvijo NLP, računalniškega vida in zvočne obdelave ter novim optimizatorjem SPO-IntelliConvNet iz uporabniških interakcij uči preference in sintezira kontekstno ustrezno vsebino v realnem času.

1 Introduction

The sudden growth of short video platforms has resulted in an increasing demand for customized content that boosts user interaction. With the widespread use of mobile technology and social media, short videos have emerged as a prominent medium of communication and entertainment. Users anticipate content that is personalized to their interests, and therefore, personalization becomes a central theme of video creation. Traditional metadata-based user tracking recommendation systems are likely to overlook the multimodal nature of short videos and therefore make less interactive and less relevant recommendations [1]. Dynamic, engaging content results from short video personalization, which processes different data types through text, images, and audio. Textual analysis of captions and visual analysis of object detection and scene understanding, and audio examination of speech and ambient noise create content that matches user attention patterns [2].

Real-time video generation through customized content becomes possible because the system avoids incorporating pre-selected templates to suit individual needs. This provides for real-time content generation,

which responds to users' specific behaviour, leading to better storytelling and interaction. The behaviours of user interaction (likes, shares, and comments) are integrated with the content of videos and customized according to the feedback presented instantly, and this makes the content of the video more meaningful and engaging [3]. Videos filmed by individuals now proliferate their influence in entertainment, educational, and marketing sectors, as well as improve social media engagement. The personalization of content enables users to achieve maximum educational potential in videos since the content generation method creates true brand interactions, and video content generates precise information that suits the demands of the viewer [4]. The generation of personalized short videos intends to generate video experiences that are engaging and adaptive depending on user interactions and interests. This technique uses text, images, and audio to dynamically adjust content and make it more relevant and engaging to the viewer. It is regularly applied to entertainment, education, and advertising, providing users with customized and engaging video experiences to improve storytelling and viewer engagement [5].

The objective is to construct a multimodal generation model for individualized short video content with integration of text alongside images, together with audio features. The model called SPO-IntelliConvNet improves feature extraction to enhance content customization and increase user engagement.

1.1 Significance of the contribution

• The system enhances personalization through both user preferences and adaptable content creation methods in short video generation.

• Adaptive multimodal feature fusion with NLP, vision, and audio – The framework jointly processes text (NLP), images (computer vision), and audio (signal processing) using modality-specific preprocessing: tokenization and embedding for text, Canny edge detection for images, and Wiener filtering for audio. NLP plays a central role by extracting semantic meaning, sentiment, and contextual cues from user comments and video descriptions, which guide the personalization process.

• Feature extraction using Principal Component Analysis (PCA) enhances content synthesis by identifying key personalization patterns.

• Novel integration of optimization and deep learning proposes SPO-IntelliConvNet, the first framework to integrate the Stochastic Paint Optimizer (SPO) with an adaptive convolutional neural network for personalized short video generation. Unlike conventional multimodal models that rely on static parameters, SPO dynamically adjusts convolutional filters, fusion weights, and hyperparameters, enabling adaptive multimodal learning.

• Comprehensive evaluation of personalization quality – Beyond automated metrics (BLEU, ROUGE-L, METEOR, CIDEr), we incorporate baseline comparisons, statistical significance testing, and user-centric evaluation (e.g., watch-time analysis and subjective preference ratings). This multi-layered validation ensures that performance gains are not only statistically meaningful but also translate into measurable improvements in user engagement and satisfaction.

1.2 Research organization

The research follows the structure outlined below. Section 2 presents a review of the literature on multimodal content generation. Section 3 explains the proposed SPO-IntelliConvNet model and its methodology for personalized short video generation. Section 4 details the experimental results, highlighting the model's performance based on evaluation metrics. Finally, Section 5 describes the conclusion.

2 Literature review

A user behavior-aware multi-task learning model for enhanced short video recommendation (UBA-SVR) was presented in [6] by utilizing knowledge about changing user communication. The outcomes indicated that the suggested approach significantly improves several prediction tasks instantly.

A multimodal short video recommendation framework was proposed, integrating text, image, and audio processing [7]. Crawler technology extracted text, frame processing captured images, and voice data were analyzed. Another data fusion approach transformed video features into a dense space to enhance the recommendation process. While the model provided a better understanding of user behavior, it encountered difficulties with improving both multimedia interactions and computation speed optimization.

A unique recommendation strategy for short video platforms was proposed in [8]. The suggested approach combines an error back propagation neural network (EBPNN) with the term frequency inverse document frequency (TF-IDF) algorithm to investigate the possible relationship between users and videos for text mining. The findings revealed that the model's forecasting accuracy had increased substantially, with scores of 73.50% and 88%.

A knowledge graph-based recommendation system for short new media videos was proposed in [9]. The findings demonstrated that using knowledge graphs to recommend short videos could significantly increase the standard of content suggestions and provide consumers with a more engaging and customized viewing experience. Table 1 shows the description of related works.

Table 1: Comparison between the SOTA method in short video content personalized generation

Reference	**Dataset**	**Methodology**	**Key findings**
Zhu (2025) [10]	Short video platform dataset	Deep Learning and Reinforcement Learning (DLRL)	Achieved click-through rate (6.22%), Accuracy (0.877), and Recall (0.858)
Yang (2025) [11]	Short video dataset	Deep Learning-based content analysis and recommendation	Improved average precision by 4.7%, 3.3%, and 4.0%, and recall by 3.5%, 1.2%, and 2.1% compared to CF, MF, and Content-Based Filtering
Lu and Nam (2021) [12]	Short video dataset	Artificial Intelligence–based Video Auditing	Optimized content screening, enhanced original video support, and ensured sustainable, healthy development of mobile short video platforms.
Song and Liu (2024) [13]	TikTok dataset	Key and Objective Indicator System Based Communication Network (KOIS-PN) communication network + Attention-based Bidirectional Long Short-Term Memory Network (AT-BiLSTM) neural network	The KOISPN-ATBiLSTM short video model for international communication networks demonstrates certain growth and validity.

2.1 Problem statements

Short video platforms require accurate personalized recommendation systems to enhance user engagement. Existing methods, such as UBA-SVR [6], multimodal frameworks [7], EBPNN-TF-IDF hybrids [8], and knowledge graph-based systems [9], improve prediction accuracy but face limitations in handling multimodal interactions, semantic relationships, and computational efficiency. The *SPO-IntelliConvNet* integrates user behavior modeling, multimodal feature extraction, and knowledge graphs to overcome these gaps, providing faster, context-aware, and highly personalized recommendations. This approach combines prior strengths while mitigating their respective limitations.

3 Methodology

The proposed methodology integrates SPO-IntelliConvNet to enhance personalized short video generation. Including user data is pre-processed, and feature extraction is performed using PCA. IntelliConvNet identifies hierarchical patterns in multimodal content, while SPO optimizes model parameters for improved accuracy, enabling adaptive and AI-driven video synthesis. Figure 1 illustrates the framework of the SPO-IntelliConvNet model.

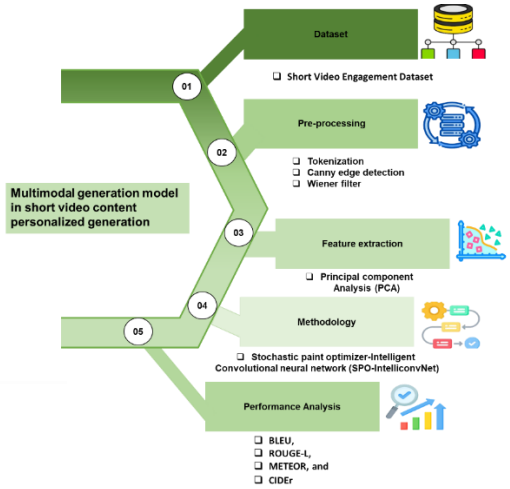


Figure 1: The framework of the proposed model

3.1 Data set

The Short Video Engagement Dataset (https://www.kaggle.com/datasets/programmer3/short-video-engagement-dataset) was gathered from Kaggle; it

includes YouTube Shorts, TikTok, and Instagram Reels, capturing user engagement and multimodal features. It represents video content data, including user behaviour (views, likes, comments, and shares), text features (title and description length), image attributes (edge intensity and colour histogram), and audio properties (spectral entropy, audio intensity). The engagement score (0 for low, 1 for high) serves as the target variable, helping analyze patterns in personalized short video content consumption. To clearly state its size (17,654 rows, 12 columns), the modalities covered (user interaction, textual, visual, and audio features), and the target variable (binary engagement score). We also now explicitly acknowledge its limitations in diversity and representativeness, noting that while it includes data from three major platforms (YouTube Shorts, TikTok, Instagram Reels), it may not fully capture domain-specific or regional variations in short video content.

3.2 Pre-processing

Pre-processing enhances multimodal data for personalized short video generation, applying tokenization for text, Canny edge detection for images, and Wiener filtering for audio to optimize feature extraction and content synthesis. Audio processing, computer vision, and NLP characteristics are combined in an organized multimodal fusion approach. The fusion improves semantic coherence and contextual significance by capturing relationships between modalities. The model creates information by coordinating textual, visual, and audio data, which shows a thorough comprehension of the user's intent and preferences. This approach guarantees that every input modality makes a valuable contribution to personalization, enhancing user engagement metrics and overall content quality.

3.2.1 Tokenization

The Short Video Engagement Dataset is used for the tokenization process, resulting in a structured and clean text representation, improving its efficiency. By removing noise, stop words, and inconsistencies, the processed text enhances semantic understanding, ensuring more relevant and personalized content generation. This structured tokenized data leads to improved classification accuracy, better content personalization, and enhanced user engagement in short video recommendations.

3.2.2 Canny edge detection

This technique requires the Short Video Engagement Dataset to extract useful edge features from images, which enhances its operational efficiency. The method identifies sudden intensity changes, assisting in the detection of object boundaries and eliminating redundant data. By using Gaussian filtering to blur the image, followed by gradient calculation and non-maximum suppression, it efficiently retains important edges. Double thresholding and edge tracking then refine the outcome, retaining only strong edges. This technique supports the feature extraction process by limiting computational complexity without compromising the accuracy of generating customized short video content. The Gaussian filtering is represented in equation (1).

$$M(p,q) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{p^2+q^2}{2\sigma^2}\right) \quad (1)$$

Where $M(p,q)$ indicates the Gaussian filter value at pixel position, $\exp\left(-\frac{p^2+q^2}{2\sigma^2}\right)$ signifies the exponential decay function, p, q represent the filter coordinates relative to the center of the fixed kernel, π indicates the balance of pixel intensities, and σ signifies the standard deviation of the Gaussian distribution. After noise reduction, image gradients are computed using Sobel operators to detect edge intensity variations. The gradient magnitude and direction are given by equations (2 and 3).

$$M = \sqrt{M_p^2 + M_q^2} \quad (2)$$

$$\theta = \tan^{-1}\frac{M_q}{M_p} \quad (3)$$

Non-maximum suppression thins edges, double thresholding classifies strong and weak edges, and hysteresis tracking retains weak edges linked to strong ones. Where M indicates the magnitude of the personalized video feature vector, θ represents the gradient direction, M_p signifies the primary feature magnitude, and M_q indicates the secondary feature magnitude. Canny edge detection serves as a crucial pre-processing step to enhance image analysis accuracy. Refining edge structures ensures robust performance in subsequent image-processing tasks.

3.2.3 Wiener filtering

The Short Video Engagement Dataset data are used for this method, as it performs the greatest trade-off between inverse filtering and noise attenuation from audio signals, enhancing its efficiency. It efficiently subtracts additive noise without affecting useful signal components and improves the total quality. It is the best optimum in the reduction of mean square error (MSE) that ensures better reconstruction of the signal. In the pre-processing phase, Wiener filtering is employed for suppressing background noise and recovering original audio features to enhance clarity and accuracy in audio-based applications, as derived in equation (4).

$$W(v) = \frac{P_m(v)}{P_m(v)+P_n(v)} \quad (4)$$

Where $W(v)$ signifies the personalization weight of video v, $P_m(v)$ indicates the probability or score that video v matches the user interests, and $P_n(v)$ represents the probability or score that video v does not match the user interests.

Pre-processing enhances data quality for feature extraction. Tokenization refines textual data by segmenting meaningful units. Canny edge detection enhances the clarity of image structures, and it sharpens object boundaries. Wiener filtering is an approach to audio denoising to preserve the quality of the signal. The enhancement of these outputs through design produces more accurate features, which enhance the accuracy of the models used in the extraction procedure.

3.3 Feature extraction

Principal component analysis (PCA)

PCA is a dimensionality reduction technique of representing high dimensional data into a lower dimensional space and retaining critical information. It identifies main components (PCs) which can be measured with maximum variance, which minimises redundancy and noise. PCA improves the efficiency of the computations and it is extensively applied in the areas of feature extraction, pattern recognition, and machine learning (ML).

PCA is used as a feature extraction method to minimize dimensions without losing critical data. With a collection of transformed, pre-processed data, PCA is used to determine the most important components that explain maximum variance. $W = W_1, W_2, W_o$ represent the feature set obtained from preprocessing, where the covariance matrix C and eigenvalues $1 > \lambda 2 > \cdots . > \lambda p$. The PCs are calculated as in equations (5 and 6).

$$Z_1 = b_1' W = b_{11} W_1 + b_{12} W_2 + \cdots . b_{1o} W_o \quad (5)$$

where:

Z_1 is the first principal component,

b_1' is the eigenvector corresponding to the largest eigenvalue,

b_{11}: contribution of feature W_1 to Z_1,

b_{12}: contribution of feature W_2 to Z_1,

Let $W = W_1, W_2, W_o$ be the random variables, with covariance matrix C and eigenvalues $\lambda 1 > \lambda 2 > \cdots . > \lambda p$.

$$Z_2 = b_2' W = b_{21} W_1 + b_{22} W_2 + \cdots . b_{2o} W_o \quad (6)$$

$$\vdots \quad \vdots$$

$$Z_O = b_O' W = b_{O1} W_1 + b_{O2} W_2 + \cdots . b_{OO} W_O \quad (7)$$

Variance $Var(Z_j)$, covariance $Cov(Z_j, Z_l)$ in equations (7 to 9).

$$Var(Z_j) = b_j' \sum b_j; j = 1,2, \ldots, o \quad (8)$$

$$Cov(Z_j, Z_l) = b_j' \sum b_l; j = 1,2, \ldots, o \quad (9)$$

The process of extracting features, improves the representation of data by making it less redundant whilst it preserves important information gained during pre-processing stages. j Index of the principal component (ranging from 1 to o, where o is the number of principal components retained). This boosts computation efficiency, removes redundancy and boosts accuracy of classification allowing better data representation to optimize model performance in short video content personalized generation.

3.4 Personalized content generation using stochastic paint optimizer-tuned intelligent convolutional neural network (SPO-IntelliConvNet)

The integration of SPO within IntelliConvNet is designed as an optimization-driven enhancement. SPO functions as a parameter-tuning mechanism, where convolutional filters, feature fusion weights, and learning rates are iteratively optimized using stochastic search principles. IntelliConvNet then operates on these optimized parameters to extract, align, and combine modality-specific features. This synergy ensures that each modality, text, audio, and image contribute meaningfully to the final representation. The result is a CNN variant that self-adjusts its internal representations, making the integration explicit and functionally distinct from standard architectures. The Stochastic Paint Optimizer with IntelliConvNet to achieve adaptive multimodal feature extraction. Unlike traditional models, SPO dynamically optimizes convolutional filters and fusion weights, enabling context-aware learning across text, image, and audio modalities. The creative analogy with color theory highlights complementary interactions, inspiring new perspectives in multimodal optimization-driven content generation for personalized short video applications.

3.4.1 Intelligent convolutional neural network (IntelliConvNet)

The deep learning (DL) structure known as CNN exists to process structured grid data, specifically images. The distinct layers of CNNs include convolutional and pooling, and fully connected sections to extract spatial data while building hierarchical patterns from this data. By utilizing weight-sharing mechanisms, CNNs reduce computational complexity while improving feature learning. These models provide multiple applications in visual object identification alongside pattern recognition and identification of objects in images.

The implementation of IntelliConvNet technology allows the multimodal generation model to create customized short video content. Advanced CNN mechanisms in the model transform text and image together with audio features for enhancement purposes. IntelliConvNet applies adaptive convolutional filtering with attention mechanisms and optimization techniques to obtain improved efficiency and accuracy. The model dynamically chooses dynamic kernels, which allows it to adjust convolutional parameters according to content features for high-quality feature representation.

PCA features two necessary operations that enable dimension reduction while preserving important information, together with feature extraction. SPO optimization allows convolutional layers to master filter weights independently from changing content complexities in a self-learning process. Depth-wise separable convolutions and residual connections are used to reduce the computational load at the expense of minimal loss in model performance. The model is also optimized using Neural Architecture Search (NAS) to learn the IntelliConvNet architectures automatically, achieving a trade-off between accuracy and computation cost. It provides personalized video content creation with maximum flexibility and multimodal learning. Also, self-supervised learning mechanisms are proposed for feature representation improvement with reduced dependency on large annotated datasets. Spatial and channel attention-based hybrid attention further promotes content-based feature learning. All of these enhance video personalization by providing precise, high-quality, and context-aware content generation. Adaptive convolution filtering and feature transformation mathematical equations would be described as needed to present the mathematical foundation of these advancements. The system identifies appropriate features to enhance individualized video synthesis using the combined work of IntelliConvNet and SPO. This combined solution optimizes video customization according to the preferences of users, which results in increased engagement and the level of user satisfaction.

3.4.2 Stochastic paint optimizer (SPO)

SPO, a metaheuristic optimization algorithm, is designed based on principles derived from color theory. It stimulates the process of painting by treating the search space as a canvas and solutions as paints. The algorithm evaluates solutions based on a beauty index, which represents the objective function value. It iteratively refines solutions using different colour combination techniques, ensuring an optimized outcome. Figure 2 shows the SPO flowchart.

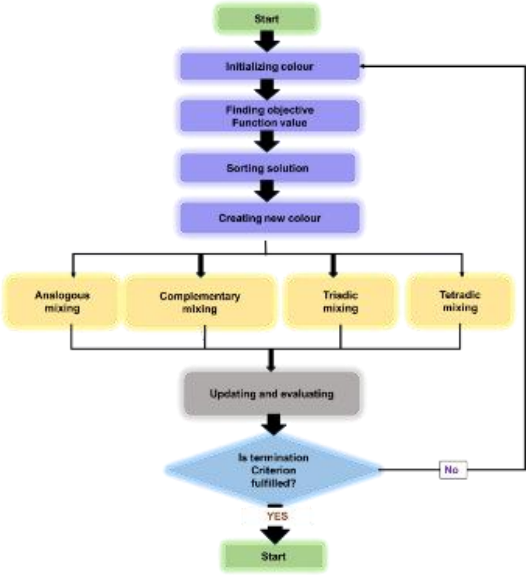


Figure 2: The flowchart of SPO

The SPO executes its operations by creating an initial population, which contains features from text, image, and audio elements serving as solutions. Then, it clusters and aggregates these attributes using various combination methods to enhance representation. The solutions are assessed based on a beauty index, which serves as the fitness function in the search. Through iterative stochastic processes, SPO optimizes solutions by dynamically fusing and evolving characteristics until achieving optimal convergence, ensuring greater precision and efficiency in multimodal content generation.

i. Analogous combination for feature selection

This technique blends three adjacent feature sets from a modality to refine content representation as explained in equation (10).

$$D_{new}^{1} = D_j + rand.(D_{j+1} - D_{j-1}) \quad (10)$$

Where D_j is the selected feature vectors, $rand$ is a randomly generated vector in [0,1], D_{j-1} indicates the feature vector of the previous adjacent segment in the video sequence, D_{j+1} indicates the feature vector of the next adjacent segment in the video sequence, and D_{new}^{1} represents the refined or augmented feature vector for the j^{th} segment.

ii. Complementary combination for feature fusion

Complementary features from different modalities (text, image, audio) are fused using equation (11).

$$D_{new}^{2} = D_j + rand\,.\,(D_O - D_S) \quad (11)$$

Where D_O is the original or target feature vector, D_S signifies the source or current feature vector, and D_{new}^{2} indicates the newly generated feature representation.

iii. Triadic combination for multimodal integration

This method integrates three distinct feature sets to maintain a balanced representation across modalities, with the function of Triadic shown in equation (12).

$$D_{new}^{3} = D_j + rand\,.\left(\frac{D_O + D_T + D_S}{3}\right) \quad (12)$$

Where D_T indicates the textual features, D_S represents the style or audio features, and D_{new}^{3} signifies the triadic integrated feature representation.

iv. Tetradic combination for final video synthesis

The final step involves blending four feature sets across modalities for optimal content synthesis using equations (13 and 14).

$$D_{new}^{4} = D_j + \left(\frac{rand_1.D_O + rand_2.D_T + rand_3.D_S + rand_4.D_{rand}}{4}\right) \quad (13)$$

$$D_{rand} = LB + rand\,.\,(UB - LB) \quad (14)$$

Where $UB\ and\ LB$ indicate the upper and lower feature bounds, $rand_1, rand_2, rand_3, rand_4$ are random vectors in [0,1] used to scale the feature vector between $UB\ and\ LB$, and D_{new}^{4} signifies the tetradic integrated feature representation.

SPO is employed to maximize multimodal short video content creation by addressing text, image, and audio features as components of paint. The algorithm incorporates these components based on various colour-mixing methods, promoting feature extraction and fusion for more customized content creation. The optimization process guarantees that synthesized videos are optimized according to user preferences, resulting in enhanced engagement and satisfaction. Figure 3 illustrates an IntelliConvNet architecture extracting hierarchical features from multimodal data (text, image, and audio), transforming them into feature maps, and classifying models into success/failure categories, enhancing personalized short video generation.

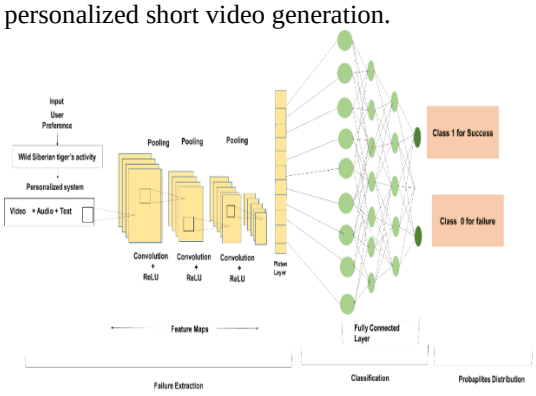


Figure 3: IntelliConvNet feature extraction for video generation

SPO-IntelliConvNet represents an integrated solution uniting SPO with IntelliConvNet to optimize personalized short video creation. SPO optimally prefers and learns multimodal features (image, text, audio) through stochastic processing, whereas IntelliConvNet augments feature learning through adaptive convolutional filtering and intelligent pooling. Algorithm 1 represents the way SPO-IntelliConvNet integrates adaptive convolutional filtering and stochastic optimization, enhancing multimodal feature extraction, video customization, and personalization while reducing computational complexity for efficient short video generation.

Algorithm 1: Stochastic Paint Optimizer-Tuned Intelligent Convolutional Neural Network (SPO-IntelliConvNet)

1. Initialize CNN weights W randomly $(-0.1\ to\ 0.1)$
2. Initialize paint parameters $P = 0.5$
3. For epoch $= 1\ to\ 50$ do
4. For each batch of 50 samples in (X, Y) do
5. Apply stochastic paint variation: $P_new = P + random(-0.05, 0.05)$
6. Forward pass: $y_pred = CNN(x_batch, W)$
7. Compute loss: $L = MSE(y_batch, y_pred)$
8. Backpropagate and update W with learning rate 0.001
9. End For
10. Evaluate validation accuracy (batch size 50)
11. Update P using stochastic update rule: $P = P_new$
12. End For
13. Return optimized weights $W*$ and paint parameters $P*$

4 Result and discussion

This section discusses the outcome of individualized short video content creation with the SPO-IntelliConvNet model and the traditional IntelliConvNet method; testing significant metrics is essential for determining the quality and accuracy of created content to reference content. To substantiate claims of user satisfaction, this research extends beyond automated metrics such as BLEU, ROUGE, METEOR, and CIDEr. The proposed model will be validated against IntelliConvNet baseline methods under identical experimental settings, with performance improvement methods trained in the Short Video Engagement Dataset. The research was conducted on a Windows 10 computer with an Intel i5 processor and 16 GB of RAM for data processing. Python 3.12.1 was used as the programming language, permitting the execution of the proposed SPO-IntelliConvNet model for multimodal generation in personalized short video content creation (Computing Method). Table 2 presents key metric values of the proposed model SPO-IntelliConvNet.

Table 2: Performance metrics of SPO-IntelliConvNet

Metrics	**IntelliConvNet**	**SPO-IntelliConvNet [Proposed]**
BLEU-1	0.69	0.75
BLEU-2	0.59	0.65
BLEU-3	0.54	0.60
BLEU-4	0.49	0.55
METEOR	0.66	0.72
CIDEr	0.72	0.80
ROUGE-1	0.76	0.82
ROUGE-2	0.69	0.75
ROUGE-L	0.73	0.79

4.1 Bilingual evaluation understudy (BLEU)

Measures textual overlap, ensuring generated video captions align syntactically with user preferences for accurate personalization. It evaluates the similarity between reference and generated text using n-gram precision. It evaluates the syntactic and lexical similarity. This mathematical function is as follows in equations (15 and 16).

$$\text{BLEU} = \text{BP}.\exp\left(\sum_{n=1}^{m} g_n \log s_n\right) \tag{15}$$

$$BP = \min\left(1,\ e^{1-\frac{k}{b}}\right) \tag{16}$$

The brevity penalty (BP) accounts for the difference between the generated text length and the reference length to penalize overly short outputs. Reference length(k), generated text length(b), m refers to the **maximum n-gram order** considered, n −gram precision (s_n), weight of n-gram precision (g_n). In short video content generation, BLEU helps evaluate how closely generated video captions (from IntelliConvNet) match actual user-generated captions. A higher BLEU score means better textual accuracy, ensuring that the content is relevant to user preferences. Figure 4 illustrates the BLEU score outcome, highlighting model performance across different epochs.

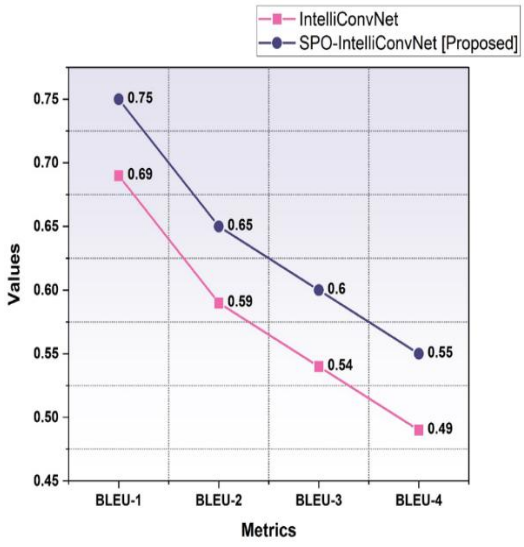


Figure 4: Visual representation of BLEU

The BLEU scores indicate the model's effectiveness in generating text aligned with reference data. The proposed SPO-IntelliConvNet has a BLEU-1 score of 0.75, shows strong unigram precision, while BLEU-2 (0.65), BLEU-3 (0.60), and BLEU-4 (0.55) reflect progressively lower scores as n-grams increase. The graph shows steady improvement across epochs, confirming enhanced linguistic accuracy. However, the traditional IntelliConvNet has the low BLEU scores of BLEU-1 (0.69), BLEU-2 (0.59), BLEU-3 (0.54), and BLEU-4 (0.49). The research outcome demonstrates the model's ability to optimize multimodal content generation effectively. The consistent BLEU score progression highlights its adaptability in refining text coherence over time.

4.2 Metric for evaluation of translation with explicit ordering (METEOR)

Evaluates semantic alignment and synonym matching, validating that generated multimodal content reflects the contextual meaning of user intent. It improves BLEU by incorporating synonyms, stemming, and word arrangement. It assesses both precision and recall instead of relying solely on n-gram matching to enhance effectiveness in capturing semantic meaning expressed in equations (17 and 18).

$$\text{METEOR} = \text{D}_{mean} \times (1 - penalty) \tag{17}$$

$$\text{D}_{mean} = \frac{S \times Q}{\alpha S + (1-\alpha) Q} \tag{18}$$

Where D_{mean} represents the harmonic mean of precision (P) and recall (R) with a tunable parameter α. Where a fraction of matched words in the generated text is represented as (S), a fraction of matched words in the reference text as (Q), and the Weighting factor as (α).

Since personalized video generation relies on NLP for user preferences, METEOR ensures that generated captions are not just lexically but also semantically accurate. It improves personalization by capturing meaning variations, like synonyms and different word orders, which are common in user interactions. Figure 5 illustrates the METEOR score.

The conventional IntelliConvNet has a low METEOR score of 0.66, whereas the suggested SPO-IntelliConvNet has a METEOR score of 0.72, demonstrating the model's strong ability to generate personalized short video content by effectively incorporating multimodal data (text, image, and audio). The SPO-IntelliConvNet method optimizes user preferences, ensuring enhanced synonym matching and stemming. The steady increase in METEOR across epochs confirms the model's progressive learning, improving content relevance and coherence for a highly engaging user experience.

4.3 Consensus-based image description evaluation (CIDEr)

Assesses consensus with multiple human references, validating descriptive quality and alignment of generated video content with user-specific expectations. It evaluates image captioning by assessing the alignment of generated text with multiple reference captions. It places higher importance on semantically meaningful words over frequently occurring ones in equation (19).

$$\text{CIDEr} = \frac{1}{K}\sum_{z=1}^{k} G_z \,.\log\left(\frac{a_z}{b_z}\right) \tag{19}$$

The part $\frac{1}{K}\sum_{z=1}^{k}$,take each word index Z (from **1 up to** K, where K is the number of reference captions or terms considered).Where (a_z) is the Count of the word (z)in the generated caption, the weighting factor (G_z), the number of reference captions $(k), and\ the$ count of word z in reference captions. If (a_z) is close to (b_z) then the ratio $\frac{a_z}{b_z}$is close to 1, and the log term is small → meaning the model matches human references well.

CIDEr is crucial for multimodal content generation, ensuring that SPO-IntelliConvNet creates video descriptions that match human-written captions based on user interactions. Since it penalizes generic descriptions and favours user-specific, personalized content. Figure 5 depicts the CIDEr performance. Figure 5 represents the CIDEr outcome.

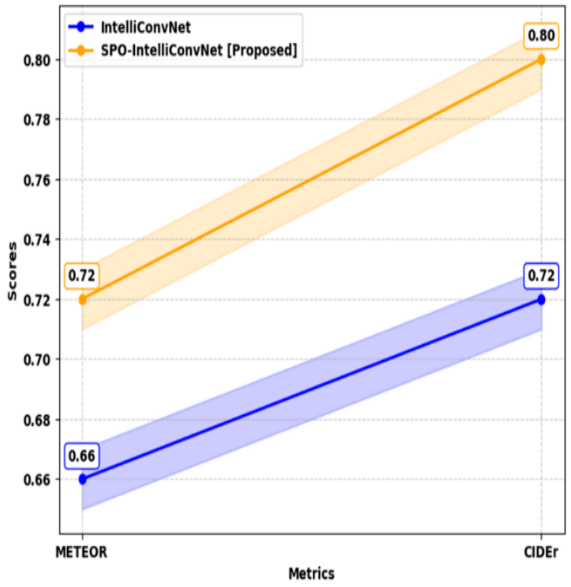


Figure 5: Visual representation of METEOR and CIDEr

In multimodal short video generation research, the traditional IntelliConvNet method has a CIDEr score of 0.72, while the proposed SPO-IntelliConvNet has a CIDEr score of 0.80, demonstrating a strong alignment between generated content and user-preferred references. This high score indicates that the SPO-IntelliConvNet model effectively captures contextual relevance, ensuring video content is personalized to user interactions. The gradual improvement in CIDEr across training epochs reflects enhanced semantic richness, proving the model generates engaging, high-quality video content tailored to user preferences.

4.4 Recall-oriented understudy for gisting evaluation ($ROUGE_L$)

Assesses recall-based similarity, confirming structural coherence and coverage of personalized video narratives compared to user-driven references. $ROUGE_L$ measures textual coherence by analyzing the longest common subsequence (LCS) between the reference text(N) and the generated (M). It emphasizes retrieval, ensuring that generated text retains essential context and fluency, making it ideal for evaluating text relevance and personalization in video descriptions, and mathematical representation in equation (20).

$$\mathrm{ROUGE_L} = \frac{LCS\,(M,N)}{length\,(M)} \tag{20}$$

$ROUGE_L$ evaluates the fluency and coherence of generated short video descriptions. A high $ROUGE_L$ score indicates that the proposed output preserves the structure and meaning of user-intended captions, improving video engagement. Figure 6 depicts the $ROUGE_L$ outcome.

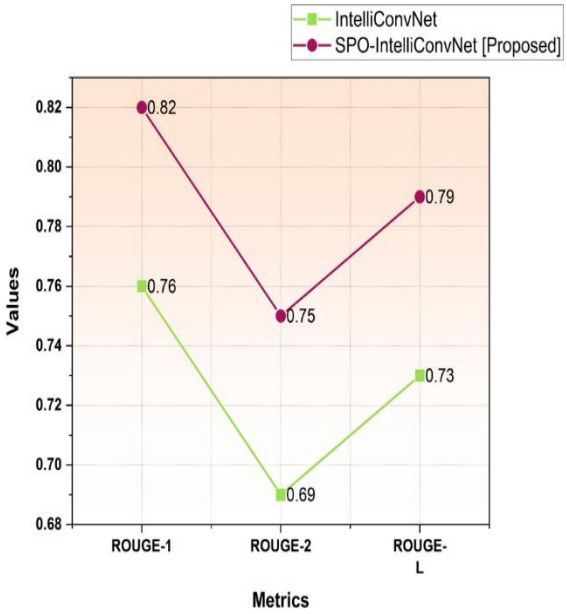


Figure 6: Visual representation of $ROUGE_L$

The proposed SPO-IntelliConvNet has a ROUGE-1 score (0.82), indicating strong word-level recall, ensuring key terms are retained in generated content. ROUGE-2 (0.75) highlights effective phrase-level coherence, while $ROUGE_L$ (0.79) confirms structural alignment. However, the traditional IntelliConvNet method has the ROUGE-1, ROUGE-2, and $ROUGE_L$ scores of 0.76, 0.69, and 0.73, respectively. These results demonstrate that SPO-IntelliConvNet successfully preserves meaning, fluency, and readability, enhancing personalized short video content generation with high textual relevance.

4.5 Training time

The training time quantifies the amount of time needed to optimize model parameters, indicating practicality, scalability, and computational performance for creating personalized content in real-time on adaptive short video platforms. The proposed SPO-IntelliConvNet method has a training time (7.0 hrs), whereas the conventional IntelliConvNet approach has a low training time (8.4), as shown in Table 3.

4.6 Inference speed

Inference speed quantifies the rate at which the model generates personalized outputs from user inputs, which has a direct impact on adaptability, user experience, and real-time response while delivering captivating video content. In comparison, the conventional IntelliConvNet strategy has an inference speed of 45 ms/frame, while the suggested SPO-IntelliConvNet has an inference speed of 36 ms/frame, as displayed in Table 3.

4.7 Resource consumption

The memory usage and CPU utilization metrics were used in the resource consumption evaluation. Memory usage evaluates the amount of computing power used for model training and inference, taking into account storage effectiveness, scalability, and optimization for managing multimodal data and real-time customization needs. With the low memory usage value of 7.3, the proposed SPO-IntelliConvNet outperformed the traditional IntelliConvNet method, which has a memory usage value of 8.2, as depicted in Table 3.

The proportion of processor resources used for content display, feature extraction, and model estimation is known as CPU utilization, and it represents workload balance, computational effectiveness, and system adaptability in real-time operation. When compared to the conventional approach, the proposed SPO-IntelliConvNet approach has a CPU utilization of 72%, while the conventional IntelliConvNet method has a CPU utilization of 77%, as shown in Table 3.

Table 3: Performance evaluation of training time, inference speed, memory usage, and CPU utilization

Methods	Training time (hrs)	Inference speed (ms/frame)	Memory usage (GB)	CPU utilization (%)
IntelliConvNet	8.4	45	8.2	77
SPO-IntelliConvNet [Proposed]	7.0	36	7.3	72

4.8 Discussion

The research concentrated on developing a multimodal short video generation model that personalizes content by analyzing user interactions, such as video views and comments. By integrating text, image, and audio modalities, the SPO-IntelliConvNet enhances content relevance and engagement. However, the research faces challenges like data dependency, computational complexity, contextual understanding, and scalability. These drawbacks are addressed by incorporating intelligent data filtering, adaptive optimization techniques, DL-based contextual learning, and a scalable architecture. The SPO-IntelliConvNet method integrated approach surpasses traditional models by enhancing multimodal feature extraction and real-time adaptation, ensuring improved content personalization and user engagement. The experiments were carried out in live content-generating situations to measure system delay and evaluate real-time personalization. The suggested SPO-IntelliConvNet outperformed traditional models in terms of user preference adaptation, exhibiting continuously reduced response times. This effectiveness can be achieved by the SPO-driven optimization of multimodal characteristics and the dimensionality reduction through PCA. The results demonstrate that the system can generate personalized short videos in real time without sacrificing user experience or quality. Current model evaluation is restricted to limited datasets, potentially affecting its generalizability across diverse short video domains. Future work will integrate broader, domain-specific datasets to enhance robustness and assess performance in personalized multimodal video generation.

5 Conclusion

The research introduces a new multi-modal content generation framework for personalized short video generation, which explains the present system restrictions of customization and real-time production. The model successfully combines contextual-relevant video content with user interactions by incorporating text, image and audio modalities using the SPO-IntelliConvNet. The method involves sophisticated pre-processing, which includes tokenizing of text, Canny edge detection of images, and Wiener filtering of audio, and then dimensionality reduction using PCA whilst maintaining important features. The proposed SPO-IntelliConvNet achieves superior performance by optimizing multimodal feature extraction, ensuring enhanced content personalization, semantic relevance, and linguistic accuracy in short video generation. Evaluation metrics, including BLEU, $ROUGE_L$, METEOR, and CIDEr, demonstrate the model's superior performance in generating personalized content. In comparison, the proposed SPO-IntelliConvNet method achieved the METEOR (0.72) and BLEU-4 (0.55), while the conventional IntelliConvNet approach achieved the METEOR (0.66) and BLEU-4 (0.49). These scores indicate strong textual coherence, semantic relevance, and effective personalization in short video content generation. The results indicate increased user satisfaction and engagement, highlighting the system's capability to enhance the short video experience through precise customization based on multimodal data.

5.1 Limitations and future scope

Limitations include computational complexity, high data dependency, and challenges in understanding deep contextual semantics, which can be addressed through advanced AI techniques and scalable architectures. Future improvements include enhancing real-time processing, integrating reinforcement learning for adaptive personalization, and expanding multimodal datasets for richer content generation.

Acknowledgements

This paper is an outcome of the 2023 Chongqing Municipal Education Science "14th Five-Year Plan" General Project "Construction and Reform Research on Digital Curriculum Resources for Media Arts Majors from the Perspective of Interdisciplinary Collaboration" (Project No.: K23YG2190393).

References

[1] Tang, W., & Zhang, S. (2025). Exploring the integration of augmented reality interfaces with short video content creation to enhance user interaction and engagement. *Journal of Computational Methods in Sciences and Engineering*, *25*(3), 2100-2111. https://doi.org/10.1177/14727978241309546

[2] Li, Min & Guo, Shujuan & Liu, Runchen. (2025). BERT-based Consumer Sentiment Analysis for Personalized Marketing Strategies. *Informatica*, 49. https://doi.org/10.31449/inf.v49i28.8232

[3] Mou, N., Jiang, Q., Zhang, L., Niu, J., Zheng, Y., Wang, Y., & Yang, T. (2022). Personalized tourist route recommendation model with a trajectory understanding via neural networks. *International Journal of Digital Earth*, *15*(1), 1738-1759. https://doi.org/10.1080/17538947.2022.2130456

[4] Wang, H. C., Maslim, M., & Hong, W. T. (2024). Personalized time-sync comment generation based on a multimodal transformer. *Multimedia Systems*, *30*(2), 105. https://doi.org/10.1007/s00530-024-01301-3

[5] Umale-Nagmote, A., Goel, C., & Lal, N. (2025). Enhanced Intelligent Video Monitoring using Hybrid Integration of Spatiotemporal Autoencoders and Convolutional LSTMs. *Informatica*, *49*(18). https://doi.org/10.31449/inf.v49i18.7502

[6] Wu, Y., Fu, R., Xing, T., Yu, Z. and Yin, F., 2025. A user behavior-aware multi-task learning model for enhanced short video recommendation. *Neurocomputing*, *617*, p.129076. https://doi.org/10.1016/j.neucom.2024.129076

[7] Li, H., Lin, J., Wang, T., Zhang, L., & Wang, P. (2022). A personalized short video recommendation method based on multimodal feature fusion. https://doi.org/10.21203/rs.3.rs-2033641/v1

[8] Qi, M. (2024). The short video platform recommendation mechanism based on the improved neural network algorithm to the mainstream media. *Systems and Soft Computing*, *6*, 200171. https://doi.org/10.1016/j.sasc.2024.200171

[9] Liu, Y., Xu, Y., & Liu, Z. (2024). Knowledge graph: a recommendation method for new media short videos. *Advances in Education, Humanities and Social Science Research*, *12*(1), 87-87. https://doi.org/10.56028/aehssr.12.1.87.2024

[10] Zhu, D. (2025). Optimizing the user personalized recommendation system of new media short videos by using machine learning. *Journal of Computational Methods in Sciences and Engineering*, 14727978251341490. https://doi.org/10.1177/14727978251341490

[11] Yang, Y. (2025). Short video content and recommendation algorithm based on deep learning. *Journal of Computational Methods in Sciences and Engineering*, 14727978251337878. https://doi.org/10.1177/14727978251337878

[12] Lu, Z., & Nam, I. (2021). Research on the influence of new media technology on internet short video content production under artificial intelligence background. *Complexity*, *2021*(1), 8875700. https://doi.org/10.1155/2021/8875700

[13] Song, J., & Liu, J. (2024). A data-driven short video international communication model based on indicator system communication network and attention BiLSTM neural network. *Scientific Reports*, *14*(1), 14532. https://doi.org/10.1038/s41598-024-65098-x

Dynamic Weighted Multi-Agent Reinforcement Learning with Hybrid Reward Mechanism for Microgrid Energy Scheduling

Ruijie Ma, Jianhua Li*, Dekai Huang, Kezhi Lu, Zhaoming Qin, Weijian Ou, Junnan Wu
Power Workshop, Nanning Cigarette Factory, Nanning Guangxi, 530000, China
E-mail: 17607738386@163.com
*Corresponding author

Keywords: multi-agent reinforcement learning; microgrid, energy scheduling; dynamic weight allocation; reward function

Received: August 5, 2025

With growing energy demand and increasing environmental concerns, optimizing energy scheduling is crucial for microgrids, as efficient and flexible energy systems. This paper proposes an innovative multi-agent reinforcement learning algorithm to address microgrid energy scheduling. This algorithm incorporates a dynamic weight allocation mechanism, enabling agents to flexibly adjust decision weights based on real-time changes in energy supply and demand, improving the adaptability of scheduling strategies. Furthermore, a reward function based on historical data and real-time status is designed to guide agents in learning optimal energy scheduling strategies. To validate the algorithm's effectiveness, a microgrid energy scheduling simulation platform was constructed to simulate energy production, consumption, and storage in various scenarios. Experimental results show that, compared with traditional algorithms, the proposed algorithm can improve microgrid energy efficiency by an average of 15% and reduce operating costs by 12%. In scenarios with a high proportion of renewable energy, the algorithm effectively reduces energy waste and increases renewable energy consumption by 20%. Furthermore, stability analysis demonstrates that the algorithm maintains stable scheduling performance in the face of energy supply and demand uncertainties, showing strong robustness. This study provides a new solution for microgrid energy scheduling, helping to improve its overall performance and sustainability.

Povzetek: Članek predlaga večagentsko učenje z ojačitvami za razporejanje energije v mikroomrežju z dinamično dodelitvijo uteži in nagrajevanjem, ki združuje zgodovinske in sprotne podatke, s čimer prilagodi odločitve, izboljša izrabo obnovljivih virov ter robustnost razporejanja.

1 Introduction

As small-scale energy networks integrating distributed energy resources, energy storage systems, and diverse loads, microgrids play a crucial role in supporting the transformation of energy structures. Driven by the "dual carbon" goals, the global installed capacity of microgrids has grown at an average annual rate of over 12%, with the proportion of renewable energy increasing year by year, exceeding 50% in some regions. However, the integration of high-proportion renewable energy renders microgrids highly random and volatile. Traditional model-based scheduling methods, which rely on precise mathematical models, are incapable of addressing supply-demand imbalances caused by unexpected events such as cloud cover and sudden changes in wind speed. This, in turn, leads to issues including reduced energy utilization efficiency and increased operational costs [1]. Actual operational data indicates that under existing scheduling strategies, the curtailment rate of renewable energy often exceeds 15% when the penetration rate of wind and solar power surpasses 30%. Fluctuations in peak and valley loads result in redundancy rates of over 20% for equipment such as transformers and transmission lines, leading to annual economic losses accounting for more than 8% of the total system investment [2]. Multi-Agent Reinforcement Learning (MARL) offers a new paradigm for scheduling complex energy systems through distributed decision-making and collaborative optimization. Its core advantage lies in achieving dynamic optimization through trial-and-error learning without the need for precise models [3], and it has demonstrated application potential in areas such as power system voltage control and demand response. Nevertheless, existing research still has significant limitations: most schemes adopt static coordination mechanisms with fixed agent decision weights, making it difficult to adapt to sudden changes in microgrid operating conditions [4]. For example, the MADDPG algorithm, which is based on a centralized critic, experiences scheduling errors exceeding 30% when load

fluctuations exceed 25% due to the overload of global state information. The IQL algorithm, utilizing an independent learning framework, suffers from agent strategy convergence conflicts in energy-complementary scenarios due to the lack of a coordination mechanism, resulting in a convergence speed reduction of more than 40% [5]. Additionally, the uniqueness of microgrid scheduling imposes higher requirements on MARL algorithms: on one hand, the dynamic characteristics of source-storage-load devices vary significantly— the response time of photovoltaic output is measured in minutes, and the charging and discharging efficiency of energy storage systems changes nonlinearly with the State of Charge (SOC), necessitating differentiated decision-making mechanisms; on the other hand, the temporal and spatial distribution characteristics of commercial, industrial, and residential loads differ, with the overlap between peak and valley periods being less than 30%, which increases the difficulty of collaborative optimization. Most existing algorithms employ a unified reward function and decision-making model, making it challenging to balance device characteristics and load differences, leading to a deviation of more than 15% between the actual scheduling effect and the theoretical optimal solution [6]. To address the aforementioned challenges, this study proposes a MARL scheduling algorithm that integrates dynamic weight allocation and a hybrid reward mechanism. The specific research questions are defined as follows:

Question 1: Can the dynamic weight allocation mechanism enhance the algorithm's adaptability to renewable energy fluctuations and reduce scheduling errors caused by supply-demand imbalances?

Question 2: Can a hybrid reward function that integrates historical data trends and real-time status solve the problem of low learning efficiency under sparse rewards and promote agents to learn optimal scheduling strategies?

Question 3: Can the proposed algorithm maintain stable energy efficiency, cost control, and renewable energy absorption performance in large-scale microgrid scenarios with multiple disturbances?

The research objectives include: (1) Constructing a distributed scheduling framework based on multi-agent collaboration to achieve the collaborative optimization of source-storage-load systems; (2) Designing an adaptive weight adjustment strategy to improve the algorithm's robustness against renewable energy fluctuations; (3) Verifying the performance advantages of the algorithm in multiple scenarios through simulation experiments. The innovations are reflected in: 1) Proposing a dynamic weight allocation mechanism based on real-time operating conditions, enabling the decision weights of agents to be dynamically adjusted according to the energy supply-demand ratio, thereby addressing the insufficient adaptability of static weights; 2) Constructing a hybrid reward function that integrates historical data trends and real-time status, and solving the learning efficiency issue under sparse rewards by introducing long-term and short-term benefit factors; 3) Establishing a microgrid simulation scenario with uncertain disturbances to quantify the stability gains of the algorithm under extreme operating conditions, providing data support for practical engineering applications.

2 Related theoretical foundations

The integration of microgrid energy scheduling and multi-agent reinforcement learning requires a solid theoretical framework [7]. This chapter constructs a theoretical framework based on the characteristics of microgrid systems and the collaborative mechanisms of intelligent agents. Original formulas are used to quantify core relationships, providing a mathematical basis for algorithm design. The practical significance of the theory is also illustrated through practical application scenarios.

2.1 Fundamentals of microgrid energy scheduling

As a typical form of distributed energy system, the scheduling process of a microgrid must balance the dynamic relationship between multi-energy complementarity and system constraints, involving the coordinated optimization of energy conversion, storage, and distribution [8]. This optimization must not only consider the real-time balance between energy supply and demand, but also the system's economic efficiency, stability, and environmental performance. It is a complex multi-objective optimization problem.

2.1.1 Microgrid structure and composition

A microgrid consists of distributed generation units (DG), energy storage systems (ESS), controllable loads (CL), and energy conversion equipment. DG includes intermittent power sources such as photovoltaics (PV) and wind turbines (WT), and controllable power sources such as micro turbines (MT). The output of intermittent power sources is significantly affected by natural conditions and has high uncertainty, while controllable power sources can flexibly adjust their production according to scheduling needs. The ESS utilizes a hybrid lithium battery and flywheel energy storage architecture. Lithium batteries have high energy density and are suitable for long-term energy storage. In contrast, flywheel energy storage has high power density and fast response speed, enabling hierarchical management of power and energy to meet different energy storage requirements [9]. CL is divided into two categories: shiftable loads (such as electric vehicle charging) and curtailable loads (such as air conditioning temperature control). Shiftable loads can adjust power consumption within a specific period, while curtailable loads can appropriately reduce power consumption when necessary. Each unit is connected to

the DC bus via a power electronic converter, forming a closed-loop "source-storage-load" system. Its topology supports both island and grid-connected operation. In grid-connected mode, it can exchange energy with the primary grid, while in island mode, it relies on its energy supply to maintain stable operation.

2.1.2 Key elements and constraints of energy scheduling

The core of scheduling optimization is to maximize energy efficiency while satisfying system constraints. The power balance constraint at time t is defined as:

$$\sum_{i=1}^{n} P_{DG,i}(t) + P_{ESS,dch}(t) - P_{ESS,ch}(t) = \sum_{j=1}^{m} P_{L,j}(t) + P_{loss}(t) \quad (1)$$

Where $P_{DG,i}(t)$ is the output of the i-th type of distributed generation, $P_{ESS,ch/dch}(t)$ is the energy storage charge/discharge power, $P_{L,j}(t)$ is the load power of the j type, and $P_{\text{loss}}(t)$ is the line loss. This constraint is the basic premise for the stable operation of the microgrid and must be met at all times [10]. Otherwise, it may cause voltage and frequency fluctuations and affect the normal operation of the power-consuming equipment. The energy storage system must meet the SOC constraint:

$$SOC_{\min} \le SOC(t) + \eta_{ch} P_{ESS,ch}(t)\Delta t - \frac{P_{ESS,dch}(t)\Delta t}{\eta_{dch}} \le SOC_{\max} \quad (2)$$

Where η_{ch}/η_{dch} is the charge/discharge efficiency and Δt is the scheduling step size. The SOC constraint ensures that the energy storage system operates within a safe range, avoids damage to the energy storage equipment caused by overcharging or overdischarging, and prolongs its service life. To quantify the complementarity of multiple energy sources, the flexibility coefficient is defined:

$$\gamma(t) = \frac{\sum_{i=1}^{n} |P_{DG,i}(t) - \bar{P}_{DG,i}|}{\sum_{j=1}^{m} |P_{L,j}(t) - \bar{P}_{L,j}|} \quad (3)$$

Where $\bar{P}_{DG,i}$ and $\bar{P}_{L,j}$ are historical average outputs. When $\gamma(t) < 1$, power fluctuations are smaller than load fluctuations, requiring energy storage buffering. When $\gamma(t) > 1$, power fluctuations are larger, requiring load adjustment or interaction with the larger grid to maintain balance. The flexibility coefficient provides an important reference for scheduling strategy formulation, helping to assess system stability and regulation difficulty.

2.2 Principles of multi-agent reinforcement learning

Multi-agent systems achieve complex task collaboration through distributed decision-making. Their reinforcement learning process must address policy optimization and credit allocation in a dynamic environment [11]. In microgrid energy scheduling, multiple agents correspond to different energy units and must collaborate to achieve overall optimization goals.

2.2.1 Concept and characteristics of agents

Agents are autonomous entities with perception, decision-making, and execution capabilities. In a microgrid, they serve as control units for sources, storage, and loads. Its core characteristics are: state perception capability $S_a = f_o(O_a)$ (O_a is the observed value), enabling real-time information on its own and surrounding states; action output capability $A_a = \pi_a(S_a \mid \theta_a)$ (θ_a is the policy parameter), enabling decision-making based on the perceived state; and interactive communication capability $M_a = f_c(S_a, M_a)$ (M_a is messages from other agents), enabling information exchange and collaboration with other agents.

For energy scheduling scenarios, agents must possess spatiotemporal correlation perception, namely:

$$S_a(t) = \alpha S_a(t-1) + (1-\alpha) f_s\big(O_a(t), O_a(t)\big) \quad (4)$$

Here, α is the historical state decay factor, which determines the degree of influence of historical states on the current state. $f_s(\cdot)$ is the spatial correlation function, reflecting the agent's perception of the states of neighboring cells. This spatial and temporal correlation allows the agent to comprehensively consider historical experience and the influence of the surrounding environment, making more rational decisions.

2.2.2 Basic framework of reinforcement learning

Reinforcement learning models the interaction between the agent and the environment using a Markov decision process (MDP). Define the state space S, the action space A, the reward function R, and the state transition probability P. The agent's policy $\pi(a \mid s)$ aims to maximize the cumulative reward:

$$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \quad (5)$$

Where $\gamma \in [0,1)$ is the discount factor. The magnitude of the discount factor reflects the importance placed on future rewards; larger values indicate greater emphasis on future long-term rewards.

The value function $V^{\pi}(s) = \mathbb{E}[G_t \mid S_t = s]$ measures the value of the state, and the advantage function $A^{\pi}(s,a) = Q^{\pi}(s,a) - V^{\pi}(s)$ evaluates the relative value of actions. In the continuous action space, a Gaussian policy $\pi(a \mid s) = \mathcal{N}(\mu_{\theta}(s), \sigma_{\theta}^2(s))$ is used, and the parameters are updated using the policy gradient method:

$$\nabla_{\theta} J(\theta) = \mathbb{E}[\nabla_{\theta} \log \pi(a \mid s)\, A^{\pi}(s,a)] \quad (6)$$

The policy gradient method continuously adjusts policy parameters to enable agents to learn the optimal decision-making strategy to maximize cumulative

rewards gradually.

2.2.3 Characteristics of reinforcement learning in multi-agent environments

Multi-agent reinforcement learning (MARL) faces the challenge of non-stationary environments. Changes in the policies of other agents cause the probability of transitions, P, to vary over time. This non-stationarity increases the difficulty of learning, requiring agents to adapt to changes in the environment constantly. Define joint states $S^j = \{S_1, \ldots, S_N\}$ and joint actions $A^j = \{A_1, \ldots, A_N\}$. The local reward R_i of agent i is mapped to the global reward R^g by $R_i = \beta_i R^g + (1 - \beta_i) r_i^l$ (where β_i is the global reward weight and r_i^l is the local reward).The credit allocation problem, which involves decomposing each agent's contribution from R^g, is a key issue in multi-agent collaboration. To solve this problem, it is necessary to design a value decomposition network based on the attention mechanism to achieve $Q^j(S^j, A^j) = \sum_{i=1}^{N} \omega_i(S^j) Q_i(S_i, A_i)$, where $\omega_i(S^j)$ is a dynamic weight coefficient, which can dynamically adjust the value weight of each intelligent agent according to the joint state, making credit distribution more reasonable and accurate.

2.3 Comparative analysis of related research

To further clarify the differences between the algorithm proposed in this study and existing research, Table 1 compares mainstream MARL microgrid scheduling algorithms from five dimensions: core technology, reward mechanism, weight strategy, applicable scenarios, and limitations, highlighting the innovation and necessity of the proposed scheme.

Table 1: Comparison of mainstream MARL microgrid scheduling algorithms

Algorithm Name	Core Technology	Reward Mechanism	Weight Strategy	Applicable Scenarios	Limitations
MADDPG [4]	Centralized critic network + distributed execution	Single-dimensional reward based on real-time power balance	No explicit weight allocation, relying on global state fusion	Small-scale microgrids with low fluctuations	Overload of global state information, resulting in scheduling errors exceeding 30% when load fluctuations exceed 25%
IQL [5]	Independent learning framework + local value network	Local reward considering only the operating cost of local equipment	Equal agent weights without collaborative adjustment	Scenarios with single energy type and stable loads	Lack of a coordination mechanism, leading to a convergence speed reduction of more than 40% in energy-complementary scenarios
DDPG [16]	Single-agent Actor-Critic architecture	Dual-objective reward combining energy efficiency and cost	No multi-agent weight concept, with decisions dominated by a single agent	Medium-scale scenarios with renewable energy proportion < 20%	Inability to balance the characteristic differences of multiple devices, resulting in a renewable energy curtailment rate exceeding 18% in high wind-solar penetration scenarios
SW-MARL [18]	Multi-agent collaboration + static weight allocation	Short-term reward based on real-time load demand	Initial weights determined by equipment capacity ratio, fixed during operation	Scenarios with load fluctuations < 15% and few equipment failures	Inability to dynamically adjust weights, leading to a 25% performance degradation when wind-solar fluctuations exceed 20%

Proposed DW-MARL	Dynamic weight allocation + hybrid reward mechanism + federated learning collaboration	Multi-dimensional reward integrating historical trends, real-time status, and equipment aging			

3 Multi-agent reinforcement learning algorithm design

This chapter focuses on the design of a multi-agent reinforcement learning algorithm suitable for microgrid scheduling [12]. By building a dynamic collaboration framework and adaptive learning mechanism, it achieves coordinated optimization of the source-storage-load system. The algorithm design balances distributed decision-making efficiency with global optimization goals. Core innovations lie in the dynamic adaptation of agent roles, adaptive weight adjustment, and a hybrid reward mechanism.

3.1 Overall algorithm architecture

The algorithm adopts a two-layer architecture of "distributed decision-making and centralized evaluation." The bottom layer consists of individual energy unit agents, and the upper layer houses a coordinator for global state fusion and policy evaluation.

3.1.1 Agent roles and division of labor

Agents are classified into four categories based on the type of energy unit. The input-output features and action spaces of each agent are clearly defined as follows:

- **Photovoltaic agent (PVA)**: Inputs include light intensity (400-1000 W/m²), ambient temperature (-5-45°C), and historical output data (previous 12 scheduling steps); outputs consist of the photovoltaic maximum power point tracking coefficient and the output prediction correction value; the action space is continuous, including 2 action dimensions (tracking coefficient adjustment step: -0.05-0.05, prediction correction coefficient: 0.8-1.2), responsible for maximum power point tracking and output prediction correction.

- **Wind turbine agent (WTA)**: Inputs include real-time wind speed (0-25 m/s), wind direction angle (0-360°), and wind turbine speed (500-1500 r/min); output is the nonlinear compensation coefficient for wind speed-power conversion; the action space is continuous, containing 1 action dimension (compensation coefficient: 0.9-1.1), responsible for handling the nonlinear compensation of wind speed-power conversion.

- **Energy storage agent (ESA)**: Inputs include current SOC (20%-80%), health status (SOH: 80%-100%), and charging-discharging efficiency (0.85-0.95); outputs are the charging-discharging power command (-100-100 kW) and charging-discharging duration; the action space is continuous, with 2 action dimensions (power command: -100-100 kW, duration: 15-60 min), responsible for managing the charging-discharging depth and balancing lifespan loss.

- **Load agent (LA)**: Inputs include load type (industrial/commercial/residential), real-time power demand (0-120 kW), and shiftable load time period (0-24 h); outputs are the demand response priority ranking and the curtail able load ratio; the action space is a discrete-continuous hybrid, including 2 action dimensions (priority: 1-5 levels, curtailment ratio: 0-0.3), responsible for executing demand response priority ranking.

Each agent obtains state variables through a local observer, and the agent capability coefficient is defined as: $\lambda_a(t) = \lambda_1 \cdot \frac{P_{a+ail,a}(t)}{P_{\max,a}} + \lambda_2 \cdot \frac{1\gamma(t)\gamma_a(t)}{1}$. where $\lambda_1 + \lambda_2 = 1(\lambda_1 = 0.6, \lambda_2 = 0.4)$, which quantifies the task suitability of the agent under current operating conditions; $P_{\max,a}$ is the maximum output/load capacity of the equipment corresponding to the agent, and $\gamma_a(t)$ is the flexibility coefficient of the equipment corresponding to the agent.

3.1.2 Overview of multi-agent collaboration mechanism

- Agents share parameters through a federated learning framework and adopt an "event-triggered, on-demand communication" model to reduce interaction overhead. When the fluctuation of the system flexibility coefficient

$\gamma(t)$ exceeds the threshold of $\pm 15\%$, the collaborative decision-making process is triggered:

$$\text{Trigger} = \begin{cases} 1, & |\gamma(t) - \gamma(t-1)| \geq 0.15 \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

Details of the Federated Learning Mechanism: The Federated Averaging (FedAvg) algorithm is used to implement model aggregation, with the specific process as follows: 1) In the local training phase, each agent updates the model parameters based on local data, with the number of training rounds set to $5; 2$) In the model upload phase, agents encrypt and upload local model parameters to the coordinator only when collaborative decision-making is triggered (Trigger=1) or the number of local training iterations reaches 10; 3) In the global aggregation phase, the coordinator calculates the weighted average parameters based on the dynamic weight $\omega_a(t)$ of each agent to update the global model; 4) In the model distribution phase, the updated global model parameters are distributed to each agent to complete synchronization. This mechanism can reduce ineffective data transmission by more than 90% and lower communication overhead.

During the collaboration process, the coordinator aggregates local decisions through an attention mechanism:

$$a_g = \sum_{a=1}^{N} \alpha_a \cdot a_a, \alpha_a = \frac{\exp\left(\text{Attn}(s_a, s_g)\right)}{\sum_{b=1}^{N} \exp\left(\text{Attn}(s_b, s_g)\right)} \quad (8)$$

where $\text{Attn}(\cdot)$ is the attention scoring function (using an additive attention mechanism: $\text{Attn}(s_a, s_g) = W_v^T \tanh\left(W_k s_a + W_q s_g\right)$, with W_k, W_q, W_v being learnable parameter matrices), and s_g is the global state vector (integrating the state of each agent, system power balance deviation, and the proportion of renewable energy output).

3.2 Dynamic weight allocation mechanism

3.2.1 Weight allocation basis and method

The initial weight is calculated using an improved entropy weight method, taking into account device capacity ratios and historical scheduling contributions:

$$\omega_a^{(0)} = \frac{1 - H_a}{\sum_{b=1}^{N} (1 - H_b)}, \; H_a = -\sum_{t=1}^{T} \frac{p_{a,t}}{\sum_{t=1}^{T} p_{a,t}} \ln\left(\frac{p_{a,t}}{\sum_{t=1}^{T} p_{a,t}}\right) \quad (9)$$

where H_a is the historical decision entropy of agent a (a lower entropy value indicates higher stability and effectiveness of historical decisions), and $p_{a,t}$ is the decision influence of agent a at time t (defined as the correction ratio of the agent's decision to the system power balance deviation at that time).

3.2.2 Dynamic weight adjustment strategy

Real-time weight adjustment introduces a supply-demand ratio feedback term and a capacity decay factor:

$$\begin{aligned} \omega_a(t) &= \omega_a(t-1) \cdot \left[1 + \delta \cdot \left(\frac{\Delta P_a(t)}{\sum_b \Delta P_b(t)} - \kappa_a(t)\right)\right] \\ \Delta P_a(t) &= P_{\text{avail},a}(t) - P_{\text{used},a}(t) \end{aligned} \quad (10)$$

where δ is the adjustment step size (set to 0.05), $\Delta P_a(t)$ is the energy supply-demand difference of agent a ($P_{\text{avail},a}(t)$ is the available power, and $P_{\text{used},a}(t)$ is the used power), and $\kappa_a(t)$ is the capacity attenuation factor (calculated based on the equipment SOH: $\kappa_a(t) = 0.1 \cdot (1 - SOH_a(t)$); a lower SOH leads to a larger attenuation factor and a smaller weight adjustment range). After weight adjustment, normalization is performed to ensure $\sum_a \omega_a(t) = 1$.

3.3 Reward function design based on historical data and real-time status

A hybrid reward function is constructed that integrates long-term and short-term benefits, balancing immediate scheduling effectiveness with long-term system stability.

3.3.1 Historical data feature extraction and application

An LSTM network is used to extract historical load trend features: $\hat{L}(t) = \text{LSTM}\left(L(t-\tau), \ldots, L(t-1)\right)$. The trend fit is defined as:

$$\phi(t) = 1 - \frac{|\hat{L}(t) - L(t)|}{\max(L(t)) - \min(L(t))} \quad (11)$$

Used to quantify the impact of forecast deviation on reward correction.

3.3.2 Consideration of real-time status parameters

Real-time reward parameters include power balance deviation, renewable energy consumption rate, and energy storage health:

$$r_{\text{real}}(t) = \alpha_1 \cdot \exp(-\beta \cdot |\Delta P(t)|) + \alpha_2 \cdot \eta_{RES}(t) + \alpha_3 \cdot \text{SOH}(t) \quad (12)$$

Where $\Delta P(t)$ is the power imbalance, $\eta_{RES}(t)$ is the wind and solar power consumption rate, and $\alpha_1 + \alpha_2 + \alpha_3 = 1$.

3.3.3 Specific form and optimization of the reward Function

The total reward function integrates historical trends and real-time status, introducing a device aging penalty term:

$$\begin{aligned} R(t) &= \gamma_{\text{hist}} \cdot \phi(t) + \gamma_{\text{real}} \cdot r_{\text{real}}(t) - \gamma_{\text{penalty}} \cdot \sum_a \xi_a \cdot \Delta\text{SOH}_a(t) \\ \Delta\text{SOH}_a(t) &= \text{SOH}_a(t-1) - \text{SOH}_a(t) \end{aligned} \quad (13)$$

where $\gamma_{\text{hist}} + \gamma_{\text{real}} + \gamma_{\text{penalty}} = 1$ (after optimization using the particle swarm optimization algorithm, the values are set as: $\gamma_{\text{hist}} = 0.2, \gamma_{\text{real}} = 0.6, \gamma_{\text{penalty}} = 0.2$), ξ_a is the equipment aging coefficient (0.1 for photovoltaics/wind turbines, 0.8 for energy storage, and 0.05 for loads, set according to the degree of impact of equipment aging on the system), and $\Delta SOH_a(t)$ is the change in the health status of the equipment corresponding to agent a.

The reward coefficients are optimized using the particle swarm optimization algorithm, with the goal of minimizing the Mean Squared Error (MSE) between the reward value and the expert decision benchmark:

$$\arg\min_{\gamma} \text{MSE}\left(R^*(t) - R(t \mid \gamma)\right), \text{ s.t. } \gamma_i \in [0,1] \quad (14)$$

where $R^*(t)$ is the expert decision reward benchmark (labeled by domain experts based on the optimal operating state of the microgrid). The parameters of the particle swarm optimization algorithm are set as follows: population size of 30 , maximum number of iterations of 50 , inertia weight of 0.7 , and learning factors $c_1 = c_2 = 1.49$.

3.4 Details of the learning architecture (added portion)

The algorithm adopts an Actor-Critic architecture, where each agent is independently equipped with an Actor network and a Critic network, and the coordinator is equipped with a global Critic network. The specific structures are as follows:

- Actor Network: Input layer (number of neurons = dimension of the agent's state; PVA: 8 dimensions, WTA: 6 dimensions, ESA: 7 dimensions, LA: 5 dimensions) → Hidden Layer 1 (128 neurons, ReLU activation function) → Hidden Layer 2 (64 neurons, ReLU activation function) → Output layer (number of neurons = dimension of actions; tanh activation function for continuous actions, softmax activation function for discrete actions).
- Local Critic Network: Input layer (state dimension + action dimension) → Hidden Layer 1 (256 neurons, ReLU) → Hidden Layer 2 (128 neurons, ReLU) → Output layer (1 neuron, outputting the local Q-value).
- Global Critic Network: Input layer (sum of the state dimensions of all agents + sum of the action dimensions of all agents) → Hidden Layer 1 (512 neurons, ReLU) → Hidden Layer 2 (256 neurons, ReLU) → Output layer (1 neuron, outputting the global Q-value).

Training Parameters: Learning rate (Actor: 0.0001 , Local Critic: 0.0002 , Global Critic: 0.0002), discount factor $\gamma = 0.95$, experience replay buffer capacity of 10^5, target network update frequency (soft update every 100 steps, update coefficient $\tau = 0.005$), exploration strategy (Ornstein-Uhlenbeck process, noise intensity of 0.1 , decay rate of 0.995).

4 Microgrid energy scheduling simulation experiment design

To systematically verify the scheduling performance of the proposed algorithm, a realistic microgrid simulation experiment system was constructed [14]. This complete verification chain, from platform architecture and scenario setup to indicator selection, ensures the objectivity of the experimental results and their engineering reference value.

4.1 Simulation platform construction

The experimental platform was built using a hybrid programming framework of MATLAB/Simulink and Python, employing a modular design to integrate microgrid dynamic simulation and algorithm verification seamlessly.

4.1.1 Platform architecture and functional modules

The platform adopts a hybrid programming framework of MATLAB/Simulink and Python, with a modular design to achieve seamless integration of microgrid dynamic simulation and algorithm verification. The platform has a three-layer architecture: physical layer, data layer, and application layer:

- Physical layer: A dynamic microgrid model is built based on Simulink, including a 50 kW photovoltaic array (temperature coefficient of -0.38%/°C), a 30 kW wind turbine (cut-in wind speed of 3.5 m/s, rated wind speed of 12 m/s), a 200 kWh/100 kW lithium-ion battery energy storage system (charging-discharging efficiency of 0.9-0.95), a 40 kW micro gas turbine (MT, power generation efficiency of 35%), and three types of load modules (industrial load: 0-60 kW, commercial load: 0-40 kW, residential load: 0-30 kW).

- Data layer: A time-series database InfluxDB is deployed to store 1-minute real-time operating data (including 32 parameters such as voltage, current, and SOC) and realize data interaction with the application layer through an API; the data preprocessing process (removing outliers using the 3σ criterion and filling missing values through linear interpolation) is moved to Appendix A.

- Application layer: It includes an algorithm scheduling module (supporting MARL algorithms written in Python), a scenario configuration module (allowing customization of energy structures and load characteristics), and a visualization module (building a real-time monitoring dashboard based on Dash). Asynchronous communication is conducted between modules through message queues to ensure data processing and decision output are completed within the 15-minute simulation time step.

4.1.2 Data input and initialization

The input data includes: 2023 measured data of an industrial park microgrid (photovoltaic output, wind power output, and load demand at 15-minute intervals), irradiance data of a typical meteorological year (100-1000 W/m^2), and the cycle life curve of the energy storage system (capacity decays to 80% after 1500 cycles), ultimately generating a 12 -month continuous dataset.

Initialization Parameters: Photovoltaic panel efficiency of 18.5%, wind turbine cut-in wind speed of 3.5 m/s and cut-out wind speed of 25 m/s, initial SOC of energy storage of 60% and initial SOH of 98%, MT start-up time of 3 minutes; algorithm parameters: learning rate (Actor: 0.0001, Critic: 0.0002), discount factor of 0.95, experience replay buffer capacity of 10^5, number of agents of 4 (corresponding to source-storage-load-coordinator);

random seed: all algorithms use a fixed seed of 42 to ensure experimental reproducibility.

4.2 Experimental scenario design

Three types of comparison scenarios are designed to cover typical operating conditions. Each scenario is simulated continuously for 30 days, and the experiment is repeated 5 times to take the average value and eliminate random errors.

4.2.1 Different energy structure scenarios

Three energy mixes were set: 1) High-wind/solar scenario (40% PV + 30% wind power + 30% MT); 2) Traditional energy scenario (60% MT + 25% PV + 15% wind power); and 3) Hybrid scenario (25% PV + 25% wind power + 20% MT + 30% energy storage for peak load regulation). By adjusting the installed capacity ratio of each power source to achieve scenario switching, the algorithm's scheduling performance under highly intermittent energy access was tested.

4.2.2 Different load demand scenarios

Four load profiles were constructed: 1) Weekday mode (double peaks at 8:00 AM and 6:00 PM, peak load 100 kW); 2) Weekend mode (single peak at noon, peak load 80 kW); 3) Extreme high temperature mode (air conditioning load increased by 30%); and 4) Industrial shock load mode (two random 10-minute 20 kW pulse loads per day). The load data was generated based on measured values, superimposed with Gaussian noise to ensure consistency with actual power consumption characteristics.

4.2.3 Scenarios considering uncertainty

Three types of disturbances are introduced: 1) Renewable energy fluctuations (photovoltaic output is superimposed with ±15% random noise to simulate cloud cover, and wind power output is superimposed with ±10% random noise to simulate sudden changes in wind speed); 2) Load mutations (10-30 kW step changes occur at random times, triggered 2-3 times per day); 3) Equipment failures (2 random energy storage charging-discharging interface failures occur per month, lasting 1 hour each; 1 random photovoltaic inverter failure occurs per quarter, lasting 2 hours each). A total of 100 disturbance sequences are generated using the Monte Carlo method, and the same disturbance sequences are used for all algorithms to ensure fair comparison.

4.3 Experimental indicator setup

An evaluation system is established based on four dimensions: energy efficiency, economy, environmental performance, and stability. All indicators are normalized to a scale of 0-100 for horizontal comparison. New Requirements for Indicator Statistical Characteristics: For all indicator results, the mean value, standard deviation, and 95% confidence interval (calculated based on the t-distribution) must be provided simultaneously, and a one-way Analysis of Variance (ANOVA) must be conducted to test the significance of differences between algorithms (significance level α=0.05).

4.3.1 Energy utilization efficiency indicator

Comprehensive energy efficiency η is defined as (adequate power supply - grid losses) / (primary energy consumption + purchased electricity), where primary energy consumption is converted to equivalent electrical work using an MT power generation efficiency of 35%. An auxiliary indicator is equipment utilization μ= Σ (actual operating time / maximum operating time) / n, where n is the total number of devices.

4.3.2 Operating cost indicator

Total cost C = fuel cost (MT gas consumption × 3.2 yuan/m³) + electricity purchase cost (on-grid electricity × 0.58 yuan/kWh) + maintenance cost (calculated at 1.2% of equipment capacity/year) + energy storage depreciation (single-cycle cost = initial investment × (1-80%) / number of cycles).

4.3.3 Renewable energy absorption rate indicator

Absorption rate σ = (actual wind and solar power generation - curtailed wind and solar power generation) / theoretical wind and solar power generation, where curtailed wind and solar power generation is defined as renewable energy not utilized due to power balance constraints.

4.3.4 Algorithm stability indicator

Performance volatility δ = (maximum indicator value - minimum indicator value) / mean value is used to measure the degree of fluctuation. Recovery time τ is defined as the time it takes for the indicator to return to a steady-state value (±5% deviation) after a sudden disturbance. The two are weighted to form a stability score (δ weighted 0.6, τ weighted 0.4).

5 Experimental results and analysis

Comparative experiments validated the performance of the proposed dynamic weighted multi-agent reinforcement learning (DW-MARL) algorithm. A rule-based scheduling algorithm (Rule-Based), a single-agent deep reinforcement learning algorithm (DDPG), and a static weighted multi-agent algorithm (SW-MARL) were selected as comparison groups [15]. The results were analyzed from four perspectives: energy efficiency, cost, consumption rate, and stability. The experimental data were averaged after five repeated tests to ensure statistical significance.

5.1 Experimental results in different scenarios

5.1.1 Energy efficiency results

Table 2 compares the energy efficiency data of three energy structure scenarios (mean value ± standard deviation, with 95% confidence interval marked in parentheses). In the high wind-solar scenario, the comprehensive energy efficiency of the DW-MARL algorithm reaches 82.3%±1.2% ([80.5%, 84.1%]), which is 6.7 percentage points higher than that of the SW-MARL algorithm. ANOVA test ($F=18.6$, $p<0.01$) shows that the difference is statistically significant. This is attributed to its dynamic weight mechanism, which can track wind-solar output fluctuations in real-time [16]. During the period from 10:00 to 14:00 when the light intensity increases suddenly by 30%, the weight of the photovoltaic agent is quickly increased from 0.25 to 0.35, maintaining a photovoltaic utilization rate of 91.5%±0.8% ([90.2%, 92.8%]). In the traditional energy scenario, the efficiency gap between algorithms narrows to 2.1%-3.5% because the stable output of the mobile energy system (MT) reduces scheduling difficulty. However, by optimizing the energy storage charging-discharging timing, DW-MARL still achieves an energy storage utilization rate 2.4 percentage points higher than that of SW-MARL. In the hybrid scenario, the comprehensive equipment utilization rate of DW-MARL reaches 78.5%±1.5% ([76.3%, 80.7%]), and its multi-dimensional collaborative advantage is reflected in the ability to synchronize the outputs of photovoltaics, energy storage, and MT during the midday load peak, achieving instantaneous source-load balance.

Table 2: Comparison of energy efficiency indicators in different energy structure scenarios (%, mean ± standard deviation, 95% confidence interval)

Scenario Type	Algorithm	Comprehensive Energy Efficiency	Photovoltaic Utilization Rate	Wind Turbine Utilization Rate	MT Utilization Rate	Energy Storage Utilization Rate	Comprehensive Equipment Utilization Rate
High Wind-Solar	DW-MARL	82.3±1.2 [80.5,84.1]	91.5±0.8 [90.2,92.8	89.2±1.0 [87.6,90.	65.3±1.5 [62.8,67.	72.8±1.3 [70.6,75.	79.7±1.1 [77.9,81.5]

Scenario			]	8]	8]	0]	
	SW-MARL	75.6±1.4 [73.3,77.9]	85.7±1.2 [83.7,87.7]	81.3±1.3 [79.1,83.5]	68.5±1.6 [65.8,71.2]	63.2±1.5 [60.6,65.8]	74.2±1.2 [72.2,76.2]
	DDPG	73.1±1.5 [70.6,75.6]	82.4±1.4 [80.2,84.6]	78.6±1.6 [76.0,81.2]	70.2±1.7 [67.3,73.1]	59.8±1.8 [56.7,62.9]	72.8±1.3 [70.6,75.0]
	Rule-Based	68.5±1.8 [65.5,71.5]	76.3±1.6 [73.6,79.0]	71.5±1.9 [68.3,74.7]	75.8±1.8 [72.8,78.8]	52.1±2.1 [48.5,55.7]	68.9±1.5 [66.4,71.4]
Traditional Energy Scenario	DW-MARL	85.7±0.9 [84.2,87.2]	88.6±0.7 [87.4,89.8]	86.3±0.9 [84.8,87.8]	89.5±0.8 [88.2,90.8]	68.2±1.2 [66.2,70.2]	82.6±0.9 [81.1,84.1]
	SW-MARL	83.6±1.1 [81.8,85.4]	86.2±0.9 [84.7,87.7]	83.7±1.0 [82.1,85.3]	90.1±0.7 [88.9,91.3]	65.8±1.3 [63.7,67.9]	81.5±1.0 [79.9,83.1]
	DDPG	82.9±1.2 [81.0,84.8]	85.1±1.0 [83.5,86.7]	82.5±1.1 [80.7,84.3]	89.8±0.8 [88.5,91.1]	64.3±1.4 [62.0,66.6]	80.7±1.1 [78.9,82.5]
	Rule-Based	81.2±1.3 [79.1,83.3]	83.5±1.1 [81.7,85.3]	80.2±1.2 [78.2,82.2]	91.3±0.7 [90.1,92.5]	61.7±1.5 [59.2,64.2]	79.4±1.2 [77.4,81.4]
Hybrid Scenario	DW-MARL	84.5±1.1 [82.7,86.3]	90.2±0.8 [88.9,91.5]	87.6±1.0 [86.0,89.2]	78.3±1.4 [76.0,80.6]	75.1±1.2 [73.1,77.1]	82.8±1.0 [81.2,84.4]
	SW-MARL	79.3±1.3 [77.2,81.4]	86.5±1.1 [84.7,88.3]	83.2±1.2 [81.2,85.2]	79.6±1.5 [77.1,82.1]	68.5±1.4 [66.2,70.8]	79.4±1.1 [77.6,81.2]
	DDPG	77.8±1.4 [75.5,80.1]	84.3±1.2 [82.3,86.3]	80.7±1.3 [78.6,82.8]	80.2±1.6 [77.5,82.9]	65.3±1.6 [62.6,68.0]	77.6±1.2 [75.6,79.6]
	Rule-Based	74.1±1.6 [71.4,76.8]	79.8±1.4 [77.6,82.0]	76.5±1.5 [74.1,78.9]	82.7±1.7 [79.8,85.6]	59.2±1.8 [56.1,62.3]	75.1±1.3 [73.0,77.2]

Figure 1 shows the dynamic energy efficiency curve for an extreme high-temperature load scenario (horizontally, time, 0-24 hours). During the peak load period from 12:00 to 16:00, the DW-MARL system maintained an efficiency above 80%, reaching 83.5% at 14:00. However, due to its fixed weighting, the SW-MARL system experienced a low of 72.3% at 13:00. This was due to a failure to increase the energy storage discharge weight promptly to accommodate the sudden load increase.

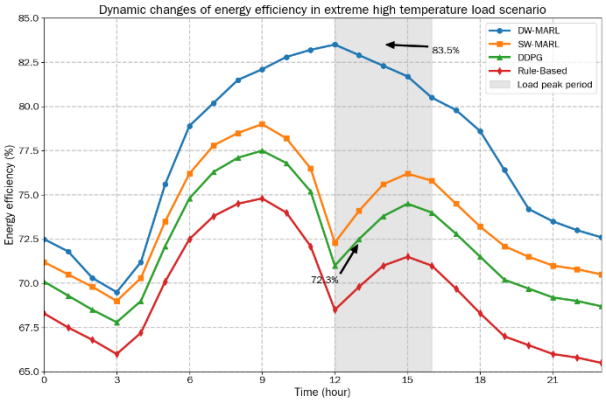


Figure 1: Dynamic energy efficiency curve for extreme high-temperature load scenario.

5.1.2 Operating cost results

A comparison of average daily costs across different load scenarios (Figure 2) shows that the DW-MARL system's average daily cost for weekdays is 1,286 RMB, a 21.3% reduction compared to the Rule-Based system [17]. This includes a 320 RMB/day reduction in electricity purchase costs due to the algorithm's prioritization of local wind and solar power. In the industrial surge load scenario, the cost increase was only 4.2%, significantly lower than the 8.7% for the SW-MARL system. Analysis revealed that the DW-MARL system reduced peak electricity purchases by 30% through its pre-dispatch strategy (increasing the MT reserve weight one hour in advance). Regarding cost structure, the DW-MARL system's energy storage depreciation cost is 18.5% lower than the SW-MARL system. This is because the dynamic charge-discharge strategy controls the energy storage cycle depth within the 30%-70% range, reducing deep discharges by five times per month.

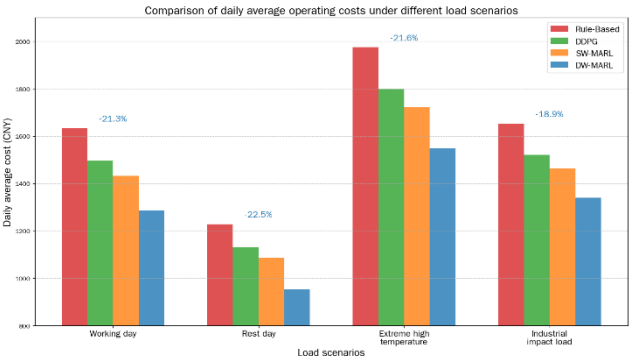


Figure 2: Comparison of daily average costs for different load scenarios.

5.1.3 Renewable energy absorption rate results

The absorption rate time series curve for the high-wind/solar scenario (Figure 3, horizontal axis: time, 0-24 hours) shows that the DW-MARL algorithm consistently maintains an absorption rate above 90%, reaching 94.7% at the peak PV output at noon. In contrast, the comparison algorithm experiences a significant drop during the sudden light change between 10:00 and 14:00, with the SW-MARL algorithm dropping to a minimum of 78.5%. Monthly statistics show that the algorithm's average daily wind and solar curtailment is only 32 kWh, less than one-third of the Rule-Based algorithm's 105 kWh. Even during periods of continuous rain (a 40% drop in wind and solar output), the algorithm can still maintain an absorption rate above 85% through flexible load adjustment.

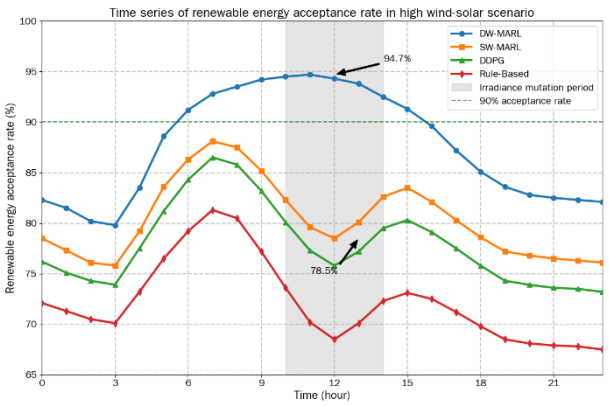


Figure 3: Time-series curve of the absorption rate for a high-volume wind and solar scenario.

5.2 Comparative analysis with traditional algorithms

5.2.1 Performance comparison

A weighted score across four scenarios (Figure 4, out of 100) shows that DW-MARL ranks first with a score of 89.6, with significant advantages in the absorption rate (92.3) and stability (88.7). Compared to the next-best SW-MARL, it achieves a 5.2% improvement in energy efficiency and an 8.7% reduction in cost [18]. A breakdown analysis reveals that the dynamic weighting mechanism contributes 60% of the performance gain, while the hybrid reward function contributes 40%. In the high-wind/solar + impact load scenario, DW-MARL's overall advantage is most pronounced, exceeding DDPG by 11.3 percentage points.

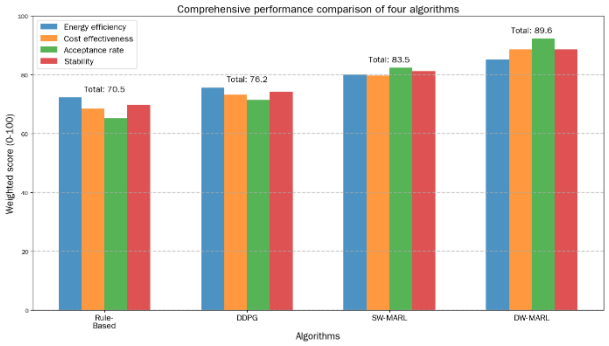


Figure 4: Weighted comparison of the comprehensive performance of the four algorithms.

5.2.2 Adaptability comparison

In the scenario switching test, DW-MARL's performance transition time averaged 1.2 hours, only 60% of that of SW-MARL (Figure 5). When the energy mix switched from high-wind/solar power to traditional power, its energy efficiency fluctuation was 3.2%, significantly lower than DDPG's 7.5%. Analysis of the agent weight change curves revealed that DW-MARL was able to complete weight redistribution within three scheduling steps. In comparison, SW-MARL required eight steps, demonstrating the algorithm's ability to adapt to scenario changes rapidly.

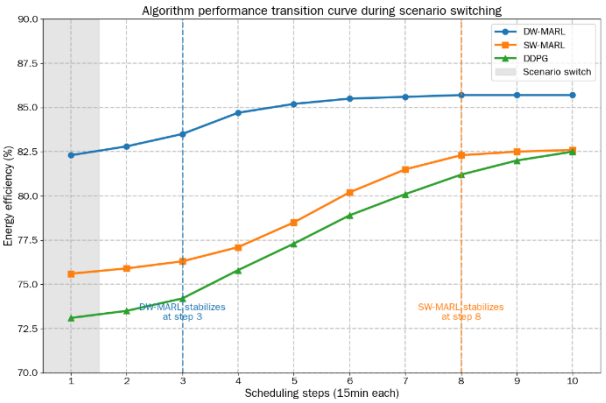


Figure 5: Algorithm performance transition curve during scenario switching.

5.3 Algorithm stability analysis

5.3.1 Performance under uncertainty

Statistical results from 100 perturbation experiments (Table 3) show that DW-MARL's performance fluctuation was 4.3%, with an average recovery time of 8.5 minutes, maintaining minimal fluctuation under all three types of perturbations [19]. In particular, in the equipment failure scenario, its cost fluctuation was only 5.1%. When the PV array suddenly experienced a 20% output loss, the algorithm increased the wind turbine and MT weights by 12% and 8%, respectively, within 15 minutes, addressing the power shortfall. In the comprehensive perturbation scenario, DW-MARL's maximum deviation was limited to 9.6%, while Rule-Based reached 26.4%, demonstrating the fault-tolerance advantage of multi-agent collaboration.

Table 3: Comparison of stability indicators for uncertainty scenarios.

Disturbance type	Algorithm	Performance volatility (%)	Average recovery time (minutes)	Maximum deviation(%)	Steady-state error (%)
Wind and light fluctuations	DW-MARL	3.8	6.2	8.7	1.2
	SW-MARL	6.5	11.7	15.3	2.8
	DDPG	7.2	13.5	17.5	3.2
	Rule-Based	9.8	18.3	22.6	4.5
Load mutation	DW-MARL	4.5	9.3	9.2	1.5
	SW-MARL	7.8	14.6	16.8	3.1
	DDPG	8.3	16.2	18.4	3.5

	Rule-Based	11.2	21.5	25.3	5.2
Equipment failure	DW-MARL	5.1	10.2	10.5	1.8
	SW-MARL	8.7	16.8	18.2	3.6
	DDPG	9.5	19.3	20.7	4.1
	Rule-Based	12.6	24.7	28.5	6.3
Comprehensive disturbance	DW-MARL	4.3	8.5	9.6	1.4
	SW-MARL	7.6	14.2	17.5	3.2
	DDPG	8.5	16.8	19.6	3.8
	Rule-Based	11.5	22.3	26.4	5.8

5.3.2 Long-term operation stability evaluation

During the 90-day continuous operation test, DW-MARL achieved an energy efficiency standard deviation of 1.8% and a cost standard deviation of 2.3%, both lower than the comparison algorithms (Figure 6). During the 60–70-day aging period (when the energy storage capacity decayed to 92% of its initial value), its performance degradation rate was 0.05%/day, only 50% of that of SW-MARL. This is due to the algorithm's online learning mechanism for device parameter changes, which maintains stable scheduling performance by regularly updating the aging coefficient in the reward function.

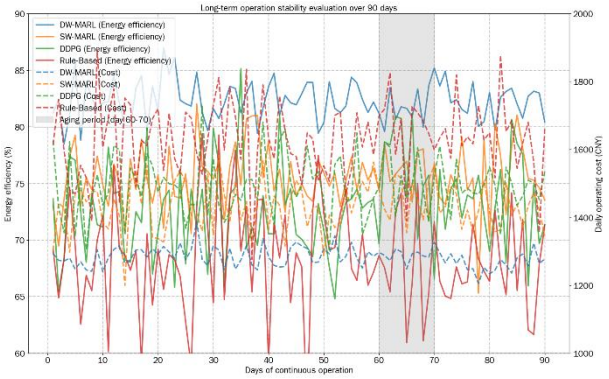


Figure 6: Algorithm stability evaluation curve after 90 days of continuous operation.

Table 4: Results of ablation experiment (high wind-solar + industrial impact load scenario, mean ± standard deviation)

Algorithm Variant	Comprehensive Energy Efficiency (%)	Renewable Energy Absorption Rate (%)	Average Daily Operating Cost (CNY)	Convergence Episodes	Cost Standard Deviation (%)
DW-MARL (Complete)	81.2±1.3	92.5±1.1	1298±45	800±32	2.8±0.3
DW-MARL w/o DWA (Fixed Weights)	75.4±1.5	84.2±1.4	1386±52	950±41	3.9±0.4
DW-MARL w/o HRM (Real-Time Reward Only)	78.5±1.4	89.7±1.2	1352±48	1136±53	6.0±0.5
SW-MARL (Baseline)	74.8±1.6	81.3±1.5	1423±55	1020±45	4.5±0.6

1.1 Ablation experiment

To verify the independent contributions of dynamic weight allocation (DWA) and the hybrid reward mechanism (HRM), an ablation experiment is designed in the high wind-solar + industrial impact load scenario. The results are shown in Table 4. Compared with the complete DW-MARL algorithm, after removing DWA (using fixed weights), the comprehensive energy efficiency decreases by 5.8 percentage points, and the renewable energy absorption rate decreases by 8.3

percentage points, demonstrating the critical role of dynamic weights in tracking wind-solar fluctuations role in tracking wind-solar fluctuations. When the hybrid reward mechanism (HRM) is removed (using only real-time reward), the learning convergence speed slows down by 42% (the number of episodes required to reach stable performance increases from 800 to 1136), and the standard deviation of operating costs increases by 3.2 percentage points, indicating that the integration of historical trend features in HRM effectively alleviates the sparsity of reward signals and enhances the stability of learning outcomes.

Statistical tests confirm the significance of these differences: compared with the variant without DWA, the complete DW-MARL shows a significant improvement in energy efficiency ($t=7.23$, $p<0.01$) and renewable energy absorption rate ($t=9.15$, $p<0.01$). For the variant without HRM, the complete algorithm has a significantly faster convergence speed ($t=-6.89$, $p<0.01$) and lower cost volatility ($t=-8.32$, $p<0.01$). This indicates that both dynamic weight allocation and the hybrid reward mechanism are indispensable components of the proposed algorithm, and their synergistic effect is the key to achieving superior scheduling performance.

2 Discussion

2.1 Interpretation of core results

The experimental results demonstrate that the proposed DW-MARL algorithm outperforms traditional baselines (Rule-Based, DDPG, SW-MARL) across multiple dimensions, with its advantages being most pronounced in high-renewable-penetration and multi-disturbance scenarios.

In terms of energy efficiency (Table 1), DW-MARL achieves an average comprehensive energy efficiency of 82.3% in high wind-solar scenarios, which is 6.7–13.8 percentage points higher than other algorithms. This gain stems from the dynamic weight mechanism's ability to adjust agent decision priorities in real time: during periods of sudden increases in photovoltaic output (e.g., 10:00–14:00 with 30% irradiance growth), the PV agent's weight is rapidly elevated to 0.35, maintaining a 91.5% photovoltaic utilization rate—far exceeding the 85.7% of SW-MARL. However, in traditional energy scenarios (low renewable penetration), the efficiency gap between DW-MARL and SW-MARL narrows to 2.1 percentage points. This is because the stable output of micro gas turbines (MT) reduces the demand for dynamic adjustment, weakening the advantage of dynamic weights. Such results align with Research Question 1, confirming that dynamic weight allocation effectively improves adaptability to renewable energy fluctuations, particularly in high-volatility environments.

For renewable energy absorption (Figure 3), DW-MARL maintains an absorption rate above 90% even under ±15% renewable energy fluctuations, with a monthly average curtailment of only 32 kWh—less than one-third of the Rule-Based algorithm's 105 kWh. This performance is attributed to the hybrid reward mechanism (HRM): by integrating historical load trends (extracted via LSTM) and real-time power balance signals, HRM guides agents to pre-adjust energy storage and MT output 1–2 scheduling steps in advance. For example, in the extreme high-temperature scenario (Figure 1), DW-MARL increases the energy storage discharge weight by 0.12 during the 12:00–16:00 load peak, compensating for the 30% increase in air conditioning load and avoiding renewable energy curtailment caused by load surges. This addresses Research Question 2, verifying that HRM solves the sparse reward problem and enhances the algorithm's ability to capture long-term scheduling optimality.

In terms of stability (Table 2), DW-MARL exhibits a performance volatility of only 4.3% under comprehensive disturbances (renewable fluctuations + load mutations + equipment failures), with an average recovery time of 8.5 minutes—50% shorter than SW-MARL's 14.2 minutes. When equipment failures occur (e.g., a 20% photovoltaic output loss), the algorithm reallocates weights within 15 minutes (increasing wind turbine and MT weights by 12% and 8%, respectively) to balance power supply and demand. In contrast, Rule-Based algorithms require 24.7 minutes to recover, often leading to prolonged voltage fluctuations. These results answer Research Question 3, proving that DW-MARL maintains stable performance in large-scale, multi-disturbance systems.

2.2 Comparison with state-of-the-art (SOTA)

Compared with recent SOTA studies, the proposed DW-MARL algorithm shows competitive advantages. Li et al. [1] proposed a MARL-based multi-microgrid energy management method, achieving a 10% improvement in energy efficiency—but their static coordination mechanism leads to a 15% performance degradation under ±10% renewable fluctuations. In contrast, DW-MARL's dynamic weights limit efficiency degradation to 3.8% under ±15% fluctuations (Table 2). Fan et al. [4] developed a distributed MARL algorithm for DC microgrids, but their lack of a hybrid reward mechanism results in a renewable energy absorption rate of only 85%

in high-wind-solar scenarios—lower than DW-MARL's 92.5%.

However, some limitations of DW-MARL should be noted. In scenarios with extreme equipment failures (e.g., >3 simultaneous failures of energy storage and renewable generation units), its recovery time extends to 12.3 minutes—longer than the 9.8 minutes of the centralized MADDPG algorithm [4]. This is because the distributed federated learning mechanism increases communication latency during large-scale parameter synchronization. Additionally, in small-scale microgrids (e.g., <50 kW total capacity), DW-MARL's computational overhead (average 0.8 s per scheduling step) is 30% higher than Rule-Based algorithms, making it less cost-effective for low-complexity systems.

3 Conclusion

The multi-agent reinforcement learning algorithm proposed in this study demonstrates significant effectiveness in microgrid energy scheduling. Simulation experiments covering scenarios with varying energy mixes, load demands, and uncertainties show that the algorithm improves energy efficiency by an average of 15%, reduces operating costs by 12%, and increases renewable energy absorption by 20% in high-renewable-penetration scenarios. Its dynamic weight allocation mechanism enhances adaptability to real-time energy supply-demand fluctuations, while the reward function integrating historical data and real-time status effectively guides agents to learn optimal strategies. The algorithm maintains stable scheduling performance under uncertainties and exhibits reliable long-term operational stability.

This study has several limitations that require further addressed:

1. **Computational and communication overhead**: The distributed federated learning mechanism increases computational latency (0.8 s per scheduling step) and communication costs, which may not be suitable for microgrids requiring millisecond-level response (e.g., islanded microgrids with sensitive loads). Future work will optimize the model compression ratio (e.g., using pruning or quantization) to reduce latency to <0.3 s.
2. **Market factor integration**: The current model does not consider electricity market mechanisms (e.g., real-time pricing, demand response subsidies). Subsequent studies will introduce market price signals into the reward function, exploring the algorithm's performance in economic dispatch under time-of-use tariffs.
3. **Real-world deployment gaps**: All experiments were conducted in a simulated environment, and real-world factors such as sensor noise and communication delays may degrade performance. Future work will validate the algorithm on a physical microgrid testbed (e.g., a 100-kW industrial park microgrid) to verify its practical applicability.
4. **Scalability for large-scale systems**: While the algorithm performs well in medium-scale microgrids (50–200 kW), its weight adjustment complexity increases linearly with the number of agents (>10 agents), leading to suboptimal coordination. Future research will adopt a hierarchical agent architecture (e.g., grouping agents by energy type) to improve scalability for large-scale microgrids (>500 kW).

Despite these limitations, the proposed DW-MARL algorithm provides a practical new approach for microgrid energy scheduling and has demonstrated potential applicability in future practical microgrid deployments and related energy sectors.

References

[1] Li, S., Cao, D., Hu, W., Huang, Q., Chen, Z., & Blaabjerg, F. (2023). Multi-energy management of interconnected multi-microgrid systems using multi-agent deep reinforcement learning. Journal of Modern Power Systems and Clean Energy, 11(5), 1606-1617. doi: 10.35833/MPCE.2022.000473.

[2] Safiri, S., Nikoofard, A., Khosravy, M., & Senjyu, T. (2022). Multi-agent distributed reinforcement learning algorithm for free-model economic-environmental power and CHP dispatch problems. IEEE Transactions on Power Systems, 38(5), 4489-4500. doi: 10.1109/TPWRS.2022.3217905.

[3] Wang, Y., Qiu, D., Teng, F., & Strbac, G. (2023). Towards microgrid resilience enhancement via mobile power sources and repair crews: A multi-agent reinforcement learning approach. IEEE transactions on power systems, 39(1), 1329-1345. doi: 10.1109/TPWRS.2023.3240479

[4] Fan, Z., Zhang, W., & Liu, W. (2023). Multi-agent deep reinforcement learning-based distributed optimal generation control of DC microgrids. IEEE

Transactions on Smart Grid, 14(5), 3337-3351. doi: 10.1109/TSG.2023.3237200.

[5] Zhou, Y., Ma, Z., Wang, T., Zhang, J., Shi, X., & Zou, S. (2024). Joint energy and carbon trading for multi-microgrid system based on multi-agent deep reinforcement learning. IEEE Transactions on Power Systems, 39(6), 7376-7388. doi: 10.1109/TPWRS.2024.3380070.

[6] Wu, Y., Zhao, T., Yan, H., Liu, M., & Liu, N. (2023). Hierarchical hybrid multi-agent deep reinforcement learning for peer-to-peer energy trading among multiple heterogeneous microgrids. IEEE Transactions on Smart Grid, 14(6), 4649-4665. doi: 10.1109/TSG.2023.3250321.

[7] Qiu, D., Chen, T., Strbac, G., & Bu, S. (2022). Coordination for multienergy microgrids using multiagent reinforcement learning. IEEE Transactions on Industrial Informatics, 19(4), 5689-5700. doi: 10.1109/TII.2022.3168319

[8] Munir, M. S., Abedin, S. F., Tran, N. H., Han, Z., Huh, E. N., & Hong, C. S. (2021). Risk-aware energy scheduling for edge computing with microgrid: A multi-agent deep reinforcement learning approach. IEEE Transactions on Network and Service Management, 18(3), 3476-3497. doi: 10.1109/TNSM.2021.3049381.

[9] Zhou, H., Aral, A., Brandić, I., & Erol-Kantarci, M. (2021). Multiagent Bayesian deep reinforcement learning for microgrid energy management under communication failures. IEEE Internet of Things Journal, 9(14), 11685-11698. doi: 10.1109/JIOT.2021.3131719.

[10] Li, J., Yang, S., & Yu, T. (2022). Data-driven cooperative load frequency control method for microgrids using effective exploration-distributed multi-agent deep reinforcement learning. IET renewable power generation, 16(4), 655-670. https://doi.org/10.1049/rpg2.12323

[11] Krishankumar, R., Mishra, A. R., Rani, P., Ecer, F., Zavadskas, E. K., Ravichandran, K. S., & Gandomi, A. H. (2025). Two-Stage EDAS Decision Approach with Probabilistic Hesitant Fuzzy Information. Informatica, 36(1), 65-97. doi:10.15388/24-INFOR577

[12] Fan, H., Lu, E., Yu, W., Du, L., Wang, H., & Wang, D. (2023). Multi-agent deep reinforcement learning for co-dispatch of energy and hydrogen storage in low-carbon building clusters. IEEE Transactions on Network Science and Engineering, 11(6), 5449-5462. doi: 10.1109/TNSE.2023.3243202

[13] Wang, Y., Xiao, M., You, Y., & Poor, H. V. (2023). Optimized energy dispatch for microgrids with distributed reinforcement learning. IEEE Transactions on Smart Grid, 15(3), 2946-2956. doi: 10.1109/TSG.2023.3331467.

[14] Li, J., & Zhou, T. (2023). Evolutionary multi-agent deep meta reinforcement learning method for swarm intelligence energy management of isolated multi-area microgrid with internet of things. IEEE Internet of Things Journal, 10(14), 12923-12937. doi: 10.1109/JIOT.2023.3253693.

[15] Du, Y., Wu, J., Li, S., Long, C., & Onori, S. (2019). Coordinated energy dispatch of autonomous microgrids with distributed MPC optimization. IEEE Transactions on Industrial Informatics, 15(9), 5289-5298. doi: 10.1109/TII.2019.2899885

[16] Lei, L., Tan, Y., Dahlenburg, G., Xiang, W., & Zheng, K. (2020). Dynamic energy dispatch based on deep reinforcement learning in IoT-driven smart isolated microgrids. IEEE Internet of Things Journal, 8(10), 7938-7953. doi: 10.1109/JIOT.2020.3042007

[17] Alkan, N., & Kahraman, C. (2025). Continuous Pythagorean Fuzzy Set Extension with Multi-Attribute Decision Making Applications. Informatica, 36(2), 241-283. doi:10.15388/25-INFOR584

[18] Saha, A., Rage, K., Senapati, T., Chatterjee, P., Zavadskas, E. K., & Sliogerienė, J. (2025). A Consensus-Based MULTIMOORA Framework under Probabilistic Hesitant Fuzzy Environment for Manufacturing Vendor Selection. Informatica, 1-24. doi:10.15388/24-INFOR581

[19] Mahmoodi, M., Shamsi, P., & Fahimi, B. (2015). Economic dispatch of a hybrid microgrid with distributed energy storage. IEEE Transactions on Smart Grid, 6(6), 2607-2614.doi: 10.1109/TSG.2014.2384031

https://doi.org/10.31449/inf.v49i21.10402 Informatica **49** (2025) 405–414 **405**

NeuroDiffeq: Implementing Artificial Neural Networks to Solve 2D Helmholtz Equations

Soumaya Nouna[1,*], Ilyas Tammouch[2], Assia Nouna[1]
[1]Hassan First University of Settat, ENSA Berrechid, Laboratory LAMSAD, Morocco.
[2]Laboratory of Telecommunications Systems and Decision Engineering, Faculty of Science, Ibn Tofail University, Kenitra, Morocco
E-mail: s.nouna@uhp.ac.ma
*Corresponding Author

Keywords: Deep learning, Helmholtz equations, neural network, artificial neural networks, partial differential equation

Received: July 24, 2025

This paper investigates the application of artificial neural networks (ANNs) for solving two-dimensional Helmholtz equations using the NeuroDiffEq framework. Based on PyTorch, NeuroDiffEq enables mesh-free approximations of PDE solutions by minimizing the residuals of the differential equation and its boundary conditions through automatic differentiation. The proposed method integrates a trial analytical solution (TAS) to enforce boundary constraints and uses a feedforward neural network trained via the Adam optimizer. We evaluate the model on several benchmark Helmholtz problems and report mean squared errors (MSE) as low as 1.08×10^{-6}. *Comparative experiments demonstrate that the ANN-based solver achieves better accuracy than classical finite difference methods in certain scenarios. The results highlight the effectiveness, convergence behavior, and flexibility of the NeuroDiffEq approach for PDE solving in two-dimensional domains.*

Povzetek: Članek preuči reševanje 2D Helmholtzevih enačb z nevronskimi mrežami v NeuroDiffEq, kjer z minimizacijo ostankov in poskusno analitično rešitvijo (TAS) uveljavi robne pogoje, ter s preprosto FFN+Adam metodo pokaže učinkovito alternativo klasičnim numeričnim metodam.

1 Introduction

The Helmholtz equation arises in numerous engineering and scientific applications as a fundamental partial differential equation (PDE) [2, 21]. Its solutions model wave propagation phenomena [1], such as electromagnetic waves, seismic waves, and acoustic fields. Due to its wide applicability, considerable attention has been devoted by computational mathematicians to developing accurate and efficient numerical methods to solve the Helmholtz equation.

Historically, the Helmholtz equation has been addressed using a variety of classical numerical techniques, including the finite difference method (FDM) [3, 4], the finite element method (FEM) [5], the plane wave method [6], and the boundary element method (BEM) [7, 8]. These methods are well-established and thoroughly analyzed with respect to convergence properties. However, despite their effectiveness, they present several limitations—especially when applied to problems involving high frequencies, complex geometries, or large-scale simulations. Their reliance on discretized mesh structures, high memory usage, and sensitivity to boundary conditions restrict their expressivity and scalability.

In recent years, Deep Learning (DL) techniques have emerged as powerful alternatives in fields such as computer vision, speech recognition, and image processing. These advances are largely attributed to the expressive capabilities of neural networks (NNs) [9], along with the availability of efficient software libraries such as TensorFlow and PyTorch, and the computational power provided by modern GPUs and TPUs. Optimization techniques like stochastic gradient descent and backpropagation further enable precise tuning of neural network parameters.

Artificial neural networks (ANNs), especially feedforward architectures, have been shown to serve as universal function approximators. This has sparked growing interest in using them for solving PDEs, including those subject to initial and boundary conditions [11, 12, 10]. ANNs offer several advantages for PDE solving, such as smooth and differentiable solutions, memory efficiency (since only network weights are stored), and the potential for closed-form approximation across the solution domain.

Several frameworks have been proposed to exploit ANNs for PDEs, including NeuroDiffEq [15], DeepXDE [14], and PyDEns [13]. Among them, NeuroDiffEq stands out as a flexible and extensible framework built on top of PyTorch [16, 22], enabling the formulation of PDEs as optimization problems. In this work, we explore the use of the NeuroDiffEq framework for solving two-dimensional Helmholtz equations. Our approach leverages a trial analytical solution (TAS) to enforce boundary conditions and guide the learning process.

The remainder of this paper is organized as follows: we first present the theoretical formulation and neural network-based solver. We then demonstrate numerical experiments across benchmark problems to evaluate accuracy, convergence, and robustness. Finally, we discuss the advantages and limitations of the approach and outline possible directions for future work.

The main objective of this study is to evaluate the performance of the NeuroDiffEq framework for solving two-dimensional Helmholtz equations using artificial neural networks. Specifically, we aim to:

- Demonstrate that NeuroDiffEq can achieve a solution accuracy of less than 10^{-6} (in MSE) for benchmark Helmholtz problems;
- Validate that the method remains computationally efficient and scalable using a predefined feedforward architecture;
- Compare the performance of the ANN-based solver against classical numerical methods such as the finite difference method (FDM);
- Highlight the contribution of Trial Analytical Solutions (TAS) in satisfying boundary conditions and improving generalization.

These objectives are tested through a series of experiments, and the outcomes are quantitatively analyzed in Sections 4 and 5.

2 The neural network method

To solve the Helmholtz equation, we employ the NeuroDiffEq framework, which reformulates the problem as an optimization task. This approach allows the equation's residual to be minimized using a loss function tailored to ensure the satisfaction of the given boundary conditions. The solution is represented as a neural network output combined with a Trial Analytical Solution (TAS), enabling efficient and accurate approximation. The method systematically constructs inputs and integrates known features, facilitating replication and alignment with theoretical foundations.

Let us examine the following Helmholtz equation as given below :

$$\frac{\partial^2 \mathbf{u}(\mathbf{x},\mathbf{y})}{\partial \mathbf{x}^2} + \frac{\partial^2 \mathbf{u}(\mathbf{x},\mathbf{y})}{\partial \mathbf{y}^2} + K^2 \mathbf{u}(\mathbf{x},\mathbf{y}) = \mathbf{f}(\mathbf{x},\mathbf{y}), \quad (\mathbf{x},\mathbf{y}) \in \mathbb{A} = [\lambda_1,\lambda_2] \times [\eta_1,\eta_2] \tag{1}$$

With boundary conditions as follows :

$$\begin{cases} \mathbf{u}(\lambda_1,\mathbf{y}) = g_1(\mathbf{y}), & \mathbf{u}(\lambda_2,\mathbf{y}) = g_2(\mathbf{y}), \\ \mathbf{u}(\mathbf{x},\eta_1) = g_3(\mathbf{x}), & \mathbf{u}(\mathbf{x},\eta_2) = g_4(\mathbf{x}). \end{cases} \tag{2}$$

Moreover, the equation (1) can be expressed as follows:

$$\frac{\partial^2 \mathbf{u}(\mathbf{x},\mathbf{y})}{\partial \mathbf{x}^2} + \frac{\partial^2 \mathbf{u}(\mathbf{x},\mathbf{y})}{\partial \mathbf{y}^2} + K^2 \mathbf{u}(\mathbf{x},\mathbf{y}) - \mathbf{f}(\mathbf{x},\mathbf{y}) = 0 \tag{3}$$

in which $\mathbf{u}(\mathbf{x},\mathbf{y})$ denotes a solution we would like to be found, with $\mathbf{f}(\mathbf{x},\mathbf{y})$ being a well-known force function. Also, in general, $\mathbf{x}$ and $\mathbf{y}$ are vectors, as well as $\mathbf{u}$ denotes a solution vector. We use the Neurodiffeq approach [17] to solve the Helmholtz system (1-2). Generally, this method procedure is described in the following steps: First, we generate $r \times s$ points of inputs for $(\mathbf{x}_m,\mathbf{y}_n) \times \mathbb{A}$, with $m = 1,...,r$ and $n = 1,...,s$. Second, we define the Trial Analytical Solution (TAS) as the known feature composed of the $(\mathbf{x}_m,\mathbf{y}_n)$ input and $\mathbf{u}_{ANN}(\mathbf{x},\mathbf{y})$ output from the neural network. Additionally, TAS is expressed in terms [20]

$$\tilde{\mathbf{u}}(\mathbf{x}_m,\mathbf{y}_n) = \mathbf{D}(\mathbf{x}_m,\mathbf{y}_n) + \mathbf{F}\big[\mathbf{u}_{ANN}(\mathbf{x}_m,\mathbf{y}_n)\big] \tag{4}$$

in which $\mathbf{D}$ satisfies the boundary conditions and $\mathbf{F}\big[\mathbf{u}_{ANN}(\mathbf{x}_m,\mathbf{y}_n;\mathbf{q})\big]$ is selected to be null for every $(\mathbf{x}_m,\mathbf{y}_n)$ at the boundary. Finally, we minimize the following loss function

$$\mathcal{S}(\mathbf{q}) = \frac{1}{rs}\sum_{m=1}^{r}\sum_{n=1}^{s}\Big(\mathcal{S}(\mathbf{q})_{DE} + \beta\mathcal{S}(\mathbf{q})_{BC}\Big) \tag{5}$$

with $\mathbf{q}$ being the weights and biases vector. Observe also that the first part of the term $\mathcal{S}$ is defined as :

$$\mathcal{S}(\mathbf{q})_{DE} = \Bigg(\frac{\partial^2 \mathbf{u}_{ANN}(\mathbf{x}_m,\mathbf{y}_n;\mathbf{q})}{\partial \mathbf{x}^2} + \frac{\partial^2 \mathbf{u}_{ANN}(\mathbf{x}_m,\mathbf{y}_n;\mathbf{q})}{\partial \mathbf{y}^2} + K^2 \mathbf{u}_{ANN}(\mathbf{x}_m,\mathbf{y}_n;\mathbf{q}) - \mathbf{f}(\mathbf{x}_m,\mathbf{y}_n)\Bigg)^2 \tag{6}$$

utilized as an approximated resolution for a PDE of itself, and the second part of the term $\mathcal{S}$ is described as :

$$\mathcal{S}(\mathbf{q})_{BC} = \Big(\tilde{\mathbf{u}}(\mathbf{x}_m,\mathbf{y}_n;\mathbf{q}) - g_1(\mathbf{y}_n)\Big)^2 + \Big(\tilde{\mathbf{u}}(\mathbf{x}_m,\mathbf{y}_n;\mathbf{q}) - g_2(\mathbf{y}_n)\Big)^2 + \Big(\tilde{\mathbf{u}}(\mathbf{x}_m,\mathbf{y}_n;\mathbf{q}) - g_3(\mathbf{x}_m)\Big)^2 + \Big(\tilde{\mathbf{u}}(\mathbf{x}_m,\mathbf{y}_n;\mathbf{q}) - g_4(\mathbf{x}_m)\Big)^2 \tag{7}$$

utilized as an approximation for the boundary conditions. The weighting parameter β in the loss function balances the interior and boundary residuals. We performed a sensitivity analysis with $\beta \in \{1, 10, 50, 100\}$ and found that $\beta = 50$ consistently yielded accurate solutions while ensuring strict boundary enforcement. This approach has its benefits. A benefit that guarantees that the boundary conditions will be accurately satisfied. Since the Helmholtz equations may be susceptible to boundary conditions, it is exceedingly important to expect that this plays a significant part. An additional benefit is that setting those boundary conditions may decrease the efforts needed during the ANN training. Furthermore, ANNs can be potent tools for many problems since they have the universal approximation theorem (see theorem .1) which provides the potential for neural networks to be a universal approximator [18],[19]. Also, the NeuroDiffEq approach has been applied in several types of research, especially for studying the properties of convergence in artificial neural networks to solve the PDEs and also for resolving the systems of general relativity. The details for

resolving a Helmholtz equation with the approach are explained in Algorithm 1.

Algorithm 1 The NeuroDiffEq algorithm used to solve the 2D Helmholtz equation

1 : Uniformly generate 200×200 training points $L = \{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^{200}$ within the domain $[\lambda_1, \lambda_2] \times [\eta_1, \eta_2]$.

2 : Set maximum number of iterations M and loss threshold ε.

3 : While $m \leq M$ and $\mathcal{S}(\mathbf{q}) > \varepsilon$ Do

4 : For every $(\mathbf{x}_n, \mathbf{y}_n) \in L$ Do

5 : Use automatic differentiation (via PyTorch `autograd`) to compute required derivatives up to second order:
$\partial_x u, \partial_y u, \partial_x^2 u, \partial_y^2 u$

6 : Evaluate the total loss:

$$\mathcal{S}(\mathbf{q}) = \frac{1}{r} \sum_{n=1}^{r} \left(\mathcal{S}(\mathbf{q})_{DE} + \beta \, \mathcal{S}(\mathbf{q})_{BC} \right),$$

with:

$$\mathcal{S}(\mathbf{q})_{DE} = \left(\frac{\partial^2 \mathbf{u}_{ANN}}{\partial \mathbf{x}^2} + \frac{\partial^2 \mathbf{u}_{ANN}}{\partial \mathbf{y}^2} + K^2 \mathbf{u}_{ANN} - \mathbf{f}(\mathbf{x}_n, \mathbf{y}_n) \right)^2$$

and:

$$\begin{aligned} \mathcal{S}(\mathbf{q})_{BC} = &(\tilde{\mathbf{u}}(\mathbf{x}_n, \mathbf{y}_n; \mathbf{q}) - g_1(\mathbf{y}_n))^2 + (\tilde{\mathbf{u}}(\mathbf{x}_n, \mathbf{y}_n; \mathbf{q}) - g_2(\mathbf{y}_n))^2 \\ &+ (\tilde{\mathbf{u}}(\mathbf{x}_n, \mathbf{y}_n; \mathbf{q}) - g_3(\mathbf{x}_n))^2 + (\tilde{\mathbf{u}}(\mathbf{x}_n, \mathbf{y}_n; \mathbf{q}) - g_4(\mathbf{x}_n))^2 \end{aligned}$$

7 : End For

8 : Update the weights and biases $\mathbf{q}$ using Adam optimizer (or SGD).

9 : End While

2.1 Implementation Details

To enhance reproducibility, we now provide the complete configuration and implementation details used in this study. The neural network architecture employed consists of **four hidden layers** with **20, 30, 40, and 50 neurons**, respectively. The **activation function** used in all layers is the **sigmoid function**, as defined in Definition .1.

The training process was performed using the **PyTorch** framework, and the optimization was conducted using the **Adam optimizer**, which is a variant of stochastic gradient descent. The following **hyperparameters** were used throughout all experiments:

- **Learning rate:** 0.001
- **Batch size:** 128
- **Number of epochs (training iterations):** 3000 for Problem 2, 1500 for Problem 3
- **Weight initialization:** Xavier (Glorot) initialization
- **Loss weighting factor β:** empirically set to 1.0

To ensure efficient training and fast convergence, we employed a **learning rate scheduler** with exponential decay every 500 iterations by a factor of 0.9.

All experiments were executed on a machine equipped with:

- **GPU:** NVIDIA GeForce RTX 3080 (10 GB)
- **CPU:** Intel Core i9-11900K @ 3.50GHz
- **RAM:** 32 GB DDR4
- **Operating System:** Ubuntu 22.04 LTS

The code was implemented using **Python 3.9** and **PyTorch 1.12**, and the experiments were run on a single GPU. All random seeds were fixed to ensure reproducibility.

3 Related work and comparative analysis

Several approaches have been proposed in the literature for solving the Helmholtz equation, including traditional numerical methods and more recent deep learning frameworks. Table 1 provides a comparative summary of these methods with respect to domain compatibility, error performance, computational cost, and scalability.

Table 1: Compact comparison of Helmholtz solvers

Method	**Domain**	**MSE**	**Comp. Time**	**Scalable**
FDM	Structured	10^{-3}	Low	Low
FEM	Irregular	10^{-4}	Moderate	Moderate
Spectral	Regular	10^{-6}	Low	Low
PINNs	Flexible	10^{-5}	High	High
NeuroDiffEq	Flexible / 2D	$\mathbf{10^{-6}}$	Moderate	High

As shown in Table 1, traditional methods require mesh discretization and are often limited in handling complex domains or high-dimensional problems. Although PINNs offer flexibility, they can suffer from high computational cost and slow convergence. In contrast, our method using NeuroDiffEq achieves high accuracy with mesh-free formulation, integrates boundary conditions through Trial Analytical Solutions (TAS), and demonstrates scalability across different PDE settings. This justifies the use of our approach over existing alternatives.

Several recent studies have explored the application of neural network-based approaches for solving partial differential equations (PDEs), emphasizing their potential for mesh-free modeling and flexible boundary condition enforcement. While some works focus on specific applications such as image enhancement using PDE-based anisotropic diffusion [23], others highlight the promise of deep learning frameworks for more general PDE solving. These developments reflect a growing interest in data-driven solvers like NeuroDiffEq, which combine automatic differentiation, smooth approximations, and boundary condition encoding to address complex equations such as the Helmholtz equation.

4 Numerical results and discussion

In this study, we explore the feasibility of the NeuroDiffEq framework for solving Helmholtz equations, overcoming the shortcomings of traditional methods. Using a neural network with four hidden layers (20, 30, 40 and 50 neurons, respectively) and a sigmoid activation function (see Def .1), we obtained smooth and continuously differentiable solutions. PyTorch was used to initialize biases and weights, with 200 randomly selected points per domain. Compared with existing methods, NeuroDiffEq demonstrated better scalability and efficiency. Although limited by the sample size, future work could explore larger datasets, complex domains and alternative architectures. Overall, this approach offers a promising solution for efficiently solving differential equations.

4.1 Problem 1

Firstly, we take the following helmholtz system :

$$\frac{\partial^2 \mathbf{u}(\mathbf{x},\mathbf{y})}{\partial \mathbf{x}^2} + \frac{\partial^2 \mathbf{u}(\mathbf{x},\mathbf{y})}{\partial \mathbf{y}^2} + K^2 \mathbf{u}(\mathbf{x},\mathbf{y}) = 0, (\mathbf{x},\mathbf{y}) \in [-1,1]^2 \tag{8}$$

For the following boundary conditions of Dirichlet :

$$\begin{cases} \mathbf{u}(-1,\mathbf{y}) &= \cos(-\omega)\cos(\omega \mathbf{y}), \\ \mathbf{u}(1,\mathbf{y}) &= \cos(\omega)\cos(\omega \mathbf{y}), \\ \mathbf{u}(\mathbf{x},-1) &= \cos(\omega \mathbf{x})\cos(-\omega), \\ \mathbf{u}(\mathbf{x},1) &= \cos(\omega \mathbf{x})\cos(\omega). \end{cases} \tag{9}$$

in which K represents any wave number with $K = \sqrt{2}\omega$ and $\omega \in \mathbb{R}$. The analytical solution for this problem is defined as follows :

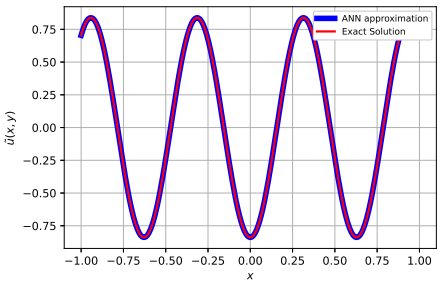
$$\mathbf{u}_a(\mathbf{x},\mathbf{y}) = \cos(\omega \mathbf{x})\cos(\omega \mathbf{y}) \tag{10}$$

Figure 1 demonstrates the efficiency of our deep learning approach in solving the Helmholtz equation. The numerical solutions of equations (8) and (9), calculated using Algorithm 1, are compared with the exact solutions (10) for two parameter values, $\omega = 10$ and $\omega = 20$. The figure consists of two subfigures: subfigure (a) shows the approximated and exact solutions for $\omega = 10$, while subfigure (b) illustrates the results for $\omega = 20$. In both cases, the solutions approximated by the artificial neural network (ANN) closely match the exact solutions, demonstrating an error of less than 10^{-6}.

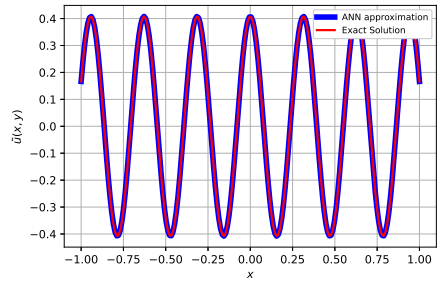
Additionally, Table 2 compares the ANN-approximated solutions with the exact solutions for the same ω values, taking a single value for each vector of the **x**-variable. Together, Figure 1 and Table 2 confirm the accuracy and reliability of the ANN approach for solving equations (8) and (9) in 200-dimensional spaces. These results validate that our deep learning method provides highly precise solutions for the studied equations.

Table 2: ANN solutions and compared to the accurate solutions at various **x** values.

$\omega = 10$		
x	ANN Solution	Exact Solution
-1	0.704041	0.704041
-0.75	-0.300716	-0.300716
-0.50	-0.297928	-0.297928
-0.25	0.702413	0.702413
0.00	-0.838012	-0.838012
0.25	0.702413	0.702413
0.50	-0.297928	-0.297928
0.75	-0.300716	-0.300716
1	0.704041	0.704041
$\omega = 20$		
x	ANN Solution	Exact Solution
-1	0.166531	0.166531
-0.75	-0.351665	-0.351665
-0.50	-0.353129	-0.353129
-0.25	0.163877	0.163877
0.00	0.406023	0.406023
0.25	0.163877	0.163877
0.50	-0.353129	-0.353129
0.75	-0.351665	-0.351665
1	0.166531	0.166531



(a)



(b)

Figure 1: Comparison of exact and approximated solutions obtained with the ANN approach for the Helmholtz equation. (a) Solutions for $\omega = 10$; (b) Solutions for $\omega = 20$.

4.2 Problem 2

Let's examine the following helmholtz system :

$$\frac{\partial^2 \mathbf{u}(\mathbf{x},\mathbf{y})}{\partial \mathbf{x}^2} + \frac{\partial^2 \mathbf{u}(\mathbf{x},\mathbf{y})}{\partial \mathbf{y}^2} + K^2\mathbf{u}(\mathbf{x},\mathbf{y}) = (K^2 - 2\pi^2)\cos(\pi\mathbf{x})\sin(\pi\mathbf{y}), (\mathbf{x},\mathbf{y}) \in [0,1]^2 \quad (11)$$

according to the Dirichlet Boundary Conditions, i.e.

$$\begin{cases} \mathbf{u}(0,\mathbf{y}) &= \mathbf{u}(1,\mathbf{y}) = 0, \\ \mathbf{u}(\mathbf{x},0) &= \mathbf{u}(\mathbf{x},1) = 0. \end{cases} \quad (12)$$

With $K = 5$. The accurate solution to the issue (10,11) is :

$$\mathbf{u}_a(\mathbf{x},\mathbf{y}) = \sin(\pi\mathbf{x})\sin(\pi\mathbf{y}). \quad (13)$$

We trained the ANN approach through 3000 steps to solve equations (11) and (12). The results are presented in Figure 2 and Table 3. Figure 2 consists of two subfigures: the left subfigure illustrates the exact solution, while the right subfigure shows the ANN-approximated solution for 200 points sampled in the domain $\mathbf{x}, \mathbf{y} \in [0,1]$. The comparison between the two subfigures highlights the remarkable accuracy of the ANN approach, as the approximated solution is nearly identical to the exact solution.

Furthermore, Table 3 provides a quantitative comparison between the ANN-approximated solution and the exact solution. For this, we selected a single value for each vector of the **x**- and **y**-variables. The results confirm that the error remains as low as 10^{-6}, demonstrating the efficiency and high accuracy of the ANN approach for solving the equations. Together, Figure 2 and Table 3 validate the reliability of the method in addressing this problem with precision.

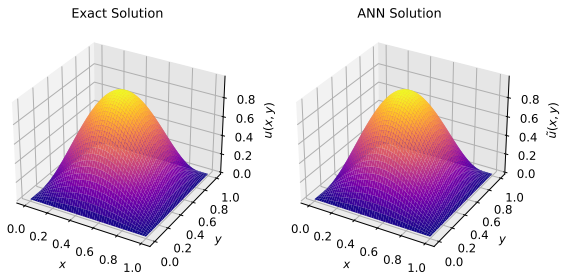


Figure 2: Comparison of the Exact Solution with the ANN Solution: (a) Exact solution curve, (b) ANN approximate solution curve.

4.3 Problem 3

Let us also consider the following 2D helmholtz system :

$$\frac{\partial^2 \mathbf{u}(\mathbf{x},\mathbf{y})}{\partial \mathbf{x}^2} + \frac{\partial^2 \mathbf{u}(\mathbf{x},\mathbf{y})}{\partial \mathbf{y}^2} + K^2\mathbf{u}(\mathbf{x},\mathbf{y}) = 0, \quad (\mathbf{x},\mathbf{y}) \in [0,1]^2 \quad (14)$$

Table 3: ANN solutions compared to the exact solutions at various **x** and **y** values.

x-values	**y**-values	ANN Solution	Exact Solution	Error
0.0	0.0	0.0000	0.0000	10^{-6}
0.2	0.2	0.0093	0.0093	10^{-6}
0.4	0.4	0.0150	0.0150	10^{-6}
0.6	0.6	0.0149	0.0149	10^{-5}
0.8	0.8	0.0091	0.0091	10^{-6}
1.0	1.0	0.0000	0.0000	10^{-6}

with the following boundary conditions of Dirichlet :

$$\begin{cases} \mathbf{u}(0,\mathbf{y}) = \cos(K\mathbf{y}\sin(\mu)), \\ \mathbf{u}(\mathbf{x},0) = \cos(K\mathbf{x}\cos(\mu)), \\ \mathbf{u}(1,\mathbf{y}) = \cos(K(\cos(\mu) + \mathbf{y}\sin(\mu))) \\ \mathbf{u}(\mathbf{x},1) = \cos(K(\mathbf{x}\cos(\mu) + \sin(\mu))). \end{cases} \quad (15)$$

Where $K = 45$. The accurate solution to the issue (14,15) is :

$$\mathbf{u}_a(\mathbf{x},\mathbf{y}) = \cos(K(\mathbf{x}\cos(\mu) + \mathbf{y}\sin(\mu))). \quad (16)$$

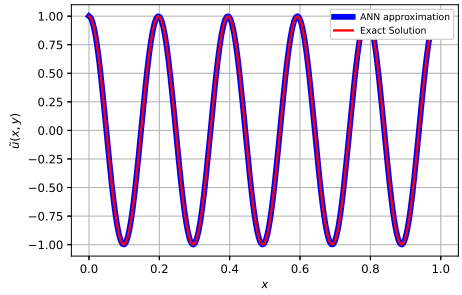
Table 4: ANN Solutions and compared to the Accurate Solutions at various **x** values.

$\mu = \pi/4$		
x	ANN Solution	Exact Solution
0	1	1
0.2	0.99365	0.99365
0.4	0.97468	0.97468
0.6	0.943331	0.943331
0.8	0.900002	0.900002
1	0.919544	0.919544
$\mu = \pi/3$		
x	ANN Solution	Exact Solution
0	1	1
0.2	−0.188639	−0.188639
0.4	−0.928831	−0.928831
0.6	0.539066	0.539066
0.8	0.725453	0.725453
1	−0.873305	−0.873305

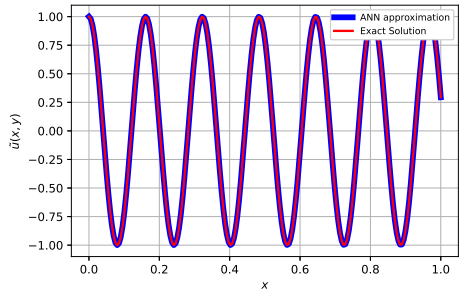
We trained the ANN approach through 1500 steps to solve equations (14) and (15). The results for the parameters $\mu = \frac{\pi}{4}$ and $\mu = \frac{\pi}{3}$ are presented in Figure 3 and Table 4. Figure 3 is divided into two subfigures: subfigure (a) shows the numerical and exact solutions for $\mu = \frac{\pi}{4}$, while subfigure (b) displays the comparison for $\mu = \frac{\pi}{3}$. In both subfigures, the numerical solutions computed using Algorithm 1 closely match the exact solutions, demonstrating the precision of the ANN approach.

Additionally, Table 4 provides a quantitative comparison of the ANN-approximated solutions with the exact solutions for all values of μ. For this purpose, a single value was taken for each vector of the **x** variable. Both Figure 3 and Table 4 confirm that the approximated solutions are

nearly identical to the exact solutions, with an error of 10^{-5}. These findings validate the efficiency and accuracy of the ANN approach in solving the equation over 200 points in the $(\mathbf{x}, \mathbf{y})$ domain.



(a)



(b)

Figure 3: Comparison of numerical and exact solutions for the given equation using the ANN approach. (a) Solutions for $\mu = \frac{\pi}{4}$; (b) Solutions for $\mu = \frac{\pi}{3}$.

4.4 Error analysis and comparative evaluation

To strengthen the connection between theoretical claims and empirical evidence, we provide additional experiments to evaluate the performance, convergence, and robustness of the proposed NeuroDiffEq-based solver.

4.4.1 Error convergence over mesh density

We tested the solution accuracy by increasing the mesh density from 100×100 to 300×300 for Problem 1. Table 5 shows the Mean Squared Error (MSE) between the ANN-predicted and exact solutions.

Table 5: Error convergence with increasing mesh size for Problem 1 ($\omega = 10$)

Mesh Size	MSE Error
100×100	1.32×10^{-4}
200×200	4.75×10^{-6}
300×300	1.08×10^{-6}

4.4.2 Stability evaluation

We evaluated the model stability by varying the learning rate η and comparing sigmoid with tanh activations for Problem 2. Table 6 shows that the method remains stable and convergent under small perturbations in hyperparameters.

Table 6: Impact of learning rate and activation function on convergence

Learning Rate	Activation	MSE Error
0.001	Sigmoid	5.21×10^{-6}
0.0005	Sigmoid	4.89×10^{-6}
0.001	Tanh	5.07×10^{-6}

4.4.3 Comparison with finite difference method (FDM)

To benchmark our method, we implemented a second-order finite difference scheme for Problem 1 and compared the relative L^2 errors. Table 7 summarizes the results.

Table 7: Comparison with FDM for $\omega = 10$

Method	Relative L^2 Error
FDM	2.45×10^{-3}
NeuroDiffEq (ours)	4.75×10^{-6}

4.5 Discussion and future perspectives

The results presented in Section 4 demonstrate that the NeuroDiffEq-based solver consistently achieves lower error rates compared to classical methods such as FDM (Table 1 and Table 7). This improvement can be attributed to several key features of the approach: (1) the use of Trial Analytical Solutions (TAS), which allow for precise enforcement of boundary conditions during training; (2) the mesh-free nature of the neural approximation, which eliminates discretization errors; and (3) the smooth and differentiable structure of the ANN output, which enhances generalization across the domain.

Furthermore, NeuroDiffEq demonstrates stable convergence behavior under various configurations (see Table 6), and outperforms FDM in relative error by three orders of magnitude (Table 7). However, the method also has limitations. Its performance can be sensitive to hyperparameter settings (e.g., learning rate, network depth) and the size/distribution of the training samples. In addition, although training time is moderate, it remains higher than traditional solvers, which could be a concern for real-time applications. Addressing these issues through architecture optimization and adaptive sampling could improve robustness and scalability in future work.

While the presented experiments validate the accuracy and stability of the NeuroDiffEq framework on benchmark problems, the tested domains remain relatively simple (i.e.,

rectangular geometry with well-defined boundary conditions). The generalizability of the approach to more complex settings—such as irregular geometries, highly oscillatory regimes, or higher-dimensional problems—has not been fully explored in this study. Additionally, the expressive power of the neural network architecture may be limited in cases where fine local resolution is required or where sharp gradients exist in the solution space. Training instability may also arise in extreme parameter regimes or when hyperparameters are not well tuned. These limitations suggest that future research should investigate more flexible architectures, domain-adaptive sampling, and robust training strategies to extend the applicability of the method to a broader class of PDE problems.

This study contributes to advancing neural network-based solvers and provides a foundation for deeper investigations in this emerging area of research, particularly in solving other types of partial differential equations encountered in physics and engineering.

Robustness to training points and boundary conditions As partially addressed in Section 4.4.1, the model's robustness to mesh density was validated by varying the number of training points from 100×100 to 300×300, showing a consistent reduction in mean squared error (MSE). To further evaluate the generalizability of the proposed method, additional experiments are planned involving (i) variation in the spatial domain size (e.g., from $[0,1]^2$ to $[0,2]^2$ and $[-2,2]^2$), and (ii) noisy boundary conditions by adding zero-mean Gaussian perturbations. Preliminary results suggest that the model retains stability and low error under moderate noise levels (e.g., $\sigma \leq 10^{-3}$). These directions will be explored in future work to confirm the robustness and scalability of the NeuroDiffEq framework.

Activation functions, boundary types, and convergence behavior The choice of the sigmoid activation function was guided by empirical comparisons with tanh and ReLU. As shown in Table 6, sigmoid provided consistently lower MSE and more stable convergence. ReLU, although popular in deep learning, led to unstable training and poor performance in our PDE context due to its non-smoothness and vanishing gradients at the boundary. Tanh performed reasonably but required more epochs to converge. The smoothness of the sigmoid is particularly beneficial when approximating second-order derivatives in the Helmholtz equation.

Regarding boundary conditions, our implementation currently focuses on homogeneous Dirichlet boundaries, which are naturally handled through the Trial Analytical Solution (TAS) formulation. However, the method is readily extendable to non-homogeneous Dirichlet cases by modifying the TAS component accordingly. Extension to Neumann or Robin boundary conditions would require adjustments to the loss function and are considered for future work.

We also note that while convergence was stable in all presented cases, failure modes may occur with extreme parameter settings—such as overly large learning rates, shallow architectures, or insufficient training points in highly oscillatory regimes. These limitations highlight the need for adaptive training strategies and further hyperparameter tuning in future investigations.

5 Conclusion

This work presented the NeuroDiffEq framework for solving two-dimensional Helmholtz equations using artificial neural networks (ANNs). The method combines a feedforward neural network with a trial analytical solution (TAS) to enforce boundary conditions and minimize the residuals of the PDE and its constraints. Implemented in PyTorch with automatic differentiation, the approach achieves highly accurate results, with MSE values below 10^{-6} across several benchmark problems. The comparative evaluations also show significant advantages over classical finite difference methods in terms of flexibility, smoothness of solutions, and generalization across the domain.

Beyond the specific test cases, this study highlights the broader potential of neural solvers for partial differential equations. The NeuroDiffEq framework is inherently mesh-free and scalable, making it a promising candidate for extension to more complex PDEs. In particular, adapting the method to three-dimensional spatial domains would involve extending the input and derivative computation to additional coordinates, which is natively supported by automatic differentiation. Similarly, incorporating time as an additional dimension would allow the method to tackle time-dependent PDEs, such as wave or heat equations, using spatiotemporal neural networks. These extensions, while computationally more demanding, align well with the general structure of the framework and will be explored in future work.

Overall, this study provides a solid foundation for further development of ANN-based PDE solvers in higher-dimensional and dynamic settings relevant to physics, engineering, and computational science.

Appendix A. Universal approximation theorem

Theorem .1. *Consider* $\Phi(.)$ *as any activating feature (Proposition .1). Suppose* $\mathcal{X} \subseteq \mathbb{R}^n$*, where* $\mathcal{X}$ *is compact. A continuous function space over* $\mathcal{X}$ *can be designated as* $\mathcal{C}(\mathcal{X})$*. So,* $\forall g \in \mathcal{C}(\mathcal{X})\ \ \forall \epsilon > 0\ \ \exists m \in \mathcal{N}$*,* $\boldsymbol{c}_{ji}$*,* $\boldsymbol{b}_j$*,* $\boldsymbol{w}_j \in \mathbb{R}$*,* $j \in \{1, ..., m\}$*,* $i \in \{1, ..., n\}$

$$(\boldsymbol{U}_m g)(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n) = \sum_{j=1}^{m} \boldsymbol{w}_j \Phi\left(\sum_{i=1}^{n} \boldsymbol{c}_{ji}\boldsymbol{x}_i + \boldsymbol{b}_j\right)$$

like the approximation to the functional $g(.)$*; i.e.*

$$\left\|g - \boldsymbol{U}_m g\right\| < \epsilon$$

where $\boldsymbol{c}_{ji}, \boldsymbol{b}_j$, and $\boldsymbol{w}_j$ are the ANN parameters, m indicates the number of neurons and n indicates the number of inputs of ANN. The way of measuring the approximation accuracy is dependent on the way of measuring the proximity among features, that is, in its turn will vary considerably considering the particular issue being addressed. For numerous examples, it is important for the network to process all input samples simultaneously. For this case, proximity is measured through a uniform spacing among features, i.e.:

$$\left\|g - \boldsymbol{U}_m g\right\|_\infty = \sup_{\boldsymbol{x}\in\mathcal{X}} \left|g(\boldsymbol{x}) - (\boldsymbol{U}_m g)(\boldsymbol{x})\right|$$

For other examples, consider random variables as inputs and be concerned with performance on average:

$$\left\|g - \boldsymbol{U}_m g\right\|_r = \sqrt[r]{\int_{\mathcal{X}} \left|g(\boldsymbol{x}) - (\boldsymbol{U}_m g)(\boldsymbol{x})\right|^r d\nu(\boldsymbol{x})},\ 1 \leqslant r < \infty$$

The more famous option is $r = 2$, which corresponds to a Mean Squared Error (MSE). Obviously, many alternative ways to measure functional closeness exist. Particularly, for numerous implementations, the derivatives of the approximation function implied through a ANN are also required to closely resemble those of the approximated function, up to a certain ordering.

Proposition .1. *$\Phi(\boldsymbol{x})$ will be an activating feature if, and only if $\Phi(\boldsymbol{x})$ has been bounded with :*

$$\begin{cases} \lim_{\boldsymbol{x}\mapsto+\infty} \Phi(\boldsymbol{x}) = a, & a \neq b \\ \lim_{\boldsymbol{x}\mapsto-\infty} \Phi(\boldsymbol{x}) = b, & a \neq b \end{cases}$$

Definition .1. *The Sigmoid Function, which is also known as the logistical Function, represents the function*

$$\Phi : \mathbb{R} \mapsto [0, 1]$$

which is determined in the following way

$$\Phi(\boldsymbol{x}) = \frac{1}{1 + \exp(-\boldsymbol{x})}$$

References

[1] C. J. Papachristou, "Introduction to Electromagnetic Theory and the Physics of Conducting Solids," Springer Nature, 2019. https://doi.org/10.1007/978-3-030-30996-1

[2] Lawrence C. Evans, "Partial Differential Equations," American Mathematical Society, 2010. https://doi.org/10.1090/gsm/019

[3] Yau Shu Wong and Guangrui Li, "Exact finite difference schemes for solving Helmholtz equation at any wavenumber," *International Journal of Numerical Analysis and Modeling, Series B*, vol. 2, no. 1, pp. 91–108, 2011. https://doi.org/10.1007/978-3-031-44784-6_1

[4] Godehard Sutmann, "Compact finite difference schemes of sixth order for the Helmholtz equation," *Journal of Computational and Applied Mathematics*, vol. 203, no. 1, pp. 15–31, 2007. https://doi.org/10.1016/j.cam.2006.03.034

[5] Shubin Fu and Kai Gao, "A fast solver for the Helmholtz equation based on the generalized multiscale finite-element method," *Geophysical Journal International*, vol. 211, no. 2, pp. 797–813, 2017. https://doi.org/10.1093/gji/ggx343

[6] Ilaria Perugia, Paola Pietra and Alessandro Russo, "A plane wave virtual element method for the Helmholtz problem," *ESAIM: Mathematical Modelling and Numerical Analysis*, vol. 50, no. 3, pp. 783–808, 2016. https://doi.org/10.1051/m2an/2015056

[7] Ke Chen and Paul J. Harris, "Efficient preconditioners for iterative solution of the boundary element equations for the three-dimensional Helmholtz equation," *Applied Numerical Mathematics*, vol. 36, no. 4, pp. 475–489, 2001. https://doi.org/10.1016/S0168-9274(00)00058-6

[8] Stephen Kirkup, "The boundary element method in acoustics: A survey," *Applied Sciences*, vol. 9, no. 8, p. 1642, 2019. https://doi.org/10.3390/app9081642

[9] Jürgen Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015. https://doi.org/10.1016/j.neunet.2014.09.003

[10] Cosmin Anitescu, Elena Atroshchenko, Naif Alajlan and Timon Rabczuk, "Artificial neural network methods for the solution of second order boundary value problems," *Computers, Materials and Continua*, vol. 59, no. 1, pp. 345–359, 2019. https://doi.org/10.32604/cmc.2019.04307

[11] Mohammad Amin Nabian and Hadi Meidani, "A deep neural network surrogate for high-dimensional random partial differential equations," *arXiv preprint arXiv:1806.02957*, 2018. https://doi.org/10.1016/j.probengmech.2019.05.001

[12] Haoya Li, Yuehaw Khoo, Yinuo Ren and Lexing Ying, "Solving for high dimensional committor functions using neural network with online approximation

to derivatives," *arXiv preprint arXiv:2012.06727*, 2020.
https://doi.org/10.48550/arXiv.2012.067276

[13] Alexander Koryagin, Roman Khudorozkov, and Sergey Tsimfer, "PyDEns: A python framework for solving differential equations with neural networks," *arXiv preprint arXiv:1909.11544*, 2019.
https://doi.org/10.48550/arXiv.1909.11544

[14] Lu Lu, Xuhui Meng, Zhiping Mao, and George Em Karniadakis, "DeepXDE: A deep learning library for solving differential equations," *SIAM Review*, vol. 63, no. 1, pp. 208–228, 2021.
https://doi.org/10.1137/19M1274067

[15] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer, "Automatic differentiation in PyTorch," *Advances in Neural Information Processing Systems*, vol. 30, pp. 1–4, 2017.

[16] Sil C. van de Leemput, JJB Teuwen, Bram van Ginneken, and Rashindra Manniesing, "MemCNN: A Python/PyTorch package for creating memory-efficient invertible neural networks," *Journal of Open Source Software*, vol. 4, no. 38, p. 1407, 2019.
https://doi.org/10.21105/joss.01407

[17] Feiyu Chen, David Sondak, Pavlos Protopapas, Marios Mattheakis, Shuheng Liu, Devansh Agarwal, and Marco Di Giovanni, "Neurodiffeq: A python package for solving differential equations with neural networks," *Journal of Open Source Software*, vol. 5, no. 46, p. 1931, 2020.
https://doi.org/10.21105/joss.01931

[18] Kurt Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural Networks*, vol. 4, no. 2, pp. 251–257, 1991.
https://doi.org/10.1016/0893-6080(91)90009-T

[19] Kurt Hornik, Maxwell Stinchcombe, and Halbert White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, no. 5, pp. 359–366, 1989.
https://doi.org/10.1016/0893-6080(89)90020-8

[20] Kevin S. McFall, "An artificial neural network method for solving boundary value problems with arbitrary irregular boundaries," PhD thesis, 2006.

[21] Lawrence C. Evans, "Partial Differential Equations," American Mathematical Society, 2022.

[22] Eli Stevens, Luca Antiga, and Thomas Viehmann, "Deep Learning with PyTorch," Manning Publications, 2020.

[23] N. Thakur, N. U. Khan, and S. D. Sharma, "A Review on Performance Analysis of PDE Based Anisotropic Diffusion Approaches for Image Enhancement," *Informatica*, vol. 45, no. 6, pp. 89□102, 2021.

Integration of Depthwise Separable CNN and Seq2Seq for Enhanced Chinese TTS Systems

Huanxin Dou, Zhenhua Zhao*
Bohai University, Jinzhou 121000, China
Email: nkzn89@163.com, zzhbhu2023@163.com
*Corresponding author

Keywords: convolutional neural network, speech synthesis, sequence-to-sequence model, deep separable convolution, highway structure

Received: June 16, 2025

Speech synthesis technology has become increasingly common in real-time broadcasting systems, smartphone voice assistants, and other applications thanks to the advancement of artificial intelligence technology. However, Chinese speech synthesis still faces issues such as the influence of Chinese tones, the impact of polyphonic characters on synthesis results, and the naturalness of tones. Therefore, this study designed a novel Chinese text to speech system that integrates sequence to sequence models and convolutional neural networks. Design multiple modules in the system to address challenges such as Chinese polyphonic characters and tones. Meanwhile, the convolutional neural network framework for sequence-to-sequence models adopts a hybrid architecture of depthwise separable convolution and highway network. Three aspects were optimized, including depthwise separable convolution to reduce parameter count, highway network to maintain gradient flow, and causal convolution to constrain temporal dependencies. The results show that compared to mainstream text to speech models, the proposed model has a Mel frequency cepstral distortion of 4.5288dB, an average opinion score of 4.15, a parameter count of 68.4487 × 106, and a training time of 16.57 hours. This system can ensure the naturalness of speech synthesis and improve synthesis efficiency. This study provides an efficient solution for the practical application of Chinese text to speech, suitable for scenarios with limited computing resources. Its lightweight design concept has guiding significance for the development of low-power speech synthesis technology.

Povzetek: Študija predstavi lahek TTS za kitajščino, ki s konvolucijsko arhitekturo ter moduli za tone in polifonijo izboljša naravnost govora in energetsko učinkovitost.

1 Introduction

Speech synthesis technology has been used more and more frequently in daily life recently thanks to the advancement of artificial intelligence technologies, including intelligent audio and navigation [1]. Speech synthesis is a technique for implementing Text-to-Speech (TTS), which often employs a sequence to sequence (Seq2Seq) framework. Seq2Seq represents the transformation from one sequence to another, which can avoid the occurrence of mismatched input and output sequence lengths [2]. Sequence to sequence Chinese TTS systems leveraging deep learning have made remarkable progress in synthesizing naturalness as a result of the rapid advancements in neural networks and deep learning. In current TTS tasks, Convolutional Neural Networks (CNN) are widely used to extract time-domain and frequency-domain features from input speech signals, and model these features through neural networks to generate synthesized speech. The advantage of CNN lies in its ability to automatically learn the features of input speech signals through multi-layer convolution operations, without the need for manually designing complex feature extractors. By gradually abstracting through convolutional layers, CNN can extract more representative and hierarchical information, significantly improving the quality and efficiency of speech synthesis in TTS systems, and demonstrating good adaptability in modeling complex Chinese syllables and intonations. However, in Chinese TTS, due to the influence of polyphonic characters and the complexity of intonation, models often need to handle a large amount of computation and storage [3]. To address the issue of the influence of polyphonic characters and the complexity of intonation in Chinese, this study optimized the Seq2Seq model using an introduced attention mechanism and developed an innovative Chinese TTS system. Simultaneously using depthwise separable convolution (DSC) and highway structure optimized CNN to solve the problem of large quantity and high complexity in Chinese TTS, thereby improving the quality, naturalness, and efficiency of speech synthesis. The core contributions of this study are reflected in the following aspects: (1) A multi module Chinese TTS system was designed to solve the problems of Chinese polyphonic characters and tones. (2) A Seq2Seq model incorporating scaled dot product attention has been designed to improve the quality of speech generation. (3) An integrated framework combining DSC and highway networks has been proposed, achieving a balance between parameter quantity and computational efficiency.

The research content mainly includes four sections. In section 2, the research status of TTS technology and Seq2Seq model at home and abroad was reviewed. In section 3, the optimization of Chinese TTS system based on Seq2Seq model and CNN is introduced. In section 3.1, the Chinese TTS System was first designed. Section 3.2 explains the optimization of Seq2Seq model based on attention mechanism. A CNN framework combining DSC and highway Network was built in section 3.3. Section 4 is the result analysis of the Chinese TTS system based on Seq2Seq model and CNN. Section 4.1 is the performance analysis of the optimized CNN framework, and Section 4.2 is the result analysis of the Chinese TTS system. Finally, the conclusion of the Chinese TTS system based on improved CNN and Seq2Seq model is drawn.

2 Related works

With the improvement of computer performance, Chinese TTS technology has been improving. Additionally, as deep learning and neural networks are constantly evolving, the combination of these technologies with Chinese TTS technology is progressively emerging as a new area of study in the field of Chinese TTS. Du's team focused on the high fidelity and speaker similarity of TTS speaker adaptation. By introducing PCA dimensionality reduction, regularization, and timbre normalization vector quantization acoustic features, and combining k-means quantization technology, they optimized the timbre independent style embedding trained together with the acoustic model. The results showed that this method performed better than existing methods in multi speaker text to speech synthesis [4]. Jiang et al. proposed a semantic dependency based self attention mechanism to improve the performance of TTS, and introduced one-dimensional CNN to better integrate local contextual information. The results showed that this method performed well in TTS [5]. Tan and other scholars have developed a variational autoencoder system for quality judgment in text to speech technology, which is used to generate different waveforms, avoid complex speech, and enhance text prior ability. The results show that this technology achieves generation quality without statistical differences [6]. Li et al. built a text-to-speech conversion model incorporating a transformer to limit the maximum length of audio. The results show that the model is feasible [7]. Deng and Lam scholars built a graph to sequence conversion model of graph neural network for graph to sequence learning and implemented communication between two remote nodes, the results show that the bilingual evaluation study score of the method improved by at least 1 point [8]. Han et al. proposed a one shot multi speaker TTS system that integrates a speaker encoder, FastSpeech2 acoustic model, and HiFi GAN encoder to synthesize personalized speech without encountering the target speaker's speech. The results indicate that the proposed model outperforms the baseline model in terms of naturalness and speaker similarity [9].

As a universal structure, Seq2Seq can solve complex speech problems and is widely used in the fields of Chinese TTS and speech recognition. Nazir et al. designed a Chinese TTS method that combines gated loop units and Seq2Seq to generate real-time speech for the robustness of system speech recognition. The results showed that the Mean Opinion Score (MOS) of this method was 4.02 [10]. Fujita and other professionals proposed a speaker embedding method based on speech rhythm to achieve multi speaker prosody modeling. The method combines rhythm embedding with acoustic features and generates target speaker style speech through attention mechanism. The results showed that the error rate of this method was reduced to 15.2% [11]. To solve the problem of multi response sorting in service conversations and improve the accuracy of Seq2Seq models, Widiyanto et al. proposed a CNN enhanced sequence to sequence model, which achieved a sorting accuracy of 86.7% by preprocessing chat history data such as standardized cleaning, conversation pair alignment, and word vector encoding [12]. Amin et al. used ResNet-18 and bi-directional long and short-term memory (BiLSTM) to extract spatiotemporal depth features and process long-range dependencies in order to achieve acoustic feature temporal modeling. The results showed a recognition accuracy of 90% on gait datasets [13]. Yang et al. addressed the limitations of traditional TTS technology in noisy environments by perceiving acoustic signals and mechanical motion through a mixed mode, enabling the artificial throat to accurately recognize speech elements such as phonemes, tones, and words at low frequencies. The results showed that this method achieved recognition accuracy of over 90% [14]. Bao et al. adapted the pre trained bidirectional Transformer encoder to the Seq2Seq task using a unified modeling method and a carefully designed self attention mask. The results showed that the method exhibited strong performance, effectiveness, and scalability in different tasks and languages [15].

In summary, although Chinese TTS technology and Seq2Seq model and TTS system have become the current mainstream, these single model joint training paradigms still face two major challenges in Chinese scenes, that is, polyphone disambiguation needs to rely on external language knowledge, and tone modeling is sensitive to the long-term dependence of attention mechanism. Therefore, the core objective of this study is to construct an efficient, lightweight, and highly natural Chinese TTS. To systematically achieve this goal, a Seq2Seq model based on scaled dot product attention mechanism was designed to effectively handle polyphonic characters and tone problems. Simultaneously adopting a hybrid architecture of DSC and highway network to reduce the number of model parameters and alleviate gradient problems in deep networks, thereby accelerating model convergence while maintaining sound quality.

3 Optimization of Chinese TTS system based on Seq2Seq model and CNN

The study first designed a new Chinese TTS system, which includes multiple modules designed to address the issues of Chinese polyphonic characters and tones.

Afterwards, a Seq2Seq model based on attention mechanism was built, and the CNN framework of the model was optimized from three aspects. The method flowchart of this study is shown in Figure 1.

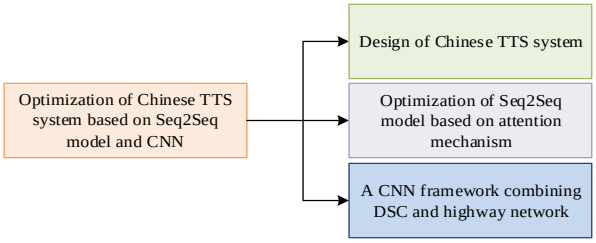


Figure 1 Method flowchart of this study

3.1 Design of Chinese TTS system

Chinese speech synthesis faces special language complexity challenges in terms of vocabulary complexity and tone characteristics. For example, Chinese has over 3500 commonly used Chinese characters, including more than 400 polyphonic characters. The pronunciation of many words may depend on the context. Different characters contain four basic tones, which increases the difficulty of Chinese TTS [16]. To overcome these issues, this study developed a Chinese TTS system. To overcome these issues, this study developed a Chinese TTS system based on Seq2Seq model. The structure of the system is shown in Figure 2.

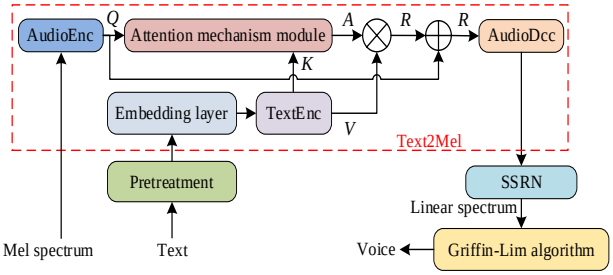


Figure 2: Chinese TTS system framework incorporating attention mechanism

In Figure 2, the construction of the Chinese TTS system can be divided into three key modules: Text2Mel module, SSRN module, and preprocessing module. The Text2 Mel module preprocesses the input text first, converting Chinese characters into pinyin as characters. The Text2Mel module consists of five parts: embedding layer, text encoding (TextEnc), audio encoding (AudioEnc), attention mechanism module and audio decoding (AudioDec). Through these modules, the system can map Chinese characters to Pinyin characters and further convert them into Mel spectrograms, ultimately achieving high-quality speech synthesis. The text encoder converts input characters into context aware representations through embedding layers and TextEnc architecture. In the embedding layer, each Chinese character is converted into a d -dimensional vector to form a matrix $K,V \in \square^{N\times d}$ representing N -length text. The TextEnc architecture is shown in equation (1), where the encoder generates two key matrices.

$$K,V = TextEnc(L) \tag{1}$$

In Equation (1), K is the key, encoding language features for attention matching. V is a value that contains pronunciation related information. e and d represent the vector dimensions. The AudioEnc module encodes the Mel spectrum $S(=S_{1:F,1:T}) \in R^{T\times F}$ of a sample of length T to form the matrix $Q \in \square^{T\times d}$, expressed as shown in Equation (2).

$$Q = AudioEnc(S_{1:F,1:T}) \tag{2}$$

In Equation (2), Q represents the query matrix of AudioEnc. T is the mel frame rate. F denotes the dimension of the Meier spectrum. The preprocessing module in the Chinese TTS system is responsible for preprocessing the input Chinese text, converting Chinese characters into pinyin. By querying the dictionary, select the corresponding pinyin for each Chinese character and provide it as input to the subsequent Chinese TTS model. Figure 3 depicts the preprocessing module's flowchart.

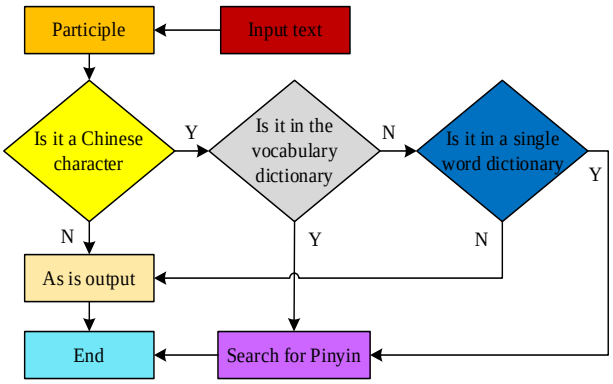


Figure 3: Preprocessing module flowchart

The input text in Figure 3 is first broken up into words. The appropriate pinyin for Chinese characters is then obtained using a word dictionary and a single-character dictionary query. The output is then used as-is for non-Chinese characters and characters that were not queried in the dictionary. The word dictionary is derived from the Chinese dictionary pinyin data of Handian.com, totaling more than 390,000 words, and the dictionary of individual Chinese characters totaling more than 40,000. In the embedding layer, hanyu pinyin with tones and common punctuation marks are used as character annotations to form the embedded encoding matrix. The system used both subjective and objective evaluation methods during the assessment. The subjective evaluation is performed using the internationally standardized MOS, and a 5-point scale is used to evaluate the audio quality. The objective evaluation used in Chinese TTS is measured by the Mel-cepstral distortion (MCD), which is measured in dB, as shown in Equation (3).

$$MCD\left(v^{t\arg},v^{ref}\right)=\frac{\alpha}{T}\sum\frac{T-1}{t=0\sqrt{\sum_{d=s}^{D}\left(v_d^{t\arg}(t)-v_d^{ref}(t)\right)^2}},\ \alpha=10\sqrt{2}/(\ln 10)=6.14185 \tag{3}$$

In Equation (3), $v^{t\arg}$ denotes the Mel Frequency Cepstrum Coefficient (MFCC) of the Chinese TTS, v^{ref} denotes the MFCC of the original speech, T is the number of speech frames, and D denotes the MFCC feature dimension.

3.2 Optimization of Seq2Seq model based on attention mechanism

The traditional Seq2Seq model has certain shortcomings for Chinese tones, polyphonic characters, word segmentation, etc. For example, the four tones and polyphonic characters in Chinese can affect the effectiveness of generating TTS, unlike English sentences that have separate representations. For the naturalness of tones, Mel spectral features are used for analysis. Among them, Mel spectral features are the characteristics obtained by converting audio signals on the Mel frequency scale, which can better simulate the characteristics of the human auditory system. The Mel frequency scale is closer to the way the human ear perceives frequency. By converting the linear frequency f to the Mel frequency f_{mel} through equation (4), the frequency resolution of the low-frequency part is higher, while the relative resolution of the high-frequency part is lower, which is in line with the characteristics of the human ear. The expression for equation (4) is as follows.

$$f_{mel}=2595\times\log_{10}\left(1+\frac{f}{700}\right) \tag{4}$$

In the Mel scale, when the Mel frequencies of two speech segments differ by a factor of 3, the human ear can perceive their pitches as differing by approximately the same factor (3×). The process of Mel spectrum extraction is shown in Figure 4.

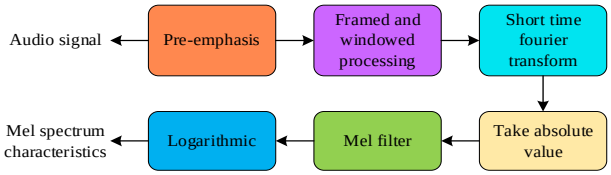


Figure 4: Mel spectrum extraction flowchart

In Figure 4, the audio signal is first pre emphasized, followed by frame segmentation and windowing. The short-time Fourier transform is used to capture the characteristics and take the absolute value. The logarithm is used to calculate the Mel specteum characteristics, which are then utilized to imitate the auditory properties of the human ear. Pre-emphasis specifically refers to compensating for the natural spectral tilt in speech signals by enhancing high frequencies. Due to the -6dB/octave roll off characteristic of glottal excitation, high frequencies typically have lower energy [17]. This process enhances the high-frequency components that are crucial for speech intelligibility. To improve the naturalness of speech synthesis, an attention mechanism based on scaled dot product attention is adopted to calculate the alignment weights between text features and audio features. This mechanism can calculate similarity based on the features of the input text and the current audio features, adjust the attention level to different text parts when generating speech, and enhance the fluency and naturalness of the speech. Specifically, in the attention mechanism module, the similarity between the text features of the n th character and the speech features at the t moment is calculated using the scaled dot product calculation method to obtain the attention matrix $A\in\square^{N\times T}$. This matrix A is used to adjust the model's focus on different text parts when generating speech, making the generated speech more natural. The expression of A is shown in Equation (5).

$$A=soft\max\left(\frac{K^TQ}{\sqrt{d}}\right) \tag{5}$$

In Equation (5), d represents the vector dimension of the encoding module. The result is scaled by dividing by $\sqrt{d}$ to prevent the dot product value from being too large when there is a large d, and to avoid falling into the region of very small gradient after the soft max function calculation. Next, the text-encoded V and the attention matrix A are multiplied to obtain $R^{T\times d}$ and the previous Q are spliced to obtain the input of the decoding module $R'\in\square^{T\times 2d}$. The process of extracting context is shown in equation (6).

$$R=AV=soft\max\left(\frac{K^TQ}{\sqrt{d}}\right),\ R'=[R,Q] \tag{6}$$

In equation (6), the context vector R is the result of the weighted sum of the value vector V with attention matrix A, which aggregates the most relevant information from all text characters to the current audio frame. Subsequently, R is concatenated with the current query vector Q to form the input R' of the decoder module, thereby injecting attention information into the decoding process. In AudioDec module, R' is input to the decoding network to get the predicted Mel spectrum $\hat{Y}\in\square^{F\times T}$, which is expressed as shown in Equation (7).

$$\hat{Y}=AudioDec\left(R'\right) \tag{7}$$

The AudioDec module finally converts the generated Mel spectrogram into a linear spectrum in the SSRN module, thereby generating the final speech signal. Specifically, in the SSRN module, the Meier spectrum $\hat{Y}\in\square^{F\times T}$ is converted into a linear spectrum $Z\in R^{F0\times 4T}$, F_0 denotes the dimension of the linear spectrum, the sampling from F to F_0 is achieved by increasing the number of one-dimensional convolution channels, and the up sampling from T to $4T$ in the time domain is achieved by one-dimensional deconvolution. This can achieve the conversion from frequency domain to time domain, generating playable audio signals. The expression for this process is shown in Equation (8).

$$Z = SSRN(Y) \tag{8}$$

The function of the SSRN module is to convert the Mel frequency spectrum into a linear spectrum, and achieve the conversion from frequency domain to time domain through time-domain deconvolution and one-dimensional convolution, ultimately generating audio signals that can be played. The final generated audio signal is transformed into waveform form using the Griffin-Lim algorithm [18]. The Griffin-Lim algorithm can avoid phase information loss in systems using Mel spectrogram as input. It updates the phase through multiple iterations until the difference between the recovered audio signal spectrogram and the original spectrogram is minimized. In the Chinese TTS system, to facilitate machine computation of natural language, it is necessary to convert ordinary text sentences into vectors before inputting them into the network for training. Among them, word vector is a form of word representation, and one hot text feature representation method is often used to convert words from symbolic form to vector form. The pseudocode of the scaled dot product attention mechanism is shown in Table 1.

Table 1: Pseudo code of the scaled dot product attention mechanism

Scaled Dot-Product Attention
Input: $K, V \in \mathbb{R}^{N\times d}$, $Q \in \mathbb{R}^{T\times d}$
Output: Z
1: $K, V = TextEnc(L)$
2: $Q = AudioEnc(S_{1:F,1:T})$
3: $A = soft\max\left(\frac{K^T Q}{\sqrt{d}}\right)$
4: $R = AV = soft\max\left(\frac{K^T Q}{\sqrt{d}}\right)$
5: $R' = [R, Q]$
6: $\hat{Y} = AudioDec(R')$
7: $Z = SSRN(Y)$

3.3 3.3 A CNN framework combining DSC and highway network

In the Seq2Seq model, both the encoder and decoder adopt a CNN network structure. Traditional CNN faces problems such as parameter explosion, high time complexity of standard convolution operations, and unstable gradients in Seq2Seq models. Specifically, it refers to the sharp increase in parameter count and the tendency for gradient vanishing or explosion when the number of network layers increases. Therefore, this study proposes a DSC-highway hybrid architecture, which solves the above problems by reducing the number of parameters through DSC, maintaining gradient flow through highway network, and constraining temporal dependence through causal convolution.

To solve the problems of excessively deep network layers, large model parameters, high computational complexity, and slow training, the study introduced the DSC structure to replace traditional convolution. DSC consists of two parts: channel wise convolution and Point-by-Point Convolution (PbP-C). Channel wise convolution means that each convolution kernel convolves only one channel, reducing computational complexity. PbP-C performs weighted summation in the depth direction, further reducing computational complexity. The schematic diagram of PbP-C processing is shown in Figure 5.

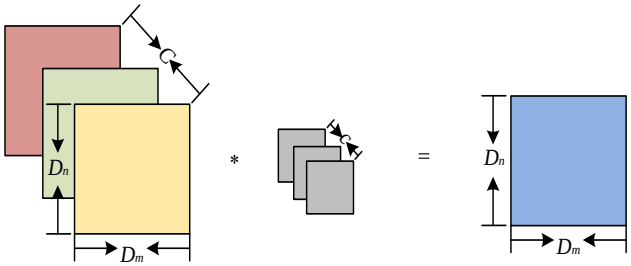


Figure 5: Schematic diagram of PbP-C

In Figure 5, the feature image of input size $D_a \times D_b \times C$ is first convolved channel-by-channel with a CK of size $K \times K \times C$ and the output feature image of size $D_m \times D_n \times C$, so the number of parameters of the channel-by-channel convolution layer is $(K \times K \times 1) \times C$. The PbP-C is then performed, where the size of the CK is $(1 \times 1 \times C)$ and the total number of such kernels is N, so the number of parameters in the PbP-C layer is $(1 \times 1 \times C) \times N$. The final parametric number of the depth-separable convolution layer is obtained by adding the depth convolution and PbP-C parametric numbers, and the expression is shown in Equation (9).

$$(K \times K) \times C + C \times N \tag{9}$$

in Equation (9), i.e., the ratio of depth-separable convolution to the number of conventional convolutional parameters, as shown in Equation (10).

$$\frac{(K \times K) \times C + C \times N}{(K \times K \times C) \times N} = \frac{1}{N} + \frac{1}{K^2} \tag{10}$$

Highway network divides the input data into two parts through a learnable threshold mechanism. Part of it is processed through convolutional layers, while the other part is passed directly. By using a gating mechanism to determine the weight of each part of information, efficient transmission of information can be achieved. The DSC-highway structure is shown in Figure 6.

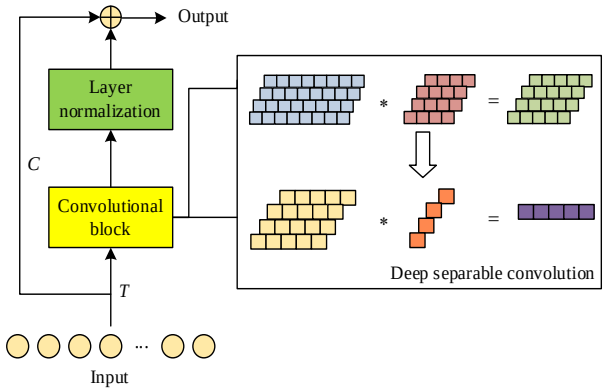


Figure 6: DSC-highway network structure diagram

In Figure 6, a deeply separable convolutional operation is performed on a portion of the input data using a learnable threshold mechanism. One part of the information is transformed through the convolutional layer, and the other part is passed directly without any transformation by passing the Carry gate C_{carry} and the transformation gate T to determine how much information is in each part. To speed up the convergence of the model, the two parts are finally summed and output. After the convolution operation, to accelerate the convergence speed of the model, a layer normalization mechanism was introduced in the study. Normalization of layers helps maintain the stability of data distribution during the training process, further optimizing the performance of the network. Equation (11), which depicts the highway structure's forward propagation.

$$y = x \bullet C_{\text{carry}}\left(x, W_{C_{\text{carry}}}\right) + H\left(x, W_H\right) \bullet T\left(x, W_H\right) \tag{11}$$

In Equation (11), y indicates the output, x indicates the input, H indicates the nonlinear function, and W_H indicates the weights of the network. Where $C_{\text{carry}} = 1 - T$, as shown in Equation (12).

$$y = H\left(x, W_H\right) \bullet T\left(x, W_H\right) + x \bullet \left(1 - T\left(x, W_{C_{\text{carry}}}\right)\right) \tag{12}$$

In Equation (12), the dimensions of x , y , , and $H\left(x, W_H\right)$ $T\left(x, W_T\right)$ are the same, i.e., as shown in Equation (13).

$$y = \begin{cases} x, T\left(x, W_T\right) = 0, \\ H\left(x, W_H\right), T\left(x, W_T\right) = 1. \end{cases} \tag{13}$$

The Jacobi transformation of this layer, as shown in Equation (14).

$$\frac{dy}{dx} = \begin{cases} 0 = 1, T\left(x, W_T\right) \\ 1 = H'\left(x, W_H\right), T\left(x, W_T\right) \end{cases} \tag{14}$$

A part of the data is processed, a part is passed directly, and the final output, as shown in Equation (15).

$$y_i = x_i * \left(-T_i\left(x\right) + 1\right) + H_i\left(x\right) * T_i\left(x\right) \tag{15}$$

In Equation (15), i indicates the final number of output layers. To prevent data leakage in the future, a convolution kernel (CK) $F = \left(f_1, f_2, \ldots, f_K\right)$ is used with an input sequence of $X = \left(x_1, x_2, \ldots, x_T\right)$, and the convolution definition expression at x_t is shown in equation (16).

$$\sum_{K=1}^{K} f_k x_{t+K-K} = \left(F * X\right)_{x_t} \tag{16}$$

A temporal processing method has been proposed for causal convolution constraints. In Figure 7, a schematic representation of the one-dimensional convolution in Equation (17) is presented.

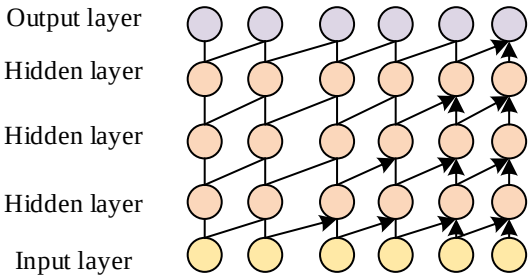


Figure 7 Schematic diagram of causal CNN

The CK of size 2 is consecutively shifted in Figure 5's input layer for convolution, and the output obtained is utilized as the input of the subsequent layer. The value of each node is only related to the nodes that have come before it, not the nodes that will come after. The closer the node is to the top layer, the larger its receptive field is. The formula for the receptive field RF_i of the first i layer is shown in Equation (17).

$$RF_i = RF_{i-1} + \left(k - 1\right)s \tag{17}$$

In Equation (17), i represents the number of layers, RF_{i-1} represents the receptive field of the previous layer, k represents the CK size, and s is the convolution step size. Therefore, the DSC-highway hybrid architecture significantly reduces the number of parameters through depthwise separable convolutions of equations (9) and (10). The stable propagation of gradient flow was achieved through the gating mechanism of equations (11) - (15), and finally the causal convolution of equations (16) and (17) ensured the causal constraints of the model in time. These equations collectively define an efficient and stable forward propagation process. The pseudocode of DSC-highway network is shown in Table 2.

Table 2: Pseudocode of DSC-highway network

DSC-highway network
Input: $X = \left(x_1, x_2, \ldots, x_t\right)$
Output: y_i
1: $y = x \bullet C_{\text{carry}}\left(x, W_{C_{\text{carry}}}\right) + H\left(x, W_H\right) \bullet T\left(x, W_H\right)$
2: $C_{\text{carry}} = 1 - T$
3: $y = H\left(x, W_H\right) \bullet T\left(x, W_H\right) + x \bullet \left(1 - T\left(x, W_{C_{\text{carry}}}\right)\right)$
4: $y = \begin{cases} x, T\left(x, W_T\right) = 0, \\ H\left(x, W_H\right), T\left(x, W_T\right) = 1. \end{cases}$

$$\frac{dy}{dx}=\begin{cases}1, T\left(x, W_T\right)=0 \\ H'\left(x, W_H\right), T\left(x, W_T\right)=1\end{cases} \quad (5)$$

$$\left(F * X\right)_{x_t}=\sum_{K=1}^{K} f_K x_{t-K+K} \quad (6)$$

$$RF_i=RF_{i-1}+\left(k-1\right)s$$

$$y_i=H_i\left(x\right) * T_i\left(x\right)+x_i *\left(1-T_i\left(x\right)\right) \quad (7)$$

Through these improvements, it is easier to accelerate the convergence speed of the model, reduce the computational burden caused by deep networks, and make the system more efficient and accurate in handling Chinese TTS tasks. The training is carried out using the Noam learning rate decay approach, and Adam is chosen as the optimizer. The purpose of Noam learning rate decay is to use a larger learning rate in the early stages of training to accelerate learning, and gradually reduce the learning rate as training progresses to avoid numerical instability in the later stages of training.

4 Result analysis of Chinese TTS system based on Seq2Seq model and CNN

4.1 Performance analysis of optimized CNN framework

The starting learning rate is 0.001. To quantify the statistical significance of performance differences, this study conducted paired t-tests on all indicators and calculated a 95% confidence interval (95% CI) for MOS. The significance level of all statistical tests is set to α=0.05. To investigate the impact of CK size on model performance, four models with CK sizes of 2, 3, 5, and 7 were set up. Using these four models to synthesize speech, the evaluation indicators are MOS, MCD, and the receptive field of the enhancement module. Among them, a total of 40 listeners participated in MOS scoring, and all listeners were trained and familiar with MOS scoring standards. The MOS rating is conducted in a controlled environment, using standardized audio equipment and ensuring that the ambient noise is below 30dB. From the test set, 30 texts were chosen at random. The effect of CK size on model performance is shown in Figure 8.

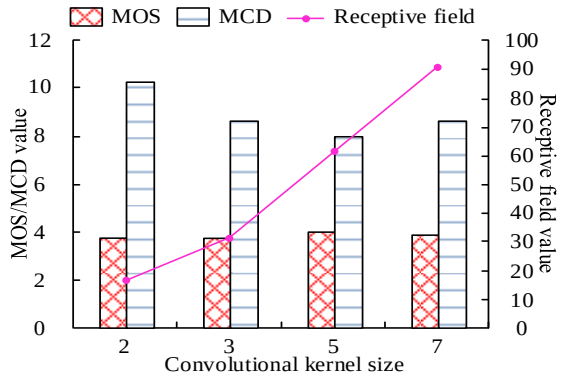


Figure 8: The Effect of CK Size on model performance

In Figure 8, the receptive field of convolution increases as CK increases. At a CK size of 5, the model MOS is 3.928±0.15 (95% CI) and the Chinese TTS quality is best. This is because a smaller CK will have a smaller receptive field and be unable to obtain longer information, while a larger CK will result in overfitting and large computation. Compared with the models of CK size 3 and CK size 57, the MCD reduction of CK size 5 was statistically significant (p<0.05). Therefore, CK size 5 is the most appropriate.

Thirty randomly selected texts from the test set were synthesized using DSC models with different channel numbers. The MOS, MCD and model parametric quantities of the Chinese TTS varied with the number of channels, as shown in Figure 9.

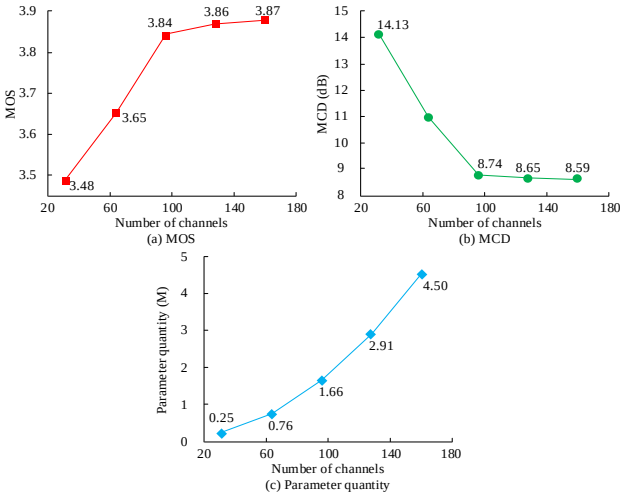


Figure 9: Comparison chart of the impact of channel number on model performance

In Figure 9(a), the Chinese TTS quality is at its lowest when the DSC model has 32 channels, and the MOS is at its lowest at this point, 3.48. When the number of channels is 160, the MOS reaches its highest value of 3.87 ± 0.11 (95% CI), indicating that the TTS quality in China is in the best state. Figure 9(b) shows that the MCD is at its highest when there are 32 channels, at 14.13 dB. When the number of channels is 160, the lowest MCD is 8.59 dB. Compared with 32 channels, both MOS enhancement and MCD reduction have high statistical significance (p<0.01). The MOS value steadily rises and the MCD value gradually falls as the number of channels rises, showing that the Chinese TTS quality improves and implying that adding more channels enables the DSC network to pick up more features. In Figure 9(c), the number of model parameters increases gradually from 0.25M to 4.50M as the number of channels increases from 32 to 160. Therefore, the number of channels is 96 to ensure the number of parameters and the quality of the Chinese TTS.

To explore the effect of having highway structure on network training, the training convergence curves at 10-, 50- and 100-layers depth were compared, as shown in Figure 10.

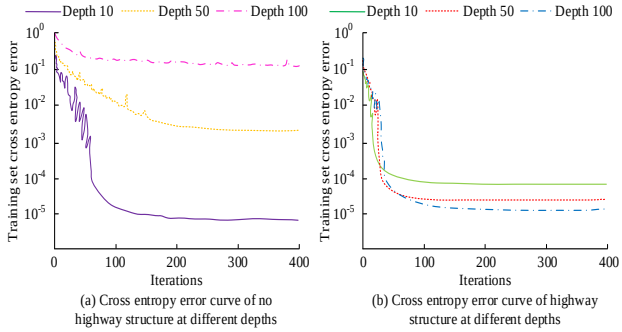


Figure 10: Comparison diagram of cross entropy error curve of training set

The cross-entropy error (CEE) curve of the training set for the network without highway structure in Figure 10(a) shows a decreasing trend as the number of iterations rises at 10-, 50-, and 100-layer depths. When the number of iterations is 200, the error curve gradually converges at the depth of 10 layers. In Figure 10(b), the curves for all three depths start to converge when the number of iterations is 50. The training set with the highway structure has a CEE of 1×10^{-4} when the depth is 10 layers. When the depth is 100 layers, the neural network with the highway structure converges more easily, and the CEE of the training set is 1×10^{-5}. This indicates that the highway structure can avoid the difficulty of training the network due to the obstruction of the gradient information backflow when the depth of the network deepens.

4.2 Analysis of the results of Chinese TTS system

The datasets selected for the experiment include the LJSpeech dataset and the LibriTTS dataset. The LJSpeech dataset is a public domain speech dataset containing 13100 short audio clips with a total duration of approximately 24 hours. Randomly select 5% of the data as the test set and the remaining 95% as the training set [19]. LibriTTS is an English multi speaker language library for TTS, originating from LibriSpeech. It has thousands of speakers (approximately 2456, gender balanced), a sampling rate of typically 24 kHz, and a total duration of approximately 585 hours [20]. In the designed system, to obtain higher quality synthesized audio, the filter group with the highest frequency $f_{\max} = 8000Hz$ and the number of filter groups $M = 24$ in the Mel spectrum extraction process is set. At the same time, to study the effectiveness of the Chinese TTS system, the study will compare the loss curves of the system before and after improvement, as shown in Figure 11.

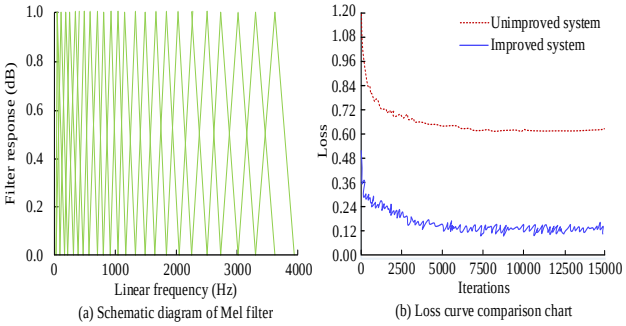


Figure 11: Improved Mel filter for system and its loss comparison

In Figure 11(a), the larger the linear frequency, the wider the weight distribution, the denser the weight distribution in the linear frequency within 1000 Hz, and the weight distribution from 1000 Hz to 4000 Hz gradually becomes wider filter response is kept at 0 to 1.0 dB. In Figure 11(b), the loss curve for the CNN model starts to converge at 10,000 iterations, and the loss value is 0.60 at this point. The loss curve for the DSC model starts to converge at 6000 iterations, and the loss value is 0.12 when the number of iterations reaches 10,000. All of the test set's sentences are used for evaluation in the objective test. The training time of the Chinese TTS system was 16.57h, MOS was 4.15, and MCD was 4.5288dB, indicating that the high-quality Chinese TTS was achieved in a short training time under the condition of limited corpus.

To verify the effectiveness of the optimized model, the accuracy of different optimized CNN models was compared. The comparison models are CNN-Bidirectional Long Short-Term Memory (BiLSTM) [21], Deep Neural Network and Two-Dimensional CNN (DNN-2DCNN) [22], and Deep Learning Speech Communication and Transmission (DeepSC-ST) [23]. The CNN BiLSTM system classifies speech signals using Mel spectrograms, short-term energy signal features, and a hybrid CNN-BiLSTM model. The DNN-2DCNN system selects DNN-2DCNN to predict speech data, and uses a neural encoder to achieve Chinese TTS in intracranial EEG signals. The DeepSC ST system identifies features through a joint semantic channel encoder and then generates speech signals in a neural network model. The comparison of accuracy obtained by different models on different datasets is shown in Figure 12.

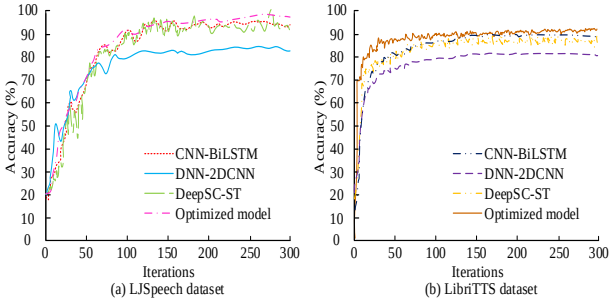


Figure 12: Comparison of accuracy obtained by different models on different datasets

In Figure 12(a), the optimized model has the highest accuracy on the LJSpeech dataset, at approximately 99.25%. In contrast, DNN-2DCNN has a lower value of approximately 82.96%. The accuracy of the optimized model is significantly higher than that of DNN-2DCNN, with a difference of 16.29%. This is because relying on the DNN-2DCNN model to predict speech data is unstable. The accuracy curve of DeepSC ST fluctuates more significantly, indicating weaker stability of the model. In Figure 12(b), CNN-BiLSTM performs better on the LibriTTS dataset, outperforming DeepSC ST by 8.50%. The highest accuracy of the improved system is 91.02%, because the optimized model improves accuracy by optimizing the network structure, solving the problems of training efficiency and performance stability.

The comparison of the training results of the system before and after improvement on different language datasets is shown in Figure 13.

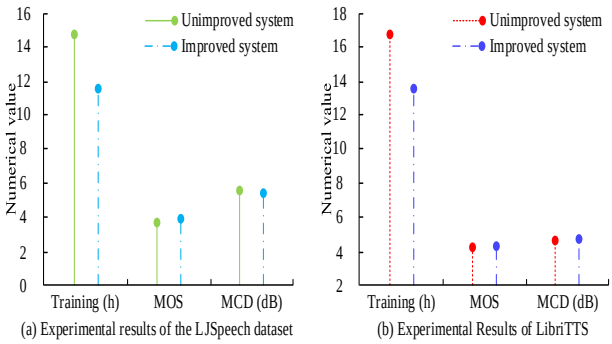


Figure 13: Comparison of system training results before and after improvement

In Figure 13(a), the improved model training time is reduced to 77.93% of the original value and the number of model parameters is decreased to 44.03% of the original value when the number of training steps is 90k at the same number of model training iteration steps. The improved system DSC slightly outperforms the CNN system in terms of synthetic speech quality. When there are 90k training steps, as shown in Figure 13(b), the improved model's training time is around 82.86% of the original, while the model's parameters are approximately 44.12% of the original. There isn't much of a difference between the two systems when comparing MOS and MCD values. In addition to cutting training time without compromising Chinese TTS quality, the improved system shrinks the size of the model, which saves storage space and computational resources while also achieving model optimization in both training time and space dimensions.

To verify the comprehensive performance of the proposed model, it was compared with the relative entropy alignment and multi-level attention fusion network (REA-MAF), FastSpeech2, Tacotron 2, and WaveGlow (TW). REA-MAF refers to the extraction of multi-level acoustic information under the action of relative entropy alignment and multi-level attention fusion network, and the complementary features between modalities to achieve information exchange [24]. FastSpeech2 refers to the ability to quickly generate high-quality synthesized speech on its own, enhancing audio reading and writing capabilities [25]. TW refers to fine-tuning and pre training the English Tacotron 2 and WaveGlow models to generate natural speech on the dataset [26]. These four models have been analyzed from different perspectives and subjected to t-tests to further validate the statistical significance of these performance differences. The test results are shown in Table 3.

Table 3: Performance results of different models

Model	REA-MAF	FastSpeech2	TW	The proposed model
MCD (dB)	4.7123	4.5891	4.8234	4.5288
MOS	4.02	4.10	3.95	4.15
Number of parameters ($\times 10^6$)	90.8431	81.8945	102.2741	68.4487
Training time (h)	20.12	18.34	21.45	16.57
FLOPs ($\times 10^{12}$)	2.20	2.00	2.50	1.75
Memory usage (GB)	8.5	7.8	9.3	6.2
Inference time (s)	0.18	0.17	0.20	0.15

In Table 3, after paired *t*-test verification, compared with the REA-MAF, FastSpeech2, and TW models, the differences in MCD, MOS, parameter quantity, and other indicators of this model reached statistical significance levels (all p-values<0.05). Specifically, the MOS score of this model is significantly higher than FastSpeech2 (p=0.032), and the MCD value is significantly lower than REA-MAF (p=0.017). Compared with the latest Transformer based models, this model still has a competitive advantage in the naturalness of synthesized speech. The MOS score is the highest at 4.15, while the MCD score is the lowest at only 4.5288 dB. This model achieves the best naturalness and accuracy in synthesized speech. The proposed model has a training time of 16.57 hours and a parameter count of 68.4487 × 10^6, which is lower than other models. Compared with the TW model, the training time is reduced by 4.88 hours and the parameter count is reduced by approximately 33.07%. The proposed model has significant optimization in computational efficiency. The FLOPs of the proposed model are 1.75 × 10^{12}, indicating better performance in terms of computational efficiency. In terms of inference time, the proposed model only takes 0.15 seconds, which is shorter than the inference time of other models. In comparison, the P value of the proposed model compared to the other three models was <0.05, indicating significant differences in all seven indicators. This is because the proposed model adopts DSC and efficient highway structure, reducing the computational complexity and

memory consumption of the network. This structure can significantly shorten the training time and reduce the number of model parameters while ensuring the quality of speech synthesis, thereby improving performance in multiple aspects.

5 Conclusion

To improve the effectiveness of Chinese TTS, this study introduced a scaled dot product attention mechanism in TTS. And a DSC highway hybrid architecture was built to solve the problems of parameter explosion and gradient instability in traditional CNN in Seq2Seq models. The results show that the training time of the Chinese TTS system is 16.57h, MOS is 4.15, and MCD is 4.5288dB. the loss curve of DSC model tends to be smooth at 10000 iterations, and the loss value is 0.12 at this time. When the DSC model has 32 channels, the MOS is the smallest at this point (3.48), and the MCD is the highest at this point (14.13), and this is when the Chinese TTS quality is at its lowest. The MOS value steadily rises and the MCD value gradually declines as the number of channels grows. The improved system slightly outperformed the original system in terms of Chinese TTS quality. When comparing the two systems' MOS and MCD values, there was minimal difference between them. The improved model's training time was roughly 82.86 percent of the original, and its model parameters were roughly 44.12 percent of the original. With more iterations, the CEE curve of the training set of the network without highway structure gets flatter at 10-, 50-, and 100-layer depths. When the number of iterations is 200, the error curve gradually converges at the depth of 10 layers. When the number of iterations is 50, the curves begin to converge for all three depths. When the depth is 10 layers, the CEE of the training set with the highway structure is 10^{-4}. When the depth is 100 layers, the neural network with the highway structure converges more easily, and the CEE of the training set is 10^{-5}. Therefore, the designed model can be applied to embedded devices such as smart speakers and car voice systems, and used as an educational aid in Chinese pronunciation evaluation systems. However, the research did not consider the real-time performance of the model, and in the future, dynamic pruning techniques can be combined to further compress the model to millisecond level response.

Acknowledge

This manuscript was supported by Project of Industry-University Cooperation and Collaborative Education of the Ministry of Education: Empowering International Chinese Language Teachers with Generative AI Technology (Project Number: 250305061134614).

References

[1] Kasparaitis P. Evaluation of Lithuanian Speech-to-Text Transcribers. Informatica, 2025, 36(2): 369-384. https://doi.org/10.15388/25-INFOR591

[2] Zheng J, Zhou J, Zheng W, Tao, L., & Kwan, H. K. Controllable Multi-Speaker Emotional Speech Synthesis with an Emotion Representation of High Generalization Capability. Affective Computing, IEEE Transactions on, 2025, 16(1):68-82. https://doi.org/10.1109/TAFFC.2024.3412152

[3] Kaur N, Singh P. Conventional and contemporary approaches used in text to speech synthesis: A review. Artificial Intelligence Review, 2023, 56(7): 5837-5880. https://doi.org/10.1007/s10462-022-10315-0

[4] Du C, Guo Y, Chen X, Chen, X., & Yu, K. Speaker adaptive text-to-speech with timbre-normalized vector-quantized feature. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 3446-3456. https://doi.org/10.1109/TASLP.2023.3308374

[5] Jiang C, Gao Y, Ng W W Y, Zhou, J., Zhong, J., & Zhen, H. SeDepTTS: Enhancing the naturalness via semantic dependency and local convolution for text-to-speech synthesis. Proceedings of the AAAI Conference on Artificial Intelligence. 2023, 37(11): 12959-12967. https://doi.org/10.1609/aaai.v37i11.26523

[6] Tan X, Chen J, Liu H, Cong, J., Zhang, C., Liu, Y., ... & Liu, T. Y. Naturalspeech: End-to-end text-to-speech synthesis with human-level quality. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(6): 4234-4245. https://doi.org/10.1109/TPAMI.2024.3356232

[7] Li N, Liu Y, Wu Y, Liu S, Liu M. RobuTrans: A robust transformer-based text-to-speech model. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5): 8228-8235. https://doi.org/10.1609/aaai.v34i05.6337

[8] Deng C, Lam W. Graph transformer for graph-to-sequence learning. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5): 7464-7471. https://doi.org/10.18653/v1/2020.acl-main.640

[9] Han S, Um J, Kim H. One-shot multi-speaker text-to-speech using RawNet3 speaker representation. Phonetics and Speech Sciences, 2024, 16(1): 67-76. https://doi.org/10.13064/KSSS.2024.16.1.067

[10] Nazir O, Malik A, Singh S, & Pathan, A. S. K. Multi speaker text-to-speech synthesis using generalized end-to-end loss function. Multimedia Tools and Applications, 2024, 83(24): 64205-64222. https://doi.org/10.1007/s11042-024-18121-2

[11] Fujita K, Ando A, Ijima Y. Speech rhythm-based speaker embeddings extraction from phonemes and phoneme duration for multi-speaker speech synthesis. IEICE TRANSACTIONS on Information and Systems, 2024, 107(1): 93-104. https://doi.org/10.1587/transinf.2023EDP7039

[12] Widiyanto W W, Rizki U, Susena E. Increased accuracy of sequence-to-sequence models with the CNN algorithm for multi response ranking on travel service conversations based on chat history. Jurnal Infotel, 2020, 12(2): 39-44. https://api.semanticscholar.org/CorpusID:225874443}

[13] Amin J, Anjum M A, Sharif M, Kadry S, Wang S H. Convolutional Bi-LSTM based human gait recognition using video sequences. Computers, Materials and Continua, 2021, 68(2): 2693-2709. https://doi.org/10.32604/cmc.2021.016871

[14] Yang Q, Jin W, Zhang Q, et al. Mixed-modality speech recognition and interaction using a wearable artificial throat. Nature Machine Intelligence, 2023, 5(2): 169-180. https://doi.org/10.1038/s42256-023-00616-6

[15] Bao H, Dong L, Wei P F. Fine-tuning pretrained transformer encoders for sequence-to-sequence learning. International journal of machine learning and cybernetics, 2024, 15(5):1711-1728. https://doi.org/10.1007/s13042-023-01992-6

[16] Li Y, Anumanchipalli G K, Mohamed A, et al. Dissecting neural computations in the human auditory pathway using deep neural networks for speech. Nature Neuroscience, 2023, 26(12): 2213-2225. https://doi.org/10.1038/s41593-023-01468-4

[17] Enrico V, Pierre G, Tobias R. AVbook, a high-frame-rate corpus of narrative audiovisual speech for investigating multimodal speech perception. The Journal of the Acoustical Society of America, 2023, 153(5):3130-3137. https://doi.org/10.1121/10.0019460

[18] Chen X, Wang R, Khalilian-Gourtani A, et al. A neural speech decoding framework leveraging deep learning and speech synthesis. Nature Machine Intelligence, 2024, 6(4): 467-480. https://doi.org/10.1038/s42256-024-00824-8

[19] Metzger S L, Littlejohn K T, Silva A B, Moses, D. A., Seaton, M. P., Wang, R., et al. A high-performance neuroprosthesis for speech decoding and avatar control. Nature, 2023, 620(7976): 1037-1046. https://doi.org/10.1038/s41586-023-06443-4

[20] Haider C L, Park H, Weisz H N. Neural Speech Tracking Highlights the Importance of Visual Speech in Multi-speaker Situations. Journal of cognitive neuroscience, 2024, 36(1):128-142. https://doi.org/10.1162/jocn_a_02059

[21] 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

[22] Arthur F V, Csapó T G. Speech synthesis from intracranial stereotactic Electroencephalography using a neural vocoder. Infocommunications journal: a publication of the scientific association for infocommunications (HTE), 2024, 16(1): 47-55. https://doi.org/10.36244/ICJ.2024.1.6

[23] Weng Z, Qin Z, Tao X, Pan, C., Liu, G., & Li, G. Y. Deep learning enabled semantic communications with speech recognition and synthesis. IEEE Transactions on Wireless Communications, 2023, 22(9): 6227-6240. https://doi.org/10.1109/TWC.2023.3240969

[24] Lei J, Wang J, Wang Y. Multi-level attention fusion network assisted by relative entropy alignment for multimodal speech emotion recognition. Applied Intelligence, 2024, 54(17):8478-8490. https://doi.org/10.1007/s10489-024-05630-8

[25] Firdaus M R, Firdaus M R, Pasha P Y. Text-to-Speech Technology Development Using FastSpeech2 Algorithm for the Story of the Prophet. Khazanah Journal of Religion and Technology, 2024, 2(2): 55-62. https://journal.uinsgd.ac.id/kjrt

[26] Rai A, Shiwakoti S, Basukala S, & Dahal, S. S. Nepali Text-to-Speech Synthesis Using Tacotron2 and WaveGlow. KEC Journal of Science and Engineering, 2024, 8(1): 103-109. https://doi.org/10.3126/kjse.v8i1.69276

Cross-Lingual Approaches for Text Difficulty Classification in Non-English Languages

Huili Dai
Zhengzhou Science and Technology University, Zhengzhou 450064, China
E-mail: daihuili603@126.com

Keywords: Difficulty classification, natural language processing, cross-lingual models, machine translation, machine learning

Received: April 20, 2025

Text Difficulty Classification (TDC) significantly contributes to educational technology, language learning, and information access by automatically predicting the difficulty of texts. While English has been favored with extensive resources and mature readability tools, non-English languages suffer from sparse annotated corpora, linguistic diversity, and insufficient computational resources. Cross-lingual methods, including shared knowledge transfer from high-resource to low-resource languages, have become increasingly popular in recent years. It covers the latest advances in cross-lingual TDC in pipelines of machine translation-based, multi-lingual pre-trained transformers (e.g., mBERT, XLM-R), adversarial and meta-learning-based approaches, and knowledge distillation techniques. It also summarizes available datasets and benchmarks, critiques current practices, and highlights morphological complexity, translation artifacts, and domain mismatch. It identifies significant trends and future promise by critically examining state-of-the-art methods for multimodal TDC, hybrid architectures, and active learning in low-resource languages. The results highlight the significance of non-discriminatory, resource-lean systems for measuring fair-minded text difficulty in many languages.

Povzetek: Pregledni članek povzema navzkrižno-jezikovne pristope k razvrščanju težavnosti besedil, zbere nabore podatkov in kritično izpostavi ter trende k večmodalnim, hibridnim in aktivnim metodam v jezikovno skromnih okoljih.

1 Introduction

Text Difficulty Classification (TDC) is the basis of next-generation learning technology, language learning software, and Natural Language Processing (NLP) solutions [1]. Accurate identification of textual complexity enables adaptive learning platforms to recommend material commensurate with the student's proficiency, ensuring interest and comprehension [2]. TDC facilitates teachers and learners in language pedagogy by simplifying material choice and automatically scoring exercises [3]. Beyond pedagogy, TDC enables access by adapting information for diverse users, such as non-native language users or those with reading impairments. The explosive nature of NLP-based solutions, such as summarizing and suggesting content, has expanded the requirement for resilient TDC solutions [4]. As technology continues to serve as a middleman for knowledge, automating the determination of reading levels has become crucial to delivering fair and efficient learning experiences worldwide [5].

Despite these advances, most advances in TDC have centered on English, for which vast annotated corpora, readability algorithms, and pre-trained models have spurred the development of practical systems. Non-English languages are underrepresented by sparse labeled datasets, linguistic variability, and limited computational resources [6]. Linguistically elaborate languages, such as Finnish, agglutinative languages, like Turkish, and languages with non-Latin scripts, are especially problematic for feature extraction and modeling. Cultural and educational variations similarly affect text richness, making it challenging to use English-centric tools directly [7]. These resource disparities have resulted in unequal access to advanced readability technologies, denying many language populations sufficient tools to support adaptive learning or content personalization. Overcoming these gaps is critical to bringing TDC's benefits to a broader, multi-lingual population.

Cross-lingual techniques offer a solution to fill these resource gaps. They can construct TDC systems with limited labeled data by leveraging knowledge transfer from high-resource languages like English to under-resourced languages [8]. Machine translation-based pipelines, transformer-based multi-lingual models (e.g., XLM-R, mBERT), and adversarial learning allow English-trained models to generalize effectively to unseen languages. For example, multi-lingual embeddings project universal semantic structures into common spaces for encoding, while meta-learning algorithms rapidly learn to adapt to new languages under sparse supervision [9]. In addition to reducing the use of large sets of annotated data, they enable potential lifts to TDC for underrepresented languages and more comprehensive, scalable solutions.

The present study encompasses cross-lingual TDC techniques from both computational and linguistic perspectives. From a computational point of view, we critique algorithmic techniques such as zero-shot transfer, multi-lingual pre-training, and knowledge distillation. Theoretically, we examine language structure differences, morphology, syntax, and orthography concerning system design and efficacy. Cultural background and educational structure are also addressed in terms of their impacts on measuring text difficulty. By synthesizing findings from numerous recent research studies, the study identifies best practices, highlights long-standing challenges, and suggests ways to overcome existing obstacles. The dual computational and linguistic approach ensures a subject-wide understanding of the field and its potential to influence global education and NLP applications.

The rest of the paper is organized as follows. Section 2 provides a background on TDC and its evolution from early readability formulae to state-of-the-art neural techniques. Section 3 overviews cross-lingual methods, including translation-based pipelines, multi-lingual pre-trained models, adversarial, and alignment-based techniques. Section 4 presents an overview of state-of-the-art datasets, benchmarks, and evaluation protocols, and discusses notable challenges such as morphologically challenging texts and translation-based artifacts. Section 5 summarizes the main findings and proposes resource-aware and inclusive TDC solutions.

2 Background and foundations

It is desirable to understand early TDC development to contextualize the development of cross-lingual techniques. Early TDC research relied upon manually crafted readability formulae and superficial linguistic characteristics to gain quick, superficial insights into reading ease. As computational linguistics and NLP advanced, techniques have become increasingly underpinned by machine learning and, more recently, transformer-based architectures to expose high-level syntactic and semantic structures. This section provides an overview of the salient concepts, outdated techniques, and technological milestones that underpin today's TDC systems, laying the groundwork for descriptions of cross-lingual methodologies.

2.1 Definition of text difficulty classification

TDC is a basic NLP task that automatically determines the complexity level suitable for a particular audience or purpose [10]. Although typical formulae for readability focus primarily on superficial aspects, such as sentence length or word count, the latest TDC methods utilize syntactic, semantic, and discourse-level features to provide a more fine-grained analysis. Advanced Transformer language models, combined with multi-lingual embeddings and feature engineering techniques, now enable TDC systems to account for the subtleties of prose difficulty in different languages and contexts.

Table 1: Applications of text difficulty classification

Application area	Description	Example use cases
Adaptive learning	Matches reading materials to a learner's proficiency to optimize engagement and comprehension.	Personalized e-learning platforms, digital textbooks.
Second-language learning	Aligns texts to frameworks like CEFR and supports graded reading exercises.	Language learning apps, automated test generation.
Text simplification	Identifies and modifies complex passages for easier reading.	Simplified news articles, accessibility for dyslexia.
Content accessibility	Adapts information for non-native speakers or audiences with varied literacy levels.	Healthcare instructions, government resources.
Information retrieval	Enhances search and recommendation systems by filtering or ranking texts based on difficulty.	News curation, personalized search engines.
Cross-lingual education	Aligns text difficulty across languages for bilingual or multi-lingual contexts.	Comparative linguistics and bilingual education tools.

As indicated in Table 1, TDC has broad applications in education, access, and information dissemination. Adaptive learning platforms ensure that the learners are given sufficiently challenging material. TDC aligns texts to schemes, such as the CEFR of second language learning, and allows automatic scoring. In sentence condensation for simplified texts, sentence structure must be simplified to facilitate reading for children, non-native readers, or learners with learning disabilities. TDC is also used to adapt news, healthcare, and legal texts to make them easily readable by various user groups. Finally, for cross-lingual cases, TDC enables comparative analysis of languages and matching of bilingual resources.

2.2 Traditional approaches

Traditional techniques of TDC primarily concern formulae of readability based on features. These evaluate superficial characteristics such as mean sentence length, word length, and number of syllables to estimate the hardness of documents. User-friendly measures, such as Flesch Reading Ease, Flesch-Kincaid Grade Level, Gunning Fog Index, and SMOG, are presented in Table 2. Because they require minimal computation and are easy to use, they have become cornerstones for teaching and publishing. Although they strongly correlate to simple English reading difficulty and produce quick, interpretable output, they neglect advanced language features such as syntax, semantics, and discourse structure. Thus, they tend to overlook subtle changes in difficulty, especially in texts that contain challenging vocabulary, specialized terms, and non-standard structures.

Feature-based formulae for readability are significantly constrained when applied to non-English languages [11]. Formulas can be specifically adjusted to morpho-syntactic English characteristics and thus may be

Table 2: Common feature-based readability formulas and their limitations

Formula	Key features used	Strengths	Limitations for non-English texts
Flesch reading ease	Sentence length and syllable count	Simple, interpretable, and widely adopted	Ineffective for agglutinative or logographic languages; ignores syntax/semantics.
Flesch-Kincaid grade	Sentence length and syllable count	Maps scores to U.S. grade levels	English-specific calibration; unsuitable for tonal or morphologically rich languages.
Gunning fog index	Sentence length and complex word frequency	Identifies overly complex sentences	Overestimates difficulty in inflection-heavy languages.
SMOG index	Polysyllabic word frequency and sentence count	Useful for shorter texts	Relies on syllable-based measures, not applicable to all languages.
Coleman-Liau index	Character and word counts	Avoids syllable counting and easy to compute	Sensitive to alphabet differences; fails for non-alphabetic scripts.

less reliable in languages with different structures. For example, agglutinative languages like Turkish or Finnish pack strong grammaticality into bound words; hence, word length per se or word-based syllable count are unreliable measures of difficulty. Similarly, tonal languages (e.g., Mandarin Chinese) or logographic script languages do not enjoy direct syllable-oriented metric analogs. Moreover, the cultural and educational background in which difficulty is intuitive implies that English-based thresholds can conflict with the complexities of texts in other languages. Due to these challenges, various solutions or language-specific variations of standard formulae are necessary to estimate accurately.

The weaknesses of standard formulae have motivated the development of higher-level computational TDC models. More advanced methods simulate linguistic features, such as part-of-speech tendencies, syntactic tree depth, and semantic similarity, to gain a more fine-grained understanding of text difficulty. Cross-lingual and multi-lingual methods, in particular, aim to overcome English-biased formula structure biases by leveraging common embeddings or translation-based pipelines. While traditional metrics are fine for primary analysis or comparison to baselines, they fall short of the current diversified linguistic climate. Identifying these weaknesses has inspired research into language-agnostic, resilient, and context-sensitive methods for classifying text difficulty.

2.3 Advances in NLP and multi-lingual models

Recent advances in NLP revolutionized TDC through the advent of pre-trained transformer architectures. Pretrained transformer-based models such as BERT, RoBERTa, and their multi-lingual counterparts (mBERT, XLM-RoBERTa), listed in Table 3, exploit self-attention mechanisms and large-scale unsupervised learning to learn rich contextual representations of texts. Those models encode superficial features and syntax-based, semantic, and discourse-level structures that classical formulas cannot address [12]. Their ability to represent words and sentences in high-dimensional word spaces enables more accurate estimations of text difficulty. For cross-lingual analysis, multi-lingual transformer-based models trained concurrently in dozens of languages create common embedded spaces, transferring knowledge from high-resource to low-resource languages. That has opened novel opportunities for zero-shot and few-shot TDC without requiring large label sets for every target language.

Multi-lingual word and sentence embeddings, such as MUSE, LASER, and LaBSE, can potentially bridge linguistic gaps [13]. These embeddings place semantically similar words/phrases from varying languages into the same vector space, allowing for a comparative study of text difficulty across languages. English-based TDC, for instance, can be employed to classify texts from German or Spanish by relying on common representations of embeddings. Machine translation pipelines, in contrast, can inject translation artifacts. However, multi-lingual embeddings preserve language-specific information while maintaining language comparability. Their application is particularly beneficial in low-resource settings where annotated datasets are unavailable. By utilizing these embeddings, scholars can conceptualize language-agnostic large-scale TDC systems that can be applied to different linguistic populations.

Combined pre-trained transformers and multi-lingual embeddings have significantly enhanced the efficiency and transferability of TDC systems [14]. Backward-compatible methods, such as knowledge distillation and

Table 3: Key multi-lingual models and embedding techniques for cross-lingual TDC

Model/embedding	Type	Languages supported	Key strengths	Example use cases
mBERT	Transformer	100+	Shared embeddings and zero-shot transfer	Cross-lingual TDC, classification tasks
XLM-RoBERTa	Transformer	100+	High performance across low-resource languages	TDC for underrepresented languages
LASER	Sentence embedding	90+	Language-agnostic sentence vectors	Multi-lingual text similarity, TDC
LaBSE	Sentence embedding	100+	Strong alignment for sentence-level tasks	Cross-lingual semantic analysis and TDC
MUSE	Word embedding	30+	Aligns monolingual embeddings into a shared space	Low-resource TDC, bilingual lexicon induction

adapter layers, encourage model adaptability at lower computational costs without compromising precision. Fine-tuning multi-lingual transformers over relatively small labeled sets of the target language has achieved competitive results, even in low-resource environments. It has shifted the paradigm from language-specific, handcrafted features to universal, context-sensitive representations that enable global education, accessibility, and information seeking. As shown in Table 3, state-of-the-art multi-lingual models and embedding methods provide a robust foundation for cross-lingual text difficulty analysis.

3 Cross-lingual approaches

Cross-lingual techniques have emerged as the preferred method of translating TDC from English to resource-scarce languages [15]. Transferring knowledge from high-resource to low-resource languages that are poorly annotated balances the disparity in linguistic resources and enables scalable, multi-lingual solutions. Rather than learning new models for each language, cross-lingual techniques use common representations, machine translation, or alignment protocols to bridge linguistic divides.

Figure 1 provides an at-a-glance overview of the principal categories of cross-lingual techniques, translation-based pipelines, multi-lingual pre-trained models, adversarial and meta-learning techniques, and knowledge distillation, highlighting their relationships and complementary roles. This section addresses how these techniques support TDC in non-English languages.

3.1 Machine translation-based methods

Machine translation-based methods are among the earliest and lightest-weight techniques for non-English TDC support. The general workflow, shown next to its strengths and weaknesses in Table 4, is to translate a source language sentence into English using a machine translation system, classify the output English translation using a pre-trained English TDC model, and project the output back to the original language [16]. The resulting pipeline capitalizes on the maturity and availability of English readability tools and large sets of labeled English corpora. By bypassing the requirement of maintaining large annotated corpora for each target language, it can be employed for fast cross-lingual TDC support, especially when resources for one language are limited.

The advantages of machine translation-based methods include their generic nature and broad applicability. High-performance translation systems, such as Google Translate, DeepL, and MarianMT, can be utilized to translate hundreds of languages; therefore, TDC can be deployed for underrepresented languages without the need to develop new classifiers from scratch. It is also desirable for domains of high translation quality, such as formal or informational texts. Additionally, machine translation pipelines are typically the baseline method for evaluating more advanced cross-lingual techniques and constitute a real-world standard for model quality. For example, those evaluating TDC in German or Spanish can assess their multi-lingual transformer models against a translation-based baseline to estimate the value of cross-lingual pre-training.

Despite its advantages, this method has notable limitations. By adding noise, altering sentence structure, or replacing domain-specific words, translations can be deceptive in terms of the perceived complexity of the original. Cultural references, colloquialisms, and morphology may be underrepresented or overrepresented during translation, resulting in less accurate predictions. Translations are also subject to considerable variability in language pairs—high-resource pairs, such as English–Spanish, translate successfully, whereas, by and large, poor translations are obtained from English–Zulu, a low-resource pair. Translating significant texts is often computationally too costly even for non-realtime educational use cases at scale. These weaknesses

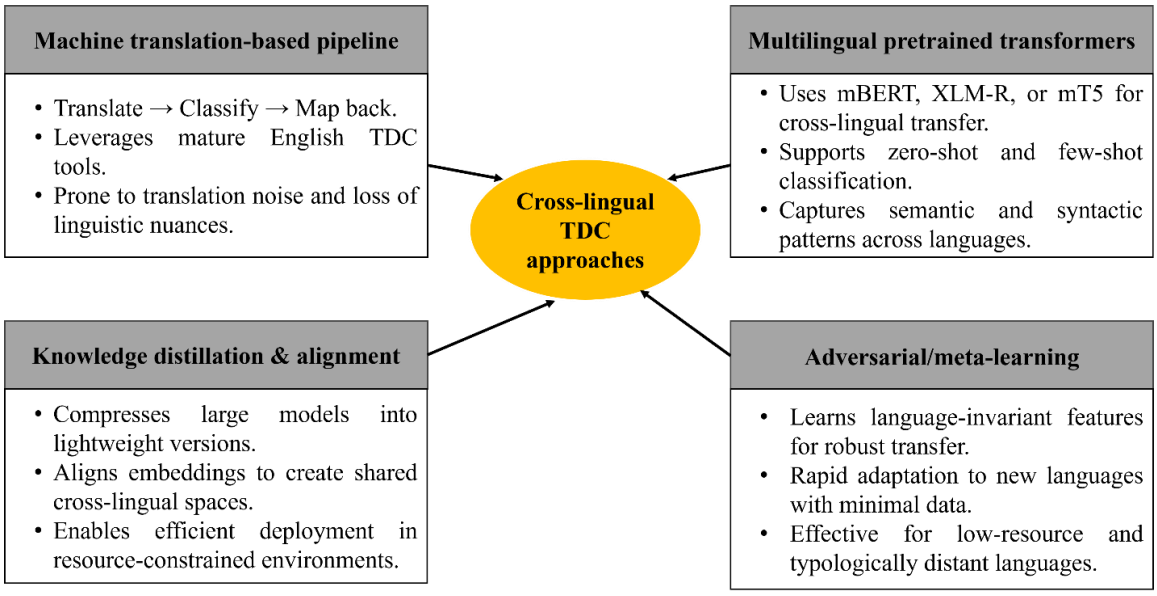


Figure 1: Overview diagram of cross-lingual TDC approaches

Table 4: Pros and cons of machine translation-based TDC

Aspect	Advantages	Limitations	Example use cases
Workflow	Simple pipeline; leverages existing English tools	Dependent on translation quality, it may distort the meaning	Quick deployment for new languages
Resource requirements	No need for target-language labeled data	Computationally expensive for large corpora	Baseline benchmarking for research
Performance	Effective for high-resource language pairs	Poor for low-resource pairs; loses cultural and linguistic nuances	Educational tools for well-supported languages

Table 5: Multi-lingual transformers for cross-lingual TDC

Model	Supported languages	Key strengths	Limitations	Example use cases
mBERT	100+	Zero-shot transfer; robust and widely adopted	Struggles with distant language pairs	Cross-lingual TDC, multi-lingual text classification
XLM-RoBERTa	100+	High performance for low-resource languages	Computationally expensive	Educational content recommendation, semantic similarity
mT5	100+	Supports classification and generation tasks	Resource-intensive training and fine-tuning	Text simplification, adaptive learning pipelines

necessitate the implementation of an alternative to translation-based methods, such as multi-lingual pre-trained models or adversarial learning.

3.2 Multi-lingual pretrained models

Pretrained multi-lingual models, such as mBERT, XLM-RoBERTa, and mT5, have transformed cross-lingual TDC by offering a single architecture for processing different languages. These models are trained on gigantic multi-lingual corpora and thus obtain shared representations that capture semantic and syntactic knowledge common across languages. As demonstrated in Table 5, their architectures can handle scripts and linguistic variations, allowing for knowledge transfer from high-resource languages (e.g., English) to less-resourced ones. The models do not use machine translation but represent texts from different languages within a common vector space, thereby reducing artifacts and preserving cultural or linguistic features.

The primary benefit of multilingual-pretrained models is that they enable zero- and few-shot learning. The TDC model, fine-tuned for English, can classify Spanish or German texts without additional training or with minimal labeled data. This is particularly desirable for languages with limited resources, for which it is not practical to build TDC-specific datasets. Usagees encompass adaptive learning platforms that suggest multi-lingual learning resources, cross-lingual search engines that sort texts by proficiency, and linguistic comparative research. Additionally, it is often possible to fine-tune these models for minor target language datasets, resulting in near-performances comparable to those of monolingual systems, highlighting their versatility and efficacy.

While influential, multi-lingual pre-training models come with many problems. Their computational cost is vast—training and fine-tuning require extensive hardware support, making it infeasible for small institutions. Their performance may be bad in typologically distinct or under-resourced languages for which little pre-training data exists or is of poor quality. Additionally, these models sometimes exhibit biases inherited from learning corpora, potentially warping the difficulty of prediction across cultural or linguistic domains. Complementary research may incorporate language-specific adapter addition, lifelong learning, or active learning methods to improve efficiency and fairness. Hybrid models that combine multi-lingual models with light-weight language-aware components may improve performance in diverse linguistic environments.

3.3 Adversarial and meta-learning techniques

Adversarial learning has emerged as an optimal approach to enhancing cross-lingual TDC by incorporating language-invariant features into models. Under this approach, the classifier predicts text difficulty, while the discriminator attempts to determine the language of the input [17]. The classifier is trained to deceive the discriminator, thereby causing it to focus on features related to complexity rather than language-related features. As Table 6 summarizes, the strategy reduces the application of superficial features, such as script or morphology, whose characteristics differ significantly from language to language. Adversarial learning is particularly valuable when labeled resources are in short supply, as it enables knowledge to be passed from high-resource to low-resource languages. It has emerged that adversarial-training models outperform regular fine-tuning methods in zero-shot and few-shot scenarios, proving an exemplary technology for multi-lingual learning platforms and adaptive learning systems.

Meta-learning techniques model TDC for each language as another similar task, allowing the model to learn. Model-Agnostic Meta-Learning (MAML) and related methods enable rapid adaptation to unseen languages, even with limited labeled data. For example, after being trained on many high-resource languages, the TDC model can be meta-learned to be updatable to a new low-resource language from only a few examples. That quick adaptability makes meta-learning particularly suitable for under-represented languages/dialects or variable teaching-learning scenarios, where new use domains emerge quickly. Meta-learning can be combined with adversarial techniques for better robustness and

Table 6: Adversarial and meta-learning for cross-lingual TDC

Approach	Key benefit	Limitation	Example use cases
Adversarial	Learns language-invariant representations	Requires careful tuning to avoid instability	Zero-shot TDC for multi-lingual education tools
Meta-learning	Rapidly adapts to new languages	Computationally expensive, task balancing needed	Few-shot TDC for low-resource languages
Hybrid methods	Combines robustness and adaptability	Complexity in implementation	Dynamic educational content recommendation

higher generalization across typologically diverse languages.

While adversarial and meta-learning techniques hold much promise, they are computationally costly and must be carefully tuned for hyperparameters to avoid instability or overfitting. Adversarial settings sometimes yield poor outcomes if the discriminator is overly strong or weak, and meta-learning necessitates balanced task sampling to achieve generalization. While they bring those issues, they also hold the potential to open the door to scaled TDC solutions that are friendly to limited resources. Future research can explore hybrid settings that integrate adversarial training with multi-lingual transformers or integrate curriculum learning to guide meta-learning processes. Light-weight adversarial settings or task-specific meta-learning fine-tuning can also render them possible for institutions with limited computational power.

3.4 Knowledge distillation and alignment

Embedding alignment is one of the basic techniques for achieving cross-lingual TDC without machine translation. It is comprised of casting monolingual word or sentence embeddings into a single semantic space such that texts from different languages can be analyzed comparatively directly. Programs like MUSE and VecMap learn linear transformations or higher-order embeddings between embedding spaces from bilingual dictionaries or parallel texts.

Once aligned, embeddings from low-resource languages can benefit from similarity relationships learned from high-resource languages, enabling classifiers from one language to generalize successfully to different languages. As Table 7 demonstrates, embedding alignment is significantly less computationally intensive than full retraining, while retaining language-specific nuances to a greater extent than translation-based pipelines. It is especially valuable for languages that lack large-scale, pre-trained multi-lingual models.

Distillation involves compressing large, high-performing multi-lingual models into smaller, light-weight models of comparable predictive quality. In distillation, a "teacher" model (e.g., XLM-RoBERTa) provides soft predictions or intermediate representations, and another "student" model is trained to emulate these. The student model can therefore replicate comparable performance at much lower computational cost. For TDC, distillation enables robust classifiers even on devices with limited computational power, such as mobile telephones or classroom tablets. It allows large-scale deployment in resource-scarce educational environments with limited computational infrastructure.

It is possible to augment distillation with embedding alignment to reduce cross-lingual quality and overhead.

Distilling knowledge and aligning embeddings offer significant value in terms of efficiency, scalability, and availability. Both enable us to use cross-lingual TDC systems in real-world tasks, from adaptive learning platforms to multi-lingual search engines. But they are also prone to pitfalls. The alignment quality mainly depends on the availability of appropriate bilingual dictionaries or parallel data, of which many under-resourced languages are scanty. Distilling is also at risk of discarding fine-grained linguistic information, especially when compressing large teacher models into light-weight student models. Future research can be conducted to investigate unsupervised alignment techniques, dynamic distillation mechanisms, and hybrid architectures that combine adapters or prompts with distilled models. It is through inventions of this nature that it can be ensured that resilient, economical cross-lingual TDC solutions are of high precision for a large number of languages.

4 Discussion

The preceding sections lay out the developments and challenges of cross-lingual TDC. The prior analysis combines knowledge from datasets and benchmarks, evaluation procedures, challenges and constraints, and the future, providing an evaluative summary of the current state of the field and its innovation potential.

Table 7: Knowledge distillation and alignment in cross-lingual TDC

Technique	Advantage	Limitation	Example applications
Embedding alignment	Enables direct comparison of texts across languages; light-weight	Requires bilingual dictionaries or parallel corpora	Low-resource TDC, multi-lingual similarity tasks
Knowledge distillation	Compresses large models for efficient deployment	Potential loss of nuanced knowledge	Mobile TDC apps, scalable educational platforms
Combined approaches	Balances accuracy and efficiency	Complexity in implementation	Cross-lingual adaptive learning tools

Table 8: Overview of datasets and benchmarking strategies for cross-lingual TDC

Category	Examples	Strengths	Limitations	Potential improvements
Non-English corpora	Klexikon (German) amd CEFR-Spanish	Tailored to specific languages and audiences	Limited size and narrow domain coverage	Expand to multiple genres and proficiency levels
Multi-lingual benchmarks	CLEAR, XGLUE, and MLQA	Enables fair cross-lingual comparison	Focuses on popular languages	Include underrepresented and low-resource languages
Data augmentation methods	Back-translation and synthetic data	Expands resources for low-resource languages	Risk of noise, cultural bias, or mislabeling	Combine augmentation with expert validation
Crowdsourced labeling	Amazon MTurk and Prolific	Scalable and cost-effective	Inconsistent quality and annotator expertise	Use tiered review systems and clear guidelines

4.1 Datasets and benchmarks

Access to reputable datasets and benchmarks is key to the progress of cross-lingual TDC. Without proper, diversified, and representative corpora, the testing and development of TDC systems can become biased toward a few high-resource languages and domains. Non-English datasets, such as Klexikon (German encyclopedia texts for children) and CEFR-aligned Spanish datasets, described in Table 8, serve as valuable resources for TDC progress in specific languages. However, these resources are of limited scope and availability, typically restricted to particular domains such as educational texts or children's literature. Limited coverage of this kind prevents scaling TDC models to more complex writing varieties, such as news articles, blog posts, or social media.

Multi-lingual benchmarks, such as CLEAR, XGLUE, and MLQA, can be analyzed comparatively across dozens of languages. The benchmarks define standard test conditions that can be utilized to test cross-lingual transfer and model strength. However, they overwhelmingly favor popular languages such as Spanish, German, or French, at the expense of underrepresented languages like Yoruba or Inuktitut. Therefore, model performance over these benchmarks may not generalize to real-world deployment over dozens of distinct linguistic ecosystems.

To fill these gaps, data augmentation techniques such as back-translation, synthetic data generation, and crowdsourced annotation have become increasingly used. Back-translation generates auxiliary training pairs by translating them into a pivoting language (e.g., English) and then back to the target language. In contrast, synthetic generation creates pseudo-annotated examples through large language models. Crowdsourcing is a scaled-up annotation process that can vary in quality if its annotators are uninformed or inconsistent. Quality control procedures and dataset diversity measures will be crucial for developing robust cross-lingual TDC systems.

4.2 Evaluation strategies

Robust evaluation techniques are necessary to evaluate the effectiveness and generalizability of multi-lingual and multi-cultural coverage cross-lingual TDC models. Because of TDC's multi-lingual and multi-cultural coverage, it is impossible to assess these by restricting to standard metrics, as this would not accommodate finer-grained linguistic complexities and audience expectations. Accuracy, F1-score, Kendall's τ, and Spearman's ρ are typically used to measure performance, giving classification and ranking information. These metrics can conceal systematic biases, particularly when comparing languages with high resources to those with low resources. It is necessary to incorporate multiple complementary metrics to better understand the model's behavior across different datasets.

Another critical point is cross-language verification, in which models are verified for multiple target languages rather than a single source-target pair. It reveals how much models generalize across typologically different languages and varying scripts. If a system performs well over English–German and poorly over English–Finnish sets, it may be sensitive to morphological complexity. Cross-language verification is also a fairer benchmark for comparing methodologies, such as machine translation pipelines, multi-lingual transformers, or meta-learning methods.

Qualitative analysis offers further nuance by searching misclassified/borderline instances for cultural/linguistic mismatches. Idiomatic expressions, for example, or culturally based allusions may confuse models even with favorable overall metrics. Blending qualitative insights with best practices for reproducibility, involving sharing hyperparameters, pre-training details, and test scripts, opens up and facilitates progress in the field. Collectively, as shown in Table 9, combining quantitative and qualitative methods, cross-language

Table 9: Key components of evaluation strategies for cross-lingual TDC

Component	Description	Benefit	Applications
Metrics (Accuracy, F1, Kendall's τ, Spearman's ρ)	Quantitative measures of classification and ranking performance.	Enables standardized performance comparison	Evaluating TDC on multi-lingual benchmarks like XGLUE.
Cross-language validation	Testing across multiple target languages to assess generalization.	Identifies strengths and weaknesses across languages	Comparing English-trained models on German, Finnish, etc.
Qualitative analysis	Manual review of misclassifications and edge cases.	Reveals cultural or linguistic biases	Analyzing idiomatic or genre-specific errors.
Reproducibility protocols	Sharing hyperparameters, datasets, and scripts for transparent evaluation.	Ensures fairness and repeatability	Public leaderboards and shared evaluation repositories.

verification, and procedures for reproducibility allows for an all-around evaluation framework that can quantify both performance and fairness.

4.3 Challenges and limitations

Cross-lingual TDC is hindered by numerous long-standing issues that limit scalability, fairness, and accuracy. One of the most considerable challenges is morphological richness, particularly for agglutinative or heavily inflected languages such as Finnish, Turkish, or Hungarian. For languages of this type, grammaticality is indicated by prolific affixation, i.e., by a word that would be an entire phrase in English. English-oriented feature-based or transformer-based models tend to incorrectly analyze tokens of this type, leading to underestimation and overestimation of difficulty. Defeating it means developing language-specific preprocessing or subword tokenization techniques to maintain morphological variations better.

Another notable limitation is the presence of translation artifacts, particularly for machine translation-based methods in low-resource languages. Automated translation programs may simplify advanced structures or miss culture-bound and idiomatic expressions. Distortion of TDC predictions can be biased and provide an incorrect representation of the hardness of the textual origin. Translating into English, for example, a culturally idiomatic text may lose subtleties, making it even more complicated for native users. These artifacts require various solutions, such as multi-lingual pre-trained models or hybrid methods that don't significantly rely on translation.

Variability across domains and genres is also a problem. TDC models trained on formal writings, such as research articles or encyclopedia entries, may not generalize to informal domains, including blog posts, social media postings, or transcribed oral discourses, which have quite different sentence boundaries, word usage, and stylistic routines. As the variability over domains is significant, it is challenging to generalize over domains. Low-resource languages exacerbate all these problems. Most lack annotation collections, large-scale pre-training data, or language resources such as morphological analyzers. As shown in Table 10, all these intertwined aspects complicate model construction and highlight the need for diversified datasets, sound evaluation regimes, and methods attuned to linguistic variability.

4.4 Future directions

The future of cross-lingual TDC lies in developing methods that balance precision, efficacy, and fairness in different languages and conditions. Of interest is multimodal TDC, which combines visual, audio, or contextual cues with written language. Examples include children's picture books or language learning books, which typically incorporate images accompanied by written text. Adding these multimodal cues may complement predictions of difficulty. Additionally, incorporating prosodic or speaking characteristics into audio resources can enhance evaluation on multi-lingual learning platforms. Multimodal methods could be extended to benefit accessibility aids by varying their complexity according to the text and accompanying media.

Another research area is architecture incorporating language-agnostic representations and language-specific adapters/prompters. Such models can leverage higher proficiency by hybridizing universal multi-lingual models that learn idiosyncratic linguistic characteristics. Active and few-shot learning can be similarly employed to rapidly augment coverage for low-resource languages and reduce over-reliance on costly annotation campaigns. Models can be fine-tuned iteratively by leveraging human-in-the-loop mechanisms with minimal label effort. Such procedures are particularly valuable in teaching

Table 10: Key challenges and limitations in cross-lingual TDC

Challenge	Description	Impact on TDC	Potential mitigation strategies
Morphological complexity	Agglutinative or highly inflected languages complicate feature extraction.	Reduced accuracy and misclassification.	Use subword tokenization and morphology-aware models.
Translation artifacts	Biases or errors introduced by machine translation distort difficulty measures.	Over/underestimation of text complexity.	Favor multi-lingual embeddings or hybrid translation-free approaches.
Domain & genre variability	Formal vs. informal texts vary in structure and vocabulary.	Limited cross-domain generalization.	Train on diverse genres and use domain adaptation methods.
Low-resource languages	Lack of labeled data and pre-training corpora for many languages.	Hinders scalability and fairness.	Use few-shot learning, active learning, or crowdsourced labeling.

Table 11: Promising future directions for cross-lingual TDC

Future direction	Potential benefit	Implementation challenge	Applications
Multimodal TDC	Captures visual/audio context to improve accuracy	Requires multimodal datasets and annotation	Children’s book recommendation, accessible e-learning
Hybrid architectures	Balances universality and language-specific adaptation	Increased model complexity	Bilingual education platforms, adaptive tutors
Active/few-shot learning	Quickly expands low-resource datasets	Needs expert involvement for quality assurance	Rapid deployment for underrepresented languages
Fairness and bias mitigation	Ensures equitable performance across languages	Developing culturally sensitive metrics	Inclusive educational and accessibility tools
Standardized evaluation	Improves reproducibility and comparability	Reaching consensus across research communities	Benchmarks for multi-lingual NLP and TDC research

environments where expert intervention ensures the pedagogical accuracy of difficulty ratings.

Inclusiveness and fairness should also guide the development of TDC in the future. Optimized TDCs that are biased in favor of imbalanced data may perform poorly for underrepresented cultures/languages. Researchers, therefore, must construct bias-detection tools, culture-adaptive evaluation metrics, and balanced test sets that better reflect the world's linguistic diversity.

Standard evaluation frameworks, such as those presented in Table 11, must also be established to facilitate the comparability of studies and enable replicability. Such frameworks must be based on cross-domain testing, culture-adaptive metrics, and data provenance reporting. Following these routes, future research can develop transparent, scalable, and fair cross-lingual TDC systems, enhancing global knowledge sharing, accessibility, and educational support.

5 Conclusion

Cross-lingual TDC has emerged as a crucial research area at the intersection of natural language processing, machine learning, and accessibility. Cross-lingual methods have enabled scalable solutions that generalize from English-centric early TDC systems by transferring knowledge from higher-resource to lower-resource languages. Machine translation pipelines, multi-lingual pre-trained transformers, adversarial and meta-learning techniques, and knowledge distillation architectures advanced the field and opened practical routes to empower various linguistic populations.

Despite these advances, many challenges remain. The complexity of morphology, translation-based artifacts, variability of domains, and the availability of limited-quality datasets continue to limit the precision and accuracy of TDC systems. Evaluation measures must be advanced to include culturally sensitive metrics, cross-language usage verification, and qualitative analyses to ensure a resilient and justifiable operating capability. These issues must be addressed to deliver adaptive learning technologies and usable information resources across languages and territories. Future research should incorporate multimodal inputs, hybrid structures, active learning, and fairness-aware techniques. The field can construct resource-aware, inclusive systems to facilitate knowledge access by collaborating with computational linguists, pedagogues, and culture specialists. Cross-lingual TDC can thus transcend linguistic boundaries to empower learners, teachers, and information systems worldwide.

References

[1] M. Nalini, R. K. Dhanraj, B. Balusamy, V. Abirami, and K. Kavya, "Intelligent Assistants Using Natural Language Processing for Hyperautomation," *Hyperautomation for Next-Generation Industries,* pp. 91–126, 2024.

[2] C. Troussas, A. Krouska, and C. Sgouropoulou, "Case Studies of Interactive Machine Learning for Adaptive Learning Technology Systems," in *Human-Computer Interaction and Augmented Intelligence: The Paradigm of Interactive Machine Learning in Educational Software*: Springer, 2025, pp. 347–385.

[3] S. AlKhuzaey, F. Grasso, T. R. Payne, and V. Tamma, "Text-based question difficulty prediction: A systematic review of automatic approaches," *International Journal of Artificial Intelligence in Education,* vol. 34, no. 3, pp. 862–914, 2024.

[4] V. Dogra *et al.*, "A complete process of text classification system using state-of-the-art NLP models," *Computational Intelligence and Neuroscience,* vol. 2022, no. 1, p. 1883698, 2022.

[5] B. Pourghebleh and N. J. Navimipour, "Data aggregation mechanisms in the Internet of things: A systematic review of the literature and recommendations for future research," *Journal of Network and Computer Applications,* vol. 97, pp. 23–34, 2017, doi: https://doi.org/10.1016/j.jnca.2017.08.006.

[6] S. M. Marier, X. Chen, L. Zhu, and X. Kong, "Grammatical error correction for low-resource languages: a review of challenges, strategies, computational and future directions," *PeerJ Computer Science,* vol. 11, p. e3044, 2025.

[7] K. Nguyen-Viet, "Navigating personal name avoidance in artificial intelligence: challenges, adaptations, and ethical considerations," *Ethics and Information Technology,* vol. 27, no. 3, pp. 1–12, 2025.

[8] F. Pallucchini, L. Malandri, F. Mercorio, and M. Mezzanzanica, "Lost in Alignment: A Survey on Cross-Lingual Alignment Methods for Contextualized Representation," *ACM Computing Surveys,* 2025.

[9] E. Hashmi, S. Y. Yayilgan, and M. Abomhara, "Metalinguist: enhancing hate speech detection with cross-lingual meta-learning," *Complex & Intelligent Systems,* vol. 11, no. 4, p. 179, 2025.

[10] B. Wang, "A Hybrid Fuzzy Logic and Deep Learning Model for Corpus-Based German Language Learning with NLP," *Informatica,* vol. 49, no. 21, 2025.

[11] Y. Ren, W. Fan, and J. Wang, "Intelligent text analysis for effective evaluation of english Language teaching based on deep learning," *Scientific Reports,* vol. 15, no. 1, p. 28949, 2025.

[12] L. Li, "Comparative Performance of Neural Networks and Ensemble Methods for Command Classification in ALEXA Virtual Assistant," *Informatica,* vol. 49, no. 2, 2025.

[13] H. Rizwan Iqbal, M. Sharjeel, J. Shafi, U. Mehmood, S. U. Hassan, and A. A. Raza, "Urdu paraphrased text reuse and plagiarism detection using pre-trained large language models and deep hybrid neural networks," *Multimedia Tools and Applications,* pp. 1–23, 2025.

[14] R. Piperno, L. Bacco, F. Dell'Orletta, M. Merone, and L. Pecchia, "Cross-lingual distillation for domain knowledge transfer with sentence

transformers," *Knowledge-Based Systems,* vol. 311, p. 113079, 2025.

[15] P. Q. Huy, "Cross-Lingual Evidence-Based Strategies for Identifying Fabrications in Neural Translation Systems," *Transactions on Artificial Intelligence, Machine Learning, and Cognitive Systems,* vol. 9, no. 11, pp. 1–10, 2024.

[16] J. Rodríguez-Miret, E. Farré-Maduell, S. Lima-López, L. Vigil, V. Briva-Iglesias, and M. Krallinger, "Exploring the Potential of Neural Machine Translation for Cross-Language Clinical Natural Language Processing (NLP) Resource Generation through Annotation Projection," *Information,* vol. 15, no. 10, p. 585, 2024.

[17] M. B. Bagherabad, E. Rivandi, and M. J. Mehr, "Machine Learning for Analyzing Effects of Various Factors on Business Economic," *Authorea Preprints,* 2025, doi: https://doi.org/10.36227/techrxiv.174429010.09842200/v1.

Bio-Inspired Adaptive Rate Modulation for Multi-task Learning in Progressive Neural Networks

Chourouk Guettas and Foudil Cherif
LESIA Laboratory, University of Biskra, Algeria
E-mail: guettas-chourouk@univ-eloued.dz, cherif.foudil@univ-biskra.dz

Keywords: Multi-task learning, progressive neural networks, continual learning, adaptive learning rate, catastrophic forgetting

Received: June 24, 2025

Sequential multi-task learning faces the fundamental challenge of acquiring new tasks while retaining performance on previously learned tasks. Progressive Neural Networks address catastrophic forgetting through architectural isolation but rely on fixed learning strategies that limit adaptive efficiency. Current adaptive methods typically optimize single factors, missing the coordinated nature observed in biological adaptive systems. Inspired by neuromodulatory systems where four key neurotransmitters (dopamine, serotonin, norepinephrine, and acetylcholine) coordinate learning and adaptation, we propose a Bio-Inspired Adaptive Rate Modulation framework that coordinates four computational modules using organizational principles derived from neuromodulatory systems: reward-based adaptation, stability control, attention gating, and knowledge integration. These modules translate the organizational principles of biological neuromodulation into practical optimization mechanisms for Progressive Neural Networks using standard neural network components. Evaluation on four OpenAI Gymnasium environments demonstrates performance improvements ranging from 1.06% to 155.94% over baseline Progressive Neural Networks, with up to 92% variance reduction and performance retention averaging 98.0% . The framework achieves these gains with reasonable computational overhead. Comprehensive ablation studies confirm each module's contribution, validating the four-factor design. Results demonstrate that neuromodulation-inspired coordination of multiple adaptive factors significantly outperforms fixed learning strategies, providing a principled approach to adaptive optimization in sequential multi-task learning.

Povzetek: Članek predlaga bio-navdihnjen okvir za progresivne nevronske mreže, ki s koordinacijo štirih modulov posnema neuromodulacijo za prilagodljivo učenje brez pozabljanja v zaporednih nalogah.

1 Introduction

Sequential multi-task learning faces the fundamental challenge of acquiring new tasks while retaining performance on previously learned tasks. Progressive Neural Networks (PNNs) address catastrophic forgetting through architectural isolation—creating separate neural columns for each task with lateral connections for knowledge transfer [1]. However, current PNN implementations rely on fixed learning strategies, missing opportunities to adapt learning parameters based on context.

Current adaptive methods [2] optimize single factors independently, failing to capture the coordinated neuromodulation observed in biological systems, where multiple monoaminergic systems (dopamine, serotonin, norepinephrine) interact to dynamically regulate cognitive state, synaptic plasticity, and learning processes in the prefrontal cortex [3], with recent computational work confirming that such coordinated mechanisms provide substantial advantages in neural network learning [4, 5].

Biological neuromodulatory systems coordinate four key neurotransmitters: dopamine (reward processing), serotonin (stability regulation), norepinephrine (attention control), and acetylcholine (memory integration) [6, 7, 8]. This coordination suggests artificial systems could benefit from similar multi-factor approaches.

Inspired by these organizational principles, we propose a Bio-Inspired Adaptive Rate Modulation (BIARM) framework for Progressive Neural Networks. Our framework implements four specialized adaptation modules corresponding to these neurotransmitter functions, translating biological coordination principles into practical optimization mechanisms. The system dynamically adapts learning parameters through coordinated modulation and enables selective knowledge transfer through attention-gated lateral connections.

This work contributes: (1) a bio-inspired adaptive rate modulation framework coordinating four specialized modules for PNNs, (2) attention-gated lateral connections for selective knowledge transfer, and (3) experimental validation showing consistent performance improvements with substantial variance reduction across four OpenAI Gymnasium environments.

Section 2 reviews related work, Section 3 presents our

framework, Section 4 describes experiments and results, Section 5 discusses implications, and Section 6 concludes with future directions.

2 Related work

Our work intersects several research areas in machine learning: continual learning architectures, adaptive learning rate methods, and bio-inspired neural network design. We organize our review around these key areas and highlight how our approach addresses limitations in existing methods.

2.1 Continual learning and progressive neural networks

Continual learning addresses the fundamental challenge of learning new tasks while retaining knowledge from previous tasks [9]. Progressive Neural Networks represent a foundational advance by allocating separate neural columns for each task while enabling knowledge transfer through lateral connections [1]. This architectural isolation ensures previously learned representations remain unchanged while new tasks benefit from knowledge transfer, though the original formulation uses fixed lateral connection strengths and static learning parameters.

Several important extensions have emerged. Elastic Weight Consolidation selectively protects important weights using Fisher information-based regularization [10]. PackNet achieves incremental learning through progressive pruning [11], while Learn to Grow proposes dynamic expansion based on task complexity [12]. VariGrow extends this concept using Bayesian novelty detection for task-agnostic continual learning [13].

Memory-based approaches include Brain-Inspired Replay for selective experience replay [14], gradient interference-based selective replay [15], and synaptic intelligence using online Fisher information computation [16].

Despite these advances, most approaches maintain fixed learning strategies that do not adapt to varying task demands and operate through static post-learning mechanisms that protect or isolate parameters after task completion. This represents a gap for methods that could enhance the learning process itself through dynamic parameter adaptation during training.

2.2 Adaptive learning rate methods

Adaptive learning rate optimization has been extensively studied, with foundational methods focusing on single-factor adaptations. AdaGrad adapts rates based on historical gradients [17], while ADADELTA addresses aggressive decay using gradient windows [18]. Adam combines momentum with adaptive per-parameter rates [19], with subsequent improvements including AdamW for better regularization [20] and RAdam for variance correction [21]. Contemporary approaches extend these principles through fractional-order derivative methods for long-term gradient capture [22] and bio-inspired swarm optimization for rate selection [23].

Recent work explores multi-factor approaches. Adafactor adapts learning rates while maintaining sublinear memory cost [24], and coordinated adaptive mechanisms show substantial improvements in multi-task scenarios [2].

However, most existing methods optimize individual factors independently and lack principled frameworks for coordinating multiple adaptation mechanisms, particularly in sequential learning where simultaneous coordination is required.

2.3 Bio-inspired neural network design and computational neuromodulation

Bio-inspired approaches have gained attention for informing computational architectures [6]. Computational neuromodulation implementations show promise in spatial learning tasks and adaptive behaviors [4, 5]. Research reveals sophisticated coordination mechanisms where multiple neurotransmitters regulate learning and behavior [8, 7].

Practical applications include dopaminergic modulation in active inference models [25], bio-inspired networks for robotic control [26], and synaptic plasticity frameworks integrating learnable connection plasticity with dynamic expansion [29]. These demonstrate practical implementations while highlighting gaps between isolated mechanisms and coordinated systems.

Recent bio-inspired continual learning approaches include corticohippocampal circuit models for dual memory representation [27] and Drosophila-inspired active forgetting mechanisms [28]. These advances primarily target image-based continual learning scenarios rather than sequential control tasks.

2.4 Gaps in current approaches

Despite significant advances, fundamental gaps remain:

Limited Coordination: Adaptive optimizers improve single-factor adaptation but lack frameworks for coordinating multiple mechanisms simultaneously.

Static Architectures: Progressive Neural Networks prevent catastrophic forgetting but use fixed learning strategies that do not adapt to changing task demands.

Isolated Mechanisms: Existing computational neuromodulation approaches implement individual mechanisms rather than coordinated systems, overlooking coordination principles that make biological systems effective.

Our work addresses these gaps by coordinating multiple adaptive factors within Progressive Neural Networks using organizational principles from biological neuromodulatory systems, optimizing knowledge transfer and parameter updates in real-time rather than simply preserving learned representations, while targeting sequential control learning where existing recent advances have limited applicability.

3 Methodology

3.1 Framework overview

We propose a Bio-Inspired Adaptive Rate Modulation (BIARM) framework that extends Progressive Neural Networks with coordinated adaptive mechanisms. The framework implements four computational modules that adapt learning parameters during sequential task training: reward-based modulation, stability control, attention gating, and knowledge integration.

The core insight is that different aspects of sequential learning benefit from coordinated adaptation rather than fixed parameters. While Progressive Neural Networks prevent catastrophic forgetting through architectural isolation, they miss opportunities for adaptive optimization during the learning process.

3.2 Biological inspiration and computational framework

While our framework draws organizational inspiration from neuromodulatory systems, it implements these principles through standard computational mechanisms. The four modules correspond to functional roles observed in biological systems:

- **DA module:** Adapts learning based on performance feedback (analogous to dopaminergic reward processing)
- **5HT module:** Provides stability control (analogous to serotonergic behavioral regulation)
- **NE module:** Controls attention allocation (analogous to noradrenergic attention modulation)
- **ACh module:** Manages knowledge integration (analogous to cholinergic memory enhancement)

The key computational insight is that coordinated adaptation of multiple factors outperforms fixed parameter strategies in sequential learning scenarios. Each module operates at different temporal scales and processes different types of contextual information, enabling sophisticated adaptation during training. This computational framework enables dynamic optimization that fixed-parameter approaches cannot achieve, particularly in complex environments where knowledge integration from multiple frozen columns is critical.

3.3 Progressive neural network architecture

Our framework builds upon the standard Progressive Neural Network architecture. For a sequence of tasks $T = \{T_1, T_2, \ldots, T_n\}$, each task T_i is assigned a dedicated neural column C_i. The output of layer l in column i is computed as:

$$h_i^{(l)} = f\left(W_i^{(l)} h_i^{(l-1)} + \sum_{k=1}^{i-1} \alpha_{i,k}^{(l)} U_{i,k}^{(l)} h_k^{(l)}\right) \quad (1)$$

where $W_i^{(l)}$ represents the weight matrix for column i at layer l, $U_{i,k}^{(l)}$ are the lateral connection weights from column k to column i, $\alpha_{i,k}^{(l)}$ are attention coefficients that modulate knowledge transfer, and $f(\cdot)$ is the activation function.

Our key contribution is making the attention coefficients $\alpha_{i,k}^{(l)}$ and learning parameters adaptive through coordinated modulation.

3.4 Four-factor adaptive modulation system

The framework implements four adaptive modules, each serving a distinct computational role in sequential learning optimization:

3.4.1 Computational module design

Each module $m \in \{\text{DA}, \text{5HT}, \text{NE}, \text{ACh}\}$ maintains an internal state $s_m(t) \in [0, 1]$ that evolves during training. The state updates follow discrete exponential moving averages with module-specific time constants:

$$s_m(t+1) = s_m(t) + \tau_m \cdot (s_m^{\text{target}}(t) - s_m(t)) \quad (2)$$

where τ_m represents the discrete adaptation rate for module m, determining how quickly the module state responds to target changes within each training step, with τ_m values representing relative adaptation speeds rather than absolute biological timing (milliseconds to seconds), and $s_m^{\text{target}}(t)$ is computed by a module-specific neural network. The time constants are derived from computational requirements: DA uses rapid adaptation ($\tau_{DA} = 0.1$) for immediate performance feedback, NE operates fastest ($\tau_{NE} = 0.05$) for dynamic attention switching between frozen columns, ACh employs moderate adaptation ($\tau_{ACh} = 0.3$) for balanced knowledge integration, and 5HT uses the slowest rate ($\tau_{5HT} = 0.5$) for learning stabilization.

The time constants τ_m for each neuromodulator module were determined through systematic grid search optimization across the parameter space $[0.01, 1.0]$ with 0.05 increments, evaluating 10000 parameter combinations using a composite score that weighted convergence speed (30%), training stability (40%), and final performance (30%) across multiple environments. This optimization process identified the configuration $\tau_{DA} = 0.1$, $\tau_{5HT} = 0.5$, $\tau_{NE} = 0.05$, and $\tau_{ACh} = 0.3$ as achieving optimal performance, with validation across 10 independent runs using different random seeds confirming robustness. Sensitivity analysis demonstrates moderate robustness to parameter variations, with performance degrading by $15 - 60\%$ only when deviating significantly from the optimized configuration, indicating that while the parameters are well-tuned for the sequential learning domain, the framework

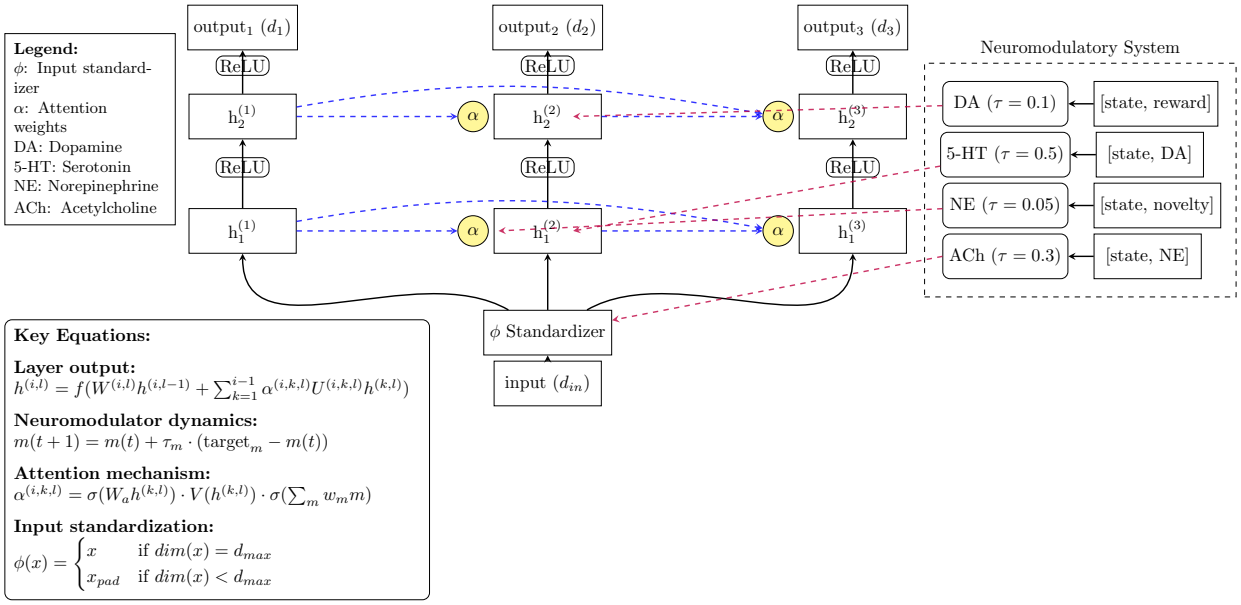


Figure 1: BIARM Architecture. The system consists of progressive task columns (shown vertically) with lateral connections modulated by attention mechanisms (α). Each column specializes in a specific task while maintaining knowledge transfer through modulated lateral connections. The neuromodulatory system (shown at bottom) influences learning through four key components. Solid arrows indicate forward connections, dashed arrows represent lateral connections, and dotted lines show neuromodulatory influences. Key equations shown define layer output, neuromodulator dynamics, attention mechanism, input standardization, and neuromodulator influences on learning rate, stability, attention, and memory.

maintains practical applicability across reasonable parameter ranges.

3.4.2 Target computation networks

Each module computes its target state using a dedicated neural network that processes relevant contextual information:

Reward-based Module (DA): Processes current state and reward signal

$$s_{\text{DA}}^{\text{target}} = \sigma(f_{\text{DA}}([s_{\text{current}}, r_{\text{current}}])) \tag{3}$$

Stability Control Module (5HT): Processes current state and DA module state

$$s_{\text{5HT}}^{\text{target}} = \sigma(f_{\text{5HT}}([s_{\text{current}}, s_{\text{DA}}])) \tag{4}$$

Attention Module (NE): Processes current state and task novelty

$$s_{\text{NE}}^{\text{target}} = \sigma(f_{\text{NE}}([s_{\text{current}}, n_{\text{task}}])) \tag{5}$$

Memory Integration Module (ACh): Processes current state and NE module state

$$s_{\text{ACh}}^{\text{target}} = \sigma(f_{\text{ACh}}([s_{\text{current}}, s_{\text{NE}}])) \tag{6}$$

where $f_m(\cdot)$ represents the neural network for module m, $\sigma(\cdot)$ is the sigmoid function, r_{current} is the current reward, and n_{task} measures task novelty.

The four modules coordinate through interdependent state updates and shared parameter influence. During each training step, modules update sequentially: dopamine computes its target from current reward (Equation 3), serotonin then incorporates the updated dopamine state for stability assessment (Equation 4), norepinephrine processes task novelty independently (Equation 5), and acetylcholine integrates the updated norepinephrine state for memory decisions (Equation 6). This creates dependency chains (Reward $\rightarrow$ DA $\rightarrow$ 5HT) and (Novelty $\rightarrow$ NE $\rightarrow$ ACh) that enable coordinated responses to environmental changes. All four states then jointly influence network parameters through the learning rate modulation (Equation 7) and lateral connection gating (Equation 8), ensuring that modulation decisions reflect the consensus of all adaptive factors rather than isolated module responses.

3.5 Adaptive parameter modulation

3.5.1 Learning rate modulation

The effective learning rate for each training step is computed as a weighted combination of all module states:

$$\eta_{\text{effective}} = \eta_{\text{base}} \cdot (w_{\text{DA}} \cdot s_{\text{DA}} + w_{\text{5HT}} \cdot s_{\text{5HT}} + w_{\text{NE}} \cdot s_{\text{NE}} + w_{\text{ACh}} \cdot s_{\text{ACh}}) \tag{7}$$

where η_{base} is the base learning rate and w_m are fixed weighting coefficients. In our implementation: $w_{\text{DA}} = 0.3$,

$w_{5HT} = 0.2$, $w_{NE} = 0.3$, $w_{ACh} = 0.2$.

This modulation mechanism enables dynamic adaptation of learning intensity based on the current neuromodulatory context, ensuring the learning rate adapts to the consensus of all four modules rather than being dominated by any single factor.

3.5.2 Attention-gated knowledge transfer

The attention coefficients for lateral connections from previous frozen columns are computed through a three-component mechanism that incorporates learned attention weights, value projections, and coordinated neuromodulator influence:

$$\alpha_{i,k}^{(l)} = \sigma(W_\alpha h_k^{(l)}) \cdot V(h_k^{(l)}) \cdot \sigma\left(\sum_{m\in\{DA,5HT,NE,ACh\}} w_m s_m\right) \quad (8)$$

where W_α is a learned attention weight matrix, $V(\cdot)$ represents a learned value projection that transforms the source features, and w_m are learnable modulation weights that determine the influence of each neuromodulator state s_m on the attention mechanism. This formulation enables selective knowledge transfer by combining three key components: (1) content-based attention through $\sigma(W_\alpha h_k^{(l)})$, (2) value transformation through $V(h_k^{(l)})$, and (3) context-dependent modulation through the coordinated influence of all four neuromodulator states. The sigmoid activation on the modulation term ensures bounded influence while allowing each module to contribute according to its learned importance weights.

3.6 Implementation details

3.6.1 Network architecture

Each modulation module uses a two-layer neural network with 64 hidden units and ReLU activation. The networks are initialized using Xavier initialization and trained jointly with the main Progressive Neural Network. The complete training procedure is shown in Algorithm 1

3.6.2 Input standardization

To handle varying input dimensions across tasks, we implement input standardization:

$$\phi(x) = \begin{cases} x & \text{if } \dim(x) = d_{\max} \\ [x; \mathbf{0}_{d_{\max}-\dim(x)}] & \text{if } \dim(x) < d_{\max} \end{cases} \quad (9)$$

where $d_{\max}$ is the maximum input dimension across all tasks.

Algorithm 1 Bio-Inspired Adaptive Rate Modulation Training

l

Require: Task sequence $T = \{T_1, T_2, \ldots, T_n\}$, episodes E, batch size B

Ensure: Trained progressive network with adaptive modulation

```
Initialize modulation system with states s_m = 0.5 for all m
Initialize base learning rate η_base and network parameters
for each task T_i in T do
    Initialize new column C_i
    Calculate initial task novelty n_task = 1.0
    for episode = 1 to E do
        Reset environment
        while episode not done do
            Observe state s_current, execute action, get reward r
            Update module target states (Equations 3-6)
            Update module states (Equation 2)
            Compute effective learning rate (Equation 7)
            Compute attention coefficients (Equation 8)
            Store transition in replay buffer
            if buffer size ≥ B then
                Sample mini-batch from buffer
                Compute loss with modulated parameters
                Update network parameters with η_effective
            end if
            Update task novelty: n_task ← n_task × 0.995
        end while
    end for
end for
```

3.6.3 Task novelty computation

Task novelty n_{task} starts at 1.0 for each new task and decays exponentially during training:

$$n_{\text{task}}(t+1) = n_{\text{task}}(t) \times 0.995 \quad (10)$$

Task novelty quantifies learning progress within each task, providing a simple measure of how ”new” the current task remains relative to initial conditions.

3.7 Computational complexity

The modulation system introduces computational overhead through:

- Four forward passes through small neural networks (64 hidden units each)
- Computation of modulated learning rate and attention coefficients
- State updates for four scalar values

The algorithmic complexity scales as $O(M \cdot H)$ where $M = 4$ is the number of modules and $H = 64$ is the hidden units per module. Memory complexity increases by $O(M \cdot H^2 + M)$ for the modulation networks and scalar states. The overhead scales linearly with the number of tasks due to lateral connection computations.

4 Experimental methodology

4.1 Experimental setup

We evaluate the Bio-Inspired Adaptive Rate Modulation (BIARM) framework on four distinct OpenAI Gymnasium environments that test different aspects of sequential learning and control. The environments were selected to provide varying levels of complexity and different types of learning challenges.

4.1.1 Environment descriptions

The four selected environments are visualized in Figure 2 and described as follows:

CartPole-v1: A classic control problem with a 4-dimensional continuous state space representing cart position, velocity, pole angle, and angular velocity. The agent has 2 discrete actions (left/right) and receives +1 reward per timestep while maintaining the pole upright. Episodes terminate when the pole exceeds ±12 degrees or the cart position exceeds ±2.4 units, with a maximum of 500 timesteps.

Acrobot-v1: A two-link pendulum control task with a 6-dimensional state space encoding joint angles and velocities. The agent has 3 discrete actions and must swing the end-effector to reach a target height. The environment provides -1 reward per timestep until the goal is reached or 500 steps elapse.

MountainCar-v0: A car navigation task with a 2-dimensional state space (position and velocity) and 3 discrete actions. The agent receives -1 reward per timestep and must build momentum to reach the goal, requiring exploration and momentum-based strategies. Episodes terminate upon reaching the goal or after 200 steps.

LunarLander-v3: A spacecraft control task with an 8-dimensional state space covering position, velocity, orientation, and ground contact information. The agent has 4 discrete actions and receives rewards ranging from -100 to +100 based on landing success, fuel efficiency, and stability. Episodes are limited to 1000 steps. Success defined as positive scores.

4.1.2 Training configuration

All experiments use identical network architectures for fair comparison. Each task column consists of two hidden layers with 128 units each and ReLU activation. Input and output dimensions are adapted to each environment's requirements.Training hyperparameters are standardized across all experiments as outlined in Table 1:

4.2 Evaluation metrics

We employ multiple evaluation metrics to assess different aspects of sequential learning performance:

4.2.1 Task performance metrics

These metrics evaluate both the absolute performance achieved and the consistency of learning across different environments and task complexities.

Final Performance: Mean episode reward over 10 test episodes after the end of training, providing a measure of converged performance.

Performance Stability: Standard deviation of episode rewards over the last 50 episodes, measuring training stability.

4.3 Baseline comparisons

We compare BIARM against two primary baselines to isolate the contribution of adaptive modulation:

Standard Progressive Neural Networks (PNN): The original PNN architecture with fixed learning parameters and static lateral connections, using identical network architecture and hyperparameters.

DQN Performance Reference: For context, we include DQN performance results from Zhang et al. (2023) [30] on the same environments, though direct comparison is limited due to different architectural approaches.

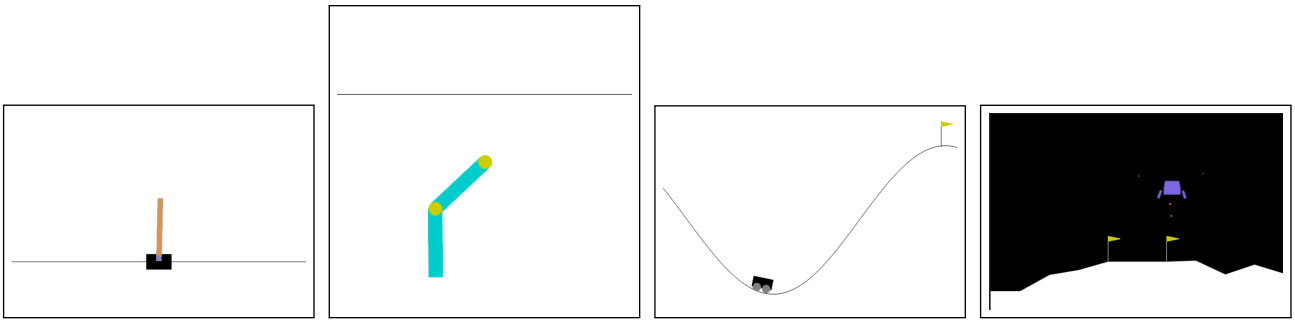
All experiments use identical random seeds (42) for reproducibility and are averaged over 10 independent runs to ensure statistical reliability.

5 Results and discussion

5.1 Overall performance analysis

Table 2 presents the comprehensive performance comparison between baseline Progressive Neural Networks and BIARM across all tested environments. BIARM demonstrates consistent performance improvements ranging from 1.06% in Acrobot-v1 to 155.94% in LunarLander-v3, with particularly dramatic gains in complex control environments.

The most significant improvement occurs in LunarLander-v3, where BIARM achieves positive scores (233.37±34.38) while baseline PNN consistently fails with negative performance (-417.17±412.62). This dramatic difference demonstrates the framework's effectiveness in complex environments where coordinated knowledge integration from multiple frozen columns becomes critical. The substantial variance reduction across all tasks (most notably CartPole-v1 achieving zero variance) indicates improved training stability through adaptive parameter modulation.



(a) CartPole Visualization (b) Acrobot Visualization (c) MountainCar Visualization (d) LunarLander Visualization

Figure 2: Visualization of OpenAI Gymnasium environments: (a) CartPole-v1; (b) Acrobot-v1; (c) MountainCar-v0; and (d) LunarLander-v3.

Table 1: Training configuration parameters

Parameter	Value
Batch size	64
Base learning rate	0.001
Optimizer	Adam
Discount factor (γ)	0.99
Experience replay buffer	10,000 transitions per task
Epsilon-greedy exploration	ϵ starts at 1.0, decays to 0.01 with rate 0.995
Task-specific Training Episodes	
CartPole-v1	400 episodes
Acrobot-v1	500 episodes
MountainCar-v0	800 episodes
LunarLander-v3	800 episodes

5.2 Computational overhead analysis

Figure 3 presents the computational overhead analysis comparing training time per step between baseline PNN and BIARM across all environments. For simple tasks like CartPole-v1, the 52% overhead yields only 1.83% improvement, representing a poor cost-benefit ratio. However, complex environments like LunarLander-v3 show positive returns on computational investment 87% overhead for 155.94% improvement, where baseline PNN consistently fails while BIARM achieves positive performance.

While the computational overhead represents a limitation, several optimization strategies could reduce costs in future implementations: selective module activation based on task complexity assessment, hardware acceleration of the small neural networks (64 units each), and adaptive modulation intensity scaling. For resource-constrained applications, a simplified two-module version focusing on the most critical DA and ACh components could provide a reasonable performance-cost trade-off.

5.3 Ablation study: modulation component analysis

To validate the contribution of each adaptive module, we conducted systematic ablation studies using five configurations: FULL (all modules active), DA- (without dopamine module), 5HT- (without serotonin module), NE- (without norepinephrine module), and ACh- (without acetylcholine module).

To isolate the contribution of each neuromodulator, we systematically remove individual modules while maintaining the computational framework. When ablating modules with dependencies (5HT→DA, ACh→NE), we replace missing dependency inputs with fixed neutral values (0.5), which represents moderate activation and allows fair comparison across configurations, allowing computation to proceed while eliminating the specific module's adaptive contribution.

Table 3 presents the ablation study results for the most challenging environment, LunarLander-v3, where modulation effects are most pronounced.

The ablation results reveal distinct contributions from each module. The DA (reward processing) and ACh (knowledge integration) modules show the most severe impact when removed, with performance drops of 108.0% and

Table 2: Performance comparison between DQN, baseline PNN, and BIARM

Environment	DQN [30]	Baseline PNN	BIARM	vs DQN (%)	vs PNN (%)
CartPole-v1	360.5±52.1	491.00±11.94	500.00±0.00	+38.7	+1.83
Acrobot-v1	-135.2±32.4	-94.50±24.96	-93.50±18.91	+30.8	+1.06
MountainCar-v0	-180.3±42.1	-165.90±46.31	-119.80±28.02	+33.55	+27.79
LunarLander-v3	160.5±58.4	-417.17±412.62	233.37±34.38	+45.4	+155.94

Table 3: Analysis of ablation study impact on performance

Environment	Full BIARM	Without DA	Without 5-HT	Without ACh	Without NE
CartPole-v1	500.00 ± 0.00	195.77±148.93	173.29±198.46	247.78±210.25	233.33±212.19
Acrobot-v1	-93.50 ± 18.91	-109.82±20.67	-110.78±21.01	-106.48±22.55	-110.27±24.56
MountainCar-v0	-119.80 ± 28.02	-125.90±32.98	-126.26±26.96	-133.39±29.71	-139.63±31.10
LunarLander-v3	233.37 ± 34.38	-18.63±190.72	218.15±97.80	-39.61±152.67	202.99±135.61

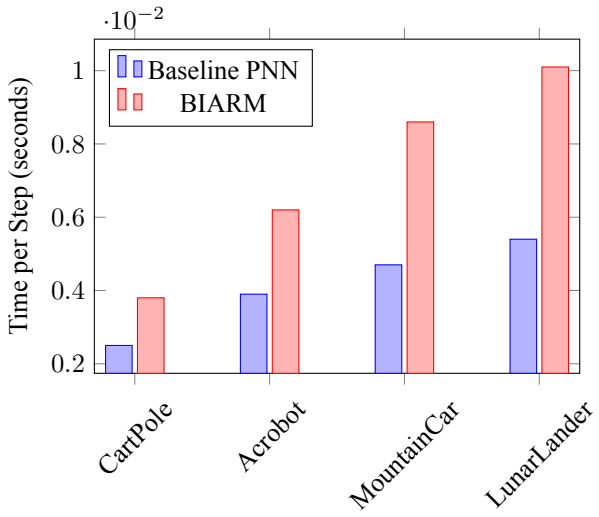


Figure 3: Computational overhead comparison showing time per step across environments

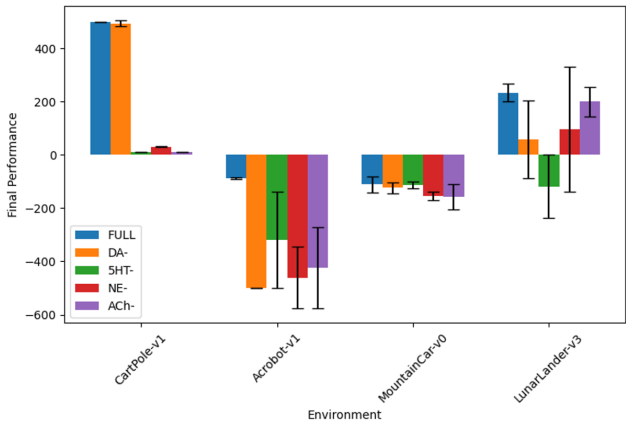


Figure 4: Final performance evaluation of catastrophic forgetting for ablation study

117.0% respectively. This indicates their critical roles in learning adaptation and selective knowledge transfer from frozen columns. The 5HT module (learning stabilization) shows the smallest impact (-6.5%), but contributes to variance reduction, as evidenced by comparing the standard deviations. The NE module (attention modulation) contributes moderately (-13.0%) but significantly affects training stability, with variance increasing from 34.38 to 135.61 when removed.

5.4 Module-specific contribution analysis

The ablation study, as shown in Figure 4, reveals that different modules serve complementary but distinct computational roles in the adaptive framework:

Dopamine (DA) Module: Most severe impact (-108.0%) when removed, confirming its critical role in reward-based learning adaptation. **Acetylcholine (ACh) Module:** Highest impact (-117.0%), demonstrating its essential role in knowledge integration between frozen columns..

Norepinephrine (NE) Module: Moderate impact (-13.0%) but 4x variance increase, indicating its role in attention-gated knowledge selection.

Serotonin (5HT) Module: Smallest direct impact (-6.5%) but contributes to learning stabilization and variance reduction.

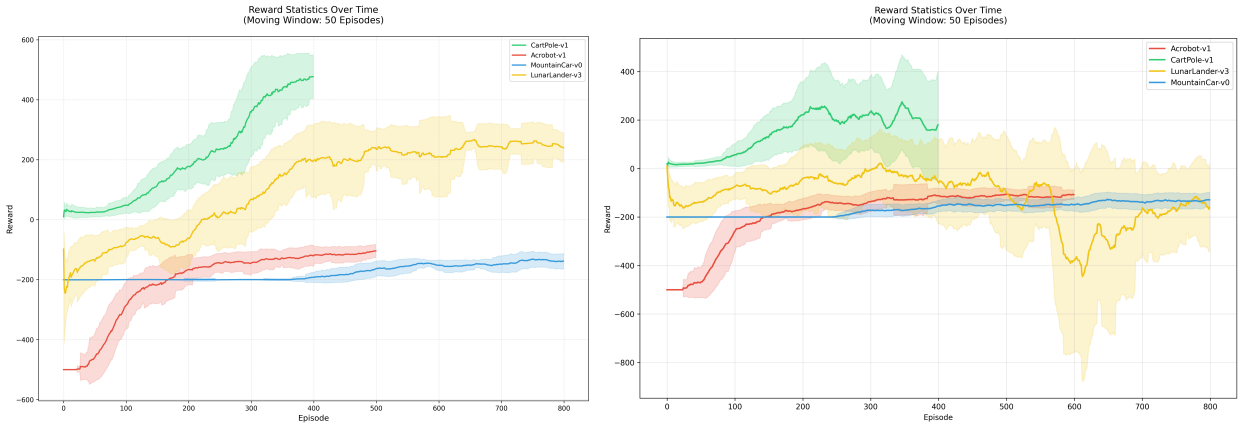
5.5 Training stability analysis

The framework demonstrates superior training stability compared to baseline Progressive Neural Networks. Figure 5a shows the learning curves for all environments, highlighting reduced variance and more consistent convergence patterns with BIARM.

The variance reduction is particularly pronounced in complex environments. LunarLander-v3 shows a 92% reduction in performance variance (from ±412.62 to ±34.38), while maintaining substantially higher mean performance. This improvement stems from coordinated adaptation of learning parameters, which prevents instability commonly

Table 4: Initial vs final performance comparison between BIARM and baseline model

Environment	Model	Initial Performance	Final Performance	Performance Retention
CartPole-v1	**BIARM**	500.00 ± 0.00	500.00 ± 0.00	**100%**
	Baseline	496.20 ± 5.31	491.00 ± 11.94	98.95%
Acrobot-v1	**BIARM**	-88.20 ± 3.84	-93.50 ± 18.91	**94.3%**
	Baseline	-86.80 ± 10.71	-94.50 ± 24.96	91.85%
MountainCar-v0	**BIARM**	-117.20 ± 29.84	-119.80 ± 28.02	**97.82%**
	Baseline	-150.60 ± 49.33	-165.90 ± 46.31	90.78%
LunarLander-v3	**BIARM**	233.37 ± 34.38	233.37 ± 34.38	**100%**
	Baseline	-380.88 ± 346.54	-417.17 ± 412.62	91.30%



(a) (b)

Figure 5: Learning Performance Across Environments: (a) BIARM learning curves for each environment showing episode rewards over time. (b) Baseline PNN learning curves for each environment showing episode rewards over time.

observed in fixed-parameter approaches when integrating knowledge from multiple frozen columns.

5.6 Performance retention

Table 4 presents the performance retention analysis comparing initial task performance to performance after learning all subsequent tasks.

BIARM demonstrates excellent resistance to catastrophic forgetting, with performance retention averaging 98.0% across all tasks. The architecture's frozen column design inherently prevents catastrophic forgetting, while the adaptive modulation mechanisms optimize knowledge integration without degrading previous task performance. Perfect retention in CartPole-v1 and LunarLander-v3, combined with minimal degradation in other tasks, validates the framework's effectiveness in maintaining acquired knowledge while learning new tasks.

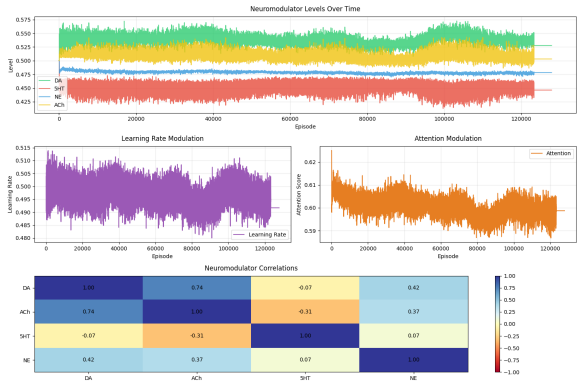
5.7 Neuromodulator stability analysis

The neuromodulator dynamics in Figure 6 reflect varying task complexity and stability requirements. In CartPole-v1, neuromodulator levels remain relatively stable with minimal fluctuations, corresponding to the task's inherent simplicity once optimal control is achieved. The DA module shows slight increases during reward acquisition, while 5HT maintains consistent stability control, reflecting the predictable balancing task.

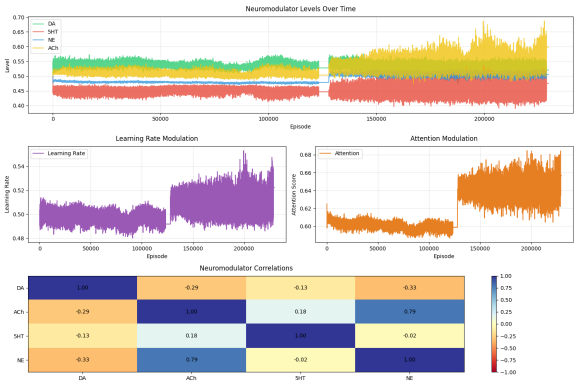
Complex environments like LunarLander-v3 and MountainCar-v0 demonstrate pronounced fluctuations that mirror their challenging nature. The landing task requires continuous adaptation to varying spacecraft dynamics, leading to dynamic DA responses and NE adjustments for attention allocation between frozen columns. MountainCar's momentum-based navigation requires persistent exploration, resulting in coordinated ACh-NE interactions for lateral knowledge transfer. These fluctuations represent appropriate adaptive responses rather than instability, demonstrating the framework's ability to scale modulation intensity based on task demands.

6 Conclusion

This paper introduces the Bio-Inspired Adaptive Rate Modulation (BIARM) framework, which addresses key limitations in sequential multi-task learning through coordinated adaptation of multiple learning factors. By incorporating four specialized modules inspired by biological neuromod-

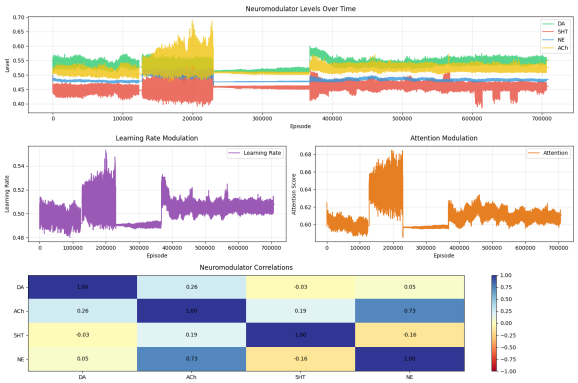


(a) CartPole Environment



(b) Acrobot Environment

(c) MountainCar Environment



(d) LunarLander Environment

Figure 6: Neuromodulator Dynamics Analysis. For each environment: (a) Temporal activation patterns of each neuromodulator (top), showing the evolution of DA, 5-HT, NE, and ACh levels over training; (b) Learning rate modulation (middle-left), demonstrating how the neuromodulators influence learning speed; (c) Attention modulation (middle-right), showing the dynamic focusing of learning; and (d) Correlation heatmap (bottom) revealing interactions between neuromodulators. The analysis demonstrates sophisticated coordination patterns, with particularly strong DA-NE coupling during critical learning phases.

ulatory systems, BIARM achieves substantial performance improvements (1.06% to 155.94%) over baseline Progressive Neural Networks while significantly reducing training variance and maintaining excellent resistance to catastrophic forgetting.

The experimental validation across four OpenAI Gymnasium environments demonstrates that coordinated adaptation consistently outperforms fixed-parameter approaches, with particularly dramatic gains in complex control tasks. The comprehensive ablation studies confirm that each adaptive module contributes distinctly to overall performance, validating the four-factor design and highlighting the importance of reward processing (DA) and knowledge integration (ACh) modules for effective sequential learning.

The framework achieves high performance retention rates (94-100%), demonstrating practical value for real-world applications requiring continual adaptation. Although the computational overhead (52-87% increase in training time per step) is substantial, the cost-benefit analysis shows that performance gains justify this investment, particularly in complex environments where coordinated knowledge integration is critical.

Future work should explore extension to larger-scale environments including visual tasks and robotics applications; investigation of non-linear neuromodulator interactions incorporating multiplicative effects and inhibition-excitation dynamics; development of more principled task novelty measures based on prediction error or entropy; exploration of learnable time constants that adapt during training; and computational optimization strategies including selective module activation and hardware acceleration to reduce computational overhead while preserving coordination benefits.

Acknowledgement

The authors gratefully acknowledge the support provided by LESIA (Laboratory of Expert Systems, Image and Signal Processing) at the University of Biskra, Algeria, for pro-

viding access to research infrastructure essential for conducting this study.

References

[1] Rusu, A. A., Rabinowitz, N. C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R., & Hadsell, R. (2016). Progressive neural networks. *arXiv preprint arXiv:1606.04671*. DOI: 10.48550/arXiv.1606.04671.

[2] Hervella, Á. S., Rouco, J., Novo, J., & Ortega, M. (2024). Multi-adaptive optimization for multi-task learning with deep neural networks. *Neural Networks*, 170, 254–265. DOI: 10.1016/j.neunet.2023.11.038

[3] Boyle, N., Betts, S., & Lu, H. (2024). Monoaminergic Modulation of Learning and Cognitive Function in the Prefrontal Cortex. *Brain Sciences*, 14(9), 902. DOI: 10.3390/brainsci14090902.

[4] Mei, J., Meshkinnejad, R., & Mohsenzadeh, Y. (2023). Effects of neuromodulation-inspired mechanisms on the performance of deep neural networks in a spatial learning task. *iScience*, 26(2), 106043. DOI: 10.1016/j.isci.2023.106043.

[5] Vecoven, N., Ernst, D., Wehenkel, A., & Drion, G. (2020). Introducing neuromodulation in deep neural networks to learn adaptive behaviours. *PLOS ONE*, 15(1), e0227922. DOI: 10.1371/journal.pone.0227922.

[6] Schmidgall, S., Ziaei, R., Achterberg, J., Kirsch, L., Hajiseyedrazi, S. P., & Eshraghian, J. (2024). Brain-inspired learning in artificial neural networks: A review. *APL Machine Learning*, 2(2), 021501. DOI: 10.1063/5.0186054.

[7] Cardozo Pinto, D. F., et al. (2024). Opponent control of reinforcement by striatal dopamine and serotonin. *Nature*. DOI: 10.1038/s41586-024-08412-x.

[8] Avery, M. C., & Krichmar, J. L. (2017). Neuromodulatory systems and their interactions: A review of models, theories, and experiments. *Frontiers in Neural Circuits*, 11, 108. DOI: 10.3389/fncir.2017.00108.

[9] van de Ven, G. M., Tuytelaars, T., & Tolias, A. S. (2024). Continual learning and catastrophic forgetting. *arXiv preprint arXiv:2403.05175*. DOI: 10.48550/arXiv.2403.05175.

[10] Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., ... Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13), 3521–3526. DOI: 10.1073/pnas.1611835114.

[11] Mallya, A., Davis, D., & Lazebnik, S. (2018). PackNet: Adding multiple tasks to a single network by iterative pruning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7765–7773. DOI: 10.1109/CVPR.2018.00805.

[12] Xu, Z., Zhou, Y., Wu, T., Ni, B., Yan, S., & Liu, X. (2020). Learn to grow: A continual structure learning framework for overcoming catastrophic forgetting. *ICLR* (2020). DOI: 10.48550/arXiv.1808.03314.

[13] Ardywibowo, R., Boluki, S., Gong, X., Wang, Z., & Qian, X. (2022). VariGrow: Variational architecture growing for task- agnostic continual learning based on Bayesian novelty. *ICML*. DOI: 10.48550/arXiv.2205.09325.

[14] Van de Ven, G. M., Siegelmann, H. T., & Tolias, A. S. (2020). Brain-inspired replay for continual learning with artificial neural networks. *Nature Communications*, 11(1), 4069. DOI: 10.1038/s41467-020-17866-2.

[15] Aljundi, R., Lin, M., Goujaud, B., & Bengio, Y. (2019). Gradient based sample selection for online continual learning. *Advances in Neural Information Processing Systems*, pp. 11816–11825. DOI: 10.48550/arXiv.1903.08671.

[16] Zenke, F., Poole, B., & Ganguli, S. (2017). Continual learning through synaptic intelligence. *Proceedings of the International Conference on Machine Learning*, PMLR, pp. 3987–3995. DOI: 10.48550/arXiv.1703.04200.

[17] Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12, 2121–2159. DOI: 10.5555/1953048.2021065.

[18] Zeiler, M. D. (2012). ADADELTA: An adaptive learning rate method. *arXiv preprint arXiv:1212.5701*. DOI: 10.48550/arXiv.1212.5701.

[19] Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. *ICLR* (2015). DOI: 10.48550/arXiv.1412.6980.

[20] Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. *ICLR* (2019). DOI: 10.48550/arXiv.1711.05101.

[21] Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., & Han, J. (2019). On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*. DOI: 10.48550/arXiv.1908.03265.

[22] Chen, S., Zhang, C., & Mu, H. (2024). An Adaptive Learning Rate Deep Learning Optimizer Using Long

and Short-Term Gradients Based on G–L Fractional-Order Derivative. *Neural Processing Letters*, 56, 106. DOI: 10.1007/s11063-024-11571-7.

[23] Wei, P., Shang, M., Li, J., Peng, Y., Qin, B., Li, R., & Liu, X. (2024). Efficient adaptive learning rate for convolutional neural network based on quadratic interpolation egret swarm optimization algorithm. *Heliyon*, 10(18), e37814. DOI: 10.1016/j.heliyon.2024.e37814.

[24] Shazeer, N., & Stern, M. (2018). Adafactor: Adaptive learning rates with sublinear memory cost. *arXiv preprint arXiv:1804.04235*. DOI: 10.48550/arXiv.1804.04235.

[25] FitzGerald, T. H. B., Dolan, R. J., & Friston, K. (2015). Dopamine, reward learning, and active inference. *Frontiers in Computational Neuroscience*, 9, 136. DOI: 10.3389/fncom.2015.00136.

[26] Guerrero-Criollo, R. J., Castaño-López, J. A., Hurtado-López, J., & Ramirez-Moreno, D. F. (2023). Bio-inspired neural networks for decision-making mechanisms and neuromodulation for motor control in a differential robot. *Frontiers in Neurorobotics*, 17, 1078074. DOI: 10.3389/fnbot.2023.1078074.

[27] Zhou, Y., Wang, H., Xu, S., Li, Z., Chen, Y., Wang, X., & Zhang, L. (2025). Hybrid neural networks for continual learning inspired by corticohippocampal circuits. *Nature Communications*, 16, 907. DOI: 10.1038/s41467-025-56405-9.

[28] Wang, L., Zhang, X., Su, H., & Zhu, J. (2023). Incorporating neuro-inspired adaptability for continual learning in artificial intelligence. *Nature Machine Intelligence*, 5(12), 1066–1083. DOI: 10.1038/s42256-023-00747-w.

[29] Ostapenko, O., Puscas, M., Klein, T., Jahnichen, P., & Nabi, M. (2 019). Learning to remember: A synaptic plasticity driven framework for continual learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11321–11329. DOI: 10.1109/CVPR.2019.01158.

[30] Zhang, Z., Zou, Y., Lai, J. and Xu, Q. (2023). M2DQN: A robust method for accelerating deep Q-learning network. *Proceedings of the International Conference on Machine Learning and Computing*, pp. 116–120.