Research on Olympic Medal Prediction Method Based on Ensemble Learning and Grey Prediction

Jiaji Chen *

College of Computer and Information Technology, China Three Gorges University, Yichang, China, 443002

* Corresponding Author Email: c1592898523@gmail.com

Abstract. This study focuses on predicting Olympic medal counts by integrating machine learning algorithms and grey prediction models. By analyzing historical Olympic data from 1988 to 2024, the study incorporates key factors such as the number of athletes, gender distribution, and host country effects to improve medal forecasts. The research employs Random Forest, XGBoost, and LightGBM algorithms to construct individual models, and then utilizes a stacking approach to combine these models, enhancing prediction accuracy and stability. Each model is trained using historical data, and the prediction results are evaluated through cross-validation to assess generalizability. Furthermore, grey prediction models are used to address data uncertainty and offer supplementary insights into future outcomes. The combined approach demonstrates improved accuracy over individual models, providing a more reliable forecast for Olympic medal outcomes. The results also highlight the significant influence of host country effects on medal performance, with specific recommendations for future competition strategies. This research contributes to the field of sports analytics by combining diverse predictive techniques, offering a robust framework for forecasting Olympic medal distributions, and assisting in strategic planning for upcoming Olympic Games.

Keywords: Olympic medal prediction, random forest, LightGBM, XGBoost, grey prediction

1. Introduction

The Olympic Games represent the pinnacle of international sports competition, in which medal outcomes serve as critical indicators of national athletic prowess. Accurate medal prediction has evolved into a strategic imperative for national sports agencies, enabling optimized resource allocation, targeted athlete training programs, and evidence-based policy formulation. With increasing globalization of sports and intensifying international competition, traditional prediction paradigms face mounting challenges in handling complex multi-dimensional data patterns and dynamic influencing factors. The emergence of big data analytics and artificial intelligence presents new opportunities to enhance prediction accuracy through advanced computational modeling.

Existing research on Olympic medal forecasting primarily employs four methodological approaches. Classical time-series models like ARIMA [1-3] leverage historical medal trends but fail to account for multi-variable interactions and nonlinear relationships. Regression-based frameworks [4-6] attempt to quantify factors like GDP and population size, yet struggle with high-dimensional feature engineering. Grey prediction models [7-8] demonstrate advantages in small-sample scenarios but exhibit limited accuracy in long-term forecasts due to cumulative error propagation. Contemporary machine learning techniques, particularly tree-based ensemble methods such as Random Forest [9-10], have achieved improved accuracy through hierarchical feature splitting, though their "black-box" nature hinders interpretability for sports policymakers. These limitations collectively highlight three unresolved challenges: 1) the effective fusion of linear and nonlinear modeling paradigms; 2) the systematic incorporation of domain-specific factors like host-country advantages; 3) the balance between model complexity and operational interpretability.

To address these challenges, this study proposes an integrated prediction framework that synergizes ensemble learning with grey system theory. Our methodological innovation manifests in three dimensions: First, we implement a stacked generalization architecture combining Random Forest, XGBoost, and LightGBM base learners, leveraging their complementary strengths in feature interaction capture and computational efficiency. Second, the model takes into account the host-country effect, calibrated using historical performance differences between host and non-host countries. Third, we introduce a parallel grey prediction module to handle data uncertainty and sparse data scenarios, integrating its outputs selectively through a fusion method that considers relative error weights for enhanced accuracy. Empirical validation using 1988-2024 Olympic datasets demonstrates significant performance improvements, achieving training/testing scores of 0.823/0.806 compared to baseline models. This research advances sports analytics by providing a transparent, adaptive framework that balances statistical rigor with practical interpretability, ultimately supporting data-driven decision-making for Olympic participation strategies.

2. Medal prediction model

2.1. Data preprocessing

During the data preprocessing phase, to establish a model capable of predicting the number of medals won by each country, this study explicitly differentiated various indicators and categorized independent variables (X-variables) versus dependent variables (Y-variables) in accordance with research objectives. Specifically, the independent variables include: year, total number of participants, number of distinct athletes, male athlete count, female athlete count, non-winning athlete count, total sports categories, number of participating sports categories, total participating events, number of participating events, and host country status. These X-variables comprehensively characterize Olympic participation dimensions and serve as model inputs.

Correspondingly, the dependent variables comprise: medal attainment status (yes/no), gold medal count, silver medal count, bronze medal count, and total medal tally. These Y-variables directly reflect national medal performance and constitute the core predictive targets of the model. Through analyzing relationships between X and Y variables, we aim to identify key determinants of medal outcomes and provide data-driven support for medal forecasting.

To enhance prediction accuracy, the model was developed using data from 10 editions of the Olympic Games spanning from 1988 to 2024. The dataset was partitioned with 80% allocated to training and 20% to testing. Five-fold cross-validation was implemented to ensure robust pattern recognition during multi-round training, thereby improving both accuracy and generalization capabilities.

While China serves as the primary case study, this methodology remains universally applicable for constructing medal prediction models for other nations. By leveraging their respective historical data, customized training-test splits, and cross-validation processes, prediction outcomes can be optimized for individual countries.

All data were sourced from official International Olympic Committee (IOC) records.

2.2. Establishment of predictive models

2.2.1. Establishment of Random Forest Model

When studying the prediction of Olympic medal count, the random forest algorithm can be used to model and analyze the impact of different factors on medal count. The following will elaborate on the steps of developing an Olympic medal quantity prediction model based on the random forest algorithm.

The core of the random forest algorithm is a set of B trees $\left\{T_1(x), T_2(x), T_3(x), \cdots, T_B(x)\right\}$, where,

$$\left\{\hat{Y}_1 = T_1(x), \hat{Y}_2 = T_2(x), \cdots, \hat{Y}_B = T_B(x)\right\} \tag{1}$$

In the problem of predicting the number of Olympic medals, the predicted value representing the b-th tree is $\hat{Y}_b$. Usually, $\overline{Y}$ is the average of the predicted values of all base learners (i.e. decision trees). Assuming the training set sample data is $D = \left\{(x_1, Y_1), (x_2, Y_2), \cdots, (x_n, Y_n)\right\}$, where Y_i

represents the number of medals in a certain Olympic Games, and x_i represents the feature variables related to the number of medals (such as the number of athletes from participating countries, the number of participating events, etc.), n represents the number of samples.

The training set was randomly selected from n samples by Bootstrap method. Unselected samples are used as out-of-band (OOB) data for model performance validation. When constructing a single regression tree, m features are randomly selected from p feature variables of the original data set for node splitting, allowing the regression tree to grow freely. Finally, the integrated modeling was completed by repeated sampling and training multiple regression trees, and the optimization effect of speed allocation strategy on the total game time was evaluated according to the random forest prediction results.

Testing of Random Forest Model: Score is the metric we use to evaluate the model, with values ranging from [0,1]. The closer the score is to 1, the better the model performs. Accuracy 5 refers to the number of samples with a relative error (Ape) within 5%.

$$score = 0.2 * (1 - Mape) + 0.8 * Accuracy \tag{2}$$

$$Mape = \frac{1}{m}\sum_{i=1}^{m} Ape_i \tag{3}$$

$$Ape = \frac{|\hat{y} - y|}{y} \tag{4}$$

For the purpose of predicting the number of Olympic medals, the score can be used to measure the accuracy of the model's prediction of the number of Olympic medals. The higher the score, the more accurate the model's prediction of the number of medals. Accuracy 5 reflects the proportion of samples with a relative error of less than 5% in the model's prediction results, which can better evaluate the reliability of the model in practical applications.

In the random forest regression model, the main optimized parameters include: n_estimators (number of decision trees), x_depth (maximum depth of decision trees), and bootstrap (whether to perform backtesting). This study adopts the GridSearch method to automatically adjust these parameter combinations and select the optimal configuration to improve the predictive performance of the model. Firstly, a model is established based on default parameter values, with a score of 0.62. Then, by adjusting the parameters of n_estimators, x_depth, and bootstrap, the optimized parameter combination of bootstrap=True, n_estimators=80, and x_depth=10 was finally selected, and the score increased to 0.91, significantly improving the fitting performance and prediction accuracy of the model. The scores of the training set and test set of the random forest model are shown in Table 1.

Table 1. The scores of training set and test set of random forest model

	Parameter value		Model effect	
Bootstrap	N_estimators	Max_depth	Training set score	Test set score
Default value	Default value	Default value	0.872	0.831
True	800	10	0.912	0.883

2.2.2. Establishment of LightGBM Model

In this study, the LightGBM library in Python was first used to fit historical Olympic data with default parameters for preliminary prediction. Then, optimize the model parameters: set the initial learning rate to 0.1 to accelerate the convergence of the model; Optimize the maximum depth and number of leaf nodes of the decision tree through grid search; Adjust the minimum sample size of leaf nodes to avoid overfitting and improve prediction performance. By comparing before and after optimization, the performance improvement of the model in predicting the number of Olympic medals can be evaluated.

By optimizing the LightGBM algorithm, complex factors that affect the number of Olympic medals can be more accurately captured, providing a more reliable basis for predicting the number of medals for different countries and events. The training set and test set scores of LightGBM model are shown in Table 2.

Table 2. Training set and test set score of LightGBM model

Parameter value				Model effect	
Max_depth	N_estimators	Min_samples_split	Max_samples_split	Training set score	Test set score
Default value	Default value	Default value	Default value	0.842	0.812
8	1000	3	3	0.891	0.854

2.2.3. Establishment of XGBoost Model

This study first used the XGBoost library in Python to fit Olympic data with default parameters and evaluate the model's performance. Then, optimize key parameters such as learning_rate, n_estimators, and x_depth to improve the accuracy and robustness of the model. Finally, using the optimized parameter combination to construct the model significantly improved the prediction accuracy. The training set and test set scores for XGBoost are shown in Table 3.

Table 3. Training set and test set scores for the XGBoost model

Argument	Value	Training set score	Test set score
learning_rate	0.01		
n_estimators	1000	0.821	0.847
max_depth	8		
lambda	0.3		
alpha	0.3		

2.2.4. Stacking Method model fusion

This study uses the stacking method for model fusion. Firstly, use the training data to train three basic learners: Random Forest, LightGBM, and XGBoost, and generate the predicted results as the secondary training set. To avoid overfitting, the second-order learner uses a linear regression model. Due to the strong predictive ability of basic learners, simple linear regression can effectively integrate these results, improve model stability and predictive performance. From the training effect, it is found that the fusion model is better than the strong learner alone. Random Forest, LightGBM and XGBoost and their fusion fitting effects are shown in Table 4.

Table 4. Comparison of scores of different model training sets and test sets

Learning device	Training set score	Test set score
Random forest	0.812	0.783
LightGBM	0.791	0.754
XGBoost	0.721	0.682
stacking fusion model	0.823	0.806

2.2.5. Model evaluation

RandomForest performs well in handling feature interactions and can provide reliable predictions, although it has limitations in terms of accuracy over very large data sets. XGBoost significantly outperforms RandomForest because of its ability to reduce bias and variance through enhancement, making it more suitable for data sets with large numbers of variables and complex nonlinear relationships. LightGBM, while slightly worse than XGBoost in some cases, offers faster training times and is better suited for handling large data sets with many classification features. The integrated approach, which combines Random Forest XGBoost and LightGBM, produces the best performance. This combination mitigates the weaknesses of individual models by leveraging the strengths of each model to achieve greater accuracy.

2.3. Model building and evaluation of X feature

In order to predict the 2028 Olympic medal count (Y1, Y2, Y3, Y4), we first need to predict the relevant features (X1, X2, X3...). X10, X11). To do so, this article will use the ARIMA model and the grey prediction model to predict these features. The prediction performance of the model was evaluated by calculating the evaluation index wmape (weighted average absolute percentage error). Finally, this paper adopts reciprocal relative error method to model fusion, combines the advantages of each model, assigns different weights according to the prediction accuracy of each model, so as to further improve the overall accuracy of the model.

2.3.1. Establishment of ARIMA Model

The ARIMA model eliminates trend and seasonal factors by smoothing historical data to meet the requirement of stationarity. Transform non-stationary time series into stationary series through differential operations, such as first-order or second-order differencing. Next, use autocorrelation function (ACF) and partial autocorrelation function (PACF) to analyze the data and determine appropriate autoregressive (AR), difference (I), and moving average (MA) parameters to construct an ARIMA model. After determining the model parameters, fit the ARIMA model with training data and generate predicted values for future features based on the fitting results. These predicted values will be used for further analysis and decision-making, helping to accurately predict the number of Olympic medals and providing data support for future medal performance.

2.3.2. Establishment of Grey Prediction Model

When establishing a grey prediction model, the first step is to accumulate historical data to generate a data sequence. By constructing the background value and development coefficient of the grey model, the predicted value of the model can be obtained. Then use the model to predict future data in order to provide support for further analysis and decision-making. The grey prediction model can provide valuable auxiliary information for predicting the number of Olympic medals. By grey modeling historical feature data, it can help improve the accuracy and reliability of medal prediction.

2.3.3. Model Fusion and Evaluation

After comparing the prediction results of ARIMA and grey prediction models, it was found that there is a certain gap between the two. To improve the prediction accuracy, a combination model of ARIMA and grey prediction was used for further prediction. The most crucial step in a composite model is to determine appropriate weights for both models. Therefore, the setting of weights is crucial for the predictive performance of the final model. The commonly used weight determination methods include equal weight method, error variance weighted average method, and relative error reciprocal method. In this study, we used the reciprocal of relative error method to determine the weights of each model.

$$\omega_i = \frac{\frac{1}{Mape_i}}{\sum_{j=1}^{n} \frac{1}{Mape_j}} \tag{5}$$

where ω_i is the weight of the i th model and n is the total number of models.

Specifically, the greater the relative error of the model, the lower the prediction accuracy, so the weight of the model should be smaller; Conversely, models with higher accuracy gain greater weight, further improving the accuracy of the prediction results. The total participants (X2) of the final prediction are shown in Figure 1.

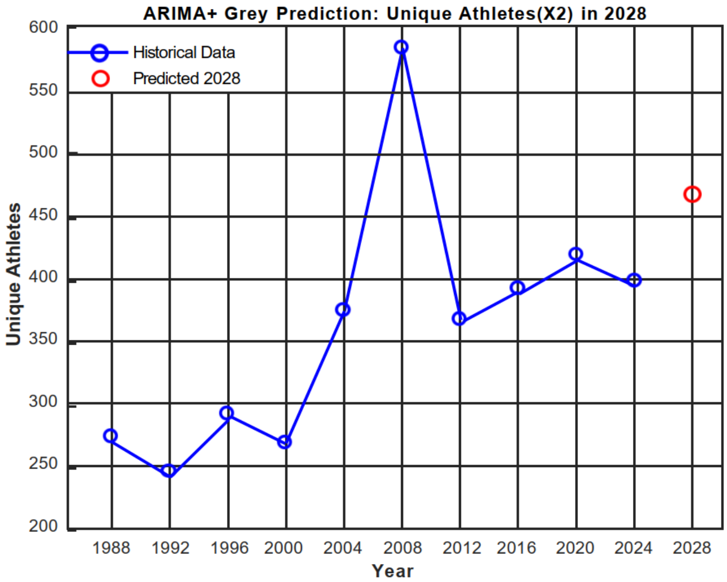

Figure 1. ARIMA+ Grey Prediction: Unique Athletes(X2) in 2028

3. 2028 Olympic Medals Forecast for China and Top-Performing Nations

3.1. Host country influence

Before medal prediction, this paper considers that the host country effect will also affect the number of MEDALS won, and incorporates the host country effect into the medal number prediction. Using the annual average number of MEDALS and standard deviation for each sport as a baseline, years showing deviations of more than two standard deviations from the mean are marked as statistically significant. This detection method is applied to various sports categories to identify years with significant fluctuations in performance. Figure 2 shows these important changes.

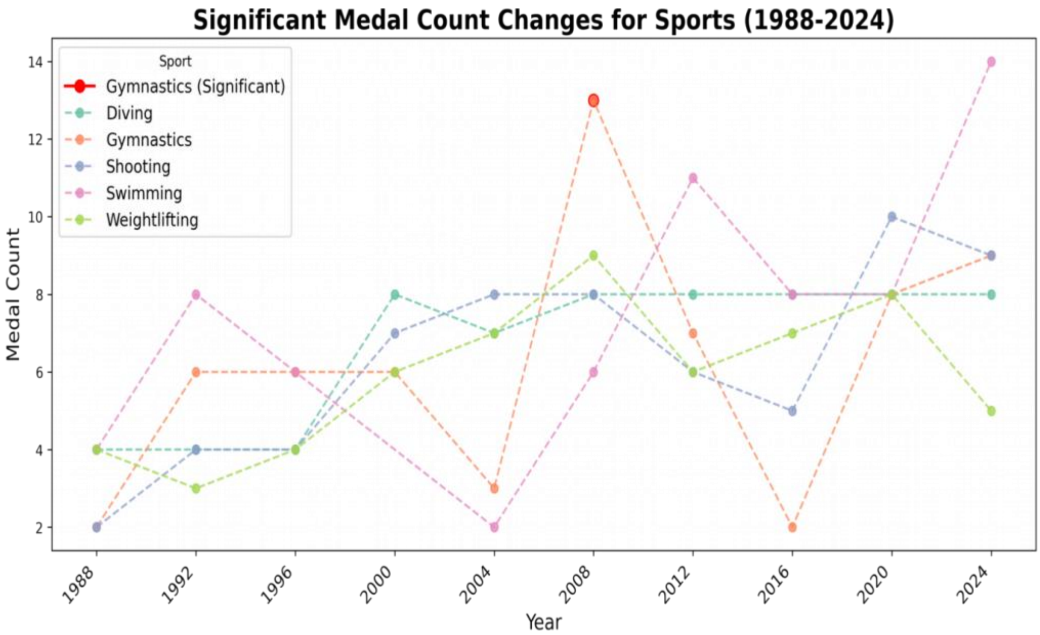

Figure 2. Significant Medal Count Changes for Sports (1988-2024)

According to the chart analysis, the number of MEDALS won by China in sports competitions changed significantly in 2008, reaching 13 MEDALS. This figure exceeds the annual average number of MEDALS for the event and is more than twice the standard deviation, indicating a significant abnormal increase in the number of MEDALS for the year. This phenomenon is closely related to the host country's local advantage.

3.2. China's medal count forecast for 2028

China is expected to win 44 gold, 31 silver and 26 bronze MEDALS, for a total of 101 MEDALS. The model fusion method significantly improves the accuracy of the prediction, and the results are visually represented in Figure 3.

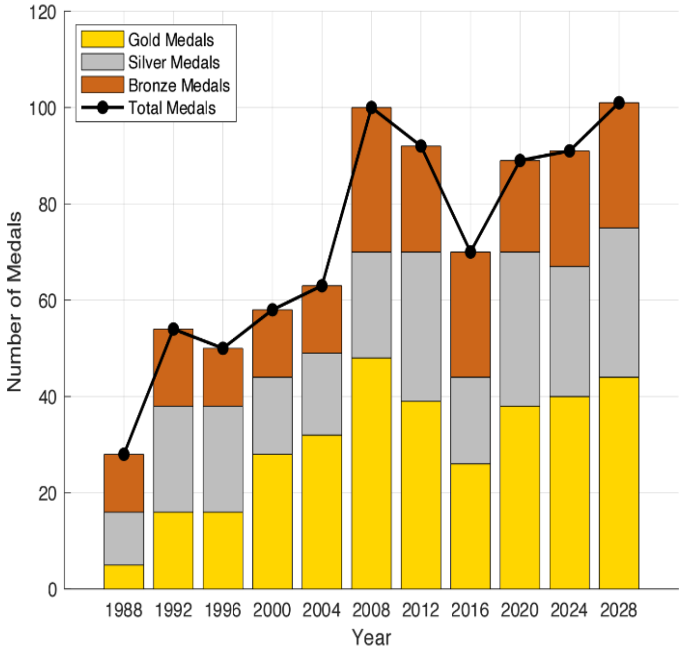

Figure 3. Forecast Results of Chinese Medals in 2028

3.3. Performance change analysis

Based on the established model, further analysis can be conducted to analyze which countries are likely to perform better or worse in 2028, and this paper will use the established "RandomForest+XGBoost+LightGBM" combination model to predict and analyze the top seven countries in the 2024 medal table. The comparison and prediction results of the total number of MEDALS are shown in Figure 4, and the comparison and prediction results of the number of gold MEDALS are shown in Figure 5.

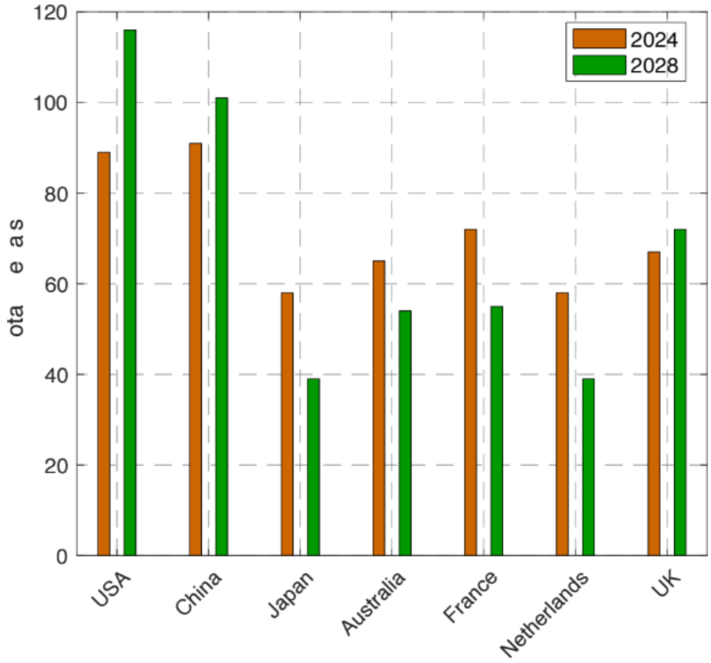

Figure 4. Comparison of total MEDALS in 2024 and 2028

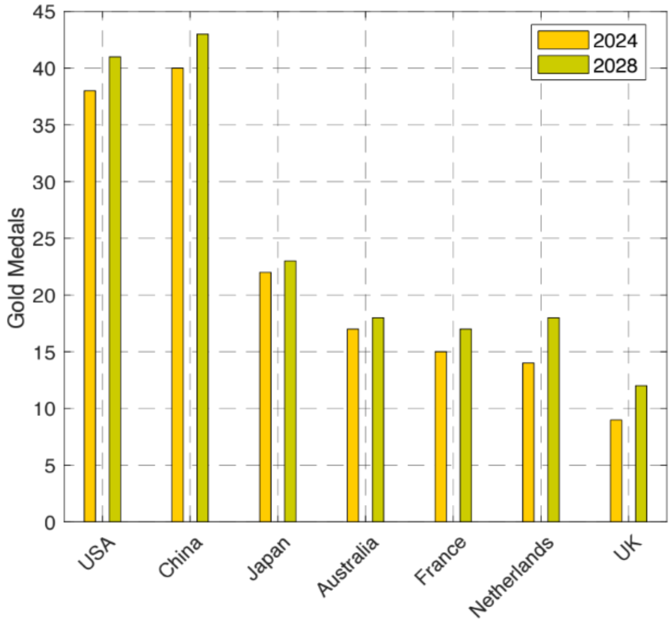

Figure 5. Comparison of gold MEDALS in 2024 and 2028

We can see that the United States and China will still hold the lead in the medal table. At the same time, some small and medium-sized countries stand out in emerging sports, which is likely to break the monopoly of traditional powers and bring about a new medal pattern. This paper will predict the countries that have not yet won Olympic MEDALS based on the established model, and estimate the probability of these countries winning their first medal in the future Olympic Games. According to the predicted results of the model, the top five countries with the highest probability of winning the first prize are shown in Table 5.

Table 5. The five countries with the highest probability of winning the first prize

Country	Estimated Probability	Potential Events
Monaco (MON)	83%	Motorsport, Sailing
Bahamas (BAH)	81%	Athletics (Sprints)
Brunei (BRN)	76%	Athletics (Sprints)
Saint Kitts and Nevis (SKN)	74%	Long-distance Running

Djibouti (DJI)	72%	Swimming, Shooting

4. Conclusions

This study significantly enhances Olympic medal prediction accuracy by integrating ensemble learning and grey prediction models. By leveraging Random Forest, XGBoost, and LightGBM algorithms, and employing a stacking method to combine these models, the prediction performance and stability are improved. The model systematically considers the host-country effect and introduces a parallel grey prediction module to address data sparsity and uncertainty. Empirical results show that the proposed model achieves training/testing scores of 0.823/0.806 on the 1988-2024 Olympic dataset, significantly outperforming baseline models.

Although this study has achieved remarkable results in Olympic medal prediction, there is still room for further improvement in prediction accuracy and model adaptability. Future research could explore optimizing feature selection and model fusion methods, particularly when dealing with more complex and dynamic datasets. Additionally, considering the diversity of Olympic athletes and the evolving nature of sports events, incorporating more dimensions of data, such as athletes' physiological and psychological states, into prediction models will be an important direction for future development. With the continuous advancement of artificial intelligence, exploring deep learning-based models for more refined and intelligent Olympic medal prediction could also be a promising area for future research.

References

[1] LI Xiang. Olympic Games Medals Prediction Based on Time Series Analysis [J]. Computer & Digital Engineering, 2018, 46 (03): 533-536+545.

[2] Luo Yubo, Cheng Yanfang, Li Mengyao, et al. Prediction of the Number of Medals and Overall Strength of the Beijing Winter Olympics -- Based on the Host Effect and Grey Prediction Model [J]. Contemporary Sports Technology, 2022, 12 (21): 183-186.

[3] Wen Jing, Li Wei Ping, Lei Fu Min. Multi-method prediction research on gold medals and total medals for Chinese team in Beijing Winter Olympics [C]. // China Sports Science Society. *Abstracts of the 12th National Sports Science Conference - Special Report (Sports Statistics Branch)*. School of Sports and Health, Hangzhou Normal University; Department of Statistics, Xi'an Physical Education University, 2022: 20-22.

[4] Tian, H., et al. Medal Forecasts and Competition Strategies for Chinese Athletes in the Beijing 2022 Winter Olympic Games -- Based on Olympic Home Advantage [J]. CHINA SPORT SCIENCE, 2021, 41 (02): 3-13+22.

[5] Yuan Junjie. Preliminary Study on the Prediction Model of Olympic Gold Medals in the Big Data Era -- A Case Study on the World Athletics Championships [J]. BULLETIN OF SPORT SCIENCE& TECHNOLOGY, 2021, 29(06): 132-134.

[6] Cao Jingbo. Medal prediction and preparation strategy for China in the 2022 Beijing Winter Olympics based on host country effect [C]. // General Administration of Sport of China, Sports Culture Development Center; China Sports Science Society, Sports History Branch. *Abstracts of 2021 Academic Conference on Cultural Resources Facilitating the Dissemination and Heritage Development of Beijing Winter Olympics Culture*. Shenyang Sport University, 2021: 73-74.

[7] LONG Jiayong, WEI Zhuohong. Research on the Prediction of Men's 100m Gold Medal Results of the Olympic Games Based on GM(1,1) Grey Model [J]. Journal of Southwest China Normal University (Natural Science Edition), 2023, 48(07): 123-128.

[8] Li Yang. Research on the competitiveness of Chinese track and field athletes in Paris Olympic Games [D]. Wuhan Sports University, 2024.

[9] Christoph S, L. S S, Dominik S, et al. Forecasting the Olympic medal distribution – A socioeconomic machine learning model [J]. Technological Forecasting & Social Change, 2022, 175.

[10] SHI Huimin, ZHANG Dongying, ZHANG Yonghui. Can Olympic Medals Be Predicted? :Based on the Interpretable Machine Learning Perspective [J]. Journal of Shanghai University of Sport, 2024, 48(04): 26-36.