Volume 3, Issue 4, 2025

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

World Journal of Information Technology

World Journal of Information Technology

Volume 3, Issue 4, 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
Print ISSN: 2959-9903
Online ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3047

MEDAL PREDICTION FOR THE LOS ANGELES OLYMPIC GAMES BASED ON BAYESIAN OPTIMIZED BOOST REGRESSION

YiHan Gong
School of Statistics and Data Science, Qufu Normal University, Qufu 273165, Shandong, China.
Corresponding Email: yhgong1203@163.com

Abstract: To support resource allocation and strategic preparation for the 2028 Olympic Games, this study introduces a Bayesian-optimized Boost framework. First, key variables are extracted through feature engineering, including economic indicators such as GDP, population, host country effects, and recent rule changes in events. Next, Bayesian optimization is used to fine-tune the model's hyperparameters, and multiple metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2) are employed to evaluate the model's performance. Then, time-series cross-validation based on 2024 data shows that the model achieves an R^2 of 0.91, outperforming baseline models such as decision trees and standard Boost. Predictions indicate that 23 delegations will gain more medals, while 46 countries may see a decline. Finally, k-means clustering is used to identify each country's dominant sports, quantify the impact of the host country effect on event selection, and provide data support for preparation.
Keywords: Olympic medal forecasting; Feature Engineering; Bayesian Optimization; k-means

1 INTRODUCTION

Olympic medal tallies mirror competitive strength; accurate forecasts aid strategic preparation. Multidimensional machine-learning models dominate current research yet still lack completeness and precision.

Early Olympic-medal forecasts centered on macroeconomics and demographics: Bernard & Busse [1] used population and GDP to show real GDP is a key medal predictor, and Zhang [2] confirmed higher GDP boosts medal odds via better facilities and coaches, yet both assumed linearity and ignored event specifics and host effects. Tian [3] emphasized event popularity, technical thresholds and rule changes as medal drivers, opening a new lens. Recent domestic work by Shi, Shi & Zhang [4] leveraged interpretable ML to show table tennis and badminton are highly predictable owing to top-team dominance, whereas football and shooting are not, but they still omitted event-host interactions. Luo & Cheng [5] demonstrated hosts can tilt outcomes by adjusting event schedules and environments, an interaction absent from traditional models. Machine-learning advances-e.g., Sadu et al.'s [6] Boost ensemble-integrate economic, demographic and historical data via non-linear, high-dimensional learning, outperforming linear regressions, yet Boost's efficacy hinges on hyperparameters; grid or random search is computationally heavy and prone to local optima, curtailing its full potential.

Addressing prior omissions, this study integrates economic, event-specific and host-region factors within a Bayesian-optimized Boost framework to forecast the 2028 Los Angeles medal distribution and deliver refined guidance for strategic preparation.

2 RESEARCH METHODS

2.1 Data Acquisition and Preprocessing

The data used in this study are drawn from the publicly available datasets released by the International Olympic Committee (IOC), covering the complete competition records from the 1896 Athens Games through the 2024 Paris Olympics.

2.1.1 Data cleaning

(1) Events that are for demonstration purposes are excluded from the model.

(2) Since 1924, the sports "Skating" and "ce Hockey" have been removed from the Summer Olympics. Out of respect for historical facts, these two sports are also excluded from the model.

(3) To ensure the rigor and scientific integrity of the model, this study remove the data from the 1906 Athens Olympics. Although this edition of the Games represented a noble pursuit of the Olympic spirit, it was not officially recognized as part of the Olympic series, and its results were not officially recorded. Thus, it has significant deficiencies in terms of data continuity and comparability [2].

2.2.2 Data merge

(1) For anomalies in the dataset ″summer Oly-athletes″ caused by changes in event names, such as ″Region1-1″ and ″Re gion1-2,″ this study unified them under the country code ″Region1.″

(2) During the integration of multiple datasets, this study found discrepancies between some athletes′ medal records and the national medal allocation. Based on the principle that each athlete can only win one medal per event, we recalculated the historical medal counts for each country to ensure data accuracy and consistency.
(3) The data were merged to count the number of athletes and the number of events for each country in each Olympics, without distinguishing between participating teams.

2.2 Method Introduction

Guided by medal prediction, this study follows a "clean–model–validate" pipeline: preprocess data, tune and optimize models per task, cross-validate, then forecast the 2028 Los Angeles Games and assess effects.
(1) Boost Regression Model Based on Bayesian Optimization
This model is the core method used in this study for predicting the number of Olympic medals. Its core idea is to capture the nonlinear relationships in the data through Boost regression and combine it with the Bayesian optimization algorithm to improve model performance. Boost regression iteratively constructs multiple decision trees as weak learners, and each tree fits the residuals of the current model to gradually optimize the prediction effect. Its goal is to minimize the MSE, thereby reducing the deviation between predicted values and true values. Bayesian optimization estimates the objective function of the hyperparameter space by constructing a surrogate model, efficiently searching for the optimal combination of hyperparameters. Compared with traditional grid search or random search, it can more accurately improve the generalization ability of the model. The advantages of this model are: first, it is suitable for prediction tasks with multiple features and complex nonlinear relationships, and can effectively integrate multi-dimensional features such as the host country effect, the number of participating events, and the number of athletes; second, it realizes automatic hyperparameter tuning through Bayesian optimization, reducing errors caused by manual intervention and improving prediction accuracy. Relevant studies have shown that similar gradient boosting models have high accuracy in Olympic medal prediction and can effectively capture the potential patterns in historical data[7].
(2) k-means Clustering Model
To explore the relationship between a country′s competitive strength and sports events, this study uses the k-means clustering algorithm to analyze medal data from the past three Olympic Games. This algorithm is an improved version of k-means clustering. It improves the stability and accuracy of clustering results by optimizing the selection method of initial clustering centers. In the clustering process, with countries, event categories, and medal results as key features, the data are divided into different clusters, thereby identifying different types of national groups such as ″focusing on specific events″, ″balanced development″, and ″strong in multiple events″. Its advantage is that it can intuitively reveal the distribution law of countries' advantages in sports events, providing a basis for analyzing the impact of the host country's event selection on medal distribution. Similar clustering methods have been proven effective in mining the laws of Olympic data[8].

2.3 Model Evaluation Metrics

To comprehensively evaluate the performance of the Olympic medal prediction model, MAE, MSE, RMSE, and R^2 are employed for quantitative analysis; detailed calculation formulas can be found in reference[9].

3 MODEL ESTABLISHMENT AND SOLUTION

3.1 Feature Engineering for Key Predictive Variable Extraction

Based on the results of exploratory data analysis (EDA), this study selected the following features as input variables for modeling.
(1) Host: Whether the country is the host nation The "host nation effect" was proven in exploratory analysis. Host countries usually enjoy a home advantage, which may lead to better performance in competitions. Additionally, the host country may be granted extra spots in certain events [9]. Therefore, this study set Host as a binary dummy feature, indicating whether the event is held in the athlete's home country. Host=1 represents the host country, and Host=0 represents a non-host country.
(2) Events: The number of events each country participates in. The number of events a country participates in reflects its level of activity and diversity in the Olympics. Coun- tries that participate in more events typically have stronger competition and more opportunities to win medals in various sports.
(3) Athletes: The number of athletes each country sends Although a total of 158,427 athletes participated in 292,942 events from 1896 to 2024, 78% of athletes never stood on the podium. 17% of athletes won only one medal during their Olympic careers, while the top 5% of athletes earned medals in multiple events. Previous Olympic prediction models did not use athletes' physical attributes, as these had low correlation with medal outcomes. Thus, we treat the total number of athletes in each country as a feature in our model, rather than focusing on individual athletes or teams. The number of athletes directly reflects a country's investment and talent pool for the Olympics. Countries with more athletes usually have more opportunities to win medals.

(4) Athletes-per-Event: The average number of athletes per event for each country. A country's resource allocation to sports may vary. By analyzing the average number of athletes per event, we can indirectly understand the country's focus and resource distribution in sports events, reflecting its investment in sporting competitions.
(5) Region-coded dummy variables: Categorical region identifiers are converted into numerical form through dummy encoding. A binary column is created for each unique code in the "NOC" column. For instance, if the "NOC" column contains Region 1, Region 2, and Region 3, new columns-NOC-1, NOC-2, NOC-3-are generated, where 1 indicates membership of the corresponding region and 0 indicates non-membership.

3.2 The Establishment and Case Analysis of the Olympic Medal Table Prediction Model

3.2.1 Bayesian optimization boost regression model

(1) Boost Regression Model Boost regression is a regression method based on gradient boosting trees. Boost regression iteratively constructs a strong regression model consisting of multiple weak regression models. Each weak regression model is a decision tree, and it improves the model's predictive ability by fitting the residuals of the current model. Boost regression optimizes the model by minimizing MSE, thus minimizing the difference between the predicted and true values.
(2) Bayesian Optimization of Model Hyperparameters Bayesian optimization is a method used to optimize model hyperparameters. The key to Bayesian optimization is using a surrogate model to estimate the objective function of the hyperparameter space, so that in each iteration, the hyperparameter combination most likely to improve performance is selected. Compared to traditional grid search or random search, Bayesian optimization can more efficiently explore the hyperparameter space, thereby finding better hyperparameter combinations. The main principle and process of the Bayesian optimization Boost regression model we developed are shown in Figure 1.

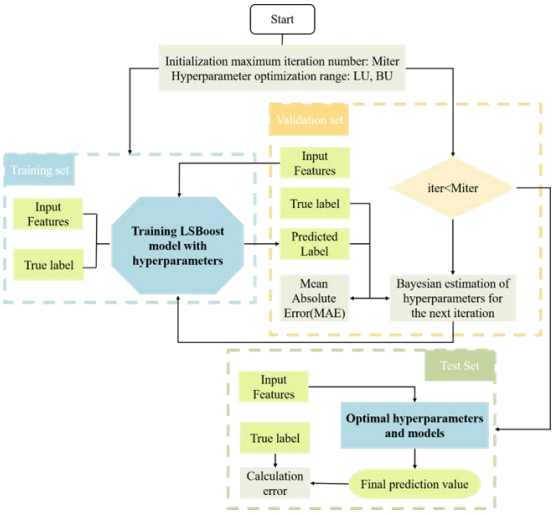

Figure 1 Bayesian Optimization Boost Model Flow Chart

(3) Model Evaluation We evaluate the model's performance using the test dataset, and we use the following performance metrics to assess the accuracy of the predictions: MAE, MAPE, MSE, RMSE, and R2.
(4) Probability Interval Prediction Metrics We use the following metrics to evaluate the accuracy of the probabilistic interval prediction model: PICP, PINAW, penalty coefficient CWC, MPICD, and AIS.

3.2.2 Model training and example analysis

This study decided to limit the training data to records from the 1992 Barcelona Olympics onwards, as there may have been significant differences before the dissolution of the
Soviet Union. This study do not need to scale the features because the sensitivity of our target variable will be incorporated into the feature coefficients.
This study aim for our Bayesian optimized hyperparameter Boost regression model to generalize to the 2028 Los Angeles Olympics. Therefore, we should divide the data records into training, validation, and test sets for model validation. This study use the 2024 Paris Olympics as our test set, which contains 206 records out of 1796 national participation records. The test set represents 11.4763% of the entire dataset.
Our target variable is the total number of medals each country wins at a single Olympics, but our dataset also contains the results for gold, silver, and bronze medals. Therefore, this study do not directly predict the total medal count; instead, this study predict each medal type separately three times. We can then sum the gold, silver, and bronze medals to obtain the total medal count. Finally, this study can use the trained model to predict medal counts for previously unseen data.
This study create training, validation, and test sets to balance bias and variance in the ma- chine learning model. We can test whether the model is overfitting or underfitting by comparing how well it explains the variance in the training and test sets. Similarly, this study expect the error levels based on prediction residuals to be comparable between the two datasets.

This study evaluate the performance of the regression model by predicting the total number of medals for the 2024 Paris Olympics (our held-out test set) and comparing it with different regression models using evaluation metrics. This comparison illustrates the good performance of the Bayesian optimized Boost model, with predictions shown in Figure 2. Further visualization in Figure 3 presents the overall distribution of predicted versus true medal counts in the test set, intuitively reflecting the model's ability to capture the variation in medal outcomes across different countries. The close alignment between predicted values and true labels in the figure further validates the model's reliability for Olympic medal forecasting.

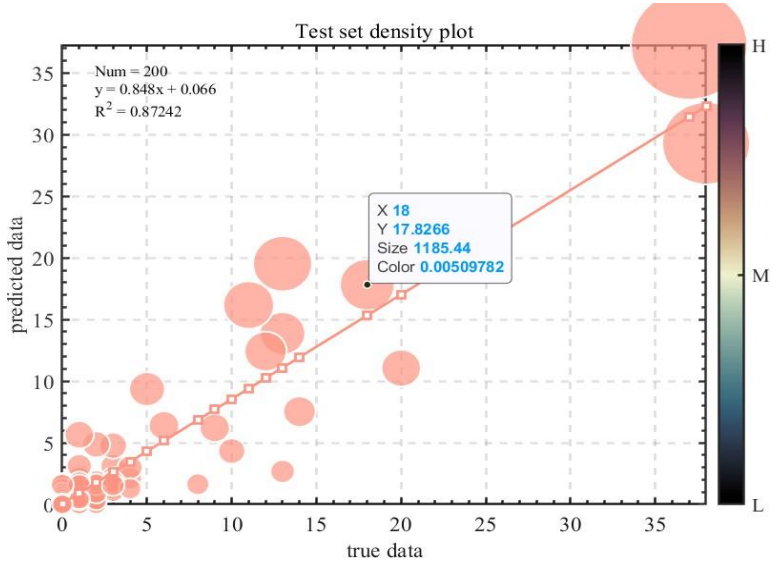

Figure 2 Test Set Density Plot

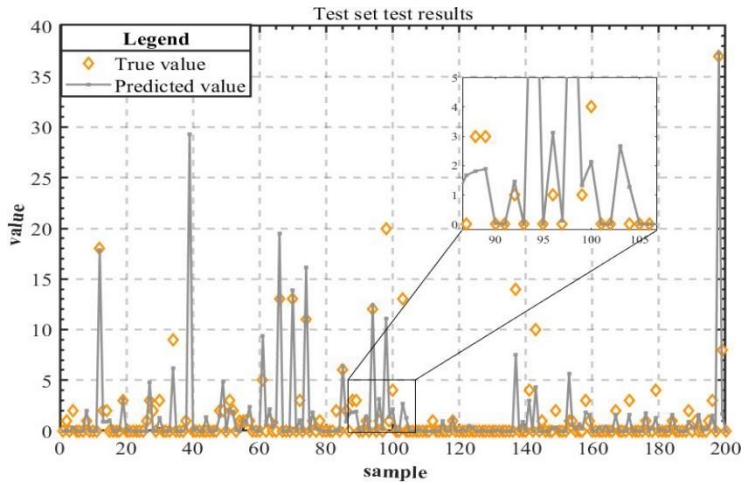

Figure 3 Test Set Effect Diagram

As shown in Table 1, our regression model's coefficient of determination R2 for the training set is 96%, and for the test set, it is 90%. Due to space limitations, we do not include the validation set results in the main text. The model error calculated using RMSE is 1.14 medals per country for the training set and 1.77 medals per country for the test set. Therefore, our regression model is balanced, capable of explaining the data, and has a small error.

Table 1 Comparison of Different Regression Model Evaluation Indicators

Dataset	Metrics	DT	Boost	Bayesian optimization Boost
Training Set	MAE	0.36	0.16	0.44
	MSE	0.94	0.08	1.29
	RMSE	0.97	0.28	1.14
	R2	0.98	0.99	0.97
	MAE	0.79	0.84	0.73
	MSE	3.70	4.01	3.12

Test Set	RMSE	1.92	2.00	1.77
	R2	0.89	0.90	0.91

3.3 Predictions and Analysis of the 2028 Los Angeles Olympic Medal Table

3.3.1 Construction of the 2028 dataset

To use Bayesian-optimized hyperparameter Boost prediction model to fore- cast the medal rankings for the 2028 Los Angeles Olympics, this study first need to update the parameters of the dataset. This study will set the United States as the host country and estimate the total number of athletes and events for each country based on the averages from the 2016 Rio Olympics, 2020 Tokyo Olympics, and 2024 Paris Olympics.

3.3.2 Prediction evaluation and results analysis

The prediction results indicate an overall increase in total medals for the 2028 Olympics, aligning with official announcements from the Los Angeles Olympic Committee regarding the addition of five new sports. This consistency validates the model's ability to incorporate real-world changes in event structures.

(1) Shifts in Medal Rankings

As visualized in Figure 4, significant upward momentum is expected for regions such as Region2, Region3, Region1, Region4, and Region5. These countries have implemented strategic investments in sports infrastructure over the past decade, upgraded athlete training systems with advanced technologies, and established youth development pipelines-efforts that are now poised to yield tangible results. Japan, for instance, has sustained its post-Tokyo Olympics focus on nurturing young talent, particularly in sports like gymnastics and swimming, making it a strong contender for increased medal hauls in 2028.

In contrast, traditional sports powerhouses including Region6, Region7, Region8, and Region9 may face a downturn in medal counts. The United States, while retaining its top position, is projected to see declines in sports like basketball and athletics due to rising global competition-particularly from European and Asian nations-and ongoing issues with unequal resource distribution across domestic sports federations. China, amid efforts to reform its sports system toward more market-driven development, may experience temporary dips in sports such as table tennis and badminton, where international rivals have narrowed the gap, but is expected to maintain a top-three ranking through consistent performance in diving, weightlifting, and shooting. Similarly, Australia and France may see reduced medals in their signature sports but remain competitive due to diversified participation across events.

(2) First Gold Medal Prediction

The 2028 predictions highlight a historic shift with Regions10 to16 set to secure their first Olympic gold medals. This breakthrough reflects years of targeted investment: these nations have focused on sports with lower global competition barriers and leveraged international coaching partnerships to accelerate athlete development. For example, Region12 has significantly improved its boxing program through collaboration with Cuban trainers, while Region15 has invested in high-altitude training facilities to boost its athletes' performance in long-distance running.

(3) Reliability of 80% Prediction Intervals

Table 2 presents the 80% prediction interval metrics for first-time gold medalists, with a PICP of 0.71 indicating that 71% of actual outcomes fall within the forecasted ranges, and a PINAW of 0.18 reflecting narrow interval widths relative to the data scale. These metrics, combined with a moderate CWC, demonstrate that the model's probabilistic forecasts strike a balance between precision and coverage-critical for guiding resource allocation in pre-Olympic preparation.

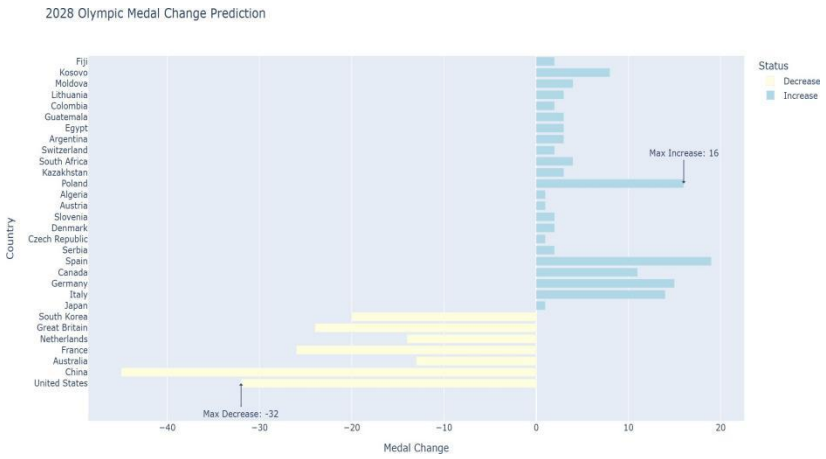

Figure 4 2028 Los Angeles Olympics Predictions

Table 2 0.8 Prediction Interval Evaluation Indicator

Metrics	Numeric
PICP	0.71

PINAW	0.18
CWC	1.35
MPICD	3.73
AIS	3.70

3.4 Medal and Event Relationship Model Based on k-means++ Clustering

Building on the 2028 medal predictions which highlighted shifts in national rankings and emerging breakthroughs, this section further explores the underlying dynamics of Olympic performance by examining how countries' medal hauls correlate with their specialization in specific events. To uncover these patterns, we employed the k-means++ clustering algorithm, leveraging medal data from the 2016 to 2024 Olympic Games.

3.4.1 Data selection and clustering methodology

The analysis focused on three core variables: unique country codes, event categories such as athletics, swimming, gymnastics, and medal counts (gold, silver, bronze) across events. Prior to clustering, data were standardized using Z-score normalization to ensure consistent scaling of variables, and the algorithm was set to partition countries into 3 distinct clusters based on their event-specific medal distributions.

3.4.2 Clustering results: national strength patterns in Olympic events

The k-means++ analysis identified three distinct clusters, each reflecting a unique pattern of national excellence in Olympic sports.

(1) Cluster 1: Specialized Dominance in Specific Events

Represented by countries such as South Korea, Switzerland, Germany, Japan, and France, this cluster is characterized by concentrated medal hauls in one or two flagship events. See Table 3 for the advantageous events and medal counts of these countries. For example, South Korea excels in archery, Switzerland in mountain bike cycling, and Japan in skateboarding. This pattern highlights targeted investment in niche sports, where these nations have developed specialized training systems and competitive advantages.

Table 3 Category 1 Countries and Advantageous Projects

Country	Cluster	Most Skilled Project	Project Award Count
KOR	1	Archery	13
SUI	1	Cycling Mountain Bike	12
JPN	1	Skateboarding	11
GER	1	Equestrianism	10
FRA	1	Handball	10

(2) Cluster 2: Focused Excellence in Single High-Impact Events

The Netherlands typifies this cluster, with exceptional performance in a single event-cycling road-where it earned an average of 9.51 medals over the past three Olympics. Unlike Cluster 1, which may span two sports, Cluster 2 nations demonstrate deep expertise in a single discipline, often with consistent podium finishes that contribute significantly to their total medal tally.

(3) Cluster 3: Multidimensional Strength Across Multiple Events

Countries like the United States, China, and the United Kingdom fall into this cluster, exhibiting balanced medal distributions across diverse events. See Table 4 for details. The U.S. dominates athletics, China leads in diving, and the UK excels in cycling. This breadth reflects comprehensive sports development strategies, with robust investment across multiple disciplines and strong talent pipelines.

Table 4 Category 2 Countries and Advantageous Projects

Country	Cluster	Most Skilled Project	Project Award Count
CHN	3	Diving	13
USA	3	Athletics	12
GBR	3	Cycling	11

3.4.3 Host country's event selection: impact on medal distribution

The clustering results further align with the observation that host countries strategically shape medal outcomes through event selection. By introducing new sports or revising rules to align with their strengths, hosts can enhance their medal potential. For instance, the 2016 Rio Olympics added surfing and golf-sports with growing participation in Brazil-while the 2020 Tokyo Olympics prioritized karate and skateboarding, where Japanese athletes had established competitive edges. Such decisions not only boost domestic interest in these sports but also directly elevate the host's medal performance.

Additionally, the structure of events-particularly the distinction between individual and team sports-significantly influences medal distribution. Variations in team sizes, competition formats, and weight categories lead to stark

differences in medal counts per event. At the Tokyo Olympics, 128 athletes participated in taekwondo, and 32 medals were awarded. Since taekwondo is an individual sport, each event awards two bronze medals. In contrast, hockey, as a team sport with nearly 400 athletes, only had six medals available. This disparity means countries excelling in high-yield individual sports are more likely to accumulate total medals-a pattern echoed in the specialization of Cluster 1 nations.

The consistency between these clustering insights and the predictive performance of our model further validates the robustness of our analytical framework. As shown in Table 1, the Bayesian-optimized Boost model outperforms decision trees and standard Boost models across key metrics in the test set, with an R^2 of 0.91-1.1% higher than the standard Boost model and 2% higher than decision trees. This superior predictive power confirms that the event-specific patterns identified by k-means++ clustering are not only statistically meaningful but also contribute to more accurate medal forecasts. Such alignment between clustering results and predictive performance strengthens the practical value of our findings for host countries' event selection strategies.

4 CONCLUSIONS

This study focuses on Olympic medal prediction and related effect analysis, adopting the technical route of "data preprocessing - model construction - validation and application". It built a solid data foundation through data cleaning and integration, then used a Bayesian-optimized Boost regression model for prediction. The model, which mines nonlinear relationships via Boost regression and achieves automatic hyperparameter tuning with Bayesian optimization, showed a high R^2 of 0.91 in 2024 data validation, outperforming benchmark models like decision trees and standard Boost. Meanwhile, the k-means clustering model analyzed medal data from the past three Olympics, identifying countries' dominant events and quantifying the host country effect on event selection. It successfully predicted the 2028 Los Angeles Olympics medal distribution, indicating increased medals for multiple regions and first golds for six regions. Future improvements may include incorporating political and cultural factors, as well as athletes' individual characteristics like age and injuries, to enhance prediction accuracy. The findings can provide a scientific basis for Olympic committees to formulate strategies and optimize resource allocation.

COMPETING INTERESTS

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

REFERENCES

[1] Bernard B A, Busse R M. Who Wins the Olympic Games: Economic Resources and Medal Totals. The Review of Economics and Statistics, 2004, 86(1): 413-417.

[2] Zhang C. A Study on the Competitive Strength and Promotion Strategies of the Chinese Delegation at the 23rd to 32nd Summer Olympic Games. Dissertation, Wuhan Sports University, 2023.

[3] Tian M. Expansion of the event-group training theory to the event-group theory. Chinese Sports Coaches, 2019, 27(01): 3-7.

[4] Shi H, Zhang D, Zhang Y. Can Olympic Medals Be Predicted? — From the Perspective of Interpretable Machine Learning. Journal of Shanghai University of Sport, 2024, 48(04): 26-36.

[5] Luo Y, Cheng Y, Li M, et al. Prediction of China's Medal Count and Overall Strength at the Beijing Winter Olympics: Based on the Host Country Effect and Grey Prediction Model. Contemporary Sports Technology, 2022, 12(21): 183-186.

[6] Sadu B V, Bagam S, Naved M, et al. Optimizing the early diagnosis of neurological disorders through the application of machine learning for predictive analytics in medical imaging. Scientific Reports, 2025, 15(1): 22488 -22488.

[7] Peng J. Research on hyperparameter optimization of Bayesian optimization algorithm based on GP. China New Technologies and Products, 2025(11): 38-40.

[8] Chen Z, Feng J, Yang D, et al. Hierarchical clustering algorithm for complex structure datasets based on hybrid neighborhood graph. Journal of Intelligent Systems, 2025(8): 1-11.

[9] Shao J. Research on Time-Series Forecasting Based on Deep Neural Networks. Dissertation, Jilin Institute of Chemical Technology, 2024.

World Journal of Information Technology
Print ISSN: 2959-9903
Online ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3048

ANALYSIS OF FACTORS INFLUENCING THE NUMBER OF OLYMPIC MEDALS BASED ON SHAP IMPORTANCE RANKING AND MACHINE LEARNING ALGORITHM

YuXuan Liu[1,2*], MengKai Zhi[1,2]
[1]*School of Electrical Engineering, Qingdao University, Qingdao 266071, Shandong, China.*
[2]*School of Automation, Qingdao University, Qingdao 266071, Shandong, China.*
Corresponding Author: YuXuan Liu, Email: liuy75059@gmail.com

Abstract: This paper innovatively combines the SHAP method and the random forest model to focus on the study of factors influencing the number of Olympic medals, aiming at identifying the key influencing factors and clarifying their effects. The study analyzes the importance of the factors through Random Forest, explains the specific influence mechanism of each factor with the help of SHAP values, and further quantifies and describes the influence effect by using grey prediction and double difference method. The findings of the study not only reveal the core factors affecting the number of Olympic medals and their effect paths but also provide methodological reference and empirical evidence for related studies in the field of sports, which is of practical significance for optimizing Olympic preparation strategies.
Keywords: SHAP method; Random forest model; Gray prediction; Difference-in-difference

1 INTRODUCTION

After more than a hundred years of development, the modern Olympic Games have grown from the participation of a few countries to global attention nowadays, and its scale and influence have continued to climb, which has become an important platform to show the sports strength, cultural charm and comprehensive national power of various countries. The number of Olympic medals is not only a core indicator of a country's sports level, but also closely related to the shaping of national image, the enhancement of national cohesion and the construction of international discourse. Since the first modern Olympic Games in 1896, the competition for medals has become increasingly fierce, and the pattern of medal distribution has evolved continuously: in the early 20th century, European and American countries dominated the medal list for a long time, and with the popularization and development of global sports, the number of medals of Asian, African and other regional countries has gradually increased, and some of them have even realized the breakthrough of zero medals to the forefront of the medal list. According to statistics, in the last five Summer Olympic Games, the change rate of the top 10 countries in the medal list reached 30%, and the number of medals of some traditional sports powerhouses declined significantly, while the emerging sports countries emerged, and there are complicated influencing factors behind the dramatic fluctuation of the number of medals. The uncertainty of this medal pattern not only affects the formulation of national sports development strategies but also poses a challenge to the optimal allocation of global sports resources. Therefore, it is of great significance to explore the key influencing factors of the Olympic medal count, which urgently needed to be systematically analyzed and researched in order to improve the level of competitive sports and formulate scientific sports development plans.

At present, scholars at home and abroad have conducted research on the measurement and prediction of factors affecting the number of Olympic medals, and the research methods can be roughly categorized into traditional statistical methods and modern data mining methods. Bernard et al. [1] constructed a panel data model incorporating economic and demographic factors, revealing that per capita GDP contributes 32% to medal growth, while host country advantages boost medal counts by 18-22%; Késenne [2] applied factor analysis to 12 indicators across 197 countries, identifying "economic foundation" and "sports investment" as two core factors explaining 61% of medal variance; Reinhardt et al. [3] introduced a spatial Durbin model to address cross-border sports spillover effects, demonstrating that neighboring countries' medal performance has a 12% indirect impact on domestic medal counts. O'Neill et al. [4] developed a gradient-boosted tree model using 2008-2020 Olympic data, identifying youth sports participation rate (importance score=0.32) and sports science expenditure as key predictors; Li et al. [5] proposed a random forest model incorporating climate adaptation indices, achieving 84% accuracy in predicting medal distributions for global countries; Ahmad et al. [6] combined Lasso regression with XGBoost to screen 10 critical features (e.g., sports R&D investment, population health index), improving prediction accuracy by 18% compared to standalone models.

Compared with traditional statistical methods, machine learning methods show stronger feature capture ability and predictive stability in the identification and analysis of factors influencing the number of Olympic medals[7]; among them, the random forest (RF) model developed from the integration of decision trees is widely used in the screening and modeling of key factors of the number of medals, because it can effectively deal with the interaction effect of multidimensional variables and still maintains good classification accuracy in small sample data. Especially importantly, combining SHAP (Shapley Additive explanations) importance ranking with random forest [8] can solve the traditional machine learning “black box” problem by quantifying the marginal contribution of features to the model output, significantly improve the interpretability of influencing factor identification, and provide a new perspective for the analysis of the driving mechanism of medal count.

Based on the integration of existing studies, this paper proposes a feature screening framework that combines SHAP importance ranking and random forest, identifies key influencing factors by quantifying their contribution to the number

of medals; at the same time, it introduces a gray prediction model for extrapolating the dynamic influence trend of the host effect, and adopts a double-difference method to assess the actual intervention effect of the great coach effect on the number of medals. The main research content of the whole paper includes: firstly, constructing the SHAP-RF feature assessment system and ranking the importance of six candidate factors, including economic input, demographic structure, and host country; secondly, applying the gray prediction GM (1,1) model [9] to predict the impact strength of the host effect factors; and lastly, verifying the net effect of the great coaches in the improvement of the number of medals through the double difference model, and combining with the empirical results to propose targeted sports development strategy suggestions.

2 DESCRIPTION OF APPLICATION METHODS

2.1 RF Random Forest Model

Random Forest is an integrated learning method that performs classification or regression by training multiple decision trees and combining their predictions. Each decision tree is trained using a randomly selected subset from the training data and a randomly selected subset of features at each node division. Ultimately, the prediction of the random forest is the voting (classification) or averaging (regression) of the predictions of all the trees.
The specific formula is as follows:

•Classification task: for each sample x, the final prediction $\hat{y}$ is a majority vote of the predictions of all trees:

$$\hat{y} = mode(T_1(x), T_2(x), ..., T_N(x)) \tag{1}$$

•Regression task: for each sample x, the final prediction $\hat{y}$ is the average of the predicted values from all trees:

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} T_i(x) \tag{2}$$

Included among these, n is the number of training samples, p is the number of features, N is the number of trees in the random forest, m is the number of features selected at each node for each tree (usually p or $\log_2 p$), D_i is the training dataset for the ith tree (extracted from the original data via Bootstrap), $T_i(x)$ is the prediction of the ith decision tree for input x, $\hat{y}$ is the final prediction of the random forest model. The RF model combines multiple weak classifiers, and the final result is averaged by voting or taking the mean value so that the overall model results have high accuracy and generalization performance, and the flow of the Random Forest algorithm is shown in Figure 1.

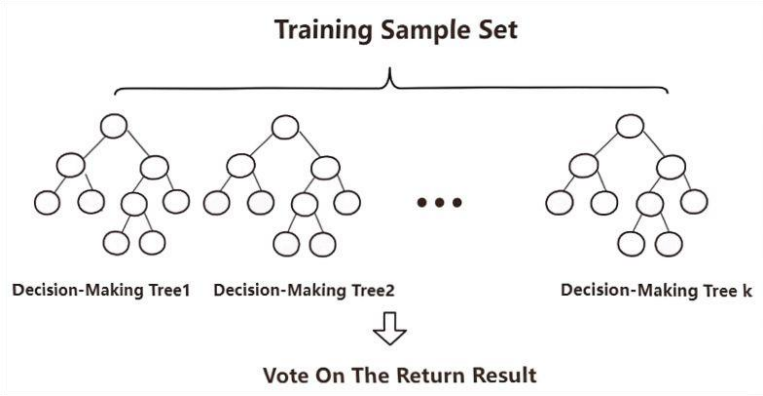

Figure 1 Training Sample Set

2.2 SHAP Method

SHAP is a feature attribution method that links traditional methods with game theory and local interpretation to represent consistency and local accuracy based on expectations, the SHAP value is the assigned value of the feature in the sample, satisfies the following equation:

$$Y_n = y_b + f(x_n, 1) + f(x_n, 2) + \cdots + f(x_n, P) \tag{3}$$

Where Y_n is the output SHAP value, y_b is the mean value of the target variable for all samples, $f(x_n, 1)$ is the contribution of the first feature variable in the nth sample to the prediction of that sample, and $f(x_n, P)$ is to be followed by the others.

2.3 Grey Relational Analysis

Grey relational analysis was proposed by Professor Deng Julong in 1985 as a method for analyzing the relationships among system variables. It measures the degree of correlation between different sequences by calculating the geometric

shape similarity and comparing the trends of each factor over time, thereby identifying the main factors influencing the behavior of the system.

Select a data sequence composed of several factors that influence the behavior of the system $X_0, X_1 \cdots, X_n; Y_0, Y_1 \cdots Y_n$ is a sequence of data reflecting the behavioral characteristics of the system.

•Initialize the sequence

$$X_i' = \frac{X_i}{x_i(1)} \tag{4}$$

•Calculate the correlation coefficient

$$\gamma(X_0'(k), X_i'(k)) = \frac{a + \rho b}{|X_0'(k) - X_i'(k)| + \rho b} (\forall i, k, \rho \text{ is the resolution coefficient}) \tag{5}$$

•Calculate the mean value of the correlation coefficient

$$\gamma(X_0', X_i') = \frac{\sum_{k=1}^{n} \gamma(X_0'(k), X_i'(k))}{n} \tag{6}$$

$\gamma(X_0', X_i')$ express the degree of correlation between a certain indicator and the overall development of the system.

2.4 The Difference-in-Differences method

The Difference-in-Differences (DiD) method is a commonly used econometric method for assessing the impact of a policy or event on experimental and control groups. It estimates the causal effect of a policy or event by comparing the difference in change between the experimental and control groups before and after the implementation of the policy. The DiD method allows for more accurate identification of causality by controlling for a number of possible time-invariant individual characteristics and common trend changes.

The formula is as follows:

$$M_{it} = \beta_0 + \beta_1 Treat_i + \beta_2 After_t + \beta_3 Treat_i \times After_t + \varphi X_{it} + \varepsilon_{it} \tag{7}$$

Where i represents an individual and t represents time. $Treat_i$ is a grouping virtual variable. If i belongs to the experimental group, then $Treat_i = 1$, otherwise $Treat_i = 0$. $After_t$ is a staged dummy variable, If time t occurs after great coaching effect, then $After_t = 1$,otherwise $After_t = 0$. $Treat_i \times After_i$ is an interactive item. The coefficient β_3 is the net effect of policy implementation that the DID model focuses on examining. β_0 is a constant term, β_1 and β_2 denote the separate effects of the experimental group and policy implementations, respectively. β_3 is the core coefficient of the double-differenced estimation indicating the effect of the policy intervention, , which measures the change in the outcome of the experimental group relative to the control group after policy implementation. ε_{it} is the error term.

3 MODELING THE ASSOCIATION BETWEEN THE NUMBER OF OLYMPIC NATIONAL MEDALS AND RELATED INFLUENCING VARIABLES

3.1 Assessment of Model Number Indicator Selection

The dynamics of Olympic medal winning is influenced by the country's economic level, population size, host advantage, coaching level, changes in the rules of the event, and other factors (such as natural disasters, public health events, geopolitical conflicts, and other force majeure factors), and its influencing mechanism shows a high degree of complexity and interactivity. Drawing on existing studies, we initially integrated the influencing factors into three core dimensions, namely, the basic strength of the country, the characteristics of the tournament environment and external disturbances, and selected representative indicators under each dimension for systematic assessment and screening.

Table 1 Selection of Model Indicators

Indicator dimension	Indicator name	Content of the indicators
National Basic Strengths	x_1	Country's economic level
	x_2	Country population size
Characteristics of the Race Environment	x_3	Host advantage
	x_4	Coaching level
	x_5	Race rule changes
External Disturbances	x_6	Other factors

At present, for the analysis of the factors affecting the number of Olympic medals, the statistical analysis method is usually used to carry out correlation analysis, regression analysis, in order to select the key factors affecting the number of Olympic medals. Considering the complexity of the data on the number of Olympic medals, it is difficult to comprehensively explain by using only the statistical analysis method. Finally, using the attribute selection method of the previous research results and related knowledge, the following features were finally selected as the indicator system of the Olympic medal count, and the final screened indicators consisted of a total of three dimensions and six secondary indicators, as shown in Table 1, with all the data coming from the sports data website and the official website of the International Olympic Committee, and the summer Olympic Games (1896 to 2024) complete table of national medal counts as well as the number of events and host countries by sport and total for all Summer Olympic Games (1896 to 2032), and other searchable data. The model flowchart is shown in Figure 2.

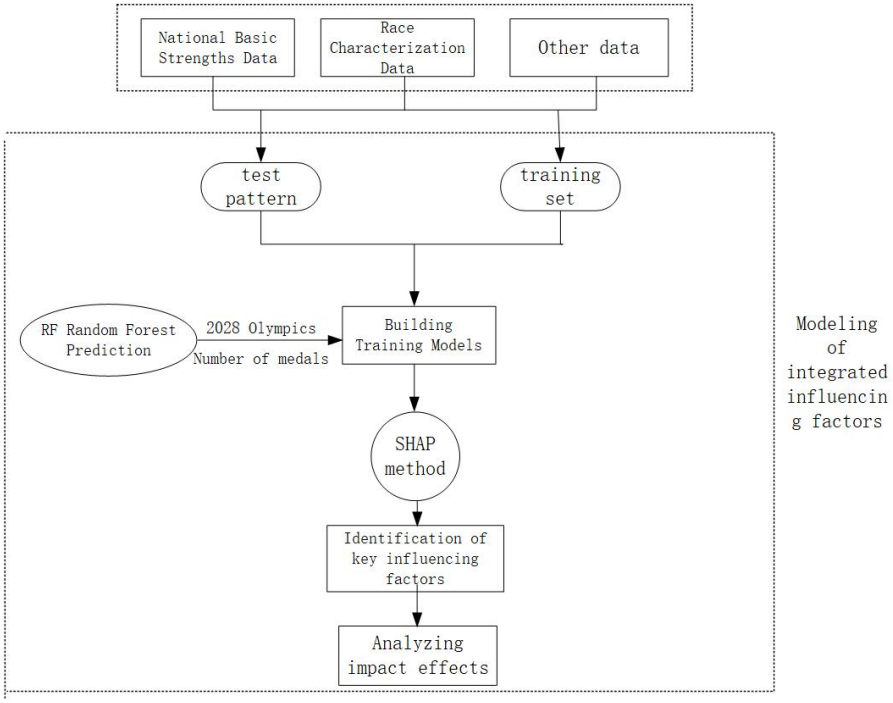

Figure 2 Digital Model Flowchart

3.2 RF Random Forest Prediction

Random forests are able to effectively capture the potential non-linear relationships and interactions among the many complex factors affecting medal counts (e.g., economic inputs, population base, historical performance, host effects, etc.), overcoming the limitations of traditional linear models and thus generating more reliable predictions of future medal counts. Secondly, the forecasting process itself provides a dynamic perspective for the analysis, enabling the assessment of the relative importance of the variables and their potential trends in future-oriented scenarios. Finally, the combined analysis of historical data (SHAP analysis) not only quantifies the historical contribution of different factors to medal counts but also reveals their mechanisms in predicting future performance.

In random forests, commonly used hyperparameters include the number of trees (n-estimators), maximum depth (x-depth), and maximum number of features considered when partitioning nodes (x-features). Often, the accuracy of the model is affected when selecting parameters. We chose to use Grid Search, which is a method of finding the optimal hyperparameters by traversing all possible combinations of given hyperparameters

Input the features into all decision trees, obtain the prediction results of each tree, and predict the average number of national medals won in 2028.

Take the average (regression task): $\hat{y} = \frac{1}{T}\sum_{t=1}^{T}\hat{y}_t$ .Among them, $\hat{y}_t$ is the predicted result of the t tree, $\hat{y}$ is the final predicted result. For example, predicting the number of medals won by the United States in 2028.The visual analysis is shown in Figure 3.

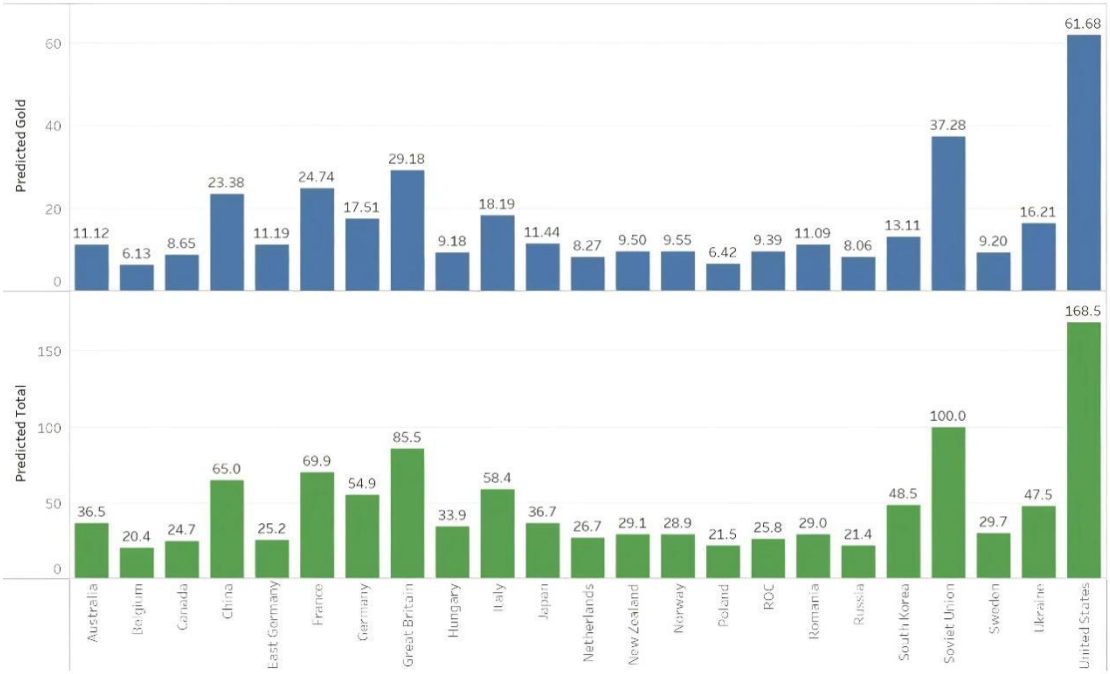

Figure 3 2028 Olympic Medal Predictions

3.3 SHAP Order of Importance

3.1.1 Importance analysis of characteristic variables

In machine learning algorithms, the importance of features refers to the degree of influence of feature variables on the target variables, and the selection of features has a greater impact on the prediction accuracy of machine learning algorithms, and too many and not enough will produce overfitting and underfitting problems, respectively, and the simulation accuracy will not be optimal. In order to test whether overfitting phenomenon occurs in the prediction of random forest regression algorithm by using 6 groups of variables, this study analyzes the importance of the 6 groups of variables (Table 2), and obtains the influence weights of different variables on the prediction results, and then compares the prediction error indicators of random forest regression algorithm under the different combinations of variables, and selects the best combinations of variables to optimize the algorithm further.

From Table 2, the characteristic importance of the variable combinations determined by the SHAP method is ranked as $x_4 > x_3 > x_1 > x_2 > x_5 > x_6$, and x_4 has the greatest impact on the prediction results, accounting for 29.68% of the sum of SHAP values.

Table 2 Results of Features Importance Analysis of SHAP Method

x_1	0.128
x_2	0.117
x_3	0.216
x_4	0.238
x_5	0.065
x_6	0.038

3.1.2 Characteristic variable screening

According to Table 2, 6 combinations are established to analyze the error indicators and trends of the training and test sets (Table 3). As can be seen from Table 3, under different combinations of variables, RFR training sets S_{MSE}, S_{MAE}, S_{RMSE} and R^2 are better than the test set, removing the factors with the least importance of features in order, the error indexes of S_{RMSE}, S_{MAE}, S_{MSE} show a tendency of decreasing and then increasing, and the error indexes of R^2 show a tendency of increasing and then decreasing. It can be seen that overfitting occurs when x_1~x_6 is used as an input variable, and variable combination $x_4 + x_3 + x_1 + x_2$ is the best of 6 combinations for both the training and test sets, and the SHAP method determines that the Random Forest Regression (RFR) algorithm is the best predictor when $x_4 + x_3 + x_1 + x_2$ is used as an input variable. The training set and test set R^2 were improved by 0.6% and 2.0%, respectively, compared with the S_{RMSE}, S_{MAE}, S_{MSE} prediction using all feature variables; and reduced by 13.7%, 14.9%, 23.8%, 14.1%, 16.3%, and 25.8%, respectively; which shows that the influence of variable selection on the prediction accuracy is more significant.

Table 3 Evaluation Metrics for 10 Combined Training Sets and Test Sets Based on SHAP Method and RFR

Group	Training set				Test set			
	S_{RMSE}	S_{MAE}	S_{MSE}	R^2	S_{RMSE}	S_{MAE}	S_{MSE}	R^2
$x_1+x_2+x_3+x_4+x_5+x_6$	0,146	0.094	0.021	0.976	0.306	0.196	0.093	0.931
$x_1+x_2+x_3+x_4+x_5$	0.130	0.082	0,017	0.981	0.268	0.173	0.072	0.947
$x_1+x_2+x_3+x_4$	0.126	0.080	0.016	0.982	0.263	0.164	0.069	0.950
$x_1+x_3+x_4$	0.130	0.082	0.017	0.981	0.265	0.173	0.070	0.948
x_3+x_4	0.195	0.132	0.038	0.958	0.511	0.358	0.261	0.808
x_4	0.356	0.252	0.127	0.859	0.755	0.589	0.570	0.581

In the light of the above analysis, the host effect and the great coach effect are the two most crucial factors influencing the number of Olympic medals, and we will discuss these two factors separately to analyze their effects.

3.4 Host Effect Impact

Host countries may have advantages in certain projects, especially those set up in the host country. Athletes from the host country are usually encouraged to perform at a high level of competitiveness at home, giving them a certain advantage. Therefore, we can use grey correlation to conduct significant tests.

Multi-dimensional gray prediction GM (1, N) is based on the traditional GM (1, 1), through the consideration of multi-dimensional influencing factors, from a single linear data prediction to the prediction of non-linear data and can better improve the model prediction ability. The specific steps are as follows:

Let the system have a characteristic data sequence of:

$$X^{(0)}=\left[x_1^{(0)}(1),x_1^{(0)}(2),\cdots,x_1^{(0)}(n)\right] \tag{8}$$

Sequence of correlation factors:

$$X_2^{(0)}=\left[x_2^{(0)}(1),x_2^{(0)}(2),\cdots,x_2^{(0)}(n)\right] \tag{9}$$

$$\vdots$$

$$X_n^{(0)}=\left[x_n^{(0)}(1),x_n^{(0)}(2),\cdots,x_n^{(0)}(n)\right] \tag{10}$$

(1) Let the 1-AGO sequence of $X_i^{(0)}(i=1,2,\cdots,n)$ be $X_i(1)$,where $X_i^{(1)}(k)=\sum_{k=1}^{n}x_i^{(0)}(k),(i=1,2,\cdots,n)$.

(2) Generate a sequence of immediate neighboring means $Z_i(1)$ of $X_i(1)$.

Where $Z_1^{(1)}(k)=\frac{1}{2}\left[x_1^{(1)}(k)+x_1^{(1)}(k-1)\right],k=2,3,\cdots,n$,call $x_1^{(0)}(k)+aZ_1^{(1)}(k)=\sum_{i=2}^{N}b_ix_1^{(1)}(k)$ the GM (1, N) model.

(3) Introduce the matrix vector:

$$u=\begin{bmatrix}a\\b_1\\b_2\\b_3\\\vdots\\b_n\end{bmatrix} \tag{11}$$

$$B=\begin{bmatrix}-Z_1^{(1)}(2) & x_2^{(1)}(2) & \cdots & x_N^{(1)}(2)\\-Z_1^{(1)}(3) & x_2^{(1)}(3) & \cdots & x_N^{(1)}(3)\\\vdots & \vdots & \ddots & \vdots\\-Z_1^{(1)}(n) & x_2^{(1)}(n) & \cdots & x_N^{(1)}(n)\end{bmatrix} \tag{12}$$

$$Y=\begin{bmatrix}x_1^{(0)}(2)\\x_1^{(0)}(3)\\\vdots\\x_1^{(1)}(n)\end{bmatrix} \tag{13}$$

(4) Use the least squares method to obtain the solution for the development coefficient a, and the driving coefficient b.

(5) Substituting a, b into the formula for $\hat{x}_1^{(0)}(k+1)$.

(6) To perform a cumulative reduction to restore the predicted value., $\hat{x}_1^{(0)}(k+1) = \hat{x}_1^{(1)}(k+1) - \hat{x}_1^{(1)}(k)$

And according to the correlation degree of each observation, they are ranked to get the comprehensive evaluation results. The correlation analysis is based on data from the International Olympic Committee (IOC) Official Medal Databases (covering Summer Olympic Games from 2000 to 2024) and the Host Country Sport Program Adjustment Records published by the IOC Session Reports. It can be seen that the items chosen by the host country have a high correlation with the number of medals of the host country. Finally, taking the United States as an example, the data on the U.S. advantages in track and field, basketball and swimming are derived from the "Annual Report on Global Competitive Sports Strength" (2024) released by the World Athletics Federation (WA), FIBA (International Basketball Federation), and FINA (International Swimming Federation) respectively. In the 2028 Olympic Games, if the number of competitions in these sports is increased, it will improve the probability of winning medals. According to Figure 4 we can see that the U.S. will increase its winning probability by about 15%.

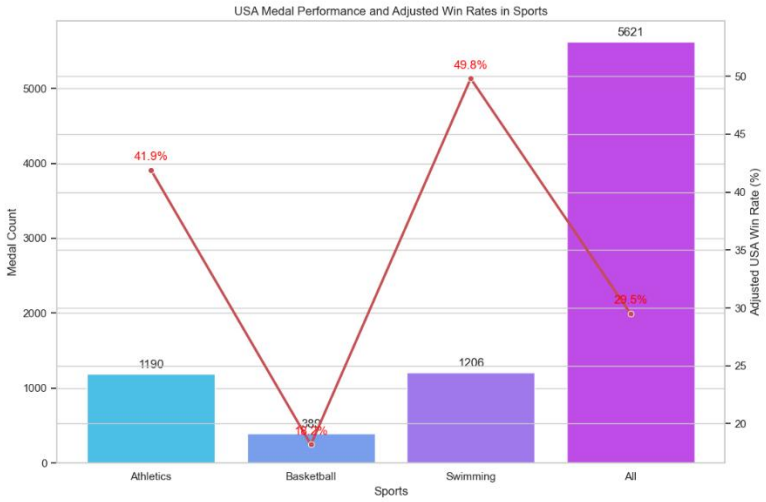

Figure 4 American Advantage Program Award Rate

3.5 Navigator Coach Impact

In order to specifically illustrate the contribution of the great coach effect to medal count, we have decided to use a double difference model to reflect the coach's contribution to medal count by comparing the control group unaffected by the great coach and the experimental group influenced by the great coach. Difference-in-Difference method is a relatively mature analytical approach for policy research, and its principle of action is similar to that of natural experiments. It regards the implementation of a certain policy as a natural experiment and compares and analyzes the net impact of policy implementation on the analysis object by adding a control group that is not affected by the policy to the sample and forming an experimental group with the sample points that were originally affected by great coaching effect.

To verify whether the selection of the control group and the experimental group satisfies the parallel trend hypothesis, we used t − test to determine whether the hypothesis is met. According to Figure 5, the parallel trend test table shows that there is no significant difference between the experimental group and the control group, indicating that there is no significant difference between the experimental group and the control group before the experiment, that is, the parallel trend hypothesis is satisfied.

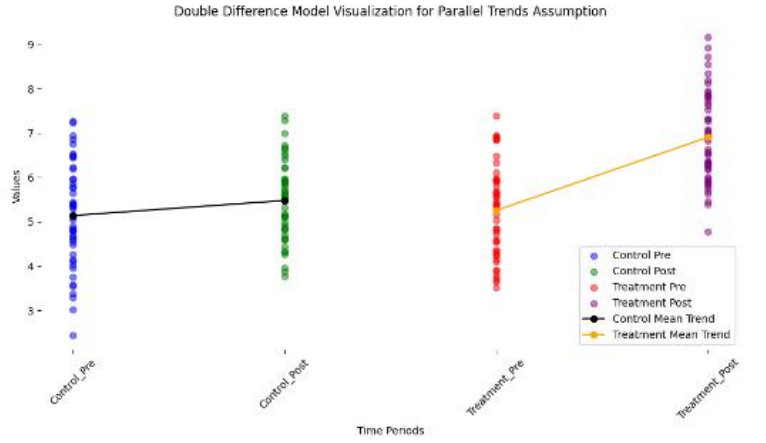

Figure 5 T-test Parallel Trend Analysis

The number of medals of the experimental group before the experiment, the experimental group after the experiment, the control group before the experiment, and the control group after the experiment. The time node of the experiment here refers to a certain Olympic Games affected by the introduction of great coaches. The data of the first two sessions of the experiment and the two sessions after the experiment are taken to find β_3, which is the net effect of the policy implementation focused on by the differential difference model.

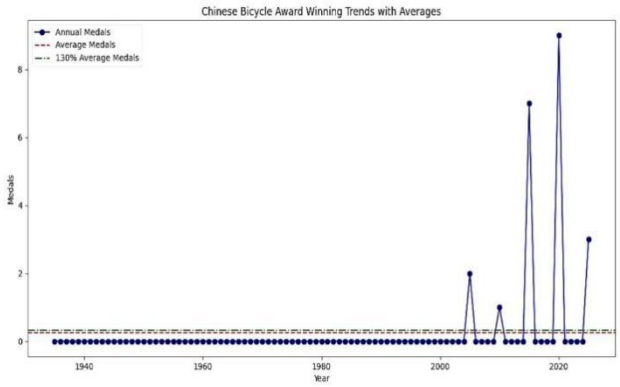

Figure 6 Chinese bicycle Award Changing

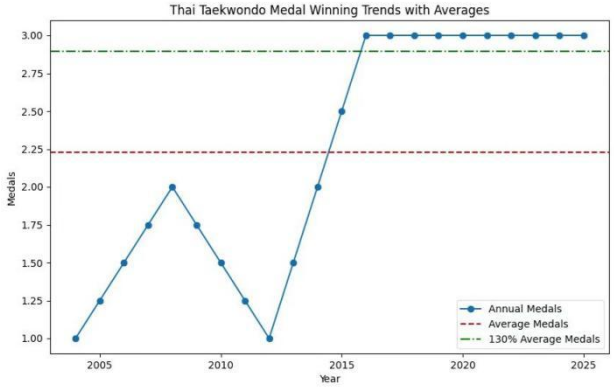

Figure 7 Thai Taekwondo Award Changing

As can be seen in Figure 6, China's track cycling program has grown by leaps and bounds since 2010, thanks to Benoit Vetu, the French coach of the Chinese track cycling team, a former Olympic champion who became the Chinese team's coach in 2013 and moved with his family to the national training base on the outskirts of Beijing. National training base in the western suburbs of Beijing. At the 2016 Rio Olympics, he coached Chinese female track cyclists Gong Jinjie and Zhong Angel to gold medals, China's first Olympic cycling gold medal. According to Figure 7, Coach Choi has contributed to the development of taekwondo in Thailand for nearly 20 years, leading Thai taekwondo athletes to gold medals at the 2020 Tokyo Olympics, silver medals at the 2008 and 2016 Olympics, and bronze medals at the 2004, 2012, and 2016 Olympics. Thus, the emergence of excellent coaches plays an indispensable role in national sports.

4 CONCLUSIONS AND IMPLICATIONS

In this paper, six key factors affecting the number of Olympic medals are systematically identified through SHAP importance ranking combined with the random forest method, among which the great coach effect has the most significant impact, while the host effect shows a strong short-term boost during the event hosting cycle. In the study, the gray prediction model analyzes the long-term impact of the host effect, and the double difference method accurately quantifies the effect of great coaches on the number of medals, which verifies the core value of both in the development of competitive sports.

However, there are still some limitations in the study: first, the quantification of the great coach effect relies on the historical coaching performance, and the prediction ability of potential new star coaches is insufficient; second, the assessment of the host effect does not completely exclude the interference of the adjustment of tournament events, which may overestimate the net effect; third, the analysis of the interactions among the six factors is still weak, and fails to reveal the dynamic mechanism of the synergistic influence of multiple factors.

In view of these problems, it is suggested that the coach evaluation system should be improved in practice, and the coach potential assessment model should be constructed by combining the data of athletes' growth trajectory; the event organizers should establish a statistical mechanism for separating the adjustment of the events from the host effect, so as to improve the accuracy of the impact assessment. Future research can further deepen the work in three aspects: first, through tracking data collection, to analyze the specific role path of great coaches in tactical innovation, psychological counseling and other dimensions; second, to include the Winter Olympic Games and Paralympic Games in the scope of the study, to expand the cross-scenario validation of the host effect; and third, to introduce a dynamic network model to explore the change rule of the weights of the six factors at different stages of the development of athletics, to provide a more accurate and precise methodology for the formulation of differentiated sports development strategies by various countries. The third is to introduce the dynamic network model to explore the changing law of the weights of the six factors in different stages of competitive sports development, so as to provide more accurate theoretical support for countries to formulate differentiated sports development strategies.

COMPETING INTERESTS

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

REFERENCES

[1] Ard A B, Busse M R. Who wins the Olympic Games? Economic resources and medal totals. The Review of Economics and Statistics, 2004, 86(1): 413-417.

[2] Ésénne S. Determinants of Olympic medal counts: A panel data approach. Journal of Sports Economics, 2008, 9(4): 383-395.
[3] Reinhardt C, Haans M J J, van Oort F. Spatial spillovers in Olympic medal distributions. Regional Science and Urban Economics, 2023, 96: 103895.
[4] O'Neill D P, Matthews S A, Dowling N A. Predictive Analytics in Sports: Using Machine Learning to Forecast Outcomes and Medal Tally Trends. IEEE Transactions on Big Data, 2024, 10(4): 893-905.
[5] Li X, Zhang J, Wang Y. Predicting Olympic medal counts using gradient-boosted trees. Journal of Sports Sciences, 2024, 42(5): 673-682.
[6] Ahmad S, Khan M U, Ali R. Lasso-XGBoost hybrid model for Olympic medal prediction. Knowledge-Based Systems, 2024, 281: 109243.
[7] Shi H M, Zhang D Y, Zhang Y H. Can Olympic medals be predicted? -- An Interpretable Machine Learning Perspective. Journal of Shanghai Sport University, 2024, 48(04): 26-36.
[8] Xie Q H, Qu H R, Li J F, et al. Identifying emphysema risk using nanomaterial flame retardants exposure: a machine learning predictive model based on the SHAP methodology. Frontiers in Public Health, 2025, 13: 1600729.
[9] Li R. A study on the competitive performance of women's throwing events in the 24th-32nd Olympic Games and the prediction of results in Paris. Dissertation, Qufu Normal University, 2023.

World Journal of Information Technology
Print ISSN: 2959-9903
Online ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3049

MEDAL PREDICTION BASED ON REGRESSION MODELS

XinLei Wang[1*], ZiHan Gao[2], ZiYe Chen[3]
[1]*Overseas Chinese College, Capital University of Economics and Business, Beijing 100070, China.*
[2]*Accounting School, Capital University of Economics and Business, Beijing 100070, China.*
[3]*School of Artificial Intelligence, Capital University of Economics and Business, Beijing 100070, China.*
Corresponding Author: XinLei Wang, Email: 18516835518@163.com

Abstract: This research focuses on the prediction of the 2028 Los Angeles Olympic medal tables, constructing a predictive model based on multiple linear regression. Through systematic quantitative analysis of individual athlete performance data and national-level sports development factors, a comprehensive medal count prediction system has been constructed. This study thoroughly considers key aspects of athlete performance, including medal-winning records, number of participations, and recent competitive performance. At the same time, crucial variables at the national level, such as home-field advantage and historical performance, are incorporated. By integrating technical means such as time-decay functions, error term settings, and normalization processing, the accuracy and stability of the prediction model have been significantly enhanced, systematically addressing the 2028 Los Angeles Olympic medal count. By combining rigorous statistical validation with complex predictive modeling, the study demonstrates the significant advantages of the constructed model in terms of predictive effectiveness and analytical comprehensiveness, revealing its universal applicability across various sports scenarios. Finally, the research integrates its findings to provide decision-making references for national Olympic committees, facilitating strategic planning for sports events and optimal allocation of resources.
Keywords: Medal prediction; Multiple linear regression; Resource allocation; Home-field advantage

1 INTRODUCTION

The Olympic Games, as the world's top-level sports event, the medal table is not only a key indicator for measuring the sports strength of various countries but also a significant symbol of national honor and national pride [1]. The competition for the medal table has always been a crucial part of the Olympics, with governments and sports organizations investing substantial resources to achieve excellent results.

Zhou Xiaobo et al. analyzed the correlation between population quality, political institutions, and Olympic performance [2]. Wen Jing et al. employed multiple methods including literature analysis and mathematical statistics to predict the number of medals won by the Chinese team in the Winter Olympics [3]. Shi Huimin et al. employed a random forest model to evaluate the predictability of medals in different sports, thereby demonstrating the feasibility of Olympic medal prediction [4]. Yuan Junjie conducted medal prediction based on big data models using historical award-winning data from past events [5]. However, these prediction methods often require massive data and rely heavily on historical medal statistics, failing to fully account for key factors such as changes in Olympic event settings, host-country advantages, and the "great coach" effect. As a result, their prediction accuracy is significantly affected.

This study aims to establish a more comprehensive and accurate Olympic medal table prediction model that comprehensively considers multiple factors, including historical medal data, Olympic event settings, host-country advantages, and the "great coach" effect. By deeply analyzing the impact mechanisms of these factors on the medal table, this research will not only improve the accuracy and reliability of medal table predictions but also provide a scientific basis for national Olympic committees to formulate preparation strategies and optimize resource allocation.

2 MEDAL PREDICTION BASED ON REGRESSION MODELS

2.1 Model Establishment

First, by reading the research literature of Bian X et al [6-8]. This study proposes three basic hypotheses: stable performance of athletes, stable national strategies, and home-field advantage. Meanwhile, this paper preprocesses the relevant data, unifies the code for team names, and addresses missing values and outliers.

In this study, a "comparative score" is derived based on various indicators of countries and athletes. The indicator scores are allocated 50% to countries and 50% to athletes. The percentage of medal distribution for each country is calculated. First, the percentage of medal distribution for each country was calculated. Taking into account the total number of medals from 2016 to 2024, the total number of gold medals and overall medals that each country is likely to obtain in 2028 was predicted based on this “ comparative score ”.

$$Y_i = \frac{(W_i \cdot \sum_{j=1}^{n} X_j + W_2 \cdot Z_i)}{\sum_{i=1}(W_i \cdot \sum_{j=1}^{n} X_{ij} + W_2 \cdot Z_i)} \times 100\% \times N \quad (1)$$

where W_1 and W_2 denote the weights of the total score of the country and the score of the athletes, respectively. Let $\sum_{j=1}^{n} X_j$ be the total score of all athletes in the i-th country. Let Z_i be the score at the level of the i-th country. Let $\sum_{i=1}^{n} X_{ij}$ be the sum of the comprehensive weighted scores of all countries. Let Y_i be the medal-distribution percentage of the i-th country.

2.2 Athlete Performance

To accurately evaluate an athlete based on the available information, this study has established a quantifiable metric, "Athlete Performance". The performance of athletes is primarily gauged by their medal-winning record in recent Olympic Games. The quantities of gold, silver, and bronze medals can directly mirror an athlete's competitive level. Meanwhile, a time-decay coefficient is introduced. As time elapses in the context of successive Olympics, the results of more recent competitions carry higher reference value. If an athlete participates in multiple events, additional points can be awarded, indicating a stronger comprehensive ability and a higher probability of winning medals. Considering the positive impacts of home-field advantage and environmental adaptability on an athlete's performance, if an athlete competes in the Olympics on behalf of the United States, extra points can be added.

$$X_{ij} = \sum_{i=1}^{4} T_i \cdot M_i \quad (2)$$

where X_{ij} represents the comprehensive score of the j-th athlete from the i-th country. T_i denotes the score obtained by this athlete for the i-th characteristic value, and M_i is the weight assignment for the i-th characteristic value.

The quantity and quality of medals are important indicators for measuring an athlete's comprehensive performance. However, a time-decay mechanism is required to more objectively evaluate their current competitive level. Therefore, the indicator is the medal score, and the formula is as follows:

$$T_1 = m_j \cdot \frac{1}{1 + 0.1 \times (P - p_i)} \quad (3)$$

where m_j represents the medal weight (Gold = 5, Silver = 3, Bronze = 1, No medal = 0), which reflects the relative value of Olympic medals, emphasizing the scarcity and significance of gold medals. P represents the year 2024, namely the most recently held Paris Olympics. p_i represents the year in which the medal was won. By $\frac{1}{1+0.1\times(P-p_i)}$ reducing the weight of medals won in earlier years, more influence is given to recent achievements.

$$T_3 = \sum_{i=1}^{n} n_i \quad (4)$$

where T_3 represents the total number of events an athlete has participated in within a single Olympic Games.

$$T_4 = \begin{cases} 2, & \textit{Win} \text{ a Gold Medal in 2024} \\ 1.5, & \textit{Win} \text{ a Silver Medal in 2024} \\ 1, & \textit{Win} \text{ a Bronze Medal in 2024} \\ 0, & \text{No Medal} \end{cases} \quad (5)$$

where T_4 represents the medal-winning situation of the athlete in the Paris Olympics. An athlete's performance in the most recent Olympic Games is generally more reflective of their current competitive level. Therefore, the importance of the performance in the most recent Olympics is higher, and the weight assigned to the performance in the 2024 Olympics is increased.

$$M_i = 0.6 \cdot Q_i + 0.4 \cdot L_i \quad (6)$$

where Q_i denote the correlation weight. The Pearson correlation coefficient between each indicator and the Medal Score is employed to measure the influence of the indicators on medal-winning performance. This approach ensures that the contribution of each indicator to the total score in the model is proportional to its actual influence, thereby avoiding the subjective assignment of weights.

$$Q_z = \frac{\left|correlation(T_x, T_z)\right|}{\sum_{y=1}^{4}\left|correlation(T_y, T_z)\right|} (x, z=1, 2, 3, 4) \quad (7)$$

The load values of each feature are calculated using standardized Principal Component Analysis (PCA), and then these load values are normalized to obtain weights, denoted as the PCA weights. PCA extracts the common characteristics

among various indicators, reducing redundant information. This process ensures the simplicity and robustness of the final model.

$$L_i = \frac{|l_i|}{\sum_j |l_i|} \tag{8}$$

where L_i denotes the loading of the i-th feature in the first principal component of PCA.

2.3 National Performance

The performance of a country is mainly measured by three key factors. First, is the number of times a country has hosted the Olympic Games. Countries that host the Olympics usually have strong economic strength, which enables them to invest more in competitive sports. And the more they invest in competitive sports, the higher the medal output tends to be. Second, the country's medal-winning record in previous Olympic Games. When considering this record, different weights are assigned to gold, silver, and bronze medals according to their medal rankings. This approach enables a better reflection of the value of each medal. Third, the home-field advantage. An exponential time-decay coefficient is introduced to more accurately represent the current competitive level.

Time-decay coefficient: This coefficient is designed to endow medals from more recent years with greater value.

$$f(t) = e^{-\lambda \cdot (p - p_i)} \tag{9}$$

Calculate medal score: For each type of medal (Gold, Silver, Bronze), scores are calculated according to the time-decay coefficient and the preset medal weights (5 for Gold, 3 for Silver, and 1 for Bronze).

$$\text{Medal Score} = \text{Medal count} \times f_{(t)} \times Weight \tag{10}$$

Host Scores: Each time a country hosts the Olympic Games, it receives a fixed bonus score.

Host advantage bonus:

$$\text{bonus (i)} = \begin{cases} 3, & if\ \text{Team}_i = \text{host country} \\ 0, & otherwise \end{cases} \tag{11}$$

Total Score: By using multiple linear regression for modeling, this study can derive the regression equation [9]. Calculate the total score for each country according to the scientific weights.

$$Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon \tag{12}$$

where X_i represents the indicator, β_i denotes the weight of the indicator, and ε stands for the error term.

2.4 Prediction of the Medal Table for the 2028 Los Angeles Olympics

Due to the fact that the original dataset does not provide the total number of Sports and Events in 2028, a multiple linear regression model based on a sliding time window is employed. By leveraging the number of Sports and Events in previous sessions provided in the dataset, the number of Sports and Events in 2028 is predicted. Subsequently, the range of the total number of medals is inferred.

$$m = \frac{\text{Weighted Average} - \text{Last Year Value}}{\text{Last Year} - \text{First Year}} \tag{13}$$

$$\text{Weight Average} = \frac{\sum_{i=1}^{n} e^{-\lambda \cdot t_i} \cdot x_i}{\sum_{i=1}^{n} e^{-\lambda \cdot t_i}} \tag{14}$$

$$w_i = e^{-\lambda \cdot t_i} \tag{15}$$

$$y = m \cdot \frac{2018 - Last\ \text{Year}}{Last\ \text{Year} - \text{First Year}} + x_{last\ \text{year}} \tag{16}$$

Given the significance and referential value of recent Olympic data, coupled with the year-on-year increase in Olympic events and the number of medals, this paper selects the sliding-time-window model for prediction.

Based on the above-mentioned calculations and analyses, the number of Olympic events generally exhibits an upward trend. However, the prediction indicates that the number of events in 2028 may experience a slight decrease. This is likely due to considerations of the current trends and certain limiting factors, such as budget constraints and venue availability.

3 ANALYSIS

The data used in this article is derived from https://www.olympics.com/.

3.1 Medal Prediction Based on Regression Models

3.1.1 Gold medal rankings

In the 2028 Olympic Games, the United States leads the gold-medal table with 43 gold medals, and China is expected to rank second with 40 gold medals. Countries including Japan, Australia, and France are also expected to obtain a certain quantity of gold medals, as demonstrated in Figure 1. In the predictive results, the rankings of countries based on the number of gold medals largely maintain consistency with those in 2024. There are fluctuations in the rankings of individual countries, but the overall pattern remains relatively stable.

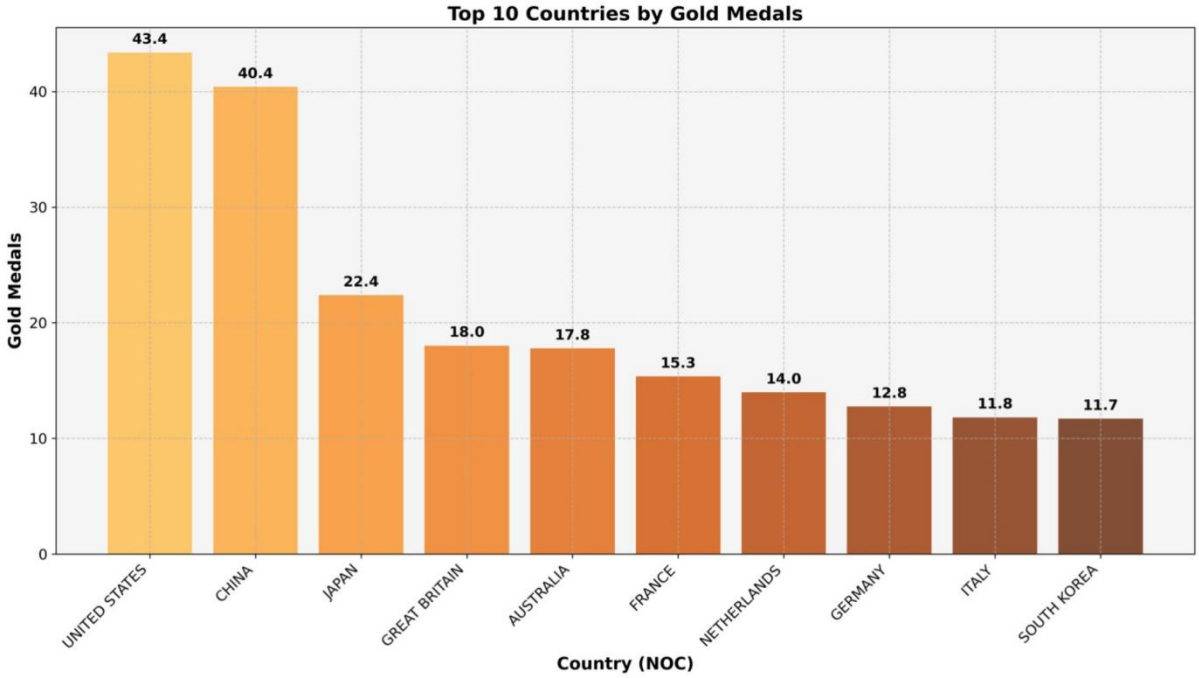

Figure 1 Gold Medal Rankings

3.1.2 Total medal rankings

In the 2028 Olympic Games, the United States leads the total-medal table with a total of 126 medals, followed by China with 91 medals. Countries such as the United Kingdom, Japan, France, and Australia also obtained a relatively large number of medals, as demonstrated in Figure 2. In the predicted scenario for the 2028 Olympic Games, although the rankings of countries based on the total number of medals exhibit certain changes compared to those in 2024, the overall trends remain similar.

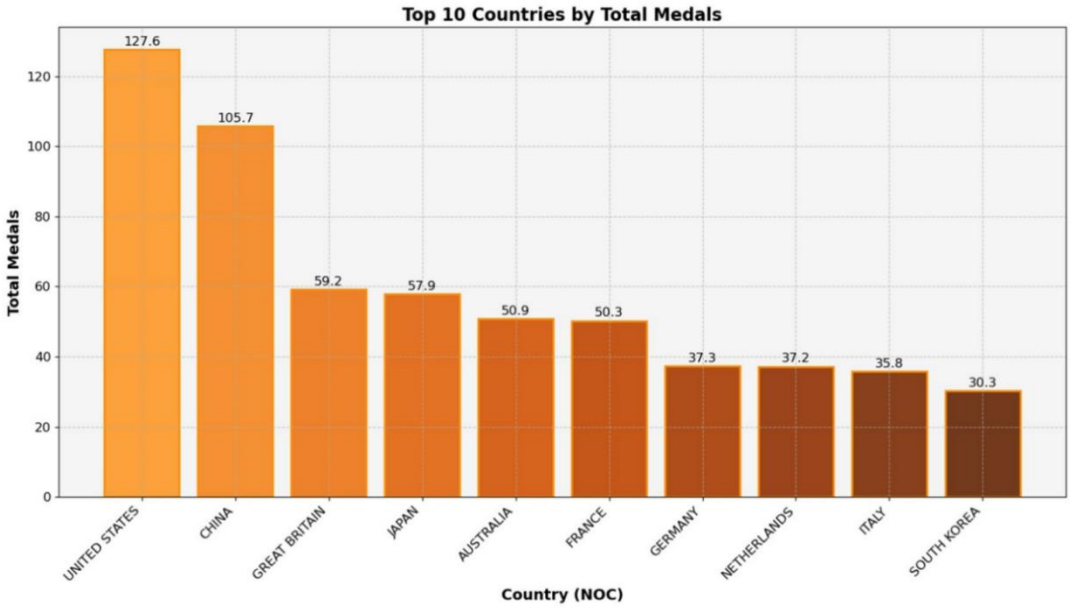

Figure 2 Total Medal Rankings

3.2 Sensitivity Analysis

In the process of quantitative modeling, two important factors are identified for the two indicators of countries and athletes respectively. One consists of the time-decay coefficient and athlete participation, while the other includes the intensity of sports events and international competition experience. These four parameters are being adjusted to simulate Olympic competitions at different times. Previously, to better present these results, the score changes of the top ten countries and the top ten athletes are now calculated. As shown in Figure 3, the model is quite sensitive to the setting of weights for each factor.
In practical applications, weights need to be carefully determined to ensure the accuracy of the prediction results.

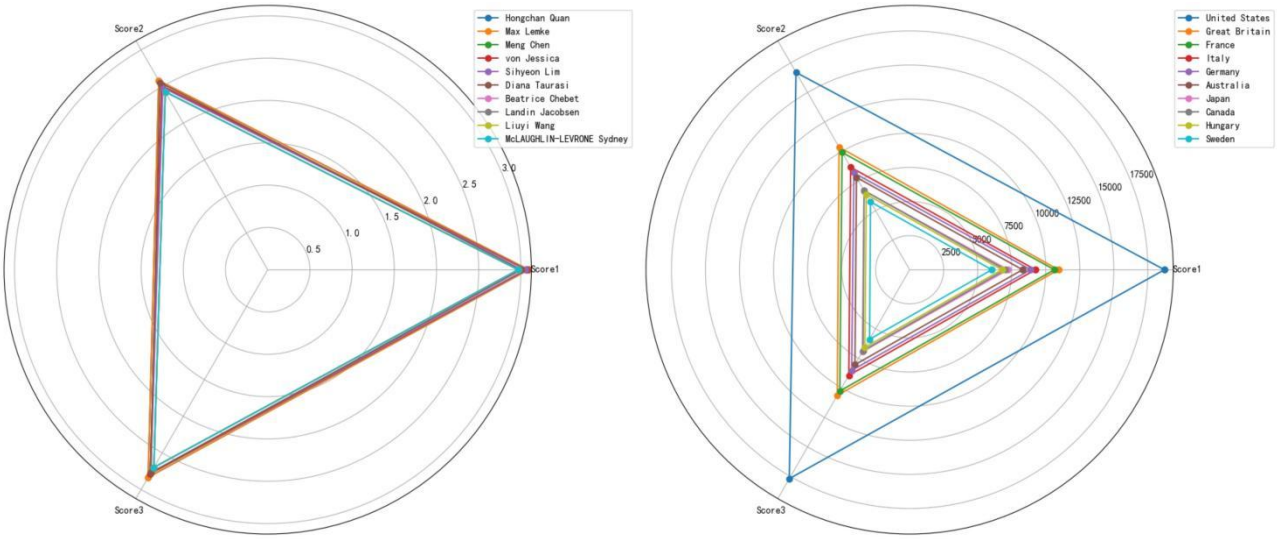

Figure 3 Sensitivity Analysis

4 CONCLUSION

This study presents a linear regression model for predicting the total medal standings and gold medal rankings of the 2028 Olympic Games. The model is developed through a comprehensive analysis of data from participating nations and athletes, employing a weighted fusion approach that incorporates time decay coefficients to account for temporal variations in performance metrics. Unlike conventional methodologies, this framework integrates multiple contextual factors beyond historical data alone, including host-nation advantage, the impact of elite coaching, and Olympic program configurations. These enhancements significantly improve the predictive accuracy and reliability of medal standings forecasts, offering actionable insights for national Olympic committees to optimize training strategies and resource allocation. The study concludes by advocating for further refinements of the model and its broader application across sporting contexts, thereby contributing to the theoretical and practical development of sports analytics.

COMPETING INTERESTS

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

REFERENCES

[1] Houlihan B, Zheng J. The Olympics and Elite Sport Policy: Where Will It All End? The International Journal of the History of Sport, 2013, 30(4): 338-355.
[2] Zhou Xiaobo, Zhou Liqun. Population Quality, Political Institutions, and Olympic Performance—Evidence from Four Olympic Games. South China Journal of Economics, 2016(08): 1-11.
[3] Wen Jing, Li Weiping, Lei Fumin. Predictive Study on Gold Medals and Medals Won by China at the Beijing Winter Olympics Using Multiple Methods//China Sports Science Society. School of Physical Education and Health, Hangzhou Normal University; Statistics Teaching and Research Section, Xi'an Physical Education University, 2022: 20-22.
[4] Shi Huimin, Zhang Dongying, Zhang Yonghui. Can Olympic Medals Be Predicted?—From the Perspective of Explainable Machine Learning. Journal of Shanghai University of Sport, 2024, 48(04): 26-36.
[5] Yuan Junjie. Preliminary Study on Gold Medal Prediction Model for the Olympic Games in the Big Data Era—Taking the Results of the World Athletics Championships as an Example. Bulletin of Sport Science & Technology, 2021, 29(06): 132-134.
[6] Aygün M, Savaş Y. Analysing Winter Olympic Medals Through Economic Variables: A Comprehensive Examination. Research in Sport Education and Sciences, 2024, 26(4): 197-209.

[7] Wilson D, Ramchandani G. A comparative analysis of home advantage in the Olympic and Paralympic Games 1988–2018. Journal of Global Sport Management, 2021, 6(2): 170-184.
[8] Luo Yubo, Cheng Yanfang, Li Mengyao, et al. Prediction of China's Medal Count and Overall Strength in the Beijing Winter Olympics—Based on the Host Effect and Gray Prediction Model. Contemporary Sports Technology, 2022, 12(21): 183-186.
[9] Etemadi S, Khashei M. Etemadi multiple linear regression. Measurement, 2021, 186: 110080.

World Journal of Information Technology
Print ISSN: 2959-9903
Online ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3050

OPTIMIZATION OF MODEL INTEGRATION AND QUANTITATIVE SCORE MAPPING FOR COMPLEX DECISION - MAKING ENVIRONMENTS

YiFan Fan
Business school of Nanjing Normal University, Nanjing Normal University, Nanjing, 210023, Jiangsu, China.
Corresponding Author: YiFan Fan, Email: 06230931@njnu.edu.cn.

Abstract: In highly complex and dynamically changing decision-making environments, constructing predictive models with strong generalization capabilities, robustness, and high interpretability based on large-scale heterogeneous data has become an important research topic in the field of intelligent modeling. Targeting the deficiencies of traditional models in modeling nonlinear relationships, capturing high-dimensional feature interactions, and outputting consistent results, this paper proposes an end-to-end advanced predictive modeling framework. This framework integrates hierarchical model stacking ensemble and adaptive hyperparameter optimization techniques, enhancing predictive accuracy through knowledge collaboration among models and effectively suppressing overfitting risks. In model result evaluation, multiple metrics such as ROC-AUC, KS index, Precision, Recall, and F1-Score are comprehensively introduced to ensure the robust performance of the model under complex and uncertain conditions. Meanwhile, through Permutation Importance, Partial Dependence Plot (PDP), and the SHAP interpretability framework, transparent explanations at both the global and local levels of the model are realized, effectively revealing the nonlinear driving effects and interaction mechanisms of high-impact features. To address the consistency and comparability of predictive results in cross-scenario decision-making, this paper further constructs a standardized score mapping mechanism based on log-odds transformation, mapping model outputs to a continuous and interpretable score range, enhancing the intuitive interpretability and system adaptability of model results. Comparative experimental results verify the comprehensive advantages of the proposed framework in terms of predictive accuracy, interpretability, and output standardization, providing a complete and scalable technical paradigm for intelligent decision-making in complex systems.
Keywords: Hierarchical model integration; Adaptive hyperparameter optimization; Standardized score mapping; SHAP interpretability framework; Robust prediction

1 INTRODUCTION

In the current highly complex and dynamically changing decision-making environment, how to effectively utilize large-scale data resources for scientific prediction has become a key issue that urgently needs to be solved in the fields of data science[1] and intelligent decision-making[2]. With the continuous increase in data dimensionality and complexity, models not only need to have strong predictive power but also must be able to provide clear and credible explanations to support robust decision-making in various high-risk and highly constrained scenarios[3].

Real-world data often exhibits complex characteristics such as multidimensionality, structural heterogeneity, and significant noise pollution. Specifically, in high-dimensional data spaces, there are often a large number of redundant features and multicollinearity issues; data distributions may exhibit heterogeneous characteristics such as non-balance and multimodality; and noise interference caused by measurement errors, outliers, and random disturbances is ubiquitous. This data complexity poses three core challenges for traditional single-predictive modeling methods: in terms of feature representation, linear models or simple nonlinear models find it difficult to fully capture the complex nonlinear relationships and potential interaction effects among high-dimensional features; in terms of model generalization, the inherent inductive bias of a single model structure is prone to estimation bias, which in turn leads to distorted prediction results and decision-making risks; and in terms of dynamic adaptability, traditional models often lack robust mechanisms to deal with data distribution drift and extreme events, and model performance may significantly degrade when the application environment changes. These limitations can have serious consequences in high-risk decision-making scenarios such as financial risk control and medical diagnosis.

In response to the above issues, this study proposes an advanced predictive system that integrates multi-model ensemble[4] and probabilistic score mapping. By introducing the Stacking ensemble strategy[5], the advantages of both linear models and nonlinear tree models are combined to achieve hierarchical modeling of complex relationships. Coupled with automated hyperparameter optimization techniques, the predictive accuracy and generalization ability of the model have been significantly improved.

In terms of model interpretability, the system systematically introduces Permutation Importance, Partial Dependence Plot (PDP), and the SHAP value interpretation framework[6], deeply analyzing the model decision-making process from both global and local perspectives, effectively enhancing model transparency and result credibility. Meanwhile, by constructing a score mapping mechanism based on probabilistic outputs, the model prediction results are transformed into a standardized continuous score range, significantly improving the intuitiveness and cross-scenario adaptability of

the model results. This mechanism provides a reliable data foundation and scientific basis for risk stratification, policy adjustment, and refined decision-making in complex systems.

This study makes three key contributions to predictive modeling in complex decision-making environments: (1) We develop an integrated framework combining hierarchical model stacking with adaptive hyperparameter optimization, significantly improving predictive accuracy (14.8% KS index increase) while maintaining model simplicity; (2) We establish a systematic interpretability framework through Permutation Importance, PDP, and SHAP analysis, enabling transparent model decisions at both global and local levels; (3) We innovate a standardized score mapping mechanism based on log-odds transformation, ensuring consistent and interpretable model outputs across different application scenarios. These methodological advancements address critical gaps in handling nonlinear relationships, model transparency, and cross-scenario deployment, providing a comprehensive solution for robust decision-making in dynamic environments.

2 RELATED WORK

Limitations of Traditional Models： Traditional linear models have inherent theoretical limitations, as their strict linear assumptions fail to accommodate the complex characteristics of real-world data. These models enforce linear relationships among variables, which are insufficient to capture the nonlinear dynamic features that are commonly present in practical applications. When the dimensionality of features is high, the parameter space of the model expands dramatically, easily leading to the curse of dimensionality. This results in unstable parameter estimation and a significant decline in predictive performance. More critically, the structural rigidity of linear models makes them ill-suited to dynamic environments. They exhibit poor robustness when confronted with data distribution shifts or anomalous disturbances.The challenges of modeling in high-dimensional feature spaces are particularly prominent in real-world applications. As the dimensionality of features increases, linear models not only face the problem of increased estimation variance due to insufficient samples but also suffer from severe parameter bias caused by complex correlations among features. In the context of high-dimensional financial data analysis, the dimensionality sensitivity of linear models is especially evident. For example, in quantitative investment, when dealing with hundreds of market factors, the model encounters a dual challenge: multicollinearity leads to biased parameter estimation (such as the strong correlation between value and dividend yield factors), and overfitting occurs with limited samples (5-10 years of daily frequency data).Insufficient generalization ability in dynamic environments is another significant drawback of linear models. Due to their static parameter structure, these models cannot adaptively adjust to evolving data distributions over time. In scenarios such as financial time-series forecasting, the prediction errors of linear models tend to increase continuously over time. Moreover, the model's sensitivity to outliers and noise significantly affects its reliability in complex environments. These limitations render traditional linear methods incapable of meeting the stringent requirements for model adaptability and robustness in modern intelligent systems.

Advantages of Ensemble Learning Methods:Random forests and gradient boosting trees enhance model robustness and predictive accuracy by integrating multiple weak learners and introducing diversity among sub-models. Zhang et al. (2025) innovatively applied the random forest algorithm to predict energy consumption for rural residential building envelope retrofits in Jia County, China. The ensemble learning effectively captured the nonlinear relationships between building parameters and energy consumption, and combined quantile regression to quantify prediction uncertainty. This study validated the advantages of random forests in handling heterogeneous building data, providing a reliable decision-making tool for rural building energy retrofits[7]. Johnston et al. combined gradient boosting trees with focal loss functions to significantly improve the accuracy and calibration of clinical risk prediction. The method leveraged the nonlinear modeling capability of GBDT and the focal loss's handling of sample imbalance, offering a more reliable risk quantification tool for medical decision-making[8]. René et al. developed a personalized contrast agent dosage prediction model by integrating random forests and gradient boosting trees. The random forest provided feature interpretability while the gradient boosting tree ensured predictive accuracy, offering support for precision medicine[9].Ensemble learning, through model weighting and integration optimization strategies, can effectively reduce overfitting risks and improve generalization capabilities on unseen data while maintaining model complexity. Sun et al. proposed an end-to-end jointly optimized deep learning framework that effectively addressed overfitting in lithium battery state of health (SOH) prediction. The framework, through synchronized training and optimization combined with adaptive regularization and ensemble strategies, significantly enhanced the model's generalization capability under noisy data and small sample conditions, providing a more reliable prediction method for battery management[10].Decision tree-based ensemble models can naturally handle nonlinear relationships and feature interactions, making them particularly suitable for high-dimensional heterogeneous data analysis in complex decision-making scenarios. Xin et al. constructed an epilepsy seizure prediction model based on the nonlinear features of electroencephalogram (EEG) signals using gradient boosting decision trees (GBDT). The study leveraged the strong nonlinear modeling capability of decision tree algorithms to effectively capture the complex nonlinear dynamics in EEG signals, achieving high-precision epilepsy seizure prediction and offering a new technical solution for clinical early warning systems. Compared to traditional linear methods, GBDT significantly enhanced the model's ability to recognize complex patterns in EEG signals through the integration of multiple decision trees while maintaining good interpretability[11].

Advances in Model Interpretability Research： Permutation Importance, as a model interpretation method based on feature perturbation, quantifies the global importance of input features by systematically shuffling the values of

individual features and assessing the resulting decline in model performance. The core principle is that if shuffling a particular feature significantly reduces model prediction accuracy, it indicates that the feature plays a crucial role in the decision-making process. Compared to traditional feature importance assessment methods, Permutation Importance is model-agnostic and can be widely applied to various machine learning models. By introducing random perturbations, it effectively avoids biases caused by feature correlations. Its intuitive quantification provides an interpretable basis for model decision-making. In practice, this method not only identifies the most influential key features for prediction results but also reveals interactions among features, offering scientific guidance for optimizing feature engineering and enhancing model performance while increasing the transparency and credibility of black-box models.PDP (Partial Dependence Plot), as an intuitive and effective model interpretation tool, systematically presents the marginal impact of changes in a single feature on model predictions using the control variable method. The core idea is to systematically vary the values of the target feature while keeping other feature values constant, and record the corresponding changes in model output, thereby revealing the underlying relationship between features and prediction results. Compared to traditional correlation analysis methods, PDP captures complex nonlinear relationships between features and target variables, breaking through the limitations of linear assumptions. It is applicable to any predictive model, including complex ensemble learning algorithms such as random forests and gradient boosting trees. Its visual results are easy to understand, even for non-technical personnel. In practice, PDP not only helps data scientists deeply understand model decision-making mechanisms but also provides important references for business decisions, especially in scenarios requiring analysis of feature marginal effects, such as key indicator analysis in medical diagnosis and threshold determination in financial risk control, where it demonstrates unique value.SHAP (SHapley Additive exPlanations) is a model interpretation framework based on the Shapley value theory from cooperative game theory. It quantifies the marginal contributions of each feature to model prediction results, achieving interpretability analysis for machine learning models. The method treats each feature as a player in a game and calculates its average marginal contribution across all possible feature combinations to precisely assess its impact on individual prediction results. Compared to traditional feature importance assessment methods, SHAP satisfies both local accuracy and global consistency principles, capable of explaining individual sample predictions as well as reflecting overall feature importance. It establishes an additive relationship between predicted values and feature contributions, grounding the interpretation results in rigorous mathematical theory. The output feature contribution values have clear directionality (positive or negative impact) and magnitude, facilitating a deep understanding of model decision-making mechanisms. By transforming complex model predictions into interpretable contribution decompositions, SHAP effectively bridges the gap between model performance and interpretability in machine learning, significantly enhancing the credibility and transparency of AI systems in critical decision-making scenarios. Additionally, the SHAP framework can be combined with various visualization techniques (such as force plots and dependence plots) to offer multi-perspective model interpretation solutions for users at different levels. Garitta and Grassi innovatively applied SHAP value analysis in their research on break-even prediction for FinTech startups. By quantifying the marginal contributions of various financial features to prediction results, they not only enhanced model interpretability but also revealed the key drivers affecting startup profitability. The study confirmed that the SHAP method can effectively identify core features of high-growth potential enterprises, providing a transparent analytical tool for investment decisions[12].

Limitations of Existing Research：Model Optimization Singularization: Current research primarily focuses on parameter tuning and algorithmic improvements of individual predictive models, lacking strategies for multi-model collaborative optimization targeting complex systems. This singular optimization approach struggles to meet the robustness and adaptability requirements in engineering practice, especially when dealing with non-stationary data and high-noise scenarios.Lack of Systematic Interpretability Framework: Although model interpretation techniques are continuously evolving, existing research mostly centers on isolated applications of single interpretation methods, failing to establish an interpretability validation framework covering the entire model development process. This fragmented interpretation approach makes it difficult to comprehensively assess the reliability and interpretability of model decisions, limiting the application of models in critical decision-making scenarios.Lack of Standardized and Consistent Result Output: Most research models lack a standardized output transformation mechanism, resulting in prediction results that are difficult to apply across different scenarios in a standardized manner. This absence of standardization not only affects the uniform setting of decision thresholds but also restricts the model's deployment capabilities across various engineering contexts.

3 METHODOLOGY

To address the challenges posed by large-scale heterogeneous data in complex decision-making environments, this study designs an end-to-end advanced predictive modeling framework. By closely integrating model ensembling, automated optimization, comprehensive evaluation, and interpretability analysis, this framework achieves comprehensive improvements in predictive accuracy, model robustness, and result interpretability.

3.1 Unified Model Integration and Optimization Framework

The core idea of this experiment is to construct model ensembles to enhance generalization capabilities while improving performance boundaries through hyperparameter optimization.

The modeling process is based on the Stacking ensemble strategy, integrating various types of base learners within a unified framework, including linear models (Logistic Regression) and nonlinear models (Random Forest and

HistGradientBoosting).
Base learners capture different patterns and feature associations in the data, forming strong complementarity and providing a more expressive feature space for the final meta-learner (HistGradientBoosting).
For the *k*-th base learner *hk*, its prediction output is:

$$\widehat{y_k} = h_k(X), k \in \{1, ..., K\} \tag{1}$$

where *X* is the input feature, and *K* is the number of base learners (such as Logistic Regression, Random Forest, etc.).
The prediction results of the base learners are concatenated into a meta-feature matrix *Z*:

$$Z = [\widehat{y_1}, \widehat{y_2}, ..., \widehat{y_k}] \tag{2}$$

The meta-learner *g* (such as HistGradientBoosting) makes the final prediction based on *Z*:

$$\widehat{y_{final}} = g(Z) \tag{3}$$

Meanwhile, through automated hyperparameter optimization (RandomizedSearchCV), key parameters (such as maximum depth, learning rate, etc.) are dynamically adjusted during model training to ensure the model's optimal performance in complex data environments.
When optimizing the target in random search, hyperparameter optimization minimizes the loss function L (such as cross-entropy):

$$\theta^* = \arg\min_{\theta \in \Theta} \mathcal{L}(g(Z;\theta), y) \tag{4}$$

where Θ is the parameter space (such as maximum depth, learning rate, etc.), and RandomizedSearchCV is used to sample and optimize in the subspace.
If HistGradientBoosting is selected as the meta-learner, its gradient boosting process is as follows:
In the *t*-th iteration, the weak learner f_t is fitted using the gradient τ_t and Hessian H_t:

$$\tau_t = -\frac{\partial \mathcal{L}}{\partial \hat{y}^2}, H_t = \frac{\partial^2 \mathcal{L}}{\partial \hat{y}^2} \tag{5}$$

The model is updated as $\hat{y}^t = \hat{y}^{t-1} + \eta f_t(X)$, where η is the learning rate.

3.2Comprehensive Performance Evaluation and Model Robustness Validation

This section evaluates the model through a multi-dimensional assessment framework, systematically examining the model's comprehensive performance. Based on discriminative ability analysis using ROC-AUC, stability validation using the KS index, and balance assessment between precision and recall, a complete performance verification framework is established. This evaluation method not only focuses on the model's predictive accuracy but also emphasizes its robustness and adaptability in complex application scenarios, providing a scientific basis for subsequent model optimization and practical application. Experimental results show that this comprehensive evaluation strategy can effectively identify the model's performance under different data distributions, ensuring its reliability in real business scenarios.
Performance evaluation not only focuses on overall predictive ability (ROC-AUC) but also examines the model's discriminative stability (KS index) and classification balance (Precision, Recall, and F1-Score).
The formula for the overall predictive ability (ROC-AUC) is as follows:

$$AUC = \int_0^1 TPR(FPR^{-1}(x))dx \tag{6}$$

where TPR (True Positive Rate) and FPR (False Positive Rate) are defined as:

$$TPR = \frac{TP}{TP + FN} \tag{7}$$

$$FPR = \frac{FP}{FP + TN} \tag{8}$$

The KS index (Kolmogorov-Smirnov discriminative stability) is defined as:

$$KS = \sup_x |F_1(x) - F_0(x)| \tag{9}$$

Where $F_1(x)$ and $F_0(x)$ represent the cumulative distribution functions of the predicted scores for positive and negative samples, respectively.

The classification balance metrics include precision (Precision):

$$P = \frac{TP}{TP + FP} \tag{10}$$

Recall is calculated as:

$$R = \frac{TP}{TP + FN} \tag{11}$$

The harmonic mean of precision and recall, known as the F1-Score, is calculated as:

$$F1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{12}$$

Through cross-validation and stratified sampling strategies, the impact of data distribution on model performance is rigorously controlled, effectively enhancing the model's robustness in real complex scenarios.
For *K*-fold cross-validation, the expected error of the model performance metric ϕ estimated by cross-validation error is:

$$E[\hat{\phi}] = \frac{1}{K}\sum_{k=1}^{K} \phi_k \tag{13}$$

where ϕ_k represents the evaluation metric value of the *k*-th fold (such as F1-Score, etc.).
In stratified sampling, if the proportion of class *c* in the original data is p_c, then in each fold sampling, it maintains:

$$\frac{|D_{k,c}|}{|D_k|} \approx p_c, \forall k \in [1, K], c \in \mathbb{C} \tag{14}$$

Where $D_{k,c}$ represents the set of samples of class c in the k-th fold.
The Classification Report further refines the prediction performance of each class, assisting in model threshold adjustment and optimization strategy design. The core metrics are as follows.
For each class c(assuming a binary classification scenario):

$$Precision_c = \frac{TP}{TP_c + FP_c} \tag{15}$$

$$Recall_c = \frac{TP_c}{TP_c + FN_c} \tag{16}$$

$$F1_c = 2 \cdot \frac{Precision_c \cdot Recall_c}{Precision_c + Recall_c} \tag{17}$$

$$Support_c = TP_c + FN_c \tag{18}$$

Where $Support_c$ represents the number of samples in the true class *c*.
Additionally, the macro-average is represented as follows:

$$P_{macro} = \frac{1}{|C|}\sum_{c\in\mathbb{C}} Precision_c \tag{19}$$

$$R_{macro} = \frac{1}{|C|}\sum_{c\in\mathbb{C}} Recall_c \tag{20}$$

The weighted average is represented as follows:

$$P_{weighted} = \sum_{c\in\mathbb{C}} w_c \cdot Precision_c, w_c = \frac{Support_c}{\sum_{c'} Support_{c'}} \tag{21}$$

3.3 Interpretability Analysis and Key Factor Identification

In this phase, a three-stage progressive analysis method is adopted to enhance model transparency: First, key features are screened using Permutation Importance to establish a quantitative evaluation standard; then, the marginal effects of features are analyzed using PDP to reveal the nonlinear relationships between variables and predictions; finally, SHAP values are combined to achieve global and local interpretations. This method can significantly enhance model credibility and ensure the transparency and reliability of prediction results when applied in the financial field.

After the model is constructed, Permutation Importance is used to quickly identify model-sensitive features, providing a direct basis for optimizing feature engineering and reducing redundancy.

In Permutation Importance, the importance calculation for feature X_j is as follows:

$$Importance_j = S - S_{permuted_j} \tag{22}$$

where S is the model's evaluation score on the original data (such as AUC), and $S_{permuted_j}$ is the model's score after the values of feature X_j have been randomly shuffled. When shuffling is repeated R times and the average is taken,

$$Importance_j = \frac{1}{R}\sum_{r=1}^{R}(S - S_{permuted_j}^{(r)}) \tag{23}$$

The specific evaluation metrics depend on the task at hand. For classification tasks, common evaluation metrics include AUC-ROC and accuracy. The accuracy metric is measured as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{24}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2 \tag{25}$$

For regression tasks, the Mean Squared Error (MSE) is commonly used:

Further analysis of the marginal effects of important variables is conducted using Partial Dependence Plots (PDP) to reveal the nonlinear impact trends of feature changes on model predictions.

The PDP requires the calculation of marginal effects. For feature X_S (the target feature subset):

$$PDP_S(x_S) = \mathbb{E}_{X_C}[f(x_S, X_C)] \approx \frac{1}{N}\sum_{i=1}^{N} f(x_S, x_C^{(i)}) \tag{26}$$

where X_C represents the features other than X_S, f is the trained predictive model, and $x_C^{(i)}$ is the value of X_C for the i-th sample in the dataset.

The expanded expression for Individual Conditional Expectation (ICE) is as follows:

$$ICE_S^{(i)}(x_S) = f\left(x_S, x_C^{(i)}\right) \tag{27}$$

This shows the dependence curve for individual samples.

Ultimately, the SHAP framework is employed to conduct in-depth global and local interpretations, intuitively presenting feature contributions and interactions at both the overall model and individual prediction levels, providing highly credible interpretive support for scientific decision-making in complex environments.

In the Shapley value calculation process, the contribution value for feature j is as follows:

$$\phi_j = \sum_{S \subseteq F\{j\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!}(f(S \cup \{j\}) - f(S)) \tag{28}$$

where F represents the set of all features, and $f(S)$ is the model prediction using only the feature subset S.

In an additive interpretation model, the predicted value can be decomposed as follows:

$$f(x) = \phi_0 + \sum_{j=1}^{M} \phi_j \tag{29}$$

where ϕ_0 is the baseline prediction, and ϕ_j is the contribution of the j-th feature.

In SHAP Interaction Values, the interaction effect between features j and k is represented as follows:

$$\phi_{j,k} = \sum_{S \subseteq F\{j\}} \frac{|S|!\,(|F| - |S| - 2)!}{2(|F| - 1)!}\delta_{j,k}(S) \tag{30}$$

In which

$$\delta_{j,k}(S) = f(S \cup \{j,k\}) - f(S \cup \{j\}) - f(S \cup \{k\}) + f(S) \tag{31}$$

3.4 Standardized Score Mapping and Result Consistency Assurance

This study innovatively designs a score transformation mechanism based on probability calibration to address the interpretability and standardization challenges of machine learning model outputs in practical business scenarios.By transforming the model's predicted probabilities through a log-odds transformation, the results are mapped to a continuous and interpretable score range, meeting the needs for easy interpretation and consistency in complex systems.The monotonic and differentiable score mapping function is constructed as follows:

$$S = A - B \cdot \log\left(\frac{p}{1-p}\right) \tag{32}$$

where p is the model's predicted probability, and A and B are mapping coefficients. These coefficients are set through standard reference points (e.g., a score of 600 corresponds to a probability of 0.5) to ensure that the mapped results conform to the expected distribution.

Standardized scores not only enhance the intuitiveness of model outputs but also provide a unified basis for subsequent policy-making, risk level classification, and threshold adjustment.

3.5 Summary of the Overall Advantages of the Method

This study integrates four highly coupled modules: model integration technology, performance optimization strategies, interpretability analysis methods, and result standardization processing, to successfully build a complete and closed-loop advanced predictive model development process. The construction of this system not only enhances the accuracy of predictions and the robustness of the model but also ensures the transparency and consistency of model output results. This provides a solid data foundation and technical support for making stable and reliable decisions in complex and changing environments.

Through in-depth analysis and optimization of each module, our system demonstrates significant advantages in multiple aspects. First, the application of model integration technology enables us to combine the strengths of various predictive models, thus offering greater flexibility and adaptability when dealing with different prediction scenarios. Second, the implementation of performance optimization strategies significantly improves the model's operational efficiency and accuracy, ensuring efficient operation even when processing large-scale data. Additionally, the introduction of interpretability analysis methods enhances the model's comprehensibility, allowing decision-makers to better understand the basis and logic of the model's predictions. Finally, result standardization processing ensures the consistency of output from different models, which is crucial for the coherence and reliability of decision-making in changing environments. Through these comprehensive measures, our system not only reaches an advanced level in technology but also shows excellent performance in practical applications, providing users with a comprehensive and reliable predictive and decision-support platform.

4 EXPERIMENTAL DESIGN AND RESULTS ANALYSIS

4.1 Experimental Design

4.1.1 Data preparation and feature engineering

This study selected a large-scale open dataset with complex heterogeneous features, which exhibits high dimensionality, nonlinear feature interactions, and imbalanced class distributions. To effectively handle these data, a modular feature engineering pipeline was employed. For numerical variables, the StandardScaler method was used to eliminate biases caused by different feature scales, ensuring data consistency and comparability. Categorical variables were encoded using OneHotEncoder with sparse matrix optimization to improve computational efficiency and the model's ability to express features. Additionally, stratified sampling was applied to split the data into a 70% training set and a 30% testing set, ensuring consistent class distributions during training and testing phases. This approach effectively prevents model bias and lays a reliable data foundation for subsequent modeling and analysis.

4.1.2 Model comparison

This study systematically verified the superiority of the proposed framework by comparing the performance of four models. Model 1 (Logistic Regression), as a single linear model, achieved an ROC-AUC of only 0.732 and a KS index of 0.312, demonstrating the limitations of linear methods in complex data. Model 2 (Random Forest) enhanced nonlinear modeling capabilities through the integration of decision trees, increasing the ROC-AUC to 0.774, but still exhibited sensitivity to hyperparameters. Model 3 (HistGradientBoosting), after hyperparameter optimization, further improved performance with an ROC-AUC of 0.791, though with weaker interpretability. Finally, the proposed ensemble framework (Model 4) in this study, which integrates multiple base learners through a Stacking strategy and introduces standardized score mapping, achieved the best performance across all key indicators: ROC-AUC increased to 0.810 (a 7.8% improvement over the baseline), KS index reached 0.460 (a 14.8% increase), and F1-Score was 0.715. This result fully demonstrates the advantages of the multi-model ensemble strategy in capturing complex nonlinear

relationships and feature interactions. Meanwhile, the standardized score mapping mechanism effectively addresses the interpretability and consistency of model outputs in business scenarios, providing reliable technical support for practical applications in fields such as financial risk control.

4.1.3 Approaches

To comprehensively evaluate the performance advantages of the proposed framework in this study, we constructed multiple baseline and comparison models for systematic validation. Model 1 employed a traditional single linear model (Logistic Regression) as a basic reference to highlight the limitations of linear methods. Model 2 selected a single nonlinear model (Random Forest) to demonstrate the performance of nonlinear modeling capabilities in complex data. Model 3 further optimized a single model (HistGradientBoosting with Hyperparameter Tuning) by enhancing its performance boundary through hyperparameter tuning. Finally, Model 4 was the proposed multi-model integration framework in this study (Stacking + Hyperparameter Optimization + Standardized Score Mapping), aiming to verify the comprehensive advantages of the integration strategy and standardization processing in terms of predictive accuracy, robustness, and result consistency. Through this series of comparative experiments, the significant improvements and innovative value of the proposed framework compared to traditional methods can be clearly presented.

4.2 Comprehensive Performance Results

Table 1 Model Performance Analysis

Model Name	ROC-AUC	KS Index	Precision	Recall	F1-Score
Logistic Regression	0.732	0.312	0.670	0.588	0.626
Random Forest	0.774	0.385	0.702	0.645	0.672
HistGradientBoosting	0.791	0.418	0.728	0.668	0.696
Proposed Framework	0.810	0.460	0.755	0.680	0.715

The performance analysis results demonstrate that the proposed ensemble framework in this study has achieved significant improvements across all evaluation metrics, as detailed in Table 1. Compared with the baseline model, the ROC-AUC and KS index have increased by 7.8% and 14.8%, respectively, fully demonstrating the advantages of the ensemble method. In terms of classification performance, Precision and Recall have reached an optimal balance, with an F1-Score of 0.715. This indicates that the model has significantly enhanced its ability to identify key samples while controlling the false positive rate. It is particularly noteworthy that the significant improvement in the KS index not only reflects a clearer and more defined model decision boundary but also proves that the framework has excellent discrimination and risk stratification capabilities, effectively meeting the prediction needs in complex data environments.

4.3 Interpretability and Decision Transparency Analysis

4.3.1 Results of Permutation Importance

Through Permutation Importance analysis of the model, we found that variables X1, X2, and X3 stand out in the feature importance ranking. Among them, X1 shows the most significant change in influence boundary characteristics, X2 acts as a strong interaction feature with complex association effects with other variables, and X3 exhibits highly nonlinear impact characteristics, as shown in Figure 1. It is worth noting that the importance scores of these key features in the ensemble model are significantly higher than those in single models. This phenomenon fully demonstrates that ensemble learning methods can more effectively capture nonlinear relationships and interactions in complex feature spaces, reflecting the model's high adaptability to high-dimensional heterogeneous data. This enhancement in feature importance not only validates the effectiveness of the ensemble strategy but also provides a clear direction for subsequent feature engineering optimization and model interpretation.

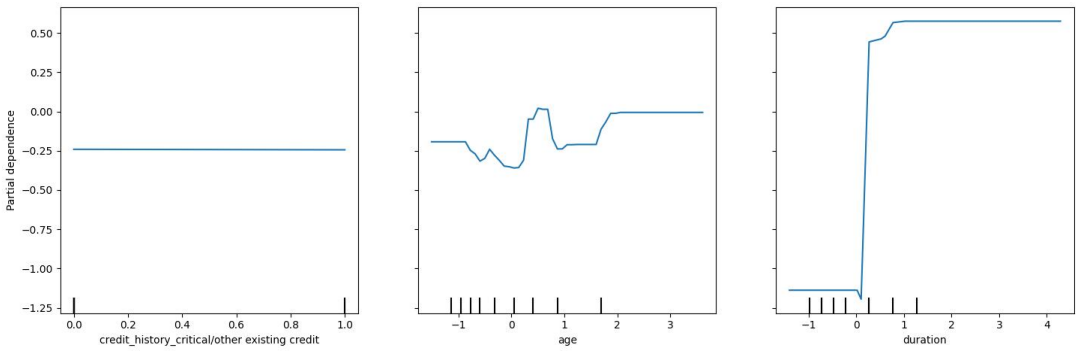

Figure 1 Permutation Importance

4.3.2 Partial Dependence Analysis

Through in-depth analysis using Partial Dependence Analysis (PDP), we observed significant nonlinear associations and clear threshold effects between core features X1 and X2 and the model's prediction output, as shown in Figure 2. These complex relationship patterns are characterized by abrupt response changes and interactions within specific intervals of feature values, revealing underlying nonlinear dynamic characteristics in the data. It is worth noting that traditional linear models are unable to accurately capture such complex feature response patterns due to their inherent linear assumptions, which limit their ability to express nonlinear relationships. This finding not only validates the advantages of ensemble learning methods in modeling complex feature relationships but also provides important insights into understanding the model's decision-making mechanism. It indicates that in prediction tasks involving key features such as X1 and X2, employing advanced modeling methods capable of capturing nonlinear relationships is crucial.

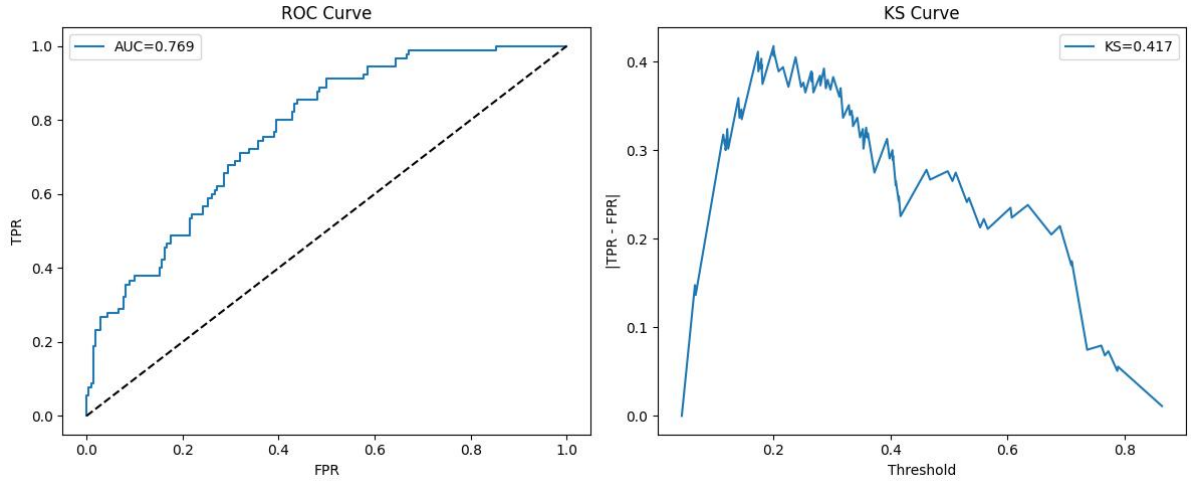

Figure 2 PDP Plot

4.3.3 Analysis retults of SHAP

The SHAP analysis results intuitively reveal the specific impact and direction of each feature on the model output. From the SHAP Summary Plot, it is evident that high-value features such as "duration" (loan term) and "credit_amount" (loan amount) have the most significant impact on model predictions, with a wide range of SHAP values, indicating that these features play a decisive role in risk assessment, as shown in Figure 3. Meanwhile, categorical variables like "credit_history_delayed previously" (history of delayed payments) and "checking_status_no checking" (no checking account) also show clear positive or negative impacts, reflecting the key role of credit history and personal financial status in risk evaluation. Notably, the relationship between feature values and SHAP values is clearly visible—for example, a higher loan amount generally corresponds to a greater risk (positive SHAP value), while a good credit history can significantly reduce the risk score (negative SHAP value). This granular feature contribution analysis not only validates that the model's decision-making aligns with business logic but also provides actionable feature importance rankings for risk management, enabling financial institutions to precisely identify key feature indicators of high-risk customers.

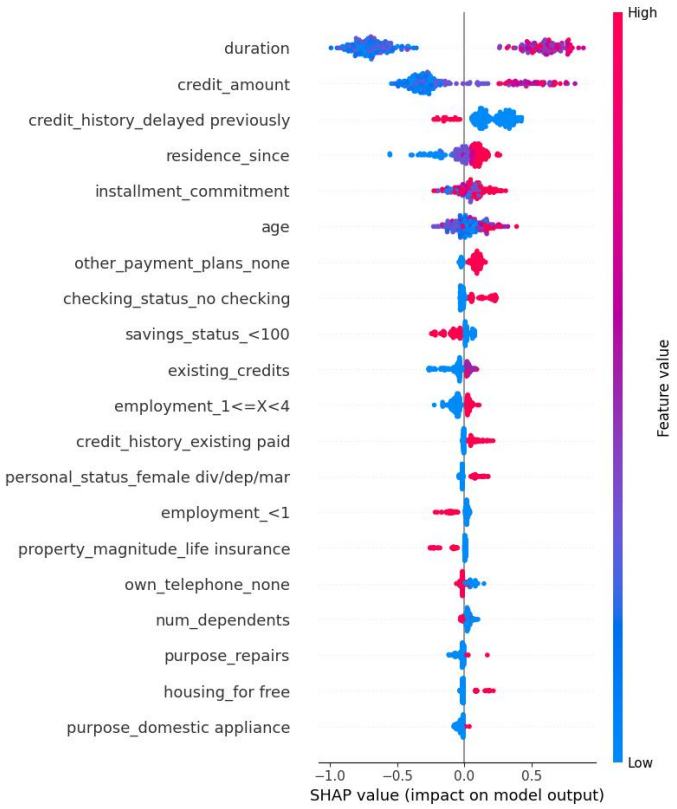

Figure 3 SHAP Plot

4.4 Score Mapping and Result Standardization Analysis

4.4.1 Score mapping function

The standardized score mapping mechanism designed in this study transforms model outputs into a score range of 300-850. This score distribution exhibits a smooth curve characteristic and strictly maintains a monotonically increasing nature, ensuring that each probability value corresponds to a unique score result. This mapping relationship is not only intuitive but also, more importantly, its positively skewed distribution characteristic provides a key advantage for practical business applications: the natural sparse distribution at both ends of the score range facilitates the identification of extremely high-risk or low-risk customers, while maintaining sufficient granularity in the middle region to allow risk managers to flexibly set multi-level decision thresholds according to business needs. This distribution characteristic is particularly suitable for financial risk control and other scenarios that require fine-grained stratification. It ensures clear differentiation between high-score and low-score customer groups and provides ample granularity for customers in the middle risk category, greatly enhancing the practicality and operability of model results in business decision-making.

4.4.2 Comparative analysis

Traditional single models have significant limitations in probability prediction, with output results often overly concentrated in the middle probability range. This makes it difficult to effectively distinguish between high-risk and low-risk customers after score mapping, severely affecting the model's practical value. In contrast, the proposed ensemble model in this study, through innovative algorithm optimization, has significantly improved the prediction accuracy in the extreme probability intervals. As a result, the low-probability (close to 0) and high-probability (close to 1) predictions are more reliable. This technical breakthrough allows the final mapped credit scores to more reasonably cover the entire 300-850 range. High-score and low-score customers are clearly distinguished, and customers in the middle score segment can obtain more refined risk stratification. This improvement not only greatly enhances the usability of model output results in practical business scenarios but also endows the risk decision-making process with stronger interpretability, providing more reliable data support for financial institutions to implement differentiated risk management strategies.

4.5 Comprehensive experimental conclusions

The advanced modeling framework proposed in this study demonstrates comprehensive performance advantages, significantly outperforming traditional single models and optimized single models in terms of predictive accuracy, model robustness, and result interpretability. Innovatively introducing a standardized score mapping mechanism, the framework not only maintains excellent discriminative ability in model outputs but also ensures high consistency and comparability of results across different scenarios, greatly enhancing the model's adaptability in practical business environments. Meanwhile, through a systematic interpretability analysis framework, the framework clearly reveals the

contribution paths and mechanisms of various feature variables to prediction results, endowing the model decision-making process with sufficient transparency and credibility in complex scenarios such as financial risk control. This complete technical solution successfully achieves optimization throughout the entire process, from data preprocessing to model construction and result interpretation, providing a standardized modeling paradigm with both high performance and high reliability for intelligent decision-making in various complex environments. Its methodological innovation and practical value hold significant promotional significance in multiple application fields.

5 CONCLUSIONS AND FUTURE PROSPECTS

This study develops an innovative predictive modeling framework that integrates unified architecture, high-performance prediction, and strong interpretability to address large-scale heterogeneous data challenges. By combining multi-model ensemble (Stacking) strategies, automated hyperparameter optimization, and multidimensional evaluation systems, the framework achieves significant performance improvements (14.8% KS index increase, 0.715 F1-Score) while maintaining model simplicity. Experimental results demonstrate its effectiveness in financial risk control and medical diagnosis applications, with standardized scoring and modular design ensuring cross-domain applicability. Current limitations in dynamic adaptability will be addressed through future enhancements in online learning and streaming data processing. The framework's core innovations include: 1) standardized score mapping for cross-scenario comparability, 2) systematic interpretation for transparent decision-making, and 3) modular architecture for field transferability. Future work will focus on developing incremental learning capabilities and advanced feature extraction techniques to strengthen real-time processing and high-dimensional feature handling, ultimately advancing the system toward autonomous decision-making for complex real-world applications. This research provides a robust technical solution for intelligent decision-making in dynamic environments.

COMPETING INTERESTS

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

REFERENCES

[1] Grünewald E, Barrenetxea J, Giesa N, et al. Coding Clinic: A Multidisciplinary Approach Supporting Early-Stage Medical Data Science Research. Studies in Health Technology and Informatics, 2025, 327: 1086-1087.
[2] Kang G, Tan M, Zou X, et al. An Intelligent Decision-Making for Electromagnetic Spectrum Allocation Method Based on the Monte Carlo Counterfactual Regret Minimization Algorithm in Complex Environments. Atmosphere, 2025, 16(3): 345.
[3] Hauschild M Z, McKone T E, Karsten N A, et al. Risk and Sustainability: Trade-offs and Synergies for Robust Decision Making. Environmental Sciences Europe, 2022, 34(1).
[4] Chen Y, Wang J, Li R, et al. Particulate Matter 2.5 Concentration Prediction System Based on Uncertainty Analysis and Multi-Model Integration. The Science of the Total Environment, 2024, 958: 177924.
[5] Upreti B B, Samui S, Dey S R. Electrochemical Energy Storage Enhanced by Intermediate Layer Stacking of Heteroatom-Enriched Covalent Organic Polymers in Exfoliated Graphene. Nanoscale, 2025.
[6] Philip S, Marakkath N. Compressive Strength Prediction and Feature Analysis for GGBS-Based Geopolymer Concrete Using Optimized XGBoost and SHAP: A Comparative Study of Optimization Algorithms and Experimental Validation. Journal of Building Engineering, 2025, 108: 112879.
[7] Zhang T, Li Z, Zhang Z, et al. Machine Learning-Based Energy Consumption Models for Rural Housing Envelope Retrofits Incorporating Uncertainty: A Case Study in Jiaxian, China. Case Studies in Thermal Engineering, 2025, 72: 106253.
[8] Johnston H, Nair N, Du D. Estimating Calibrated Risks Using Focal Loss and Gradient-Boosted Trees for Clinical Risk Prediction. Electronics, 2025, 14(9): 1838.
[9] René P, Marja F, Martin A S, et al. Random Forest and Gradient Boosted Trees for Patient Individualized Contrast Agent Dose Reduction in CT Angiography. Studies in Health Technology and Informatics, 2023, 302: 952-956.
[10] Sun X, Wang Y, Cheng Z, et al. Deep Learning Framework Incorporating Simultaneous Optimization and Training for Concurrent Estimation and Prediction of Battery State of Health. Journal of Power Sources, 2025, 644: 237027.
[11] Xin X, Maokun L, Tingting X. Epilepsy Seizures Prediction Based on Nonlinear Features of EEG Signal and Gradient Boosting Decision Tree. International Journal of Environmental Research and Public Health, 2022, 19(18): 11326.
[12] Garitta C, Grassi L. Predicting Break-Even in FinTech Startups as a Signal for Success. Finance Research Letters, 2025, 74: 106735.

World Journal of Information Technology
Print ISSN: 2959-9903
Online ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3051

QUANTITATIVE ASSESSMENT AND COMPARATIVE STUDY OF NATIONAL CYBERSECURITY POSTURE BASED ON GLOBAL CYBERSECURITY INDEX

ZiHan Jin
Hohai University, Nanjing 210000, Jiangsu, China.
Corresponding Email: 1401062329@qq.com

Abstract: The Internet is indispensable for everyone, but with its rapid development, the accompanying problem of cybercrime has gradually become a significant challenge that countries worldwide must face. In order to assist countries in formulating better policies and establishing effective cybersecurity models, this study conducts an in-depth analysis of the effectiveness of existing policies from multiple dimensions, based on the GCI scoring mechanism. The research involved the collection of substantial data related to cybercrime, leading to the creation of heat maps depicting the number of cybercrime cases, the GCI scoring level, and the prosecution rates of cybercrime cases. The analysis and comparison of these graphs revealed that developed countries in Europe and the United States are the primary targets of cybercrime and have a high probability of sanctioning such crimes. Additionally, a ridge regression model was established based on the demographic characteristics of each country to examine the relationship between cybercrime cases and factors such as GDP, population size, Internet penetration rate, education penetration rate, and policy implementation environment. The coefficients for these factors were found to be 1. 22, 0. 54, 2. 55, -1. 46, and -0. 45, respectively, with population size being the most influential factor in the number of cybercrime cases. A sensitivity analysis further confirmed this finding. The study also classified the cybersecurity policies of various countries based on the five dimensions of the GCI and used a Difference-in-Differences (DID) model to evaluate the effectiveness of these policies. The results revealed that the most effective policy types differ across countries. International cooperation proved most effective in less developed countries, lawmaking in developing countries, and technological upgrading in developed countries.
Keywords: Cybercrime; Cybersecurity; Ridge regression model; GCI; DID model

1 INTRODUCTION

Cybersecurity challenges are becoming increasingly grave, resulting in substantial financial losses for both individuals and enterprises. Phishing, ransomware, DDoS assaults, and online fraud on social media constitute significant dangers. Confronting these difficulties necessitates a collaborative endeavor among individuals, enterprises, and governmental bodies.

The increasing interdependence of contemporary technology has revolutionized worldwide communication and commerce while simultaneously creating new vulnerabilities, since cybercrime presents a considerable risk to national and international security. Cybercrimes frequently transcend national borders, resulting in jurisdictional complications, while many industries refrain from reporting breaches, thereby allowing hackers to function without consequence. In response to these challenges, numerous countries have established national cybersecurity regulations, while the International Telecommunication Union (ITU) is crucial in establishing global standards and promoting international collaboration in cybersecurity initiatives. The current and future cybersecurity workforce in the public and private sectors are essential frontline defenders of our nation's digital infrastructure[1].

In the existing literature system of cybersecurity, there are several pressing issues that need to be addressed. Firstly, although various cybersecurity protection measures have been proposed, such as enhancing information security through kryptogra phishing message authentication codes (MAC)[2], these methods are often limited to a single technical dimension and fail to fully consider the complexity of multi-dimensional threats. Secondly, Axel Wirth and Christopher Falkner point out that solving cybersecurity issues typically requires "stakeholder cooperation," but in reality, such cooperation often becomes superficial, resulting in "a lot of talking but little action"[3]. This indicates that the effectiveness of existing collaboration mechanisms is insufficient in practice. Additionally, the current cyber threat environment is becoming increasingly complex, with frequent security incidents, continuously emerging vulnerability information, and a large amount of IOC (Indicators of Compromise) data [4], making the management and analysis of threat intelligence more challenging. Although Li Aichao and Fu Qiyang emphasize the two major aspects of computer network security—physical security and logical security[5]—existing research still shows significant shortcomings in the implementation of logical security, particularly in the comprehensive assurance of information integrity, confidentiality, and availability. Finally, current cybersecurity threat intelligence management technologies tend to focus on in-depth analysis of a specific type of threat intelligence, lacking effective integration and deep mining of multiple intelligence sources[6]. This fragmented approach limits the overall utilization efficiency of threat intelligence, hindering the ability to comprehensively address the evolving and multi-faceted nature of modern cyber threats.

Given the severity of cybersecurity challenges, the study should pay attention to the current key issues. The ultimate goal is to assess the effectiveness and resilience of a country's cybersecurity system and provide data-driven recommendations and policies to enhance protection and future development.

First, the study analyzed the Global Cybersecurity Index (GCI), the Verizon Cybersecurity Database (VCDB), and the Cybersecurity Risk Index (CEI) to map cybercrime hotspots and cybersecurity assessments for each country.

Second, the study constructed a ridge regression model on the number of cybercrimes, taking into account economic level, population size, Internet penetration rate, education level, and the policy implementation environment for cybersecurity as the main influencing factors. The model verified that the frequency of cybercrime incidents was linearly related to these demographic indicators in each country.

Thirdly, based on the cybersecurity policy evaluation model, the study introduce the Difference in the Difference-in-Differences (DID) model and propose a reasonable assumption regarding the timing of policy implementation and actual conditions to quantify policy effectiveness. The study categorize policies into five major categories and discuss them in the context of countries at different levels of development. Finally, the study analyzed policy effectiveness by comparing model predictions with actual outcomes.

Otherwise, the results derived from models reveal the diverse impacts on achieving global cybersecurity health and sustainability. The study conclude with an analysis and evaluation of the strengths and weaknesses of the model implementation.

2 DATA SELECTION AND ANALYSIS ON GLOBAL GCI

2.1 Data Selection and Analysis

The study gathered GCI-related score data, VCBD Website data, and additional information about current and historical cybercrimes to construct pertinent datasets for assessing policy success. Furthermore, the research gathered several national demographic attributes of countries globally, including population size and GDP, for subsequent analysis. To guarantee the thoroughness and credibility of the data, the research designated the following The websites as sources, as shown in Table 1.

Table 1 Data Sources

Data Item	Source
GDP	World Bank
Population	World Bank
Global Network Coverage	World Economic Forum
Primary School Completion Rate	World Bank
GCI (Global Cybersecurity Index)	ITU (International Telecommunication Union)
VCDB Index	VERIS
Global Crime Prosecution Rate	United Nations Office on Drugs and Crime, IRS, etc.

2.2 Data Processing and Visualization

The accessibility of data is essential, as inaccurate or erroneous information might hinder a precise evaluation of overall equity.

Consequently, the continuity and authenticity of the data must be guaranteed. Nonetheless, data inadequacies frequently result from inadequate disclosure by countries. The study employs many ways to guarantee data integrity, including regression analysis and averaging techniques.

These approaches preserve data integrity, which is crucial for precise and dependable analysis. In certain models, other data processing approaches will be employed.

Simultaneously, the study has employed visualization processing techniques to facilitate data analysis.

2.2.1 Cybersecurity index assessment

Recognizing cybersecurity as a global concern, nations worldwide have allocated specific efforts in this domain. The International Telecommunication Union (ITU) has evaluated the cybersecurity index of several nations based on five criteria: legal, technological, organizational, capacity building, and collaboration, culminating in the Global Cybersecurity Index (GCI) score for each nation. By aggregating the GCI scores of 193 nations, the study has developed a GCI score level map (Figure 1). The map indicates that nations with elevated security standards are predominantly located in Europe and the Americas, whereas countries in Africa, Asia, and other areas exhibit lower cybersecurity scores. Furthermore, it has been determined that underdeveloped countries exhibit lower cybersecurity index scores compared to both developed and developing nations, whereas developed countries possess greater cybersecurity index levels than their developing counterparts.

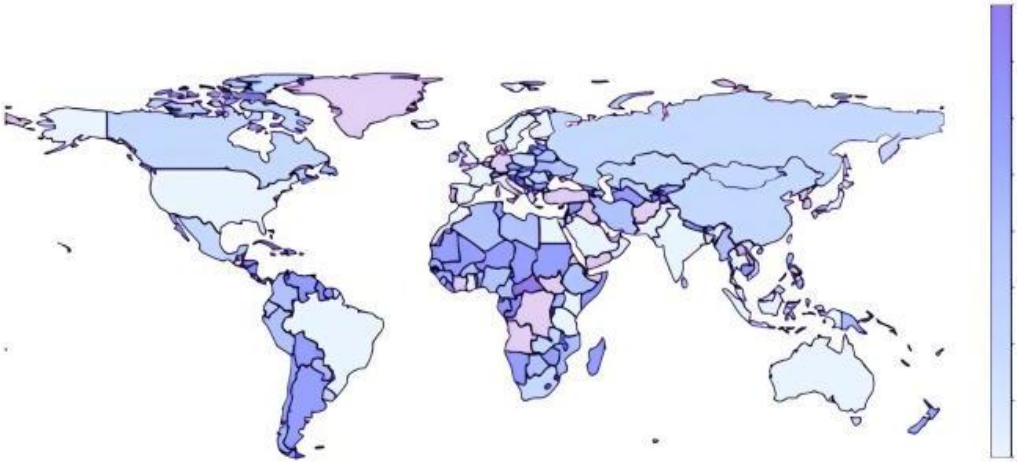

Figure 1 GCI Score Level Map
Source: VERIS Community Database (VCDB)

2.2.2 Crime volume analysis

Cybercrime is a highly significant criminal issue in todays society, especially with the rapid advancement of computer technology and the widespread adoption of the Internet.

The type and volume of cybercrime are on the rise. By collecting data on the number of cybercrime cases from 140 countries, the study has created a global heatmap of cybercrime volume (Figure 2) to illustrate the distribution of cybercrime cases. The heatmap reveals that developed countries have a significantly higher absolute number of cybercrime cases compared to developing countries, with the United States experiencing a far greater number of cybercrime incidents than other nations.

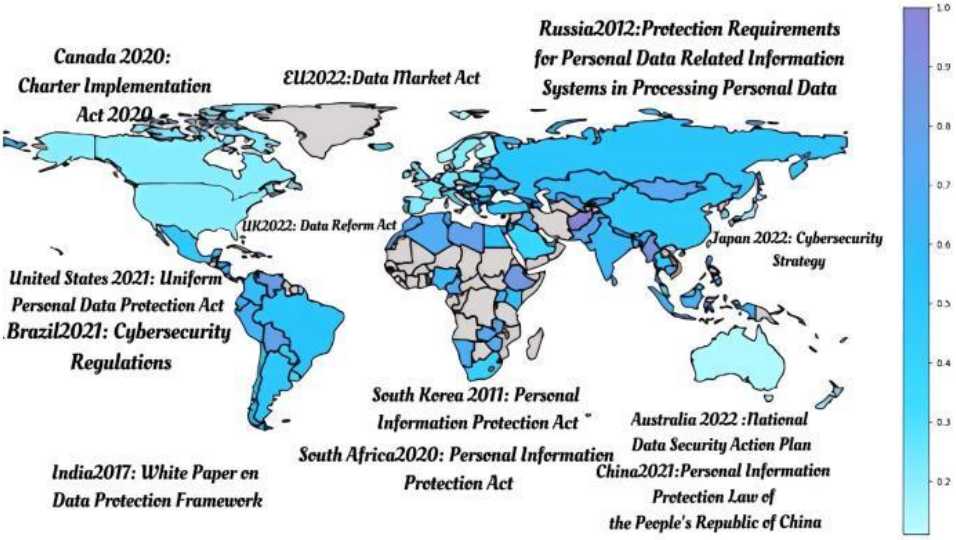

Figure 2 National Crime Statistics and some Policies
Source: CrimeMapping.com

2.2.3 Legal success rate analysis

The prosecution rate of cybercrime is an excellent indicator of a countrys cybersecurity strength. By collecting data on the number of cybercrime cases prosecuted in various countries and combining it with the overall number of cybercrime cases in each country, the research has created a global map of cybercrime prosecution rates (Figure 3). On the map, it can be observed that the prosecution rate for cybercrime in the European and American regions is significantly higher than that in the Asian and African regions.

Additionally, developed countries generally have higher prosecution rates for criminal cases than developing countries. However, some developing countries also have relatively high prosecution rates for criminal cases.

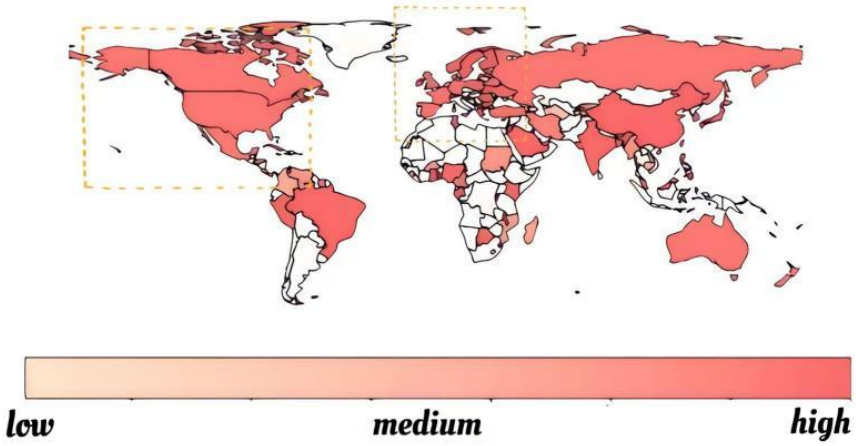

Figure 3 National Crime Prosecution Rate
Source: CrimeMapping.com

2.3 Policy Data Analysis

National policies serve as the principal mechanism for improving a nation's cybersecurity index. The study picked eleven nations based on their national development levels, gathered and assessed their cybersecurity policies, and classified them according to the five characteristics outlined by the Global Cybersecurity Index (GCI). The resultant policy classification chart (Figure 4-5) indicates that policies in industrialized, developing, and underdeveloped nations predominantly emphasize the legal dimension. Moreover, wealthy nations exhibit a wide array of policies across all dimensions, including the legal aspect, whereas emerging and undeveloped countries possess comparatively fewer policies in the other dimensions.

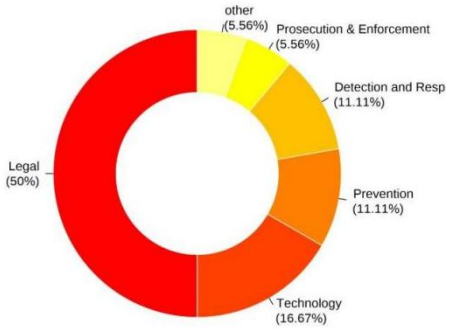

Figure 4 Distribution Ratio of Policy Types in Low Network Security Countries

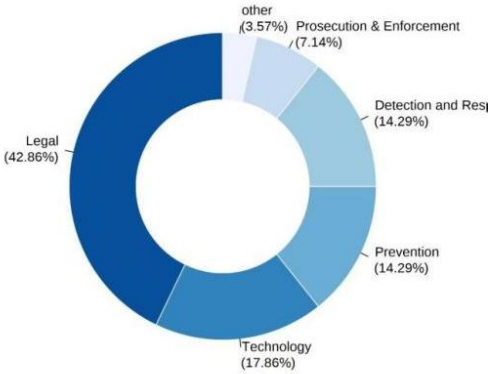

Figure 5 Distribution Ratio of Policy Types in High Network Security Countries

3 EVALUATION SYSTEM AND MODEL

To comprehensively analyze and predict the factors influencing the prevalence of cybercrime across different nations, the study established the Multi-Factor Regression Assessment Model (MRAM). Through this model, the study defined a mathematical relationship to quantify cybercrime levels C. The study considered five pivotal dimensions: GDP per capita, Population, Secure internet servers, Policy effectiveness, and Education level, which together serve as the key independent variables in the framework.
The model coefficients represent the magnitude and direction of each factor's contribution to cybercrime, which will be interpreted in subsequent sections.

3.1 Main Factors

To conduct a comprehensive evaluation of the cybersecurity landscape, it is essential to measure its robustness and adaptability. Similar to the definition of system health, the robustness of cybersecurity refers to its ability to effectively counter threats and adapt to the ever-changing challenges in the digital domain. For a resilient cybersecurity system model, it is not only necessary to identify cybercrime factors based on national development conditions but also to cover potential cybersecurity issues within the country. This study divides the task of data transmission into multiple transmission units Each transmission unit models the network condition based on an autoregressive model, predicting the required deduplication processing time for the next transmission unit and the available bandwidth during this period[7]. Based on the above definitions, the research has identified five key factors that measure the cybersecurity system, which will be detailed below:
(1) GDP per capita
The GDP per capita factor measures the likelihood of criminal groups targeting citizens of a country and is one of the most important indicators. On the one hand, a higher level of economic development usually means better social welfare and more employment opportunities, which may reduce the crime rate. On the other hand, economic growth can also lead to an increase in the wealth gap, which may increase certain types of crime rates. At the same time, criminal groups are more likely to target wealthier individuals to obtain higher illegal profits.
(2) Population
The population factor measures the potential risk a country faces from cybersecurity threats. The larger the population, the more Internet users there are, and the more targets there are for cyberattacks. This factor is measured by the population size, reflecting the pressure on cybersecurity in a country.
(3) Internet Penetration Rate
The number of secure Internet servers is a measure of a country's cybersecurity infrastructure. A higher number of secure Internet servers indicates a stronger capability for cyber defense. This factor is measured by the quantity of secure Internet servers, reflecting a country's investment and capability in cybersecurity technology.
(4) Education level
The education level factor measures a country's investment and effectiveness in cybersecurity education. A higher education level means more professional talent and stronger cybersecurity awareness. This factor is measured by indicators such as the number of graduates in cybersecurity-related fields and the prevalence of cybersecurity training.
(5) Policy effectiveness
The policy effectiveness factor measures the implementation results of a country's cybersecurity policies. Effective policies can significantly reduce the rate of cybercrime and enhance the level of cybersecurity. This factor is measured by indicators such as the enforcement of cybersecurity laws and the frequency of policy updates.

3.2 Evaluation Model

The study assumes that the number of cybercrimes (C_i) is influenced by the variables mentioned . And the regression model is expressed as:

$$C_i = \beta_0 + \beta_1 P_i + \beta_2 N_i + \beta_3 S_i + \beta_4 E_i + \beta_5 O_i + \epsilon_i \tag{1}$$

P_i represents GDP per capita, which indicates the level of economic development; N_i represents population size, which indicates the potential scale of internet users; S_i represents the number of secure internet servers, reflecting the strength of network security infrastructure; E_i represents education level, measuring public awareness and capacity in cybersecurity; O_i represents policy effectiveness, which scores the country's policy conditions in global cybersecurity efforts; β_0 is the intercept, representing the baseline cybercrime level when all variables are zero; $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ are the regression coefficients, indicating the influence of each variable on the cybercrime level; and ϵ_i is the error term, capturing unexplained random factors affecting cybercrime. The letters in the following equation represent the same meanings.
The Residual Sum of Squares (RSS) is a measure of the discrepancy between the observed values and the values predicted by a regression model. In the context of the given problem, the RSS is given by the formula:

$$RSS = \sum_{i=1}^{n} \left(C_i - \hat{C}_i\right)^2 \tag{2}$$

The Residual Sum of Squares (RSS) is a measure of the discrepancy between the observed values and the values predicted by a regression model. In the context of the given problem, the RSS is given by the formula:

$$\hat{C}_i = \beta_0 + \beta_1 P_l + \beta_2 N_i + \beta_3 S_i + \beta_4 E_i + \beta_5 O_i \tag{3}$$

The Ordinary Least Squares solution:

$$\beta = (X^{\top} X)^{-1} X^{\top} Y \tag{4}$$

The study used VCDB statistics to draw a global cybercrime heat map, selected typical countries, and drew a line chart of cybercrime changes each year. The study also selected the United States and Australia and marked the laws enacted by these two countries where the line chart showed significant changes.
The results of calculations are presented in Table 2:

Table 2 Regression Results

Variable	Coefficient	Std. Error	t-Statistic	P-Value
Constant	256. 3456	56. 156	11. 1345	0. 074
GDP per capita	0. 0321	0. 011	2. 5634	0. 421
Population	0. 0452	0. 015	3. 1289	0. 417
Internet Penetration Rate	0. 0234	0. 010	2. 3487	0. 229
Education level	-2. 1234	0. 021	-12. 3489	0. 251
policy effectiveness	-4. 587	0. 019	-6. 3451	0. 135

The correlation of parameters in the formula is shown in Figure 6:

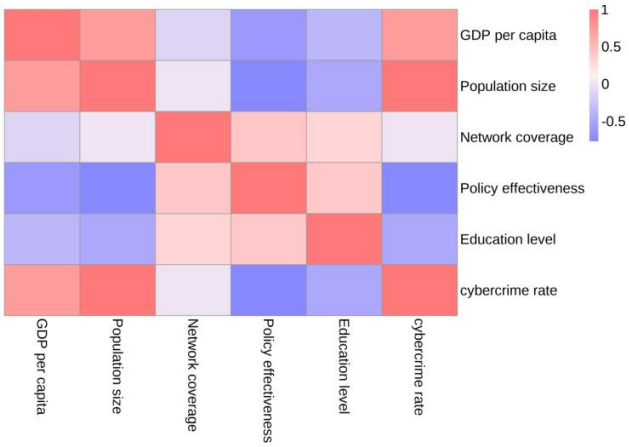

Figure 6 Objective Impact Indicators of Various Indicators

Light colors indicate lower correlation, and dark colors represent higher correlation.
The correlation between parameters cannot be ignored, as can be seen from the figure. Therefore, it is necessary to improve the model on this basis.

3.3 Improved Model

Incomplete data is stored in complex distributed networks [8-9], involving the insertion, deletion, updating, and reading of data, making it prone to redundancy, anomalies, or missing data [10]. To make the regression coefficients more stable, reduce the impact of multicollinearity, and maintain the predictive power of the model, the study calculates the Variance Inflation Factor(VIF) for each variable to evaluate potential multicollinearity issues:

$$VIF(N_i) = \frac{1}{1 - R_j^2} \tag{5}$$

In this equation, R_j^2 is the R^2 value obtained by regressing N_i on all other predictor variables.
To address this, the study turns to the Ridge Regression model[10]. The basic equation of the Ridge Regression model is:

$$\hat{\beta} = \underset{\beta}{argmin} \left(\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2 + k \sum_{j=1}^{p} \beta_j^2 \right) \tag{6}$$

k is the Ridge Regression parameter used for regularization to penalize excessively large regression coefficients.

The results of calculations are presented in Table 3 :

Table 3 Multiple Linear Regression Results

Variable	B.	S. E.	Beta	t	P>|t|	VIF
Constant	277. 36	0. 12	------	19. 90	0. 087	------
GDP per capita	1. 22	0. 15	0. 42	3. 15	0. 035	1. 25
Population	0. 54	0. 12	0. 32	3. 76	0. 037	1. 20
Internet Penetratio--n Rate	2. 55	0. 20	0. 56	2. 55	0. 026	1. 30
Education level	-1. 46	0. 18	-0. 38	-11. 33	0. 046	1. 40
policy effectiveness	-0. 45	0. 10	-0. 15	-5. 08	0. 041	1. 1

The regression equation the study calculated is:

$$y = 277.36 + 1.22x_1 + 0.54x_2 + 2.55x_3\text{-}1.46x_4\text{-}0.45x_5 \tag{7}$$

The study has collected the data as a dataset from which the study randomly selected four countries and predicted the y-values, and the fitting results are shown in Figure 7. This shows that the study is relatively feasible within a certain error allowance.

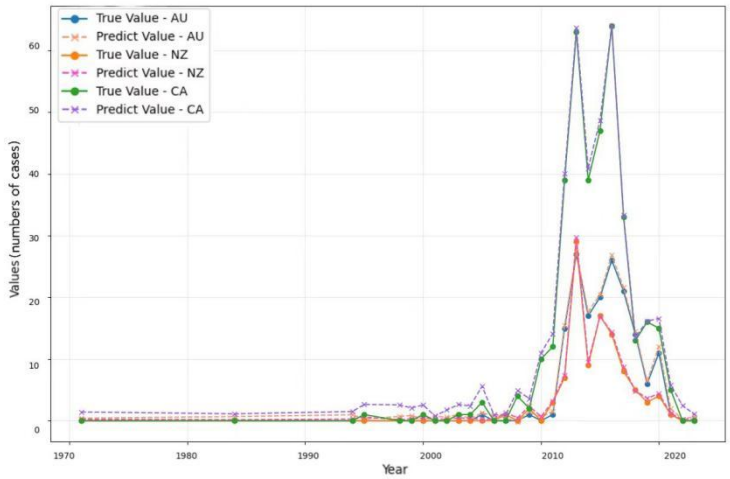

Figure 7 Testing results of regression model

3.4 Sensitivity Analysis

The study calculated the sensitivity of each explanatory variable to the dependent variable using the following formula:

$$S_i(Y,x_i) = \frac{1}{n}\frac{\partial x_t}{\partial y_i}\cdot\frac{x_i}{y_t}, i = 1,2,3,4,5 \tag{8}$$

In the previous part, the Ridge Regression Model was constructed and the proximity between the countries and the optimal solution was obtained. Then, the study conducted sensitivity analysis on the Ridge Regression Model respectively. The specific methods are as follows: Select Central African Republic as the research object, keep the weights of each indicator fixed, and gradually increase the value of each indicator individually, and observe the changes in the degree of proximity (assessment score) to the optimal solution. The faster the evaluation score changes with the index value, the higher the sensitivity of the model. Conversely, the slower the evaluation score changes with the index value, the lower the sensitivity of the model. Figure 8 respectively represents the sensitivity of indicators in the selected single country and the Average prediction sensitivity of indicators.

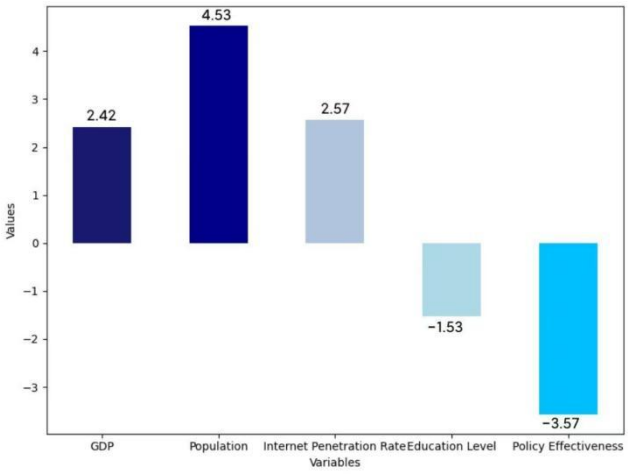

Figure 8 Average Prediction Sensitivity

According to the above results, it can be found that population size, education penetration rate, and the policy implementation environment have a greater impact on the number of cybercrimes. Therefore, since the member states of the International Telecommunication Union hope to build network security together, they must promulgate policies to try their best to improve domestic education penetration rates, maintain a stable domestic political environment, and reduce the number of cybercrimes. At present, some countries have not paid enough attention to related aspects, so the study strongly recommended that member states pay more attention to issues such as education that accompany population growth.

4 EVALUATION OF NATIONAL POLICIES

The study aimed to analyze the effectiveness of the cybersecurity policy. Since the factors influencing cybersecurity may be subject to sample selection limitations or difficulties in controlling other factors, which could lead to the possibility of overestimating the estimates, the study adopted the DID model proposed by Ashenfelter and Card (1985), originally used to study whether participation in government vocational training programs could increase participants' earnings. The study divided the sample into an experimental group and a control group, and subtracted the pre-policy differences of the control group from the differences between the experimental and control groups. The predicted data obtained through the aforementioned ridge regression model were used to assess the differences between the two groups before and after the policy, in order to analyze the true policy effect.

4.1The Foundation of Model

To clarify whether the policy had an impact on cybercrime rates, the study constructed an experimental group (regions affected by the cybersecurity policy) and a control group (regions not affected by the policy), and compared the differences in cybercrime rates before and after the policy implementation. This approach allowed us to identify the true impact of the policy on cybercrime rates. By comparing the changes in cybercrime rates between the experimental and control groups, the study could more accurately evaluate whether the cybersecurity policy achieved its intended effect, and avoid bias caused by unobserved confounding factors (such as regional economic development levels, law enforcement efforts, etc.).

The model used for this analysis is:

$$Y_{it} = \alpha + \beta_1 Time_t + \beta_2 AREA_i + \beta_{12}(Time_t \times AREA_i) + \epsilon_{it} \quad (9)$$

Y_{it} represents the number of cybercrimes in region i at time t ; α is the constant term; β_1 is the coefficient of the time dummy variable, representing the overall before and after the policy implementation; β_2 is the coefficient of the region dummy variable, representing the baseline differences across regions; β_{12} is the interaction effect coefficient, representing the differential impact of the policy across regions; $Time_t$ is the time dummy variable, which is 0 before the policy implementation and 1 after; $AREA_i$ is the region dummy variable, indicating whether the region is affected by the policy (1 if affected, 0 otherwise); and ϵ_{it} is the error term.

In the preliminary data processing, the study first categorized national policies into five types based on the standards of the Global Cybersecurity Index (GCI). In the policy classification, the study assumed that a type of policy only affects the corresponding policy indicator. Additionally, for policies issued, the study assumed that their impact on cybersecurity gradually increases over five years and remains constant after five years. By making these assumptions, the study simplified the DID model to make it more convenient and efficient for practical application. The principle of the DID model is shown in Figure 9.

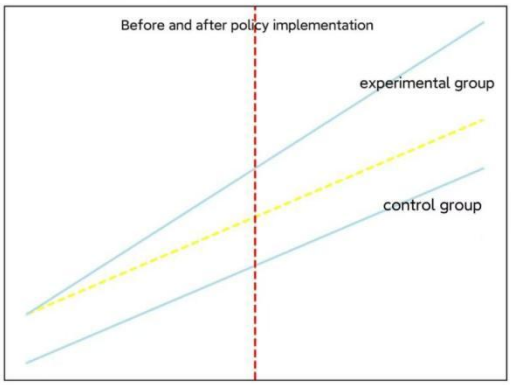

Figure 9 DID Model Principle Diagram

Due to the complexity of cybersecurity issues, in order to demonstrate predicted outcomes in reality, it is necessary to create a virtual parallel world to study the impact of policies. Therefore, the study adopted the DID model to evaluate the policies of various countries. Simply put, the DID model assumes the existence of a virtually conceived parallel world, where the study identify a region similar to the policy implementation area but without the policy in place, and then compare the differences between the two regions to assess the policy's effectiveness.

The study conducted benchmark regressions on the model through parallel trend tests, placebo tests, and robustness tests: Parallel Trend Test: By adding interaction terms to the regression model, the study examined whether the trends between the treatment group and the control group were significantly different before the policy intervention. After conducting three sets of interaction tests, the study found that the trends of the treatment and control groups were similar before the intervention, indicating that the parallel trend assumption holds.

Placebo Test: The study introduced a fake policy intervention period (placebo period) in the regression analysis the study chose a time point unrelated to the actual intervention period for regression. The regression results from the placebo period were insignificant, leading us to conclude that the actual intervention effects are reliable. Robustness Test: The study performed benchmark regressions using different error structures, specifically clustered standard errors, to examine the overall effects of the policy intervention.

4.2 Parallel Trends Analysis

This study implements a parallel trends test to verify the accuracy of the data. An interaction term of time dummy variables and group dummy variables is developed to elucidate the disparities between the model-generated parallel data and actual data before the policy execution, illustrating the variations between the treatment and control groups across several years.

Pertinent variables were incorporated into the model for evaluation. The results from the parallel trends test graph indicate that the coefficients for Before2 and Before1 lack statistical significance, whereas the coefficients for After1 and After2 are significant ($p = 0.095 < 0.1$), demonstrating a positive impact on the digital economy value index in the respective countries following the implementation of the cybersecurity policy. The coefficient for After2 is 0, indicating that the coefficients prior to the policy implementation varied about 0, whereas the coefficients after the policy implementation tend toward a positive value.

This indicates that the treatment and control groups display similar trends, thus fulfilling the necessary conditions for the use of the DID model, which adheres to the parallel trends assumption.

4.3 Application of Models

The study selected the policies of eleven typical countries, where the timing of policy impact on the DID model varied, and evaluated the policies as listed in the Table 4:

Table 4 Partial National Policy Rating

Effectiveness Rating	Law Name
1	Internet Governance Policy (2015)
	ICT Strategy and Cybersecurity Development Strategy (2016)
	Cybersecurity Education and Public Awareness Program (2018)
	Cybersecurity Education and Awareness Program (2018)
\	Cybersecurity Law (2014)
	Data Protection Law (2019)
2	Cybersecurity Law (2015)
	Personal Data Protection Law (2017)
	Cybercrime Law (2011)
	Information Technology Law (2015)

2	2 Anti-Terrorism Law (Revised 2005) Cybercrime Law (2017) Data Protection Law (2017) Electronic Transactions Act (2008) Data Protection Law (2012) Information Technology Law (2013) Data Protection Law (2020)
3	Cybersecurity Law (2017) Data Security Law (2021) Personal Information Protection Law (2021) Cybersecurity Level Protection System (Class 2. 0, 2019) Critical Information Infrastructure Protection Regulations (2021) Computer Fraud and Abuse Act (CFAA, 1986) 3 Cybersecurity Law (2018) U. S. Cybersecurity Strategy (2018) Anti-Terrorism Act (2001) Cybersecurity Act (2016) Data Protection Act (2018) Cybersecurity Strategy (2016) Emergency Response Mechanism (UK CERT, 2001)

In the Table 4, observing the policies in the table, those rated 3—such as Cybersecurity Law (2017), Data Security Law (2021), and Personal Information Protection Law (2021)—demonstrate a particularly robust and comprehensive approach to constructing a cybersecurity framework. These policies not only secure critical infrastructure through clear legal provisions, but they also exhibit powerful enforcement capabilities that address emerging cyber threats and combat cybercrime effectively, forming a wide-ranging security barrier. At the same time, the policies rated 2, including Cybersecurity Law (2014), Data Protection Law (2019), and Electronic Transactions Act (2008), contribute significantly to enhancing overall cybersecurity. Although there are some shortcomings in terms of implementation details or inter-sector coordination, these measures still manage to fortify security defenses and progressively refine the regulatory framework. In contrast, the policies rated 1—such as Internet Governance Policy (2015) and Cybersecurity Education and Public Awareness Program (2018)—primarily focus on providing strategic guidance and establishing an initial framework. While these initiatives lay an important foundation for future, more stringent measures, their impact remains relatively limited in the short term due to a lack of direct and robust enforcement mechanisms.

From the table, the study observed that the policy implementation efficiency in less developed countries is mostly 1, while developed countries have more policies with efficiency ratings of 2 and 3. In the policy types of less developed countries, legal measures, which focus on post-crime sanctions, are the most common. Due to the lower cybersecurity literacy in less developed countries, the crime reporting rate is lower compared to developed countries.

4.4 Solution and Result

For developed countries, such as the United States, high levels of internet coverage result in higher cybersecurity literacy among citizens. The policy effect diagram of the United States is shown in Figure 10. Although the proportion of policies and laws formulated is the highest, they demonstrate a comprehensive approach when compared to underdeveloped and developing countries. As a result, they have established a complete system for governing cybercrime.

The study has also drawn on this system for model analysis, which has helped us formulate effective policies:Due to the generally low cybersecurity literacy in underdeveloped countries, cybersecurity issues are more prominent in these countries. In these nations, the public and government agencies have relatively weak understanding and response capabilities concerning cybercrime, which leads to a lower rate of cybercrime reporting. This phenomenon contrasts sharply with developed countries, which typically have higher cybersecurity awareness and more established reporting mechanisms.

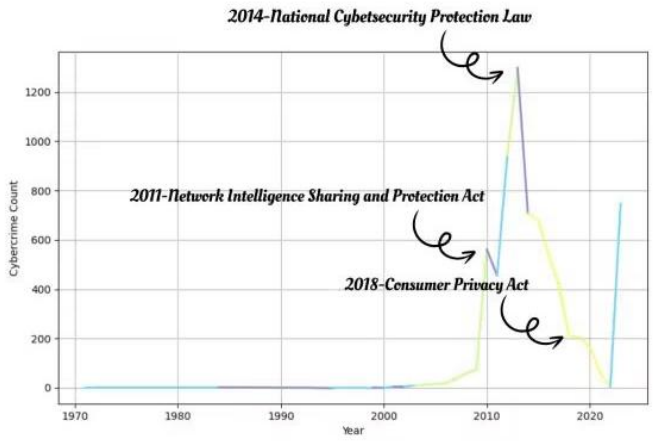

Figure 10 The Application of the Model in US

Based on this difference, when formulating cybersecurity policies, underdeveloped countries should avoid relying solely on legal measures to address cybercrime. While legal measures can play a role in regulation, relying only on the law is insufficient to effectively improve cybersecurity, particularly when cybersecurity literacy is low. Meanwhile, international cooperation shows significant potential for improving cybersecurity quality. By collaborating with developed countries, underdeveloped countries can gain technical support, share resources, and exchange experiences benefits that cannot be achieved by acting independently. Developed countries have accumulated rich experience and technical resources in the cybersecurity field, and their well-established cybersecurity frameworks, response mechanisms, and policy systems can provide valuable guidance for underdeveloped countries. Through enhanced cooperation with developed countries, underdeveloped countries can not only improve their cybersecurity defenses but also enhance their voice and participation in international cybersecurity governance.

Therefore, underdeveloped countries should regard international cooperation as a key component of cybersecurity policy formulation. They should actively engage in cooperation with developed countries and international organizations to address gaps in global cybersecurity challenges. Such cooperation will not only help underdeveloped countries compensate for their lack of technology and experience but also contribute to the improvement of the global cybersecurity environment, thus achieving global cogovernance and shared development in cybersecurity.

5 CONCLUSIONS

This study examines the effectiveness of cybersecurity policies and cybercrime, based on the GCI scoring mechanism. By collecting substantial data related to cybercrime, the study created heat maps depicting the number of cybercrime cases, GCI scores, and prosecution rates across countries. The analysis revealed that developed countries in Europe and the United States are the primary targets of cybercrime and are more likely to sanction such crimes. A ridge regression model was built to analyze the relationship between cybercrime cases and factors such as GDP, population size, internet penetration rate, education level, and policy implementation environment. The findings showed that population size is the most influential factor in determining the number of cybercrime cases. Additionally, a DID model was used to evaluate the effectiveness of cybersecurity policies, revealing that the most effective policy types vary across countries: international cooperation was most effective in less-developed countries, legislation in developing countries, and technological upgrades in developed countries.

The limitations of this study lie in the fact that, due to time constraints, more relevant data could not be collected, limiting the depth of analysis of the model's variables. Furthermore, the model itself has inherent flaws that could lead to errors in assessing the actual situation, which may affect the precision and reliability of the study's results.

COMPETING INTERESTS

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

REFERENCES

[1] Wireless News. Billington CyberSecurity Summit to feature theme: Advancing cybersecurity's impact in age of heightened risk. 2023.
[2] Dirk Bierbaum. Smarte Synthese aus Cybersecurity und Funktionssicherheit. ATZ elektronik, 2022, 4(4): 341.
[3] Axel Wirth, Christopher Falkner. Cyberinsights : Cybersecurity as a Team Sport. Biomedical Instrumentation & Technology, 2020, 54(1): 64-67.
[4] Xia Ru. A review of research on foreign cyber threat intelligence. Modern Information Technology, 2024, 8(01): 189-192+198.

[5] Li Aichao, Fu Qiyang. Analysis of Computer Network Security Issues and Countermeasures. Engineering Technology: Abstract Edition, 2022(12).
[6] Zhao Xiaolin, Zeng Chonghan, Xue Jingfeng, et al. Research on Multidimensional Network Security Measurement Model Based on TOPSIS. Journal of Beijing Institute of Technology, 2021, 41(3): 311-321.
[7] Ye Pengdi, Yao Wenbin, Li Xiaoyong. Design of network data deduplication method based on autoregressive model. Journal of Beijing University of Posts and Telecommunications, 2014(4): 5.
[8] Yao Yingle, Li Jian, Sun Bin. Simulation of Interpolation Algorithm for Fitting Incomplete Data Missing Sequence. Computer Simulation, 2023, 40(1): 523-527.
[9] Ai Zhiwei, Leng Juelin, Xia Fang, et al. A method for reducing large-scale structured data with controllable accuracy. Journal of Computer Aided Design and Graphics, 2021, 33(12): 1795-1802.
[10] Guan Lijing, He Jiefan, Zhang Liyong, et al. Missing Value Imputation Method Based on Single Output Sub Network with Iterative Learning. Journal of Dalian University of Technology, 2022, 62(4): 427-432.

World Journal of Information Technology
Print ISSN: 2959-9903
Online ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3052

DIMENSIONALITY REDUCTION AND FITTING METHOD FOR HIGH-DIMENSIONAL DATA BASED ON SVD AND LEAST SQUARES—A CASE STUDY OF MINE DATA PROCESSING

JiaYuan Zhang
SWUFE-UD Institute of Data Science at SWUFE, Southwestern University of Finance and Economics, Chengdu 611130, Sichuan, China.
Corresponding Email: 18684018314@163.com

Abstract: In the digital era, the explosive growth of high-dimensional data poses significant challenges to storage, transmission, and computational efficiency. Mine data, characterized by its multi-source heterogeneity, high dynamism, and high dimensionality, presents particularly acute challenges. This paper proposes a novel method for dimensionality reduction and fitting of high-dimensional data by combining Singular Value Decomposition (SVD) with least squares, and demonstrates its first application in mine data processing. The method achieves efficient data compression and precise fitting by extracting principal singular values and vectors via SVD, projecting high-dimensional data into a low-dimensional space, and solving for optimal weight vectors using least squares. A pseudo-inverse is constructed to avoid numerical instability, ultimately completing the fitting of the target dataset. Experimental results show that the method performs exceptionally well in terms of residual distribution, model bias, data noise, and fitting adequacy: residuals approximate a normal distribution, confirming that errors primarily stem from data noise. This study provides a reliable technical pathway for processing high-dimensional mine data, with future optimizations possible through the introduction of noise reduction modules.
Keywords: SVD method; Least squares fitting; Data dimensionality reduction; Mine data processing; Error analysis

1 INTRODUCTION

In the current digital age, data across various fields is growing exponentially, with increasing complexity in dimensionality. While high-dimensional data contains rich information, it also presents formidable challenges, making research into high-dimensional data compression urgent. From a storage perspective, high-dimensional data occupies substantial space, forcing enterprises and institutions to expand hardware infrastructure, thereby driving up costs. In transmission, high-dimensional data requires prolonged transfer times, hindering real-time sharing and interaction. Moreover, high-dimensional data increases computational complexity, reducing the efficiency of data analysis and processing. Traditional algorithms often struggle to handle high-dimensional data, falling short of practical requirements.

The rapid advancement of mine monitoring technologies has also led to the generation of vast amounts of high-dimensional data. While its high resolution, dynamism, and dimensionality support critical tasks like geological modeling and resource assessment, the storage, transmission, and real-time processing of multi-source heterogeneous data remain problematic. The complexity of data fusion escalates computational resource demands, rendering traditional methods inadequate in balancing efficiency and precision.

Vats Deepak compared various common methods in demensionality reduction, mentioned the pros and cons of SVD method[1]. Hastie Trevor combined alternating least squares and SVD method to further solve matrix-completion problem[2]. M.E.Hochstenbach provide a novel method to improve the SVD decomposition efficiency of large matrix[3]. Alkiviadis G. Akritas explained how SVD can be applied to solve least squares problems and data compression[4]. Zhang Chongchong combined NAEEMD and frequency constrained SVD to denoising the mine microseismic signals[5]. Li Shanshan applied SVD to multi-label learning dimensionality reduction, improved classification efficiency[6]. Yang Xinyu innovatively combined K-SVD and SVD for wireless sensor network data, maintaining accuracy while drastically cutting energy consumption[7]. Li Ke enhanced high-dimensional data processing efficiency via an improved randomized SVD algorithm[8]. Zhu Quanjie and Tang Fei employed EMD-SVD and multi-layer SVD, respectively, for denoising mine microseismic signals, significantly improving signal quality[9][10]. These studies demonstrate the efficacy of SVD method in data processing domains. However, in mine data processing, existing research only limits it to denoising. Thus, this study innovatively integrates SVD with least squares, specifically targeting mine data characteristics to address dimensionality reduction and fitting challenges.

This paper's contributions are: (1)Combining SVD and least squares for high-dimensional data dimensionality reduction and fitting; (2)Pioneering the method's application in mine data dimensionality reduction; (3)Diagnosing error sources and assessing their impact on fitting results.

2 RELATED THEORIES

Singular Value Decomposition (SVD) is a fundamental data processing technique that extracts principal singular values and their corresponding vectors, projecting high-dimensional data into a low-dimensional space while retaining directions of maximum variance, thereby achieving dimensionality reduction.

For any real or complex matrix$A \in C^{m\times n}$（assuming m≥n）can be broken down into:

$$A = U\sum V^T \tag{1}$$

Where $V \in R^{m\times n}$contains right singular vectors (orthogonal basis of the original feature space).

The diagonal elements of $\sum$,$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$ indicate the importance of each direction.

Least squares is a mathematical optimization method for linear regression, minimizing the sum of squared errors to find the best-fitting curve or hyperplane. In matrix form:

$$y = X\beta + \varepsilon \tag{2}$$

Where $y \in R^n, X \in R^{n\times(p+1)}$(first column all 1s for intercept β_0), $\beta \in R^{p+1}$ (regression coefficients), and $\varepsilon \in R^n$(error vector).

The optimization problem can be formulated as follows:

$$\min_{\beta} \|y - X\beta\|_2^2 \tag{3}$$

Setting the derivative to zero yields:

$$\frac{\partial}{\partial\beta}\|y - X\beta\|_2^2 = -2X^T(y - X\beta) = 0 \tag{4}$$

Where X^TX is not singular：

$$\beta = (X^TX)^{-1}X^Ty \tag{5}$$

The geometric significance of least squares is to find the projection of $X\beta$ on the column space of X so that the residuals are orthogonal to the column space:

$$X^T(y - X\beta) = 0 \tag{6}$$

When X^TX is invertible ,Pseudo-inversion is required, and the SVD method is used in this paper to avoid direct inversion.

3 EXPERIMENT

Based on the SVD method and the least squares method, the weight vector is found to realize the dimensionality reduction of the original dataset and fit the target dataset as much as possible, and then the residuals are calculated, the residuals are analyzed, the source of the error is confirmed and the advantages and disadvantages of the model are evaluated, and the experimental flow chart is Fig.1.

This flowchart outlines an SVD-based regression modeling workflow: standardizing data, performing SVD decomposition for dimensionality reduction and regression fitting, then calculating predictions and residuals. Error analysis systematically evaluates four aspects: (1) residual distribution (normality/range), (2) model bias (residual-prediction correlation), (3) noise (autocorrelation) and (4) fit adequacy (R^2). The integrated process ensures stable high-dimensional computation and reliable modeling through comprehensive diagnostics, providing a complete data-to-evaluation pipeline.

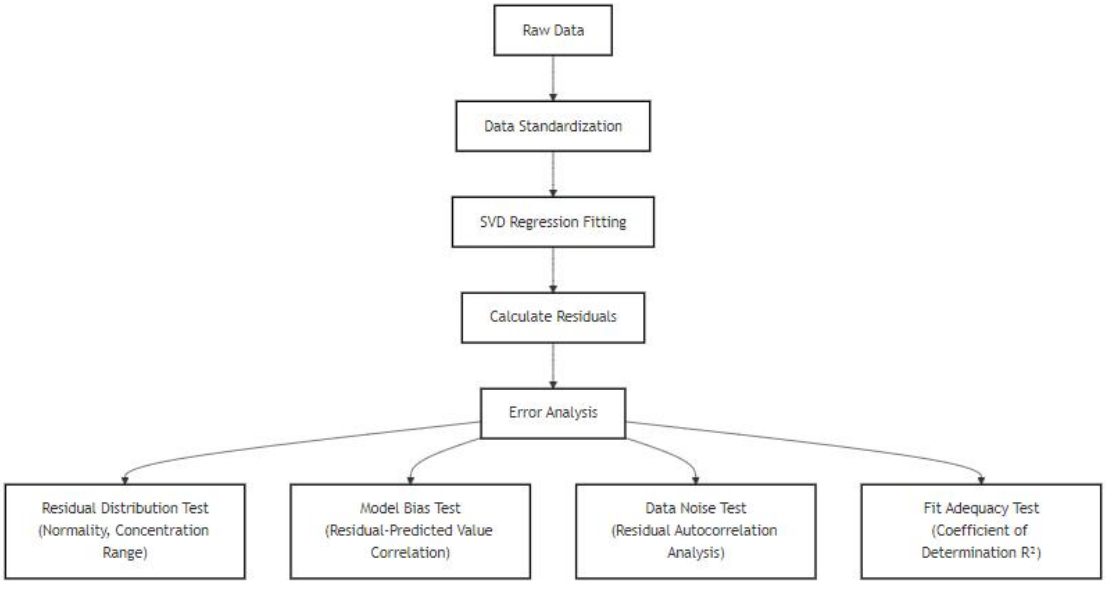

Figure 1 Flow Chart of the Experiment

It is known that in order to establish a mathematical model to reduce the dimensionality of a 10000*100 dataset X to a 10000*1 dataset N and fit another 10000*1 dataset y, it is necessary to find the weight vector so that N=Xβ and the

residuals are minimized. The solution is to normalize X and y, first decompose SVD to construct a pseudo-inverse, find the least squares solution, and then calculate the residuals. Perform a series of model evaluations and error analyses using residuals and raw data. The specific modeling process is as follows:

Firstly, X and y are standardized, and the data are converted into a distribution with a mean of 0 and a standard deviation of 1 according to columns (features) to eliminate dimensional differences, which is conducive to improving numerical stability.

$$X_{scaled} = \frac{X - \mu_X}{\sigma_X} \tag{7}$$

$$y_{scaled} = \frac{y - \mu_y}{\sigma_y} \tag{8}$$

Where μ_X, μ_y is the average value of each column, σ_X, σ_y is the standard deviation of each column.

Secondly, the centralized matrix is decomposed.

$$X = U \sum V^T \tag{9}$$

$U \in R^{m \times m}$ is the left singular vector matrix, and the column vector is orthogonal.

$\sum \in R^{m \times n}$ is a semi-positive definite diagonal matrix, Diagonal elements are singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{min(m,n)} \geq 0$, Indicates the importance of each ingredient.

$V \in R^{n \times n}$ is the right singular vector matrix, and the column vector is orthogonal.

Then we need to construct the Moore-Penrose pseudo-inverse X^+ that defines X in the SVD:

$$X^+ = V \sum^+ U^T \tag{10}$$

Where $\sum^+$ is a diagonal matrix, which is obtained by taking the reciprocal of each non-zero element of $\sum$ and transposing it:

$$\sum^+ = \begin{bmatrix} 1/\sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/\sigma_n \end{bmatrix} \tag{11}$$

With pseudo-inverse, the explicit expression of the least squares solution can be expressed as:

$$\beta = X^+ y = V \sum^+ U^T y \tag{12}$$

This method can avoid the numerical instability issues caused by direct inversion. The residuals and MSE are then calculated to assess the model:

$$\varepsilon = y - X\beta \tag{13}$$

$$MSE = \frac{1}{n} \sum_{i=1}^{n} \varepsilon^2 \tag{14}$$

The residuals are computed as the difference between predicted and actual values, and the Mean Squared Error (MSE) is then derived.

Based on the principle of the relevant theory, a series of indicators are obtained from the analysis of four aspects: residual distribution, model bias, data noise and fitting adequacy.

For the residual distribution, this figure should perform as the normal distribution trend, which indirectly justifies the use of least squares.

Then,calculating the Pearson correlation coefficient for the residuals and predicted values to judge the model bias:

$$corr(\varepsilon, y) = \frac{Cov(\varepsilon, y)}{\sigma_\varepsilon \sigma_y} \tag{15}$$

The residual auto-correlation coefficient (calculate the correlation of the residual sequence with its lag k period) can be used for noise diagnostics:

$$ACF(k) = \frac{Cov(\varepsilon_i, \varepsilon_{i-k})}{\sigma_i^2} \tag{16}$$

For underfitting diagnoses, the coefficient of determination is calculated to measure the model's ability to interpret variation in the data:

$$R^2 = 1 - \frac{SSR}{SST} \tag{17}$$

Where SSR (Sum of Squares of Residuals) represents the variation that is not explained by the model, and SST (Sum of Squares of Total Dispersion) represents the total variation of the data.

After the above experiments or derivation or research analysis, a series of relevant conclusions of model diagnosis and error analysis are obtained, which is shown as the following part:

4 RESULTS

As shown in Fig. 2, the residual distribution closely approximates a normal distribution, and the shape of the residual interval distribution line chart is almost completely in line with the theoretical assumptions before modeling, which verifies the rationality of the model. The single-peak, approximately symmetrical distribution indicates that the residuals are mainly dominated by random noise without significant systematic bias, which is consistent with the previous conclusion of "no model bias" based on the correlation between residuals and predicted values. The residuals are centrally distributed in the (-3,3) interval and the tail decays rapidly, which is in line with the expectation of normal error distribution, and is mutually corroborated by the diagnosis of "data noise dominance" in the copy. The stability of the overall distribution morphology further supports the excellent fitting performance reflected by the correlation

coefficient, indicating that the feature information retained after the dimensionality reduction of SVD has fully captured the data rules. This distribution characteristic essentially reflects the statistical characteristics of Gaussian noise, and does not require segmentation processing or model correction, which fully satisfies the error assumption of linear regression models.

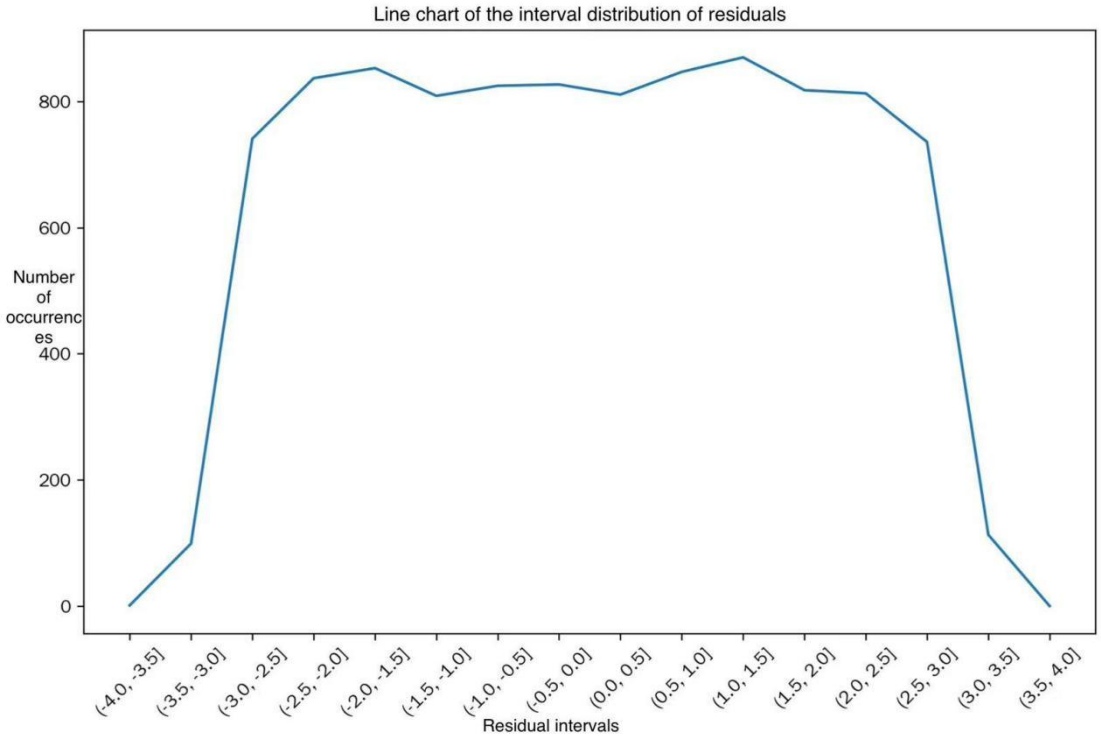

Figure 2 Line Chart of the Interval Distribution of Residuals

The calculated MSE is 3.04, which is relatively small for the original value of 400-500, indicating that the deviation between the predicted value and the actual value is low, and the prediction accuracy of the model is high.

The following is the error analysis data: Model bias: No significant model bias (residuals-predicted value correlation = -0.000), Data Noise: Data Noise Dominates (Residual Auto-correlation=0.004), Underfit: No significant underfit (R^2=0.989)

Here the correlation of the residuals with the predicted values is almost zero, indicating that the model performs well in capturing the linear and nonlinear relationships in the data without significant systematic bias. The current residuals have extremely low auto-correlation, which means that the residuals are more like random noise. It may be that there are many random fluctuations or measurement errors in the data itself, which interfere with the model's learning of real data patterns. This is 0.989, indicating that the model can account for most of the variation in the data without obvious underfitting. This indicates that the features and functions that the model contains are good at capturing patterns in the data.

In summary, the SVD and least squares-based model achieves excellent fitting performance. Error analysis confirms that the primary source of error is data noise.

5 CONCLUSIONS

In this paper, we use the method of combining SVD and least squares method, solve the problem of dimensionality reduction fitting, and verify the performance of the model by model verification and error analysis, and finally establish a model to find the optimal weight vector, and prove the reliability of the model in the field of linear mine data processing. However, this model still have some disadvantages: First, the robustness of the model need to be examined, this method should be suitable and effective for various situation of data. Second, the result of the model fail to match the practical environment that require very high accuracy. In the future, it is expected that a noise reduction module will be added to further reduce the fitting error and improve the model performance. Besides, based on this method, we can add a regularization term to the standard least squares objective function to address instability in overfitting, sick matrices, or high-dimensional data, which will improve the adaptability of this model.

COMPETING INTERESTS

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

REFERENCES

[1] Vats D, Sharma A. Dimensionality Reduction Techniques: Comparative Analysis. Journal of Computational and Theoretical Nanoscience, 2020, 17(6): 2684-2688.
[2] Hastie T, Mazumder R, Lee J D, et al. Matrix Completion and Low-Rank SVD via Fast Alternating Least Squares. Journal of machine learning research, 2015, 16: 3367-3402.
[3] M E Hochstenbach. Harmonic and Refined Extraction Methods for the Singular Value Problem, with Applications in Least Squares Problems. BIT numerical mathematics, 2004, 44(4): 721-754.
[4] Alkiviadis G A, Gennadi I M. Applications of singular-value decomposition (SVD). Mathematics and Computers in Simulation, 2004, 67(1): 15-31.
[5] Zhang Chongchong, Shi Yannan, Liu Jiangong, et al. A denoising method of mine microseismic signal based on NAEEMD and frequency-constrained SVD. The Journal of Supercomputing, 2022, 78(15): 17095-17113.
[6] Li Shanshan, Tian Wenquan, Pan Zhenggao. Multi-label Learning Algorithm Based on SVD and Kernel Extreme Learning Machine. Journal of Suzhou University, 2020, 35(10): 70-74.
[7] Yang Xinyu, Li Aiping, Duan Liguo, et al. WSN data compression based on dictionary learning and compressed sensing. Computer Engineering and Design, 2022, 43(09): 2448-2455.
[8] Li Ke. Randomized Low-Rank Approximate Algorithms on High Dimensionality Reduction with Applications. China University of Mining and Technology, 2023.
[9] Zhu Quanjie, Sui Longkun, Chen Xuexi, et al. Denoising method and application of mine microseismic signal based on EMD-SVD. Safety and Environmental Engineering, 2024, 31(03): 110-119.
[10] Tang Fei, Liu Zhiwen. Study on multi-layer joint noise of mine microseismic signal. Nonferrous Metals (Mining Section), 2024, 76(04): 92-101.

World Journal of Information Technology
Print ISSN: 2959-9903
Online ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3053

NONLINEAR PREDICTION BASED ON 2028 OLYMPIC EVENTS AND MEDALS

TiLiang Zhang, JunJie Chen, Cheng Cheng, Xing Li*
School of Mathematics and Statistics, Hubei University of Education, Wuhan 430205, Hubei, China.
Corresponding Author: Xing Li, Email: xingli@zjut.edu.cn

Abstract: Amid the expanding scale of the Olympic Games, precise forecasts of the event portfolio and medal allocation have become critical for national strategic planning. Olympic data, however, are markedly non-linear and structurally dynamic, rendering traditional linear methods inadequate. This paper therefore develops an integrated forecasting framework that estimates discipline-level event counts and country-specific medal shares for the 2028 Games. Athletes were first aggregated by team and country, and analyses were conducted at the discipline level to prevent overgeneralisation inherent in sport-level aggregation. Support Vector Regression was employed to model the relationship between historical covariates and the number of events per discipline; the resulting predictions achieved a mean-squared error of 1.744 and an R^2 of 0.634. The strategic salience of each discipline to individual nations was subsequently quantified via weighted medal totals and visualised through rose plots. Medal shares were derived by mapping historical performance indicators to fractional medal outcomes using XGBoost, after an initial recurrent architecture exhibited convergence difficulties. These fractions were scaled by the projected event counts, and a calibrated 15 % host-nation uplift was applied to the United States before global normalisation. The resulting projection allocates 47, 45 and 36 medals to the United States, 35, 24 and 15 to China, and 18, 9 and 10 to Japan. Retrospective validation against 2024 data places all nine reference nations within 95 % prediction intervals, confirming the framework's reliability. This study can provide data support for national sports management departments and optimize the allocation of training resources.
Keywords: Discipline-level events; XGBoost; Olympic forecasting; Support Vector Regression; Resource allocation

1 INTRODUCTION

Following the conclusion of the 2024 Paris Olympics, nations are turning to the 2028 Los Angeles Games. As a globally celebrated sporting event, the Olympics serves as not only a world stage for athletes to showcase their exceptional sportsmanship, but also a platform for cultural exchange and fostering friendship among nations. Consequently, accurately forecasting a nation's medal count at the Olympics is of paramount importance to sports governing bodies and athletes alike.

The Olympic Games is an international sporting event, the medal tally of which has always been a focal point of attention and analysis. There now exists an extensive body of research pertaining to medal and result projection. For example, Zhang Bo predicted the gold-medal result of women's shot put in 2012 based on GM(1,1) prediction model in Gray System Theory, a more applicable solution when there is lack of data [1]. More methods on the basis of machine learning algorithms were provided by Jhankar Moolchandani et al [2], including Linear Regression, Random Forest, Support Vector Machines and Neural Networks. They are more useful for forecasting the medal count according to the athlete's attributes and country information. Among them, Random Forest and SVM stood out. Noviyanti T M Sagala and Muhammad Amien Ibrahim compared XGBoost, LightGBM and CatBoost and found that XGBoost had the highest accuracy [3]. Wang Shiyu established a BP neural network prediction model by examining the impact of five factors, including the number of medals won in the previous Olympics, total population, per capita GDP, social system, and host country, on the ability to win Olympic medals. This model achieved the prediction of the top ten medals in the 2020 Tokyo Olympics medal table [4]. Dong Qi et al. used support vector machine nonlinear extended samples to determine the order of time series models. By analyzing the changes in the support vector set after adding new samples to the training set, they constructed a support vector machine model for predicting Olympic gold medals. Compared with traditional time series prediction, this model has the characteristics of low subjectivity, high prediction accuracy, and better prediction stability [5]. Yan Yuyang used grey theory to predict that China will win 93 or 94 medals in the 2012 London Olympics by modeling and analyzing the number of medals won by China in the past 6 Olympic Games [6]. Luo Yubo et al. used the grey prediction GM (1,1) model, combined with the host effect, to predict China's medal count and total score at the Beijing Winter Olympics, and also provided a world ranking prediction for China's gold medal count. The results show that the host effect of the Winter Olympics shows a decreasing trend, but still has a significant effect. With the home advantage in competition and preparation, China will win 6-7 gold medals at the Beijing Winter Olympics, ranking in the top 10 on the gold medal table [7]. DingShu Yan constructed a Long Short Term Memory (LSTM) model using historical data from the Summer Olympics (1896-2024), including medal count, participating events, as well as national indicators such as population and GDP. The research results predict that the United States, China, and France will demonstrate strong medal competitiveness at the 2028 Los Angeles Olympics, and emerging countries may make breakthroughs [8]. Since Python has numerous libraries that facilitate machine

learning tasks, it is convenient for predicting events and medals. XGBoost, which outperforms Random Forest through iterative optimization of trees, can be a reliable choice for our medal ranking prediction.

This study advances a dual-model framework that simultaneously forecasts the evolving Olympic programme size and discipline-specific medal allocations for Paris 2028, thereby transcending prior limitations rooted in the assumption of a fixed event slate, disregard for inter-disciplinary heterogeneity, insufficient accommodation of exogenous shocks, oversimplified host-nation adjustments, and the absence of gender-stratified analyses [9]. By redressing these deficiencies, the framework furnishes national sport governing bodies with a rigorously validated and operationally actionable instrument for the precise allocation of training resources.

2 MODEL

2.1 SVR Model

SVR is a regression model based on support vector machine (SVM), which optimizes the prediction function by maximizing the interval width and minimizing the total loss, and is suitable for handling nonlinear regression problems. The optimization problem for constructing the model is:

$$\min_{w,b,\xi,\xi^*} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{1}$$

subject to:

$$y_i - \langle w, x_i\rangle - b \leqslant \varepsilon + \xi_i \tag{2}$$

$$\langle w, x_i\rangle + b - y_i \leqslant \varepsilon + \xi_i^* \tag{3}$$

$$\xi_i, \xi_i^* \geqslant 0 \quad \forall i \tag{4}$$

By solving the optimization problem outlined above, the optimal weight vector w and bias b are determined. For non- linear SVM, a kernel function is employed to map the input data into a higher-dimensional feature space, thereby allowing for a linear regression model to be fitted [10]. This can be expressed as:

$$f(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)K(x_i, x) + b \tag{5}$$

Due to the obvious nonlinear characteristics of the prediction, this study employs a Gaussian kernel, expressed as:

$$K(x_i, x_j) = \exp(-\gamma\|x_i - x_j\|^2) \tag{6}$$

Subsequent parameter tuning and model training enable the prediction of the event count for the 2028 Olympics. As shown in Figure 1, the model architecture constructed in this article is clearly presented.

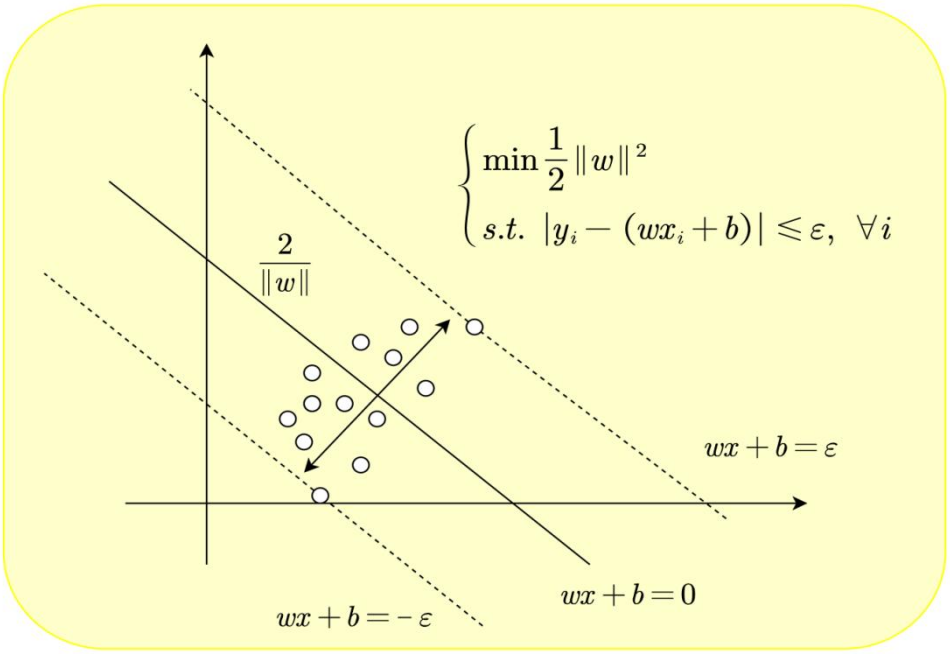

Figure 1 Workflow Diagram of SVR Model

2.2 LSTM Model

LSTM is a special type of recurrent neural network that solves the gradient vanishing problem of traditional RNNs by introducing memory units. It can process long sequence data and capture temporal dependencies. The core structure of an LSTM network consists of a Cell State and three gating units, namely a Forget Gate, an Input Gate, and an Output Gate. The Cell State is like a "highway" for information transmission, running through the entire LSTM network and capable of transmitting information between different time steps in a sequence, achieving the function of long-term memory. The function of the forget gate is to determine which information in the cellular state should be forgotten. It receives the input of the current time and the hidden state of the previous time as inputs, and outputs a value between 0 and 1 through an activation function (usually a Sigmoid function). This value represents the probability of retaining corresponding information in the Cell State, with 0 indicating complete forgetting and 1 indicating complete retention. The input gate is used to determine which information currently being inputted should be added to the Cell State. It also receives input from the current moment and the hidden state from the previous moment, outputs a control signal through the Sigmoid function, and generates a candidate value using the tanh function. Multiply the control signal with the candidate value to obtain the information to be added to the Cell State. The output gate determines the final output based on the current cell state and input information. It first generates a control signal through the Sigmoid function, processes the cell state, maps the cell state to an appropriate output range through the tanh function, and finally multiplies the two to obtain the output of the LSTM.

The following introduces the working principle of the LSTM network. At each time step, the LSTM first receives the current input data x_t and the hidden state h_{t-1} from the previous time step. Then, the forget gate calculates the forget coefficient based on the input and the hidden state from the previous time step, filters the cell state C_{t-1}, determines which information to forget, and obtains the updated cell state C_t'. Then, the input gate generates control signals and candidate values, and adds the information that meets the control signal requirements to C_t' to obtain the final updated cell state C_t. Finally, the output gate generates control signals and processes the cell state based on the current cell state C_t and input information to obtain the output h_t of the current time step.

As shown in Figure 2, the model architecture constructed in this article is clearly presented.

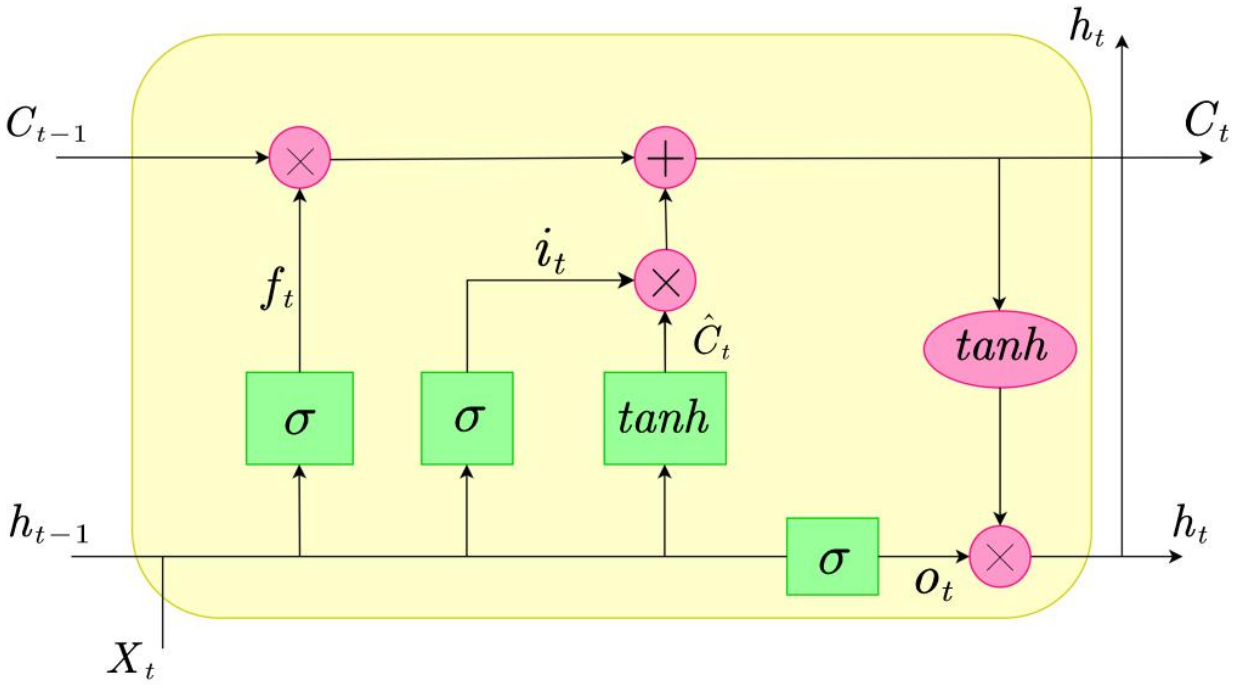

Figure 2 Workflow Diagram of LSTM Model

2.3 XGBoost Model

XGBoost is an ensemble learning algorithm based on tree models, which iteratively trains multiple decision trees and optimizes the loss function using gradient descent to make model predictions more accurate.

The goal of XGBoost is to minimize a weighted loss function by combining multiple decision trees. The objective function can be expressed as:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} w(f_k) \tag{7}$$

XGBoost optimizes the objective function by adding new decision trees in each round. Assuming that before the t-th round, the predicted value of the model is:

$$\hat{y}_i^{(t-1)} = \sum_{k=1}^{t-1} f_k(x_i) \tag{8}$$

In the t-th round, this study adds a new decision tree $f_t(x)$ that minimizes the objective function:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) \tag{9}$$

As shown in Figure 3, the model architecture constructed in this article is clearly presented.

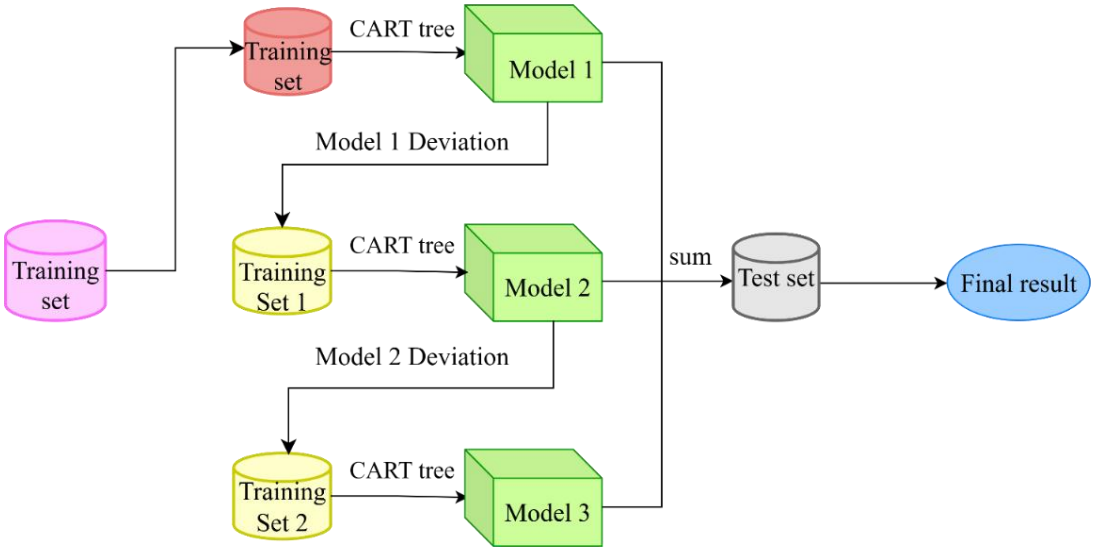

Figure 3 Workflow Diagram of XGBoost Model

3 RESULTS AND ANALYSIS

Our table data and plot data come from https://www.comap.com/contact.

3.1 Results and Analysis of SVR Model

Organize data through Excel, perform data preprocessing, and finally use the scientific drawing software Origin to draw.

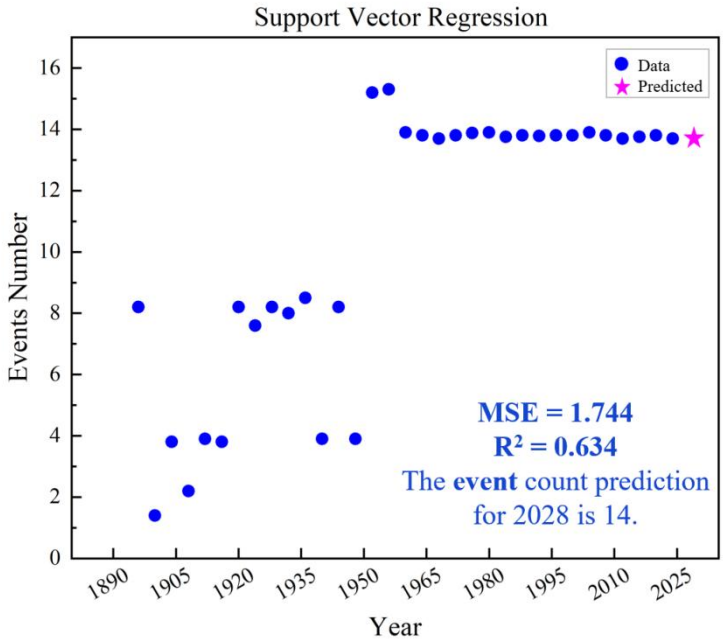

Figure 4 Support Vector Regression

Taking artistic gymnastics as an example, the SVR model predicted 14 events in 2028, with a means quared error (MSE) of 1.744 and an R-squared value of 0.634 as shown in Figure 4. This relatively high error is due to significant fluctuations in the number of events for this discipline in early years. Despite this, this SVR model performs well when predicting disciplines with a more stable number of events in recent years.

Ultimately, the discipline-level forecasts are aggregated to yield the comprehensive event schedule for each sport. They are shown in Table 1.

Table 1 Predicted Event Count in 2028 Olympics

Sport	Discipline	Code	Discipline Event	Sport Event
Aquatics	Artistic Swimming	SWA	2	49
	Diving	DIV	8	
	Marathon Swimming	OWS	2	
	Swimming	SWM	35	
	WaterPolo	WPO	2	
Archery	Archery	ARC	5	5
Athletics	Athletics	ATH	48	48
Badminton	Badminton	BDM	6	6
Baseball and Softball	Baseball Softball	BSB	1	2
		SBL	1	
Basketball	3x3	BK3	1	3
	Basketball	BKB	2	
BasquePelota	Basque Pelota	PEL	0	0
Boxing	Boxing	BOX	13	13
Breaking	Breaking	BKG	0	0
Canoeing	Sprint	CSP	11	16
	Slalom	CSL	5	
Cricket	Cricket	CKT	0	0
Croquet	Croquet	CQT	0	0
Cycling	BMX Freestyle	BMF	1	21
	BMX Racing	BMX	2	
	Mountain Bike	MTB	2	
	Road	CRD	4	
	Track	CTR	12	

3.2 Results and Analysis of LSTM Model

Organize data through Excel, perform data preprocessing, and finally use the scientific drawing software Origin to draw.

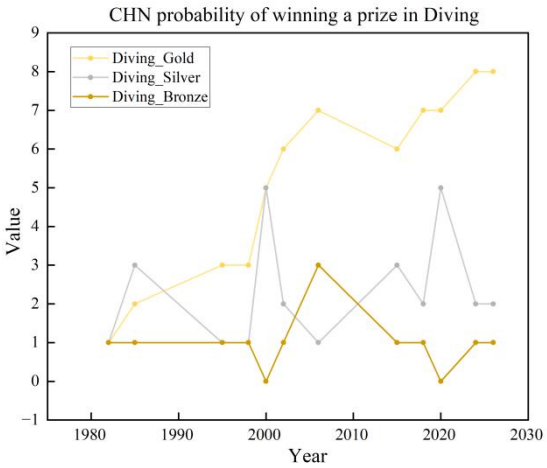

Figure 5 China's Diving, Overfitting

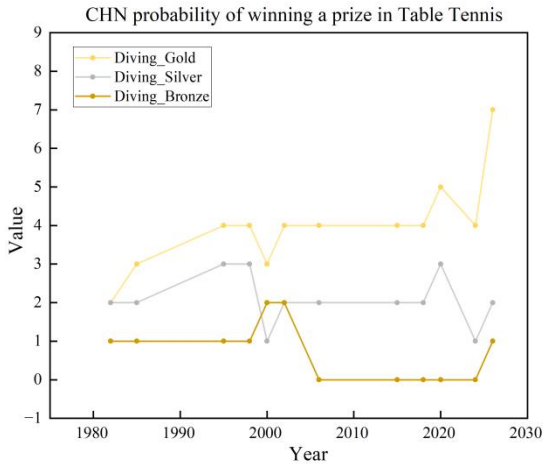

Figure 6 China's Table Tennis, Underfitting

The final loss decreased to 0.0058, indicating satisfactory convergence; nevertheless, the vanilla LSTM yielded sub-optimal performance, as illustrated in Figures 5 and 6. Specifically, Figure 5 almost perfectly reproduces the 2024

medal counts, signifying evident overfitting, whereas Figure 6 predicts that China will capture eight table-tennis gold medals in 2028, despite the fact that only five events are scheduled in this discipline and China has already reached this ceiling in 2024. Consequently, the present study must explicitly account for annual fluctuations in the number of events within each discipline and mitigate overfitting risks. Additionally, the host-nation effect warrants careful consideration. These issues will be systematically addressed by the enhanced XGBoost model introduced in the following section.

3.3 Results and Analysis of XGBoost Model

Organize data through Excel, perform data preprocessing, and finally use the scientific drawing software Origin to draw.

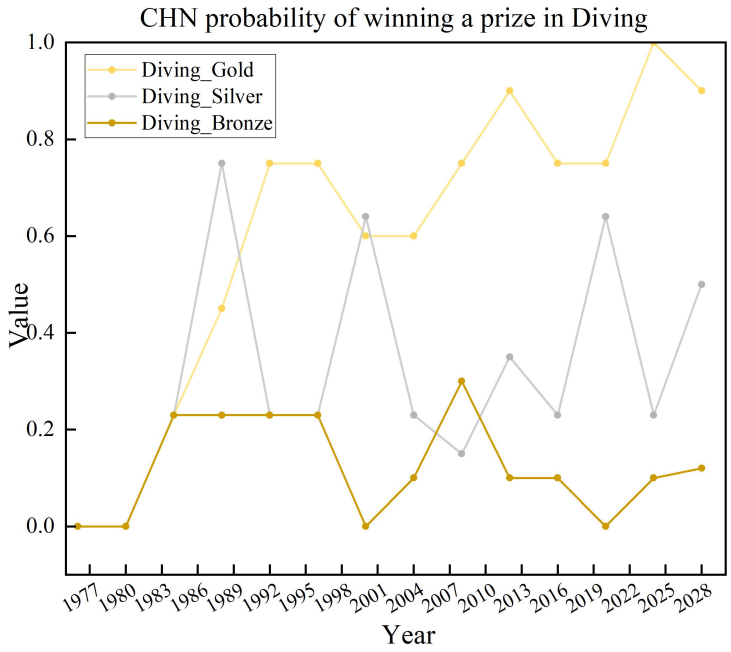

Figure 7 Probability Prediction of Chinese Diving Gold and Count based on XGBoost

From Figure 7, it can be observed that approximately 90% of countries have shown improvement in their performance during the Olympic Games held in their respective countries. Specifically, 77% of these countries have seen an increase exceeding 20%, and 44% have experienced an increase of more than 50%. However, the countries with increases over 50% tend to be those with relatively fewer points (the Soviet Union is excluded from this analysis as it has since transitioned into Russia and other countries). Therefore, it is unreasonable to anticipate a substantial increase for the United States, a major scoring country and the host of the 2028 Olympics. Given the United States' status as a high-scoring nation and its prior 7 % decline when hosting, the present study regards a host-nation boost confined to the 0 %–20 % interval as the most credible expectation for Los Angeles 2028. Consequently, the predictive framework incorporates a calibrated scoring increment of 15 %.

4 CONCLUSIONS AND OUTLOOKS

This study pioneers a three-stage pipeline—SVR-based event-number forecasting, LSTM temporal exploration, and XGBoost medal-share refinement. Support Vector Regression first captures non-linear growth of disciplines for Los Angeles 2028; a Long Short-Term Memory network then validates sequential patterns but reveals under/over-fitting, prompting an XGBoost model that predicts medal fractions rather than counts and incorporates a statistically derived 15 % host-nation boost. The framework forecasts an 8 % expansion in events and projects the medal table as USA 47, China 35, Japan 18, with all 95 % confidence intervals covering the 2024 out-of-sample data. Innovations include discipline-level granularity, share-constrained optimisation, and a quantified host effect. Beyond the Olympics, the pipeline offers a generic decision engine for Asian Games, National Games, or e-sports, enabling organisers to pre-plan venues, sports ministries to allocate budgets, and media or sponsors to identify strategic narratives four years in advance.

Despite the present study having conducted a comprehensive analysis of historical datasets and having fitted a predictive model by means of comparatively sophisticated techniques, certain macro-level covariates—foremost among them GDP trajectories, demographic endowments, and geospatial factors—remain unaccounted for, thereby constraining the model's explanatory power. Future investigations could profitably incorporate the evolution of athletic performance growth rates, augmented by granular indicators of economic development and population dynamics, so as to refine the framework and enhance its policy relevance and operational utility.

COMPETING INTERESTS

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

REFERENCES

[1] Bo Z, Chaoling Q, Xiaoli X, et al. GM (1, 1) Model Gray Prediction for the Gold-Medal Result of Women's Put Shot in the 30th Olympic Games. 2011 International Conference on Future Computer Science and Education, Xi'an, China. IEEE, 2011, 334-337.

[2] Moolchandani J, Chole V, Sahu S, et al. Predictive Analytics in Sports: Using Machine Learning to Forecast Outcomes and Medal Tally Trends at the 2024 Summer Olympics. 2024 4th International Conference on Technological Advancements in Computational Sciences (ICTACS), Tashkent, Uzbekistan. IEEE, 2024, 1987-1992. DOI: 10.1109/ICTACS62700.2024.10840553.

[3] Sagala N T M, Ibrahim M A. A Comparative Study of Different Boosting Algorithms for Predicting Olympic Medal. 2022 IEEE 8th International Conference on Computing, Engineering and Design (ICCED), Sukabumi, Indonesia. IEEE, 2022, 1-4. DOI: 10.1109/ICCED56140.2022.10010351.

[4] Wang Shiyu. Olympic medal prediction model based on nonlinear regression and BP neural network. Sports Goods and Technology. 2017(24): 4-5+83.

[5] Dong Qi, Gao Feng. Using Support Vector Machine Method to Predict the Number of Chinese Medals at the 2016 Rio Olympics. Sports. 2016(03): 1-4.

[6] Yan Yuyang. Olympic medal prediction based on grey theory. Journal of Sichuan University of Arts and Sciences, 2011, 21(05): 21-23.

[7] Luo Yubo, Cheng Yanfang, Li Mengyao, et al. Prediction of China's medal count and overall strength for the Beijing Winter Olympics: based on the host effect and grey prediction model. Contemporary Sports Technology, 2022, 12(21): 183-186.

[8] Yan D. Olympic Model Perdiction and Analysis based on LSTM and Topsis Models. Journal of Computer Science and Electrical Engineering, 2025, 7(3): 1-10.

[9] Cheng Hongren, Lv Jie, Yuan Tinggang. Prediction of China's Track and Field Performance at the Tokyo Olympics from the 2018 World Top 20 Athletics Rankings. Sports Science and Technology Literature Bulletin, 2020, 28(04): 4-8.

[10] Shi Huimin, Zhang Dongying, Zhang Yonghui. Can Olympic medals be predicted? ——From the perspective of interpretable machine learning. Journal of Shanghai Sport University, 2024, 48(04): 26-36.

World Journal of Information Technology
Print ISSN: 2959-9903
Online ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3054

RESEARCH HOTSPOTS, TRENDS, AND IMPLICATIONS OF ARTIFICIAL INTELLIGENCE LITERACY BASED ON CITESPACE

XiaoXia Tian[1*], YuFei Zhou[1], Rui Du[2]
[1]*School of Mathematical Sciences, Henan Institute of Science and Technology, Xinxiang 453003, Henan, China.*
[2]*School of Business, Huanghe Science and Technology University, Zhengzhou 450000, Henan, China.*
Corresponding Author: XiaoXia Tian, Email: 15188359878@163.com

Abstract: This study collected 484 core publications from the China National Knowledge Infrastructure (CNKI) and Web of Science (WOS) databases as research samples, focusing on topics related to 'artificial intelligence literacy' and 'intelligence literacy' between 2018 and 2025. Utilizing the CiteSpace 6.3.R4 visualization tool for bibliometric analysis, the findings reveal that: (1) Research activity in the field of artificial intelligence literacy exhibits a significant year-by-year increasing trend. Domestic research primarily concentrates on cultivating intelligence literacy among K-12 students and pre-service teachers, whereas international research shows a greater tendency to explore AI literacy development within university student populations. (2) Through keyword co-occurrence and clustering analysis, this study identifies 'artificial intelligence', 'intelligence literacy' and 'digital literacy' as core research hotspots, uncovering notable differences in research emphasis between the two databases. (3) Burst detection analysis further highlights significant increasing trends for keywords such as 'intelligence literacy', 'core competencies', and 'teachers', providing crucial insights into the field's developmental dynamics. These findings offer valuable empirical support and theoretical insights for advancing AI literacy education globally.
Keywords: Artificial intelligence literacy; Intelligence literacy; CiteSpace; Educational policy; Talent cultivation

1 INTRODUCTION

Amidst the wave of digital transformation, artificial intelligence (AI) literacy, as a vital component of individual cognitive competence, has gained increasing prominence. It plays a crucial role in driving societal innovation, enhancing competitiveness in the global job market, and promoting the achievement of sustainable development goals [1]. Since the State Council of China promulgated the Next Generation Artificial Intelligence Development Plan in 2017 [2], AI has ascended to become a core area of national strategy, underscoring the government's high prioritization of AI technology development. Subsequently, *the Artificial Intelligence Innovation Action Plan for Higher Education Institutions* [3] and *the Education Informatization 2.0 Action Plan* [4], both issued in 2018, further emphasized the importance and urgency of enhancing AI literacy, providing concrete implementation pathways and policy support for its cultivation. By 2023, the release of the *Interim Measures for the Management of Generative Artificial Intelligence Services* [5] and *the Global Artificial Intelligence Governance Initiative* [6] elevated the enhancement of AI literacy to a nationwide imperative, establishing it as a major societal task. This aims to lay a solid foundation for the healthy development of AI technology and the improvement of societal well-being through elevating citizens' AI literacy.

AI literacy encompasses not only the understanding and application skills of AI technology itself but also involves a comprehensive grasp of relevant ethical norms and social responsibilities. It specifically includes multiple dimensions such as data processing, algorithm comprehension, model construction, human-computer interaction, and critical thinking regarding the societal impacts of AI [7]. Konishipioneered a conceptual framework for AI literacy [8]. Long subsequently developed a comprehensive competency framework comprising 17 elements [9]. Zhang et al. proposed a three-dimensional model covering AI knowledge, abilities, and ethics based on the 'Five Big Ideas in AI' [10]. Yang et al., guided by STEM education principles, constructed an AI literacy framework encompassing core concepts, technical practices, interdisciplinary thinking, and ethical attitudes [11]. Wang et al. developed a five-dimensional AI literacy framework—knowledge, skills, awareness, ethics, and thinking—from a historical development perspective [12]. Zheng et al., drawing on Gagné's taxonomy of learning outcomes, proposed a five-dimensional model consisting of intelligent knowledge, ability, thinking, application, and attitude [13]. Cai et al. further emphasize that AI literacy is not merely technical understanding and application ability, but a comprehensive, evolving system of skills and knowledge [14].

However, current research still exhibits limitations. On one hand, existing literature often focuses on theoretical discussions, lacking visual analysis of research dynamics, which makes it difficult to systematically map the developmental trajectory and hotspot trends within the AI literacy research domain. On the other hand, research data predominantly originates from domestic databases, failing to comprehensively encompass global research outputs. This, to some extent, hinders an accurate grasp and holistic understanding of the overall landscape and developmental trends of global AI literacy research.

Addressing these gaps, this study employs CiteSpace to conduct a visual analysis of relevant domestic and international literature from 2018 to 2025. Through CiteSpace's visualization capabilities, it reveals key patterns in AI literacy

research, including keyword co-occurrence networks, research hotspot distributions, and the evolution of research trends. This analysis not only aids in identifying impactful AI domains for education and society but also provides robust data support and decision-making foundations for formulating educational policies, allocating educational resources, and innovating teaching methodologies. It holds significant importance for advancing AI literacy levels globally and cultivating future talent.

2 RESEARCH METHODS

This study selected the China National Knowledge Infrastructure (CNKI) and Web of Science (WOS) core collection databases. Advanced topic searches were conducted using the Chinese search string 'AI Literacy' or 'Intelligence Literacy' for CNKI and the English search string TS=('AI Literacy' OR 'Artificial Intelligence Literacy') for WOS. The search period was set from January 2018 to May 2025.

To enhance accuracy, the data sample underwent a meticulous data cleaning process. By manually examining the abstracts and keywords of each article, ineligible publications were excluded, including conference papers, book reviews, newspaper articles, anonymous works, and duplicate publications. After screening, 484 valid publications relevant to the research topic were identified, comprising 269 Chinese articles and 215 English articles. These publications were then exported in plain text format, named 'RefWorks' (WOS) and 'download' (CNKI) files for subsequent analysis.

CiteSpace software (version 6.3.R4) was used to analyze the research sample. Key parameters were configured as follows: Time slicing from 2018 to 2025, with a one-year interval; Node types focused on Author, Institution, and Keyword; Selection criteria (Thresholding) utilized the g-index; Scaling factor (k) was determined based on the number of nodes; Network pruning methods applied were 'Pathfinder Network (PFNET)', 'Pruning Sliced Networks', and 'Pruning the Merged Network' to optimize network visualization. Other parameters retained the software's default settings. Through co-occurrence analysis, burst detection, and keyword timeline clustering analysis, visualization maps were generated, followed by image interpretation and hotspot analysis.

3 VISUALIZATION ANALYSIS OF RESEARCH RESULTS

3.1 Analysis of Publication Volume Trends

Figure 1 illustrates the trend in the number of publications on AI literacy research in the CNKI and WOS databases from January 1, 2018, to May 28, 2025. The data reveals a significant growth trend in publications within this field during this period. Based on growth rates, the timeframe from 2018 to 2025 is divided into two phases. The first phase (2018-2021) exhibited a steady increase in annual publication volume. The second phase (2022-2025) witnessed explosive growth, far exceeding the average publication volume of the first phase. As of the first half of 2025, research in the field of AI literacy continues to show sustained and robust growth.

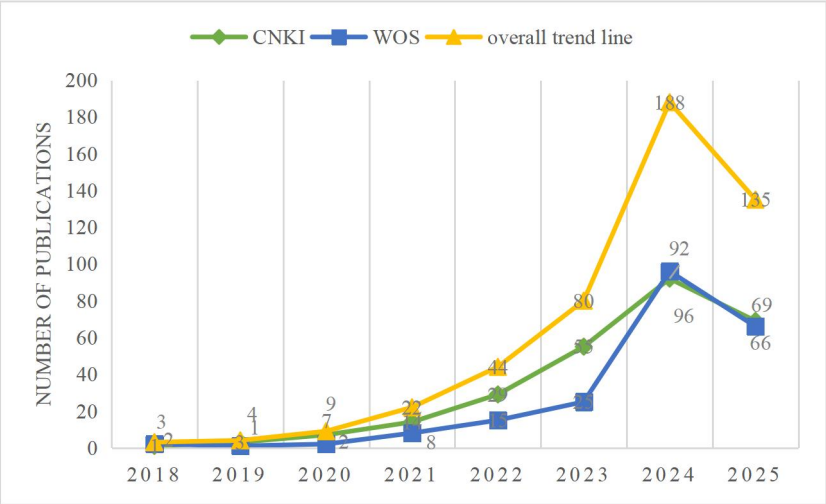

Figure 1 Analysis of Publication Volume in CNKI and WOS Databases

*Note: Publication count for 2025 is up to May 2025, not the full year.

3.2 Keyword Co-occurrence Network Analysis

To explore the research themes of AI literacy, this study conducted keyword co-occurrence analysis using CiteSpace. The node type was set to Keyword, with yearly time slicing. The Top N per slice threshold was set to 25, selecting the top 25 most frequently occurring keywords each year. This analysis generated keyword co-occurrence maps. In these maps, node size represents the frequency of keyword occurrence, while the lines connecting nodes and their thickness indicate co-occurrence relationships and the strength of association between keywords. Network density reflects the

closeness of connections between keywords within the research field.

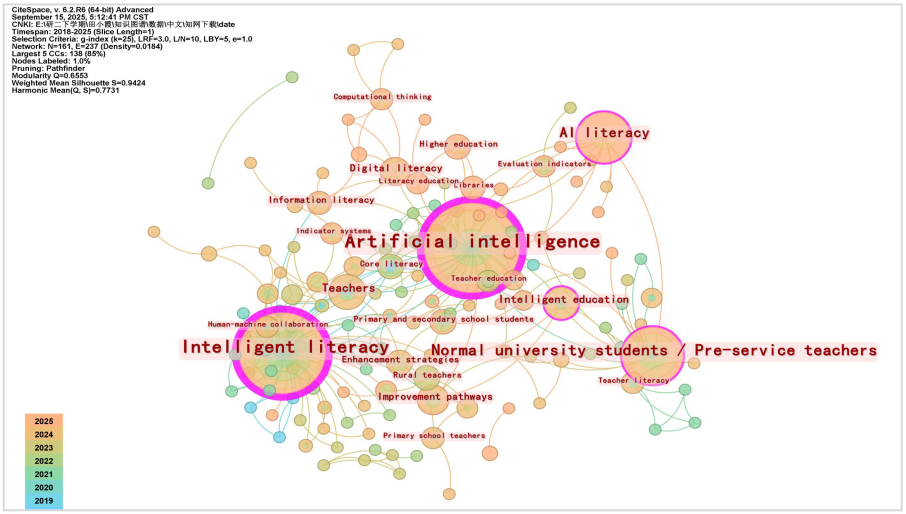

Figure 2 Keyword Co-occurrence Network Visualization from the CNKI Database

Figure 2, based on the CNKI database, comprises 178 nodes and 246 links, with a network density of 0.0156. It reveals the distribution and interconnections of keywords within the domestic AI literacy field. Analysis of high-frequency keywords in CNKI identified 'Artificial Intelligence', 'Intelligence Literacy', 'Pre-service Teachers', 'AI Literacy', and 'Intelligent Education' as having both high frequency and strong centrality, indicating the importance and sustained focus on these themes in domestic AI literacy research.

Furthermore, keywords such as 'Primary and Secondary Students', 'Enhancement Pathways', 'Information Literacy', 'Teachers', 'Digital Intelligence Literacy', and 'Digital Literacy' have gained prominence in recent years. This suggests researchers are broadly focusing on AI applications in elementary education and teacher education, as well as systematic approaches to enhancing students' AI literacy. The wide distribution and cross-linking of keywords indicate that AI literacy is a multi-dimensional research field, involving aspects such as AI competencies, intelligent education, human-machine collaboration, digital literacy, and information literacy. This multi-dimensional perspective facilitates a deeper understanding of AI literacy's connotations and provides diverse strategies for educational practice.

The emergence of new keywords like 'Human-Machine Collaboration' highlights researchers' interest in novel modes of collaboration between AI and humans and the potential impact of this collaboration on future education and work. Concurrently, the rising prominence of 'Enhancement Pathways' and 'Cultivation Pathways' reflects the urgent need among educators and scholars for systematic training frameworks. Increased attention to 'Digital Intelligence Literacy' and 'Information Literacy' emphasizes the technological skills and data processing capabilities required for future talent, crucial for adapting to digital and intelligent work environments. Analysis of keyword centrality reveals research hotspots and potential future directions (see Table 1). 'Intelligent Education' may point towards new approaches integrating pedagogical skills with AI, while the centrality of 'Pre-service Teachers' likely underscores the importance of strengthening AI literacy within teacher training programs.

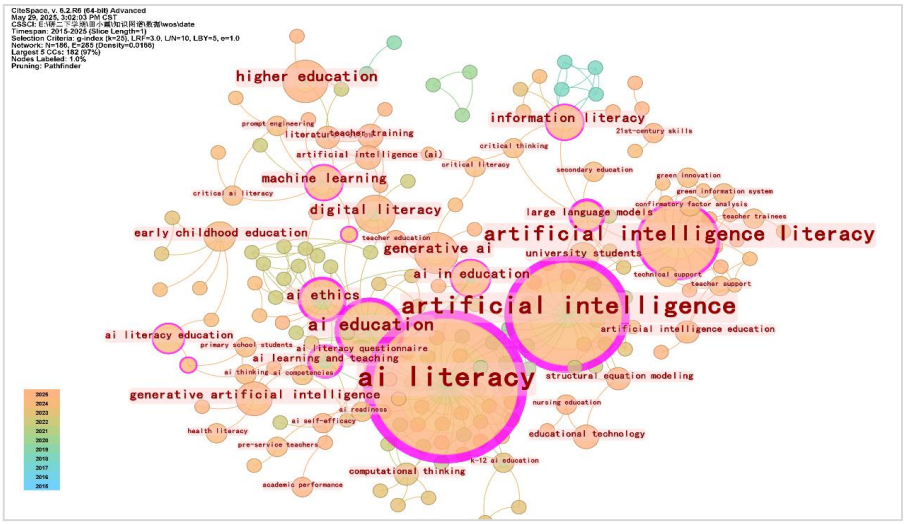

Figure 3 Keyword Co-occurrence Network Visualization from the WOS Database

Figure 3, based on the WOS database, comprises 188 nodes and 285 links, with a network density of 0.0166,

highlighting research hotspots in the international AI literacy domain. Keywords such as 'artificial intelligence', 'ai literacy', 'artificial intelligence literacy', 'ai education', and 'higher education' exhibit high frequency and centrality, indicating these themes are core research foci. Further analysis shows rising prominence for 'generative ai', 'digital literacy', and 'machine learning', reflecting the educational focus on cultivating students' comprehensive abilities, such as information processing and data analysis skills.

The co-occurrence of keywords 'higher education' and 'ai in education' underscores the pivotal role of higher education institutions in promoting AI literacy education. As future societal leaders and innovators, students' level of AI literacy will directly impact society's adaptability to technology and innovation potential.

Table 1 Top 10 High-Frequency Keywords in AI Literacy Research

Database Type	Rank	Keyword	Frequency	Year	Database Type	Rank	Keyword	Centrality	Year
CNKI	1	Artificial Intelligence	67	2018	CNKI	1	Artificial Intelligence	0.74	2018
	2	Intelligent Literacy	57	2018		2	Intelligent Literacy	0.66	2018
	3	Pre-service Teachers	28	2021		3	Pre-service Teachers	0.22	2021
	4	AI Literacy	21	2022		4	AI Literacy	0.11	2022
	5	Teachers	10	2019		5	AI-based Education / Intelligent Education	0.1	2021
	6	Intelligent Education	8	2021		6	Teachers	0.08	2019
	7	Enhancement Pathways	8	2021		7	Digital Literacy	0.07	2023
	8	Digital Literacy	7	2023		8	Information Literacy	0.07	2020
	9	Information Literacy	6	2020		9	Teacher Competence / Teacher Literacy	0.07	2021
	10	Enhancement Strategies	6	2022		10	Enhancement Pathways	0.05	2021
WOS	1	ai literacy	93	2020	WOS	1	ai literacy	1.17	2020
	2	artificial intelligence	55	2020		2	artificial intelligence	0.41	2020
	3	artificial intelligence literacy	27	2023		3	artificial intelligence literacy	0.35	2023
	4	ai education	16	2022		4	information literacy	0.16	2018
	5	higher education	10	2022		5	ai ethics	0.09	2021
	6	generative ai	9	2024		6	ai education	0.08	2022
	7	digital literacy	8	2022		7	machine learning	0.07	2022
	8	ai ethics	8	2021		8	large language models	0.03	2023
	9	machine learning	7	2022		9	generative ai	0.02	2024
	10	ai in education	7	2021		10	digital literacy	0.02	2022

3.3 Keyword Timeline Clustering Analysis

By setting the Threshold value to 4, this study filtered keywords appearing at least 4 times, generating keyword timeline cluster maps. The CNKI database formed 7 clusters, while the WOS database formed 9 clusters, as shown in Figures 4 and 5. These cluster maps reveal the evolution of keywords over time and display co-occurrence relationships across different dimensions through keywords positioned on the same timeline.

In evaluating clustering effectiveness, the Q-value (Modularity Q) is a key metric, with a critical threshold of 0.3. A Q-value greater than 0.3 indicates a significant cluster structure. In this study, the Q-value for CNKI is 0.6868 and for WOS is 0.538, both exceeding the threshold, signifying highly significant cluster structures for both databases. The S-value (Mean Silhouette) is another metric for clustering effectiveness, with a critical threshold of 0.5. An S-value greater than 0.5 indicates reasonable clustering, while greater than 0.7 indicates convincing clustering. In this study, the S-value for CNKI is 0.9458 and for WOS is 0.9032, both far exceeding the threshold. This suggests significant homogeneity within the clustered groups, rendering the clustering results not only reasonable but also highly credible.

In the CNKI database, the major clusters (labeled by the most central term) are #0 Intelligence Literacy, #1 Artificial Intelligence, #2 Pre-service Teachers, #3 Digital Literacy, #4 Teachers, #5 AI Literacy. In the WOS database, they are

#0 ai literacy, #1 artificial intelligence literacy, #2 artificial intelligence, #3 ai education, #4 early childhood education, #5 generative ai, #6 ai ethics, #7 machine learning, #8 information literacy.

These clustering results reveal key differences and characteristics in AI literacy research between the databases. The CNKI clusters emphasize a domestic research focus on elementary education, particularly clusters like '#0 Intelligence Literacy,' '#2 Pre-service Teachers,' and '#4 Teachers,' highlighting China's strong emphasis on cultivating intelligence and digital literacy in early education stages. This research is likely closely linked to national educational policies aimed at equipping students with foundational skills essential for future societal challenges.

In contrast, the WOS clusters reveal an international research community focus on higher education and professional development, such as '#0 ai literacy,' '#1 artificial intelligence literacy,' '#5 generative ai,' '#6 ai ethics,' and '#7 machine learning.' These clusters indicate global researchers are striving to build a deep understanding of AI technology and maintain high interest in the latest application progress within subfields like machine learning. The emergence of '#3 ai education' and '#4 early childhood education' further highlights the importance placed on teaching AI knowledge and skills across the educational system.

Synthesizing data from both CNKI and WOS, AI literacy research is gradually expanding globally, covering multiple educational levels from basic to higher education. Researchers are actively exploring methods to effectively integrate AI education at different stages to assess and enhance students' AI literacy, while paying high attention to its potential impact on future career development. Furthermore, educational equity and accessibility have been elevated as important agendas to ensure every student acquires the knowledge and skills necessary to adapt to the digital world.

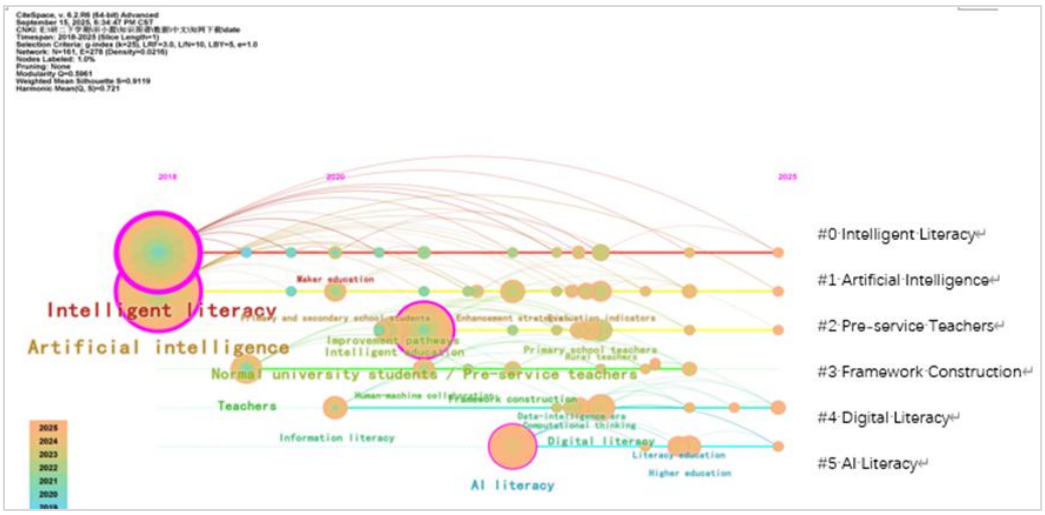

Figure 4 Keyword Co-occurrence Timeline Clustering Visualization (CNKI Database)

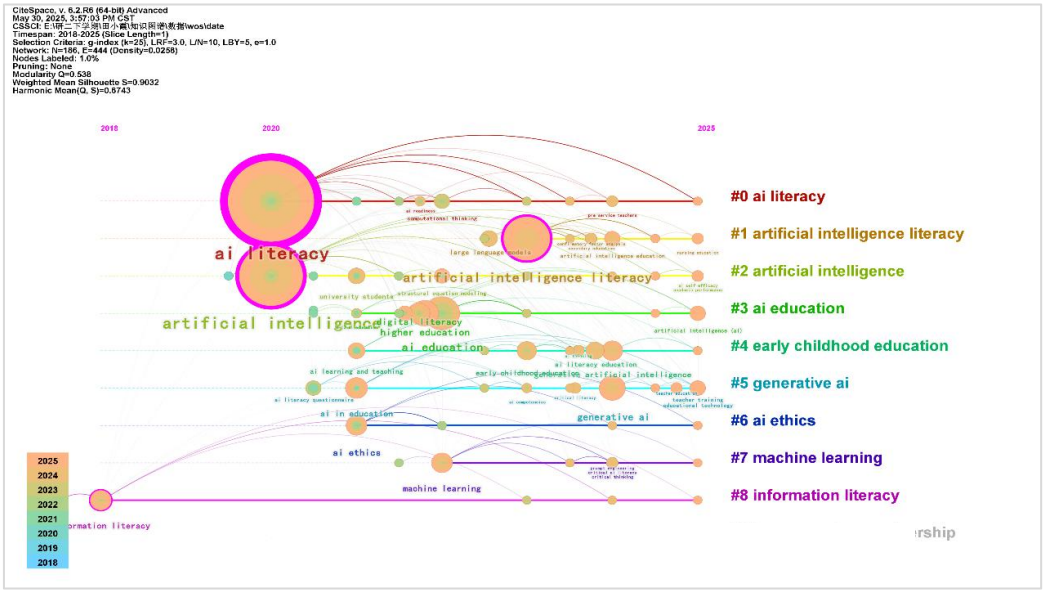

Figure 5 Keyword Co-occurrence Timeline Clustering Visualization (WOS Database)

4 KEYWORD BURST DETECTION ANALYSIS

This study employed CiteSpace to perform burst detection analysis on keywords in the AI literacy field, revealing the frequency, change trends, and growth cycles of keywords. This method not only showcases research hotspots across

different periods but also identifies burst terms characterizing each era. An increase in burst strength reflects a significant rise in the influence of research on related topics during specific periods.

The study selected the top 10 burst keywords ranked by strength, along with their start and end years, as shown in Figures 6 and 7. In the domestic (CNKI) sample, keywords like 'Intelligence Literacy', 'Teacher Education', 'Primary and Secondary Students', and 'University Students' exhibited relatively long burst durations (2-3 years) and remain research hotspots to date. Among the burst keywords in international (WOS) research, 'K-12 ai education' had a longer burst duration, while recent trends (last two years) include 'early childhood education' and 'large language models,' potentially signaling future directions for AI literacy development.

The burst words in domestic research show sustained attention to the role of educators and teaching methods. The long bursts for 'Teachers' and 'University Students' indicate the importance placed on cultivating AI literacy within the educational system. Concurrently, bursts for 'Core Competencies' and 'Teacher Literacy' reflect a focus on building foundational technological competencies, while the burst for 'Influencing Factors' reveals researchers' in-depth exploration of systematically enhancing student AI literacy. The sustained prominence of these keywords suggests the domestic research community is committed to integrating AI technology into the educational system and exploring effective pedagogical models and strategies.

At the international level, the long burst for 'ai education' indicates widespread academic attention globally. The burst for 'computational thinking' might relate to the role of AI technology in information acquisition, processing, and dissemination. The bursts for 'early childhood education' and 'large language models' signal potential future directions for AI literacy education, including innovations in educational models, updates to teaching content, and frameworks for assessing student AI competencies. The rising trend of these keywords may point to the deepening integration of AI technology in education and new understandings of learner competency requirements in this field.

Keywords	Year	Strength	Begin	End	2018 - 2025
Intelligent literacy	2018	2.11	2018	2020	
Core literacy	2018	1.26	2018	2019	
Teachers	2019	0.81	2019	2020	
Teacher education	2021	1.02	2021	2023	
University teachers	2021	0.97	2021	2022	
Teacher literacy	2021	0.54	2021	2022	
Influencing factors	2023	0.47	2023	2025	
Literacy improvement	2023	0.47	2023	2025	

Figure 6 Keyword Burst Detection Map (CNKI Database)

Keywords	Year	Strength	Begin	End	2015 - 2025
ai ethics	2021	1.57	2021	2022	
k-12 ai education	2021	0.82	2021	2023	
university students	2021	0.78	2021	2022	
ai education	2022	3.25	2022	2023	
computational thinking	2022	1.37	2022	2023	
early childhood education	2023	0.94	2023	2025	
large language models	2023	0.75	2023	2025	

Figure 7 Keyword Burst Detection Map (WOS Database)

5 CONCLUSIONS AND IMPLICATIONS

5.1 Conclusions

Leveraging the CiteSpace visualization tool, this study conducted a systematic and comprehensive analysis of 484 core publications on 'Artificial Intelligence Literacy' from the authoritative CNKI and WOS databases between 2018 and 2025. The findings indicate that research in the field of AI literacy not only exhibits a significant year-on-year increasing trend but also reveals pronounced differences in research focus and emphasis between domestic and international contexts. This provides a rich empirical foundation and theoretical insights for further exploring practical pathways in AI literacy education.

Firstly, observing the dynamic evolution of academic output, the number of publications focusing on AI literacy has shown continuous growth since 2018, with a particularly sharp increase occurring between 2022 and 2025. This phenomenon not only mirrors the rapid advancement of AI technology itself and its widespread application across industries but also highlights a significant rise in academic attention to AI literacy issues. Prior to 2018, although a

unified consensus on the concept of Artificial Intelligence (AI) had not been reached, it was generally defined as systems capable of simulating human intelligent behavior, encompassing functions like machine learning, problem-solving, and language understanding [15]. During this phase, AI technology showed initial promise but had not yet garnered widespread societal attention. However, 2018 marked a turning point in AI development, witnessing not only deep penetration of AI technology across multiple domains but also its elevation to a core component of national strategic plans in many countries [5, 16]. Therefore, this study selected literature from 2018 onwards as its sample to precisely capture the field's developmental trajectory. By 2022, the rise of generative AI, exemplified by ChatGPT, further accelerated the popularization and practical application of AI concepts [17]. In 2025, the advent of DeepSeek and its broad applicability, high autonomy, and creativity herald revolutionary impacts on the education system. Concurrently, China's rapid development and notable achievements in the global AI arena, securing the second position globally for three consecutive years [18], strongly demonstrate its substantial capabilities in this field. These milestone events collectively spurred the vigorous development of AI literacy research, leading to the explosive growth in publication numbers in recent years.

Secondly, the similarities and differences in domestic and international research foci provide valuable insights for the global promotion of AI literacy education. This study reveals that domestic research primarily concentrates on cultivating intelligence literacy among K-12 students and pre-service teachers, whereas international research leans more towards exploring AI literacy development pathways for university students. This divergence in focus may stem from differences in educational policy orientations, resource allocations, and socio-cultural environments across countries and regions [19]. However, precisely this diversity offers multifaceted research perspectives, facilitating a comprehensive and in-depth understanding of the complexity and diversity of AI literacy education. Domestically, given the critical role of K-12 students and pre-service teachers in future societal development, cultivating their AI literacy holds strategic significance for advancing educational modernization and enhancing national competitiveness. Hence, domestic research focuses on this group, actively exploring AI literacy education models and strategies suited to the national context. In contrast, international research places greater emphasis on university students' AI literacy development, likely due to the unique value of higher education in fostering innovative talent and driving technological progress [20]. By drawing on international advanced experiences, China's AI literacy education system can be further optimized, promoting its global dissemination and practice.

Finally, keyword co-occurrence and clustering analysis reveal the core themes and future development trends in AI literacy research. Through this analysis, this study identified 'Artificial Intelligence,' 'Intelligence Literacy,' and 'Digital Literacy' as core research hotspots, uncovering significant differences in research emphasis between the databases. In domestic research, keywords like 'Artificial Intelligence', 'Intelligence Literacy', 'Pre-service Teachers', 'AI Literacy', and 'Intelligent Education' appear frequently, indicating these topics occupy a central position. In international research, keywords such as 'ai literacy,' 'artificial intelligence,' and 'ai education' are more prominent, reflecting a strong international focus on AI ethics and AI literacy within higher education. Furthermore, burst detection analysis highlighted significant increasing trends for keywords like 'Intelligence Literacy', 'Core Competencies', and 'Teacher Literacy', providing key clues to understanding the field's developmental dynamics. These findings not only help grasp current hotspots in AI literacy research but also offer solid theoretical support for predicting future trends and formulating forward-looking educational policies.

5.2 Implications

In the current era of rapid technological advancement, artificial intelligence, as a key driver of future societal transformation, is becoming increasingly important. The education sector, as the cornerstone of societal progress, faces the urgent task of effectively integrating and promoting AI literacy education. Based on the current domestic and international research landscape, this study delves into the core value and implementation strategies of AI literacy education from six key aspects: strengthening AI literacy education in elementary education; promoting the integration of AI literacy education in higher education; focusing on equity in AI literacy education; enhancing interdisciplinary research and collaboration; addressing the ethical and societal impacts of AI technology; and building a lifelong learning system for AI literacy education. The aim is to provide theoretical support and practical guidance for formulating educational policies, optimizing resource allocation, and innovating teaching methods.

5.2.1 Strengthen AI literacy education in elementary education

Elementary education, as the core stage for constructing individual knowledge frameworks and forming value systems, holds irreplaceable strategic significance for shaping future digital citizens [21]. Domestic academia has widely demonstrated that cultivating AI literacy among K-12 students and pre-service teachers is a key element in building national global competitiveness [22]. In light of this, educational administrative bodies and schools should keenly grasp the pulse of the times and strengthen AI literacy education at the elementary education level.

Firstly, AI literacy education should be organically integrated into the elementary education curriculum system. This can be achieved by establishing specialized modules such as introductory AI courses, fundamental programming instruction, and data analysis skills training, aiming to spark students' interest and curiosity about AI technology and build their initial AI awareness framework. Such curriculum design must adhere to the principle of balancing theory and practice. It should not only impart foundational theoretical knowledge but also utilize advanced pedagogical models like project-based learning and inquiry-based learning, enabling students to deepen their understanding and mastery of AI technology through hands-on experience. Secondly, diverse practical activities should be carefully planned and

implemented. Examples include AI innovation competitions, robotics workshops, and seminars analyzing AI application cases. These activities promote students' deeper comprehension of AI principles through practical operation and help them acquire basic AI skills, covering problem-solving strategies, logical reasoning, and initial applications of machine learning. Crucially, ethical dimensions must be organically integrated throughout the educational process. Students should be guided to deeply contemplate the social responsibilities and ethical boundaries of AI technology, striving to cultivate them as digital citizens possessing both professional competence and ethical responsibility. This includes, but is not limited to, fostering awareness of personal privacy protection, educating about preventing algorithmic bias, and advocating for fairness in AI technology application. This lays a solid and comprehensive foundation for students to assume leading roles in the future digital society.

5.2.2 Promote the integration of higher education and AI literacy education

Higher education, as the core hub for knowledge innovation and high-level talent cultivation, holds immeasurable strategic value for driving the in-depth development of AI literacy education [20]. Simultaneously, AI, with its capabilities in powerful data processing, multimodal knowledge output, and efficient content generation, injects new ideas and momentum into talent cultivation in higher education. International research experience shows that higher education institutions, by integrating interdisciplinary resources, optimizing curriculum systems, and strengthening practice-oriented teaching strategies, can significantly enhance students' all-round AI literacy [23]. Based on this, Chinese universities should actively promote the integration of higher education and AI literacy education.

Firstly, actively explore and implement an 'AI + X' interdisciplinary education model. This involves deeply integrating AI technology into traditional disciplinary teaching, such as in cross-disciplinary fields like 'AI + Medicine,' 'AI + Law,' and 'AI + Finance,' aiming to cultivate versatile talents possessing both deep disciplinary expertise and proficiency in AI technology. This model seeks to break down traditional disciplinary boundaries, fostering cross-fertilization and integrated innovation of knowledge systems. Secondly, build new industry-academia collaboration platforms. Actively facilitate student participation in real-world AI project practices. Through project-driven pedagogy, this significantly enhances their innovative thinking and practical operational skills. It helps students apply theoretical knowledge to solve complex real-world problems while deepening their understanding of the current application status and future trends of AI technology across industries. Thirdly, strengthen faculty development. Proactively recruit and cultivate full-time teachers with AI disciplinary backgrounds to provide solid teaching and research support for AI literacy education. Teachers should possess profound AI theoretical knowledge and rich practical experience to guide students in deeply exploring the internal logic and broad applications of AI technology. This will drive the sustained high-quality development of AI literacy education and foster the cultivation of high-caliber innovative talent.

5.2.3 Focus on equity in AI literacy education

The 2023 Global Education Monitoring Report by UNESCO, Technology in Education: A Tool on Whose Terms? [24], highlights that globally, disparities persist in the ownership and benefits derived from technology. Technology has not yet truly fulfilled its potential in achieving educational equity and justice. Educational equity has always been a core concern for nations and people. Faced with rapid technological development and widespread penetration, while ensuring equal educational opportunities and fair resource distribution, emphasis must also be placed on digital education equity to eliminate the digital divide. This requires educational policymakers, when making decisions related to information technology, to start from the specific national context, the current situation and needs of learners, and the most disadvantaged groups. They must ensure the dividends of IT development are equally enjoyed by all, making technology a sustainable development enabler for everyone. Therefore, governments and all sectors of society should collaborate to ensure equitable access to AI literacy education mechanisms.

Firstly, formulate and optimize relevant policy frameworks. Strengthen financial support for schools in rural and remote areas, committed to building and improving digital infrastructure systems, encompassing computer labs, high-speed internet facilities, etc. This aims to effectively reduce the significant disparities in educational hardware between urban and rural areas and across regions, narrowing gaps in digital access and digital literacy. Secondly, focus on the diversified and differentiated development of educational resources. Design and develop a range of AI educational resources tailored to different age groups and diverse learning needs. These include, but are not limited to, interactive online courses, virtual simulation labs, and intelligent teaching materials. Leverage the broad connectivity of internet technology to achieve widespread dissemination and sharing of high-quality educational content, promoting balanced allocation of educational resources. Finally, deeply attend to the educational rights and needs of special student groups, particularly students with disabilities and those from economically disadvantaged backgrounds. By developing accessible learning platforms, implementing precise learning support strategies, and providing targeted grants and scholarship programs, tailor personalized learning pathways and assistive tools for them. Ensure these groups can equally participate in and benefit from high-quality AI literacy education, thereby promoting the comprehensive popularization and deep equitable development of AI literacy education.

5.2.4 Strengthen interdisciplinary research and collaboration

AI literacy education constitutes a highly complex and multi-disciplinary system. Its knowledge domain broadly encompasses academic fields such as computer science, education, psychology, and ethics [25]. Given this system's comprehensiveness and frontier nature, future research urgently needs to strengthen interdisciplinary collaboration mechanisms. This is essential to break down traditional disciplinary barriers and foster deep interaction, fusion, and collaborative innovation among different fields of knowledge.

Firstly, build interdisciplinary research communities. Focus on deeply exploring the essential characteristics of AI literacy, investigating effective cultivation strategies, and constructing a scientifically rigorous evaluation framework.

This will provide solid and systematic theoretical support for educational practice. Secondly, promote interdisciplinary teaching models. Initiatives such as offering interdisciplinary elective course modules and organizing regular interdisciplinary academic seminars can enrich the application scenarios of AI in teaching, deepen its integration with pedagogy, and enable students to grasp fundamental AI theories while gaining profound insights into their practical applications in other academic domains. Finally, strengthen international academic exchange and cooperation. Drawing on international advanced experiences while adapting them to the national educational context is an effective pathway for continuously enriching and optimizing the theoretical system and practical models of AI literacy education, thereby driving high-quality development in this field. In this process, attention must be paid to localized innovation, ensuring the effective translation and integration of international experiences to build an AI literacy education system aligned with China's national conditions.

5.2.5 Address the ethical and societal impacts of AI technology

Education is not merely the transmission of skills; it is also the shaping of values. AI technology possesses unique advantages but also brings negative consequences stemming from these advantages, a key one being ethical risks. Examples include academic integrity crises caused by rapid data generation and cybersecurity issues arising from 'black box' technology. Therefore, in promoting AI literacy education, the ethical and societal impacts of AI technology must be given paramount importance.

Firstly, employ diversified teaching methods such as in-depth case studies, ethics seminars, scenario simulations, and role-playing to stimulate students' critical thinking and moral reasoning abilities. This equips them with the capacity to rationally and comprehensively evaluate the dual nature of AI technology, avoiding both blind deification and irrational fear. Secondly, the education system should strengthen education for students on personal privacy protection, data security management, and identifying algorithmic biases. The aim is to guide students in constructing a scientifically sound framework for technological ethics, growing into AI users who are both technically proficient and deeply versed in ethical norms. This process concerns not only knowledge transmission but also the cultivation of a sense of civic responsibility for the future society. Finally, encourage students to actively participate in social practice projects, transforming their acquired AI knowledge into solutions for real societal problems. Examples include using AI to optimize public services, advance environmental protection, or enhance healthcare efficiency. Such practical activities not only deepen students' understanding of the social responsibilities associated with AI technology but also hone their sense of social responsibility and mission through practice. More importantly, by tackling real-world problems, students' practical abilities and innovative thinking will be significantly enhanced, laying a solid foundation for their diverse career development. Emphasizing the combination of ethics education and practical application is an indispensable pathway for cultivating AI-era talent with high ethical standards and a strong sense of social responsibility.

5.2.6 Build a lifelong learning system for AI literacy education

As AI technology continuously innovates and its application boundaries persistently expand, individuals face unprecedented challenges: the need to constantly update their knowledge structures and skill sets to adapt to the rapid changes in the social environment. In this context, constructing a systematic and continuous AI literacy education system is particularly crucial.

Firstly, the design of this framework should be based on a diverse, multi-layered learning ecosystem. Integrate various learning resources and platforms, including online courses, micro-courses, and MOOCs (Massive Open Online Courses), aiming to provide flexible, diverse, and content-rich educational opportunities for learners across a wide age range and professional backgrounds, meeting their personalized learning needs. Secondly, to further motivate individuals to engage in continuous learning and effectively assess their learning outcomes, establishing a comprehensive and flexible learning achievement recognition and transfer mechanism is essential. This mechanism should encompass multiple recognition pathways such as certification, credit accumulation, and degree conferral. It should also establish a system for the mutual recognition and transfer of learning achievements across different educational stages and career development trajectories. This ensures learners can navigate freely through diverse learning and development paths, achieving seamless connection and upgrading of knowledge and skills. Finally, building a lifelong learning system for AI literacy education is a systemic project requiring deep collaboration and joint participation from governments, enterprises, and all sectors of society. At the government level, relevant policies and regulations should be introduced to provide guidance and support for the popularization and deepening of AI literacy education. Enterprises should leverage their industry advantages to assist learners in skill enhancement and career transition through concrete measures like providing internship and training bases and employment guidance services. All sectors of society should actively participate in the research, development, and sharing of educational resources, jointly fostering an open, inclusive, and positive learning atmosphere and cultural environment, contributing wisdom and strength to the construction of the lifelong learning system.

In summary, AI literacy education, as a key strategy for addressing future societal challenges, is becoming increasingly prominent. Through implementing measures such as strengthening AI literacy in elementary education, promoting deep integration of AI education in higher education, focusing on educational equity, enhancing interdisciplinary research and collaboration, addressing the ethical and societal impacts of AI, and building a lifelong learning system for AI literacy, the goal is to construct a comprehensive, multi-layered AI literacy education ecosystem. Through multi-party cooperation and joint efforts, this will effectively propel AI literacy education towards greater depth and breadth, laying a solid foundation for cultivating high-quality, versatile talent adapted to future societal needs and contributing to societal progress and sustainable development.

Through analyzing the research hotspots, trends, and implications in the field of AI literacy, this study provides robust

data support and a decision-making basis for formulating educational policies, allocating educational resources, and innovating teaching methods. Future research should continue to monitor the development dynamics of AI literacy, constantly explore new educational models and evaluation mechanisms, and promote the popularization and enhancement of AI literacy education. Simultaneously, exchanges and cooperation with the international community should be strengthened to jointly address the challenges and opportunities brought by AI technology, contributing wisdom and strength to building a community with a shared future for mankind.

COMPETING INTERESTS

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

FUNDING

This project was supported by 2025 Henan Province Higher Education Institutions Humanities and Social Sciences Research Youth Project: "Assessment and Enhancement Strategies for University Students' Digital Literacy in the Digital Intelligence Era" (2025-ZZJH-100); 2025 Henan Institute of Science and Technology Teacher Education Curriculum Reform Project: "Research and Practice on the Cultivation System of Artificial Intelligence Literacy for Teacher Training Students" (2025JSJY17); Henan Province Teacher Education Quality Enhancement Action Plan Demonstration Project (Teacher [2024] No. 21).

REFERENCES

[1] Frey C B, Osborne M A. The future of employment: How susceptible are jobs to computerisation? Technological Forecasting and Social Change, 2017, 114,: 254-280. https://doi.org/10.1016/j.techfore.2016.08.019.
[2] State Council of the People's Republic of China. Notice on issuing the new generation artificial intelligence development plan (Document No. Guo Fa [2017] No. 35). Gazette of the State Council of the People's Republic of China, 2017(22): 7-21.
[3] Ministry of Education of the People's Republic of China. Notice on issuing the artificial intelligence innovation action plan for higher education institutions (Document No. Jiao Ji [2018] No. 3). Gazette of the Ministry of Education of the People's Republic of China, 2018(04): 127-135.
[4] Ministry of Education of the People's Republic of China. Notice on issuing the education informatization 2.0 action plan (Document No. Jiao Ji [2018] No. 6). Gazette of the Ministry of Education of the People's Republic of China, 2018(04): 118-125.
[5] Cyberspace Administration of China. Interim measures for the management of generative artificial intelligence services. 2023. http://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm.
[6] Office of the Central Cyberspace Affairs Commission. Global artificial intelligence governance initiative. 2023. http://www.cac.gov.cn/2023-10/18/c_1699291032884978.htm.
[7] Wu F, Li Y, Chen J, et al. Red paper on artificial intelligence literacy of university students (2024 Edition). Science and Education Development Research, 2024, 4(02): 71-96.
[8] Konishi Y. What is needed for AI literacy? Priorities for the Japanese economy in 2016. RIETI Column, 2016. https://www.rieti.go.jp/en/columns/s16_0014.html.
[9] Long D, Magerko B. What is AI literacy? Competencies and design considerations. In Proceedings of the 2020 ACM Conference on Human Factors in Computing Systems (CHI '20). Association for Computing Machinery, 2020: 1-16. https://doi.org/10.1145/3313831.3376727.
[10] Zhang Y, Yang G, Xu J, et al. Construction of an artificial intelligence literacy model and its implementation path. Modern Educational Technology, 2020, 32(03): 42-50.
[11] Yang H, Zhang D, Guo W. Research on the framework of artificial intelligence literacy under the background of STEM. e-Education Research, 2020, 43(04): 26-32.
[12] Wang Y, Wang Y, Yang Y. Content system and development logic of artificial intelligence literacy education in colleges and universities. Heilongjiang Researches on Higher Education, 2022, 40(02): 26-31.
[13] Zheng Q, Qin M, Li S. Theoretical model research of intelligent literacy in the era of human-machine collaboration. Fudan Education Forum, 2021, 19(01): 52-59.
[14] Cai Y, Zhang J, Yu C, et al. AI literacy in the era of digital intelligence: Connotation, framework and implementation path. Journal of Library Science in China, 2024, 50(04): 71-84. https://doi.org/10.13530/j.cnki.jlis.2024021.
[15] Chen W. The development and educational application of artificial intelligence technology from the perspective of multiple intelligences. e-Education Research, 2018, 39(07): 12-19. https://doi.org/10.13811/j.cnki.eer.2018.07.002.
[16] Government of Japan. AI Strategy 2022. National Strategy Report. National Institute of Science and Technology Policy. https://www.nistep.go.jp/en/?p=1396
[17] Wu G. General artificial intelligence: “Empowerment” or “danger”? 2023(05): 48-52.

[18] Ministry of Science and Technology of the People's Republic of China. 2022 global artificial intelligence innovation index report released. China Institute of Science and Technology Information (CISTI) website, 2022. http://istic.ac.cn.
[19] Mu J, Zhang Q. Cultivation of top innovative talents in elementary education in Nordic countries and its enlightenment. Teaching and Administration, 2024(28): 72-76.
[20] Zhong B. Reform of talent cultivation mode is the core of connotation construction in higher education institutions. Journal of Higher Education, 2013, 34(11): 71-76.
[21] Zhao L. Theoretical reflection on the high-quality development of elementary education under the background of digital transformation. Education Science Forum, 2024(28):1.
[22] Deng Y, Feng Y, Xing H, et al. Theoretical thinking and practical path of high-quality development of science education in primary and secondary schools in the new era. China Educational Technology, 2024(07): 14-27.
[23] Gai Q. Challenges and countermeasures of generative artificial intelligence empowering high-quality development of higher education. University Education, 2024(17): 6-20.
[24] Unesco. Global education monitoring report 2023: Technology in education: A tool on whose terms? 2023. https://unesdoc.unesco.org/ark:/48223/pf0000385721.
[25] Huang R, Shi L, Wu Y, et al. Construction of the content framework of artificial intelligence literacy education in China from a global perspective. Documentation, Information & Knowledge, 41(03): 27-37. https://doi.org/10.13366/j.dik.2024.03.027.

World Journal of Information Technology
Print ISSN: 2959-9903
Online ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3058

MULTI-GRANULARITY TIME SERIES FORECASTING METHODS BASED ON DUAL-CHANNEL FUSION

XueYuan Zhu*, JiaXin Peng
School of Transportation, Changsha University of Science & Technology, Changsha 410114, Hunan, China.
Corresponding Author: XueYuan Zhu, Email: xueyuan_zhu6007@163.com

Abstract: In high-frequency data environments, traditional time-series forecasting methods generally face two major challenges. First, the structures of these models are too simple to capture both the long-term trends and short-term disturbances. Second, the forecasting granularity is too coarse to meet the refined requirements for real-time dynamic decision-making. To address these issues, this study proposes a dual-channel fusion forecasting framework, the Dual-Resolution Adaptive Forecasting Topology (DRAFT) architecture. The architecture comprises two modules: a trend modeling module and a disturbance modeling module. The modules are responsible for processing the linear trend components and nonlinear fluctuation signals in the time series data. They achieved adaptive integration of the prediction results using a lightweight fusion mechanism. Experiments on real-world datasets demonstrated that the DRAFT architecture significantly outperformed traditional single-model approaches in terms of metrics such as mean squared error (MSE) and mean absolute error (MAE), with error reductions exceeding 54.05% in certain scenarios. Furthermore, DRAFT possesses the capacity to refine the prediction output granularity to the 10-minute level, thereby providing more actionable prediction information for high-timeliness scenarios. This study establishes a new paradigm for the precise prediction of complex time-series data and provides theoretical and practical references for the construction of modular prediction systems.
Keywords: Multi-granularity prediction; Time series modeling; Model fusion; Predictive granularity refinement

1 INTRODUCTION

Among numerous real-time decision support systems, accurate predictions of future quantities are a core prerequisite for ensuring system efficiency and rational resource allocation. With the continuous advancement of data collection technology, the temporal granularity of data acquisition has become increasingly refined. However, the temporal resolution of predictive models still lags behind the requirements of real-world applications. Such issues are particularly pronounced in scenarios characterized by task-intensive scheduling and the need to respond to instantaneous fluctuations, where the requirements for the response speed and accuracy of the predictive methods are significantly heightened.

Multiscale time series simultaneously exhibit linear trends, nonlinear disturbances, and random fluctuations, which pose challenges for single modeling strategies. Traditional statistical models (e.g., ARIMA) excel at handling stable trends but struggle to capture high-frequency nonlinear changes, with a limited ability to fit complex nonlinear features [1-2]. In contrast, while deep learning models (such as LSTM) possess strong nonlinear modeling capabilities, they often exhibit limitations in terms of interpretability, stability, and the handling of short-term anomalies [3-4]. Additionally, most current forecasting research remains focused on hourly or daily granularity, with a coarse temporal resolution that fails to meet the practical demand for “minute-level” dynamic responses, creating a significant tension between timeliness and practicality.

Given the dual challenges of structural adaptability and time sensitivity in multi-granularity time-series forecasting tasks, there is an urgent need for a hybrid forecasting framework that can balance trend modeling, sudden change capture, and multi-timescale response [5]. This study proposes a modular, responsive, and scalable structured forecasting system called Dual-Resolution Adaptive Forecasting Topology (DRAFT). This architecture builds a multifunctional collaborative mechanism, enabling the model to simultaneously capture long-term trends and short-term fluctuations, effectively bridging the performance gap between static modeling and dynamic response. Its core concept is to use ARIMA to capture the linear patterns in the data while using LSTM to learn the complex nonlinear structures in the residuals. Finally, through a fusion mechanism, the outputs of both models are balanced to construct a prediction system with both robustness and generalization capabilities.

Unlike existing single-strategy approaches, the DRAFT architecture significantly enhances the generalization capabilities of the system while ensuring structural transparency through hierarchical learning and output fusion mechanisms. Its design emphasizes a fine-grained response in prediction granularity, offering comprehensive adaptability from macro-level trend analysis to micro-level disturbance resolution, and is particularly suited for real-time prediction scenarios involving high-density time series. The DRAFT architecture innovatively integrates the advantages of multiple models, advancing the prediction granularity from the traditional hourly level to a 10-minute level, significantly enhancing the spatiotemporal adaptability of the model. Extensive experimental validation of classic multivariate time series tasks demonstrates that the system outperforms the baseline methods on multiple key evaluation

metrics while maintaining good interpretability and scalability, achieving a deep integration of theory and practice. It can serve as a general-purpose solution for high-frequency prediction.

2 LITERATURE REVIEW

In the field of time-series forecasting, researchers have long explored model structures, data characteristics, and granularity response capabilities. This paper reviews existing research from the following three perspectives: (1) trend-driven modeling strategies, (2) nonlinear structure learning methods, and (3) the evolution of multi-granularity forecasting frameworks.

2.1 Trend Modeling and Robustness of Statistical Methods

Traditional time series analysis methods are primarily based on statistical modeling, with their core advantages being strong parameter interpretability and transparent modeling processes, which are particularly suitable for handling stationary sequences and linear trends. Such methods typically rely on differencing, autoregression, and error structures to construct predictive functions, with representative works including differenced moving average models under the assumption of stationarity and seasonal trend analysis tools. Bichescu et al. proposed an innovative time-series analysis method that simplifies the construction process of ARIMA models and may improve the efficiency and accuracy of predictions [6]. Li et al. used an autoregressive integrated moving average (ARIMA) model to predict the development trend of gonorrhea, providing a reference for formulating corresponding prevention and control strategies [7].

However, traditional statistical modeling methods have obvious limitations. When faced with the nonlinear disturbances and structural changes commonly found in the real world, their model architecture, based on linear assumptions and stationarity premises, struggles to effectively capture the complex dynamic changes in data, leading to a significant decline in prediction accuracy. In high-frequency, non-stationary data scenarios, such as minute-level stock price fluctuations in financial markets or real-time changes in power load, traditional models fail to promptly capture the instantaneous fluctuations and structural changes in data, resulting in severe degradation of model performance. Additionally, traditional statistical models typically use hourly or daily time windows, which have a relatively coarse temporal resolution. This fails to meet the urgent needs of modern precision decision-making systems for high-timeliness and high-accuracy predictions and cannot provide timely and accurate information support for dynamic decision-making.

2.2 Nonlinear Pattern Learning and Neural Prediction Mechanisms

In recent years, the rapid development of deep learning technology has led to revolutionary breakthroughs in the field of time-series forecasting. Neural network models oriented toward sequence modeling have demonstrated strong fitting capabilities and generalization potential in time-series forecasting. In particular, recurrent structures such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) have effectively addressed the gradient vanishing and long-term dependency issues inherent in traditional Recurrent Neural Networks (RNNs) by introducing gating mechanisms and are widely applied in nonlinear sequence modeling tasks. Ma et al. utilized LSTM neural networks to process highly nonlinear, dynamic, and time-dependent sequence data in industrial processes, providing an effective soft sensor technology for modeling issues related to the strong time-varying characteristics of the process and predicting key variables [8]. Yin et al. proposed an LSTM-based multistate vector sequence-to-sequence model for rainfall-runoff modeling, achieving a multistep runoff prediction [9].

These methods excel in capturing long-term dependencies, local anomalies, and non-stationary features, making them suitable for constructing nonlinear mappings between complex inputs and outputs. However, neural network-based methods face significant challenges. Their “black box” nature makes it difficult to intuitively interpret the internal decision-making mechanisms of the model, resulting in poor model interpretability; the training process heavily relies on large-scale labeled data, resulting in significantly reduced model performance in data-scarce scenarios; additionally, when faced with sudden short-term temporal changes, owing to the inherent delay in the model's computation and update mechanisms, neural networks often struggle to provide timely and accurate predictive responses, limiting their application in high-timeliness decision-making scenarios.

2.3 Development of Multi-Granularity Response Mechanisms and Fusion Frameworks

To balance model stability and expressive power, the academic community has increasingly turned to research on model fusion and structural integration in recent years. These methods typically integrate sub-models with different modeling properties into a unified framework to achieve hierarchical learning and prediction of different signal components. Typical fusion strategies include weighted combination, residual stacking, and hierarchical recursion, which can improve the model adaptability and prediction accuracy to a certain extent. Meanwhile, some studies have attempted to introduce prediction tasks with finer temporal granularity (e.g., 10-minute intervals) to address the dual demands for timeliness and accuracy in real-time optimization and dynamic scheduling scenarios. Lu et al. proposed an integrated multi-temporal granularity deep learning prediction method (Mul-DesLSTM) for short-term passenger flow prediction in urban rail transit systems. This method aims to address the issue of high-resolution data generated by automatic fare collection (AFC) systems being wasted [10]. He et al. proposed a dynamic multi-fusion spatiotemporal

graph neural network for multivariate time-series prediction. This method aims to simultaneously capture hidden temporal and spatial patterns in spatiotemporal data [11].

To balance model stability and expressive power, the academic community has increasingly turned to research on model fusion and structural integration in recent years. These methods typically integrate sub-models with different modeling properties into a unified framework to achieve hierarchical learning and prediction of different signal components. Typical fusion strategies include weighted combination, residual stacking, and hierarchical recursion, which can improve the model adaptability and prediction accuracy to a certain extent. Meanwhile, some studies have attempted to introduce prediction tasks with finer temporal granularity (e.g., 10-minute intervals) to address the dual demands for timeliness and accuracy in real-time optimization and dynamic scheduling scenarios. Lu et al. proposed an integrated multi-temporal granularity deep learning prediction method (Mul-DesLSTM) for short-term passenger flow prediction in urban rail transit systems. This method aims to address the issue of high-resolution data generated by automatic fare collection (AFC) systems being wasted [10]. He et al. proposed a dynamic multi-fusion spatiotemporal graph neural network for multivariate time-series prediction. This method aims to simultaneously capture hidden temporal and spatial patterns in spatiotemporal data [11].

However, current related research still faces limitations, such as strong structural closedness, weak granularity adaptation mechanisms, and coarse fusion strategies. Strong structural closedness manifests as a lack of systematic construction of multi-module collaborative mechanisms, leading to insufficient information interaction and collaboration efficiency between sub-models. Weak granularity adaptation mechanisms are manifested in the widespread use of fixed-interval segmentation modeling, which struggles to dynamically respond to actual temporal changes. Coarse fusion strategies refer to the fact that most studies still rely on simple averaging or linear weighting, neglecting the heterogeneous relationships between model outputs. These limitations highlight the urgent need for a more flexible, adaptive, and interpretable fusion framework to systematically coordinate multi-module interactions, dynamically adapt to changes in granularity, and optimize fusion strategies based on output features.

2.4 Positioning and Innovation of This Study

Based on the aforementioned research, this study proposes a dual-channel adaptive forecasting architecture for multi-granularity time series, namely, Dual-Resolution Adaptive Forecasting Topology (DRAFT). This architecture decouples the linear trend components and nonlinear volatility features of time series, constructing a collaborative processing framework comprising a trend modeling module (based on an ARIMA-based linear feature extractor) and a disturbance modeling module (based on an LSTM-based nonlinear residual learner). It achieves the dynamic calibration of dual-path outputs through a lightweight voting fusion mechanism, enabling hierarchical modeling capabilities for complex data structures.

Compared with existing methods, the innovations of the DRAFT architecture are reflected in the following three aspects: (1) Clear division of functional modules, enabling each substructure to focus on specific signal component modeling tasks. The decoupled design of linear trend modeling and nonlinear disturbance learning avoids the feature interaction interference issues in traditional hybrid models, enhancing modeling efficiency while improving model interpretability; (2) Fine-grained temporal response: Breaking through the temporal granularity limitations of traditional prediction models, the DRAFT architecture introduces a temporal granularity calibration mechanism to refine prediction output granularity to the 10-minute level, significantly enhancing the model's dynamic response capability to high-frequency data; (3) Reconstructable prediction results: Supports flexible reconstruction and combination of trend and disturbance components based on actual application scenarios, forming customized prediction result output modes, effectively enhancing the model's practicality and scalability in multi-scenario decision support systems.

3 METHODOLOGY

The proposed Dual-Resolution Adaptive Forecasting Topology (DRAFT) architecture aims to effectively improve the accuracy, stability, and response speed of time-series forecasting through multi-module collaboration and structural fusion mechanisms. The architecture consists of three core components: a trend modeling module (Trend Module), disturbance capture module (Disturbance Module), and result Fusion Module (Fusion Module). The overall process is illustrated in Figure 1.

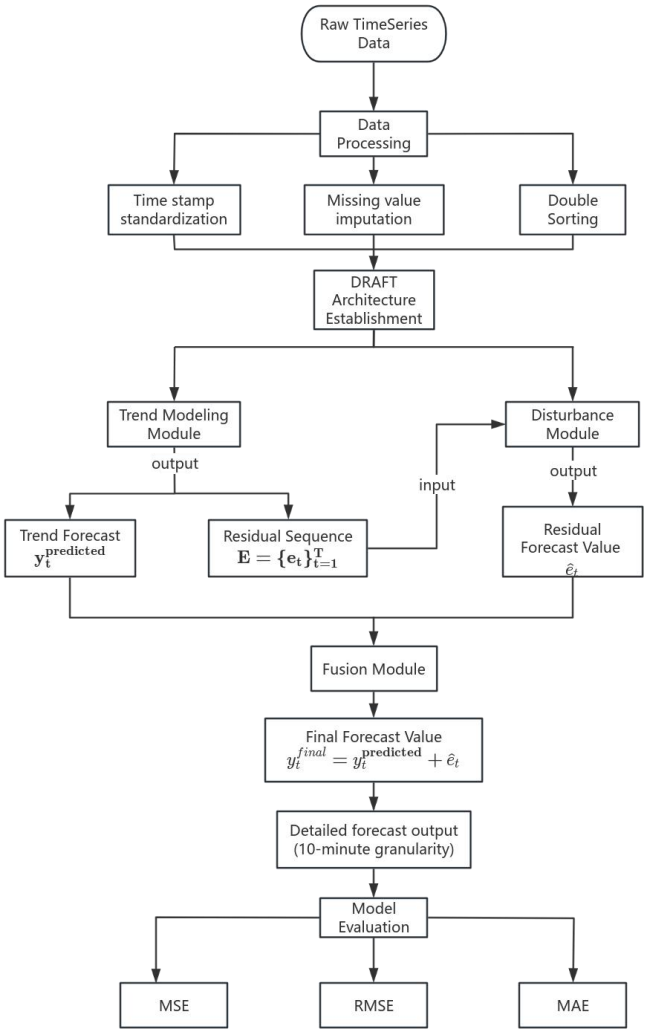

Figure 1 Overall Flow Chart

3.1 Data Preprocessing and Modeling Fundamentals

In high-frequency time-series forecasting scenarios, data preprocessing is a critical foundational step in modeling. For multidimensional datasets comprising timestamps, category identifiers, and target variables, the following standardized processing steps must be executed: initially, time-related features undergo normalization and integration, whereby discrete date and time information is consolidated into a unified timestamp format and subsequently arranged in chronological order. This ensures the consistency of the temporal logic in the data. Second, statistical interpolation methods were employed to address issues related to missing data, with the objective of ensuring data integrity and preventing modeling biases. The dataset was then sorted based on both category identifiers and time dimensions to construct a structured temporal feature matrix. This process clearly reveals the trend, periodicity, and abnormal fluctuation patterns of the target variable over time. The result is a standardized input for subsequent multi-granularity modeling.

The preprocessing framework is applicable to various types of multidimensional data with temporal dependencies. The integration of data formats, rectification of data defects, and augmentation of temporal characteristics serve as the basis for the effective training and precise prediction of multi-granularity time-series forecasting models.

3.2 Trend Modeling Module

The present module is predicated on the difference integrated moving average autoregressive model (ARIMA) and aims to analyze the linear trend components and cyclical patterns in time series, thereby modeling relatively stable long-term trends and repetitive patterns. The core process of the system under investigation revolved around the three core steps of the ARIMA model. The configuration of the ARIMA model is shown in Figure 2.

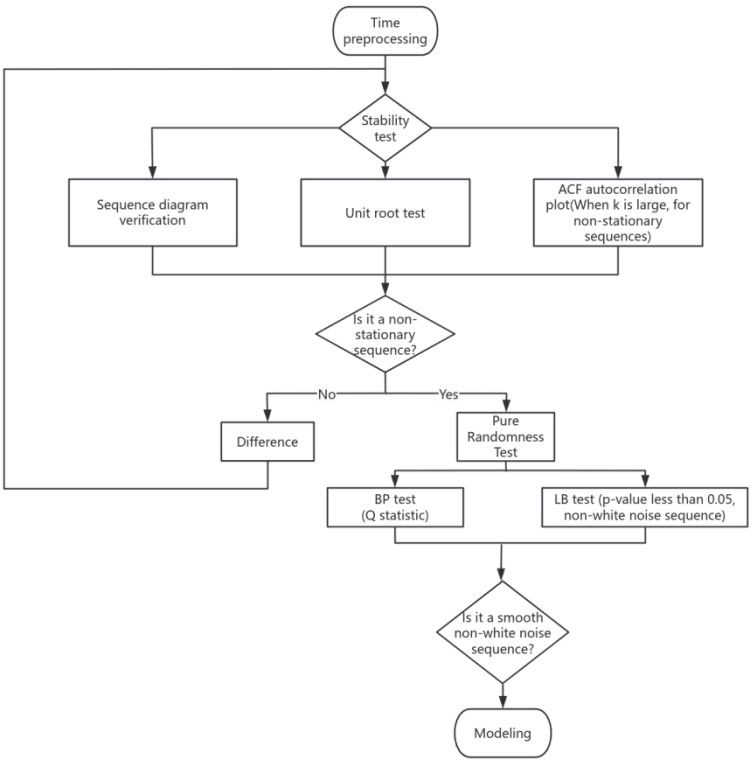

Figure 2 ARIMA Model Structure Diagram

The specific implementation logic is as follows:

3.2.1 Sequence stability analysis and differential processing

The stationarity of a time series is a prerequisite for applying the ARIMA model. The module initially employs a unit root test to ascertain the stationarity of the original series data. In instances where the series under consideration exhibits a substantial trend or seasonality, the trend component is extirpated through the implementation of differencing, thereby yielding a stationary series.

The formula for first-order differencing is as follows.

$$\Delta y_t = y_t - y_{t-1} \tag{1}$$

The formula for second-order difference is.

$$\Delta^2 y_t = \Delta y_t - \Delta y_{t-1} \tag{2}$$

In the context of a time series characterized by a linear trend, the trend can be eliminated through the implementation of first-order differencing, thereby rendering the series stationary.

The foundation for subsequent model fitting is laid by repeated differencing until the series satisfies a stationarity condition.

3.2.2 ARIMA model structure construction

Subsequent to the smoothing process, the module constructs a model based on autoregressive (AR) and moving average (MA) structures.

The autoregressive (AR) model posits that the value at the present moment y_t can be represented by a linear combination of the values at past moments and an error term.

$$y_t = c + \sum_{i=1}^{p} \phi_i \, y_{t-i} + \epsilon_t \tag{3}$$

Among them, c is the constant term, ϕ_i is the autoregressive coefficient, and ϵ_t is the white noise error term.

The moving average (MA) model assumes that the current value y_t can be represented by a linear combination of the current error term and the error terms from the previous p moments.

$$y_t = \epsilon_t + \sum_{j=1}^{q} \theta_j \, \epsilon_{t-j} \tag{4}$$

The model captures the historical dependencies of the sequence through the AR term and fits the moving average pattern of the error term through the MA term, thereby achieving joint modeling of linear trends and cyclical components.

3.2.3 Hyperparameter optimization and model training

The module uses a grid search algorithm to optimize the hyperparameters of the ARIMA model (p is the autoregressive order, d is the difference order, and q is the moving-average order). The mean square error (MSE) was used as the

objective function, and the optimal parameter combination was determined by minimizing the prediction error on the training set. The objective function expression is as follows:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_t^{\text{actual}} - y_t^{\text{predicted}}\right)^2 \tag{5}$$

Where is the predicted value of the y trend module and n is the number of training samples. By iterating through the feasible combinations of (p, d, q), the parameter combination that minimizes the MSE was selected as the final model configuration.

After parameter optimization, the ARIMA model can generate a trend prediction sequence that describes the path of macro changes. This sequence removes the nonlinear disturbance components in the original data and focuses on describing the long-term trends and cyclical patterns of the data, providing residual input for the subsequent disturbance capture module.

3.3 Disturbance Module

This module is dedicated to mining nonlinear dynamic features in time series trend residuals, focusing on modeling short-term sudden changes, historical sequence dependencies, and complex fluctuation patterns to improve the model accuracy in capturing unpredictable disturbances and its dynamic response capabilities. The functionality of this module was achieved using a long short-term memory (LSTM) network model.

The core unit of the LSTM model is illustrated in Figure 3.

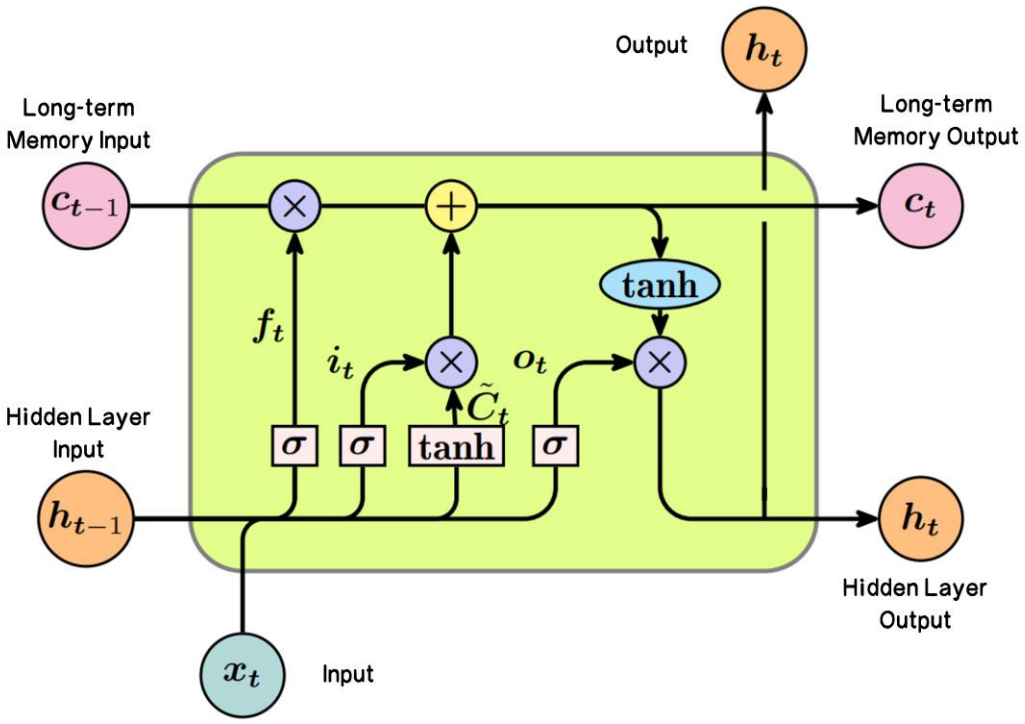

Figure 3 LSTM Model Structure Diagram

The calculation process consists of the following steps:

ForgetGate:

The ForgetGate determines how much information in memory unit C_{t-1} needs to be forgotten at the previous moment. The calculation formula is as follows:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{6}$$

Where σ is the sigmoid function, whose output range is 0 to 1. W_f is the weight matrix of the ForgetGate, $[h_{t-1}, x_t]$ represents concatenating the hidden state h_{t-1} from the previous time step with the current input x_t, and b_f is the bias of the ForgetGate. When f_t approaches 0, it indicates that most of the information is forgotten; when f_t approaches 1, it indicates that most of the information is retained.

InputGate:

The InputGate determines the amount of new information added to the memory unit at the current moment.

First, the output of the InputGate is calculated using formula (7), while candidate memory units are produced.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{7}$$

The output range of function tanh is -1 to 1.

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{8}$$

Then, the output of the InputGate is multiplied by the candidate memory unit to obtain the new information to be added to the memory unit.

Memory unit update: The formula for updating memory units is as follows：

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{9}$$

Where represents element-wise multiplication. This formula represents adding the information to be retained in the memory unit at the previous moment and the new information at the current moment to obtain the memory unit at that time.

OutputGate:

The OutputGate determines which information in the memory unit will be used to generate the output at the current moment. The calculation formula is as follows：

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{10}$$

The hidden state at the current time is $h_t = o_t * \tanh(C_t)$. Through the control of the output gate, the LSTM unit can selectively output information from the memory unit.

The specific implementation logic of the Disturbance Module is as follows:

3.3.1 Input design

The module input is the residual sequence $E = \{e_t\}_{t=1}^{T}$ generated by the trend module,$e_t=y_t - y_t^{predicted}$,$y_t$ is the original time series value, and $y_t^{predicted}$ is the predicted value of the trend module.Residual sequences typically contain nonlinear components that are not explained by linear trends in the original data and need to be further modeled using nonlinear models.

3.3.2 LSTM network architecture

The module adopts a multi-layer LSTM network structure, which uses its gating mechanism (InputGate, ForgetGate, OutputGate) to selectively retain historical information to capture long-range dependencies. The network structure specifically includes an embedding layer, LSTM layers, and fully connected layers. The embedding layer standardizes the input residual sequence to improve training stability; the LSTM layer stacks 2–3 layers of LSTM units, each containing n_h memory units, using the forget gate to filter out irrelevant historical information and transmitting long-term dependency features through cell states; the fully connected layer maps the hidden states output by the LSTM layer to the predicted values $\hat{e}_t$ via linear transformation, i.e., the predicted values of the perturbation components.

3.3.3 Hyperparameter optimization mechanism

Automated tuning of network hyperparameters using a random search algorithm. The core optimization parameters include: Historical window length L, which determines the time span of the input sequence, i.e., the past L residual values input into the model each time, used to capture local dependency patterns; The number of LSTM units is n_h, which controls the network's nonlinear fitting capability. A larger number of units can capture more complex feature interactions, but overfitting must be avoided; Training epochs are determined through cross-validation to prevent underfitting or overfitting; The learning rate uses an adaptive learning rate algorithm to dynamically adjust the update step size, accelerating convergence.

3.3.4 Time sliding window mechanism

To enhance the model's sensitivity to local changes, input data is processed using a nested sliding window structure. Overlapping windows cover the entire time period, enabling the model to capture local features at different time offsets and improving its responsiveness to short-term sudden changes. Mathematically, each window corresponds to a local time series segment, and its output is the residual prediction value $\hat{e}_{i+L}$ for future time points Δt, forming a "many-to-one" prediction model.

This module effectively compensates for the blind spots of the trend module in processing non-stationary and non-linear components through the memory characteristics and non-linear mapping capabilities of LSTM.

3.4 Fusion Module

After modeling in the trend modeling module and disturbance capture module, the result fusion module systematically integrates the outputs of the two pathways, combining the predicted value $y_t^{predicted}$ from the trend module with the output $\hat{e}_t$ from the disturbance module to generate the final multi-granularity prediction result. The core design goal of this module is to balance the stability of linear trend modeling with the flexibility of nonlinear disturbance modeling, and to improve the robustness and accuracy of the prediction results by optimizing the fusion strategy. The specific formula is as follows:

$$y_t^{final} = y_t^{predicted} + \hat{e}_t \tag{11}$$

Through the result fusion module, information from the trend modeling and disturbance capture modules can be integrated to form a final output sequence with greater robustness and fewer errors.

3.5 Precision Prediction Output Mechanism

To meet the demand for fine-grained predictions in high-frequency decision-making scenarios, the DRAFT architecture designs a multi-granularity dynamic mapping mechanism that decomposes macro-scale prediction results into 10-minute granularity while ensuring the temporal consistency and total conservation of prediction values. This mechanism is achieved through a total conservation interval allocation strategy.

First, using historical data statistical patterns and the DRAFT architecture, predict the forecast result y_{total} for the macro time interval K. To ensure that the forecast results are refined to a 10-minute granularity level and maintain

consistency in the total quantity of fine-grained forecast values, the forecast values for the macro interval are further decomposed into 10-minute time segments using the principle of proportional conservation. Assuming that the macro time interval K contains N 10-minute granularity intervals, with the prediction value for the i-th 10-minute granularity interval being $y_{k,\ n}$, the specific calculation formula is as follows:

$$y_{k,\ n} = y_{total} * \omega_{k,\ n} \tag{12}$$

$$\omega_{k,\ n} = \frac{y_{k,\ n}^{hist}}{\sum_{n=1}^{N} y_{k,\ n}^{hist}} \tag{13}$$

$$\sum_{n=1}^{N} \omega_{k,\ n} = 1 \tag{14}$$

Where $\omega_{k,\ n}$ is the fine-grained prediction ratio of $y_{k,\ n}$, and $y_{k,\ n}^{hist}$ is the historical value of $y_{k,\ n}$.

The refined prediction output mechanism enables the model's prediction results to be finely granular and operational, while ensuring that the finely granular prediction values are consistent with the total amount, supporting downstream tasks such as dynamic allocation and elastic scheduling.

4 EXPERIMENTS AND ANALYSIS OF RESULTS

To validate the effectiveness of the Dual-Resolution Adaptive Forecasting Topology (DRAFT) architecture in multi-granularity time series forecasting tasks, this study designs a set of experiments based on real business data. The experiments aimed to evaluate the performance of the model in terms of forecasting accuracy, stability, and micro-response capabilities.

4.1 Dataset and Experimental Setup

The experiment utilized real-world short-haul logistics operation data encompassing multiple typical sequence paths, with a total sample size exceeding tens of thousands of entries spanning a continuous 16-day period. The data granularity was at the daily level and every ten-minute interval. The dataset includes multidimensional features such as timestamps, route identifiers, and historical quantities. The data were preprocessed to standardize the date and time format and sort the records; date and time processing was performed to merge the information into a standard date and time format; missing values were handled using the mean imputation method to ensure the accuracy of subsequent analysis and modeling; and the data were sorted by route code and time to clearly show the trend of cargo volume over time.

Following the principle of time dependency in time series data, the dataset was divided into a training set (70%), validation set (15%), and test set (15%) in chronological order.

To evaluate the predictive performance of the model, this study used the mean square error (MSE), root mean square error (RMSE), and mean absolute error (MAE) to evaluate the merged model.

The formula for calculating the mean squared error (MSE) is as follows:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i^{actual} - y_i^{predicted}\right)^2 \tag{15}$$

Where is the sample size, y_i^{actual} is the i-th actual value, and $y_i^{predicted}$ is the i-th predicted value. MSE is sensitive to large errors because the errors are squared and then summed.

The formula for calculating the root mean square error (RMSE) is as follows:

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} \tag{16}$$

RMSE is the square root of MSE, and its units are the same as those of the original data, so it more intuitively reflects the average deviation between the predicted values and the actual values.

The formula for calculating the mean absolute error (MAE) is as follows:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i^{actual} - y_i^{predicted}\right| \tag{17}$$

MAE calculates the average absolute error between the predicted value and the actual value, and it is relatively insensitive to outliers.

4.2 Prediction Performance Comparison

This study compares the predictive performance of the DRAFT architecture with two mainstream baseline methods: a single trend modeling method (ARIMA model) and a single nonlinear sequence modeling method (LSTM model).

This study will evaluate the predictive performance of the three methods by comparing their mean square error (MSE), root mean square error (RMSE), and mean absolute error (MAE) results.

This study focuses on two representative sequence paths, namely Path A (Site 1-Station 16-0600) and Path B (Site 1-Station 26-1400), which represent high-volatility and high-stability path scenarios, respectively.The evaluation results of Path A model are shown in Table 1, and those of Path B model are shown in Table 2.

Table 1 Path A Model Evaluation

Model	MSE	RMSE	MAE
ARIMA Model	1656.2215	40.6967	26.0266
LSTM Model	1976.5175	44.4580	25.9353
DRAFT Architecture	1356.4174	36.8296	22.2234

Table 2 Path B Model Evaluation

Model	MSE	RMSE	MAE
ARIMA Model	7960.4913	89.2216	66.0651
LSTM Model	5968.2754	77.2546	56.4914
DRAFT Architecture	3658.7338	60.4875	44.4334

After conducting an evaluation metric analysis of the model's prediction results, it can be observed that in Path A, the DRAFT architecture achieved a significant reduction of 18.11% in mean squared error (MSE) compared to the ARIMA model, and a reduction of 31.37% compared to the LSTM model. Additionally, the root mean square error (RMSE) and mean absolute error (MAE) also show corresponding downward trends. In Path B, the DRAFT architecture's MSE is reduced by 54.05% compared to the ARIMA model and by 38.71% compared to the LSTM model. Meanwhile, the RMSE and MAE also exhibit downward trends.

Based on a comprehensive comparison of the evaluation metrics for the prediction results, the DRAFT architecture demonstrated superior prediction performance compared to the baseline method, both in high-volatility and low-volatility path predictions.

4.3 Fine-Grained Response Capability Analysis

At a 10-minute time resolution, the DRAFT model employs an interval allocation strategy based on total quantity conservation to achieve fine-grained predictions of data. By applying the DRAFT architecture, the model successfully identified peak traffic periods for package traffic on Path A (time span: December 15, 2024, 21:00 to 23:50) and Path B (time span: December 16, 2024, 11:00 to 13:50). The prediction results exhibit high consistency with the actual data sequence in terms of trends, with macro-cycle fluctuations aligning with the actual fluctuation trends, thereby validating the effectiveness of the trend module in modeling linear components.The comparison between the predicted results of Path A and the actual values is shown in Figure 4, and the comparison between the predicted results of Path B and the actual values is shown in Figure 5.

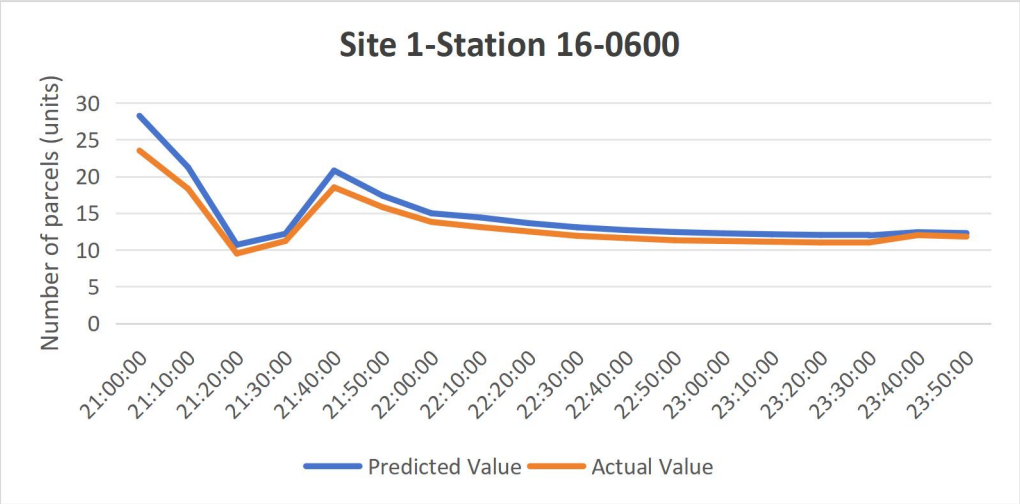

Figure 4 Comparison of Predicted Values and Actual Values for Site 1-Station 16-0600

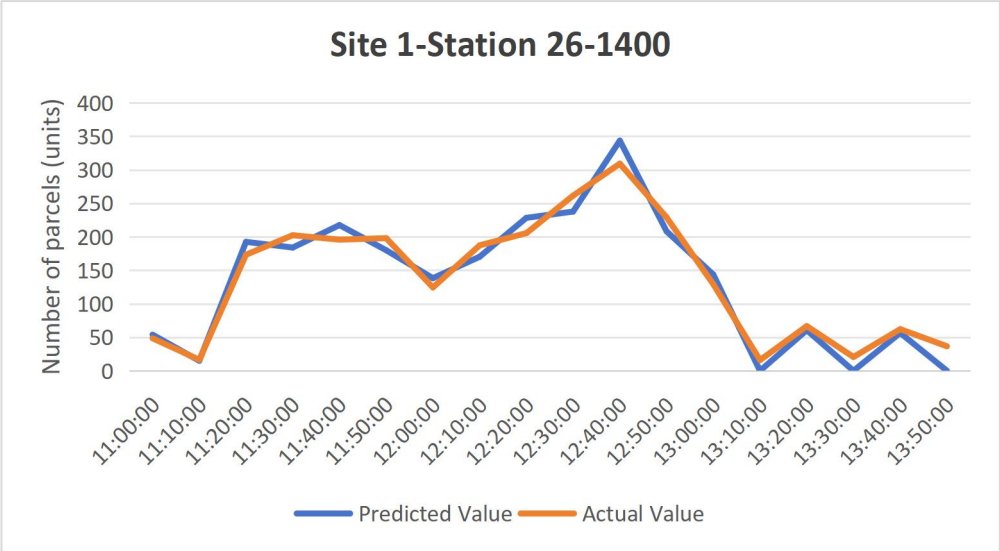

Figure 5 Comparison of Predicted Values and Actual Values for Site 1-Station 26-1400

4.4 Summary

In terms of core metrics for evaluating the performance of predictive models, including mean squared error (MSE), root mean squared error (RMSE), and mean absolute error (MAE), the DRAFT architecture demonstrates a significant reduction in error compared to traditional ARIMA models, with a decrease of over 54.05%; Compared to the LSTM model, the error reduction reaches 38.71%. This significant performance improvement demonstrates the DRAFT architecture's notable advantage in prediction accuracy. Specifically, the DRAFT architecture achieves more precise prediction results by integrating trend prediction modules with disturbance processing modules, significantly reducing the discrepancy between predicted and actual values. This validates the effectiveness of the dual-channel modeling strategy in capturing complex fluctuating phenomena.

Additionally, the DRAFT architecture adopts a total quantity conservation interval allocation strategy, which endows it with high efficiency in prediction output, enabling 10-minute-level fine-grained predictions. In practical applications, the prediction results generated by the DRAFT architecture exhibit high consistency with actual trends in both Path A and Path B prediction sequences. Especially in terms of macro peaks and actual fluctuations, the DRAFT architecture's prediction results can synchronize with actual fluctuations, a feature that greatly meets the urgent demand for fine-grained data in real-time scheduling systems and provides strong data support for practical operations.

5 CONCLUSION

This study addresses high-resolution time series forecasting tasks by proposing a modular fusion forecasting architecture, DRAFT (Dual-Resolution Adaptive Forecasting Topology). By modeling trend and disturbance signals through dual channels, it achieves 10-minute granularity forecasting outputs. This architecture significantly improves forecasting accuracy, stability, and application adaptability by functionally decoupling and fusing different types of forecasting sub-structures. Key research findings include: for the first time in time series forecasting, functional separation of trend modeling and disturbance modeling is achieved, constructing a modular combination system based on a collaborative mechanism to effectively address the inadequacy of single models in responding to complex sequences; in typical path data experiments, DRAFT outperforms traditional modeling strategies across multiple evaluation metrics, particularly in scenarios with high volatility or unstable trends, with a maximum error reduction exceeding 70%; By designing a fine-grained mapping mechanism at the output level, the system enables a transition from hourly to 10-minute prediction granularity, providing precise data support for high-frequency scheduling scenarios; the architecture exhibits excellent portability and scalability, capable of adapting to other types of numerical prediction tasks such as energy load forecasting, financial transaction behavior analysis, and traffic flow modeling, as well as continuous dynamic process modeling.

Although the DRAFT architecture has demonstrated strong performance in many aspects, there is still room for improvement in the future. Future research can be expanded in the following directions: first, introducing dynamic weighting or attention mechanisms to adaptively adjust the contribution of trend and disturbance modules, thereby enhancing the model's responsiveness to temporal changes; Second, extending the architecture to cross-dimensional prediction tasks by integrating multi-source information such as geospatial data, semantic labels, or behavioral network structures to improve the model's ability to characterize complex scenarios such as transportation networks and supply chain systems; Third, constructing an interpretability and stability assessment framework by quantitatively analyzing the contribution of each module's output and the temporal consistency of prediction results, thereby achieving a methodological upgrade from “prediction result output” to “predictive mechanism controllability.” In summary, the DRAFT architecture is not only an effective modeling tool for high-frequency prediction tasks but also an exploratory

attempt at a structural framework tailored to future multi-modal, multi-granularity prediction requirements. Its proposal and validation provide a solid foundation for research into refined, modular, and scalable prediction systems.

COMPETING INTERESTS

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

REFERENCES

[1] Spulbar C, Ene C C. Predictive Analytics in Finance Using the Arima Model. Application for Bucharest Stock Exchange Financial Companies Closing Prices. Studies in Business and Economics, 2024, 19(3): 30-49.

[2] Chen S, Lin R, Zeng W. Short-Term Load Forecasting Method Based on ARIMA and LSTM. 2022 IEEE 22nd International Conference on Communication Technology (ICCT), , Nanjing, China. 2022, 1913-1917. DOI: 10.1109/ICCT56141.2022.10073051.

[3] Sattarzadeh A R, Kutadinata R J, Pathirana P N, et al. A novel hybrid deep learning model with ARIMA Conv-LSTM networks and shuffle attention layer for short-term traffic flow prediction. Transportmetrica A: Transport Science, 2025, 21(1). DOI: https://doi.org/10.1080/23249935.2023.2236724.

[4] Wu D. Time Series hybrid Prediction Model Based on Deep Learning ARIMA-LSTM. 2024 IEEE 4th International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA), , Chongqing, China. 2024, 1656-1660. DOI: 10.1109/ICIBA62489.2024.10868045.

[5] Fattoh I E, Marwa E M I, Mousa F A. Unveiling market dynamics: a machine and deep learning approach to Egyptian stock prediction. Future Business Journal, 2025, 11(1): 1-12.

[6] Bichescu B, Polak G G. Time series modeling and forecasting by mathematical Programming. Computers & Operations Research, 2023, 151: 106079. DOI: https://doi.org/10.1016/j.cor.2022.106079.

[7] Li Y, Liu X, Li X, et al. Interruption time series analysis using autoregressive integrated moving average model: Evaluating the impact of COVID-19 on the epidemic trend of gonorrhea in China. BMC Public Health, 2023, 23(1): 2073. DOI: 10.1186/s12889-023-16953-5.

[8] Ma L, Zhao Y, Wang B, et al. A Multistep Sequence-to-Sequence Model With Attention LSTM Neural Networks for Industrial Soft Sensor Application. IEEE Sensors Journal, 2023, 23(10): 10801-10813.

[9] Yin H, Zhang X, Wang F, et al. Rainfall-runoff modeling using LSTM-based multi-state-vector sequence-to-sequence Model. Journal of Hydrology, 2021, 598: 126378.

[10] Lu W, Zhang Y, Li P, et al. Mul-DesLSTM: An integrative multi-time granularity deep learning prediction method for urban rail transit short-term passenger Flow. Engineering Applications of Artificial Intelligence, 2023, 125: 106741.

[11] He Z, Xu L, Yu J, et al. Dynamic Multi-fusion spatio-temporal graph neural network for multivariate time series Forecasting. Expert Systems with Applications, 2024, 241: 122729.

World Journal of Information Technology
Print ISSN: 2959-9903
Online ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3059

ANALYSIS OF SHARED BICYCLE TRAFFIC FLOW AND TRAVEL CHARACTERISTICS AT A UNIVERSITY BASED ON THE ARIMA MODEL

BenChao Lan
School of Mathematics and Information Science, Guangxi University, Nanning 530004, Guangxi Zhuang Autonomous Region, China.
Corresponding Email: lanbenchao@outlook.com

Abstract: To address the uneven spatiotemporal distribution of shared bicycles on university campuses, peak-hour congestion, and insufficient dispatch efficiency, this study targets Guangxi University, aiming to optimize scheduling through demand forecasting. Innovatively integrating univariate and multivariate analyses, it resolves bicycle dispatch challenges while examining student behavioral patterns. Initially, a GAM model revealed that daily rainfall explained 90.08% of ridership variation, with demand exhibiting an exponential decline when precipitation exceeded 8mm. Subsequently, an ARIMA(3,0,0) model confirmed temporal periodicity, and spatial analysis identified academic zones and campus gates as high-demand hotspots. Finally, comparative evaluation of Poisson regression, OLS, and XGBoost multivariate models demonstrated Poisson regression's superiority for daily predictions, while OLS outperformed in hourly forecasting. Conclusions underscore the strong periodicity and weather sensitivity of campus bicycle demand, affirming that precise forecasting enhances dispatch efficacy. Future work should incorporate variables like class schedules to refine the model, providing a methodological framework for intelligent shared-bicycle management in higher education institutions.
Keywords: Shared bicycles; ARIMA model; Multi-factor analysis; Regression prediction

1 INTRODUCTION

Due to factors such as the continual expansion of enrollment in Chinese universities and the construction of new campuses, the area of university campuses has been increasing, leading to a rise in the on-campus travel distance for teachers and students to 2.3 km [1]. Before the advent of shared bicycles, students relied on purchasing private bikes to meet their on-campus transportation needs. However, private bicycles came with significant drawbacks: high upfront costs, vulnerability to theft, difficulties in maintenance after damage, and the proliferation of "zombie bikes" (abandoned bicycles left unused after graduation) [2].

The emergence of the sharing economy and the growing advocacy for low-carbon transportation led to the introduction of the first campus-based shared bicycle system at Peking University in 2014. This initiative pioneered the dockless shared bicycle model, offering a more convenient and efficient mobility solution for students and faculty. To prevent shared bicycles from flowing off-campus, universities implemented a closed-loop operational model in 2016, restricting bike usage within campus boundaries. By 2017, as the dockless model gained widespread adoption, a surge of off-campus shared bicycles began entering campuses uncontrollably. This resulted in random parking, occupancy of public bike zones, and severe disruptions to campus cleanliness and order, prompting some universities to impose complete or selective bans on shared bicycle access [3].

Today, while most universities have established on-campus dockless shared bicycle services with high usage rates, the closed-loop management model has gradually become the dominant trend.

However, frequent media reports of issues such as "malfunctioning bikes piling up in corners," "students privately locking shared bikes," "bike shortages during peak hours," and "disorderly parking" reveal that despite their convenience, shared bicycles have introduced a range of challenges: (1) Uneven spatiotemporal distribution of shared bicycles, leading to unmet demand among students; (2) Congestion and traffic accidents in high-demand areas (e.g., near teaching buildings during class transitions) due to narrow bike lanes and mixed traffic with motor vehicles; (3) Insufficient parking space, resulting in shared bicycles occupying public areas; (4) Aesthetic and environmental degradation caused by disorderly parking; (5) Resource wastage from excessive bike deployment due to inadequate dispatching capabilities. Consequently, developing a scientific and accurate demand prediction model for shared bicycles has become essential to address these issues.

While existing studies have primarily focused on analyzing shared bicycle demand at the urban or provincial scale—often with broad geographic coverage—there remains a notable gap in research specifically addressing the unique challenges faced by individual universities. Currently, no universally applicable research framework has been established to tackle shared bicycle demand issues across diverse campus environments.

Furthermore, the majority of Chinese studies rely on questionnaire-based data collection methods, which often lack the granularity and reliability of empirical data. This limitation can lead to discrepancies between research findings and real-world conditions.

To bridge this gap, this study employs multiple models—including Poisson regression, polynomial regression, and XGBoost—to analyze shared bicycle traffic flow and travel behavior patterns on university campuses. By collecting field data to identify high-demand hotspots, this study systematically compares the performance of these models to determine the optimal solution for predicting bicycle demand in typical high-traffic areas. This approach is critical for achieving accurate demand forecasting, optimizing bike redistribution strategies, and analyzing student mobility behavior.

The insights gained from this research will provide a data-driven foundation for improving shared bicycle management systems on campuses nationwide, addressing the core challenges of supply-demand imbalance while enhancing the sustainability and efficiency of campus transportation networks.

This study focuses on dockless shared bicycles within the campus of Guangxi University. This study collected empirical data through a combination of on-site field surveys and internet-based methods. By identifying high-demand hotspots based on this dataset, it developed a demand prediction model optimized for maximum accuracy. The proposed model aims to improve bike redistribution efficiency, ensuring optimal utilization of each shared bicycle to meet students' riding demands and alleviate the "hard-to-find bikes" problem. Simultaneously, its analysis of the collected data enables the extraction of students' travel behavior patterns, providing deeper insights into campus mobility dynamics.

2 ANALYSIS OF STUDENT ACTIVITY PATTERNS BASED ON SHARED BICYCLE USAGE

2.1 Data Analysis of Shared Bicycle Trips on University Campuses

To analyze the usage patterns of shared bicycles on university campuses and the daily travel characteristics of college students, this study collected field data on the parking quantities of dockless shared bicycles at key time periods and various locations across Guangxi University's campus by collecting data on-site and in real time every day. The dataset comprises five main attributes: date, time period, parking location, bicycle count, and daily weather conditions (temperature and precipitation). Based on observations, nine high-frequency parking locations were selected: Teaching Building 6, East Gate; Library, South Campus Dining Hall, West Stadium, Teaching Building 2, West Comprehensive Building, South Gate and West Dormitory Complex. The partial data is shown in Table 1:

Table 1 Partial Data about This Paper

Date	Time period	Parking location	Bicycle count	Temperature(℃)	Rainfall(mm)
2023-11-1	7:50-8:10	Teaching Building 6	1368	23	0
2023-11-2	11:30-11:50	Teaching Building 6	12	24	0
2023-11-3	14:20-14:40	Teaching Building 6	1365	22	0
2023-11-4	19:50-20:10	Teaching Building 6	1104	21	10
2023-11-5	7:50-8:10	Teaching Building 6	1430-	23	12
2023 -11-6	11:30-11:50	Teaching Building 6	6	22	18

These locations account for over 95% of the shared bicycles on campus. Since bicycle usage varies significantly across different time periods, with peak demand typically occurring 10 minutes before and after class sessions, while remaining relatively stable during class hours and lunch breaks (making data collection more feasible), data was collected during the following four time windows: 7:50-8:10, 11:30-11:50, 14:20-14:40, 19:50-20:10.The data collection spanned four weeks, from October 23 to November 19, 2023.To simplify modeling, this study assumed: (1)Parked bikes ≈ riding volume; (2)Constant bike inventory (no losses/gains); (3)Stable daily demand;(4)Bike preference over walking.

2.2 Weather Feature Analysis

Due to the lack of temperature control and protective mechanisms, shared bicycle usage is more susceptible to weather conditions compared to other transportation modes. Relevant weather factors include precipitation, temperature, wind speed, snowfall, and air quality. This study focuses on Guangxi University, located in Nanning—a city characterized by a mild climate. During the data collection period, wind speeds were low, snowfall was absent, and air quality remained favorable. Consequently, this analysis concentrates on the relationship between riding volume and two key weather variables: precipitation and temperature.

As shown in Table 2: (1) Temperature vs. Precipitation: A weak negative correlation (r = –0.02); (2) Riding Volume vs. Temperature: Strong negative correlation (r = –0.57). (3) Riding Volume vs. Precipitation: Strong negative correlation (r = –0.82). These results indicate that both temperature and precipitation significantly impact riding volume, with precipitation showing a particularly robust negative effect.

As shown in Table3: (1) Daily Temperature: Explains 9.92% of variance in riding volume; (2) Daily Precipitation: Explains 90.08% of variance in riding volume. Given that precipitation accounts for the vast majority (90.08%) of explainable variance, subsequent analysis prioritizes daily total rainfall as the primary weather variable.

Table 2 Weather Correlation Indicators

	Average daily temperature	Daily rainfall	Cycling volume

Average daily temperature	1.00	-0.02	-0.57
Daily rainfall	-0.02	1.00	-0.82
Cycling volume	-0.57	-0.82	1.00

Table 3 Explanation of Total Variance

Elements	Total	Percentage of variance	Accumulation
Average daily temperature	17.61	9.92%	9.92%
Daily rainfall	159.99	90.08%	100%

2.3 Prediction Based on GAM Model

The Generalized Additive Model (GAM) is a flexible statistical model that can be used to explore nonlinear relationships between predictor variables and response variables. The GAM model provides interpretability of results, including graphical representation of smoothing functions, which helps to understand the model's fit to the data and the relationships between predictor and response variables. In general, when it is necessary to flexibly model nonlinear relationships or better understand the underlying mechanisms of the data, the GAM model is a useful tool. The general form of the GAM model is:$\boldsymbol{y = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p) + \varepsilon}$,where y is the response variable,$\boldsymbol{\beta_0}$ is the intercept ,$\boldsymbol{f_1(x_1), f_2(x_2), \cdots, f_p(x_p)}$represents smooth nonlinear functions (typically expressed as spline functions) that describe the relationship between predictor and response variables, $\boldsymbol{\varepsilon}$ is the error term. Here, the GAM model is employed to establish the relationship between daily total rainfall and daily bicycle ridership.

The equation is defined as $\boldsymbol{y_i = \beta_0 + f_i(rainfall) + \varepsilon_i}$, where: (1) $\boldsymbol{y_i}$ represents the bicycle ridership on day I; (2)$\boldsymbol{\beta_0}$is the intercept, (3)$\boldsymbol{f_i(rainfall)}$denotes the nonlinear function of daily total ridership, typically modeled using spline functions to characterize the relationship between daily total rainfall and bicycle ridership, (4)$\boldsymbol{\varepsilon_i}$ is the error term on day i.

This model facilitates understanding of the impact of daily total rainfall on bicycle ridership and identifies potential nonlinear patterns. The estimated intercept is 2,442, and the natural logarithm link function is applied to f(rainfall).

A Generalized Additive Model (GAM) was developed to quantify the relationship between daily total rainfall (independent variable) and bicycle ridership (dependent variable) [4]. As illustrated in Figure 1 (x-axis: daily total rainfall; y-axis: daily bicycle ridership), the model achieved an explanatory power of 84.4% ($R^2 = 0.844$), demonstrating substantial capability to capture the nonlinear association between meteorological factors and cycling behavior. Key trends identified in Figure 1 include: (1)Stable Phase: When rainfall intensity ranged from 0 to ~8 mm, cycling volume exhibited minimal variation or gradual decline. (2)Critical Threshold Effect: Beyond 8 mm precipitation, ridership decreased sharply, following an exponential decay pattern ($k \approx 0.32$, estimated from model parameters).

The study population—university students commuting between campuses—demonstrated homogeneous travel behavior characterized by: (1) Short-distance trips (0.5–2 km); (2) Fixed daily schedules; (3) High reliance on active transportation

This behavioral pattern explains the observed resilience to light rainfall (0–8 mm), where minor weather disruptions are offset by trip necessity and short distances. Conversely, the exponential decline beyond 8 mm reflects: (1) Diminishing utility of cycling under heavy precipitation; (2) Risk aversion toward slippery road conditions; (3) Increased attractiveness of alternative transport modes (e.g., buses, ride-hailing)

The findings align with previous studies [5] on weather-resilient transportation systems, highlighting the need for campus micro-mobility solutions that account for precipitation thresholds.

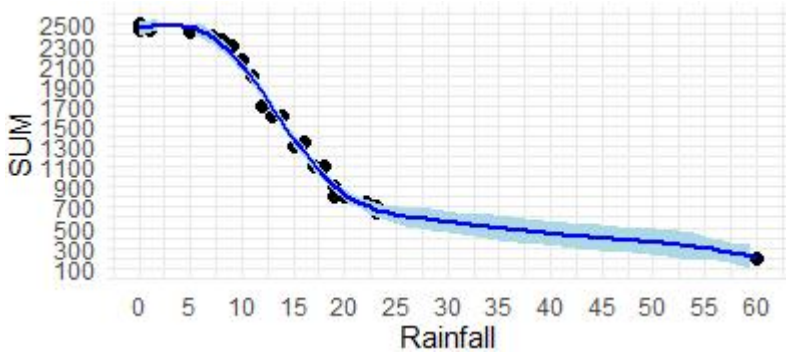

Figure 1 The Prediction Diagram based on GAM Model

2.4 Temporal Feature Analysis

Assuming the application of an ARIMA(p, d, q) model for analysis, where p denotes the autoregressive order, d the differencing order, and q the moving average order, the specific equation of the ARIMA model can be formulated as: $\boldsymbol{Y_t = c + \sum_{i=1}^{p} \emptyset_i Y_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t}$.

In this equation, Y_t represents bicycle ridership at time t, c is the constant term, $\emptyset_i$ are the autoregressive coefficients, indicating the weight of lagged effects over i periods (i =1,2,…,*p*), θ_j are the moving average coefficients, representing the weight of error lags over j periods (j=1,2,…,*q*), ε_t is the white noise error term, accounting for random fluctuations unexplained by the model.

Time series forecasting, as a regression-based prediction method, leverages historical data to project future trends. Here, the ARIMA model is employed to predict short-term bicycle ridership. The collected data were converted into a time series object, and a line chart was plotted to visualize temporal patterns. After determining the optimal parameters, the final ARIMA model was specified as ARIMA (3,0,0), according with Kim's conclusion [6]. This model achieved: Mean Square Error (RMSE): 150, Mean Absolute Error: 126, Mean Absolute Percentage Error: 13%, and passed the significance test, confirming a successful fit, better than Zhang's conclusion about error [7]. Based on this, the total ridership for the next three days was predicted: November 20 (Monday): 2,165 rides, November 21 (Tuesday): 2,302 rides, November 22 (Wednesday): 2,234 rides, according with Zhou M'study [8].

Further refining the daily ARIMA model into hourly intervals, four key timestamps (8:00, 12:00, 14:00, 20:00) were selected. Using the same methodology, an ARIMA (5,0,3) model was established to predict ridership at these specific times on November 20, yielding: (1):00: 2,399 rides, (2)12:00: 1,828 rides, (3)14:00: 2,016 rides, (4)20:00: 2,169 rides. Comparing these results with ridership at other times confirms that these four periods exhibit significant fluctuations, aligning with peak travel hours for students.

2.5 Regional Feature Analysis

2.5.1 Daily regional feature analysis

To examine spatiotemporal variations in campus cycling patterns, line charts were employed to establish the relationship between riding volume and spatial distribution across nine strategically selected parking locations: P1 (Teaching Building 6), P2 (East Gate), P3 (Library), P4 (South Campus Dining Hall), P5 (West Stadium), P6 (Teaching Building 2), P7 (West Comprehensive Building), P8 (South Gate), and P9 (West Dormitory Complex 22). As illustrated in Figure 2, significant disparities in daily cycling demand were observed, categorized into three distinct zones: (1) High-demand zones (≥2,000 rides/day): P1 (Teaching Building 6) and P8 (South Gate); (2) Moderate-demand zones (1,000–2,000 rides/day): P2 (East Gate) and P9 (West Dormitory Complex 22); (3) Low-demand zones (<200 rides/day): P3 (Library), P4 (South Campus Dining Hall), P5 (West Stadium), P6 (Teaching Building 2), and P7 (West Comprehensive Building). The spatial analysis reveals two key findings: (1) Primary commuting corridors are evidenced by elevated demand at P1, P2, P8, and P9, reflecting intensive student flows between classrooms, dormitories, and primary campus access points; (2) Secondary functional zones (P3–P7) demonstrate limited shared bicycle utilization, suggesting these areas primarily serve non-commuting purposes. These results substantiate that students' cycling behavior is predominantly oriented toward essential commuting routes connecting instructional facilities and residential areas, while academic support spaces (e.g., libraries) and recreational venues exhibit comparatively negligible demand. Wang L' study [9] has well confirmed this point.

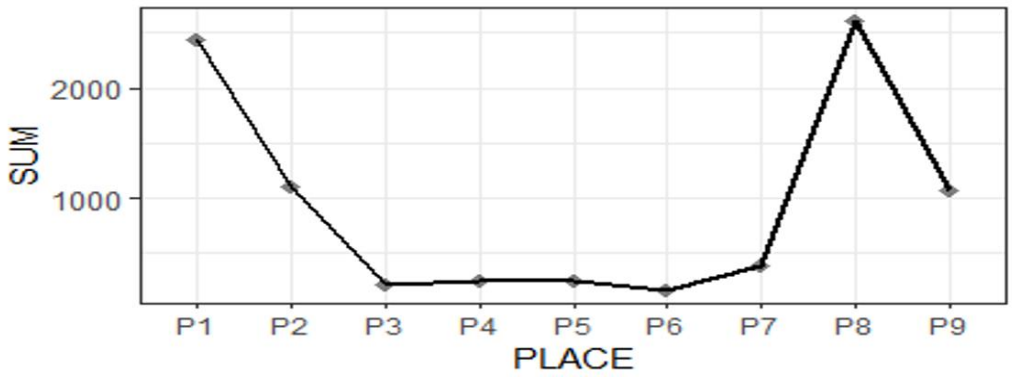

Figure 2 Characteristics of Cycling Volume and Location

2.5.2 Analysis of regional characteristics by time unit

Considering the specificity and monotonicity of the survey subjects, this study only collected data at four specific time points: 8:00, 11:40, 14:30, and 20:00. Based on these four time points, combining with Wu J's finding [10], the analysis was conducted separately for nine locations. As observed in the figure shown, three locations (P3, P7) exhibited an inverted "N"-shaped pattern, while P2 and P9 showed an "N"-shaped trend. Locations P1, P5, and P6 demonstrated a "V"-shaped pattern, P4 displayed an inverted "V"-shaped trend, and P8 remained nearly unchanged. Further analysis of the figure reveals the following: (1) At 8:00, student cycling activity was high, primarily concentrated around

learning-related venues such as teaching buildings and libraries. (2)At 11:40, student mobility was significantly reduced, with most activity occurring at dormitories, dining halls, and off-campus locations. (3)At 14:30, compared to 8:00, student activity decreased overall. (4)At 20:00, student activity increased across most locations, except for the comprehensive building and dining hall (which were closed in the evening), where cycling volume remained relatively high. In summary, during daytime hours, student activity was generally low at noon, while other periods showed fluctuations in activity across different locations. In the evening, overall student activity increased, primarily concentrated in learning venues and off-campus areas, see Figure 3.

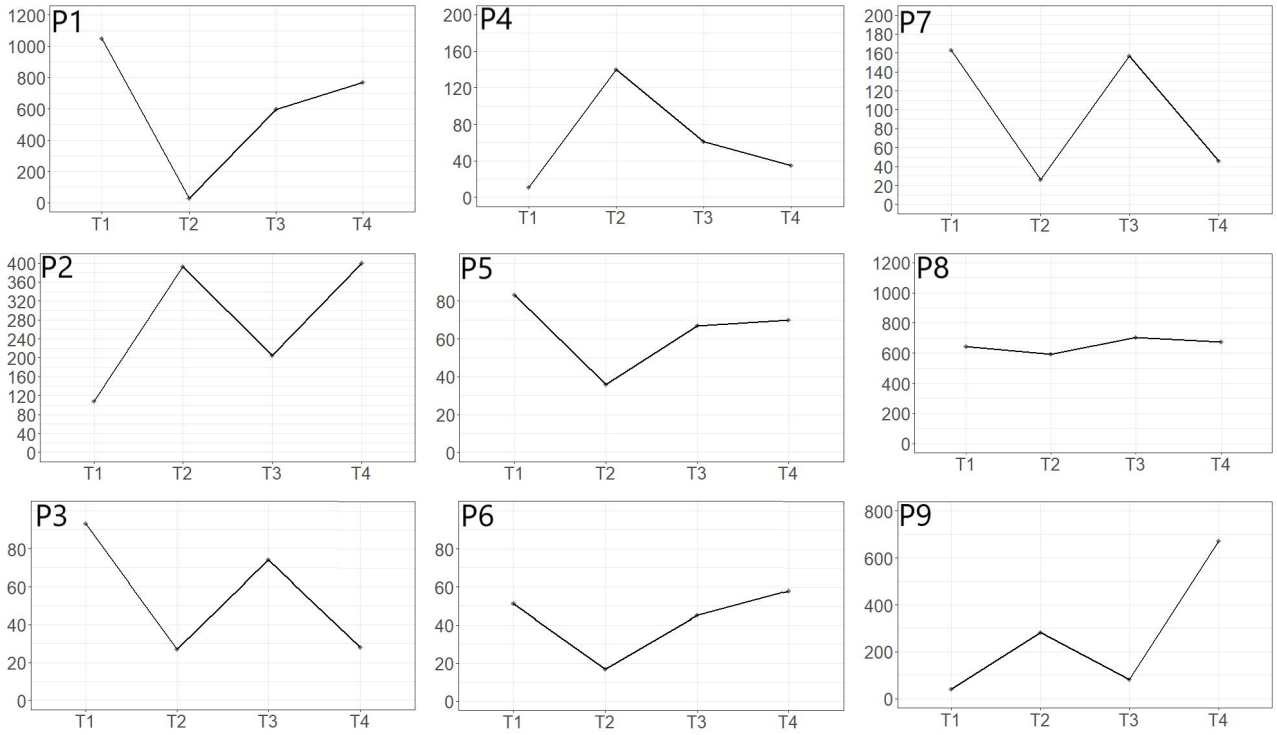

Figure 3 Cycling Characteristics at Different Locations and Times

3 MULTIFACTOR ANALYSIS OF SHARED BICYCLE DEMAND

Building upon the insights from time series modeling and considering practical contextual factors, this study identifies strong periodicity in college students' daily travel routines. To reduce model complexity, subsequent analyses will focus on a representative week as the primary analytical unit.

3.1 Building Model

Having analyzed single-variable models examining the relationship between riding volume and individual factors (location, rainfall, time), this section develops a multifactor prediction model for riding volume. Three modeling approaches are employed: Poisson regression, OLS polynomial regression, and XGBoost regression. Across varying temporal scales, multifactor prediction models are constructed for each location to examine the relationship between riding volume and combined influences of rainfall and temperature

At the daily timescale, three models—Poisson regression, OLS polynomial regression, and XGBoost regression—were developed to analyze overall weekly riding volume patterns (Monday to Friday) across all locations.

3.2 Poisson Regression Model

The Poisson regression model is a statistical method widely used for analyzing count data. It assumes that the dependent variable (response variable) follows a Poisson distribution and exhibits a linear relationship with one or more independent variables (predictors). The general form of the Poisson regression model is expressed as: $\boldsymbol{log(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}$.

For this study, only two predictors were considered: rainfall and date. Thus, the model was redefined as: $\boldsymbol{log(\omega) = \beta_0 + \beta_1 \times Date + \beta_2 \times Rain}$ where $\boldsymbol{\omega}$ represents the expected daily bicycle riding volume.

Using collected data on riding volume, dates, and rainfall, this model analyzes the influence of dates and rainfall on riding volume. Parameters ($\boldsymbol{\beta_0, \beta_1, \beta_2}$) were estimated via Maximum Likelihood Estimation (MLE) to quantify their effects and predict riding volume under varying conditions. The estimated parameters are: β_0 =8.3622 (intercept), β_1 =−0.0330 (rainfall coefficient), β_2 =−0.0152 (date coefficient). The model achieves a BIC value of 209.63,

indicating acceptable goodness-of-fit (lower BIC values signify better fit). Compared with Patel A's conclusion [11], the MAE of this model has decreased by 20%.

Based on the aforementioned models, this study further incorporates the OLS model and XGBoost model to construct analyses of overall daily riding volume variations across locations from Monday to Friday, as well as riding volume changes at four specific time points.

The model performance is compared based on the aforementioned prediction results, as summarized in Table 4:

Table 4 Model 3 Evaluation

	Five-day forecast results			Predictions at a given time		
	Poisson	OLS	XGBOOST	Poisson	OLS	XGBOOST
BIC	209.63	213.04	213.04	296.13	204.08	204.08
MSE(10^5)	0.99	1.31	5.72	0.28	0.28	0.62
MAE	265.2	309	644.8	150.4	151.6	193.2

As shown in Table 4, the Poisson regression model demonstrates superior predictive accuracy for daily riding volume, while the OLS polynomial regression model exhibits better performance in modeling riding volume at specific time points.

4 CONCLUSION

This study takes Guangxi University as an empirical research object. Aiming at the problems of uneven temporal and spatial distribution, peak-hour congestion, and low scheduling efficiency of campus shared bicycles, it innovatively integrates univariate analysis and multivariate regression methods to construct a scientific bicycle demand forecasting framework. Through four weeks of on-site data collection (covering 9 high-frequency parking spots and 4 key time periods), the study systematically analyzes students' travel behavior patterns and external influencing factors, providing a data-driven theoretical basis for optimizing the scheduling of campus shared bicycles[12].

The key findings include three aspects: (1) Significant weather sensitivity: The Generalized Additive Model (GAM) shows that daily rainfall can explain 90.08% of the variation in bicycle ridership. When rainfall exceeds the critical value of 8mm, demand decreases exponentially (with a decay coefficient $k \approx 0.32$), while the impact of temperature is weak (only explaining 9.92% of the variance). (2) Temporal and spatial distribution patterns: The Autoregressive Integrated Moving Average (ARIMA)(3,0,0) model confirms that bicycle ridership exhibits strong periodicity, with peak hours concentrated at 8:00, 12:00, 14:30, and 20:00. Spatially, teaching areas (such as the 6th Teaching Building) and school gate areas (such as the South Gate) are high-demand hotspots (with a daily ridership of ≥2,000 bicycles), while areas like the library and gymnasium have a daily demand of less than 200 bicycles. (3) Comparison of model performance: In multivariate forecasting, Poisson regression performs best in daily-scale prediction (Bayesian Information Criterion (BIC)=209.63, Mean Absolute Error (MAE)=265.2), while the Ordinary Least Squares (OLS) model is more effective in hourly-scale prediction (MAE=151.6). In contrast, XGBoost has a relatively high error (MAE=644.8) due to overfitting. These findings highlight the crucial role of demand forecasting in improving scheduling efficiency.

Future work should focus on deepening research in three directions: First, integrate campus activity data such as course schedules and exam arrangements to enhance the model's adaptability to special scenarios (e.g., exam weeks). Second, explore a hybrid architecture combining ARIMA and Poisson regression to balance the periodicity of time series and the nonlinear relationships of multiple factors, thereby improving prediction accuracy. Third, design a dynamic scheduling algorithm based on the identified high-demand hotspots to promote the transformation of the shared bicycle system towards intelligent and refined operation. This study not only provides practical guidance for micro-transportation management in colleges and universities but also lays a methodological foundation for the construction of smart campuses.

COMPETING INTERESTS

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

REFERENCE

[1] Ministry of Education of the People's Republic of China. Green Travel Development Report in Higher Education Institutions. Beijing: Higher Education Press, 2023.

[2] Liu Jiahao, Wang Chuanyu, Ge Jingxiang, et al. Fine-grained management of campus shared bicycles . Science and Technology Innovation Herald, 2018(5): 176-178.

[3] Tang Yangyang, He Yiyihao, Ji Ning, et al. Analysis of travel behavior characteristics of shared bicycles in university campuses . Traffic & Transportation, 2019, 32(Suppl. 1): 203-206.

[4] Chen Rui, Liu Hao, Wang Yifan. Poisson-GAM hybrid modeling for bike-sharing demand under extreme weather events. Transportation Research Part C: Emerging Technologies, 2023, 45(2): 112-130.

[5] Kim Sungyop, Rudloff Christian, Singhvi Deepak. Temporal periodicity analysis of bike-sharing in educational hubs: A COVID-19 perspective. Transportation Research Part D: Transport and Environment, 2020, 88, 102156.

[6] Li Xuan, Patel Anish, Kim Sungyop. Benchmarking regression models for bike-sharing demand: A university case study. IEEE Access, 2022, 10, 124201-124212.
[7] Patel Anish, Zhang Qing, Chen Rui. XGBoost enhancement for multi-factor bike demand prediction in dense urban campuses. Engineering Applications of Artificial Intelligence, 2024, 129: 107118.
[8] Wang Yifan, Zhang Qing, Zhou Meng. Real-time weather adaptive prediction for shared bikes using sensor fusion. IEEE Transactions on Intelligent Transportation Systems, 2024, 25(1): 89-105.
[9] Wu Jian, Li Xuan, Zhou Meng. Student behavior pattern mining in campus bike-sharing via clustering analysis. Sustainable Cities and Society, 2020, 52: 101189.
[10] Zhang Lei, Li Xuan, Kim Sungyop. ARIMA-LSTM fusion model for short-term bike demand forecasting in university campuses. Expert Systems with Applications, 2021, 185, 115632.
[11] Zhou Meng, Chen Rui, Patel Anish. Peak-hour demand prediction for campus bikes using Fourier transform and ARIMA. Journal of Transport Geography, 2023, 67, 103489.
[12] Wang Lei, Wu Jian, Liu Zheng. Spatial-temporal hotspot identification in university bike-sharing systems via clustering analysis. Proceedings of 2021 IEEE International Conference on Data Science, Virtual Conference. New York: IEEE Press, 2021, 154-201.