Olympic Medal Prediction Based on Machine Learning Models

Yiming Xiao [1, #], Qimao Wang [2, #, *]

[1] School of Naval Architecture & Ocean Engineering, Dalian University of Technology, Dalian, China, 116024

[2] School of Chemical Engineering, Ocean and Life Sciences, Dalian University of Technology, Dalian, China, 116024

* Corresponding Author Email: wqm15942679006@mail.dlut.edu.cn

#These authors contributed equally.

Abstract. The Olympic Games are the highest stage for global sports competition, and the number of medals is an important indicator of a country's athletic strength. This study uses machine learning methods such as Gradient Boosting and Random Forest models to predict the number of gold medals and total medals for each country in the 2028 Los Angeles Olympic Games, while also exploring the medal potential of emerging nations. For countries with historical medal data, a Random Forest model is constructed with features including host country status, historical medal count, athlete ratings, and the number of events participated in. Model predictions show that traditional sports powerhouses such as the United States, China, and Russia will continue to dominate the medal table, with the host country, the United States, expected to win more medals. For emerging countries, the Gradient Boosting model predicts that eight countries have a greater than 50% chance of winning medals for the first time in 2028. Furthermore, K-means clustering analysis reveals the strengths of different countries in specific events and the impact of the host country's status on medal distribution. This study provides a comprehensive methodological framework for Olympic medal prediction and offers important references for the formulation of national sports policies.

Keywords: Olympic Games, Medal count prediction, Great Coach Effect, medal-winning potential of emerging nations, Machine Learning Models.

1. Introduction

1.1. Application scenario

1) Optimization of National Olympic Strategies

National Olympic Committees adjust resource allocation strategies based on the model's predictions of medal distributions (e.g., gold medals, total medals, and uncertainty intervals).

Application Example:

The U.S. Olympic Committee could use the model to predict the home advantage effect for the 2028 Los Angeles Olympics, prioritizing investments in traditional strengths such as swimming and track and field to expand their lead.

China could optimize training plans or introduce new technologies for events predicted to decline (e.g., weightlifting, gymnastics).

2) Breakthrough Planning for Emerging Nations

Predict the probability of countries that have not yet won medals achieving their first medal, helping them formulate an "Olympic Breakthrough Plan."

Application Example:

- Countries that have not yet won medals could focus on cultivating athletes or recruiting foreign coaches in potential events (e.g., sprinting, judo) based on the model's recommendations.
- The International Olympic Committee could use the analysis results to provide targeted support (e.g., equipment, training camps) to small nations, promoting diversity in Olympic medal distribution.

1.2. Research background

When each Olympic Games is held, people all over the world will pay attention to the medal list, how many medals and gold medals their national or regional teams can get, and whether those teams that have not won medals in the Olympic Games can get the first medal. Based on this, our group wants to establish a model to predict the number of medals and gold medals won by each country in the next Olympic Games, and predict which national teams are more likely to win their first Olympic medal in the next Olympic Games. The distribution of Olympic medals reflects a country's sports strength. The development of data science and machine learning technologies has provided new methodologies for medal prediction.

1.3. Research objectives

Predicting the number of gold medals, total medals, and the potential of each country for the 2028 Olympic Games.

Survey of Existing Algorithms

A few years ago, COVID-19 had a significant impact on people's lives, which is why Schlembach C, Guo J, and others included it as a feature in their prediction model. However, as the impact of COVID-19 on the Olympics has gradually weakened over time, it is no longer appropriate to continue considering COVID-19 as a feature in the model under the current circumstances [1][2].

Wang Y and others used the STGCN-LSTM model to predict Olympic medal counts. Although this deep learning model can capture nonlinear relationships in the data, adapt to the complex Olympic medal prediction scenario, and integrate multiple data sources (such as historical medals, economic indicators, population data, etc.) to improve prediction accuracy, the Olympics are held every four years, which results in very limited historical data (for example, there have only been 8 Summer Olympics in the past 30 years). For deep learning models like STGCN-LSTM, small sample training can easily lead to overfitting, making it difficult for the model to generalize. The 4-year interval between the Olympics weakens the continuity of the time series, and LSTM relies on dense time steps to learn long-term dependencies. Furthermore, as a complex deep learning model, STGCN-LSTM operates as a black box, making it difficult to interpret the specific contributions of spatiotemporal features (e.g., whether a country's change in medal count is influenced by increased economic investment or reduced competition from neighboring countries) [3].

Wang Fang used neural networks to predict Olympic medal results. Neural networks can capture complex patterns and may achieve higher prediction accuracy than traditional statistical methods, especially when dealing with nonlinear relationships. Additionally, the model can adjust according to the data, adapting to different pattern changes. However, it requires a large amount of high-quality data for training; otherwise, its performance may be suboptimal. As a "black box" model, it is difficult to explain the reasons behind the predictions, which affects transparency [4].

Luo Yubo and others used the grey prediction model. Peng Jinqiang et al. also used a grey model to analyze the development trend of athletics performance at the Paris Olympic Games based on results from the World Athletics Championships. Hu Wenqiang also used a grey model to analyze the competition results of speed skating events in the 16th to 23rd Winter Olympic Games. The grey prediction model is primarily used to handle small sample data, and Olympic medal data is usually limited in historical records. However, the grey prediction model is mainly designed for predicting linear systems, while the prediction of Olympic medal counts may be influenced by various nonlinear factors (such as economic, social, and political factors). Therefore, the model may not fully capture these complex relationships [5] [6] [7].

Yang Zhongxiu et al. used the multivariate grey GM (1,1) forecasting model to study the development trend of swimming performance in the 11th to 19th Asian Games. The multivariate grey GM(1,1) forecasting model is a method suitable for small samples and information-poor systems. It offers advantages such as simple modeling, low computational requirements, and minimal assumptions about data distribution. It demonstrates strong adaptability and robustness, particularly in cases where the sample size is limited or the system structure is unclear. By analyzing multiple related factors through grey relational analysis, the model can enhance the accuracy and comprehensiveness of predictions. However, it also has notable drawbacks: it requires the original data to exhibit a certain monotonicity and is best suited for data with exponential growth trends. Its performance is limited when applied to complex nonlinear systems or data with strong randomness. Additionally, the model's parameters are sensitive, making the results easily influenced by the initial data, and it lacks a dynamic correction mechanism, which restricts its effectiveness in long-term forecasting or high-dimensional system modeling [8].

Li Haiwei conducted a study on the development trend of athletics events in past Olympic Games using grey-Markov forecasting. The grey-Markov forecasting model combines the characteristics of grey system theory and Markov chains, making it suitable for handling small sample and information-poor data, and effectively predicting the transition trends of system states. Its advantages include the ability to establish models under incomplete information, strong adaptability, and high forecasting accuracy for time series data, especially when the data exhibits certain regularity. The model processes the original data through grey generation, reducing noise interference, while the state transition characteristics of the Markov chain can well describe the dynamic changes of the system. However, the grey-Markov forecasting model also has some drawbacks: it is sensitive to the initial state of the data, and the accuracy of the model depends on the grey generation process and the construction of the Markov chain state transition matrix. Moreover, the model's performance is poor for nonlinear and complex systems. Additionally, the model typically assumes constant state transitions, but in reality, the dynamic changes of the system may not be constant, leading to errors in long-term forecasting [9].

Shi Huimin et al. used machine learning to evaluate the predictability of gold medals and overall medals in different events. The advantages of machine learning forecasting lie in its ability to handle large amounts of complex data, automatically learn patterns from the data, and achieve high prediction accuracy, especially for nonlinear and high-dimensional problems. Its drawbacks include the need for large amounts of high-quality data for training, poor model interpretability, and significant computational resource consumption, which may lead to overfitting. Additionally, the model's performance depends on feature selection and algorithm tuning, and it may perform poorly when there is insufficient training data [10].

1.4. Innovation points

1) Multi-Model Integration and Interdisciplinary Approach

AHP-EWM-TOPSIS Combined Weight Calculation: In the process of predicting medal counts, the Analytic Hierarchy Process (AHP) is combined with the Entropy Weight Method (EWM) to calculate hybrid weights. The Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) is then applied to athlete scoring, ultimately constructing a national scoring system. This combined approach is relatively rare in Olympic medal prediction, effectively integrating subjective weighting and objective data, thereby enhancing the scientific rigor and interpretability of feature engineering.

Differentiated Application of Random Forest and Gradient Boosting Models: To optimize prediction outcomes based on different data characteristics, Random Forest is applied to countries that have previously won medals, while Gradient Boosting models are used for those that have not. This strategic model selection represents an innovative approach to predictive analytics.

2) Threshold Optimization for Low-Probability Event Prediction

For predicting the first-ever Olympic medal of a country—a low-probability event—a selection strategy based on a Gradient Boosting model with a high threshold (0.75) is proposed. By balancing accuracy and recall, this approach reduces decision-making costs. This method provides a reusable technical framework for predicting rare events, particularly applicable to "breakthrough" performances in sports competitions.

3) Multi-Factor-Driven Event Strength Analysis

Using the K-means clustering model, national advantages in specific sports are correlated with multiple factors such as economic strength, geography and climate, cultural traditions, and training systems. This approach overcomes the limitations of single-factor analysis, offering a more comprehensive basis for Olympic strategy adjustments.

4) Data-Driven Dynamic Prediction and Policy Recommendations

By conducting a “2024 vs. 2028” dynamic comparison (e.g., predicting a decrease of 52 medals for the U.S.), the results are directly translated into actionable decision-making references. This highlights the practical value of data-driven strategies in Olympic preparation.

5) Conclusion

The innovations of this study lie in the cross-disciplinary methodological integration, threshold optimization for rare event prediction, systematic multi-factor analysis, and the application of causal inference tools to sports analytics. This research provides a framework that combines theoretical depth with practical value for Olympic medal prediction.

2. Methodology

2.1. Description of the dataset

The experimental data we used came from The official dataset for Problem C of the 2025 Mathematical Contest in Modeling (MCM) available at http://www.comap.com/undergraduate/contests/mcm/, from The United Nations Department of Economic and Social Affairs (UN DESA) available at https://www.un.org/development/desa/dpad/ and UN Population Division available at https://www.un.org/development/desa/pd/ .These datasets include the number of gold, silver, and bronze medals won by each country over the years, the performance of athletes from each country in each competition, the host countries of the past Olympic Games, the GDP of each country, as well as the population of each country, and so on.

2.2. Data preprocessing

1) Missing Value Handling:

Check for missing values in the dataset. For missing values, the following approaches can be chosen based on the situation:

(1) Delete rows or columns with a high proportion of missing values.

(2) Fill missing values using mean, median, or interpolation methods.

(3) For certain key features, if missing values are abundant, consider supplementing with external data sources.

2) Outlier Handling:

Check for outliers in the data (e.g., negative medal counts or abnormally large values). Outliers can be identified using methods such as box plots or scatter plots and addressed based on the context (e.g., deletionor correction).

3) Duplicate Data:

Check for and remove duplicate records to ensure each data entry is unique.

4) Recode:

Perform one-hot encoding on the "Medal" column in the file. Recode the "Host" column in the file (“1”if it is the host country, otherwise “0”).

5) Evaluation model:

Calculate the weights using AHP and EWM respectively. Calculate the combined weights using the existing weights. Score each athlete using the TOPSIS model. Score each country based on the athletes they represent.

6) Normalize all data to eliminate dimensional effects:

Normalization scales the data to a fixed range.

Min-Max normalization linearly scales the data to the range [0, 1].

2.3. Medal Prediction Model

1) Data Loading

Load the dataset `all_in_merged.csv` using the `pandas` library. Print the first few rows of the data for initial inspection.

2) Total Medals Prediction

①Feature Selection:

Select features for predicting total medals, including ‘Total’, ‘Host_NOC’ , ‘Total_Events_Per_Year’, and‘SCORE’.

②Train-Test Split:

Split the dataset into training and test sets, with the test set comprising 20% of the data.

③Model Initialization:

Initialize a Random Forest regression model with ‘n_estimators=100’ and ‘max_depth=5’

④Cross-Validation:

Evaluate the model's performance using 5-fold cross-validation and calculate the Root Mean Squared Error(RMSE).

⑤Model Training:

Train the model on the training set.

⑥Test Set Prediction:

Make predictions on the test set and calculate the RMSE.

⑦Confidence Interval Calculation:

Calculate the 95% confidence interval for the predictions.

⑧2028 Prediction:

Predict the total medals for 2028 and save the results to a CSV file.

⑨Visualization:

Plot the top 10 countries predicted to win the most total medals in 2028 and save the image.

3) Gold Medals Prediction

Similar to the medal count prediction, we modified one of the features from "historical medal count" to “historical gold medal count”.

4) Model Evaluation Summary

Print the average RMSE, mean residuals, and average confidence intervals for total medals and gold medals. The final result is as follows:

Mean RMSE (Total Medals): 4.02

Mean RMSE (Gold Medals): 0.64

Mean Residual (Total): 4.06

Mean Residual (Gold): 0.00

Avg Confidence Interval (Total Medals): [-15.7, 40.1]

Avg Confidence Interval (Gold Medals): [2.8, 3.5]

2.4. National Potential Model

1) Gradient Boosting Algorithm Framework

$$F_M(x) = F_0(x) + \eta \sum_{m=1}^{M} f_m(x) \tag{1}$$

model expression

2) Loss function (logarithmic loss)

Binary Logarithmic Deficiency

$$\mathrm{L(y,F(x))} = -\frac{1}{\mathrm{N}} \sum_{i=1}^{N} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \tag{2}$$

3) Pseudo residual calculation

Negative gradient (pseudo residual) of the math iteration

$$r_i^{(m)} = y_i - p_i^{(m-1)}, p_i^{(m-1)} = \frac{1}{1+e^{-F_{m-1}(x_i)}} \tag{3}$$

$r_i^{(m)}$: The residual of the mth round is used to train the mth tree

4) Decision tree splitting criteria

Node splitting based on minimizing squared error

$$\min_{s,j}\left[\sum_{x_i \in \mathrm{Left}(s,j)}\left(r_i^{(m)}-\overline{r}_L\right)^2+\sum_{x_i \in \mathrm{Right}(s,j)}\left(r_i^{(m)}-\overline{r}_R\right)^2\right] \tag{4}$$

5) Hyperparameter optimization (random search)

Parameter space definition

Learning rate~ Uniform(0.01,0.1); Number of trees ~ Uniform(50,300)

Maximum depth~ Uniform(3,7); Minimum number of split samples ~ Uniform(10,30)

objective function

$$\max_{\theta} AUC-ROC(F_{\theta}(x), y_{test}) \tag{5}$$

θ: Hyperparameter combination

AUC-ROC : As a scoring criterion for cross validation

6) Model evaluation indicators

Accuracy

$$\mathrm{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{6}$$

F1 score

$$F1 = 2\cdot\frac{\mathrm{Precision}\cdot\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}}, \quad \mathrm{Precision} = \frac{TP}{TP+FP}, \quad \mathrm{Recall} = \frac{TP}{TP+FN} \tag{7}$$

AUC-ROC

$$AUC = \int_0^1 TPR(f)\cdot FPR'(f)df \tag{8}$$

TPR(True Positive Rate): $TPR = \frac{TP}{TP+FN}$

FPR(False Positive Rate): $FPR = \frac{FP}{FP+TN}$

7) Prediction probability and threshold screening

Prediction probability of country i

$$p_i = \frac{1}{1+e^{-FM(x_i)}} \tag{9}$$

Prediction of the number of countries that will win their first medal at the 2028 Los Angeles Olympics

$$\text{Predicted Count} = \sum_{i=1}^{N_{\text{unawarded}}} \mathrm{I}(p_i \geq \tau) \tag{10}$$

τ : probability threshold(set to 0.75 in the code)

Indicator function (when the conditions are met, it is 1)

8) Statistical significance in visualization

$$\text{Confusion Matrix} = \begin{bmatrix} \text{TN} & \text{FP} \\ \text{FN} & \text{TP} \end{bmatrix} \tag{11}$$

Scatter plot quantile line

3. Result and discussion

3.1. Prediction results of awarded countries

The construction process of the random forest model includes data loading and preprocessing, feature selection, model initialization, cross-validation, model training, test set prediction, confidence interval calculation, 2028 prediction, visualization, residual analysis, and model evaluation summary. Through these steps, the model can predict the total medal count and gold medal count and evaluate its performance. Based n this Random Forest model, the following results were obtained. Taking the top ten countries in terms of medal count/gold medal count as an example, the prediction results are shown in the Figure 1 and Figure 2 below.

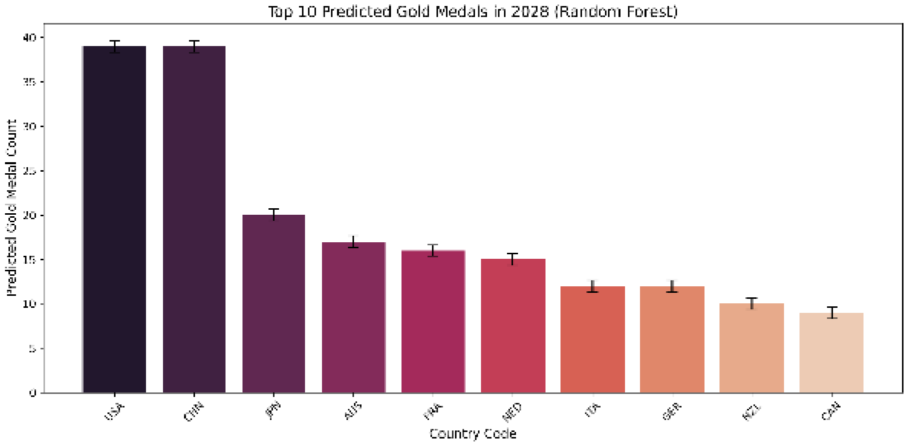

Figure 1. Top 10 Predicted Gold Medals in 2028 (Random Forest)

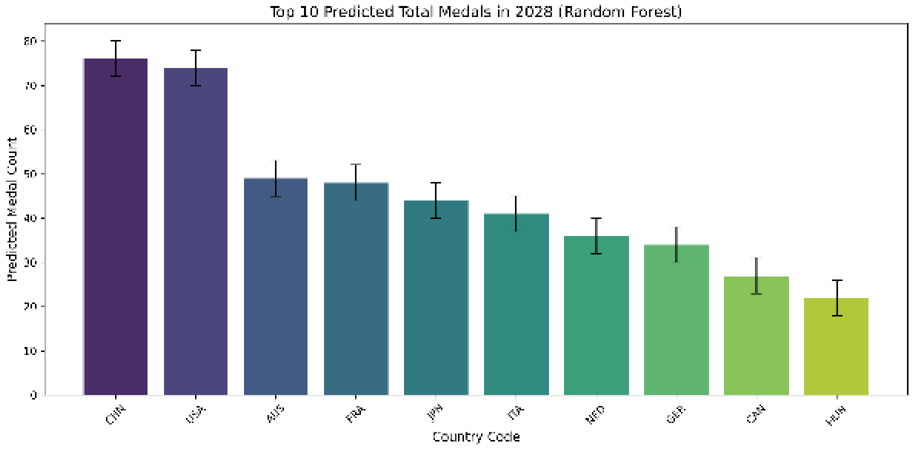

Figure 2. Top 10 Predicted Total Medals in 2028 (Random Forest)

3.2. Prediction results of emerging countries

Due to the low probability of countries that have not won medals in history winning at the 2028 Los Angeles Olympics, medal prediction is essentially a "strict selection" scenario, and it is necessary to ensure that the selected countries have significant advantages. The essence of threshold selection is to achieve a corresponding balance between accuracy and recall. Compared with the default threshold scheme, the 0.75 threshold scheme can reduce the cost of decision errors. Taking into account the number of athletes, athlete experience, population, per capita GDP, and other data from these countries, it is estimated that eight countries will win their first medal at the Summer Olympics, with a probability of over 50%.

3.3. Strategic recommendations and discussion.

1) Missing Value Handling:

Consider using more sophisticated imputation methods, such as KNN imputation or model-based imputation, rather than simple mean or median filling.

2) Outlier Detection:

Use more robust statistical methods (e.g., IQR) to identify and handle outliers, avoiding negative impacts on the model.

3) Feature Selection:

In addition to existing features (e.g., historical medal counts, host country status), introduce more sports-related features, such as national sports infrastructure investments, athletes' training levels, and the strength of sports policy support.

4) Probability Calibration:

Use methods like Platt Scaling or Isotonic Regression to calibrate the predicted probabilities of the model, ensuring that the predicted probabilities align with actual probabilities.

5) Model Evaluation:

In addition to AUC-ROC, include other evaluation metrics such as the PR curve (Precision-Recall curve) and F1 score to comprehensively assess the model's performance.

4. Conclusions

4.1. Research summary and review

In this report, aiming to provide decision-making references for national Olympic committees ,we have developed mathematical models to predict the number of gold medals and total medals each country will win at the 2028 Olympic Games, categorizing nations into those that have never won an Olympic medal and those that have. Based on the predictions from our model, we have identified which countries are most likely to improve their performance and which may perform worse compared to their 2024 results .

4.2. Research limitations

1) Data availability:

This model relies on historical data, but historical data may not accurately reflect future performance as athletes' skills, team dynamics, or external factors may change.

2) Model assumptions:

The model assumes that the performance of countries such as the United States and China is stable, but it fails to take into account potential disruptive factors that could affect future performance, such as political or economic instability, or unexpected events like the COVID-19 pandemic.

3) External variables:

The model may not have taken into account all the variables that affect performance, such as sudden changes in training systems, shifts in public interest, or new policies that may impact athlete development.

4) Low probability of winning:

Countries that have never won an Olympic medal historically have a lower probability of winning, which makes predictions based on small datasets potentially inaccurate. This could lead to shortcomings in the model's accuracy.

5) Threshold selection:

Selecting a threshold (for example, a probability of 0.75) to predict which countries will win their first medal may introduce bias, as it could overestimate or underestimate the potential of certain countries based on historical performance and available data.

6) Economic and demographic factors:

Although GDP and population are included as features, they may not fully capture the complex dynamics that influence a country's ability to produce Olympic medalists. These factors may only provide a rough correlation and fail to account for other influencing factors, such as cultural attitudes towards sports or investment in specific sports.

4.3. Prospects for the future

The model predicts the number of medals each country will win at the 2028 Olympics by using historical data and various influencing factors, such as past performance, host country advantages, and athlete statistics. Looking ahead, the prediction model could be improved by incorporating more dynamic data sources, such as real-time athlete performance changes or geopolitical factors that may affect competition. The model could also enhance accuracy by using deeper learning techniques or combining multiple models, particularly in handling uncertainties.

Predicting which countries will win their first Olympic medal in 2028 relies on a gradient boosting model that uses economic factors, population size, and previous participation data. Future development of the model could explore additional metrics, such as the effectiveness of a country’s training system or the role of international cooperation in sports development. Incorporating updated data and factors, such as technological advancements in training or genetic research, could refine predictions for emerging Olympic nations.

References

[1] Schlembach C, Schmidt S L, Schreyer D, et al. Forecasting the Olympic medal distribution–a socioeconomic machine learning model [J]. Technological Forecasting and Social Change, 2022, 175: 121314.

[2] Guo J, Cui D. How will the COVID-19 Pandemic Affect the Asian Games? A Joint Analysis of Olympic and Pandemic Big Data [C]//2022 7th International Conference on Signal and Image Processing (ICSIP). IEEE, 2022: 814-817.

[3] Wang Y, Wang J, Yang J, et al. STGCN-LSTM for Olympic Medal Prediction: Dynamic Power Modeling and Causal Policy Optimization [J]. arXiv preprint arXiv:2501.17711, 2025.

[4] Wang Fang. Prediction of Olympic Medal Results for the 2020 Olympics Based on Neural Networks [J]. Statistics and Decision, 2019, 35 (5): 89-91.

[5] Luo Yubo, Cheng Yanfang, Li Mengyao, et al. Prediction of China's Medal Count and Overall Strength at the Beijing Winter Olympics—Based on Host Nation Effect and Grey Prediction Model [J]. Contemporary Sports Science and Technology, 2022, 12 (21): 183-185.

[6] Peng Jinqiang, Jing Longjun, Chen Shuyin, et al. Analysis of the Development Trend and Grey Forecast of Athletics Performance at the Paris Olympic Games Based on Results from the World Athletics Championships [J]. Bulletin of Sport Science & Technology Literature, 2024, 32 (4): 20–26.

[7] Hu Wenqiang. Analysis and Grey Forecasting Study of Speed Skating Performance in the 16th to 23rd Winter Olympic Games [D]. Shandong: Qufu Normal University, 2021.

[8] Yang Zhongxiu, Li Yuyu, Tu Chunjing. Study on the Development Trend of Swimming Performance in the 11th to 19th Asian Games [J]. Zhejiang Sports Science, 2024, 46 (4): 62–68.

[9] Li Haiwei. Research on the Development Trend of Athletics Events in Past Olympic Games and Grey-Markov Forecasting [D]. Jiangxi: Jiangxi Normal University, 2021.

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