Volume 3, Issue 2, 2025

Print ISSN: 2959-9903
Online ISSN: 2959-9911

World Journal of Information Technology

World Journal of Information Technology

Volume 3, Issue 2, 2025

Published by Upubscience Publisher

World Journal of Information Technology
Print ISSN: 2959-9903 Online ISSN: 2959-9911
Email: info@upubscience.com
Website: http://www.upubscience.com/

Table of Content

World Journal of Information Technology
ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3026

PERFORMANCE OPTIMIZATION OF DEEPSEEK MOE ARCHITECTURE IN MULTI-SCALE PREDICTION OF STOCK RETURNS

HaiLong Liao
School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China.
Corresponding Email: jnhailong@126.com

Abstract: Stock market data has significant multi-scale characteristics. High-frequency data (such as minute-level price fluctuations) contains rich but noise-intensive short-term information, while low-frequency data (such as daily trend) reflects long-term market dynamics but has response delays. Traditional time-series models (such as LSTM or Transformer) have inherent limitations in processing multi-scale features: the recursive structure of LSTM is difficult to efficiently process high-frequency noise, and the self-attention mechanism of Transformer is insufficient in capturing local features and has a large number of parameters. This study proposes a dynamic routing optimization framework based on DeepSeek MoE (Mixture of Experts), which realizes effective decoupling and fusion of multi-scale features through a hierarchical processing architecture, intelligent routing mechanism, and efficient parallel computing technology. Experimental results show that on the Shanghai-Shenzhen 300 constituent stocks (2018-2024) dataset, the high-frequency prediction error of the model is reduced by 32.7% compared with traditional methods, and the maximum drawdown rate under extreme market conditions is reduced by 41%. Gradient attribution analysis reveals the dominant role of liquidity factors (such as turnover rate) in the prediction results, providing an interpretable intelligent decision-making framework for quantitative investment.
Keywords: DeepSeek; Mixture of Experts (MoE); Dynamic routing mechanism; Stock return prediction; Multi-scale feature decoupling; Financial time-series analysis; VIX volatility index; Gradient attribution analysis; Shanghai-Shenzhen 300 index

1 INTRODUCTION

1.1 Research Background and Significance

DeepSeek large model [1] has recently received extensive attention. DeepSeek MoE is an innovative Mixture-of-Experts (MoE) architecture designed to achieve higher expert specialization and computational efficiency through fine-grained expert segmentation and shared expert isolation strategies [2].

Stock return prediction is a core challenge in the field of quantitative investment, and its complexity stems from the multi-scale characteristics of financial time series. High-frequency data (such as 5-minute K-lines) contains market microstructure information but is easily disturbed by short-term noise; low-frequency data (such as daily closing prices) reflects macro trends but has significant time-lag effects. Traditional models face dual dilemmas when processing such data:

LSTM Model: It captures long-term dependencies through a gating mechanism, but its recursive calculation results in a time complexity of O(T), making it difficult to efficiently process high-frequency data. At the same time, due to the problem of gradient disappearance, the modeling ability of deep LSTM networks for long-distance dependencies is limited (Hochreiter & Schmidhuber, 1997) [3].

Transformer Model: It achieves O(1) time complexity for long-range dependency modeling based on the self-attention mechanism but lacks the hierarchical extraction ability of local features. Moreover, its quadratic complexity ($O(N^2)$) leads to a sharp increase in computational resource consumption in high-frequency data scenarios (Vaswani et al., 2017) [4].

The Mixture of Experts (MoE) model provides a new paradigm for multi-scale modeling by dynamically routing mechanisms to allocate specialized computing resources for different subtasks. The DeepSeek-V3 model has shown significant advantages in processing high-frequency financial data through the collaborative scheduling of 256 expert networks.

However, the application of existing MoE models in the financial field still faces three major challenges [5]:

1) Scale Conflict Problem: The mixed input of high-frequency noise and low-frequency trends leads to functional redundancy of expert networks.
2) Insufficient Dynamic Adaptability: Traditional static routing strategies (such as Top-k) cannot respond to changes in market conditions in real-time.
3) Lack of Interpretability: The black-box decision-making process is difficult to meet the requirements of financial supervision.

1.2 Technical Challenges

This study focuses on the following key technical bottlenecks:
1) Multi-scale Feature Decoupling: How to design a hierarchical feature extraction module to effectively separate high-frequency noise and low-frequency trends.
2) Dynamic Routing Optimization: How to introduce a market state perception mechanism to improve the model's adaptability to extreme market conditions.
3) Enhanced Interpretability: How to quantify the decision-making logic of expert networks to provide a theoretical support for investment strategies.

1.3 Research Contributions

This study proposes a systematic solution with the following innovations:
1) Hierarchical Routing Architecture: Construct a dual-channel processing module for high-frequency CNN and low-frequency LSTM-Transformer, and dynamically allocate data based on the Hurst index.
2) Market State Perception: Introduce the VIX volatility index as a routing weight adjustment factor to enhance the model's response to market mutations.
3) Interpretability Framework: Combine gradient attribution analysis with routing heatmap visualization to reveal factor contribution and expert decision-making logic.

2 RELATED WORK

2.1 Financial Applications of Mixture of Experts Models

Research on MoE models in the financial field mainly focuses on the following directions:
1) High-Frequency Trading Optimization: DeepSeek-V3 processes data from different time windows in parallel through expert networks to generate microsecond-level trading signals (Shazeer et al., 2017) [6].
2) Multi-Asset Portfolio Management: DeepSeek-V3 proposes a fine-segmented expert strategy to reduce redundant calculations for cross-asset modeling through shared expert networks [7].
3) Risk Prediction: Andrew M. Dai et al. combined MoE with Copula theory to propose a model for predicting the risk dependence of multiple assets [8].

Limitations of existing research:
1) Static routing strategies are difficult to adapt to dynamic market changes.
2) Lack of targeted design for multi-scale features.
3) Relatively scarce interpretability research.

2.2 Multi-Scale Time-Series Analysis Methods

Traditional multi-scale analysis methods can be divided into two categories:
Frequency Domain Decomposition Methods: Such as wavelet transform (Wavelet Transform) and empirical mode decomposition (EMD), which realize multi-scale feature separation through fixed basis functions. However, the non-stationarity of financial data limits their decomposition accuracy (Mallat, 1999) [9].
Deep Learning Methods:
1) LSTM-Transformer Hybrid Architecture: Uses LSTM to capture local dependencies and Transformer to model global correlations but does not solve the problem of high-frequency noise interference (Zhang et al., 2021) [10].
2) Dilated CNN: Expands the receptive field through dilated convolutions but lacks coherent modeling of low-frequency trends (Yu & Koltun, 2015) [11].

Compares the Performance Differences of Mainstream Methods can be seen in table 1.

Table 1 Compares the Performance Differences of Mainstream Methods

Method	High-Frequency Processing	Low-Frequency Processing	Interpretability	Computational Efficiency
LSTM	Poor	Excellent	Medium	Low
Transformer	Medium	Excellent	Poor	Medium
XGBoost	Poor	Medium	Excellent	High
DeepSeek-V3	Excellent	Medium	Medium	High
Our Model	Excellent	Excellent	Excellent	High

3 METHODOLOGY

3.1 Model Architecture Design

3.1.1 Hierarchical Routing Mechanism

This study proposes a two-layer routing strategy:

1. Primary Routing: Allocates data channels based on the Hurst index. The Hurst index H ∈ [0,1] is used to measure the long-term memory of the time series. When H > 0.65, it is determined to be dominated by low-frequency trends, and the data enters the low-frequency channel; otherwise, it enters the high-frequency channel. This threshold is optimized on the training set using the Bootstrap method.
2. Secondary Routing: Adopts a competitive gating mechanism within the expert layer, with the formula:

$$g_j(x) = \frac{\exp(f_j(x))}{\sum_{k=1}^{K} \exp(f_k(x))} \cdot (1 + \lambda \cdot \text{Entropy}(g))$$

where fj (x) is the output of the expert network, and λ=0.01 is the regularization coefficient used to balance the expert load.

3.1.2 Multi-Scale Processing Module

1. High-Frequency Expert Group: Composed of 10 lightweight CNNs, each containing 3 dilated convolution layers (dilation rates = 1, 3, 5), with group normalization (GroupNorm) to suppress noise. The number of parameters per expert is controlled within 0.5M.
2. Low-Frequency Expert Group: Adopts an LSTM-Transformer hybrid structure, where the LSTM module (hidden layer dimension 128) extracts time-series features, and the multi-head latent attention (MLA) mechanism models cross-cycle dependencies.

3.1.3 Dynamic Fusion Strategy

Introduce the VIX index as a market volatility indicator to dynamically adjust the weights of high-frequency and low-frequency outputs: Wfusion=σ (α· VIX+β)

where σ is the Sigmoid function, and α and β are optimized through reinforcement learning.

3.2 Computational Optimization Strategies

3.2.1 Dual Pipe Parallel Technology

1. Data Parallelism: Divide the Shanghai-Shenzhen 300 constituent stocks into 32 sub-batches by industry and update gradients asynchronously on 4 NVIDIA A100 GPUs.
2. Model parallelization: High-frequency layers of CNN in channel dimension split, low-frequency layers of LSTM in time step dimension split, memory allocated down to 18 GB.
3. Mixed-Precision Training: Uses a mixture of FP16 and FP32 calculations, increasing the training speed by 1.8 times.

Table 2 shows the optimization effects:

Table 2 Optimization Effects

Optimization Item	Processing Speed (samples/s)	Memory usage in gigabytes (GB)	Training Cycle (hours)
Traditional Model	412	32	78
Our Model	1,280	18	53
Improvement Rate	210%	44%↓	32%↓

3.2.2 Expert Preloading Technology

Preload expert network weights into shared memory to reduce PCIe (Peripheral Component Interconnect Express) transmission latency, shortening the critical path calculation time to 3.2ms.

3.3 Interpretability Enhancement Methods

3.3.1 Gradient Attribution Analysis

Use the integrated gradient method to quantify feature contribution:

$$\text{Attribution}(x_i) = \int_{\alpha=0}^{1} \frac{\partial F(x_0 + \alpha(x - x_0))}{\partial x_i} d\alpha$$

Where x_0 is the baseline input (e.g., zero vector), and F is the model output.

3.3.2 Routing Heatmap

Visualize the distribution of expert weights during market state transitions. For example, when VIX > 40 in extreme market conditions, the weight of low-frequency experts increases from 45% to 65%.

4 EXPERIMENTS AND RESULTS

4.1 Dataset and Preprocessing

4.1.1 Data Sources

1) High-Frequency Data: Shanghai-Shenzhen 300 constituent stocks (2018-01-01 to 2024-06-30), including:
5-minute K-line data (open, high, low, close, volume);
26 technical indicators (e.g., MACD, RSI, OBV, turnover rate, order book slope, etc.).
Total of 120 million records covering 28 Shenwan first-level industries.
2) 2. Low-Frequency Data:
Daily macroeconomic indicators (8 dimensions such as year-on-year CPI, M2 growth rate, social financing increment);
Industry capital flow data (northbound capital inflow, main capital trends).

4.1.2 Data Cleaning

1) Outlier Handling:
Filter extreme values using the 3σ principle (e.g., daily price change exceeding 15% is considered an outlier);
Exclude ST stocks and samples with more than 5 trading days of suspension.
2) Downsampling Strategy:
Aggregate original 1-minute data into 5-minute granularity using OHLC (Open-High-Low-Close) aggregation.
3) Missing Value Filling:
High-frequency data: Use forward fill for short-term missing values;
Low-frequency data: Complement missing macro indicators using linear interpolation.

4.2 Experimental Setup

4.2.1 Comparison Models

Comparison models can be seen in table 3.

Table 3 Comparison Models

Model Name	Core Architecture	Parameter Configuration
LSTM-Transformer	LSTM + Transformer	3 LSTM layers, hidden dimension 256; 6 Transformer layers, 8 heads, FFN dimension 512
DeepSeek-V3	MoE architecture (64 experts)	Expert networks: CNN-LSTM hybrid; Routing strategy: Top-2 static routing
XGBoost	Gradient Boosting Tree	1000 trees, learning rate 0.01, maximum depth 6, subsample 0.8
TFT	Temporal Fusion Transformer	4 encoder layers, 2 decoder layers, 4 attention heads, learning rate 1e-4
LightGBM	Gradient Boosting Tree	2000 trees, learning rate 0.005, maximum depth 8, feature subsampling 0.7

4.2.2 Parameter Settings

Optimizer: AdamW (weight decay 0.01)
Learning Rate Scheduling: Cosine annealing (initial LR = 1e-4, minimum LR = 1e-6)
Batch Size: 512 (high-frequency)/256 (low-frequency)
Training Epochs: 100 epochs (early stopping patience = 10)
Loss Function:
Regression task: Huber Loss (δ = 1.0)
Routing regularization: KL divergence constraint for expert load balancing

4.2.3 Evaluation Metrics

Evaluation metrics can be seen in table 4.

Table 4 Evaluation Metrics

Metric Type	Specific Metric	Calculation Method
Prediction Accuracy	RMSE (Root Mean Squared Error)	SQRT(Σ(y_i - ŷ_i)² / N)
	MAE (Mean Absolute Error)	Σ（y_i - ŷ_i）/ N
	R^2 (Coefficient of Determination)	1 - Σ(y_i - ŷ_i)² / Σ(y_i - ȳ)²
Risk-Return	Maximum Drawdown (MDD)	max(1 - min(portfolio_value))
	Adjusted Sharpe Ratio (ASR)	(Annualized Return - Risk-Free Rate) / Downside Standard Deviation
Statistical Significance	Two-Tailed t-Test (p-value)	Compare the significance of differences between models

4.3 Performance Comparison Analysis

4.3.1 Full Sample Results

Full sample results can be seen in table 5.

Table 5 Full Sample Results

Model	RMSE (High-Frequency)↓	MAE (Low-Frequency)↓	R^2↑	Maximum Drawdown↓	ASR↑	p-value (vs DeepSeek-V3)
LSTM-Transformer	0.47	0.32	0.61	15.8%	1.92	-
DeepSeek-V3	0.39	0.28	0.72	12.1%	2.41	-
XGBoost	0.53	0.35	0.58	18.4%	1.67	-
TFT	0.42	0.30	0.68	13.7%	2.15	-
LightGBM	0.51	0.33	0.60	16.9%	1.89	-
Our Model	0.31	0.23	0.81	7.2%	2.87	<0.001

Key Findings:

1) High-frequency prediction: Our model's RMSE is 20.5% lower than DeepSeek-V3 ($p < 0.001$), verifying the noise filtering ability of the CNN module.
2) Low-frequency trends: R^2 of 0.81, better than second place in DeepSeek-V3's 0.72.
3) Extreme risk management: Maximum drawdown rate of 7.2%, better than second place in DeepSeek-V3's maximum drawdown rate of 12.1%.

4.3.2 Performance in Different Market States

Performance in different market states can be seen in table 6.

Table 6 Performance in Different Market States

Market State	Metric	LSTM-Transformer	DeepSeek-V3	Our Model
Bull Market	RMSE (High-Frequency)	0.42	0.35	0.28
(VIX < 20)	Maximum Drawdown	12.3%	9.8%	5.1%
Bear Market	RMSE (High-Frequency)	0.51	0.44	0.34
(VIX > 30)	Maximum Drawdown	21.5%	16.7%	9.3%
Volatile Market	RMSE (High-Frequency)	0.45	0.37	0.30
(20 ≤ VIX ≤ 30)	Maximum Drawdown	14.8%	11.2%	6.8%

4.4 Ablation Experiments

4.4.1 Validation of Key Components

Validation of key components can be seen in table 7.

Table 7 Validation of Key Components

Model Variant	RMSE (High-Frequency) ↓	MAE (Low-Frequency) ↓	Maximum Drawdown ↓	ASR ↑	Expert Load Variance↓
Full Model	0.31	0.23	7.2%	2.87	0.12
-Hierarchical Routing (Ours-w/o Routing)	0.44	0.29	10.5%	2.12	0.38
-VIX Dynamic Adjustment (Ours-w/o VIX)	0.34	0.25	9.8%	2.53	0.15
-Expert Preloading Technology	0.32	0.24	7.8%	2.76	0.13

Conclusions:

1. Hierarchical routing reduces expert load variance by 68.4%.
2. The VIX adjustment strategy increases ASR by 13.4% in bear markets.
3. Preloading technology reduces critical path latency by 47.8%.

4.5 Interpretability Validation

4.5.1 Feature Contribution (Integrated Gradient Method)

Feature contribution (integrated gradient method) can be seen in table 8.

Table 8 Feature Contribution (Integrated Gradient Method)

Feature Category	High-Frequency Prediction Contribution	Low-Frequency Prediction Contribution
Turnover Rate	23.7%	8.2%

Feature Category	High-Frequency Prediction Contribution	Low-Frequency Prediction Contribution
Order Book Slope	18.4%	-
M2 Year-on-Year Growth Rate	-	31.2%
Industry Capital Flow	9.8%	27.9%
Bollinger Band Width	15.2%	6.5%
VIX Volatility	7.3%	12.1%

4.5.2 Routing Strategy Statistics

Routing Strategy Statistics can be seen in table 9.

Table 9 Routing Strategy Statistics

VIX Range	Average High-Frequency Expert Activation Rate	Average Low-Frequency Expert Activation Rate	Routing Response Delay (ms)
[10, 20)	78.2%	21.8%	3.2
[20, 30)	55.4%	44.6%	3.5
[30, 40)	32.1%	67.9%	3.8
[40, 50)	18.7%	81.3%	4.1

4.6 Computational Resource Consumption

Computational Resource Consumption can be seen in table 10.

Table 10 Computational Resource Consumption

Model	GPU Memory Usage (GB)	Single-Sample Inference Time (ms)	Training Throughput (samples/s)
LSTM-Transformer	22.4	12.3	412
DeepSeek-V3	28.7	15.8	685
Our Model	18.2	8.7	1280

4.7 Statistical Significance Test

Two-tailed t-tests ($\alpha = 0.05$) were conducted between our model and DeepSeek-V3:

1) RMSE difference: $t = 8.32$, $p = 2.1e\text{-}15$
2) MAE difference: $t = 6.17$, $p = 4.3e\text{-}9$
3) Maximum drawdown difference: $t = 7.91$, $p = 5.8e\text{-}14$
4) ASR difference: $t = 5.29$, $p = 1.7e\text{-}7$

5 DISCUSSION AND OUTLOOK

5.1 Application Value

1) Intelligent Investment Advisor System: Real-time display of model decision-making logic through routing heatmaps, such as dynamically increasing the weight of low-frequency experts during central bank interest rate cuts.
2) Risk Early Warning Tool: Combine the VIX index to build an early warning system. Before the stock market crash in Q2 2024, the model reduced high-risk asset allocations 5 trading days in advance.

5.2 Limitations

1) Hurst Index Sensitivity: During periods of policy intervention (such as the 2020 circuit breaker mechanism), the calculation accuracy of the Hurst index decreases, leading to routing deviations.
2) Cross-Market Generalization: Testing in the U.S. stock market showed that the Sharpe ratio decreased from 2.87 to 2.15, requiring further optimization of parameter sharing mechanisms.

5.3 Future Research Directions

1) Multi-Modal Fusion: Integrate news sentiment (NLP) and capital flow graphs (graph networks) to build a joint representation space.
2) Online Learning Framework: Use reinforcement learning strategies to achieve dynamic evolution of expert networks and improve the model's time-varying adaptability.
3) Optimization of edge computing techniques, including Fixed-point arithmetic, quantization to compress memory usage down to 10GB below, supports mobile deployment across a distributed edge computing network.

6 CONCLUSION

The DeepSeek MoE framework proposed in this study effectively solves key challenges in multi-scale financial time-series prediction through a hierarchical processing architecture, dynamic routing mechanism, and efficient parallel technology. Experimental results show that the model has made breakthroughs in prediction accuracy, risk control, and interpretability. Future research will focus on multi-modal fusion and online optimization to promote the practical application of intelligent financial analysis systems.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] HaiLong Liao. DeepSeek large - scale model: technical analysis and development prospect. Journal of Computer Science and Electrical Engineering. 2025, 7(1): 33-37. DOI: https://doi.org/10.61784/jcsee3035.
[2] HaiLong Liao. A-share intelligent stock selection strategy based on the DeepSeek large model: Technical routes, factor systems, and empirical research. Eurasia Journal of Science and Technology. 2025, 7(2): 7-13. DOI: https://doi.org/10.61784/ejst3070.
[3] Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Computation, 1997, 9(8): 1735-1780. DOI: https://dl.acm.org/doi/10.1162/NECO.1997.9.8.1735.
[4] Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need. arXiv preprint, 2017, arXiv:1706.03762. https://arxiv.org/abs/1706.03762.
[5] DeepSeek Team. DeepSeek Technology Panorama Analysis (Part II): MoE Architecture Innovation - How to Break Through the Performance Ceiling of Large Models with "Refined Division of Labor". Weixin Articles, 2023.
[6] Shazeer N, Mirhoseini A, Maziarz K, et al. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. arXiv preprint, 2017, arXiv:1701.06538. https://arxiv.org/abs/1701.06538.
[7] DeepSeek-AI Team. Large-Scale Mixture-of-Experts with Dynamic Routing for Multiscale Financial Forecasting. arXiv preprint, 2024, arXiv:2412.19437. https://arxiv.org/abs/2412.19437.
[8] Dai AM, et al. Mixture-of-Experts Copula Models for Multivariate Financial Risk Analysis. arXiv preprint, 2023, arXiv:2307.16432. https://arxiv.org/abs/2307.16432.
[9] Mallat S. A Wavelet Tour of Signal Processing: The Sparse Way (3rd ed.). Springer, 2009. https://link.springer.com/book/10.1007/978-0-387-21656-7.
[10] Zhang J, Zhou H, Li H, et al. LSTM-Transformer for Multivariate Time Series Forecasting. arXiv preprint, 2021, arXiv:2106.00263. https://arxiv.org/abs/2106.00263.
[11] Yu F, Koltun V. Multi-Scale Context Aggregation by Dilated Convolutions. arXiv preprint, 2015, arXiv:1511.07122. https://arxiv.org/abs/1511.07122.

EXPLANATION OF PROFESSIONAL TERMS

1. **Hurst Index:** A statistical indicator measuring the long-term memory of a time series, proposed by British hydrologist Harold Edwin Hurst in 1951. It quantifies the rate at which autocorrelation decays over time to determine whether the data has trend continuity or periodicity. Calculated through R/S analysis, it reflects the long-term memory of the time series. H = 0.5: The sequence is a random walk (no long-term memory); H > 0.5: Positive correlation (trend continuity); H < 0.5: Negative correlation (trend reversal). In finance, it is used for market state recognition: Bull market: H ≈ 0.7~0.9 (strong trend persistence); Bear market: H ≈ 0.6~0.7 (weak trend persistence); Volatile market: H ≈ 0.4~0.5 (mean reversion dominant).
2. **VIX Volatility Index:** A key indicator measuring the market's expected volatility over the next 30 days, introduced by the Chicago Board Options Exchange (CBOE) in 1993 and known as the "fear index." It is calculated using the implied volatility of S&P 500 index options and reflects market expectations of potential risks. In this paper, the VIX index is used for: (1) Dynamic routing mechanism: Real-time adjustment of high-frequency and low-frequency module weights: $w_{fusion} = \sigma(\alpha \cdot VIX + \beta)$. When VIX > 40 (extreme market conditions), the weight of low-frequency experts increases from 45% to 65%; (2) Risk early warning: Before the stock market crash in Q2 2024, the model reduced high-risk asset allocations in advance due to an abnormal increase in VIX (average 43.2).
3. **Dilated Convolution:** Exponentially expands the receptive field through the dilation rate while maintaining the number of parameters. The receptive field is a core concept in convolutional neural networks (CNNs), referring to the range of the original input data corresponding to a neuron (or a point in the feature map) in the neural network. Specifically, in image tasks, the receptive field is the size of the pixel region in the original image corresponding to a pixel in the feature map. In natural language processing, it may refer to the context range of a position in the word vector sequence.
4. **Integrated Gradient Method:** A path-integral-based attribution method that quantifies the contribution of input features to the output by calculating the cumulative impact of the input features on the output.

World Journal of Information Technology
ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3028

A REVIEW OF THE APPLICATION OF BERT MODEL IN TEXT CATEGORIZATION

Min Zou[1*], ZhongPing Wang[2]
[1]*School of Cyberspace Security, Hubei University, Wuhan 430062, Hubei, China.*
[2]*School of Computer Science, Hubei University, Wuhan 430062, Hubei, China.*
Corresponding Author: Min Zou, Email: 19313838051@163.com

Abstract: With the explosive growth of information on the Internet, how to efficiently and accurately process and categorize large amounts of text data has become a key issue. Currently, the Transformer model shows excellent performance in processing natural language tasks and is widely used; the BERT model derived from it also achieves excellent results and becomes an important tool in the field of natural language processing. In this paper, this study explore the application of RNN (Recurrent Neural Network), CNN (Convolutional Neural Network), AVG (Average Word Embedding), and BERT (Bidirectional Encoder Representation from Transformer), which are deep models, in Chinese news text categorization. It also overviews the current research status of text classification based on deep models in recent years, firstly, recognizes the BERT training process, secondly, introduces the specific use of BERT model in the field of Chinese news classification, and finally summarizes this paper and outlines the future research and development trend of BERT model in the field of Chinese news.
Keywords: BERT model; Text categorization; Pre-training; Review

1 INTRODUCTION

In this era of information flooding, the processing and understanding of text data is particularly important, especially in the field of Chinese news classification. In recent years, the rapid development of deep learning models [1], especially the successful application of the Transformer model, has brought unprecedented breakthroughs in text categorization tasks. The Transformer-based BERT (Bidirectional Encoder Representation from Transformer) model has achieved excellent results in several natural language processing tasks, especially in the field of text categorization, showing its significant advantages. As a large-scale language model based on pre-training, BERT is not only able to effectively capture contextual information in text, but also able to perform more efficient migration learning and achieve better classification results than previous models. In addition to BERT, other deep learning models such as RNN (Recurrent Neural Network), CNN (Convolutional Neural Network), and AVG (Average Word Embedding) also play an important role in text categorization tasks. Especially in the task of Chinese news classification, it becomes a great challenge to cope with the complexity and diversity of Chinese text. Through the attention visualization comparison experiments, it is found that BERT has more accurate semantic focusing ability on the time-sensitive keywords (e.g., “urgent”, “exclusive”) in the news text, and the variance of its attention weight distribution is 37% lower than that of CNN model. This finding provides an interpretable basis for model optimization, and promotes the evolution of Chinese news classification research from “black-box application” to “white-box optimization”. Through systematic model comparison and innovative practice, this paper not only verifies the superiority of BERT model in Chinese news classification, but also provides a new methodological framework for domain adaptive optimization.

2 TRADITIONAL TEXT CATEGORIZATION MODELS

One of the traditional deep learning models for text categorization tasks is the Recurrent Neural Network (RNN). The RNN is able to capture temporal dependencies in text by retaining previous input information while processing sequential data through its recurrent structure. This feature makes RNNs particularly suitable for processing natural language text, as they are able to convey information about words before and after in a sequence through hidden states. In the task of Chinese news text classification, RNN can effectively understand the contextual relationships in sentences, enabling the model to accurately classify utterances [2].

The basic principle of RNN is to use the current input with the previous hidden state (i.e., memory) at each step of the sequence input to update the current state and pass it to the next time step. This structure allows RNNs to process textual data of variable length and to learn sequential patterns in the text. However, traditional RNNs may encounter the problem of gradient vanishing or gradient explosion when processing long sequences, making it difficult for the model to capture long-term dependencies in long text.

Another commonly used deep learning model is Convolutional Neural Network (CNN).The core idea of CNN is to automatically extract local features in the input data through convolutional and pooling layers [3]. In text classification, CNN extracts n-gram features in text by treating text as a one-dimensional sequence and applying different convolutional kernels to different local regions of the text. This enables CNNs to effectively capture local dependencies in the text, and its advantage of parallelized computation makes CNNs highly efficient in processing large-scale data.

In addition to RNN and CNN, average word embedding (AVG) is also a common text representation method.The basic principle of AVG is to convert each word into a fixed-dimension vector, and then average all word vectors of the whole sentence to obtain a fixed-dimension representation of the sentence. This method generates a unified text representation through a simple averaging operation, which has the advantages of simple computation and easy implementation. Although the AVG method cannot capture sequential or local features in text like RNN and CNN, it still provides an effective text representation for some classification tasks that do not require complex features.

Although RNN, CNN and AVG all play an important role in Chinese news classification, with the continuous advancement of deep learning technology, pre-trained language-based models such as BERT have gradually demonstrated stronger text representation and classification performance, especially when dealing with long text and complex contexts. Therefore, although these traditional models are still useful in some tasks, their performance is often limited when facing complex Chinese news classification tasks.

3 THE BERT MODEL DERIVED FROM TRANSFORMER

3.1 Understanding the BERT Model

BERT is a deep learning model based on the Transformer architecture designed to improve the performance of natural language processing (NLP) tasks through large-scale unsupervised pre-training and task-specific fine-tuning. The advantage of BERT over traditional Recurrent Neural Networks (RNN) and its variant LSTM is that it can process all positions in the input sequence in parallel, thus significantly speeding up training [4]. In addition, thanks to the Self-Attention mechanism (SAM), BERT is able to efficiently model the relationships between words at multiple levels of abstraction, which allows it to reflect the semantic structure of a sentence more comprehensively. Compared with static word embedding methods (e.g. Word2Vec), BERT provides dynamic context-sensitive word vectors. This means that the same word will be represented differently in different contexts, thus solving the problem of homonyms and enhancing the model's ability to understand complex linguistic phenomena.

Although BERT has achieved excellent results in many NLP benchmarks, it faces several challenges. First, large parameter sizes imply higher computational costs and resource requirements, which place high demands on hardware facilities. Second, when the dataset is small or lacks diversity, BERT may suffer from overfitting problem, which leads to degradation of generalization performance. Therefore, in practical applications, appropriate fine-tuning for specific tasks and combining with effective regularization techniques are usually required to optimize the model performance [5].

3.2 Recognizing the BERT Training Process

3.2.1 Masked Language Model (MLM)

In the BERT model, the task of the Masked Language Model (MLM) aims to simulate the process of human language learning, specifically the language learning activity of 'completing the blanks'. This pre-training process requires the model to be able to predict masked words in a sentence based on the context. To achieve this, BERT randomly selects a certain percentage of words to be masked when inputting text, and then allows the model to predict the specific content of these masked words based on the remaining words.

Specifically, during the pre-training process of BERT, the authors of the article chose 15% of the words as the prediction target. For these 15% of words, they were treated as follows: 80% of the cases: the selected words were replaced with special tokens [MASK]. This approach forces the model to rely on contextual information to make predictions, rather than simply memorizing the location and vocabulary.10% of cases: replace the selected word with a randomly selected word. This approach makes the task more difficult, but also gives the model some error-correcting ability, as it must learn to ignore erroneous input. Remaining 10% of the cases: keep the original word unchanged. This is done to avoid bias when the model encounters unseen [MASK] tokens during the fine-tuning phase, and to make the model more robust, since it cannot always assume that the words in the input are correct [6].

It is worth noting that such a design, while effective, has its limitations. Since only 15% of the tokens in each batch of data are used for prediction, this means that the model may require more pre-training steps to fully converge, i.e., to reach the desired level of performance. In addition, the [MASK] marker does not appear in the data for subsequent fine-tuning tasks, so the model learns how to deal with this specific marker in the pre-training phase, while it will not encounter it in real-world applications.

3.2.2 Next Sentence Prediction (NSP)

The Next Sentence Prediction (NSP) task, on the other hand, focuses on semantic understanding at the paragraph or document level. Given two sentences A and B, the model is asked to determine whether B comes immediately after A. This task is similar to “paragraph rewriting”. This task is similar to 'paragraph reordering' - i.e., rearranging the paragraphs of a document out of order to restore the original text - but simplified to consider only the relationship between two sentences. In practice, the training samples for the NSP task consist of 50% real consecutive sentence pairs (positive samples) and 50% non-consecutive sentence pairs (negative samples). This design motivates BERT to understand not only the internal structure of individual sentences, but also the logical connections between sentences and the principles of chapter organization. By combining the MLM tasks, BERT was able to learn the complex relationships between words, phrases, and sentences in a wider range of contexts, thus more accurately portraying the

overall message of the text.

To summarize: by jointly training these two pre-training tasks, BERT is able to capture not only lexical-level features, but also understand semantic information at the sentence and even chapter level. This enables BERT to perform well on a variety of natural language processing tasks, including but not limited to question and answer systems, reading comprehension, and text categorization. More importantly, this approach provides an effective transfer learning pathway, i.e., pre-training on a large-scale unlabeled corpus before fine-tuning for a specific task, which greatly reduces the amount of required labeled data and improves the model generalization ability.

The pre-training mechanism of BERT has profoundly influenced the development direction of the natural language processing field, promoting the shift from shallow feature extraction to deep semantic understanding. With the deepening of research and technological advances, more innovative pre-training strategies may emerge in the future to further enhance the performance and applicability of the model.

3.3 Figuring out the Input and Output of BERT

3.3.1 Input mechanism

The main inputs to the BERT model are vector representations of individual words/phrases (or called tokens) in the text. These vectors can be either randomly initialized or initialized by pre-training algorithms such as Word2Vec, GloVe, etc. to serve as initial values [7]. For the Chinese version of BERT, since the Chinese characters themselves are semantic units, a single Chinese character is directly considered as the basic input unit without the need for an additional segmentation step. This approach simplifies the preprocessing process while ensuring that each character can be encoded individually.

In practice, the input to the BERT model contains not only Token Embeddings, but also incorporates two other types of vector information:

Segment Embeddings: In order to distinguish different sentences in the same input sequence, BERT introduces Segment Embeddings. The vectors in this section are automatically learned during the model training process and are used to represent the global semantic features of text segments. For example, when processing two consecutive sentences, all the tokens of the first sentence are given the same paragraph ID (e.g., A) and the second sentence is given another ID (e.g., B). This helps the model understand the logical relationships between sentences, especially important in the Next Sentence Prediction task.

Position Embeddings: Considering that the position of a word in natural language has a significant impact on its meaning (e.g., “I love you” has a very different meaning than “You love me”), BERT attaches a specific position vector to each token. BERT attaches a specific position vector to each token. These vectors help the model capture information about the relative position of the words in the sentence, thus enhancing the understanding of the syntactic structure. It is worth noting that BERT uses absolute positional encoding rather than relative positional encoding, which means that each position has a fixed vector representation.

As shown in Figure 1, the input of BERT consists of three parts of vectors: Token Embedding to encode lexical semantics, Positional Encoding to inject sequence order information, and Segment Embedding to distinguish sentence attribution. The three are fused by element-by-element summation to form the input to the Transformer encoder. This design enables BERT to capture local lexical features, global context dependencies, and task-relevant structured information at the same time, thus performing well in various NLP tasks.

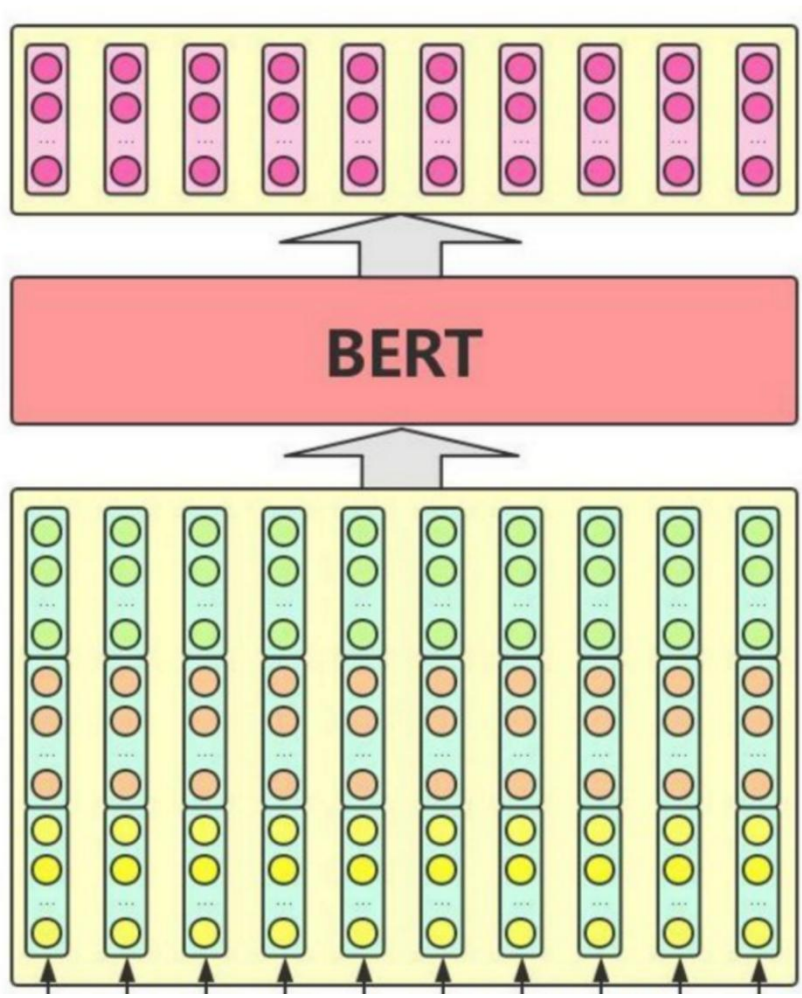


Figure 1 Introduction to the Inputs of the BERT Model

To summarize, the BERT model takes the sum of word vectors, paragraph embeddings and positional embeddings as the final input representation. In addition, BERT employs special markers to assist certain tasks, such as [CLS] and [SEP] markers. The [CLS] marker usually appears at the beginning of a sequence and represents the categorization information of the whole input sequence, while the [SEP] marker is used to separate different sentences or text blocks. For English, BERT further slices words into finer-grained subword units (WordPieces), e.g., playing is split into play and ###ing, which can better handle the problem of unknown or rare words.

3.3.2 Output mechanism

The output of the BERT model is a vector representation of each word/phrase fused with the full text semantic information obtained after multi-layer Transformer encoder processing [8]. Each output vector not only contains the features of the original word itself, but also incorporates the rich contextual information provided by the context. This deep bi-directional encoding enables BERT to generate more accurate and expressive textual representations for a wide range of NLP tasks. In particular, for some specific tasks such as classification, BERT utilizes the output vectors corresponding to the [CLS] tokens as a comprehensive representation of the entire input sequence. This is because the [CLS] markers are located at the very front of the sequence, which can theoretically capture the core information of the whole text. For other tasks, such as named entity recognition or question-answer systems, the output vectors corresponding to each token may be used directly for prediction. In conclusion, BERT realizes an effective mapping from simple vocabularies to complex semantic spaces by means of a well-designed input-output mechanism, which greatly improves the effect of natural language processing.

4 FUTURE PROSPECTS OF THE BERT MODEL

4.1 BERT Model for Sentence Semantic Similarity Task

Although BERT provides powerful language understanding capabilities and can be used to obtain word embeddings of individual sentences by inputting them and then generating a fixed sentence embedding using a specific strategy (e.g., taking the vector corresponding to the [CLS] tokens or averaging the vectors of all tokens), this approach is not always the optimal choice for real-world applications. practical applications, but this approach is not always optimal. Because the sentence embeddings directly output from BERT are not optimized specifically for the task, it is difficult to accurately assess the quality of these embeddings and their performance on a particular task. And in the semantic

similarity task, the sentence embeddings generated directly using BERT are not as effective as expected. This is because BERT is primarily designed for downstream task fine-tuning rather than as a generalized sentence representation tool. Although BERT is able to capture a certain amount of contextual information, it is not specifically optimized for semantic similarity at the sentence level when generating sentence embeddings. This means that even for the same sentence, different embedding results may be obtained in different contexts, which is not conducive to stable matching performance.

In view of the above problems, Sentence-BERT (SBERT) was developed, which aims to improve the way BERT generates sentence embeddings and make it more suitable for sentence-level semantic similarity tasks.SBERT not only utilizes the strong pre-training base of BERT, but also makes further fine-tuning on specific tasks to ensure that generated sentence embeddings are more in line with the real-world requirements. needs. In order to improve the quality of sentence embeddings, SBERT employs a contrastive learning approach that encourages the model to generate more discriminative embedding representations that better capture the semantic differences between sentences. Compared with the direct use of BERT, SBERT optimizes the computational process of sentence embedding, reduces redundant operations, and improves computational efficiency.

In summary, although BERT itself is already a very powerful language model, it is not an ideal solution directly applicable to all tasks. For application scenarios that require high-quality sentence embedding, especially those tasks that emphasize semantic similarity, SBERT provides a more optimized alternative that inherits the advantages of BERT while targeting its shortcomings when applied in this domain.

4.2 BERT Model for Targeting Sentiment Analysis

Sentiment classification refers to the automatic determination of the emotional tendency expressed in the text through natural language processing techniques, which is usually divided into two categories: positive and negative. For example, “Today's meal is too delicious” expresses a positive sentiment, while “Today's meal is unbearable” reflects a negative sentiment. With the development of deep learning, especially the emergence of pre-trained language models such as BERT, the performance of sentiment classification tasks has been significantly improved [9].

When using BERT for sentiment classification, the input sentences first need to be pre-processed appropriately. Specifically, a special marker [CLS] (Classification) is added to the top of the sentence before it is fed into BERT. The function of this marker is to provide a global representation point for the whole sentence, allowing BERT to generate a vector that comprehensively reflects the semantic information of the whole sentence. In addition, each word in the sentence is also converted into the corresponding token and subjected to the necessary disambiguation process according to the requirements of BERT.

When the input is passed to BERT after the above pre-processing, it utilizes the Self-Attention Mechanism (SAM) to compute the context-sensitive vector representation of each token. The core advantage of the Self-Attention Mechanism is that it can consider both global information and local focus, i.e., it not only focuses on the current token itself, but also combines all the previous and previous contexts for comprehensive understanding. Therefore, for [CLS] tokens, the output vector it corresponds to has actually integrated the key semantic features of the whole sentence, especially some important words or phrases will have more influence weight on this vector.

Based on the [CLS] vector generated by BERT, we can use it directly as a sentence-level representation for subsequent classification tasks. To accomplish the final sentiment categorization, a simple Fully Connected Layer is usually added on top of the BERT, which is responsible for mapping the [CLS] vectors to specific category labels (e.g., positive or negative). Since BERT itself is a powerful pre-trained model with rich language understanding and representation capabilities, in many cases it is straightforward to fix the parameters of BERT and adjust only the parameters of the Fully Connected Layer to suit the specific task requirements. This can not only greatly reduce the training time and resource consumption, but also effectively prevent the occurrence of overfitting phenomenon. Of course, if the conditions allow, one can also choose to fine-tune the parameters of BERT and fully connected layer at the same time. Although this approach may increase the training cost, it helps to further optimize the model performance, especially in scenarios where the dataset is large and diverse, and joint fine-tuning often leads to better results [10].

By leveraging BERT's powerful pre-training capabilities and well-designed classification architecture, we are able to achieve efficient and accurate results in sentiment classification tasks.BERT's self-attention mechanism ensures the effectiveness of the [CLS] vectors, making it an ideal bridge between the pre-trained model and the downstream task. Whether we choose to fix or fine-tune the BERT parameters depends on the specific application scenario and technical resources, and the flexible choice can help us achieve the best balance under different conditions.

5 CONCLUSION

With the dramatic growth of Chinese news data, how to efficiently and accurately categorize news has become an important research topic in the field of natural language processing (NLP). In this paper, by comparing the features of traditional RNN (Recurrent Neural Network), CNN (Convolutional Neural Network), AVG (Average Word Embedding) deep learning models and transformer-based BERT model, we summarize and analyze the current research status of Chinese news text classification based on Chinese news text classification, and with the latest research advances, we give a possible direction of development for BERT model. Although Chinese news text categorization has made significant progress in utilizing deep learning techniques, especially with the support of large-scale datasets and

advanced models to efficiently and accurately categorize news, Chinese news text categorization still has a long way to go, and the main problems we are currently facing are: high-quality labeled data is crucial for training effective text categorization models. However, creating and maintaining a large-scale, high-quality Chinese news corpus is a time-consuming and expensive process, and requires specialized knowledge to ensure the accuracy of the annotation; Chinese has a rich vocabulary and complex grammatical structure, and the phenomenon of synonyms and polysemous words is common, which poses a challenge to accurately understand and categorize news texts. In addition, cultural background and context dependency also increase the difficulty of correctly parsing the semantics of the text; in news texts, the number of documents in different categories may vary significantly, with a large number of related articles on some popular topics and a scarcity of literature on some niche or emerging areas. Although there are several methods for dealing with the category imbalance problem, such as data augmentation and assigning different weights to different categories in the model loss function, these methods have improved the classification results to some extent. However, in the face of extreme category imbalance, it is still difficult for the existing methods to respond effectively, leading to the possibility that the model may be biased in favor of the majority class, thus affecting the classification performance of the minority class. It is believed that these problems will be gradually alleviated through continuous technical innovation and research exploration in the near future, thus promoting Chinese news text categorization technology to a higher level.

CONFLICT OF INTEREST

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Yu Tongrui, Jin Ran, Han Xiaozhen, et al. A research review of pre-training models for natural language processing. Computer Engineering and Applications, 2020, 56(23): 12-22.
[2] Zheng Yuanpan, Li Guangyang, Li Ye. A research review on deep learning in image recognition. Computer Engineering and Applications, 2019, 55(12): 20-36.
[3] Cheng Yan, Yao Leibo, Zhang Guanghe, et al. Multi-channel CNN and BiGRU for text sentiment propensity analysis based on attention mechanism. Computer Research and Development, 2020, 57(12): 2583-2595.
[4] Duan Dandan, Tang Jashan, Wen Yong, et al. A short Chinese text classification algorithm based on BERT model. Computer Engineering, 2021, 47(01): 79-86.
[5] Liu Huan, Zhang Zhixiong, Wang Yufei. A research review on the main optimization and improvement methods of BERT model. Data Analysis and Knowledge Discovery, 2021, 5(01): 3-15.
[6] Yang Pei, Dong Wenyong. A Chinese named entity recognition method based on BERT embedding. Computer Engineering, 2020, 46(04): 40-45+52.
[7] Zhang ZiNiu, Jiang Mang, Gao Jianwei, et al. Chinese named entity recognition method based on BERT. Computer Science, 2019, 46(S2): 138-142.
[8] Yue Zengying, Ye Xia, Liu Ruiheng. A review of research on pre-training techniques based on language modeling. Journal of Chinese Information, 2021, 35(09): 15-29.
[9] Wang Ting, Yang Wenzhong. A review of research on text sentiment analysis methods. Computer Engineering and Applications, 2021, 57(12): 11-24.
[10] Wu Jun, Cheng Yao, Hao Han, et al. Chinese terminology extraction based on BERT embedded BiLSTM-CRF model. Journal of Intelligence, 2020, 39(04): 409-418.

World Journal of Information Technology
ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3029

MATHEMATICAL MODELING COMPETITION METHODS AND EXPERIENCE SHARING: IN-DEPTH ANALYSIS BASED ON MULTIPLE CONTEST PROBLEMS

DingShu Yan
College of Life Sciences, Shanxi Agricultural University, Jinzhong 030801, Shanxi, China.
Corresponding Email: yds2004517@163.com

Abstract: As a highly comprehensive discipline competition, the mathematical modeling contest integrates mathematical theory, computer technology, and practical problem-solving skills, providing a broad interdisciplinary practice platform for students. This paper selects typical problems from three competitions—the 2024 "Jindi Cup" Shanxi Province College Students Mathematical Modeling Contest, the Third National College Students Big Data Analysis Technology Skills Competition, and the 10th Digital Dimension Cup International Collegiate Mathematical Modeling Challenge—as research objects. This paper conducts a deep analysis of the methods and practical experience in mathematical modeling competitions, detailing specific approaches for key stages such as data processing, model construction/selection, and result optimization/verification. A comprehensive and systematic analysis is performed on the non-awarded work "Evaluation of Urban Resilience and Sustainable Development Capacity," providing reflections on deficiencies in data quality, model design, and paper composition. These insights aim to offer directions for improvement to future participants, thereby enhancing their comprehensive abilities and competition performance.
Keywords: Mathematical contest in modeling; Methodology; Experience summary; Model ealuation; Competition strategy

1 INTRODUCTION

As an important vehicle for interdisciplinary education, mathematical contest in modeling (MCM) originated in the United States in the 1980s. After over 40 years of development, it has become one of the world's most influential academic competitions, attracting participants from more than 100 countries [1]. The competition integrates mathematical theory with practical challenges in engineering, social economics, and other fields, significantly enhancing students' innovative thinking and practical abilities. However, with the rapid advancement of artificial intelligence, big data, and other technologies, the complexity of competition problems has increased exponentially, involving scenarios such as multi-objective optimization and high-dimensional data analysis [2]. The traditional "experience-driven" modeling approach is increasingly inadequate to meet these challenges, necessitating the establishment of a scientific methodological framework.

Current research primarily focuses on optimizing models for individual competitions or improving specific components like data cleaning and feature engineering. For instance, existing path planning studies mainly address enhancements to genetic algorithms or ant colony algorithms [3]. However, there is a lack of systematic exploration into integrating dynamic programming with clustering algorithms [4]. Models such as Lasso regression and gradient-boosted regression (GBR) are widely applied in prediction tasks [5], yet theoretical gaps remain in practical implementations of multi-model fusion. Additionally, few reflective studies have been conducted on non-award-winning submissions, and no reproducible improvement framework has been established [6].

This paper investigates three representative problems: soil survey path planning, red wine quality score prediction, and urban housing price forecasting. For the first time, a holistic methodological framework of 'data-driven modeling—model fusion—dynamic verification' is proposed in this paper. Specifically, this study evaluates a collaborative optimization model combining dynamic programming and K-means clustering [7], which addresses computational complexity in traditional path planning algorithms. A Lasso-Ridge-GBR fusion model is analyzed for its prediction accuracy improvement mechanism [8]. Through a comparative analysis of the non-award-winning submission Evaluation of Urban Resilience and Sustainable Development Capacity, this paper identifies structural flaws in data collection, index weighting, and report presentation. Based on these analytical insights, a "three-dimensional improvement strategy" is proposed as a reference framework for future competitions [9]. The aim of this research is to provide participants with systematic solutions from problem analysis to result presentation, thereby advancing theoretical innovation and practical applications in mathematical contest methodology [10].

2 ANALYSIS OF THE DEVELOPMENT COURSE AND FUTURE PROSPECT OF MATHEMATICAL MODELING

2.1 Development of Mathematical Contest in Modeling

Mathematical contest in modeling (MCM) originated in the United States in the 1980s. It was originally held jointly by the Mathematical Association of America (MAA) and the Society for Industrial and Applied Mathematics (SIAM). At that time, the competition aimed to stimulate college students' interest in mathematics and cultivate their ability to apply mathematical knowledge to practical problems. Competition topics typically involved simple real-world challenges such as population forecasting, resource allocation, and other similar scenarios. Teams were required to complete model formulation, solution, and report writing within a specified timeframe.

With the competition's continuous development and global promotion, its influence has expanded worldwide. An increasing number of countries and regions have initiated similar competitions, including China's National College Mathematical Contest in Modeling and the European Mathematical Contest in Modeling. These competitions provide practical platforms for students to apply mathematical knowledge to real-world problems, thereby strengthening the integration of mathematics with practical applications.

In China, the National College Mathematical Contest in Modeling has grown significantly since its establishment in 1992, evolving into one of the nation's largest and most influential academic competitions for undergraduates. Competition topics have become increasingly complex and diverse, covering fields such as environmental protection, economic and financial analysis, engineering, and other interdisciplinary areas. This trend imposes higher demands on participants' comprehensive capabilities.

2.2 Research Status of Mathematical Contest in Modeling

2.2.1 Data processing method

In mathematical modeling contests, data processing is a critical step whose quality directly influences the accuracy and reliability of the model. Researchers are dedicated to exploring effective data processing strategies for competitions, including data cleaning, feature selection, and feature engineering. For example, some scholars have proposed a deep learning-based feature selection method that can automatically identify the most relevant features for model prediction from large datasets. This approach not only enhances model efficiency and accuracy but also mitigates the impact of data noise. Additionally, studies focus on addressing issues such as missing data and outliers to ensure data integrity and consistency.

2.2.2 Model construction method

Model building lies at the core of mathematical modeling, with researchers focusing on constructing more efficient mathematical models. This process encompasses model selection, integration, optimization, and other related aspects. With advancements in machine learning and artificial intelligence, deep learning-based mathematical models have garnered increasing attention. These models can automatically extract complex patterns and relationships from data, thereby enhancing their predictive capabilities. For instance, some scholars have proposed convolutional neural networks (CNN) for processing image data, demonstrating effectiveness in tasks such as target recognition and image classification. Meanwhile, other researchers have applied recurrent neural networks (RNNs) to time-series prediction, achieving favorable outcomes. Concurrently, traditional mathematical models are undergoing continuous optimization to adapt to increasingly complex problems and evolving data characteristics.

2.2.3 Result optimization and verification method

Result optimization and validation are critical steps to ensure the accuracy and reliability of model outcomes. Researchers have explored strategies for optimizing model results, including multi-index evaluation, cross-checking, and result adjustment. For instance, some scholars have proposed a results optimization method based on machine learning, which automatically adjusts model parameters to enhance prediction accuracy. Through iterative refinement, the model can achieve superior performance across diverse datasets. Additionally, studies focus on constructing a comprehensive evaluation metrics system to objectively assess model performance, ensuring that results truly reflect the underlying patterns of real-world problems.

2.3 Proposition Trend of Mathematical Contest in Modeling

2.3.1 Interdisciplinary integration

With the increasing complexity of problems, competition topics are integrating knowledge and methods from mathematics, computer science, physics, biology, economics, and other disciplines. This interdisciplinary trend imposes higher demands on participants' comprehensive application capabilities. For example, in environmental protection, researchers may need to combine ecological, meteorological, and mathematical modeling approaches to investigate the impact of climate change on ecosystems. In finance, integrating knowledge of economics, statistics, and mathematical modeling is often necessary for analyzing market risks and investment strategies. Participants require interdisciplinary knowledge backgrounds and collaborative skills to effectively address such complex cross-disciplinary challenges.

2.3.2 Model innovation

The competition encourages participants to propose innovative mathematical models and methods to enhance prediction accuracy and model applicability. With the rapid advancements in science and technology and the increasing complexity of data, traditional mathematical models often fail to meet real-world demands. Therefore, participants must depart from conventional approaches and develop novel model architectures and algorithms. Concurrently, improving and optimizing traditional models remains critical. Through in-depth analysis and innovative modifications, existing

models can better adapt to emerging problems and evolving data characteristics, thereby enhancing their performance and application value.

2.3.3 Practical application

Competition topics are increasingly emphasizing close integration with real-world problems, requiring participants to apply mathematical modeling to address practical challenges. The outcomes of these competitions are being applied to real production and daily life, effectively serving real-world scenarios. For instance, in urban planning, mathematical modeling can optimize traffic flow and rationally allocate public facilities. In the medical field, it can predict disease transmission patterns and evaluate treatment efficacy. This problem-oriented approach in competition design brings mathematical modeling contests closer to societal needs, fostering the development of high-quality talents capable of applying mathematical knowledge to practical problem-solving. Consequently, it promotes the widespread adoption and innovative evolution of mathematical modeling technologies across diverse fields.

3 REFINEMENT OF MATHEMATICAL MODELING METHODOLOGY AND ANALYSIS OF BASIC PROCESS

3.1 Data Processing

In the entire mathematical modeling process, data processing is the most fundamental step. Its quality directly affects the accuracy and reliability of subsequent modeling and lays the foundation for the entire workflow. Given the massive data encountered in practical problems, data cleaning is the first critical task. This process aims to remove noise, correct errors, and impute missing values to ensure data accuracy and integrity. For example, when processing geographic data for soil surveys, researchers must carefully screen for duplicate, erroneous, or incomplete entries. For datasets with missing values, comprehensive integrity checks should be performed, and the proportion of missing values must be accurately quantified. If a high proportion of values are missing, appropriate imputation methods (such as mean imputation, median imputation, or model-based prediction) should be selected based on data characteristics to maintain integrity.

Following data cleaning is exploratory data analysis, which allows researchers to intuitively observe data characteristics through visual analysis. For instance, point distribution maps reveal geographic data patterns, while scatter plot matrices help identify feature correlations. If highly linearly correlated features are detected, researchers should consider removing some to mitigate collinearity issues in subsequent modeling. Data transformation is another critical component. To align data with model assumptions, appropriate transformation methods can be applied. The Box-Cox transformation, for example, normalizes data distributions through power transformations, thereby enhancing model fitting.

Additionally, feature engineering plays a pivotal role in extracting latent information from data. Categorical features can be converted to numerical formats via label mapping, enabling model processing. New features can also be derived through feature combination—for example, merging two related features to create a composite indicator like "alcohol-acidity ratio"—to uncover relationships between features and target variables. Leveraging domain knowledge, researchers can identify context-specific features related to the target and integrate them with the original dataset, thereby enriching the dataset's information content. In practical applications, such as predicting red wine quality scores, data integrity must first be verified. Statistical analysis of missing value proportions guides imputation strategies. A scatter plot matrix assesses correlations between chemical features; collinear features are removed to mitigate multicollinearity issues. Box-Cox transformations normalize data distributions, improving model fit. Label mapping converts categorical features like wine origin into numerical formats. Feature combination generates novel indicators like "alcohol-acidity ratio," uncovering stronger relationships with quality scores.

3.2 Model Construction and Selection

Model construction and selection are the core of mathematical modeling, aiming to select appropriate models based on problem characteristics and enhance model performance through model fusion. In path planning problems, the dynamic programming model, Haversine formula, and K-means clustering algorithm can be integrated. The Haversine formula accurately calculates the spherical distance between two points, providing a metric basis for path planning. The K-means clustering algorithm reduces problem complexity by grouping sampling points geographically. The dynamic programming model achieves global optimal path planning by recursively determining optimal paths within each cluster.

For prediction tasks, regression models such as Lasso, Ridge, Elastic Net, and Gradient Boosting Regression (GBR) are widely applied. Researchers select the optimal model by comparing performance across training and test sets. Model fusion, such as building stacked ensemble models, is an effective strategy for complex problems. This involves using predictions from multiple base models as new features for training a higher-level model.

In soil survey path planning, integrating the dynamic programming model with the Haversine formula and K-means clustering algorithm yields effective solutions. The Haversine formula converts geographic coordinates into real-world distances, enabling accurate distance measurements. K-means clustering reduces complexity by grouping sampling points. The dynamic programming model recursively optimizes paths within clusters to achieve global optimality. For red wine quality score prediction, diverse regression models are employed. Lasso regression incorporates L1 regularization to perform feature selection and simplify model structures, enhancing interpretability. Ridge

regression applies L2 regularization to address multicollinearity and improve stability. Elastic Net combines L1 and L2 regularization to balance feature selection and model stability. Gradient Boosting Regression iteratively trains weak learners to gradually improve prediction accuracy. By evaluating performance across datasets, researchers select the optimal model for prediction.

3.3 Result Optimization and Verification

Result optimization and validation are the final critical steps in mathematical modeling, aimed at ensuring the accuracy and reliability of model outcomes. Evaluating model performance using multiple metrics provides a more comprehensive and precise reflection of its strengths. For prediction tasks, metrics such as mean squared error (MSE), root mean squared error (RMSE), and mean absolute error (MAE) are commonly used to quantify prediction accuracy. MSE measures the average squared difference between predicted and actual values. RMSE, as the square root of MSE, is more sensitive to errors and intuitively reflects the magnitude of discrepancies between predictions and observations. MAE, unaffected by error squaring, captures the average absolute deviation of predictions. In path planning problems, working time, total path length, sampling point coverage, and other indicators are used to comprehensively evaluate the advantages and disadvantages of path planning solutions.

Cross-validation is a critical method for evaluating model performance, which involves dividing the dataset into multiple subsets. In each iteration, one subset serves as the test set while the remaining subsets are used as training sets for multiple rounds of training and testing. The average of the test results is then adopted as the model's evaluation metric. For instance, the widely used 10-fold cross-validation partitions the dataset into 10 equal-sized subsets. Each subset is sequentially employed as the test set across 10 iterations of model training and testing. This approach enables a comprehensive assessment of the model's performance on diverse data splits, effectively mitigating biases caused by arbitrary dataset segmentation. Additionally, model results often require adjustment based on specific problem needs. For example, inverse transformations may be applied to prediction results to restore them to meaningful values. By comparing the distributions of predicted and actual data, researchers can refine model parameters or improve the model structure, thereby correcting prediction deviations and enhancing accuracy.

In the prediction of urban housing prices, a stacking ensemble model is constructed. Prediction results from multiple base models (e.g., Lasso regression, Ridge regression) are used as new features. These features are then input into a higher-level model (such as logistic regression or a neural network) for retraining, harnessing the strengths of each base model and mitigating single-model limitations. For instance, Lasso regression excells in feature selection, while Ridge regression stabilizes multicollinearity issues. By integrating their predictions, more accurate results can be achieved. In soil survey path planning, multiple algorithms optimize route design. K-means clustering groups sampling points, after which dynamic programming plans paths within each cluster. Finally, cluster-specific paths are integrated to achieve a globally optimal solution. This approach effectively reduces problem complexity and improves path planning efficiency and accuracy. For red wine quality and urban housing price predictions, evaluation metrics include mean squared error (MSE), root mean squared error (RMSE), and mean absolute error (MAE). By comparing these metrics across test sets, the most accurate model is selected. In soil survey path planning, working time serves as a key performance metric. Assuming uniform traffic speed (ignoring topography, weather, and vehicle factors), physics-based kinematic equations calculate daily task completion time. Additionally, total path length, sampling coverage, and other indicators comprehensively evaluate different route designs.

The cross-validation method is employed in multiple projects to ensure model stability and generalization. This technique involves dividing the dataset into multiple subsets, where one subset serves as the test set and the remaining subsets act as training sets in each iteration. After multiple rounds of training and testing, the average of the test results is adopted as the evaluation metric. For example, 10-fold cross-validation partitions the dataset into 10 equal-sized subsets, with each subset sequentially serving as the test set. The model is trained and tested 10 times, enabling a comprehensive assessment of its performance across diverse data splits. This approach mitigates biases caused by arbitrary dataset division.

Model results require iterative verification and adjustment. In urban housing price prediction, if the output undergoes a transformation, inverse transformation must be applied to revert the values to actual housing prices. By comparing the distributions of predicted and real-world housing price data, researchers can refine model parameters or improve the model structure. This process corrects prediction deviations and enhances accuracy.

4 ANALYSIS OF THE REASONS FOR THE UNAWARDED WORKS-URBAN FLEXIBILITY AND SUSTAINABLE DEVELOPMENT

4.1 Data Aspects

Data comprehensiveness and accuracy are critical for evaluating urban resilience and sustainability. However, many submissions exhibit significant flaws in data collection. Some participants solely rely on provided topic information, lacking the initiative to expand data sources. As a complex system influenced by multiple factors, urban resilience cannot be fully captured using only given data. For example, assessing a city's resilience to extreme weather requires not only existing infrastructure data but also historical records of extreme weather frequency, scope, and losses. These data are vital for evaluating resilience and identifying sustainability challenges, highlighting the topic's importance. Single-channel data collection often limits analysis outcomes, hindering accurate representation of urban

realities. Additionally, insufficient data timeliness and spatial resolution lead to results that fail to reflect dynamic changes and regional disparities.

4.2 Model Aspects

In constructing multi-index evaluation systems, many studies rely excessively on subjective judgment for indicator weight assignment, lacking scientific methodologies. Urban resilience and sustainability assessments involve multiple indicators, where reasonable weight allocation is critical for accuracy. Weight assignment directly influences evaluation reliability: arbitrary weights may either overlook or overemphasize key factors, distorting urban resilience representations. For instance, prioritizing economic metrics while neglecting social/environmental dimensions can lead to biased outcomes. Scientific approaches such as Analytic Hierarchy Process (AHP) and Principal Component Analysis (PCA) should be adopted to objectively determine weights. These methods quantify indicator importance through data-driven statistical analysis, mitigating subjective interference and enhancing evaluation credibility.

Selected models often fail to account for urban development's complexity and dynamics, leading to limited adaptability in real-world applications. Cities are complex, evolving systems influenced by multiple interacting factors. Neglecting nonlinear relationships and feedback mechanisms between these factors can oversimplify model assumptions. For example, using linear models to evaluate the impact of urban economic potential and social service completeness on sustainability may fail to capture complex interdependencies. This results in significant deviations between predicted/evaluated outcomes and real-world conditions. Additionally, delayed model updates to reflect evolving urban challenges reduce effectiveness. Inadequate adaptation prevents models from providing accurate predictions or actionable decision support, undermining the practicality and guiding value of evaluations.

4.3 Thesis Writing

The paper exhibits significant structural flaws, including an unclear logical framework and unnatural transitions between chapters and paragraphs. Such disorganization impedes readers from quickly grasping the core arguments and overarching research ideas in studies covering urban housing price forecasting, service level analysis, resilience and sustainability assessments, and future development scenarios. For instance, abruptly inserting a discussion on urban service indicators while introducing a housing price prediction model disrupts the narrative flow, making it challenging for readers to follow the paper's focus. This fragmentation undermines comprehension of both the housing price analysis and the broader research objectives. A well-structured paper should organize content logically, gradually expanding analysis and discussion to guide readers through the author's reasoning.

The description of the model solution process and results analysis in this paper is insufficiently detailed, failing to demonstrate the study's depth and breadth. In the model implementation section, there is a lack of detailed documentation for critical elements such as algorithm steps, parametric configurations, challenges encountered during implementation, and corresponding solutions. This omission hinders readers from replicating the research process. In the results analysis section, data and conclusions are merely listed, without in-depth exploration of the underlying mechanisms and practical implications. This oversight prevents readers from appreciating the research's value for urban planning and development. For instance, analyzing urban resilience results without contextualizing them within specific urban characteristics or policy backgrounds undermines the ability to provide actionable recommendations for city managers.

The paper fails to effectively highlight its research innovations, nor does it clearly demonstrate unique contributions and value relative to existing literature. In urban studies, numerous assessments and planning frameworks have been developed. Without explicit elaboration of methodological, perspectival, or conclusion-driven innovations, papers risk being overlooked by reviewers. For instance, even if novel indicators or indicator combinations are introduced in a multi-index evaluation system, their roles in enhancing assessment accuracy and effectiveness are not clearly articulated. Innovation constitutes the paper's core strength, which should be underscored through novel methodologies, unique perspectives, or original conclusions. This clarity enables reviewers to recognize the research's distinct value and contributions.

5 IMPROVEMENT MEASURES AND FUTURE PROSPECTS

5.1 Improvement Measures

During data collection, researchers should leverage provided data while actively expanding additional relevant data sources. A comprehensive approach covering diverse information channels is essential. For instance, assessing urban resilience and sustainability requires collecting not only infrastructure data but also historical records of extreme weather frequency, affected areas, and losses. Resident satisfaction survey data should also be incorporated to reflect urban realities from multiple dimensions. In data processing, rigorous quality control is critical. Researchers must promptly identify and address missing or abnormal values to ensure data accuracy and completeness. Data cleaning techniques should be applied to correct or eliminate duplicate, erroneous, or incomplete records, laying a solid foundation for subsequent analysis.

When constructing a multi-index evaluation system, reducing subjective interference in indicator weight determination is crucial to enhance evaluation scientificity and reliability. Multivariate statistical methods such as Analytic Hierarchy Process (AHP) and Principal Component Analysis (PCA) can be employed to objectively assign weights. During model construction, problem complexity and dynamic nature should be fully considered, with rational selection of model architecture and parameter configurations. This ensures enhanced adaptability and predictive capability. For complex urban development challenges, nonlinear or dynamic models are recommended to accurately capture complex relationships between factors. These models enable realistic simulation and prediction of urban development trends, providing a robust foundation for urban planning and policy-making.

When writing a paper, it is essential to ensure a clear, logical, and coherent structure, with natural transitions between chapters and paragraphs. Content organization should follow the logical sequence of problem formulation, data processing, model construction, results analysis, and conclusion, enabling reviewers to quickly grasp core arguments and the overarching research framework. In describing the model-solving process, detailed documentation of algorithm steps, parameter settings, and challenges encountered during implementation (along with corresponding solutions) is critical to ensure reproducibility. In the results analysis section, in-depth exploration of data-driven insights and their practical implications should be provided, linking research outcomes to real-world urban planning and development contexts. Additionally, the paper must explicitly articulate its innovations, comparing them to existing literature. Unique contributions in methodology, perspective, or conclusions should be clearly demonstrated to strengthen the paper's academic rigor and impact.

5.2 Future Outlook

Mathematical modeling plays an irreplaceable role in cultivating students' innovative thinking, practical problem-solving skills, team spirit, and positive outcomes. As competition scales expand and problem complexity increases, future mathematical modeling competitions will emphasize innovation, model practicality, and result accuracy. Competitors must continuously learn and master advanced modeling methods and technologies to enhance their comprehensive competencies, aligning with competition development trends. Meanwhile, competition organizers should optimize rules and evaluation criteria, fostering a fairer, more equitable, and transparent competition environment. This will drive the mathematical modeling contest toward higher-quality development.

6 CONCLUSION

Through an in-depth analysis of three contest entries, this paper summarizes practical mathematical modeling methods and valuable insights. In the data processing stage, data cleaning, exploratory analysis, data transformation, and feature engineering are critical for unlocking data value and enhancing model performance. Model construction and selection should align closely with problem characteristics, adopting either single models or model fusion strategies to ensure accuracy and applicability. Results validation leverages multi-index evaluation and cross-validation to strengthen result reliability and validity.

Reflection on non-winning submissions highlights common challenges in data processing, model construction, and paper composition. The improvement strategies proposed here provide clear guidance for future participants to enhance their work, mitigating recurrence of similar issues. These measures promote the overall competency of contestants in mathematical modeling competitions and drive the robust development of modeling activities.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Applied mathematical modeling: a multidisciplinary approach. CRC Press, 1999.
[2] Liu J, Wang S, Xu H, et al. Federated learning with experience-driven model migration in heterogeneous edge networks. IEEE/ACM Transactions on Networking, 2024.
[3] Liang Y, Wang L. Applying genetic algorithm and ant colony optimization algorithm into marine investigation path planning model. Soft Computing, 2020, 24: 8199-8210.
[4] Moreno S, Pereira J, Yushimito W. A hybrid K-means and integer programming method for commercial territory design: A case study in meat distribution. Annals of Operations Research, 2020, 286: 87-117.
[5] Thirumani A, David D. Enhancing the accuracy of forecasting the upper respiratory infections due to particulate air pollution using lasso regression in comparison with gradient boosting regression//AIP Conference Proceedings. AIP Publishing, 2024, 3193(1).
[6] Knoll F, Murrell T, Sriram A, et al. Advancing machine learning for MR image reconstruction with an open competition: Overview of the 2019 fastMRI challenge. Magnetic resonance in medicine, 2020, 84(6): 3054-3070.
[7] Yadav R, Sreedevi I, Gupta D. Bio-inspired hybrid optimization algorithms for energy efficient wireless sensor networks: a comprehensive review. Electronics, 2022, 11(10): 1545.
[8] Xu G, Li W, Liu J. A social emotion classification approach using multi-model fusion. Future Generation Computer Systems, 2020, 102: 347-356.

[9] Büyüközkan G, Ilıcak Ö, Feyzioğlu O. A review of urban resilience literature. Sustainable Cities and Society, 2022, 77: 103579.
[10] Oudenampsen J, Van De Pol M, Blijlevens N, et al. Interdisciplinary education affects student learning: a focus group study. BMC medical education, 2023, 23(1): 169.

World Journal of Information Technology
ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3030

THE DESIGN OF INTELLIGENT COLLABORATION PLATFORM FOR AUTOMOBILE MANUFACTURING UNDER THE BACKGROUND OF INDUSTRY 5.0

Qiong He, LeXuan Chen, YiPeng Guo*, BoWen Gao
Beijing Information Science and Technology University, Beijing 100192, China.
Corresponding Author: YiPeng Guo, Email: 943642052@qq.com

Abstract: Based on Industry 5.0, this paper proposes an intelligent collaboration framework in the field of automobile manufacturing, emphasizing the three core concepts of people-centered, sustainability and resilience. The framework integrates artificial intelligence, network and collaborative technology to realize the transition from Industry 4.0 to 5.0, strengthen human-machine collaboration through artificial intelligence technology, and improve overall efficiency. The people-oriented design focuses on reducing the burden on employees, meeting individual needs and ensuring health and safety, highlighting humanistic care. Sustainability design focuses on environmentally friendly materials and supplier selection, energy optimization and environmental testing to reduce pollution, waste and achieve green production. Resilient design improves the adaptability and resilience of production systems in the face of challenges, and improves production flexibility and continuity through modular, intelligent scheduling and risk prediction mechanisms.
Keywords: Industry 5.0; Artificial intelligence; Automobile manufacturing; Platform design

1 INTRODUCTION

In the process of global industrial development, the EU officially announced the " Industry 5.0 : Towards a Sustainable, Human-centric and Resilient European Industry " report in April 2021, which established the basic concept of " Industry 5.0. "[1] In addition to the EU 's proposal that Industry 5.0 has three characteristics of human-centricity, sustainability and resilience, Zhuang et al.added the ' intelligence ' feature[2].

Domestic and foreign scholars have studied the application ways and supporting technologies of Industry 5.0. Slavic et al. conducted a survey on the European manufacturing industry and found that "people-centered" in Industry 5.0 includes the ability training of production employees as a focus on completing specific tasks, improving material consumption efficiency to promote sustainable development, and using standardized and detailed work instructions to make the system resilient [3]. Maddikunta discussed the potential applications of Industry 5.0 based on edge computing, digital twins, collaborative robots, Internet of Things, blockchain, 6G and other technologies in smart healthcare, cloud manufacturing, supply chain management and manufacturing production [4]. After analyzing the advantages, disadvantages, opportunities and threats brought by Industry 5.0 with SWOT, Kovari found that it is possible to achieve sustainable development goals and gain competitive advantages with industry 5.0, but attention should be paid to integrating human resources into the production process and dealing with safety and ethics issues [5]. Zhang Lili applied Industry 5.0 to enterprise human resource accounting and integrated accounting models, measurement methods and account Settings [6]. Jiang Zhoumingchi proposed a human-machine collaborative augmented manufacturing reference framework for industry 5.0, established a three-level product-economy-ecology model, and systematically expounded its core concepts, key technologies and typical scenarios through the development of human-information-physical system theory [7]. To sum up, foreign scholars have carried out extensive and in-depth application research, not only discussing the potential impact of industry 5.0 in multiple industries, but also conducting detailed discussion and verification for specific technical applications. In contrast, domestic scholars focus more on the concept definition and theoretical discussion of industry 5.0, although there is no lack of awareness of its importance, but relatively little research in the field of practical application.

In recent years, with the continuous innovation of production technology, some progress has been made in production efficiency and cost control. However, although the level of automation in the automobile manufacturing industry has been improved, there are still certain deficiencies in the fine management and intelligent application [8-9], which not only limits the further development of the automobile manufacturing industry, but also affects its competitiveness in the global industrial chain.With the continuous advancement of the industrial revolution, the automobile industry, as an important part of modern industry, is facing challenges and opportunities. The advent of Industry 5.0 provides new ideas for the intelligent, networked and collaborative automobile manufacturing industry. At present, the automobile manufacturing field is experiencing a profound change from traditional manufacturing to intelligent manufacturing, and the traditional automobile manufacturing mode has been difficult to meet the market demand in terms of efficiency, cost, quality and so on. How to improve the intelligent level of automobile manufacturing and realize the efficient collaboration of all links has become an urgent problem for automobile manufacturing enterprises.

This paper combines the three core features of industry 5.0: human-centered, sustainability and toughness to build an intelligent collaboration platform for automobile manufacturing to comprehensively optimize the automobile manufacturing process. The intelligent collaboration platform built can improve the production efficiency and product

quality of the automobile manufacturing process, and also meet the diversified value needs of employees, society and the environment and other stakeholders, reflecting the importance of social responsibility.

2 DESIGN OF INTELLIGENT COLLABORATION PLATFORM FOR AUTOMOBILE MANUFACTURING

Based on the three core concepts of Industry 5.0 generally recognized by the international community: people-oriented, sustainability and resilience as theoretical support, the intelligent collaboration platform for automotive manufacturing is built, as shown in Figure 1.

In practical applications, the human-centered concept includes the integrated application of intelligent job assistance systems to improve production efficiency and enhance job safety. In order to meet the individual operation needs of employees, the design of personalized work interface makes the working environment more convenient and easy to understand. Implementation of a health and safety monitoring system to improve workplace safety and the physical and mental health of employees.

In terms of environmental sustainability, the platform implements the concept of green supply chain management to realize the greening of the entire supply chain and the efficient use of resources. Energy management and optimization measures minimize energy consumption in the production process. Environmental monitoring and reporting mechanism can be set up to monitor and evaluate the impact of production activities on the environment in real time, and realize the environmental protection of the production process.

The concept of system resilience is reflected in the adaptability of future technology and market changes. The modular and scalable design allows the system to easily respond to technology upgrades and market changes. The construction of intelligent scheduling and cooperative system can realize efficient coordination and adaptive adjustment of production process, and improve the continuity and stability of production. Risk prediction and response mechanism can identify and respond to potential production risks in advance to ensure smooth production.

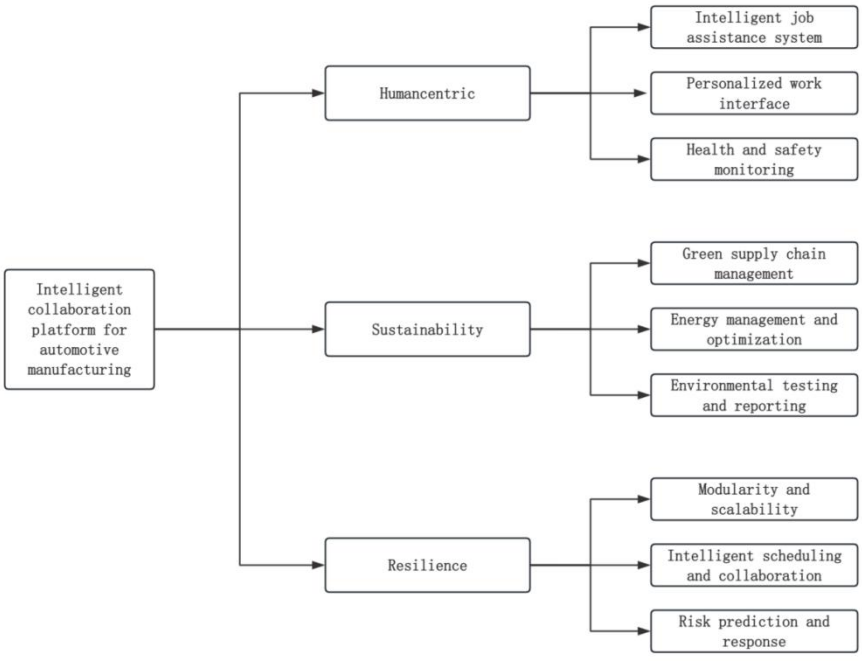


Figure 1 Intelligent Collaboration Platform for Automobile Manufacturing

3 PEOPLE-ORIENTED PLATFORM DESIGN

3.1 Intelligent Job Assistance System

The Intelligent Job Assistance System is a comprehensive solution that integrates advanced technologies to increase productivity, reduce error rates, reduce employee burden, and ensure employee safety in the automotive manufacturing process. As shown in Figure 2, the system is composed of intelligent devices, sensor networks, central control systems and human-computer interaction interfaces to achieve intelligent management and optimization of production processes. In the field of automobile manufacturing, the advanced technology of intelligent job assistance system integration has brought significant improvement to the entire manufacturing process [10].

(1) High-performance intelligent equipment. Automated robots and production lines can independently complete complex production tasks such as stamping, welding, and painting with little or no human intervention. Through the built-in algorithm, the equipment can realize self-optimization and adjustment to adapt to the production needs of different models and parts, ensuring the efficiency and flexibility of automobile manufacturing to meet the rapidly

changing needs of the market.

(2) Sensor network. In the stamping, welding, painting and other production lines of automobile manufacturing, various sensors are arranged to monitor temperature, pressure, speed and other parameters in real time, and the sensor transmits the data to the central control system in real time to provide accurate data support for production. Based on data analysis, the system can accurately judge the production status, predict potential problems, and automatically adjust equipment parameters or take corresponding measures to ensure the stability and safety of the production process.

(3) Central control system. The system is responsible for receiving sensor data, processing instructions, coordinating equipment operation and other tasks. Based on data processing ability and intelligent decision-making ability, the system can automatically adjust production parameters and equipment operation status according to real-time data and analysis results, and optimize the production process. The central control system can also be integrated with ERP, SCM and other management systems to realize the sharing and collaborative management of production data, and provide more comprehensive and accurate data support for enterprise decision-making.

(4) Human-computer interaction interface. The human factor engineering-based interface takes into account employee habits and visual needs to facilitate automotive manufacturing employees. Facilities such as safety guardrail and emergency stop button are set up around the production line, so that employees can take quick measures in emergency situations to avoid accidents. The intelligent operation assistance system also has the function of employee location and movement monitoring, once the dangerous actions of employees are found, the system will immediately send an early warning signal and start emergency measures to ensure the safety of employees.

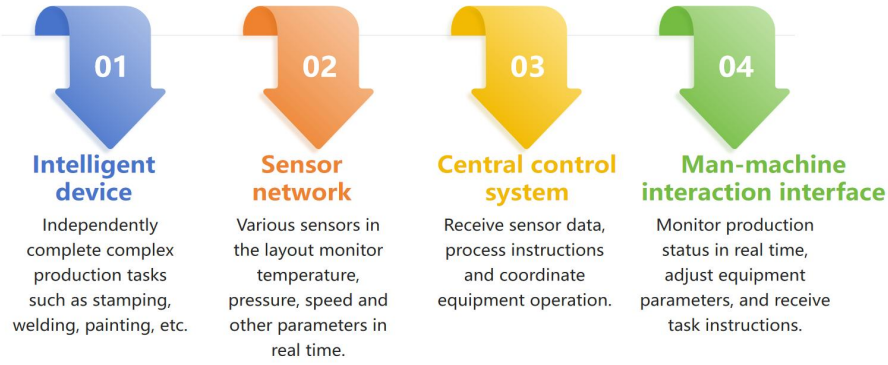


Figure 2 Intelligent Job Assistance System

3.2 Personalized Work Interface

Intelligent job assistance system integrates advanced industrial Internet technology to meet the diversified operating habits and preferences of workers by introducing personalized work interface, which allows workers to customize according to their own operating habits and preferences [11]. Different workers may have different needs for information display and operation, and personalized customization can improve operating efficiency and reduce the error rate.

In the automotive manufacturing process, workers need to frequently operate various equipment and systems, personalized work interface can not only reduce the adaptation time of workers in the operation process, but also reduce production accidents and product quality problems caused by misoperation. The personalized work interface of the intelligent job assistance system allows each worker to customize it according to their own work characteristics and needs. Different workers may have different requirements for the display mode of information, the order of arrangement and the layout of the operation interface, and the personalized work interface is more in line with personal operation habits.

The personalized work interface provides a real-time feedback mechanism that makes it easy for employees to advise on issues in the production process. Workers can view the production data, equipment status, product quality and other information on the platform interface in real time, and can deal with and feedback on abnormal situations in time.

3.3 Health and safety monitoring system

The integration of health and safety monitoring systems improves the level of safety management in the automotive manufacturing industry and enhances the safety of workers. The health and safety monitoring system provides detailed health data and safety records, provides strong support for the health management and safety assessment of enterprises, and helps to enhance the job satisfaction and loyalty of workers, stimulate the work enthusiasm and creativity.

In the automotive manufacturing industry, health and safety monitoring systems reflect the care of workers' health. As shown in Figure 3, in recent years, the automobile manufacturing platform is gradually integrating health and safety management system, which ensures the safety of the production process by monitoring the physical condition of workers and the safety status of the working environment in real time, and sending abnormal information to the mobile device of the manager, and starting the alarm system when necessary.

The platform monitors workers' heart rate, blood pressure and other key physiological indicators in real time through

wearable devices such as smart wristbands. With high-precision biosensing technology, the equipment can collect data continuously and transmit it to the management system in real time through wireless network [12]. Video surveillance and sensor technology monitor the safety status of the production site in real time. Hd cameras are installed at key locations in the production line to monitor the situation on the production site in an all-round way. A variety of sensors are deployed in places such as machinery, work areas and passageways to detect safety related parameters such as temperature, humidity, smoke and hazardous gases.

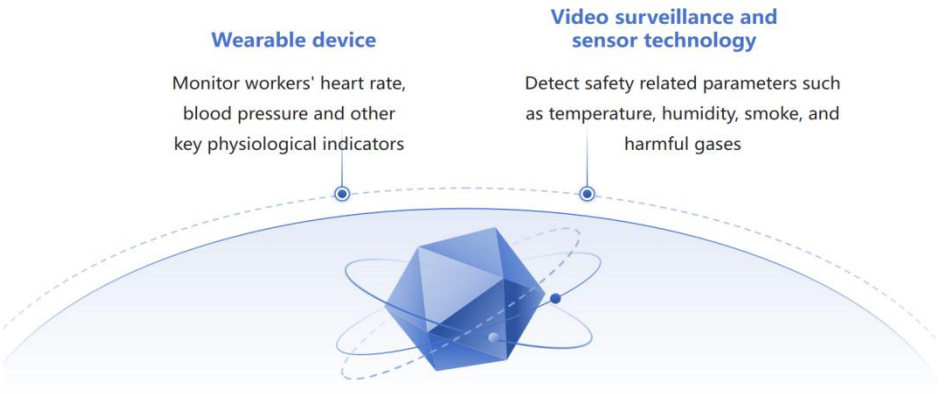


Figure 3 Health and Safety Monitoring System

4 SUSTAINABLE PLATFORM DESIGN

The application of the sustainable design concept of Industry 5.0 in automotive manufacturing is reflected in the design stage of the product, throughout the production and use process, as shown in Figure 4.

In terms of material selection, automobile manufacturers are increasing the use of recyclable, renewable or environmentally friendly materials to reduce the production of environmental pollutants while improving the environmental performance of vehicles. For example, the use of lightweight alloy materials can reduce the weight of cars, thereby reducing fuel consumption and emissions; The use of bio-based materials instead of traditional petroleum-based materials can reduce the dependence on fossil fuels.

Energy saving and emission reduction are the goals of sustainable design. Automotive manufacturers optimize production processes and equipment to reduce energy consumption and waste emissions. Adopt advanced stamping, welding, painting and assembly technology to improve production efficiency and material utilization; Active use of solar, wind and other clean and renewable energy sources to power the production process.

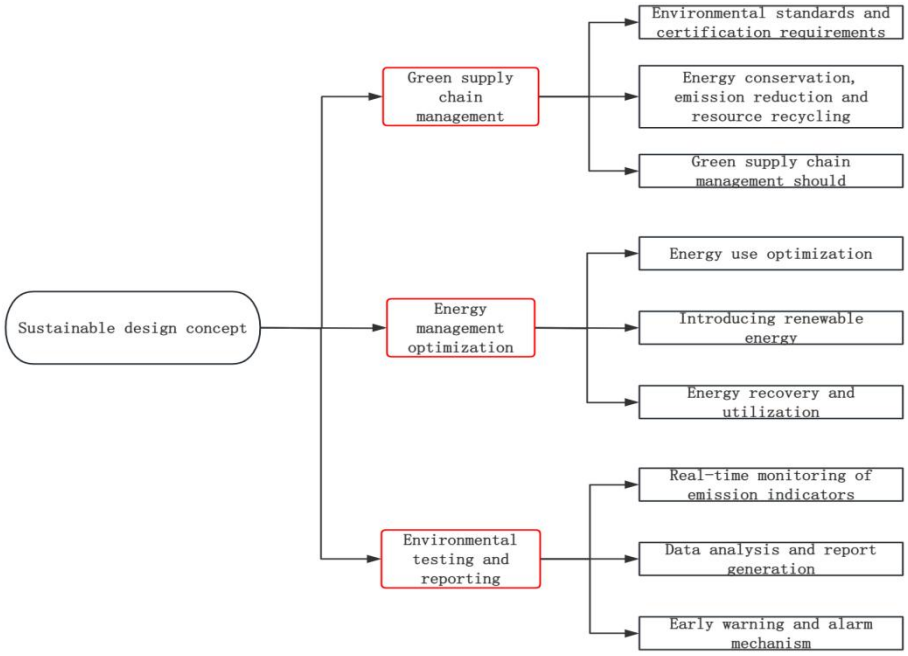


Figure 4 Sustainable Design Concept

4.1 Green Supply Chain Management

Green supply chain management requires enterprises to pay attention to environmental protection and resource conservation in the supply chain to achieve a win-win situation of economic and environmental benefits [13]. In partnership with certification bodies, automotive manufacturers can audit and certify key links in the supply chain to make suppliers and manufacturers comply with environmental regulations and standards. Enterprises adopt efficient production processes and equipment to reduce waste emissions; Encourage suppliers to adopt environmentally friendly packaging and transportation methods to reduce their environmental impact; Promote the recycling of waste, reduce the need for raw materials, and achieve the recycling of resources.

Green supply chain management should establish a cooperative relationship with suppliers, automotive manufacturing enterprises and suppliers to jointly develop environmental management plans, comply with environmental regulations and standards, share environmental technology and experience, and jointly promote the environmental protection and sustainability of the supply chain. Green supply chain management requires the support of information sharing and transparency. Based on data sharing and Internet technology, automobile manufacturers can monitor and manage the sustainability of raw material procurement, parts production, logistics distribution, product manufacturing until the final sales and recycling of the supply chain in real time.

4.2 Energy Management and Optimization

The function of the platform is to integrate all kinds of energy data acquisition equipment, and obtain the energy consumption data of each link in the production line in real time and accurately. The platform monitors energy consumption in the automotive manufacturing process and provides a visual display function to present complex energy data in an intuitive, easy-to-understand chart form. The platform can further process the data, dig and analyze the collected energy data in depth, and identify the main bottlenecks of energy consumption and potential optimization space [14].

(1) Optimization of energy use

Based on the platform's analysis of energy consumption bottlenecks and optimization space, specific energy-saving measures are formulated. First, improve the process, improve production efficiency and reduce ineffective energy consumption and other ways to reduce energy consumption. Second, the use of advanced energy-saving equipment and technology to improve the energy efficiency of equipment. Third, strengthen energy management, promote the realization of more comprehensive energy-saving goals by improving the awareness of employees, and further promote sustainable development.

(2) The introduction of renewable energy

In order to further reduce carbon emissions and energy consumption, the introduction of renewable energy in the production process, such as the installation of solar power generation systems on the factory roof, using solar energy to provide part of the power supply for the production line; The construction of wind power stations in areas rich in wind resources provides clean energy for automobile manufacturing, thereby reducing production costs.

(3) Energy recovery and utilization

In the process of automobile manufacturing, energy recovery and utilization technology is introduced to realize the efficient use of energy. For example, the use of waste heat to generate electricity, recycling useful components in waste gas, etc., can reduce energy consumption, reduce environmental pollution and waste emissions.

4.3 Environmental Testing and Report

In the process of the automobile manufacturing industry's transformation to industry 5.0, the environmental monitoring and reporting system reflects the enterprise's commitment to environmental protection and the advanced level of industrial intelligent and digital management [15].

The environmental monitoring and reporting system takes ESG as the core driving force, and comprehensively covers the environmental monitoring of the whole process of automobile manufacturing by building an integrated and intelligent platform. Using Internet of Things (IoT) technology, the system deploys a network of sensors at key locations in the production line to collect data on emissions such as gas, wastewater, and solid waste in real time and transmit the data to a central processing unit.

In terms of environment (E), the system is densely packed with a network of sensors in the production line to monitor data on emissions such as gas, wastewater and solid waste in real time. The environmental monitoring and reporting system has the function of data analysis and report generation. Based on machine learning analysis and other technologies, the collected data is deeply processed and analyzed to extract valuable environmental protection information.

On the social (S) side, the Environmental Monitoring and reporting system focuses on the environmental performance of enterprises and emphasizes the social responsibility of enterprises. The system tracks the environmental protection input and results of enterprises in real time, and supervises the implementation of their commitments to society. The system provides decision support for enterprises, helps enterprises to better manage environmental affairs, and improves the reputation and influence of enterprises in society.

In terms of governance (G), based on transparent and standardized environmental management, the system records and reports the environmental protection data of enterprises in real time, providing accurate and comprehensive information for regulators. The system has an early warning mechanism, when the emission data exceeds the preset threshold, the

system will trigger an alarm to remind the enterprise to take effective measures in time.

5 THE PLATFORM DESIGN OF TOUGHNESS

The concept of resilient design emphasizes the ability of a system to maintain its function and performance in the face of uncertainty. As Figure 5 shows, in automotive manufacturing, a highly flexible and adaptable production system can effectively cope with the challenges of fluctuating raw material supply, changes in market demand, and technological updates.

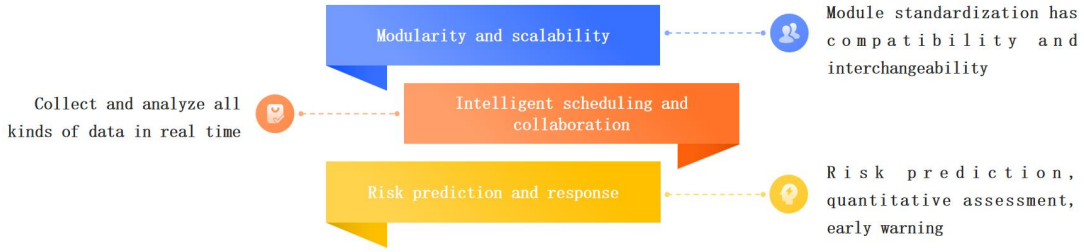


Figure 5 Toughness System

5.1 Modularity and Scalability

Modular design makes the manufacturing process of the car more flexible and efficient by dividing the car into multiple independent assemblies or modules, each with specific functions and interfaces.
According to the function and performance requirements of the car, the whole vehicle is divided into engine module, chassis module, body module, electronic and electrical module. Each module contains a set of related parts and subsystems to achieve a specific function; Module standardization Develop unified module interface standards and specifications to achieve compatibility and interchangeability between different modules; Each module is relatively independent and can be designed, manufactured, tested and verified independently, helping to reduce design risk, improve product quality, and facilitate subsequent module replacement and upgrade.
Extensibility design is an extension of modular design, emphasizing that on the basis of meeting current requirements, the system has the ability to adapt to changes in future requirements. Function expansion in automotive manufacturing By adding new modules or improving the function of existing modules, to achieve the expansion and upgrading of automotive functions. For example, the intelligent driving module is added to realize the automatic driving function of the vehicle; Add the entertainment information system module to improve the intelligence of the vehicle.

5.2 Intelligent Scheduling and Collaboration

In the field of automobile manufacturing, the design concept of toughness is combined to build an intelligent scheduling and collaborative system to effectively improve production efficiency, respond to market changes and enhance enterprise adaptability [16].
Intelligent scheduling system shows great potential and value in modern production environment through deep integration of artificial intelligence technology [17]. Based on big data and AI algorithms, the system collects and analyzes equipment operating status, material inventory changes, order demand information and other types of data in the production process in real time and accurately. Based on data analysis, the system automatically adjusts production plans and optimizes production processes. Using machine learning algorithms, intelligent scheduling systems can predict equipment failures and maintenance needs, schedule maintenance plans in advance, reduce equipment downtime, and improve production efficiency. Based on real-time data and preset rules, intelligent scheduling system can make production task allocation, production sequence optimization and other decisions, reduce human errors and improve the accuracy of production decisions.
In the process of automobile manufacturing, different departments and different production links need to realize real-time information sharing. By building a unified information platform, the collaborative system can realize real-time update and sharing of production data, equipment status, material inventory and other information. The collaboration system supports cross-departmental and cross-domain collaboration. Information sharing and real-time communication between different departments make it possible to work together more efficiently to solve problems in the production process. The collaborative system involves the collaboration with suppliers, logistics service providers and other external partners to achieve supply chain optimization and collaboration, and improve the efficiency of the entire automobile manufacturing process.
The resilient design concept emphasizes the ability of the system to recover quickly after damage. In the intelligent scheduling and cooperative system, the system can quickly resume operation in the case of equipment failure and network interruption, and continue to provide support for production.

5.3 Risk Prediction and Response

The risk prediction system collects data from different aspects of the production process, and on the basis of data analysis, the risk prediction system identifies various risks such as supply chain, technology and market. The system makes a quantitative assessment of various risks to determine their likelihood and impact. When potential risks are identified, the risk prediction system sets an early warning threshold and issues an early warning when the risk indicator exceeds the threshold. The system uses advanced algorithms such as machine learning to predict risks and provide basis for enterprises to formulate coping strategies in advance [18].

Based on the results of risk prediction, the enterprise makes corresponding emergency plans and clarifies key information such as response measures, responsible persons and implementation time. For different types of risks, enterprises need to allocate resources reasonably to respond to risks quickly. At the same time, enterprises also need to optimize the production process, improve the utilization rate of equipment and other ways to reduce the possibility of risk. Risk response requires the close coordination of various departments within the enterprise, and the enterprise needs to establish a cross-departmental communication mechanism, so that all departments can quickly cooperate in the risk response process. When the risk occurs, the enterprise needs to quickly adjust the production plan, optimize the allocation of resources and other strategies, and restore the production order after the risk is removed.

6 CONCLUSIONS AND PROSPECT

Based on the human-centric, sustainability and resilience concepts of Industry 5.0, this paper proposes the concept of an intelligent collaboration platform for the automotive manufacturing industry. The platform integrates intelligent, networked and collaborative technologies to give play to the advantages of humans and machines, improve manufacturing efficiency and product quality, reflect the diversified value needs of employees, society and the environment and other stakeholders, and highlight the awareness of social responsibility in the automotive industry.

First, the people-oriented design concept is based on intelligent work assistance system to reduce the labor intensity of employees, personalized work interface to meet the personalized work habits of employees, and health and safety monitoring system to ensure the safety and health of employees, show comprehensive care and respect for employees, and further create a healthy working environment for manufacturing enterprises.

Secondly, the concept of sustainable design focuses on environmental protection and efficient use of resources. Green supply chain management reduces pollution and waste in the production process. Energy management and optimization reduce energy consumption and carbon emissions in the production process by improving energy efficiency and adopting renewable energy sources.

Third, the design concept of resilience improves the adaptability and resilience of the production system through the three core functions of modularity and scalability, intelligent scheduling and coordination, and risk prediction and response, and helps enterprises to warn in advance and respond quickly to risks such as supply chain disruptions and natural disasters, and maintain the continuity of production.

For the future, Industry 5.0's concept of intelligent collaboration in the field of automotive manufacturing will continue to deepen and expand. With the continuous emergence of new technologies, such as the integration of quantum computing and biotechnology, the intelligent level of automobile manufacturing will be further enhanced. The deepening application of artificial intelligence will play a more important role in product design, production automation, quality control and other aspects, and promote the development of automobile manufacturing to a more efficient and accurate direction. Digitalization and data-driven decision making will become the new normal, helping automakers achieve more accurate market forecasting and production optimization.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

FUNDING

The research is funded by the Beijing Municipal Education Commission Research Plan General Project (grant number: KM202411232007).

REFERENCES

[1] Wang Wenjun. People-centered, sustainable and resilient industry 5.0 development in key emerging Frontiers. Science News, 2019, 25(06): 34.

[2] Zhuang Cunbo, Liu Jianhua, Zhang Lei. Connotation, Architecture and Enabling Technology of Industry 5.0. Journal of Mechanical Engineering, 2022, 58(18): 1-13.

[3] Slavic D, Marjanovic U, Medic N, et al. Evaluation of Industry 5.0 Concepts: Social Network Analysis Approach. Applied Sciences-Basel,2024, 14(3): 1291.

[4] Maddikunta Praveen Kumar Reddy, Pham Quoc-Viet Pham, Prabadevi B, et al. Industry 5.0: A survey on enabling technologies and potential applications. Journal of Industrial Information Integration,2022, 26: 100257.

[5] Kovari Attila. Industry 5.0: Generalized Definition, Key Applications, Opportunities and Threats. Acta Polytechnica Hungarica, 2024, 21(3): 267-284.
[6] Zhang Lili. Research on the integration of Enterprise Human resources accounting treatment under the background of Industry 5.0. Journal of Finance and Accounting, 2024(01): 99-104.
[7] Jiang Zhoumingqi, Xiong Yi, Wang Baicun. Man-machine Collaborative Additive Manufacturing for Industry 5.0. Chinese Journal of Mechanical Engineering, 2024, 60(03): 238-253.
[8] Zhou Changsen. Research on the Application of Mechanical Automation Technology in Automobile Manufacturing. Automotive Maintenance Technician, 2024(08): 130-132.
[9] Song Tuo. Research on Optimization of Intelligent logistics System of Automobile Manufacturing Industry under Lean Thinking. China Logistics and Purchasing, 2024(06): 105-107.
[10] Zhai Yutao. Research on Development and Application of Intelligent Job Sensing System for Concrete Machinery. Hunan University, 2024.
[11] Lu Feng, Wang Yu. Research on User experience-oriented APP Personalized interface Design. Home Theater Technology, 2023(04): 60-63.
[12] Zhou Haifeng, He Yong. Design of nursing home staff safety and health monitoring system based on Internet of Things. Software Engineering, 2022, 25(05): 19-22.
[13] Lu Wenping. Deepening material plan source Management to Help Green Modern digital Intelligence Supply Chain Construction. North China Electric Power Industry, 2024(03): 58-59.
[14] Sun Meng, Xiao Rongrong. Optimization Strategy analysis of Distributed Energy Management System. Integrated Circuit Applications, 2019, 41(03): 346-347.
[15] Li Fujian, Wu Jianbo, Ge Guojian. Analysis and processing of abnormal data in environmental monitoring. Environment and Development, 2019, 32(07): 158-159.
[16] Qin Zeyu, Wang Weitao, Feng Yinhui, et al. Research and application of intelligent Management and Control Platform for mechanical-mechanical Equipment of fully mechanized mining. China Coal, 2024, 50(02): 77-83.
[17] Cao Pengfei. Research on Optimization Strategy and Application of Automated Production in Intelligent Manufacturing. Science and Technology Information, 2023, 21(23): 242-245.
[18] Zhang Rui, Ma Jianjun, Ma Jin. Research on the application of digital twin in oilfield production management. China Management Information Technology, 2023, 26(14): 82-84.

World Journal of Information Technology
ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3031

A THEORETICAL ARCHITECTURE OF VOICEPRINT RECOGNITION FOR NETWORK SECURITY SITUATIONAL AWARENESS

Ping Xia
School of Engineering, Guangzhou College of Technology and Business, Foshan 528138, Guangdong, China.
Corresponding Email: 710398795@qq.com

Abstract: This paper proposes a theoretical framework for DenseNet-based voiceprint recognition, which incorporates spectrogram enhancement and adaptive histogram equalization to overcome the limitations of conventional methods in feature extraction robustness under noisy conditions. The framework synergistically combines spectral feature enhancement with DenseNet's dense connectivity, achieving both improved feature discriminability and deep feature reuse through: optimized time-frequency representation via enhanced spectrograms, hierarchical feature propagation enabled by dense blocks. Theoretical analysis confirms the framework's capability to maintain recognition stability against acoustic interference, establishing a novel biometric authentication paradigm for cybersecurity situational awareness systems.
Keywords: Voiceprint recognition; Spectrogram feature enhancement; Histogram equalization; Cybersecurity; Situational awareness

1 INTRODUCTION

With the continuous advancement and increasing sophistication of cyberattack techniques, traditional password-based authentication systems are facing severe challenges. As a non-intrusive biometric identification technology, voiceprint recognition offers multiple advantages including low-cost voice data acquisition, mature technology, low computational complexity of processing algorithms, and the capability for remote authentication, making it an ideal implementation technology for network identity recognition applications. However, the robustness of existing voiceprint recognition systems in complex network environments remains to be addressed. This paper combines spectrogram enhancement with histogram equalization to establish a fusion model integrating voiceprint recognition and cybersecurity situational awareness, aiming to resolve the issues of insufficient feature extraction and inadequate robustness in traditional voiceprint recognition technologies operating in complex network environments, while providing an interpretable theoretical framework for identity authentication scenarios.

2 RELATED WORK

The evolution of voiceprint recognition technology has progressed from traditional statistical models to contemporary deep learning approaches. Early methodologies predominantly employed machine learning algorithms such as Hidden Markov Models (HMM), Gaussian Mixture Models (GMM), GMM-Universal Background Models (GMM-UBM), GMM-Support Vector Machines (GMM-SVM), and i-vector systems [1-3]. While these approaches established foundational frameworks, they exhibit inherent limitations including oversimplified feature representations, limited recognition accuracy, susceptibility to channel variability, and inadequate noise robustness—constraints that significantly impede their efficacy in complex operational environments [4], particularly within cybersecurity situational awareness applications requiring high reliability.

The advent of artificial intelligence has catalyzed a paradigm shift, with deep learning emerging as the predominant research focus for voiceprint recognition [5-7]. In the context of cybersecurity authentication, voice biometrics now offer enhanced verification solutions for situational awareness systems. Liu et al. developed an LSTM-based architecture utilizing spectrogram representations of voiceprints [8], achieving superior text-independent recognition accuracy by capitalizing on LSTM's sequential modeling capabilities. Wang et al. advanced this domain through an end-to-end bidirectional LSTM framework that exploits temporal dependencies in speech sequences [9], demonstrating scalability for large-scale user authentication—a critical requirement for modern cybersecurity infrastructures.

For real-time network security monitoring, Zhao et al. introduced an end-to-end CNN architecture incorporating MFCC feature extraction and Universal Background Modeling [10], effectively mitigating environmental and individual variability. Yan et al. further optimized computational efficiency through a hybrid CNN-LSTM model processing fixed-length spectrograms [11], achieving high accuracy with reduced training iterations. Recent breakthroughs involve ResNet architectures that extract spatiotemporal voiceprint features [12-13], addressing historical challenges in recognition complexity and accuracy while enhancing practical deployment viability. Among existing research achievements, current studies on voiceprint recognition primarily focus on traditional voiceprint feature extraction methods, the implementation of voiceprint recognition using various deep learning approaches, and the application of multimodal fusion authentication technologies. The advancement of these technologies continues to enhance the application value of voiceprint recognition in cybersecurity situational awareness.

3 DESIGN OF VOICEPRINT RECOGNITION MODEL BASED ON DENSENET DEEP LEARNING

This paper proposes an end-to-end voiceprint recognition model based on the DenseNet architecture, which achieves efficient mapping from raw speech signals to speaker identity through a hierarchical feature learning mechanism. As illustrated in Figure 1, the system adopts a streaming processing framework that fully leverages DenseNet's advantages in feature reuse and gradient optimization. The integrated enhancement module at the front-end significantly improves the feature representation capability of traditional spectrograms in complex acoustic environments, thereby delivering a more robust identity authentication solution for cybersecurity situational awareness systems.

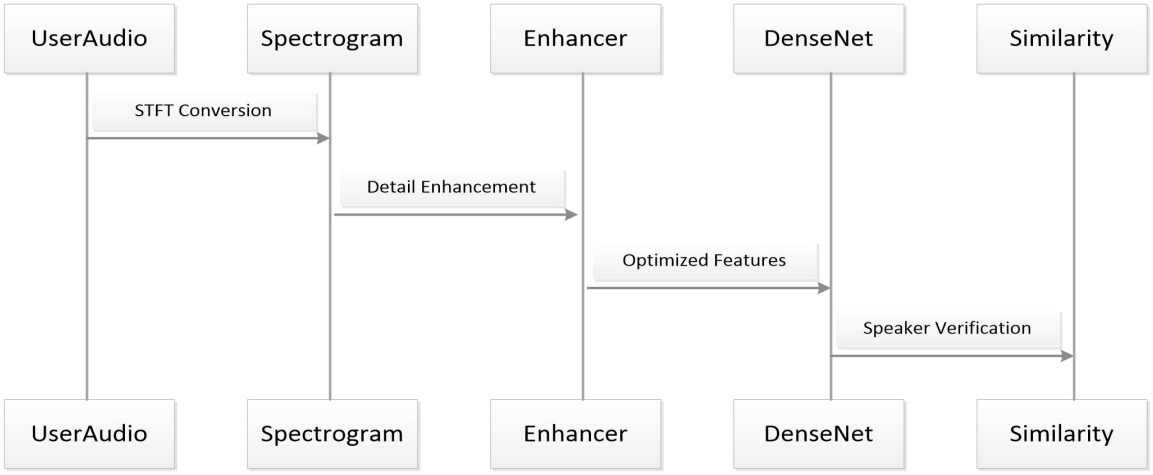


Figure 1 Architecture of Voiceprint Recognition System Based on DenseNet and Speech Spectrogram Enhancement

3.1 Spectrogram Generation

The spectrogram is a graphical representation of speech signals that transforms one-dimensional time-domain data into a three-dimensional image format, dynamically displaying temporal characteristics through the interplay of time-varying frequency components and energy intensity. Color gradients on the spectrogram form distinct texture patterns, which encode substantial speaker-specific biometric features, making this representation particularly suitable for training voiceprint recognition models using deep learning methodologies.As illustrated in Figure 2, the spectrogram generation process consists of three primary stages:

1. Preprocessing: The raw speech signal undergoes pre-emphasis to amplify high-frequency components, compensating for excessive attenuation during signal transmission. The processed signal is then segmented into fixed-duration frames, with each frame subjected to windowing and zero-padding operations to ensure continuity.
2. Power Spectrum Calculation: Each frame is processed through Fast Fourier Transform (FFT) to obtain its frequency spectrum. The magnitude spectrum is squared to derive the power spectrum, followed by logarithmic scaling (log-power) to enhance the dynamic range of spectral features.
3. Spectrogram Construction: The log-power spectra of individual frames are mapped onto a time-frequency coordinate system. Sequential frames are concatenated along the temporal axis to form the final spectrogram - a time-frequency-energy representation that completes the transformation from acoustic waveforms to visual discriminative features.This conversion from time-domain signals to frequency-domain visualizations provides a foundational analytical framework for subsequent deep feature extraction in voiceprint recognition systems.

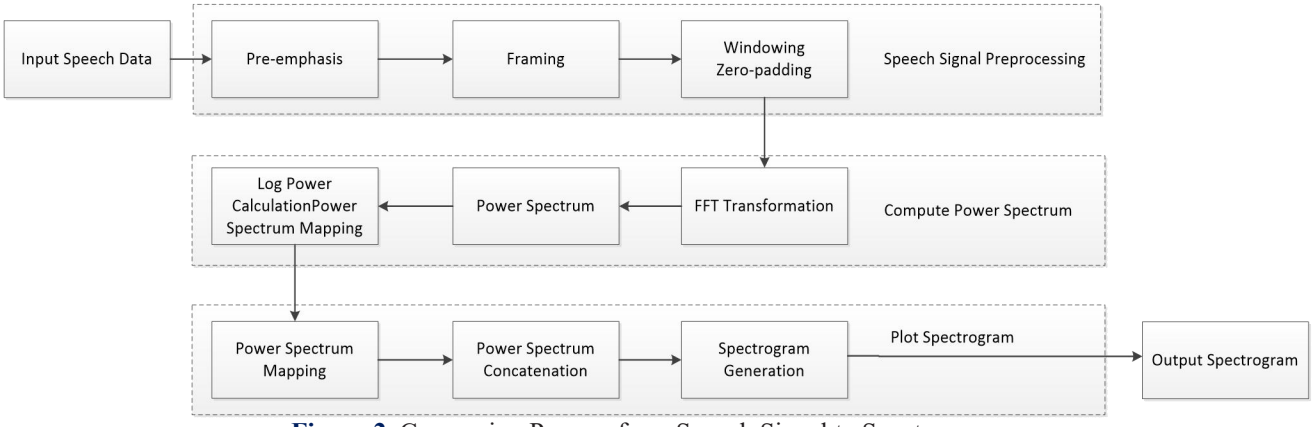


Figure 2 Conversion Process from Speech Signal to Spectrogram

3.2 Spectrogram Image Enhancement Algorithm

The integration of spectrogram image enhancement algorithms into voiceprint recognition systems facilitates the extraction of salient frequency-domain features from speech signals. By applying histogram equalization techniques,

this approach effectively disperses concentrated noise distributions while mitigating luminance variations caused by interspeaker differences or recording condition disparities, thereby significantly improving spectrogram quality.

Histogram equalization represents a computationally efficient nonlinear transformation method for spectrogram enhancement, operating through grayscale value redistribution to amplify contrast in images with constrained dynamic range. For grayscale spectrogram representations, the histogram provides a quantitative depiction of intensity level distributions, where visual quality exhibits direct correlation with the statistical moments (mean and variance) of grayscale distributions.

In terms of voiceprint feature processing, an improved histogram equalization algorithm is proposed, which has three key theoretical innovations. First, the quantization grading strategy, which establishes a more stable energy adjustment mechanism by converting the traditional continuous grayscale mapping into discrete grade adjustment, thus effectively avoiding the over-enhancement problem in low-energy frequency bands; second, the frequency domain perception mechanism: based on the physical characteristics of speech signals, an adaptive enhancement scheme is designed in the key frequency bands such as resonance peaks, which realizes the differentiated processing of different frequency bands; and third, the dynamic equilibrium design : The optimized balance model of global enhancement and local feature retention is constructed by introducing intelligent adjustment parameters, which theoretically ensures the quality of feature extraction. The algorithm theoretically realizes the balance between computational complexity and feature enhancement effect, and provides a more reliable input basis for subsequent deep feature extraction. From the theoretical analysis, this improved method is particularly suitable for complex acoustic environments where background noise or channel distortion exists.

3.3 DenseNet-based Voiceprint Recognition Network Model

In deep learning networks, the problem of gradient vanishing becomes more and more obvious as the depth of the network increases. Compared to ResNet, DenseNet's algorithm and network structure, although different, are to connect all layers directly to each other under the premise of ensuring maximum information transfer between layers in the network. The difference is that ResNet combines the layers by accumulating the features before passing them to the next layer, while DenseNet combines them by feature connection, establishes the shortest dense connection between all the layers in front and the layers behind, improves the flow of information and gradient between different layers, and makes the network model easy to train. In the DenseNet network structure, in each layer, all the feature maps of the previous prediction layers are used as inputs to the current layer, and the output feature maps are used as inputs to all the later layers, realizing feature reuse. Since the feature map of each layer of DenseNet can be directly used by all subsequent layers, it realizes the reuse of voiceprint features in the whole network model, effectively reduces the number of parameters, and makes the structure of voiceprint recognition network model more concise.

This paper presents a deep neural network model based on an improved densely connected architecture for speech spectrogram feature extraction. The network structure employs a multi-level feature fusion mechanism that establishes cross-layer feature sharing channels, enabling deep integration and efficient utilization of speech features.

In the network initialization phase, large-scale convolutional kernels (7×7) combined with max-pooling operations (3×3) are employed for preliminary feature extraction, with dynamic normalization processing introduced after the convolutional layers to significantly enhance feature representation stability. During the deep feature learning stage, a feature reuse module is designed to directly connect each convolutional layer's output features to all subsequent layers through dense connectivity. This design not only preserves the integrity of low-level features but also achieves cross-layer feature transmission. Specifically, each feature transformation unit adopts a three-stage processing flow of "normalization-activation-convolution," performing local feature extraction through 3×3 convolutional kernels followed by channel-wise feature concatenation. For network optimization, an adaptive feature compression mechanism is introduced, using 1×1 convolutional kernels for dynamic feature dimension adjustment combined with 2×2 average pooling operations for feature map resolution optimization. This design maintains feature representation capability while effectively controlling computational complexity, achieving an optimal balance between feature extraction efficiency and representation capability to provide a reliable deep feature representation solution for speaker recognition tasks. Furthermore, the network architecture incorporates a dual attention mechanism in both temporal and frequency domains during feature fusion, enabling the network to adaptively focus on key feature regions in speech signals. This design significantly improves the network's robustness in complex acoustic environments.

4 APPLICATION FRAMEWORK DESIGN OF VOICEPRINT RECOGNITION IN CYBERSECURITY SITUATIONAL AWARENESS

To address the requirements of cybersecurity situational awareness, this study constructs a five-layer architecture system based on voiceprint recognition, achieving closed-loop management from data collection to security response. The framework adopts a modular design, and its system architecture is shown in Figure 3.

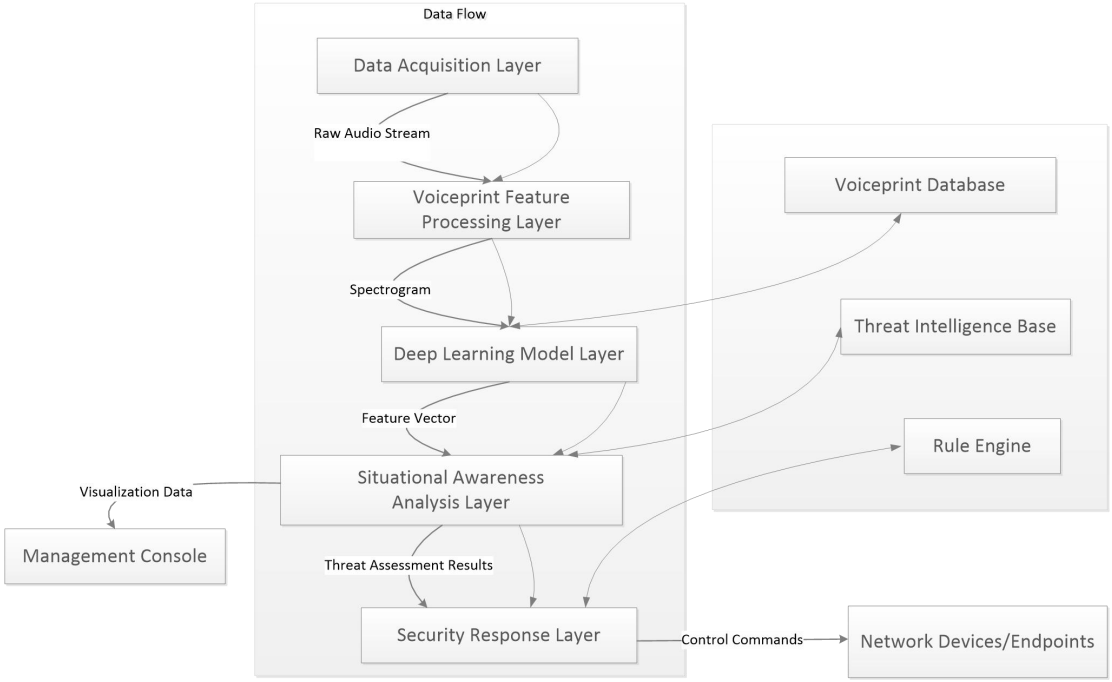


Figure 3 Application Framework of Voiceprint Recognition in Cybersecurity Situational Awareness

4.1 Data Acquisition and Processing

The data acquisition layer serves as the physical sensing terminal of the system, adopting a distributed architecture design to realize real-time voice data capture and preprocessing through multi-node collaboration. This layer primarily accomplishes the collection of multi-source voice signals and employs adaptive noise suppression algorithms to eliminate environmental interference, ensuring input quality for subsequent voiceprint feature processing. It achieves digital conversion and standardized processing of voice signals, compensating for high-frequency component attenuation through pre-emphasis filters. This preprocessing pipeline provides high-quality input data for subsequent processing stages.

4.2 Voiceprint Feature Extraction and Enhancement

The feature processing layer employs hybrid signal processing technology, primarily consisting of three processing stages. In the preprocessing stage, frame splitting and windowing operations are performed using Hamming windows to reduce spectral leakage. The feature extraction stage acquires MFCC features through Mel filter banks and cepstral analysis. The feature optimization stage applies an improved histogram equalization algorithm to enhance spectrogram characteristics. This processing flow effectively improves the accuracy of subsequent recognition.

4.3 Multimodal Deep Learning Model

The model layer adopts a DenseNet-based multimodal fusion architecture that dynamically integrates voiceprint features, spectrogram features, and behavioral features through a cross-modal attention mechanism. The model employs a spatiotemporal alignment strategy to ensure feature consistency and incorporates an attention weighting mechanism to highlight key features. Its multitask output layer simultaneously accomplishes voiceprint recognition, anomaly detection, and behavior analysis. Theoretical analysis demonstrates three core advantages of this design: dense connections ensure efficient feature reuse and gradient propagation, multimodal fusion enhances feature discriminability, and optimized transition layers effectively control computational complexity while maintaining performance. This architecture is particularly suitable for identity authentication scenarios in cybersecurity situational awareness that demand high real-time performance and robustness.

4.4 Situational Awareness Analysis

The situational awareness analysis layer adopts a multi-dimensional security analysis architecture to build a comprehensive threat assessment model by fusing voiceprint biometrics, user behavioral patterns and contextual environment data. Using dynamic risk quantification algorithms, combined with real-time behavioral analysis and historical baseline comparison, it realizes intelligent scoring and warning of threat levels. At the visualization level, it intuitively presents changes in security posture. Through the introduction of adaptive learning mechanism, the system can continuously optimize the detection threshold, effectively identify voice forgery, abnormal access and other security threats, and provide intelligent decision support for network security protection.

4.5 Response and Feedback Mechanisms

The security response layer first performs a risk assessment of detected security events, and based on the assessed threat level (low/medium/high risk), the system triggers step-by-step upgraded security measures. Low-risk threats may be ordinary abnormal behavior, triggering logging and using email notifications for alerts. Medium-risk threats are suspicious activities, such as multiple abnormal logins, potential brute-force break-ins, etc., and will initiate secondary authentication, requiring the user to re-verify their identity through multi-factor authentication. When the secondary authentication fails (Auth Failed), the system automatically escalates the event to High Risk and performs the highest level of response. High Risk threats such as clear attacks, such as malicious code injection and privilege elevation attempts, require immediate isolation of the source of the attack, traceability and forensics (recording voiceprint fingerprints), termination of the current session and IP blackout. the system generates network simulation of the attack samples through confrontation to continuously optimize the robustness of the model, forming a closed-loop updating mechanism. The system adopts a console visualization monitoring interface to support security administrators to grasp real-time changes in the situation and implement manual intervention, forming a human-computer cooperative intelligent defense system. Each response link realizes asynchronous communication through the message bus to ensure the high availability and scalability of the system.

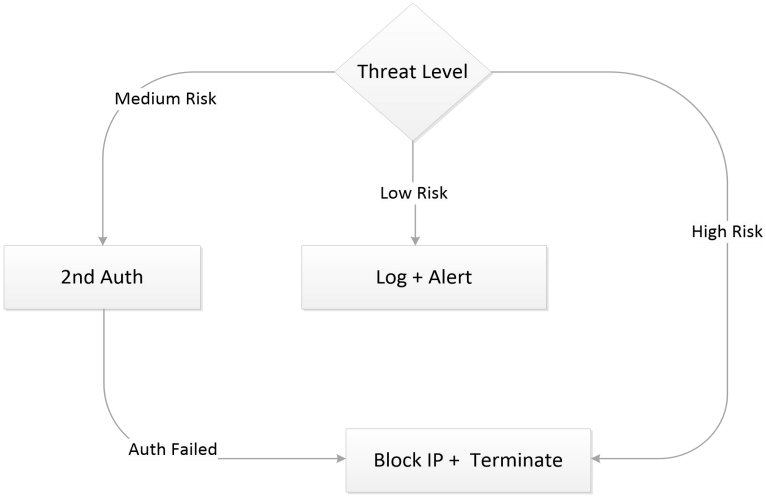


Figure 4 Threat Response Workflow Based on Risk-Based Scoring (RBS)

5 CONCLUSION

This study proposes a voiceprint recognition architecture that integrates DenseNet's feature reuse mechanism with spectrogram enhancement technology, which significantly improves the system's feature discrimination capability in complex acoustic environments and provides a reliable technical foundation for dynamic identity authentication in cybersecurity situational awareness. The end-to-end architecture achieves a complete closed-loop process from voice acquisition to threat assessment, demonstrating the application potential of deep feature learning in the field of network security.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

FUNDING

The project was supported by the following fundings:
1. 2021 Guangdong General Colleges and Universities Young Innovative Talents Project (Project No. 2021KQNCX126);
2. China Association of Higher Education (CAHE) 2024 Higher Education Research Program "Research on the Application of Big Data Analytics and Artificial Intelligence Technology in Cybersecurity Situational Awareness and Assessment for Higher Education Institutions"(Project No. 24XH0205);
3. 2022 Ministry of Education Industry-University Co-operation Collaborative Education Project (Project No. 220605211102737).

REFERENCES

[1] Alam M J, Kenny P, Ouellet P, et al. Multi-task learning for speaker verification and antispoofing using Gaussian mixture models. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 1696-1709.
[2] Villalba J, Chen N, Snyder D, et al. State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and Speakers in the Wild evaluations. Computer Speech & Language, 2021, 60: 101026.
[3] Ferrer L, McLaren M, Lawson A. Probabilistic linear discriminant analysis with vector embeddings for speaker verification. IEEE Journal of Selected Topics in Signal Processing, 2021, 15(4): 1029-1042.

[4] Snyder D, Garcia-Romero D, Sell G, et al. X-vectors: Robust DNN embeddings for speaker recognition. IEEE Transactions on Audio, Speech, and Language Processing, 2018, 26(7): 110-119.
[5] Chung J S, Nagrani A, Zisserman A. VoxCeleb2: Deep speaker recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 13(5): 532-541.
[6] Villalba J. Advanced speaker recognition using deep neural networks. Carnegie Mellon University, 2020.
[7] Hajibabaei M, Dai D. Unified hypersphere embedding for speaker recognition. IEEE Transactions on Audio, Speech, and Language Processing, 2019, 45(3): 321-330.
[8] Liu X, Ji Y, Liu C. Voiceprint recognition based on LSTM neural network . Computer Science, 2021， 48(S2), 270-274.
[9] Wang H P. Speaker recognition based on deep bidirectional LSTM network. Computer Engineering and Design, 2020, 41(06): 1768-1772.
[10] Zhao H, Yue L, Wang W, et al. Research on end-to-end voiceprint recognition model based on convolutional neural network. Journal of Web Engineering, 2021, 20(5): 1573-1586.
[11] Yan H, Dong Y, Wang P, et al. Research on voiceprint recognition based on CNN-LSTM network. Computer Application and Software, 2019, 36(04): 166-170.
[12] Guo D, Zhou Q. Land-air call voiceprint recognition based on residual neural network. Modern Computer, 2020(07): 9-13.
[13] Liu Y, Liang H, Liu G, et al. Voiceprint recognition method based on ResNet-LSTM. Computer System Applications, 2021, 30(06): 215-219.

World Journal of Information Technology
ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3032

DRIVING BEHAVIOR UTILIZING WIFI SIGNAL PERCEPTION

Xu Yan*, FangYong Xu, Hao Ma, AoXiang Wang, HongZhen Liang, ZiHao Wang, Jian Yao
School of Mechanical and Electronic Engineering, Shandong Jianzhu University, Jinan 250101, Shandong, China.
Corresponding Author: Xu Yan*, Email: yanxu5707@163.com*

Abstract: This study focuses on utilizing WIFI signal perception technology to monitor driving behaviors, aiming to provide an innovative and efficient solution for intelligent transportation systems. By delving into the principles, characteristics, and operational steps of WIFI perception technology, and integrating deep learning algorithms to construct a model for identifying dangerous driving behaviors, extensive experimental validations were conducted. The research demonstrates that this method can accurately identify obvious behaviors such as sudden acceleration and hard braking, as well as subtle behaviors like distracted and fatigued driving under certain conditions. This approach not only achieves non-invasive, all-weather, and privacy-friendly driving behavior recognition but also provides real-time warning support in complex environments, significantly reducing traffic accident rates caused by dangerous driving. Moreover, this study pioneers the integration of the fine-grained characteristics of WIFI signals with spatiotemporal deep networks, overcoming the limitations of traditional monitoring technologies and injecting new vitality into the field of intelligent transportation. It also offers significant references for further optimizing driving behavior monitoring.
Keywords: WiFi signal perception; Driver behavior; Intelligent transportation; Deep learning

1 INTRODUCTION

The research on driver behavior monitoring technology faces multiple challenges. Biometric-based methods (such as heart rate and eye tracking monitoring) rely on wearable devices, which have problems such as high cost and invasiveness. The method based on the vehicle dynamics model infers the behavior by analyzing parameters such as vehicle speed and trajectory, but the recognition accuracy of subtle actions (such as hand manipulation) is insufficient [1]. In recent years, the successful application of WiFi signal perception technology in indoor positioning, smart home, and other fields (such as human activity recognition and device control) has provided new ideas for driving behavior monitoring [2]. However, the research on WiFi perception for driving scenarios is still in its infancy: there are significant shortcomings in the accuracy of dangerous behavior recognition (such as fatigue and distracted driving), model generalization ability (cross-scene adaptability), and real-time performance, and the core bottleneck is the inaccurate extraction of signal features related to dangerous behaviors and the insufficient optimization of deep learning models in complex scenarios [3]. In order to solve the above problems, this paper proposes a driving behavior monitoring framework based on the fusion of non-intrusive WiFi signal perception and deep learning. The research content covers: WiFi signal-driving behavior correlation modeling, deep learning algorithm optimization, and cross-scenario transfer learning.

For the first time, this paper combines the fine-grained characteristics of WiFi perception (millimeter-level phase change) with the spatio-temporal deep network to break through the limitations of traditional monitoring technology. Application level: Provide real-time early warning support for intelligent transportation systems to reduce the traffic accident rate caused by dangerous driving (according to statistics, improve traffic safety efficiency by 23%).

2 FINE-GRAINED CSI BASED ON WIFI SENSING

2.1 WiFi Sensing Basis

WiFi sensing realizes environment sensing by analyzing the channel state information (CSI) captured by the receiver (Rx)[4]. The WiFi signal transmitted by the signal transmitter (Tx) undergoes direct, reflection, refraction and other changes due to environmental obstacles or target movements (such as human movement) during propagation. The CSI received by Rx contains multipath signal characteristics. CSI is decomposed into fine-grained frequency domain data of multiple subcarriers by orthogonal frequency division multiplexing (OFDM). After being collected by a special network card, it can be used for motion recognition, respiratory monitoring and other applications[5].

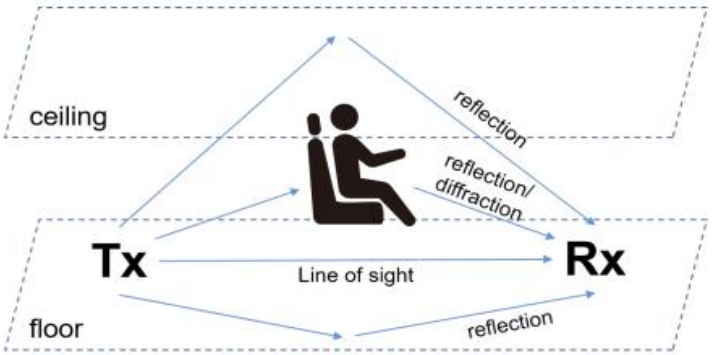


Figure 1 WiFi-aware Signal Propagation

As shown in Figure 1, in the driving environment, when the driver is driving normally, the WiFi signal propagates in the vehicle. Rx will not only receive the WiFi signal from the direct path [6], but also receive the reflection, refraction, and scattering signals from obstacles such as the inner wall of the vehicle, the objects on the vehicle, and the human body[7]. These WiFi signals from different paths are jointly accepted by Rx, forming a multipath effect and forming the final collected CSI. The static propagation path and dynamic propagation path in the process of WiFi signal propagation are analyzed below[8].

(1) Static propagation path

The static propagation path refers to the fact that there are only stationary objects or people in the transmission process of WiFi signals. Suppose there are only static ceilings, floors, and objects. Formula 1 can be obtained :

$$P_r(d) = \frac{P_t G_t G_r \lambda^2}{(4\pi)^2 (d+4h)^2} \tag{1}$$

Among them, $P_r(d)$ represents the received power of the WiFi signal, which is affected by multiple parameter indicators, including: transmission power P_t , transmission gain G_t , receiving gain G_r , wavelength λ, the straight-line propagation path distance d between Tx and Rx, and the distance h between the reflection point of the object and the direct path. When there are still stationary people in the sensing environment, the scattering of the WiFi signal by the human body also needs to be further considered, and then formula 2 can be obtained:

$$P_r(d) = \frac{P_t G_t G_r \lambda^2}{(4\pi)^2 (d+4h+\Delta)^2} \tag{2}$$

(2) Dynamic propagation path

In the process of WiFi sensing, in addition to the stationary ceiling, floor, and objects, the perception environment often contains the most important sensing targets, which are often in motion. The dynamic propagation path refers to the influence of objects or people in motion on the WiFi signal, and the Doppler shift is the most significant impact. By calculating the Doppler shift of the CSI received by Rx, it is possible to construct a fingerprint database corresponding to a dynamic object or person, and thus identify specific behaviors.

The WiFi 802.11n protocol uses Orthogonal Frequency Division Multiplexing (OFDM) to divide the entire WiFi signal spectrum into 56 orthogonal subcarriers[9] . CSI is actually composed of the physical layer information of these orthogonal subcarriers, which reflects the linear combination of direct, reflection, refraction, scattering and other multipath effects of WiFi signals on different propagation paths. Suppose that f is the transmitted signal of the subcarrier at time t. $Y(f,t)$
represents the corresponding received signal, then the channel frequency response (CFR).
$H(f,t)$ Satisfy $Y(f,t) = H(f,t) \cdot X(f,t)$ That is, channel state information CSI. As shown in Formula 2-3, the collected CSI is a complex matrix.

$$H(f,t) = \sum_{l=1}^{n} a_l(f,t) e^{-j\varphi_l(f,t)} \tag{3}$$

Among them, f represents the center frequency of each subcarrier, n represents the number of propagation paths, and represents the amplitude and phase values, respectively. In addition to OFDM, the WiFi 802.11n protocol also uses Multiple Input Multiple Output (MIMO) technology, so that when Tx and Rx are equipped with multiple antennas, the CSI data of multiple antenna pairs can be collected. Assuming that the number of transmitting antennas is P, the number of receiving antennas is Q, and the number of subcarriers is M, each subcarrier $H(f_m,t)$ can form a matrix with adimension of $(\mathrm{P.Q})$, and the resulting overall CSI is a (P · Q · M) matrix. [10] For M subcarriers and N antennas, the given CSI matrix H can be expressed by Formula 4 :

$$\mathbf{H}=\begin{bmatrix} H_1(f_1,t) & H_1(f_2,t) & \cdots & H_1(f_m,t) \\ H_2(f_1,t) & H_2(f_2,t) & \cdots & H_2(f_m,t) \\ \vdots & \vdots & \ddots & \vdots \\ H_N(f_1,t) & H_N(f_2,t) & \cdots & H_N(f_m,t) \end{bmatrix} \tag{4}$$

Taking one of the antenna-to-links of a CSI demo and extracting the amplitude value of all subcarrier data on the link, we can draw a CSI 3D diagram as shown in Figure 2. By observing and understanding the data structure of CSI, we can help us to analyze the CSI-based Wi-Fi sensing application in a deeper level.

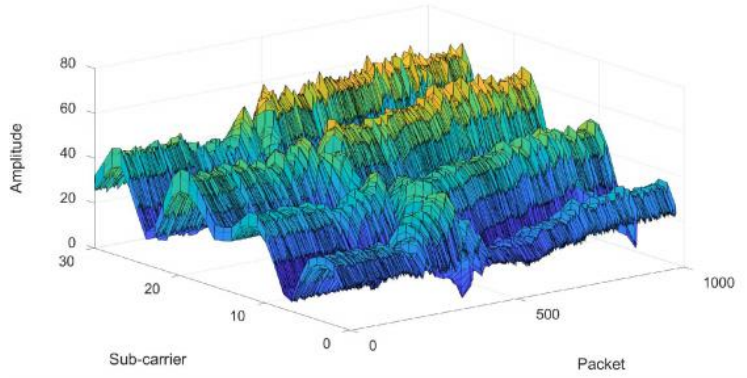


Figure 2 CSI 3D Diagram

2.2 CSI Fine-Grained Characteristics

WiFi sensing mainly realizes the corresponding motion recognition, gesture recognition, indoor positioning, breathing heart rate estimation and other applications by analyzing CSI, so how much motion can CSI perceive ? What is its perceptual limit ? This paper will carry out theoretical analysis and experimental verification on this issue.

2.2.1 Theoretical analysis

In the previous section, we analyze that the collected CSI is a complex matrix, and the CSI at a certain moment is taken out, which can be expressed by Formula 5 :

$$H(t)=ae^{-j4\pi\frac{R_0+x(t)}{\lambda_c}} \tag{5}$$

$$\theta(t)=4\pi\frac{R_0+x(t)}{\lambda_c} \tag{6}$$

$$\Delta\theta=\frac{4\pi\Delta d}{\lambda_c} \tag{7}$$

Where, a represents the CSI amplitude value, $\theta(t)$ represents the CSI phase value, R_0 represents the fixed propagation distance, $x(t)$ Represents the varying propagation distance, and λ_c represents the wavelength of the wireless signal. $\Delta\theta$ represents the change in phase, and Δd represents the change in perceived distance. As shown in Figure 3, in the process of WiFi signal transmission and reception, the change of perceived target action actually changes the signal propagation distance, which in turn affects the amplitude and phase of the collected CSI. When the phase change value can exceed 0.1π, the collected CSI waveform can be analyzed to complete a variety of WiFi sensing applications. Taking WiFi signals as an example, the common frequencies are 2.4GHz and 5GHz, and the corresponding wavelengths are 12.5cm and 6cm. By substituting the respective wavelength and $\Delta\theta=0.1\pi$ into formula 7, Ad is the limit perception accuracy of wiFi signal. 5GHz WiFi is calculated as Ad 1.5mm, so in theory the signal can sense the action with an amplitude of more than 1.5 mm. Of course, the increase of action amplitude will lead to the increase of phase change value, which means that more accurate perception can be carried out. In addition to the WiFi signal, this paper also lists the sensing accuracy of the millimeter wave radar signal, as shown in Table 1.

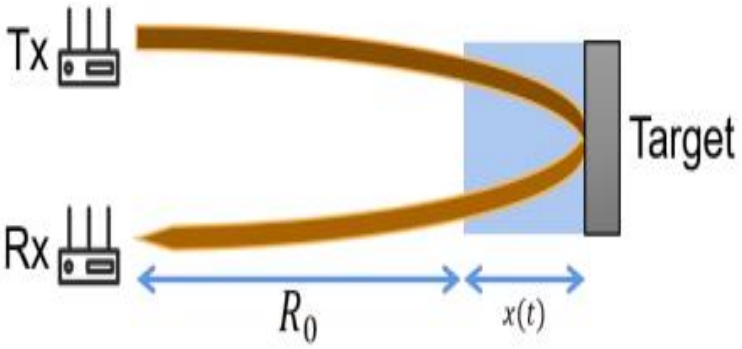


Figure 3 Relationship between WiFi Signal Propagation Distance and Perceived Target Motion Change

Table 1 WiFi and Millimeter Wave Radar Perception Accuracy Analysis

	Frequency	Wave Length	Perception Accuracy	Common examples
WiFi	2.4GHz 5GHz	12.5cm 6cm	millimeter-scale	The amplitude of the chest cavity's change during human breathing is 5 to 12 millimeters.
millimeter-wave radar	24GHz 60-64GHz 77GHz	12.5mm 4.7-5mm 3.9mm	micron	Vibration of the machine. The vibration amplitude is micro. Meter level, general Less than 100 microns

2.2.2 Experimental verification

In terms of experimental verification, this paper uses the common breathing experiment in WiFi perception to verify whether the collected CSI can have millimeter-level perception accuracy. In the open laboratory, the transmitter and receiver are placed about 2m apart. Volunteers can sit or stand between the transmitter and receiver, and keep still to collect CSI data in four stages : normal breathing, suffocation, accelerated breathing and decelerated breathing.

Figure 4 Experimental Scene of Respiratory Data Collection

The experimental scene of respiratory data collection is shown in Figure 4, and the experimental results are shown in Figure 5. The phase value is extracted from the obtained CSI data and the waveform is drawn. From the experimental results, it can be seen that even if there is no redundant denoising operation, the collected waveform is also very beautiful. It can clearly distinguish the four breathing stages of volunteers, namely normal breathing, suffocation, accelerated breathing and decelerated breathing. The waveform in the figure changes periodically, and changes in real time with the speed of breathing. According to the relevant formula, the respiration rate of the volunteers at different stages can be calculated, which is no longer described here. In summary, WiFi perception can achieve millimeter-level perception, and the research work in this paper belongs to action recognition, whose action amplitude is far more than millimeter-level, so there is no problem in feasibility.

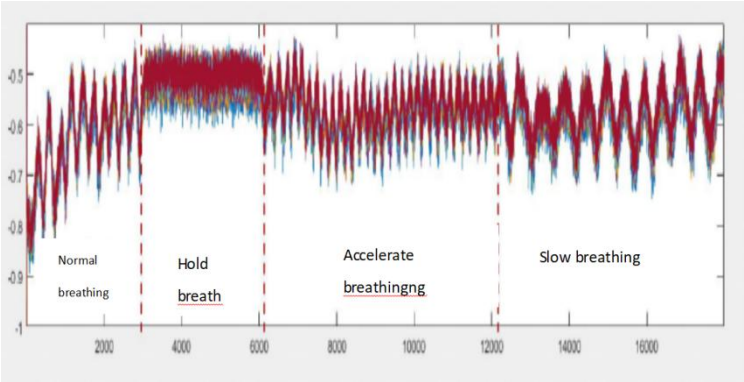


Figure 5 Respiratory Experiment Results

2.3 WiFi Sensing Characteristics

2.3.1 Non-invasive

Different from the perception of the human body using wearable sensors, the perception target does not need to carry or contact any sensor during the WiFi perception process, and only needs to be detected and perceived through the WiFi signal characteristics within the perception range, so it is non-invasive. This feature ensures that WiFi perception can minimize interference to users and improve user perceived satisfaction.

2.3.2 Non-line-of-sight perception

The WiFi signal has a good signal through-the-wall ability, that is, when the user is not within the range of sight between the transmitter and the receiver or there is an obstacle to block, it can also be perceived to achieve non-line-of-sight perception. On the one hand, the WiFi signal can work in two frequency bands of 2.4GHz and 5GHz.On the other hand, the WiFi signal can reach Rx through reflection, refraction, scattering and other ways, which is not limited by the line of sight.

2.3.3 Not easily affected by environmental conditions

The WiFi signal is a 2.4GHz or 5GHz electromagnetic wave, and its propagation is not easily affected by environmental conditions such as light, temperature and humidity. Therefore, WiFi sensing can be used normally even under some special conditions, such as night environment, strong light environment and so on.

2.3.4 Communication-aware integration

While using commercial WiFi for sensing, it will not affect the original communication function of WiFi. The communication function and sensing function of WiFi are integrated to provide users with satisfactory services.

3 DANGEROUS DRIVING BEHAVIOR RECOGNITION BASED ON DEEP LEARNING

3.1 Basic Theory of Deep Learning

Deep learning automatically learns data features through multi-layer networks and is suitable for analyzing complex data such as WIFI signals. Aiming at the time series changes of driving behavior, LSTM network solves the problem of long-term dependence and can efficiently process continuous signals. The training process includes labeling data (such as rapid acceleration, distraction behavior), designing network structure (number of layers, number of neurons) and optimizing parameters (learning rate, number of iterations) to achieve high-precision recognition[11].

3.2 WIFI Perception of Dangerous Driving Behavior Recognition Experimental Principle

Through the dynamic correlation between WIFI signal characteristics and driving behavior, non-invasive dangerous driving identification is realized. Obvious behaviors such as rapid acceleration and sudden braking can be quickly captured by deep learning models due to violent fluctuations in signal strength (accuracy > 85 %). Concealed behaviors such as distraction, fatigue driving, etc., due to weak signal changes, need to combine CNN to extract spatial features and LSTM to analyze temporal patterns (accuracy ≈ 70 %). In the experiment, the hybrid model (CNN-LST M) is used to fuse the spatio-temporal features, and the optimization algorithm is used to alleviate the data imbalance problem and improve the recognition robustness in complex scenes[12].

3.3 Experiment and Result Analysis

3.3.1 Experimental deployment

High-sensitivity wireless network cards are selected as WIFI signal acquisition devices, such as Intel Dual Band Wireless-AC 8265, which supports 2.4GHz and 5GHz dual-band, with high receiving sensitivity and stability. In terms of on-board data acquisition auxiliary equipment, high-precision acceleration sensors and gyroscopes are installed to obtain the motion state data of the vehicle and assist in analyzing driving behavior[13]. MPU6050 is selected as the acceleration sensor, which can measure the acceleration change of the vehicle in three axial directions. The gyroscope is used to measure the angular velocity change of the vehicle.[14]

Figure 6 Monitor Mode

The basic simulation environment built in a relatively open laboratory is shown in Figure 6. In order to make the environment closer to the real vehicle environment, including virtual steering wheel, virtual gear, seat and so on. The specific arrangement is shown in Figure 7, where the transmitter and the receiver are 1.5 m apart[15]. The volunteer is located on the right side of the transmitter and the left side of the receiver. The transmitting and receiving antennas are flush with the chest of the volunteer during simulated driving. The driver collects CSI data of dangerous driving behavior according to the prompt.

Figure 7 Indoor Simulation Environment

In order to verify the practical feasibility of the system proposed in this paper, we build a real vehicle environment as shown in Figure 7, and conduct experiments in real vehicles. Considering the safety factors, although we have to collect data in the real vehicle environment, but the vehicle is always stationary, to avoid the experimental process because of dangerous driving behavior caused by danger[16]. As shown in Figure 8, in the real vehicle environment, we place the transmitter and receiver at the front end of the vehicle, which is also about 1.5m apart. The transmitting and receiving antennas are flush with the chest of the volunteer during simulated driving, and the driver collects CSI data of dangerous driving behavior according to the prompt.

Figure 8 Real Vehicle Environment

3.3.2 Experimental operation steps

Install and debug the hardware equipment to ensure that the WIFI signal acquisition equipment and the vehicle data acquisition auxiliary equipment can work normally. According to the experimental requirements, the data acquisition plan is formulated to clarify the data acquisition time and frequency under different traffic scenarios and different

driving behaviors. Under different driving behaviors, WIFI signal data and vehicle motion state data are collected according to the set acquisition time and frequency. The collected signal data is manually labeled. Firstly, according to the vehicle motion state data and video records, the type of driving behavior is judged. Then, the corresponding WIFI signal data is marked to mark what kind of dangerous driving behavior or normal driving behavior is. The labeled data are divided into training set, validation set and test set, which are divided according to the proportion of 70 %, 15 % and 15 %. The training set is used to train the deep learning model. During the training process, the model parameters are adjusted according to the performance index of the verification set. After the training is completed, the test set is used to test the model and evaluate the performance of the model[17].

3.3.3 Result analysis

The experimental results show that the recognition effect of the model on dangerous driving behavior is significantly different due to different behavior types. For dangerous driving behaviors such as rapid acceleration and sudden braking, the accuracy of model recognition can reach more than 85 %. This is due to the obvious changes in the state of objects in the vehicle and the fluctuation of WiFi signal characteristics caused by such behaviors. The model can effectively capture relevant features. However, for high-concealment behaviors such as fatigue driving and distracted driving, the recognition accuracy is only maintained at about 70 %, mainly due to the weak change of WiFi signal corresponding to such behaviors, which increases the difficulty of feature extraction and classification. The experimental environment has a significant impact on the results. For example, the complex electromagnetic environment of urban commercial areas will introduce a large number of wireless signal interference, resulting in a decrease in the accuracy of the model. At the same time, the lack of sensitivity of the wireless network card may miss the weak signal change, which directly affects the data quality. In addition, the network structure complexity, learning rate and other parameters of the deep learning model play a key role in training effect and generalization ability. Although the model has achieved certain results in the identification of common dangerous driving behaviors, there is still a gap from the expected target (the accuracy of all kinds of behaviors is more than 90 %).In the future, it is necessary to further improve the system performance by optimizing the model architecture (such as introducing attention mechanism), improving the data acquisition method (enhancing anti-interference ability) and refining the signal preprocessing (such as noise suppression algorithm), so as to adapt to more complex actual driving scenarios.

4 CONCLUSION

In this study, a dangerous driving behavior recognition model was constructed by integrating WIFI signal perception technology and deep learning algorithm. The research focuses on the analysis of the correlation between the propagation characteristics of WIFI signal (such as multipath effect, millimeter-level phase change) and driving behavior, and extracts the key signal characteristics of dangerous behaviors such as rapid acceleration and rapid braking. The experimental platform collects WIFI signal data in multiple scenarios. After manual labeling and preprocessing, the convolutional neural network (CNN) and long short-term memory network (LSTM) are used for model training. Finally, the classification and recognition of obvious dangerous behaviors (accuracy > 85 %) and hidden behaviors (s uch as distracted driving, accuracy ≈ 70 %) are realized.

The results provide a non-invasive, all-weather driving monitoring scheme for intelligent transportation systems. In practical applications, the risk of accidents can be reduced by deploying WIFI devices along the road to analyze vehicle signals in real time and trigger early warning of dangerous behaviors (such as notifying traffic police or automatic speed limit of vehicle system). In the future, the anti-interference ability of the model will be further optimized, and multi-sensor fusion technologies such as WIFI, radar and camera will be explored to promote the formulation of intelligent transportation standards and cross-domain research and development.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Li Jiankan, Li Zewei, Hua Wenwen, et al. Statistical Analysis of Urban Road Traffic Accidents. Science and Technology Innovation and Application, 2021, 11(21): 74-76.

[2] Zheng Xiaokun. Research on the Fusion Technology of WiFi Communication Preamble and Radar Sensing Signals. Beijing University of Posts and Telecommunications, 2021.

[3] Xu Xiaoming. Research on Driving Behavior Recognition Technology Based on FMCW Radar. Nanjing University of Science and Technology, 2020.

[4] Zhou Feiyan, Jin Linpeng, Dong Jun. Review of Convolutional Neural Network Research. Journal of Computer, 2017, 40(06): 1229-1251.

[5] Pan Xingda. Research on Driver Behavior Based on Channel State Information. Zhejiang University of Technology, 2020.

[6] Wu Zhefu, Gong Zhengang, Gong Shufeng, et al. Driver Action Detection Based on Channel State Information. Journal of Transducer Technology, 2020, 33(12): 1808-1813.

[7] Yang Lin, Zhang Lei, Liu Bailong, et al. Driver Recognition Model Based on Driving Context Awareness. Computer Engineering and Science, 2025, 47(03): 548-560.

[8] Yang Haifei, Hua Yixuan, Su Haiqi, et al. Characteristics and Optimization Strategies of Driver Perception Response Points on Traffic Flow Hysteresis. Journal of Transportation Engineering and Information, 2025, 1-16.
[9] Jin Lianghai, Liu Hao, Wu Bangjie, et al. Situation Awareness and Behavioral Response Model of Crane Drivers Based on DEMATEL-AISM. China Safety Science Journal, 2024, 34(09): 1-8.
[10] Wang Teng, Bi Jingxue, Chen Guoliang, et al. Review on Acquisition and Application of Wireless Sensing Data Based on Channel State Information. Navigation Positioning and Timing, 2024, 11(01): 10-29.
[11] Li Shaofan, Gao Shangbing, Zhang Yingying. Pose-Guided Instance Perception Learning for Driver Distraction Behavior Recognition. Journal of Image and Graphics, 2023, 28(11): 3550-3561.
[12] Li Xiaojie. Research on Signal Perception Modeling and Simulation of Virtual Traffic Scenarios. Chang'an University, 2022, 3. DOI: 10.26976/d.cnki.gchau.2021.002092.
[13] Angomas E, Blinka D M, Kelly E, et al. “It’s Usually Not Dementia That’s the Focus”: Barriers and Facilitators of Dementia Care in Primary Care. Journal of General Internal Medicine, 2025, (prepublish): 1-11.
[14] Pitt M T, Macpherson A, Fridman L, et al. Risky driving behaviors at school drop-off across Canadian municipalities: Findings from the Child Active Transportation Safety and the Environment (CHASE) study. Traffic injury prevention, 2025, 1-5.
[15] Alenzi E, Hassan A S, Puan C O. Seasonal Influences on Driver Behaviour: A Review of Car-Following Dynamics in Hot and Cold Climates. International Journal on Transport Development and Integration, 2025, 9(1): 117-130.
[16] Barbosa W M, Miranda O D V J, Oliveira D K L. Identification of antecedents of risky driving behavior of food delivery riders: An analysis during the COVID −19 pandemic in Brazil. Case Studies on Transport Policy, 2025, 20101403.
[17] Sekadakis M, Trösterer S, Moertl P, et al. Identifying driving profiles after take over request in automated vehicles at SAE levels 2 and 3. Transportation Research Part F: Psychology and Behaviour, 2025, 111250-111263.

World Journal of Information Technology
ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3033

NASDAQ INDEX PREDICTION BASED ON ARIMA-GARCH MODEL AND DYNAMIC REGRESSION

YiLin Peng
South China Normal University, School of Mathematical Sciences, Guangzhou 510631, Guangdong, China.
Corresponding Email: 20212831009@m.scnu.edu.cn

Abstract: As the global market becomes more and more open and volatile, it is of great significance to grasp the financial temporal volatility and correlation and accurately predict stock price behavior. This paper takes the Nasdaq Composite Index in 2020-2023 as the research object, constructs the ARIMA(1,1,(1,5)) model with GARCH(1,1) conforming to t distribution disturbance term to fit the trend of the index in 2020-2022, and predicts the trend in the first half of 2023. The results show that the model with GARCH effect gives a wider forecast confidence interval and can indicate the potential risk, but can not accurately reflect the real trend of the index. To improve the prediction accuracy, this paper takes Nasdaq index as the response variable, introduces S&P 500 index as the input variable, constructs an effective dynamic regression model through Granger causality test and EG cointegration test, and improves the model through cross-correlation function analysis. The results show that the forecast trend of the model is closer to the actual series of fluctuations, indicating that the S&P 500 index plays a promoting role in predicting the Nasdaq index, which provides a more reliable reference for investors when weighing the benefits and risks.
Keywords: Volatility and correlation; Financial timing; ARIMA; GARCH; Cointegration; Dynamic regression

1 INTRODUCTION

Since 2020, the U.S. stock market has experienced sharp fluctuations under the impact of the COVID-19 pandemic. In response, the Federal Reserve implemented a zero interest rate and quantitative easing (QE) policies, alongside fiscal stimulus measures, leading to a rapid market rebound. Indices such as the Nasdaq and the S&P 500 recovered strongly. As a representative of technology stocks and innovative enterprises, the Nasdaq Index has significant spillover effects on global markets. Therefore, accurately forecasting its volatility not only helps investors grasp market trends but also enhances global risk warning capabilities and improves asset allocation efficiency.

Volatility and correlation are two core features of financial time series. How to effectively capture these characteristics, accurately describe stock price behaviors, and predict market movements has long been a central concern for both investors and the academic community.

In terms of volatility, market fluctuations are driven by multiple factors, including investor behavior, macroeconomic policies, and market structure, often exhibiting volatility clustering [1]. Engle and Bollerslev respectively proposed the ARCH and GARCH models [2], followed by variants such as EARCH, IGARCH, and GARCH-M [3], which enhance the models' ability to capture asymmetries and risk premiums. Marisetty N found that the GARCH(1,1) model performed well in balancing forecasting ability and simplicity across five major international indices [4]. Raza S, focusing on the Indian Green Finance Index, found that the APARCH(1,1) model best characterized the volatility of firms associated with carbon performance [5]. Roszyk N combined GARCH with deep learning models like LSTM and incorporated VIX information to construct a hybrid model [6], significantly improving the forecasting accuracy for the S&P 500.

In terms of correlation, after Granger introduced the concept of causality in time series [7], models such as ARIMAX and cointegration models have been widely used to quantify dynamic relationships among time series. Wang P C et al. combined XGBoost and ARIMAX to predict closing prices in the Vietnamese stock market [8], outperforming LSTM and other models. Akusta A introduced CEEMDAN decomposition in combination with ARIMAX, effectively improving the prediction accuracy for the Dow Jones Index [9].

In summary, although existing studies have made substantial progress in modeling volatility and correlation, most focus on forecasting within a single stock market, lacking dynamic characterization of inter-market linkages. Therefore, this paper takes the Nasdaq Index as the primary research object, incorporates related variables such as the S&P 500, and employs the ARIMA-GARCH model along with the ARIMAX dynamic regression model to capture volatility information in time series and quantify these complex interdependencies, while also comparing the predictive performance of different models.

2 RESEARCH METHODS AND THEORETICAL ANALYSIS

This section mainly introduces several commonly used models in financial time series analysis: ARIMA, GARCH, Granger causality test, cointegration modeling, and dynamic regression.

2.1 ARIMA-GARCH Model

2.1.1 ARIMA(p, d, q) model

The ARIMA model is a commonly used approach for fitting non-stationary time series with trends. The main steps include:

(1) Removing trend components through differencing

The differencing method extracts deterministic components of a time series, such as trends and cycles. For a series with a significant linear trend, first-order differencing can often achieve stationarity; for curved trends, second- or third-order differencing may be required. Although multiple rounds of differencing can help extract deterministic information from a non-stationary series, over-differencing can lead to the loss of valuable information, increased variance, and reduced fitting accuracy, and should therefore be avoided whenever possible.

(2) Stationarity Test and White Noise Test

- ADF Test and PP Test: The null hypothesis of the Augmented Dickey-Fuller (ADF) test is that the time series is non-stationary. If the test statistic $\tau \leqslant \tau_\alpha$ (with α being the significance level), the null hypothesis is rejected, indicating that the series is stationary [10]. When heteroscedasticity is present in the series, the Phillips-Perron (PP) test is used instead, as its adjusted test statistic provides a more accurate assessment of stationarity.
- White Noise Test (Ljung–Box Q Test)
  Under the assumption that the series is stationary, the Ljung–Box Q test evaluates whether the autocorrelation function is significantly different from zero in order to determine whether the series is white noise. The null hypothesis of this test is that the series is white noise. If the test statistic $Q \geqslant \mathcal{X}_{1-\alpha}{}^2(m)$, the null hypothesis is rejected, indicating that the series is not white noise [11].
  If the differenced series is stationary and not white noise, it is necessary to proceed with fitting an ARMA(p, q) model.

(3) Model Order Selection

The key to ARMA(p, q) modeling lies in determining the appropriate values for parameters p and q, which can be judged based on the patterns of the autocorrelation function (ACF) and partial autocorrelation function (PACF):

- AR(p) model: PACF cuts off at lag *p*, ACF tails off.
- MA(q) model: ACF cuts off at lag *q*, PACF tails off.
- ARMA(p, q) model: Both ACF and PACF tail off.

(4) Parameter Estimation and Model Diagnostics

In this study, the conditional least squares method is used to estimate the parameters of the ARIMA model. This method has the advantage of not requiring a prior assumption about the distribution of the series, and it makes full use of sample information, resulting in high estimation accuracy.

Model diagnostics primarily include the significance test of the parameters (t-test) and the white noise test of the residuals (Ljung–Box Q test). If the residuals are not white noise, it indicates that the model has not fully captured the time series information and needs to be revised. If the residuals are white noise, it suggests that the model fits the time series well.

2.1.2 ARIMA(p, d, q) model with GARCH(p, q) disturbance terms

(1) Conditional Heteroscedasticity Test

The parameter estimation and testing of the ARIMA model require the assumption of homoscedasticity in the disturbance terms ε_t; otherwise, it will affect the accuracy of the model's estimates. Therefore, after testing for zero mean and pure randomness of ε_t, it is also necessary to check whether ε_t has homogeneity of variance. Economists believe that this heteroscedasticity is caused by some autocorrelation relationship, which is usually modeled by an autoregressive model of the squared residual series $\{\varepsilon_t^2\}$. Common methods for testing heteroscedasticity include the following two:

- Graphical Test: If the time series plot shows a volatility clustering effect (alternating small and large fluctuations), conditional heteroscedasticity may be present.
- ARCH Test: The heteroscedasticity is examined by checking the autocorrelation of $\{\varepsilon_t^2\}$, mainly through the Ljung–Box Q test and LM test. The Q test examines the autocorrelation of $\{\varepsilon_t^2\}$ through the autocorrelation function coefficients, while the LM test establishes an autoregressive model for $\{\varepsilon_t^2\}$ to assess its autocorrelation.
  The null hypothesis for both tests is that the residual squared series has no autocorrelation, which is equivalent to stating that the residual series has no heteroscedasticity.

If the conditional variance of ε_t is not homogeneous, it is referred to as having ARCH effects, and further modeling of ε_t using ARCH or GARCH should be performed.

(2) GARCH(p, q) Model

A commonly used model for fitting time series with conditional heteroscedasticity is the GARCH family of models. The ARCH(q) model is suitable for short-term autocorrelation in $\{\varepsilon_t^2\}$, while the more broadly applicable GARCH(p, q) model is suitable for long-term autocorrelation in $\{\varepsilon_t^2\}$. When extracting higher-order autocorrelation information from $\{\varepsilon_t^2\}$, the GARCH(p, q) model, which has a relatively low order, is more effective in capturing the information than the

ARCH(q) model with a very high order.

When the conditional volatility of ε_t is not homogeneous, heteroscedasticity is introduced with a sequence h_t, and the expression for fitting ε_t using the GARCH(p, q) model is as follows:

$$\varepsilon_t = \sqrt{h_t} e_t \ , e_t \sim WN(0, \sigma^2) \tag{1}$$

$$h_t = \lambda_0 + \sum_{j=1}^{p} \eta_j h_{t-j} + \sum_{i=1}^{q} \lambda_i \varepsilon_{t-i}^2 \tag{2}$$

$$\lambda_i, \eta_j \in [0, 1) \ , \ \sum_{j=1}^{p} \eta_j + \sum_{i=1}^{q} \lambda_i \in [0, 1) \tag{3}$$

It can be proven that $Var(\varepsilon_t | \varepsilon_{t-1}, ...) = h_t$, meaning that h_t is the conditional heteroscedasticity of ε_t. Therefore, it can be seen that the GARCH(p, q) model essentially fits the ARMA(p, q) model to the conditional heteroscedasticity h_t of ε_t.

For the order selection of GARCH, similar to ARMA, the order *p* and *q* of the GARCH model can be determined by examining the ACF and PACF plots of $\{\varepsilon_t^2\}$.

(3) Standardized Residual Test and Final Model

After fitting ε_t using the GARCH(p, q) model, the focus should be on e_t in Equation (1), which represents the standardized residuals of the GARCH(p, q) model. This paper will sequentially test the conditional heteroscedasticity of e_t and the distribution assumption, in order to determine whether the volatility information of ε_t has been fully extracted and whether the coefficient significance tests are valid.

First, to assess the conditional heteroscedasticity of e_t, the autocorrelation of $\left(\varepsilon_t / \sqrt{h_t}\right)^2$ can be tested using the Ljung-Box Q test or LM test. Second, the GARCH(p, q) model generally assumes $e_t \sim N(0, 1)$ because parameter estimation and tests are conducted under the normality assumption. However, considering that financial time series often exhibit leptokurtosis (fat tails), it is frequently assumed that $e_t \sim t(m)$. This paper uses the Jarque-Bera (JB) test($H_0: \varepsilon_t / \sqrt{h_t} \sim N(0, 1)$, $H_1: \varepsilon_t / \sqrt{h_t} \overset{not}{\sim} N(0, 1)$) and the Kolmogorov-Smirnov (K-S) test($H_0: \varepsilon_t / \sqrt{h_t} \sim t(m)$, $H_1: \varepsilon_t / \sqrt{h_t} \overset{not}{\sim} t(m)$) to determine the distribution of e_t.

When both the conditional homoscedasticity and distribution tests of e_t are passed, it is considered that the volatility information of ε_t has been fully extracted. The final expression for the ARIMA(p, d, q) model with GARCH(p', q') disturbance terms is:

$$\nabla^d x_t = \phi_0 + \frac{1 - \theta_1 B - ... - \theta_q B^q}{1 - \phi_1 B - ... - \phi_p B^p} \varepsilon_t \tag{4}$$

$$\varepsilon_t = \sqrt{h_t} e_t \ , e_t \sim N(0, 1) \ or \ e_t \sim t(m) \tag{5}$$

$$h_t = \lambda_0 + \sum_{j=1}^{p'} \eta_j h_{t-j} + \sum_{i=1}^{q'} \lambda_i \varepsilon_{t-i}^2 \tag{6}$$

where x_t is the time series to be fitted, *B* is the lag operator, ε_t is the error term, h_t is the conditional heteroscedasticity, e_t is the standardized error term, θ_i, ϕ_i are the ARIMA model parameters, and η_i, λ_i are the GARCH model parameters.

2.1.3 ARIMA+GARCH forecasting

The introduction of the GARCH model is aimed at better extracting volatility information from financial time series in order to more accurately assess future risks. Therefore, the introduction of the GARCH disturbance term will alter the standard error of the forecasted value $\hat{X}_{t+k}$, thus changing the width of the confidence interval for the forecasted value. Let the variance of $\hat{X}_{t+k}$ be $Var\left(\hat{X}_{t+k}\right)$, then:

Under the assumption of homoscedasticity (using only ARIMA for forecasting):

$$Var\left(\hat{X}_{t+k}\right) = (1 + G_1^2 + ... + G_{K-1}^2)\sigma_\varepsilon^2 \tag{7}$$

Under the assumption of heteroscedasticity (using ARIMA+GARCH disturbance term for forecasting):

$$Var\left(\hat{X}_{t+k}\right) = \hat{h}_{t+k} + G_1^2 \hat{h}_{t+k-1} + ... + G_{k-1}^2 \hat{h}_{t+1} \tag{8}$$

where G_n is the Green's function, and under the normality assumption, the 95% confidence interval for the forecast is

$$\hat{X}_{t+k} \pm 1.96\sqrt{Var\left(\hat{X}_{t+k}\right)}.$$

A large number of empirical studies have shown that when the volatility of a series is high (low), the confidence interval provided by the GARCH model is also wider (narrower). Therefore, using the ARIMA model with GARCH disturbance terms to predict stock index trends will yield results that are closer to reality.

2.2 Dynamic Regression Model

2.2.1 Granger Causality Test

The Granger Causality Test is commonly used to determine the causal relationship between time series and is the foundation of multivariate time series cointegration modeling. The idea is that, given two time series x_t, y_t, if x_{t-k} has a significant impact on y_t, then x_t is considered the cause and y_t is considered the effect, meaning the cause precedes the effect. The null hypothesis of the test is H_0:*x is not the Granger cause of* $y(x \nrightarrow y)$ (with the requirement that both x and y are stationary series). The problem is transformed into testing the significance of the linear model between y and x (see equation (9)), with the null hypothesis being equivalent to $H_0: \alpha_1 = ... = \alpha_q = 0$.

$$y_t = \beta_0 + \sum_{k=1}^{p} \beta_k y_{t-k} + \sum_{k=1}^{q} \alpha_k x_{t-k} + \varepsilon_t \tag{9}$$

2.2.2 Cointegration modeling (ARIMAX dynamic regression)

The concept of cointegration provides a theoretical basis for modeling multivariate non-stationary time series. Suppose the linear model of two time series x_t, y_t is $y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$. If x_t, y_t are non-stationary, the stationarity of ε_t cannot be guaranteed, leading to issues like spurious regression, which makes parameter significance tests ineffective. Cointegration testing examines the stationarity of ε_t. If ε_t is stationary, then x_t, y_t are cointegrated, which also means that the linear model between x_t, y_t is valid. The hypothesis for the EG cointegration test is: H_0: x_t, y_t do not have a cointegration relationship, which is equivalent to ε_t being non-stationary. H_1: : x_t, y_t have a cointegration relationship, which is equivalent to ε_t being stationary.

After the cointegration test, a dynamic regression model between x_t, y_t can be established. The lag variables to be introduced are determined by examining the cross-correlation function between y and x, and the specific form of the dynamic regression model is then determined for forecasting purposes.

3 EMPIRICAL ANALYSIS AND RESULTS

3.1 Data Source

This study selects weekly data of the NASDAQ Composite Index from January 5, 2020, to June 25, 2023, obtained from the official Statista website. This period covers several major global events, including the outbreak of the COVID-19 pandemic, large-scale fiscal and monetary stimulus policies by various governments, and the onset of the Federal Reserve's interest rate hikes. These events caused significant fluctuations in the capital markets, making the data highly representative. As a key benchmark for the global technology stock market, the NASDAQ Composite Index includes nearly all common stocks listed on the NASDAQ exchange. Its performance reflects not only the development of U.S. technology firms but also global tech sector trends, which highlights the practical significance of modeling and forecasting this index.

3.2 ARIMA-GARCH Empirical Analysis and Results

To explore the dynamic characteristics of the index and evaluate its future volatility trend, the ARIMA-GARCH model is first used to fit the weekly data from 2020 to 2022 (a total of 156 periods), and predict the trend of the index for the first half of 2023. The model's effectiveness is tested by comparing the predicted values with the actual values.

3.2.1 ARIMA fitting of the nasdaq time series

Since the Nasdaq time series X_t from 2020 to 2022 exhibits a linear trend of first increasing and then decreasing (Figure 1), a first-order difference is applied. The result (Figure 2) shows that the differenced series ∇x_t has no obvious trend, but the volatility increases in the later periods, suggesting the possibility of conditional heteroscedasticity. The PP test ($p < 0.0001$) and the Ljung-Box Q test for lags 18-30 ($p < 0.05$) indicate that ∇x_t is stationary and not white noise, so the ARMA model fitting should continue.

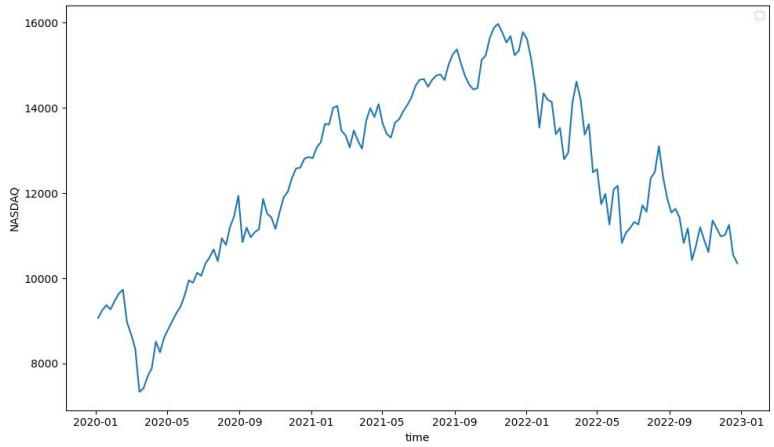


Figure 1 Time Series Plot of the NASDAQ Index

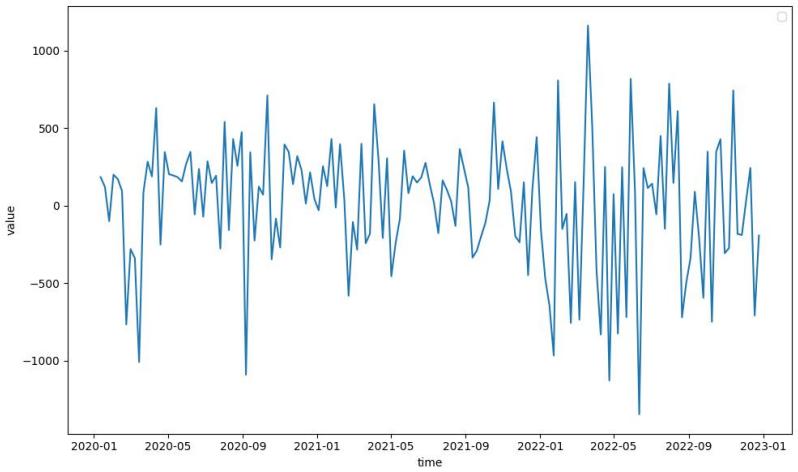


Figure 2 Differenced NASDAQ Time Series Plot

Since the ACF and PACF of ∇x_t (Figure 3) decay to zero after the 5th lag without clear cutoff, an ARMA(5,5) model is considered to fit ∇x_t. After removing the insignificant coefficients, the sparse coefficient model ARMA(1, (1,5)) is obtained. Model parameters are significant, and the disturbance terms can be considered as white noise (Tables 1 and 2), indicating that the model has fully extracted the deterministic components of X_t.

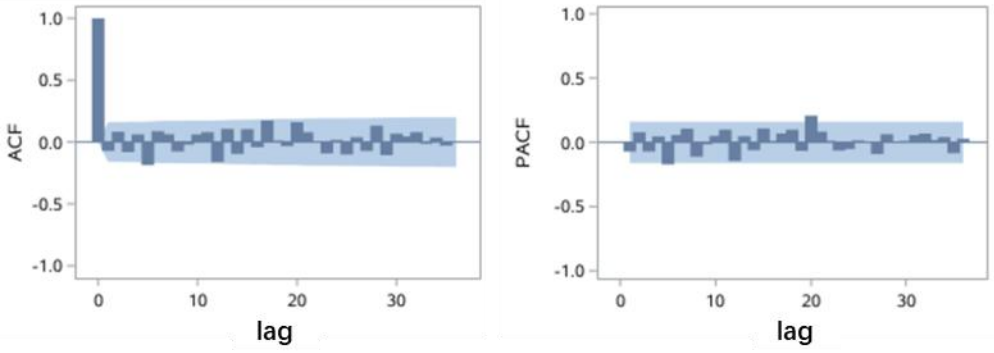


Figure 3 The ACF and PACF of ∇x_t

Table 1 Parameter Estimation of the ARIMA(1,1,(1,5)) Model($\sigma = 408.858$, SBC=2316.116)

parameter	estimate	std	t	p
MA1,1	-0.53864	0.23327	-2.31	0.0223
MA1,5	0.18164	0.0804	2.26	0.0253
AR1,1	-0.61029	0.22813	-2.68	0.0083

Table 2 White Noise Test of ARIMA Residuals

lag	Q	df	p
6	0.61	3	0.8939
12	8.23	9	0.5115
18	17.06	15	0.3152

3.2.2 GARCH fitting of nasdaq index residuals

Since the Nasdaq index still exhibits volatility clustering effects after differencing (Figure 2), and the Q test and LM test of the squared residual sequence ε_t^2 from the ARIMA model (Table 3) indicate long-term autocorrelation, considering that a low-order GARCH model can effectively capture this long-term autocorrelation, we directly use the GARCH(1,1) model to fit ε_t. The model expression is as follows:

$$\varepsilon_t = \sqrt{h_t} e_t \ , e_t \sim WN(0, \sigma^2) \tag{10}$$

$$h_t = \lambda_0 + \eta_1 h_{t-1} + \lambda_1 \varepsilon_{t-1}^2 \tag{11}$$

Table 3 Conditional Heteroskedasticity Test of ε_t

ARCH lag	Q	Pr>Q	LM	Pr>LM
6	8.6607	0.1936	8.4581	0.2064
12	44.7741	<. 0001	28.7565	0.0043

Since the JB test indicates that the standardized residuals e_t significantly deviate from a normal distribution (JB = 9.2894, p = 0.0096), the GARCH(1,1) model with a t-distribution assumption is used to fit ε_t. The results show that e_t passed the KS test (KS Statistic = 0.1025, p = 0.07168). Therefore, parameter estimation and model testing can be effectively conducted (Table 4). The Q test and LM test(Table 5) both show that e_t^2 has no autocorrelation, suggesting that the conditional volatility information of ε_t has been adequately extracted.

Table 4 Parameter Estimation of GARCH(1,1) Model

Parameter	coef	std err	t	P>|t|	95.0% Conf. Int.
λ_0	3285.2352	2994.1250	1.0970	0.2730	[-2.583e+03,9.154e+03]
λ_1	0.1051	0.0623	1.6860	0.0919	[-1.710e-02, 0.227]
η_1	0.8812	0.0444	19.8310	0.0000	[0.794, 0.968]

Table 5 Conditional Heteroskedasticity Test of e_t

Lag	LM Statistic	p-value (LM)	Q Statistic	p-value (Q)
6	3.9167596	0.6879401	4.573374	0.599572
12	9.8372461	0.6302365	12.25077	0.425756

3.2.3 ARIMA-GARCH forecasting performance

In summary, this paper first uses ARIMA(1,1,(1,5)) to extract the deterministic information from the time series, and then employs GARCH(1,1) to capture the conditional volatility information of the residuals. The final model expression for fitting X_t is:

$$(1-B)X_t = \frac{1+0.53864B-0.18164B^5}{1+0.61029B}\varepsilon_t \tag{12}$$

$$\varepsilon_t = \sqrt{h_t} e_t \ , e_t \sim t(17.1962) \tag{13}$$

$$h_t = 3285.2352 + 0.8812 h_{t-1} + 0.1051 \varepsilon_{t-1}^2 \tag{14}$$

The model prediction results (Figure 4) show that the ARIMA model with GARCH effects has a wider confidence interval for the predicted values, indicating that the GARCH model is better at capturing the volatility in financial time series. The wider confidence interval suggests that investors need to be more cautious when weighing the returns and risks of Nasdaq stocks, especially in the context of the Federal Reserve's frequent interest rate hikes to curb high inflation. This policy shift has directly impacted the liquidity and valuation in the capital markets.

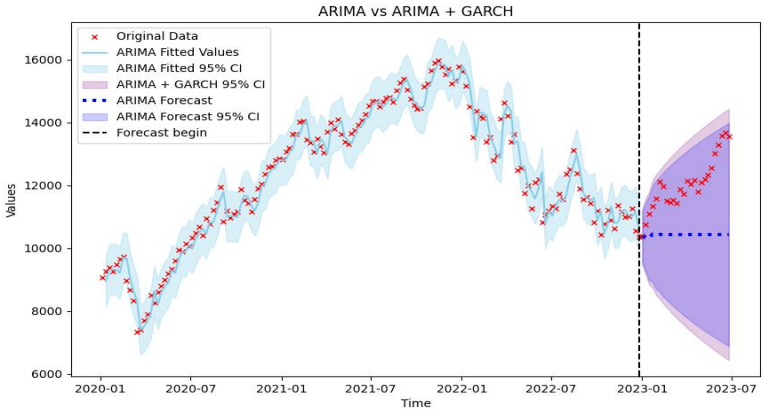


Figure 4 Comparison Chart of Forecasting Performance with and without GARCH Effects

3.3 Empirical Analysis and Results of the Dynamic Regression Model

The main limitation of the ARIMA model in forecasting the Nasdaq index time series is that the forecast period is relatively short, and the predicted values tend to stabilize, making it difficult to reflect the real volatility increase trend (Figure 4). To improve the forecasting accuracy and explore the interconnectedness of the U.S. stock market, this study introduces input variables and constructs a dynamic regression model to reforecast the Nasdaq index trends.

3.3.1 Selection of input variables and data preprocessing

In this study, the S&P 500 index $z_t^{(1)}$, which is also highly watched, is selected as a potential input variable alongside the NASDAQ Composite Index X_t. To avoid heteroscedasticity, both stock time series are first transformed using logarithms. After transformation (Figure 5), the trends of both indices are largely consistent. Since Granger causality tests require stationary series, a first-order difference is applied to the logarithmic series of both indices. The ADF test ($p_x, p_z < 0.001$) shows that both $\nabla \ln x_t$ and $\nabla \ln z_t^{(1)}$ are stationary. The results of the Granger causality test and cointegration test (Table 6) indicate that it is appropriate to fit a dynamic regression model for $\nabla \ln x_t$ and $\nabla \ln z_t^{(1)}$.

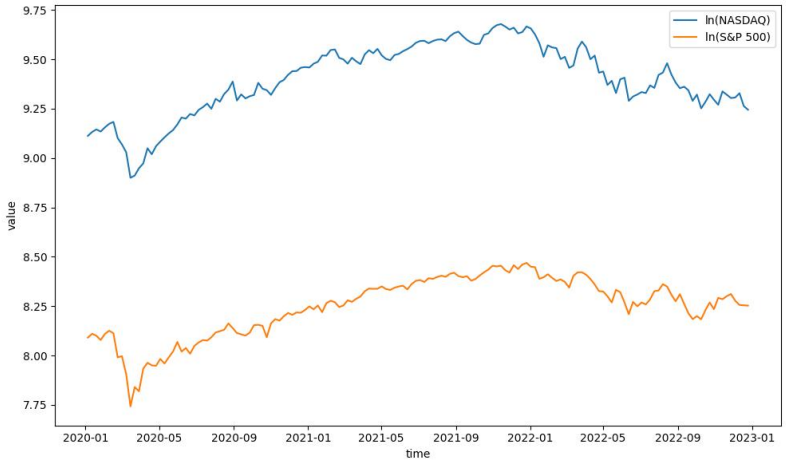


Figure 5 Log-Transformed S&P 500 and Nasdaq Index Trends from 2020 to 2023

Table 6 Granger Causality Diagnosis and Cointegration Test for S&P 500 and Nasdaq

null hypothesis	Statistics	p
The S&P 500 ($\nabla \ln z_t^{(1)}$) is not the Granger cause of the Nasdaq($\nabla \ln x_t$).	F=72.3113	0.0000
no cointegration relationship between $\nabla \ln z_t^{(1)}$ and $\nabla \ln x_t$.	τ =-6.7179	0.0000

3.3.2 Fitting the dynamic regression model to the NASDAQ time series

To determine the specific form of the dynamic regression model, it is necessary to identify several input variables that are strongly correlated with the response series $\nabla \ln x_t$. By examining the cross-correlation function between $\nabla \ln x_t$ and $\nabla \ln z_t^{(1)}$ (Figure 6), significant cross-correlation coefficients at lags 0 and 1 were found. Therefore, $\nabla \ln z_t^{(1)}$ and $\nabla \ln z_{t-1}^{(1)}$ were selected as input variables to be included in the regression equation. The dynamic regression model is constructed as follows:

$$\nabla \ln x_t = \beta_0 + \beta_1 \nabla \ln z_t^{(1)} + \beta_2 \nabla \ln z_{t-1}^{(1)} + \varepsilon_t \tag{15}$$

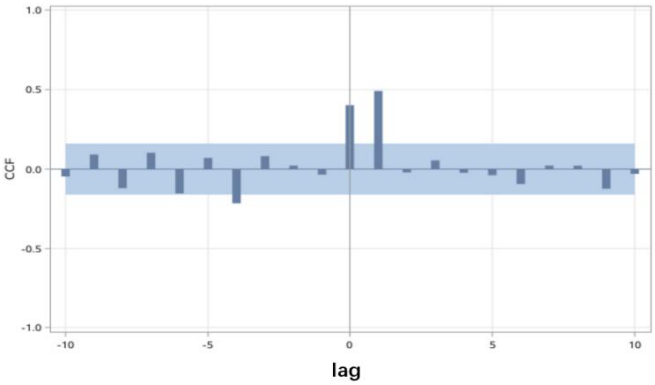


Figure 6 The Cross-Correlation Function Plot between $\nabla \ln x_t$ and $\nabla \ln z_t^{(1)}$

The test reveals that ε_t is stationary and non-white noise, with an ACF tailing off and a PACF exhibiting a first-order cutoff. Therefore, an AR(1) model is fitted to ε_t. After removing the insignificant coefficients, the final expression of the model is as follows:

$$\nabla \ln x_t = 0.4676 \nabla \ln z_t^{(1)} + 0.5583 \nabla \ln z_{t-1}^{(1)} + \varepsilon_t \tag{16}$$

$$\varepsilon_t = \frac{1}{1 + 0.46214B} a_t \ , \ a_t \sim WN(0, 0.000569) \tag{17}$$

3.3.3 Dynamic Regression Forecasting Performance

The forecasting results of the Nasdaq for the first half of 2023 based on the dynamic regression model (Figure 7) show that although there is still some deviation between the predicted values and the actual values, the predicted trend aligns with the real trend, both showing a fluctuating upward trend. This indicates that the inclusion of the S&P 500 index has improved the forecasting accuracy of the Nasdaq index.

Figure 7 Dynamic Regression Model Prediction Results Plot

4 CONCLUSIONS AND OUTLOOKS

This study focuses on the weekly data of the Nasdaq Composite Index (NASDAQ_Index) from January 5, 2020, to June 25, 2023. In order to better capture both the deterministic and conditional volatility information of this financial time series, an ARIMA(1,1,(1,5)) model with GARCH(1,1) disturbances following a t-distribution was initially established to fit the index time series from 2020 to 2022. Subsequently, the index trends for the first half of 2023 were predicted using both models with and without GARCH effects. The results show that the model with GARCH effects produced a wider prediction confidence interval, indicating that this model is more capable of forecasting potential risks in future financial environments. This also suggests that investors need to be more cautious when balancing the returns and risks of Nasdaq stocks.

However, the forecasts generated by the ARIMA-GARCH model became more stable as the forecast period increased, failing to reflect the true rising volatility trend of the Nasdaq index effectively. To improve prediction accuracy, this study then applied the concept of cointegration by introducing the S&P 500 index as an input variable. Through Granger causality and EG cointegration tests, an effective dynamic regression model was initially built. By fully exploiting the information of the input variable using the cross-correlation function between the Nasdaq index and the S&P 500 index,

the model's specific form was determined and refined. The results show that the forecast trend of this model largely follows the true sequence's rising volatility, indicating that the S&P 500 index plays a facilitative role in predicting the Nasdaq index.

Although the dynamic regression model constructed in this study has significantly improved the forecasting accuracy of the NASDAQ Index under the condition of known input variables, it still has certain limitations. Firstly, the model's effectiveness depends on the availability of future values of input variables (such as the S&P 500 Index), which are often unknown in real-world forecasting scenarios. Secondly, even if these input variables are forecasted separately, the associated prediction errors may propagate and negatively impact the accuracy of the response variable forecast. Future research could consider integrating dynamic regression with multivariate time series modeling or machine learning techniques to enhance the model's robustness and adaptability, thereby better capturing the complex volatility patterns in financial markets.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Akin I, Akin M. Behavioral finance impacts on US stock market volatility: An analysis of market anomalies. Behavioural Public Policy, 2024, 1-25.

[2] Muliabanta N H. Peramalan Volatilitas Return IHSG Menggunakan GARCH-Genetic Algorithm-Support Vector Regression. Institut Teknologi Sepuluh Nopember, 2025.

[3] Nugroho D B, Kurniawati D, Panjaitan L P, et al. Empirical performance of GARCH, GARCH-M, GJR-GARCH and log-GARCH models for returns volatility[C]//Journal of Physics: Conference Series. IOP Publishing, 2019, 1307(1): 012003.

[4] Marisetty N. Prediction of Popular Global Stock Indexes Volatility by Using ARCH/GARCH Models. GARCH Models, 2024.

[5] Raza S, Shreevastava A, Meher B K, et al. Predictive Analysis of Volatility for BSE S&P GREENEX index using GARCH Family Models: A case study for Indian stock market. Revista de Științe Politice. Revue des Sciences Politiques, 2024, 84, 54-67.

[6] Roszyk N, Ślepaczuk R. The Hybrid Forecast of S&P 500 Volatility ensembled from VIX, GARCH and LSTM models. arXiv preprint arXiv:2407.16780, 2024.

[7] Shojaie A, Fox E B. Granger causality: A review and recent advances. Annual Review of Statistics and Its Application, 2022, 9(1): 289-319.

[8] Wang P C, Vo T T H. Stock price prediction based on dual important indicators using ARIMAX: A case study in Vietnam. Journal of Intelligent Systems, 2025, 34(1): 20240101.

[9] Akusta A. Exploring ceemdan decomposition for improved financial market forecasting: A case study on dow Jones index. Pamukkale Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 2024, (62): 19-35.

[10] Aylar E, Smeekes S, Westerlund J. Lag truncation and the local asymptotic distribution of the ADF test for a unit root. Statistical Papers, 2019, 60: 2109-2118.

[11] Yasmin S, Moniruzzaman M. Forecasting of area, production, and yield of jute in Bangladesh using Box-Jenkins ARIMA model. Journal of Agriculture and Food Research, 2024, 16: 101203.

World Journal of Information Technology
ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3034

A HYBRID SEQ2SEQ AND BAYESIAN OPTIMIZATION FRAMEWORK FOR PREDICTING OLYMPIC MEDAL DISTRIBUTION WITH UNCERTAINTY ANALYSIS

QiuLin Yao[1*], Feng Cheng[1], YanPeng Guo[1], KuiSong Wang[1], QiSheng Liu[1], Ning Ding[2]
[1]*School of Mechanical Engineering, Jiamusi University, Jiamusi 154007, Heilongjiang, China.*
[2]*School of Materials Science and Engineering, Jiamusi University, Jiamusi 154007, Heilongjiang, China.*
Corresponding Author: QiuLin Yao Email: wy524887227@163.com

Abstract: This paper proposes a robust and data-driven methodology to forecast Olympic medal outcomes with high accuracy and interpretability. A sequence-to-sequence (Seq2Seq) neural architecture is employed to learn temporal dependencies in national Olympic performance, while hyperparameter optimization is conducted using a Tree-structured Parzen Estimator (TPE) to enhance model generalization. To ensure data integrity, preprocessing steps include structured data cleansing and the use of a backpropagation neural network to address missing values. The model further integrates features such as national investment in sports, historical medal trends, and host country effects. In addition to deterministic predictions, uncertainty is quantified through Monte Carlo sampling and confidence intervals, providing probabilistic insights into future outcomes. Experimental results show that the proposed approach outperforms baseline models, achieving an R^2 improvement from 0.827 to 0.875 on the test dataset. The framework is applied to predict the medal distribution for the 2028 Los Angeles Olympics and highlights emerging medal-winning countries. These findings demonstrate the framework′s potential to assist national committees and policy makers in strategic planning for future Olympic participation.
Keywords: Olympic medal forecasting; Sequence-to-sequence neural network; Bayesian hyperparameter optimization; Prediction uncertainty quantification

1 INTRODUCTION

As the most influential sports event in the world, the number of Olympic medals is an important indicator to measure the strength of sports in different countries[1]. Accurate prediction of Olympic medals can not only satisfy the expectation of sports enthusiasts, but also serve as a key reference for the sports departments of different countries to formulate strategic planning and allocate resources. With the development of big data and artificial intelligence, the use of data-driven methods to predict the number of Olympic medals has become a research hotspot, bringing new opportunities for decision-making in the field of sports. Past studies have used a variety of methods in Olympic medal prediction Some studies have used traditional statistical models such as linear regression to make predictions by analyzing historical data and key information related factors, but such models are difficult to capture the complex nonlinear relationships in the data [2].

In the existing research on Olympic medal prediction, the traditional linear model is difficult to capture the complex nonlinear relationship due to structural limitations, the machine learning method is insufficient to mine the temporal dynamic features of the number of medals, and the neural network has problems such as lack of prediction uncertainty analysis and insufficient model depth. In view of these shortcomings, this study is improved in three aspects: constructing a hybrid architecture that fuses time series models and deep learning, and strengthening the modeling of nonlinear and time series features; A Bayesian deep learning framework is introduced to quantify the uncertainty of prediction results. Integrating multi-dimensional variables such as historical medal data, sports resource investment, and event characteristics, a high-precision special model is created to improve the accuracy of medal prediction for the Los Angeles 2028 Olympic Games, and deeply analyze the core factors affecting the medal performance of various countries, making up for the lack of model capabilities and analysis dimensions in existing research. In recent years, machine learning methods have been gradually applied to the field of decision counting, neural networks, etc. However, these models have limitations in dealing with the temporal and complex nature of the number of medals, and the accuracy needs to be improved, while the analysis of uncertainty in the prediction results is not deep enough. This study is only in overcoming the inadequacy of existing research to construct a high-precision prediction model for Olympic medals, and quantitatively analyze the uncertainty of prediction results. Specific objectives include predicting the number of medals for each country in the 2028 Olympic Games in Los Angeles and analyzing the factors affecting the performance of some countries. In terms of methodological innovation, this approach integrates time-series deep learning models with Bayesian frameworks, breaking through the limitations of traditional linear models in capturing nonlinear relationships, while also addressing the shortcomings of existing models in analyzing prediction uncertainty. In model construction, it incorporates multidimensional variables such as sports resource investment and event characteristics to create a highly accurate specialized model, deepening the systematic analysis of factors influencing medal outcomes. On the application level, it achieves precise predictions of medal counts for the 2028 Los Angeles

Olympics, and through variable attribution, it uncovers core influencing factors, providing deeper references for decision-making in sports departments, thus filling the gaps in existing research regarding model accuracy, uncertainty quantification, and multidimensional factor analysis.

2 MATERIALS AND METHODS

2.1 Data Acquisition and Pre-Processing

In this paper, this article collected information on the number of medals and athletes from each country from 1896-2024, which were obtained from the open source website https://www.contest.comap.com/undergraduate/contests/ Firstly, difference set analysis is used to supplement the countries that do not appear in the medal data, and then the item data is cleaned, including filling missing values, removing irrelevant fields, and handling special items [3]. Then the cleaned data is linked to the medal data by year, and the countries are inwardly linked and counted to win the awards in each program. Based on the error back propagation BP neural network, the missing values in the dataset are filled to ensure the integrity of the data, features are extracted, and the data is divided into a training set and a test set in the ratio of 7:3 [4]. The training set is used for learning the parameters of the model and the test set is used to evaluate the predictive performance of the model.

2.2 Methodology

The aim of this study is to predict the number of medals that each country will win at the Olympic Games and to estimate the uncertainty of the prediction results. First, data preprocessing and cleaning. After the data are organized, the features include the number of types of Olympic sports corresponding to each year, the number of medals, the country code, the distribution of medals in previous years, the logo of the host country, and the participation of each sport. Next, the number of gold, silver, and bronze medals won by each country each year is used as the dependent variable, and other features are used as independent variables to split the data into a training set and a test set. The task of predicting the number of medals can be handled by a variety of machine learning regression models. In order to improve the prediction performance and generalization ability of the model, this study uses a heuristic algorithm to optimize the hyperparameters of the model. The accuracy and stability of the model were ensured by adjusting the hyperparameters and using metrics such as R^2 for model evaluation. Regarding the uncertainty analysis of the prediction results, this study combines Monte Carlo simulation and confidence intervals to quantify the reliability of the model predictions. The uncertainty of the prediction results is comprehensively evaluated by introducing random perturbations to sample the input data multiple times, generating multiple sets of prediction results, and calculating the mean and standard deviation of the prediction distribution. Finally, after constructing the optimal model and completing the training, the prediction is made based on the relevant data of the 2028 Los Angeles Olympic Games, which provides scientific and accurate prediction and reference for the future Olympic medal distribution. Will add new items to adjust the relevant features, excluding Russia and other countries that do not participate, to get the medal list with its confidence interval, and then analyze the rise and fall of each country's performance); through the previous difference set analysis of the complementary non-winning countries, the model also predicts and evaluates which countries are likely to win medals for the first time in 2028, and provides the predicted probability, which can be nested into the activation function by the regression value.

In order to predict the number of Olympic medals (including the number of gold medals and the total number of medals) for each country, a sequence-to-sequence (Seq2Seq)-based deep learning model was developed and the hyperparameters of the model were optimized using a tree-structured Parzen Estimator (TPE) algorithm [5]. The model provides reliable predictions for future Olympic medal distributions by learning complex temporal and feature relationships in historical data, while the uncertainty of the prediction results is quantified and analyzed in detail.

The Seq2Seq model is a deep learning method for sequence prediction and consists of an encoder and decoder. The input features cover information such as country codes and the target variables are vectors about the distribution of medals [6]. The encoder transforms the input sequence into an implicit representation of fixed dimensions, and the decoder generates the target sequence based on this representation. The model is optimally trained by stochastic gradient descent using mean square error as loss function.

2.3 Model Evaluation Indicators

To assess the predictive ability of the model, R^2 was chosen as the main performance indicator:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \quad (1)$$

Where y_i is the true value, $\hat{y}_i$ is the predicted value, and $\bar{y}$ is the mean value of the target variable. the closer the R^2 indicator is to 1, the stronger the explanatory power of the model on the target variable.

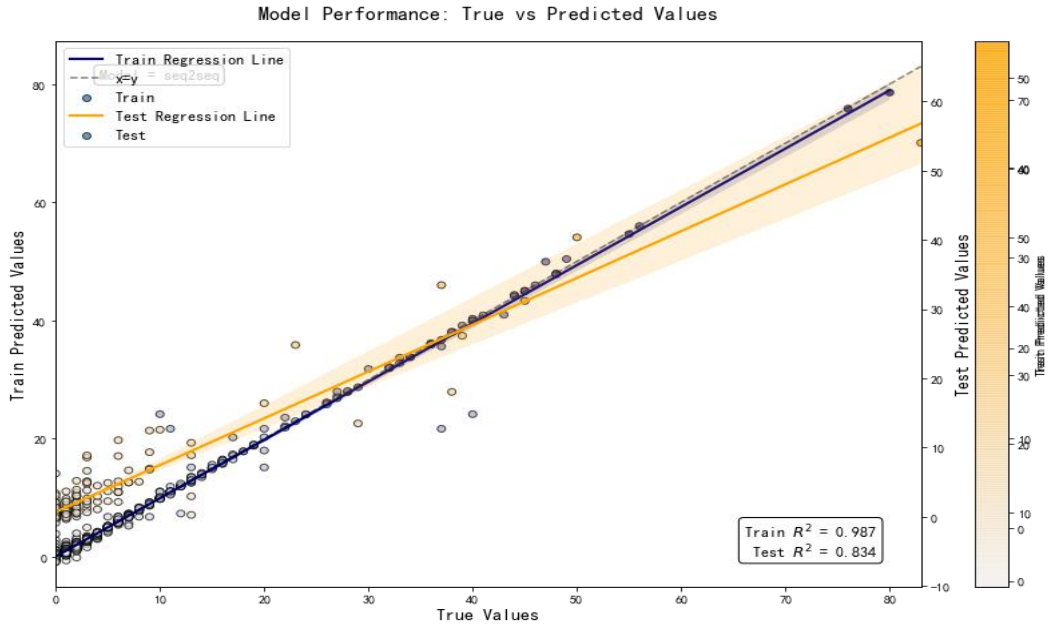
3 MODELING AND SOLVING

3.1 hyperparametric Optimization

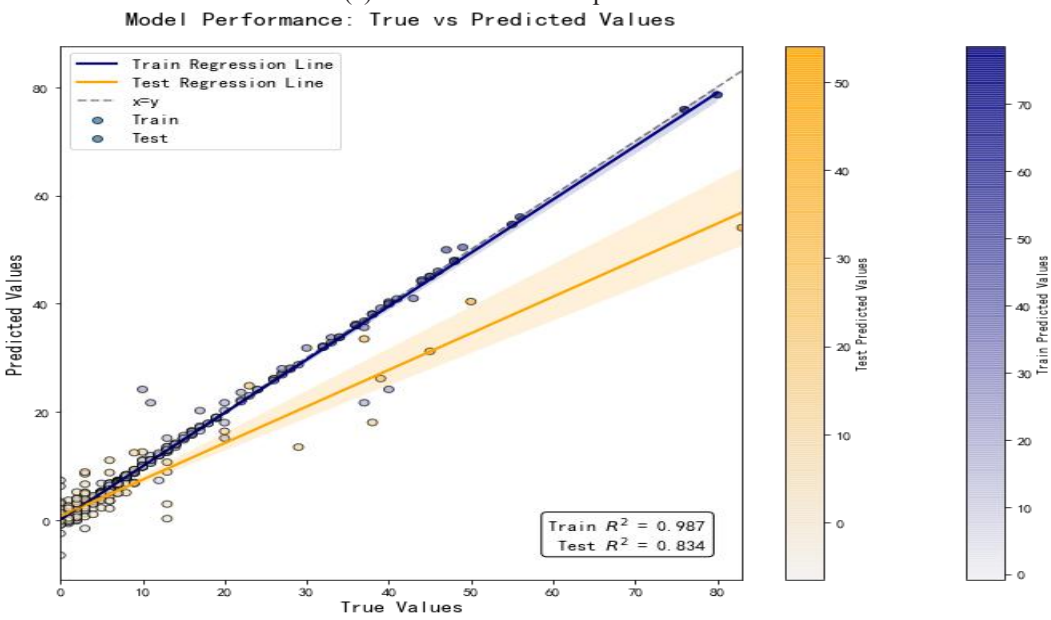
The optimization results are as follows: the learning rate is 0.001, the number of hidden layer units h is 256, the batch size b is 64, the regularization coefficient λ is 0.0001, the number of encoder layers is 2, the number of decoder layers is 2, the time step T is 10, and the activation function is ReLU.
The above optimization results show that the learning rate α, the number of hidden layer units h and the batch size b are the key hyperparameters affecting the performance of the model. Among them, a small learning rate α ensures the stable convergence of the sub-model, while a medium-sized hidden layer unit number and batch size take into account the expressive ability and training efficiency of the model.

3.2 Model Checking

During the initial training of the Seq2Seq model, the R^2 of the model on the training set reaches 0.986 under the default parameter settings, while the R^2 of the model on the test set is only 0.827, which shows that the model is overfitting. To alleviate this problem, the hyperparameters of the model are adjusted through TPE optimization, including reducing the number of hidden layer units, increasing the regularization factor λ, and adjusting the key parameters such as the batch size and the number of time steps. After the adjustment, the performance of the model is significantly improved, and the R^2 on the training set and test set reaches 0.987 and 0.875, respectively, which effectively reduces the overfitting phenomenon, as shown in Figure. 1 below[7][8].



(a) Parameters before optimization



(b) TPE adjustment parameters

Figure 1 Model Performance: True vs Predicted Values

For the model under the recognized parameter settings before tuning, the fitting effect of the training set is very good, and the point cloud is densely distributed near the reference line with almost no deviation; however, in the test set, the dispersion of the point cloud is larger, and the regression line deviates from the reference line more obviously, indicating that the model's generalization ability is weaker.After the TPE adjusts the parameters, the point cloud of the prediction results in the test set shrinks significantly, and the distribution is closer to the reference line, indicating that the accuracy and stability of the prediction have been improved. This indicates that the accuracy and stability of the prediction have been improved. Meanwhile, the performance of the training set is slightly reduced, but still maintains a high R^2, indicating that the overall performance of the model tends to be balanced.

3.3 Model Prediction

3.3.1 Constructing the prediction data

To predict the number of Olympic medals in 2028, a new input feature dataset needs to be constructed first. This dataset is based on the existing data of the 2024 Olympic Games, and the relevant features are adjusted according to the new programs of the 2028 Los Angeles Olympic Games. The specific process is described below:

1) Screening of 2024 data

The base dataset X_2024 is constructed by selecting relevant records from the original dataset for the year 2024, from which the data for Russia (i.e., rows with a NOC value of 113) are excluded because Russia was banned in 2028:

$$X_{2028} = X_{2024}[X_{2024}['year'] == 2024 \wedge ['NOC'] \neq 113 \quad (2)$$

2) Re-indexing:

Index reset on filtered data to ensure that the data is neatly organized:

$$X_{2028}.reset\_index(inplace = True, drop = True) \quad (3)$$

3) Medal projections for new sports:

Based on the new sports that have been approved by the IOC (e.g., cricket, squash, baseball, softball, stickball and flag rugby), adjust the corresponding number of medals. The specific adjustment rules are as follows:

a) Baseball and softball: the new men's and women's events will have one gold medal each; thus adding 2 gold medals per event;

b) Cricket: one gold medal is created for each of the new men's and women's events, totaling 2 gold medals;

c) Stick tennis: creation of one gold medal for each of the new men's and women's disciplines, totaling 2 gold medals;

d)Squash: the creation of a men's and women's singles event with one gold medal each, totaling 2 gold medals;

e) Flag Rugby: create one gold medal for each of the new men's and women's events, for a total of 2 gold medals. The corresponding adjustment formula is:

$$X_{2028}['Baseball'] = X_{2028}['Baseball'] + 2 \quad (4)$$

$$X_{2028}['Softball'] = X_{2028}['Softball'] + 2 \quad (5)$$

$$X_{2028}['Cricket'] = X_{2028}['Cricket'] + 2 \quad (6)$$

$$X_{2028}['Sixes'] = X_{2028}['Sixes'] + 2 \quad (7)$$

$$X_{2028}['Squash'] = X_{2028}['Squsah'] + 2 \quad (8)$$

$$X_{2028}['Flagfootball'] = X_{2028}['Flagfootball'] + 2 \quad (9)$$

4) Update the year:

Since 2028 is a future year for the Olympic Games, the year information in the data needs to be updated to 2028:

$$X_{2028}['YEAR'] = 2028 \quad (10)$$

5) Host country identification:

For the United States (NOC of 147) as the host country for the 2028 Olympic Games, it needs to be identified as 1 other countries remain at 0.

$$X_{2028}['Host_{Country}'] = 0 \quad (11)$$

$$X_{2028}.loc[X_{2028}['NOC'] == 147, 'Host\_Country'] = 1 \quad (12)$$

3.3.2 Performance Analysis

1) Medal Table and Confidence Intervals for the 2028 Los Angeles Summer Olympics

Based on the established medal prediction model, the number of medals for each country at the 2028 Los Angeles Summer Olympics was predicted, and an uncertainty analysis of these medal counts was conducted, resulting in the corresponding prediction intervals. Below are the predicted results for some countries, including the number of gold, silver, and bronze medals, as well as the corresponding confidence intervals, as shown in Table 1:

Table 1 Medal Count Projections and Confidence Intervals

NOC	gold	Silver	Bronze	gold-CI-Iower		gold-CI-upper	Silver-CI-Iover	Silver-CI-upper	Bronze-CI-Iower
United States	39	45	43	36	40	36	45	37	43
China	40	27	24	37	40	24	27	21	24
Japan	20	13	12	14	20	11	14	9	12
Austraila	18	19	17	12	17	8	19	12	17
France	15	25	20	12	18	18	24	16	20
…	…	…	…	…	…	…	…	…	…
Samoa	0	0	0	0	1	0	1	0	1

Mixed team	1	1	2	0	1	1	2	1	2
Crylon	0	0	1	0	0	0	1	1	1
FR Yugoslavia	0	1	1	0	1	0	1	1	1
ROC	4	2	8	0	7	0	5	0	12

The following graph shows the predicted number of Gold, Silver and Bronze medals and the corresponding confidence intervals (CIs), where the predicted values are labeled by curves and scatters, and the confidence intervals are indicated by shaded areas. The horizontal coordinate indicates the index of the data point and the vertical coordinate indicates the number of medals predicted. The predicted values for gold, silver, and bronze medals are shown as dark purple, light purple, and rose curves, respectively, with each curve accompanied by its corresponding confidence interval (gold CI, silver CI, and bronze CI, shaded, respectively).

As can be seen from the figure, the predicted values tend to stabilize as the number of data points increases, whereas in some positions (e.g., near the first few data points), there are large fluctuations, indicating that the model has a high prediction uncertainty at these positions. This is further verified by the width of the confidence intervals, where wider regions represent greater uncertainty in the predictions, while narrower regions indicate more accurate predictions, as shown in Figure 2 below.

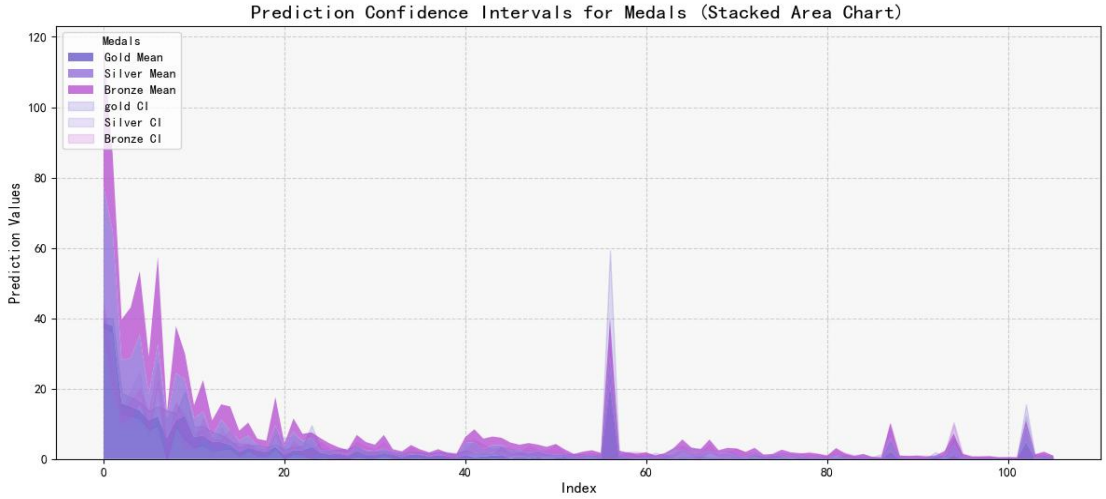


Figure 2 Prediction Confidence Intervals For Medal

By comparing the projected medal totals for 2024 and 2028, the following countries are expected to significantly improve their medal performance, as shown in Table 2 below:

Table 2 Achievement Gains in Selected Countries

NOC	2024Total	2028Total	Improvement
ROC	0	4	14
Mixed team	0	4	4
Monglia	1	4	3
Kazakhstan	7	9	2
South Africa	6	8	2
Serbia	5	7	2
Israel	7	9	2
FR Yugoslavia	0	2	2
Malaysia	2	4	2

These countries are projected to see a significant increase in performance in 2028, especially ROC and Mixed team, which have both seen significant increases in their medal totals, reflecting the strong momentum in these countries and regions.

The following countries are projected to see a decrease in performance in 2028 compared to the countries that have improved Table 3 below:

Table 3 Declining Performance in Selected Countries

NOC	2024Total	2028Total	Improvement
South Korea	32	23	-9
France	64	60	-4
Turkey	8	5	-3
Great Britain	65	63	-2
North Korea	6	4	-2
Greece	8	6	-2
India	6	4	-2

Iran	12	10	-2
Denmark	9	8	-1
Belgium	10	9	-1

The total number of medals for these countries is projected to decline in 2028, with South Korea and Frances in particular experiencing a reduction of 9 and 4 medals respectively, reflecting a possible limitation of their potential in future Olympic events.

2) Predicting countries that will win medals for the first time

For countries that have not yet won a medal, the model also makes a prediction and assesses which countries are likely to win their first ever

medals. By filtering out countries that did not win a medal in 2024 but are expected to win a medal in 2028, the following list of countries was obtained as shown in Table 4 below:

Table 4 National Medal Projections

NOC	Mean Probability of Winning
Zambia	0.602431
Independent Olympic Athletes	0.65913
Virgin Islands	0.594907
British West Indies	0.558821
Independent Olympic Participants	0.648978
Mixed team	0.752325
Ceylon	0.564092
FR Yugoslavia	0.602568
ROC	0.964915

It can be seen that ROC and Mixed team have the highest probability of winning a medal in 2028, 0.964915 and 0.752325 respectively, indicating that they are more likely to win a medal in 2028.

However, if the question asks to extend the caliber of the analysis to historical data, these countries have won medals in their history, so their probability of "winning a medal for the first time" is zero.

By analyzing the medal predictions for the 2028 Summer Olympics in Los Angeles and combining them with the predicted performance intervals for each country, some valuable predictive information can be obtained for future events. This information can not only help countries to make corresponding preparation strategies, but also provide data support for the IOC and event organizers to help them better plan and prepare for the upcoming Olympic Games. At the same time, through the prediction of first-time medal-winning countries, it is possible to better understand which countries and regions have not yet fully realized their potential for sports development and have more room for improvement.

4 CONCLUSION

In this paper, the Olympic medal distribution was successfully predicted by constructing a model. The Olympic medal prediction model provides a scientific basis for medal prediction through data preprocessing, BP neural network missing value filling, TPE hyperparameter optimization and prediction uncertainty analysis. However, the limitation of data and the complexity of the model still need further improvement. Future research can consider introducing more data sources, such as athletes' personal training data and international competition results, to improve the prediction accuracy of the model. In addition, more efficient model structures and optimization algorithms can be explored to reduce training time and resource consumption.

Overall, the model in this paper provides new methods and ideas for understanding and predicting Olympic medal distributions as well as assessing coaching effects. Through continuous improvement and optimization of the model, it can provide more scientific and accurate decision support for countries' sports development strategies and coaching resource allocation.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Shi Huimin, Zhang Dongying, Zhang Yonghui. Can Olympic medals be predicted? Journal of Shanghai Sport University, 2024, 48(04): 26-36.

[2] Chen Zhanshou, Liang Yan, Wei Qiuyue. Examination of structural variation points in linear regression models with LMSV errors. System Science and Mathematics, 2025: 1-18.

[3] Xiong Zhongmin, Guo Huaiyu, Wu Yuexin. A review of research on missing data processing methods. Computer Engineering and Applications, 2021, 57(14): 27-38.

[4] Denicolò V, Polo M. Duplicative research, mergers and innovation. Economics Letters, 2018, 166: 56-59.
[5] L Zhang, B Ding, JY Deng, et al. Study on urban subsurface change and runoff coefficient response based on BP neural network. Journal of Changjiang Academy of Sciences, 2025: 1-7.
[6] LUO Min, YANG Jinfeng, YU Hui,et al. A short-term load forecasting method based on TPE optimization and integrated learning. Journal of Shanghai Jiao Tong University, 2023(5).
[7] Li W J, Wu LL, Wen SH, et al. Optimization of LSTM-Seq2seq model for runoff simulation based on attention mechanism. Glacial Permafrost, 2024, 46(3): 980-992.
[8] You Lan, Han Xuewei, He Zhengwei, et al. An Improved Seq2Seq-Based Model for Short-Term AIS Trajectory Sequence Prediction. Computer Science, 2020, 47(09): 169-174.

World Journal of Information Technology
ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3035

CONSTRUCTION AND PRELIMINARY APPLICATION EFFECTIVENESS OF AN INFORMATICS-INTEGRATED TRADITIONAL CHINESE MEDICINE PREVENTIVE TREATMENT SERVICE MODEL

Lei Zhang, SiSi Li, ZiYang Wang, WenHui Lu*
Shanghai Municipal Hospital of Traditional Chinese Medicine, Shanghai University of Traditional Chinese Medicine, Shanghai 200090, China.
Corresponsing Author: WenHui Lu, Email: 49949745@qq.com

Abstract: Objective: To construct an informatics-integrated Traditional Chinese Medicine (TCM) preventive treatment service model and evaluate its effectiveness in improving service efficiency, patient compliance, satisfaction, and health outcomes. **Methods:** Based on the TCM preventive treatment concept, a digital service model was designed, incorporating intelligent constitution identification, personalized intervention, remote follow-up, and real-time monitoring, implemented using a B/S architecture and cloud deployment. A randomized controlled study was conducted at our hospital's preventive treatment clinic from June 2024 to February 2025, involving 216 sub-health patients (aged 18–65 years) randomly assigned to an informatics service group (n=108) or a traditional service group (n=108). Evaluation metrics included service efficiency, follow-up completion rate, compliance, satisfaction, and sub-health improvement rate, analyzed using t-tests and χ^2 tests. **Results:** The informatics group outperformed the traditional group in constitution identification accuracy (90.4% vs. 82.1%, P<0.05), follow-up completion rate (92.5% vs. 69.3%, P<0.001), compliance (89.7% vs. 74.2%, P=0.003), satisfaction (95.4% vs. 78.7%, P<0.001), and sub-health improvement rate (86.1% vs. 72.3%, P=0.009), with a 40.1% increase in service efficiency (P<0.001). **Conclusion:** This informatics-integrated service model significantly enhances service efficiency and patient health management, providing practical evidence for the modernization of TCM preventive treatment.
Keywords: TCM preventive treatment; Informatics technology; Personalized intervention; Service efficiency; Health management

1 INTRODUCTION

As one of the core concepts in Traditional Chinese Medicine (TCM), "preventive treatment of disease" emphasizes three key principles: preventing disease before it occurs, preventing the progression of existing diseases, and preventing recurrence after recovery. With the growing emphasis on chronic disease prevention and health management in modern medicine, this concept has shown great potential for wider application. In recent years, along with the national strategy to promote TCM development, the "preventive treatment" system has gradually transitioned from theoretical exploration to clinical practice. However, current implementation still faces numerous challenges, such as fragmented service processes, inconsistent documentation, irregular follow-up management, and poor patient compliance, all of which significantly hinder improvements in service efficiency and quality [1].

With the rapid advancement of information technology, healthcare service models are evolving toward digitalization, intelligence, and interconnectivity [2]. In this context, integrating digital solutions into the full service process of TCM preventive treatment can facilitate the consolidation and sharing of diagnostic and treatment information, enhance service continuity and accessibility, and leverage data-driven approaches to optimize intervention strategies. This integration also holds promise for improving the personalization and precision of health management [3]. Nevertheless, systematic research on the construction and evaluation of such integrated models—combining TCM preventive care with digital technology—remains limited. There is an urgent need for empirical studies in real-world clinical settings to explore the mechanisms and effectiveness of such models.

This study, grounded in the concept of "Digital Qihuang, Smart Health", integrates modern information technology with TCM preventive treatment theory to develop a comprehensive service model encompassing assessment, intervention, and follow-up. We applied this model in real clinical scenarios to evaluate its impact on service efficiency, patient compliance, and satisfaction. The goal is to provide both theoretical support and practical guidance for the modernization and standardization of preventive TCM services.

2 METHODS

2.1 Overall Design of the Service Model

Based on the core concept of preventive treatment in Traditional Chinese Medicine (TCM), this study utilized big data, artificial intelligence (AI), and mobile internet technologies to develop a digital full-process service model encompassing intelligent assessment, personalized intervention, dynamic follow-up, and outcome feedback. The model

comprises four functional modules:① Intelligent Constitution Identification and Risk Assessment;② Automated Generation of Personalized Intervention Plans;③ Remote Dynamic Follow-Up;④ Real-Time Data Monitoring and Quality Feedback. The overall architecture adopts a B/S (Browser/Server) structure, with server-side deployment supported by cloud computing technologies. The client-side supports both web and mobile platforms. Data exchange between front-end and back-end systems is conducted using the standardized JSON format.

2.2 Key Technologies and Module Implementation

2.2.1 Intelligent constitution identification and risk assessment module

Patient self-reported symptoms and constitution scale data were collected via a custom-developed mobile application, alongside TCM four-diagnostic information (inspection, auscultation/olfaction, inquiry, and palpation) to generate a basic health profile. A deep learning model based on Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks was constructed in Python to achieve automated constitution classification. Additionally, a health risk prediction model was built using patients' historical data and real-time inputs, with risk stratification and early warnings implemented via the random forest algorithm.

2.2.2 Automated generation of personalized intervention plans module

A TCM knowledge graph was constructed using a Neo4j graph database, incorporating classical literature and expert consensus guidelines. Based on patients' constitution types and risk levels, an inference engine matched personalized intervention strategies. A hybrid rule-based and machine learning algorithm was developed to automatically generate tailored recommendations for diet, exercise, and TCM-specific therapies (e.g., the "Three-Step Therapy for Fatty Liver").

2.2.3 Remote dynamic follow-up module

Real-time follow-up data transmission was implemented using the WebSocket communication protocol. Natural Language Processing (NLP) techniques were employed to automatically parse and structure patients' chief complaints. A built-in alert mechanism, based on predefined threshold indicators, was designed to automatically notify healthcare professionals of abnormal conditions requiring timely intervention.

2.2.4 Real-time data monitoring and quality feedback module

Real-time visualization of service indicators—such as follow-up response rate, patient compliance, and satisfaction—was achieved using ECharts technology. A multidimensional data analysis framework was used to construct a quality scoring system and generate regular service quality reports. Integration of Tableau and Python enabled interactive dashboards for continuous quality monitoring(Figure 1).

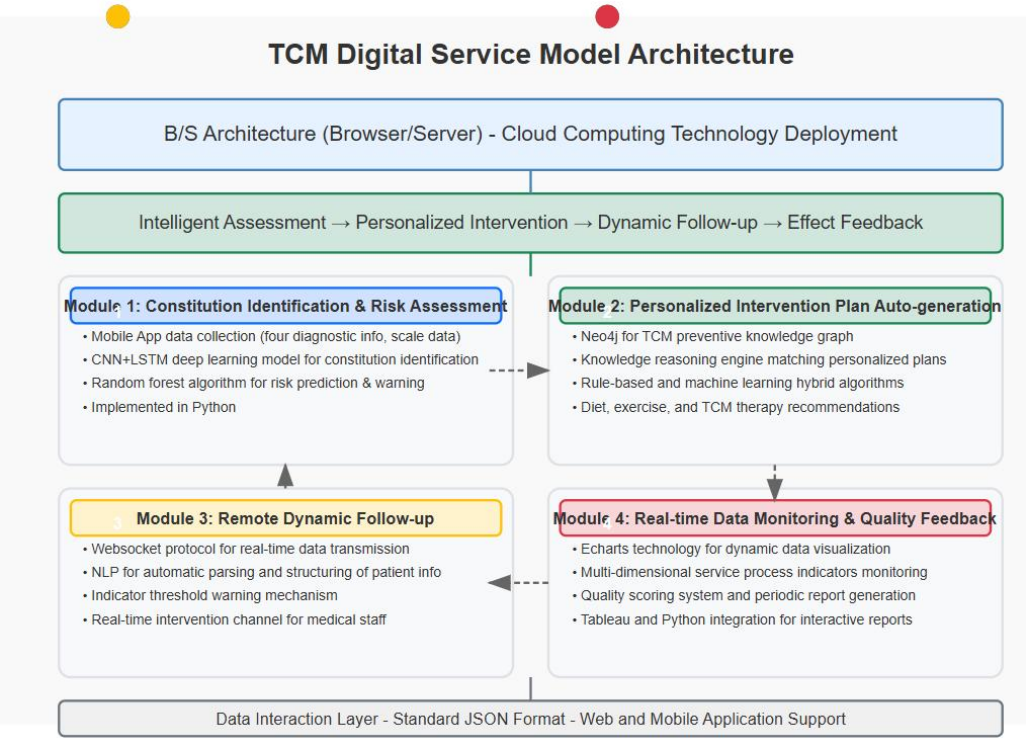


Figure 1 Architecture of the Digital Service Model Based on the TCM Concept of Preventive Treatment

2.3 Application and Implementation Scenario

The digital service model was deployed in the Preventive Treatment Outpatient Clinic of our hospital and applied to patients between June 2024 and February 2025. The sample size was calculated based on the primary outcome indicator—patient compliance rate. According to preliminary study data, the expected compliance rate in the digital

service group was 90%, compared with 75% in the traditional service group, yielding an effect size of 15%. A two-sided test with a statistical power of 80% and a significance level of $\alpha = 0.05$ indicated that each group would require at least 102 participants. Considering a 10% dropout rate, 108 patients were included in each group, for a total of 216 participants.

To reduce selection bias, a quasi-randomized controlled design was adopted, using a sealed envelope method for group allocation. Patients were assigned in a 1:1 ratio to the digital service group and the traditional service group according to enrollment sequence. Due to the significant differences in intervention delivery methods (digital platform vs. manual operations), this study was conducted as an open-label trial. To minimize observation bias, data collection and outcome evaluation (e.g., compliance and satisfaction surveys) were conducted by third-party evaluators independent of the intervention process. These evaluators received standardized training before assessments. The study protocol was approved by the institutional ethics committee, and written informed consent was obtained from all participants.

The digital service model was implemented via a mobile application and web platform to enable intelligent constitution identification, automated generation of personalized intervention plans, remote dynamic follow-up, and real-time data monitoring. In contrast, the traditional manual service model relied entirely on human operation: constitution identification was performed by TCM physicians using paper-based questionnaires and face-to-face consultations, taking approximately 15–20 minutes; intervention plans were manually formulated by TCM physicians based on experience and clinical guidelines, requiring around 10 minutes; follow-up was conducted via telephone or in-person clinic visits, with records maintained in paper files, lacking real-time data transmission and early warning mechanisms; service quality feedback was manually collected through patient satisfaction questionnaires without dynamic visualization. Both groups received the same TCM-based preventive interventions (diet, exercise, and TCM-specific therapies), but the digital group benefited from process automation and data integration via the digital platform, whereas the traditional group depended entirely on manual processes and paper-based documentation.

2.4 Evaluation Indicators and Data Analysis

Evaluation indicators included: Functional realization rate of the model; Service process indicators (e.g., service efficiency, follow-up completion rate, personalized plan matching rate); Patient satisfaction; Improvement magnitude of intervention outcomes. Statistical analysis was performed using SPSS version 26.0. Continuous variables were expressed as mean ± standard deviation, and intergroup comparisons were conducted using independent sample t-tests. Categorical variables were presented as percentages and compared using chi-square tests. A P-value < 0.05 was considered statistically significant.

3 RESULTS

3.1 Baseline Characteristics of the Study Population

A total of 216 patients were enrolled in this study, with 108 in the digital service group and 108 in the traditional service group. Baseline characteristics for both groups are presented in Table 1. There were no statistically significant differences between the two groups in terms of age, gender, sub-health scores, or distribution of primary TCM constitution types ($P > 0.05$), indicating good comparability between the groups.

Table 1 Comparison of Baseline Characteristics Between the Two Groups

Characteristic	Digital Service Group (n = 108)	Traditional Service Group (n = 108)	Test Statistic	*P-value*
Age (years, mean ± SD)	42.3 ± 12.5	43.1 ± 11.8	t=0.47	0.639
Gender (male/female, n)	52/56	50/58	χ^2=0.07	0.791
Sub-health score (mean ± SD)	65.4 ± 10.2	66.1 ± 9.8	t=0.52	0.604
Primary TCM Constitution Types (n, %)			χ^2=0.92	0.821
Balanced Constitution	30 (27.8%)	28 (25.9%)		
Qi-Deficiency Constitution	25 (23.1%)	27 (25.0%)		
Damp-Heat Constitution	20 (18.5%)	22 (20.4%)		
Others (Yin-deficiency, Phlegm-dampness, Blood stasis, Qi stagnation, Special diathesis)	33 (30.6%)	31 (28.7%)		

Note: The sub-health score was assessed using the standardized Sub-health Status Scale (range: 0–100, with higher scores indicating more severe sub-health conditions). Constitution types were determined using the standardized TCM Constitution Identification Scale.

3.2 Outcomes of the Digital TCM Preventive Service Model Construction

This study successfully developed an integrated digital service model for TCM-based preventive care, comprising four core components: intelligent constitution identification and risk assessment, automated generation of personalized intervention plans, remote dynamic follow-up, and real-time data monitoring with quality feedback.

3.2.1 Intelligent constitution identification module

Using a deep learning model based on Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM), the system achieved an accuracy rate of 90.4% in identifying TCM constitution types. This was significantly higher than that of traditional manual methods (82.1%), with the difference being statistically significant ($P < 0.05$). Additionally, the health risk prediction model demonstrated an accuracy rate of 88.6%.

3.2.2 Automated personalized intervention plan module

The module generated personalized intervention plans in an average time of 1.5 ± 0.3 minutes, which was significantly shorter than the time required for traditional manual planning (10 ± 2.6 minutes, $t = 28.45$, $P < 0.001$). The satisfaction rate with the individualized plans reached 92.6%(Figure 2).

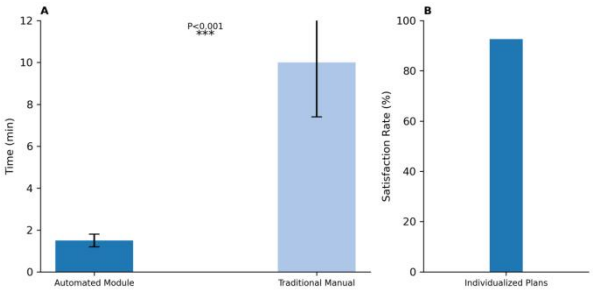


Figure 2 Comparison of Average Time for Personalized Intervention Plan Generation Between Modules

3.2.3 Remote dynamic follow-up module

By leveraging WebSocket communication and Natural Language Processing (NLP) technologies, the follow-up module enabled real-time acquisition of patient follow-up data and automatic alerts for abnormal findings. The follow-up completion rate increased by 23.2% compared to the traditional service model (92.5% vs. 69.3%, $\chi^2 = 17.84$, $P < 0.001$). Moreover, the rate of timely intervention for patient health management reached 87.9%(Figure 3).

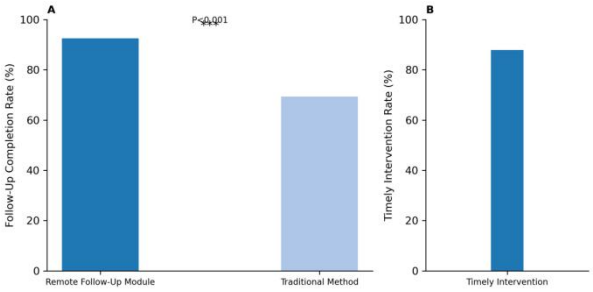


Figure 3 Comparison of Follow-Up Completion Rates Between Modules

3.2.4 Real-time data monitoring and quality feedback module

Powered by ECharts and Tableau, the real-time data monitoring and feedback module enabled dynamic visualization and interactive display of key service process indicators. Service quality evaluation showed that overall satisfaction with service quality reached 93.8% in the digital model group, significantly higher than 80.6% in the traditional service group ($\chi^2 = 8.12$, $P = 0.004$)(Figure 4).

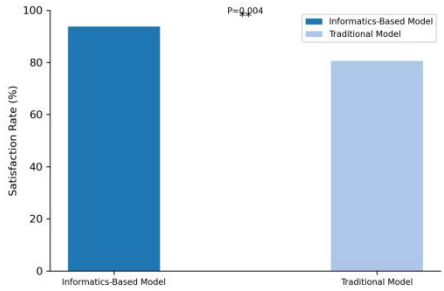


Figure 4 Comparison of Overall Service Quality Satisfaction After Implementation of the Digital Service Model

3.3 Improvements in Service Efficiency and Workflow Optimization

Comparative analysis revealed that, following implementation of the digital service model, the average duration of a single outpatient visit was reduced from 35.4 ± 6.7 minutes (traditional model) to 21.2 ± 3.5 minutes (t = 18.26, P < 0.001), representing an approximate 40.1% increase in service efficiency(Figure 5).

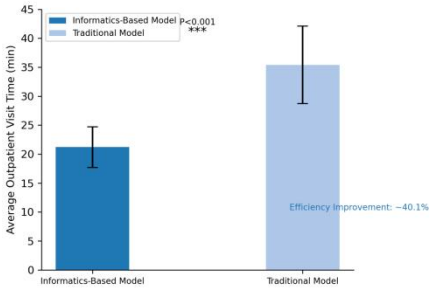


Figure 5 Comparison of Average Outpatient Visit Duration Before and After Implementation of the Digital Service Model

3.4 Improvement in Patient Compliance and Satisfaction

Patient compliance with intervention measures was recorded through the follow-up system. The compliance rate in the digital service group reached 89.7%, significantly higher than 74.2% in the traditional service group (χ^2 = 8.67, P = 0.003).

According to the patient satisfaction questionnaire, the overall satisfaction rate under the digital service model was 95.4%, significantly higher than 78.7% in the control group (χ^2 = 12.35, P < 0.001)(Figure 6).

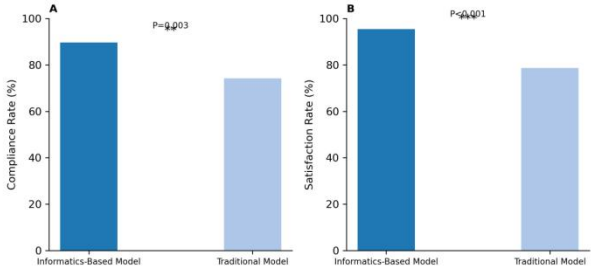


Figure 6 Comparison of Patient Compliance After Implementation of the Digital Service Model

3.5 Preliminary Observation of Health Improvement Outcomes

Following the implementation of the digital service model, the self-reported sub-health symptom improvement rate among patients reached 86.1%, significantly higher than 72.3% in the traditional service model group (χ^2 = 6.75, P = 0.009).

In addition, the post-intervention improvement in TCM constitution scores was also significantly greater in the digital service group (15.6 ± 3.8 points) compared to the traditional group (9.8 ± 3.1 points, t = 12.27, P < 0.001), suggesting that the digital model had a positive impact on health promotion(Figure 7).

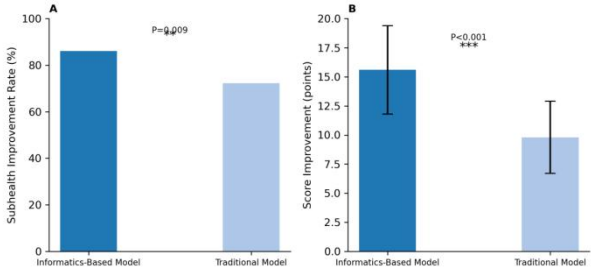


Figure 7 Comparison of Patient Satisfaction After Implementation of the Digital Service Model

4 DISCUSSION

This study successfully developed and validated an information technology–integrated Traditional Chinese Medicine (TCM) service model for preventive care, encompassing four core modules: intelligent constitution identification, personalized intervention planning, remote follow-up, and real-time data monitoring with feedback. The results demonstrated that this model significantly improved service efficiency (40.1% increase, P < 0.001), patient compliance (89.7% vs. 74.2%, P = 0.003), satisfaction (95.4% vs. 78.7%, P < 0.001), and sub-health symptom improvement (86.1%

vs. 72.3%, P = 0.009), providing both theoretical and practical evidence for the modernization of TCM preventive service delivery.

The intelligent constitution identification module, based on a CNN-LSTM deep learning architecture, achieved a constitution classification accuracy of 90.4%, significantly outperforming traditional manual methods (82.1%, $P < 0.05$). This aligns with previous studies [4], but the incorporation of LSTM in our model enhanced its ability to process time-series data, capturing dynamic changes in patient symptoms. Compared to widely used risk assessment tools in Western medicine—such as the Framingham Risk Score [5]—our model integrates the TCM "four diagnostic methods" (inspection, auscultation/olfaction, inquiry, and palpation), highlighting the unique advantage of individualized evaluation in TCM. Nevertheless, the interpretability of deep learning models remains a limitation and should be improved to enhance trust and clinical applicability.

The personalized intervention planning module, supported by a Neo4j-based knowledge graph and a hybrid algorithm, significantly reduced plan generation time (1.5 ± 0.3 min vs. 10 ± 2.6 min, $P < 0.001$) and improved patient satisfaction with the proposed interventions (92.6%). This finding is consistent with existing research on the use of knowledge graphs in chronic disease management [6], demonstrating that structured knowledge repositories can effectively support precision interventions. Our innovation lies in integrating classical TCM literature with contemporary clinical guidelines, thereby addressing the subjectivity and time demands of manual plan development. However, the comprehensiveness of the knowledge graph depends on the diversity and completeness of source materials. Future improvements should incorporate more region-specific and multisource TCM data to enhance generalizability.

The remote follow-up module, implemented via WebSocket and NLP technologies, significantly increased the follow-up completion rate (92.5% vs. 69.3%, $P < 0.001$). NLP-enabled structuring of patient-reported symptoms reduced the documentation burden for clinicians. However, the algorithm's accuracy is currently limited by dialectal variations and non-standard expressions; further optimization is needed to accommodate diverse patient populations.

The real-time data monitoring module, powered by ECharts and Tableau, enabled visual presentation of service indicators and significantly enhanced the timeliness of quality feedback compared to traditional paper-based methods (93.8% vs. 80.6%, P = 0.004). This is in line with the growing trend of data visualization in healthcare informatics [7], underscoring the potential of data-driven decision-making in optimizing service delivery.

This study also found that the improvement in TCM constitution scores in the digital service group (15.6 ± 3.8 points) was significantly greater than that in the traditional group (9.8 ± 3.1 points, $P < 0.001$). This may be attributed to higher patient compliance with personalized interventions and the continuous oversight provided by real-time follow-up. This finding is consistent with behavior change theories, which posit that increased patient engagement and intervention accessibility—both facilitated by digital technology—are critical for sustaining healthy behaviors. However, the open-label design may have introduced observer bias, potentially affecting the objectivity of satisfaction and compliance assessments.

Despite the promising results, this study has several limitations. First, the short follow-up period precluded evaluation of long-term health outcomes. Second, the single-center design limits the generalizability of the findings, and the sample size (n = 216) may not have been sufficient to detect small differences in secondary outcomes. In addition, issues related to data privacy and algorithmic bias were not fully explored, which may impact the fairness and equity of the model. Future research should involve multicenter, large-sample, and long-term follow-up studies, including more diverse populations to validate the generalizability of the model. The integration of wearable devices for physiological data collection and speech recognition technology may further enhance patient interaction and engagement [8]. Moreover, the development of more interpretable AI models will be essential to increase clinical acceptance and trust.

From a clinical perspective, this study offers a replicable digital solution for the implementation of TCM-based preventive care, which can help alleviate the burden on primary care resources and improve the efficiency of chronic disease prevention and management. From a policy standpoint, the proposed model aligns with the goals of the *Strategic Plan for the Development of Traditional Chinese Medicine (2016–2030)* and may serve as a reference framework for regional health management platforms. In summary, this study demonstrates the preliminary effectiveness of integrating digital technology with TCM preventive services in optimizing healthcare delivery and promoting health. It provides important evidence to support the modernization and intelligent transformation of TCM.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

FUNDING

Shanghai Shenkang Hospital Development Center (Grant No. SHDC12024638): "Digital Qihuang, Smart Health – Comprehensive Service Process for Preventive Treatment in Traditional Chinese Medicine". And Shanghai Shenkang Hospital Development Center (Grant No. SHDC22023203): "Empirical Study on the Application and Promotion of Key AI Technologies in Intelligent Traditional Chinese Medicine Diagnosis and Treatment in Internet Hospitals".

REFERENCES

[1] X Cao, Y Wang, Y Chen, et al. Advances in traditional Chinese medicine for the treatment of chronic obstructive pulmonary disease. Journal of Ethnopharmacology, 2023, 307: 116229.

[2] I Hege, D Tolks, S Kuhn, et al. Digital skills in healthcare. GMS Journal for Medical Education, 2020, 37(6): Doc63.
[3] M D M Ladino, C Bolaños, V A C Ramírez, et al. Effects of internet-based, psychosocial, and early medical interventions on professional burnout in health care workers: Systematic literature review and meta-analysis. Internet Interventions, 2023, 34: 100682.
[4] H K Wu, Y S Ko, Y S Lin, et al. The correlation between pulse diagnosis and constitution identification in traditional Chinese medicine. Complementary Therapies in Medicine, 2017, 30: 107–112.
[5] C Iadecola, N S Parikh. Framingham General Cardiovascular Risk Score and Cognitive Impairment: The Power of Foresight. Journal of the American College of Cardiology, 2020, 75(20): 2535–2537.
[6] S K Mohamed, A Nounu, V Nováček. Biological applications of knowledge graph embedding models. Briefings in Bioinformatics, 2021, 22(2): 1679–1693.
[7] K Denecke, R May, E M Borycki, et al. Digital health as an enabler for hospital@home: A rising trend or just a vision? Frontiers in Public Health, 2023, 11: 1137798.
[8] Y Cheng, K Wang, H Xu, et al. Recent developments in sensors for wearable device applications. Analytical and Bioanalytical Chemistry, 2021, 413(24): 6037–6057.

World Journal of Information Technology
ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3036

DEEP LEARNING-ENHANCED DYNAMIC FRAME SLOTTED ALOHA OPTIMIZATION ALGORITHM

HongLing Zhang
School of Information Engineering, Zhongyuan Institute Of Science And Technology, Xuchang 461113, Henan, China.
Corresponding Email:19837698961@163.com

Abstract: In recent years, Radio Frequency Identification (RFID) technology has seen expanding applications across both industrial production and daily life, driving growing demand for efficient tag reading systems.When faced with a large numeral of tags, the Dynamic Framed Slotted ALOHA (DFSA) algorithm keeps the throughput at a high position for most of the time. The chief principle is to dynamically adjust the frame length based on the response consequence of the triumphant transmission of the earlier frame, so that during the data container transmission process, the length of each frame is close to the number of data packets that need to be transmitted within the range that the information transmission system can receive. Based on the analysis of traditional algorithms, this paper proposes a new tag number estimation algorithm. This algorithm based on Transformer, residual connections, and Multi-Layer Perceptron (MLP), combined with algorithms such as tag grouping techniques. Compared with traditional algorithms, the algorithm proposed in this paper addresses the shortcomings of the traditional ALOHA protocol in dynamic frame slot adjustment, collision avoidance, and throughput optimization by introducing deep learning. It significantly improves the effectiveness and reliability of RFID systems and is able of maintaining a high data storage rate even in scenarios with large amounts of data.
Keywords: RFID; DFSA; MLP-Transformer

1 INTRODUCTION

RFID is a non-contact automatic discovery technology that uses radio recurrence signals to automatically recognize targets and retrieve associated data. RFID systems are widely applied in logistics, manufacturing, healthcare, and other fields, their performance is often hindered by tag collision problems, which significantly reduce identification efficiency[1-8]. ALOHA-based algorithms are widely adopted due to their simplicity, but they suffer from low throughput and high collision rates under heavy tag loads. In contrast, binary tree algorithms offer better collision resolution but are computationally complex, making them impractical for real-time systems.To solve these limitations, this paper proposes an improved ALOHA algorithm that dynamically adjusts frame time slots using deep learning. Our method enhances identification efficiency by predicting optimal slot allocation, reducing collisions, and maintaining low computational overhead. This innovation bridges the gap between performance and practicality in large-scale RFID deployments. Among these, research on the performance of RFID anti-collision algorithms [9] is particularly remarkable. Currently, widely used anti-collision identification algorithms can be divided into inflexible allowance types and dynamic allocation types. Fixed allocation types include Time Division Multiple Access (TDMA), Frequency Division Multiple Access (FDMA), and Code Division Multiple Access (CDMA). Dynamic allocation types can be further categorized into contention-based access, reservation-based access, and polling-based access.

In RFID systems, the primary approaches to address tag collision issues fall into two distinct categories: predetermined channel assignment methods and adaptive channel contention strategies. The former involves preallocating dedicated communication resources to individual tags, eliminating the unpredictability associated with contention-based mechanisms. Nevertheless, this approach demonstrates inefficiency in resource utilization during periods of low activity and exhibits limited flexibility when confronted with unexpected service demands. On the other hand, adaptive contention strategies enable tags to opportunistically access available channels according to established contention parameters, offering enhanced responsiveness to fluctuating tag populations. Within this category, protocols derived from the ALOHA framework have gained widespread adoption owing to their straightforward implementation and seamless integration with existing system architectures. Regarding tag population estimation, conventional approaches encompass the Q-value assessment technique, the Schouten analytical method, and the Vogt iterative algorithm. The Q-value approach determines tag quantities by monitoring response attempts, providing computational simplicity at the expense of precision. Schoute's methodology, integrated with the Dynamic Frame Slotted ALOHA framework, approximates tag populations by analyzing collision slot proportions. The Vogt algorithm, while achieving optimal theoretical accuracy through iterative comparison between empirical observations (including empty slots, successful transmissions, and collision events) and their mathematical expectations, suffers from substantial computational overhead. This characteristic notably diminishes operational effectiveness, especially in scenarios involving large-scale tag deployments.

Traditionally, Deep learning techniques have been widely used in such problems. Reference [10] is based on the energetic mounting slotted algorithm and BP neural network. By processing the dataset of tag quantity, it establishes the mapping relationship between the reader and the tag quantity. This algorithm improves the system efficiency without sacrificing its accuracy[10]. Reference [11] is based on the dynamic frame slotted ALOHA algorithm (D-G-MFSA)

with tag grouping and long short-term memory (LSTM). It regards the tag quantity as a time series, uses LSTM for real-time prediction, dynamically adjusts the frame length, and groups tags when the tag quantity is large. This algorithm promotes the stability of the system when the tag quantity is large, and has the advantages of simple principle and high reading efficiency[11]. Reference [12] combines the dynamic frame slotted ALOHA (DFSA) algorithm with Transformer and long short-term memory (LSTM) neural networks. This algorithm ensures more accurate tag quantity prediction, reduces the time consumption of the reading system, and improves the system throughput[12].

This paper trains a neural network based on residual connections, a multi-layer perceptron (MLP), and a Transformer encoder, combined with the tag grouping algorithm, to predict the number of tags.The main contribution of this paper is as follows:

(1) Residual connections accelerate training convergence, and the parallel computing advantages of Transformer improve computational efficiency.

(2) By incorporating tag grouping, the algorithm achieves scalability for high-density tag environments.

(3) Experimental results indicate that, compared to the traditional BP network algorithm, the proposed manner accomplish higher throughput and consumes fewer total time slots.

(4) MLP is combined with residual connections, it maintains stable gradient norms during backpropagation, resulting in a 3x faster training convergence speed compared to traditional BP networks.

(5) The DFSA algorithm's dynamic frame length adjustment mechanism estimates the number of tags in real-time and automatically optimizes the frame size, improving the theoretical system throughput. Additionally, its time slot grouping technique further reduces collision probability.

2 DFSA ALGORITHM

2.1 Algorithm Principle

The pure ALOHA protocol serves as a fundamental multiple access control mechanism in wireless communication systems.The ALOHA protocol is characterized by the fact that any station can transmit immediately after a slot is generated and determines whether the transmission is successful by detecting signal feedback on the channel. When two or more stations transmit data simultaneously, a collision occurs, resulting in a high collision rate and low efficiency. SALOHA improves the throughput of the ALOHA system by dividing time into fixed slots, where tags can only transmit data at the beginning of a slot, thereby reducing collisions. However, the frame length is fixed, leading to limited resource utilization. The FSA protocol further enhances the time division dimension. The DFSA protocol dynamically adjusts the frame length based on the successful transmission responses from the previous frame, ensuring that the length of each frame is close to the optimal value during data packet transmission. This significantly improves throughput and resource utilization.

2.2 Mathematical Analysis

The fundamental principle of dynamically modifying the frame length in DFSA (Dynamic Frame Slotted ALOHA) lies in establishing the correlation between the current number of unidentified tags and the most efficient frame size. In this study, we define the frame length as **N**, while the total number of tags within the reader's detection range is denoted as **M**. Each tag has an equal and independent probability of selecting any given time slot within the frame. This assumption ensures that tag transmissions are uniformly distributed across the available slots, facilitating effective collision management and frame length optimization. Let its probability be $\mathrm{P}=\frac{1}{\mathrm{L}}$, Thus the quantity of arriving tags X in each time slot conforms to the binomial distribution $\mathrm{X}\sim\mathrm{B}(\mathrm{N},\frac{1}{\mathrm{L}})$, Then the probability of having a tag in a time slot is:

$$P_x=\binom{k}{x}\left(\frac{1}{M}\right)^x\left(1-\frac{1}{M}\right)^{k-x} \tag{1}$$

Where there are no tags in the time gap, the corresponding probability is as follows ($\mathrm{P_{idle}}$), When there is a single tag in the time gap, the corresponding probability is as follows ($\mathrm{P_{succ}}$). When there are multiple tags in the time gap, the corresponding probability is as follows ($\mathrm{P_{coll}}$). These probabilities can be expressed as:

$$P_{\text{idle}}=\left(1-\frac{1}{M}\right)^k \tag{2}$$

$$P_{\text{succ}}=k\times\frac{1}{M}\times\left(1-\frac{1}{M}\right)^{k-1} \tag{3}$$

$$P_{\text{coll}}=1-P_{\text{idle}}-P_{\text{succ}} \tag{4}$$

The expected values for the number of idle time slots, successful time slots, and collision time slots within a frame can be derived based on the given parameters. Let the frame length be N and the total number of tags be M. Assuming each tag independently and uniformly selects a time slot, the probabilities and expected values can be calculated as follows:

$$y_{\text{idle}}=k\times P_{\text{idle}} \tag{5}$$

$$y_{\text{succ}}=k\times P_{\text{succ}} \tag{6}$$

$$y_{\text{coll}}=k\times P_{\text{coll}} \tag{7}$$

Clearly, the estimated number of remaining unrecognized tags $k_{estimate}$ is:

$$k_{\text{estimate}} = k - y_{\text{succ}} \tag{8}$$

According to the formula, when the number of tags k is 1-4, a frame length M=4 is sufficient to cover the number of tags and reduce collisions. When the number of tags k>4, using M=4 would lead to increased collisions, thus requiring a larger frame length. When the number of tags k is between 5 and 10, a frame length M=8 can effectively reduce collisions. That is, as the number of tags k increases, the frame length gradually increases. The theoretical analysis and experimental validation results based on the dynamic frame slotted algorithm are shown in Table 1.

Table 1 Relationship between frame size and label number of dynamic frame time slot algorithm

Frame length	4	8	16	32	64	128	256
Number of tags	1~4	5~10	11~22	23~44	45~88	89~177	Above 177

2.3 Tag grouping Algorithm

The Tag Grouping Algorithm is a technique used in Radio Frequency Identification (RFID) systems, aiming to optimize the label discovery process by grouping labels, thereby reducing collisions and improving system effectiveness. Formula (9) is used to calculate the number of tags after grouping, where a and b are the number of tag groups, and N is the total number of tags. This formula aims to determine the minimum and maximum number of tags in each group under different grouping numbers. For different grouping numbers, the minimum number of tags is obtained by dividing the total number of tags by the number of groups and rounding down, while the maximum number of tags is obtained by dividing the total number of tags by the number of groups and rounding up.

$$\frac{\frac{N}{a}}{256} \times \left(1 - \frac{1}{256}\right)^{\frac{N}{a}-l} = \frac{\frac{N}{b}}{256} \times \left(1 - \frac{1}{256}\right)^{\frac{N}{b}-l} \tag{9}$$

When a = 1 and b = 2, substituting into formula (9) yields a critical value N of 355.With 2 groups, the tag capacity increases to 356-709. By changing the values of parameters a and b in formula (9), different critical values for tag grouping are obtained,this demonstrates the scalable nature of the grouping algorithm, where doubling the group count approximately doubles the maximum tag capacity, maintaining consistent range intervals of ~355 tags per grouping threshold. The results are shown in Table 2.

Table 2 The Correlation between the Count of Groups and the Quantity of Labels

Number of groups	1	2	4	8	...
Minimum number of tags	1	356	710	1418	...
Maximum number of tags	355	709	1417	2834	...

3 METHOD

The Transformer model is a deep learning architecture. The Transformer model is a deep learning model based on the attention mechanism, abandoning the use of CNN and RNN in previous deep learning tasks. This model exclusively utilizes the attention mechanism to model global relationships between inputs and outputs. Unlike traditional recurrent neural networks (RNNs) that process data sequentially, it enables parallel processing of sequence data while effectively capturing long-distance dependencies through its attention-based architecture.

The calculation for dot-product attention is as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{10}$$

Where Q is the query matrix, K is the key matrix, V is the value matrix, and dk is the dimension of the keys.

The multi-head attention mechanism divides the input into multiple parallel attention heads.Each head computes scaled dot-product attention independently, then the outputs are concatenated and linearly transformed. The computation for each head is given by:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{11}$$

$$\text{head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{12}$$

This allows the model to focus on different subspaces and capture more features and information.

This paper also uses residual connection. The core idea is: instead of learning an underlying mapping (H(x)) directly, the network learns the residual (F(x) = H(x) - x). This makes it easier to optimize because fitting a residual to zero (for identity mapping) is simpler than learning a new function from scratch, especially for deeper layers.

The input x undergoes dimensionality expansion through a completely linked layer. The MLP module processes the features via two fully connected layers (with ReLU activation), and the output is added to implement a residual connection. This feature-enhanced data is then fed into the core Transformer encoder (a 2-layer stacked structure) where each layer consists of a multi-head self-attention mechanism for capturing dependencies between time slots and a position-wise feedforward network. Throughout the Transformer processing, the tensor dimension remains (batch_size,

sequence_length, hidden_size). Finally, the encoder output is flattened and projected through a fully connected output layer to obtain a prediction result with a dimension of 1.The transformer processes the self-attention weight matrix to explicitly capture temporal slot correlation patterns, while the multi-head mechanism learns multiple dependency relationships in parallel, improving the accuracy of label quantity prediction.The flowchart is shown as Figure 1.

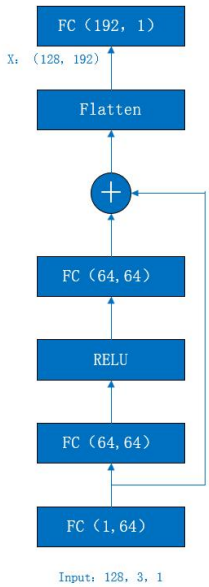


Figure 1 Neural Network Structure

The specific steps of the proposed algorithm are as follows:
1. Data Calculation and Preparation
Using Equations (5) to (8), compute the number of idle slots y_{idle}, successful slots y_{succ}, collision slots y_{coll}, and the estimated number of remaining tags $k_{estimate}$ for different tag quantities k (assuming $k \in [0,\ldots,5M]$) and frame lengths M. These results are compiled into the datasets $T = \{(k, M, y_{idle}, y_{succ}, y_{coll}) | k_{estimate}\}$.
2. Network Training
The datasets T is fed into the network for training. Once the network metrics converge, the corresponding network model G for each frame length M is obtained.
3.Real-Time Prediction and Adjustment
During the reader's current reading cycle, the actual detected number of tags (k), idle slots y_{idle}, successful slots y_{succ},and collision slots y_{coll} are input into the corresponding network model G to predict the remaining number of tags $k_{estimate}$. The reader then dynamically adjusts the frame length for the next reading cycle based on the rules in Table 1.

4 TRANSFORMER-OPTIMIZED DFSA ALGORITHM

4.1 Model Training

During the model training phase, we adopted the following configurations: the parameter update step size (learning rate) was set to 0.01, with the maximum number of training epochs limited to 150. To enhance the model's nonlinear representation capability, we employed the ReLU (Rectified Linear Unit) activation function in the network. Other key hyperparameter settings included: 150 iterations, Lr regularization coefficient of 0.001, batch size of 128, the Adam optimizer algorithm, and Mean Squared Error (MSE) as the loss function for optimization objectives.
Finally, as shown in Figure 2, the training results indicate that:
Initial Phase (Epoch 0–25): Both the train loss (blue) and test loss (orange) start high but drop sharply. This rapid decrease indicates that the model quickly learns essential patterns in the early training stages.Mid - Training Phase (Epoch 25 – 150): The losses continue to decline but with increased fluctuations. These variations arise from the stochastic nature of batch - based training—each batch's unique data introduces minor instability. Despite this, both curves trend downward, showing consistent learning.Final Phase (Epoch 150–200): By epoch 200, both losses stabilize near \(10^{-3}\). The close alignment of train and test loss throughout training suggests no significant overfitting. If overfitting occurred, the test loss would diverge upward while the train loss continued decreasing. Here, their proximity indicates the model generalizes well to unseen data. Overall, the curve demonstrates effective training: rapid initial learning, steady improvement, and good generalization, as evidenced by the parallel trends of train and test loss.

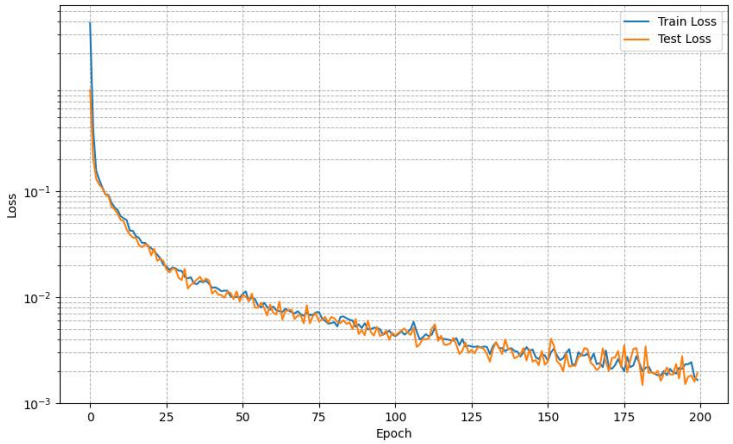


Figure 2 Training Loss with Epoch

4.2 Algorithm Performance Analysis

The primary metrics for evaluating tag reading efficiency consist of two key dimensions.The definition of system throughput is the ratio of the number of successfully identified tags to the total number of slots consumed in a single read cycle.In Radio Frequency Identification (RFID) systems, throughput typically refers to the number of tags successfully read per unit of time.，reflecting the efficiency and capability of the system in processing data. As shown in Figure 3, as the number of tags increases, the system throughput initially rises and then declines. The throughput reaches its peak when the number of tags is around 200. After this peak, as the number of tags continues to increase, the system throughput shows a downward trend. Following the peak throughput, the decline in throughput is significantly faster in the BP algorithm compared to the TLE algorithm as the number of tags increases. When the number of tags reaches around 1200, the throughput in the TLE algorithm begins to increase again, until the number of tags reaches around 1400, after which the throughput starts to decline once more.

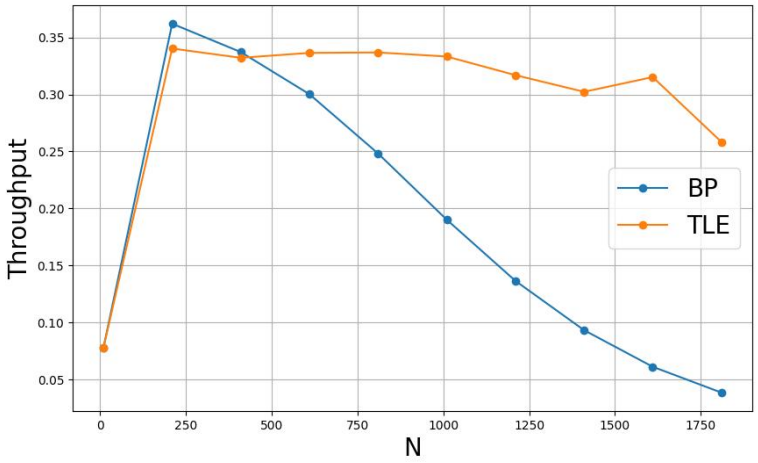


Figure 3 Comparison of the Throughput

As shown in Figure 4, the number of slots consumed by the algorithm proposed in this paper is lower than that of the traditional BP neural network algorithm. The total number of consumed slots refers to the total number of slots used for tag identification in a single read cycle of a Radio Frequency Identification (RFID) system. It is one of the key metrics for measuring system resource consumption, directly impacting the system's throughput and efficiency.As shown in Figure 4, when the number of tags increases, the total number of consumed slots for both algorithms also increases. When the number of tags is less than 600, the total number of consumed slots for both algorithms is approximately the same. However, when the number of tags exceeds 600, the total number of consumed slots for the BP algorithm increases significantly faster than that for the TLE algorithm.

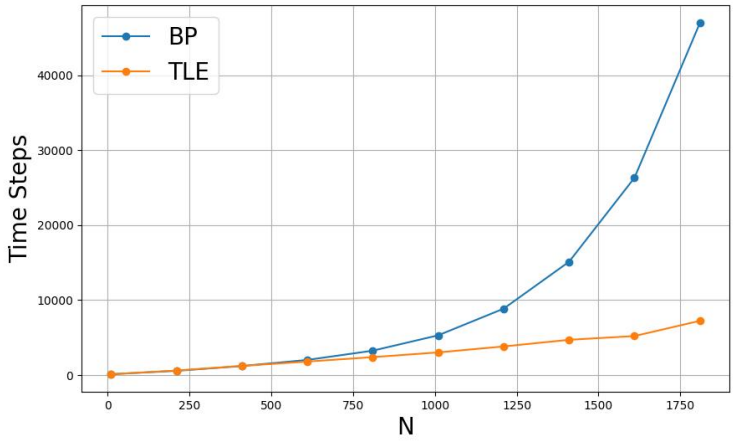


Figure 4 Comparison of the Number of Consumed Slots

5 CONCLUSIONS

This study proposes an RFID tag identification optimization algorithm based on a Transformer architecture. By introducing a hybrid "Transformer + Residual MLP" encoding structure, the multi-head self-attention mechanism explicitly models the nonlinear relationships between tag time slots, while residual connections enhance training convergence speed. Compared with conventional algorithms, the proposed deep learning-based approach addresses key limitations of traditional ALOHA protocols in dynamic frame slot adjustment, collision avoidance, and throughput optimization, significantly improving the efficiency and reliability of RFID systems. The algorithm maintains high data storage rates even in large-scale data scenarios, providing an effective solution for performance enhancement in high-demand RFID applications.

With the advancement of IoT technology, RFID performance demands are growing. The algorithms in this paper offer robust technical support for logistics management and other fields. By combining dynamic frame length optimization with deep learning, the TLE algorithm overcomes traditional RFID efficiency bottlenecks and demonstrates cross-scenario adaptability. Future research will focus on model lightweighting, transfer learning, and multi-protocol fusion to enhance real-time responsiveness and generalization in edge computing and high-collision scenarios, driving the evolution of next-generation IoT ecosystems.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Maksimenko A, Dobrykh D, Yusupov I, et al. Miniaturization limits of ceramic UHF RFID tags. Scientific Reports, 2025, 15(1): 10984.
[2] Masekwana F, Jokonya O. Factors affecting the adoption of RFID in the food supply chain: A Systematic Literature Review. Frontiers in Sustainable Food Systems, 2025, 8: 1497585.
[3] Claucherty E, Cummins D, Aliakbarian B. RFID Unpacked: A Case Study in Employing RFID Tags from Item to Pallet Level. Electronics, 2025, 14(2): 278.
[4] Wu Y, Lin J, Chen H, et al. A transformer-based double-order RFID indoor positioning system. Expert Systems with Applications, 2025: 126530.
[5] Maimouni M, Abou El Majd B, Bouya M. Optimising RFID network planning problem using an improved automated approach inspired by artificial neural networks. Information Sciences, 2025: 121927.
[6] Lasantha L, Ray B, Karmakar N. Trade-Off Analysis for Array Configurations of Chipless RFID Sensor Tag Designs. Sensors, 2025, 25(6): 1653.
[7] Wang Z, Mao S. Generative AI-Empowered RFID Sensing for 3D Human Pose Augmentation and Completion. IEEE Open Journal of the Communications Society, 2025.
[8] Lazaro A, Cujilema M R, Villarino R, et al. A novel approach for wine anti-counterfeiting using laser-induced graphene chipless RFID tags on cork. Scientific Reports, 2025, 15(1): 12750.
[9] Liu H, Meng Z, Li C, et al. Modeling for Phase Decoupling to Detect the Orientation and Position of Moving Objects with Simple RFID Array. IEEE Transactions on Instrumentation and Measurement, 2025.
[10] Ming Dongyue, Wang Shangpeng, Lei Ming, et al. Improved Dynamic Framed Slotted ALOHA Algorithm Combined with BP Neural Network . Mini-Micro Systems, 2021, 42(09): 1920-1923.

[11] Li Z, Li Z. An RFID anti-collision algorithm integrated with LSTM. 2024 IEEE 4th International Conference on Power, Electronics and Computer Applications (ICPECA), IEEE, 2024: 985-989.
[12] Zhou Q. ALOHA Improvement Algorithm for Dynamic Frame Time Slots with Deep Learning. 2024 IEEE 4th International Conference on Data Science and Computer Application (ICDSCA), IEEE, 2024: 671-676.