Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 12
ESAC 2026

Prediction and Strategic Analysis of Olympic Medals

Yihe Ma *, Xiyuan Wang

School of Information Network Security, People's Public Security University of China, Beijing, China, 100038

* Corresponding Author Email: 13223251726@163.com

Abstract. The Olympic medal table showcases the competitive level of each country, and establishing a medal prediction model based on historical data is crucial for accurately predicting each country's performance at the 2028 Los Angeles Summer Olympics. This article first preprocesses data from the past 10 Olympic Games, selecting 11 features Xi (such as total number of participants, total number of events, etc.) and 4 predictor variables Yi (such as whether medals were won, total number of medals, etc.), and encodes the classification features. Subsequently, five cross validation fits were performed on the data using random forest, LightGBM, and XGBoost models to establish the relationship between features and predictor variables. Next, the stacked model is used to integrate the prediction results of these three models, assign different weights, and finally construct an Olympic medal prediction model. On the training set (80%), the model scored 0.823, and on the testing set (20%), the model scored 0.806. Next, use ARIMA and grey prediction models to predict the feature variable Xi of the 2028 Olympics, and substitute it into the medal prediction model to obtain the medal situation Yi of each country in 2028. According to the model, China and the United States may perform better in 2028, while Japan and France may perform worse; The probabilities of Monaco, Bahamas, and Brunei winning medals for the first time are 83%, 81%, and 76%, respectively; Diving, weightlifting, shooting, and gymnastics will have a significant impact on China's total medal count; As the host country, there was a significant change in the medal count of Chinese gymnastics in 2008.

Keywords: Olympic Medal Prediction Model; Random Forest; Lightgbm; XGBoost; Stacked Model.

1. Introduction

Olympic medal prediction research has evolved from statistical analyses (e.g., Bernard & Busse's 2004 regression linking GDP to medals) to advanced machine learning, as demonstrated by R. Sayeed et al.'s (2025) evaluation of 13 ML models. Despite improved accuracy, critical gaps persist: existing feature systems over-rely on macroeconomic/historical data, ignoring micro-factors (e.g., coach quality), and studies on emerging nations or first-medal probabilities remain scarce. While methods like Lundberg and Lee's (2017) SHAP framework and dynamic feature engineering offer solutions for interpretability and trend capture, their application to Olympic predictions is understudied.

The Olympic Games is the world's highest level comprehensive sports event, and the number of medals won by countries not only reflects the athletes' competitive level, but also is influenced by multiple factors such as sports systems, training mechanisms, and event settings [1]. The medal table of the 2024 Paris Summer Olympics shows the performance of various countries, with the United States ranking first with 126 medals, China tied with the United States for first place in the gold medal table, and France ranking fifth in the gold medal table. However, there are still over 60 countries that have not yet won Olympic medals, and these countries may face breakthroughs in the future. The aim of this study is to establish a medal prediction model based on existing Olympic historical data, athlete participation information, and event settings, predict the medal table for the 2028 Los Angeles Summer Olympics [2], and evaluate possible performance changes in various countries. We will explore the advantages of each country in different events, analyze the impact of factors such as host country selection and the "great coach" effect on medal count [3]. We hope to provide data support for national Olympic committees through model prediction, help them optimize their preparation strategies, and provide future development references for countries that have not yet won medals.

2. Research on Olympic Medals Based on Random Forest Model

2.1. Establishment of Random Forest Model

The data source of this paper is https://www.mcm.edu.cn/.

Random forest is a type of ensemble learning model that constructs multiple learners to collectively accomplish specific learning tasks. The basic idea is to first generate a set of basic learners, and then combine these learners together through specific strategies to form a holistic model. Decision trees are often used as basic learners because they themselves belong to weak learners, but after integration, they typically exhibit strong predictive abilities and become typical strong learners [4]. Random forest is a set model based on decision trees, using the CART decision tree algorithm. Due to its excellent predictive performance and stability, random forests have been widely used in various tasks. When studying the prediction of Olympic medal numbers, the random forest algorithm can be used to model and analyze the impact of different factors on medal numbers. The following will elaborate on the steps of establishing an Olympic medal prediction model based on the random forest algorithm.

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

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

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

In the problem of predicting the number of Olympic medals, the predicted value representing the b-th tree is. Usually, Y - is the average predicted value of all base learners (i.e. decision trees). Assuming the training set sample data is $D=\{(x_1, Y_1), \cdots, (x_n, Y_n)\}$, where Y_i represents the number of medals in a certain Olympic Games, and xi 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.). In this way, the random forest model can be trained and predicted based on these feature variables, effectively predicting the number of Olympic medals.

(1) Sample extraction and training

There are a total of n samples in the historical dataset of the Olympic Games. Use bootstrap method to randomly select samples from the sample set for model training. Self service methods allow for replacement sampling, so some samples may be repeatedly sampled. Unsampled samples will be used as the test set for the model to evaluate the predictive performance of random forest regression. These unextracted data are referred to as' out of bag data '(OOB).

(2) Feature selection and partitioning

In the original competition dataset, there are p feature variables. When building a regression tree, it is necessary to select m feature variables from the branch nodes to construct the regression tree, which can be constrained by the parameters in the model. When using the random forest model for regression prediction, the regression tree can grow freely without pruning operations.

(3) Repeated sampling and modeling

Repeat the steps of random sampling and building regression trees until all regression trees have been trained. After all the regression tree training is completed, the modeling process based on the random forest algorithm is also completed, so that the impact of the optimal speed allocation strategy on the total time can be predicted.

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

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

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

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

As for predicting the number of Olympic medals, scores 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 optimization parameters include: n_estimators (number of decision trees), x_depth (maximum depth of decision trees), and bootstrap (whether backtesting is performed). This study uses 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 the fault 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 boot trap=True, n_estimators=80, and x_depth=10 was ultimately selected, and the score increased to 0.91, significantly improving the fitting performance and prediction accuracy of the model. The scores for the training and testing sets 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

2.1.1. Establish LightGBM model.

LightGBM is an efficient gradient boosting framework based on decision tree algorithm, widely used in regression and classification tasks. Compared with traditional algorithms, LightGBM maintains high accuracy while increasing running speed by about ten times and reducing memory consumption by about three times. It has high training efficiency, low memory consumption, supports GPU and parallel computing, and is very suitable for processing large-scale data [5]. In the task of predicting the number of Olympic medals, LightGBM can effectively handle complex features and large-scale data, improving the accuracy of model predictions.

In this study, the LightGBM library in Python was first used to fit historical Olympic medal counts with default parameters for preliminary prediction. Then, optimize the model parameters: set the initial learning rate to 0.1 to accelerate model convergence; 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, it is possible to more accurately capture the complex factors that affect the number of Olympic medals, providing a more reliable basis for predicting the number

of medals in different countries and events. The training and testing set scores of LightGBM 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

3. 2028 feature (X1, X2, X3... X10, X11) prediction model

In order to predict the number of Olympic medals in 2028 (Y1, Y2, Y3, Y4), this article first needs to predict the relevant features (X1, X2, X3... X10, X11). For this purpose, this article will use ARIMA model and grey prediction model to predict these features. Subsequently, the predictive performance of the model was evaluated by calculating the evaluation metric 1-wmape (weighted average absolute percentage error). Finally, this article adopts the reciprocal of relative error method for model fusion, combines the advantages of each model, and assigns different weights to each model based on its prediction accuracy, thereby further improving the overall accuracy of the model [6, 7].

3.1. Establish ARIMA model

The ARIMA model smooths historical data to eliminate trends and seasonal factors to meet static requirements. Convert non-stationary time series into stationary series through first-order or second-order differencing operations. Then, use autocorrelation function (ACF) and partial autocorrelation function (PACF) to analyze the data, determine appropriate autoregressive (AR), difference (I), and moving average (MA) parameters, and 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 [8]. 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.

3.2. Establish a grey prediction model

Grey prediction is a prediction method used for small sample, fuzzy, and uncertain data. This method is based on grey system theory and eliminates data instability by generating cumulative sequences for prediction. The grey prediction model generates a set of static sequences by accumulating historical data, and then models and predicts using simplified difference equations. Due to its ability to effectively handle small sample data and cope with incomplete or uncertain situations, grey prediction methods have been widely applied in many fields. When establishing a grey prediction model, the first step is to accumulate historical data, generate data sequences, and obtain the predicted values of the model by constructing the background values and development coefficients of the grey model. Then use the model to predict future data, providing support for further analysis and decision-making. The grey prediction model can provide valuable auxiliary information for predicting the number of Olympic medals. Grey modeling of historical feature data can help improve the accuracy and reliability of medal prediction.

3.3. Model Fusion and Evaluation

After comparing the prediction results of ARIMA prediction model and grey prediction model, it was found that there is a certain gap between the two. In order to improve the prediction accuracy, we adopted a combination model of ARIMA and grey prediction 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. This study used the reciprocal of relative error method to determine the weights of each model. Specifically, the larger the relative error of the model, the lower its prediction accuracy, so the weights of the model should be smaller; On the contrary, models with higher accuracy have greater weights, thereby further improving the accuracy of prediction results. The final predicted total number of participants (X2) is shown in Figure 1.

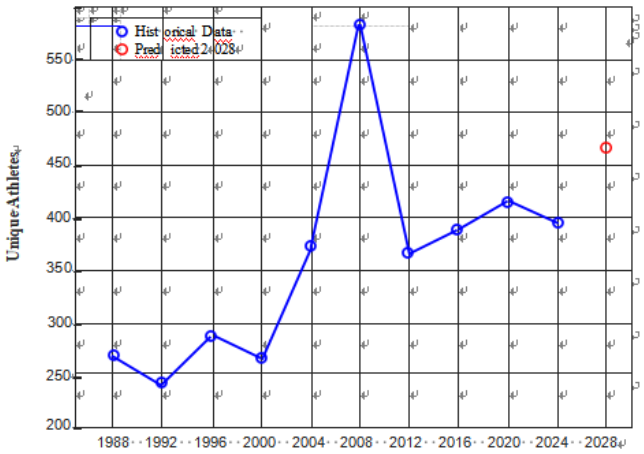

Figure 1. ARIMA+Grey Prediction: Unique Athletes in 2028 (X2)

3.4. 2028 medal forecast

In this study, the advantages of three advanced machine learning algorithms were utilized to predict China's medal count at the 2028 Olympics using model fusion methods: Random Forest, XGBoost, and LightGBM. Each algorithm is carefully selected because they have unique abilities in processing complex data and generating accurate predictions. By integrating the predictive capabilities of these models, more robust and reliable predictions have been achieved. The prediction results indicate that China is expected to win 44 gold medals, 31 silver medals, and 26 bronze medals, totaling 101 medals. This method greatly improves the accuracy of predictions, as it utilizes the individual strengths of each model while minimizing any bias or error that may arise from a single method. The integration of these algorithms effectively captures potential patterns and trends in the data, providing a more comprehensive outlook for China's potential performance in the upcoming Olympic Games. As shown in Figure 2, it demonstrates the predictive ability of the model and its potential to provide valuable insights for future sports analysis.

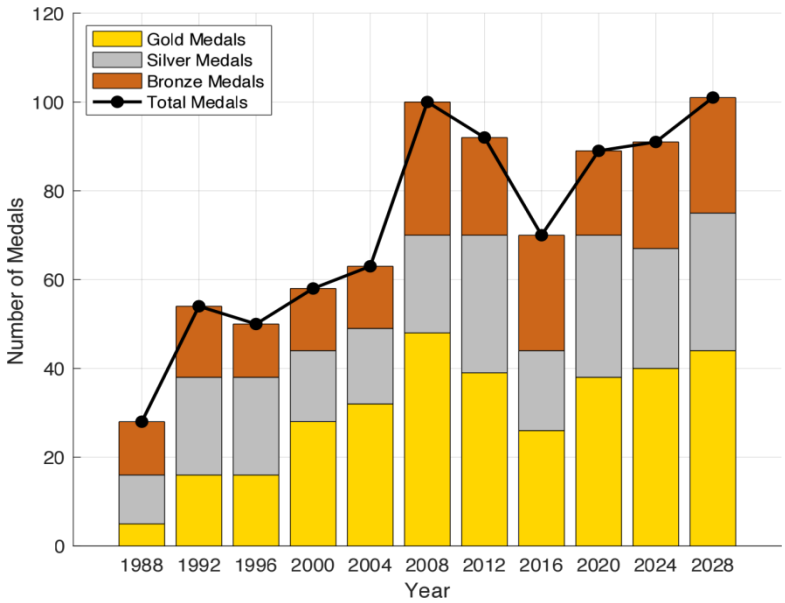

Figure 2. Predicted Results of Chinese Medals in 2028

3.5. Performance change analysis

To analyze which countries may perform better and which countries may perform worse in 2028. This article will use the established "Random Forest+XGBoost+LightGBM" combination model to predict and analyze the top seven countries in the 2024 medal table. The predicted results are shown in Figure 3.

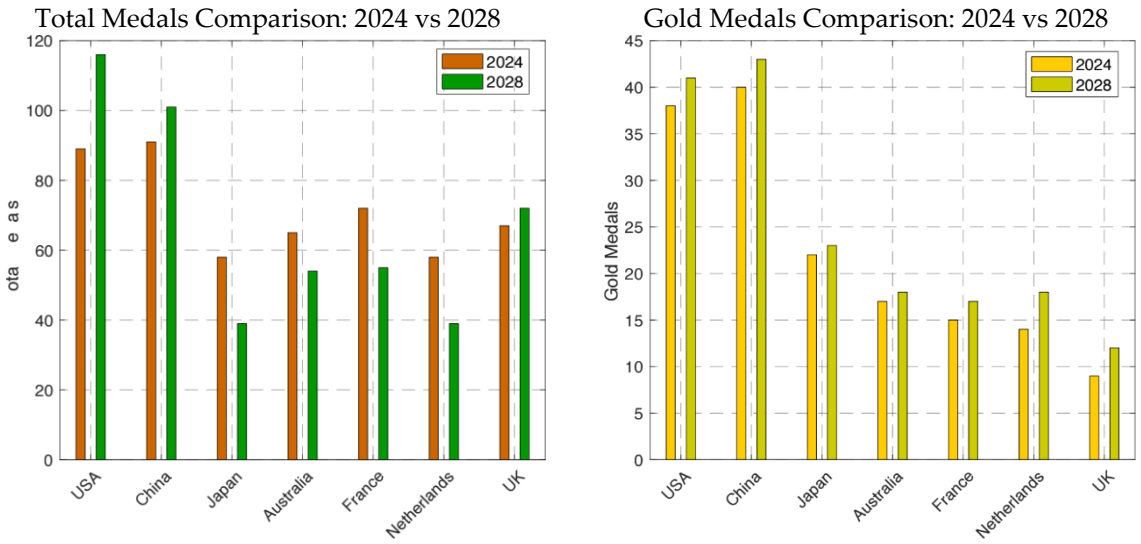

Figure 3. Comparison of medal counts between countries in 2024 and 2028

Analyzing the medal results of the 2028 Olympics, it can be seen that the United States and China will still dominate the medal table. With strong sports capabilities and a deep reserve of athletes, the two countries still have significant advantages in multiple events. However, other countries such as Japan, Australia, France, etc. may have made breakthroughs in certain projects, but still face significant challenges in overall strength, especially in some traditional advantage projects. Although countries such as the Netherlands and the United Kingdom have also experienced growth, they are facing competition pressure from emerging and traditional powers, and their overall performance may be relatively stable. At the same time, some small and medium-sized countries are expected to rise in emerging projects, which may break the monopoly of traditional strong countries and bring new medal patterns. Therefore, the medal distribution of the 2028 Olympics will become more diversified, and global competition will become increasingly fierce. Countries need to adapt to new challenges in order to make breakthroughs on the constantly changing Olympic stage. This article will predict countries that have not yet won Olympic medals based on existing models, and estimate the probability of these countries winning medals for the first time at future Olympic Games. According to the prediction results of the model, the top five countries with the highest probability of winning the championship are shown in Table 3.

Table 3. 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

3.6. The impact of the host country

In this study, we aim to analyze the changes in the number of medals in various sports events in different years through statistical methods, and identify the years with significant changes. Firstly, by calculating the annual average medal count and its standard deviation for each sports event, it serves as a benchmark for evaluating changes. Then, using statistical methods, we compare the number of

medals awarded each year with the annual average of the event [9, 10]. If the difference between the number of medals awarded in a certain year and the annual average exceeds twice the standard deviation, we consider the change in that year to be significant. Next, this article will group all sports events and use this significance detection method to screen for years with significant changes in medal counts. The significant changes in these results help determine which events have shown excellent performance or fluctuations in a specific year, and provide important basis for subsequent medal predictions and trend analysis. The following will present these significant changes in data in an intuitive way, making it easier to understand the performance fluctuations of different projects in different periods.

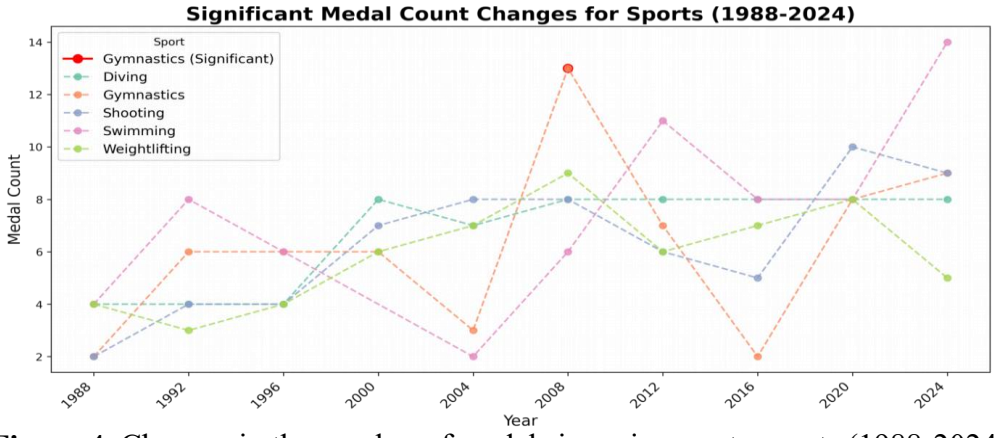

Figure 4. Changes in the number of medals in major sports events (1988-2024)

According to Figure 4, there was a significant change in the number of medals won in Chinese gymnastics in 2008, reaching 13. This data exceeds the annual average number of medals in the gymnastics event and is also more than twice the standard deviation, indicating a significant abnormal increase in the number of medals in the gymnastics event that year. This phenomenon is closely related to the host country's home advantage. As the host of the 2008 Olympic Games, China has invested a lot of resources and training support in gymnastics, providing athletes with sufficient preparation conditions and further stimulating their competitive potential in the favorable environment of home games. In addition, gymnastics, as a traditional advantage event in China, has accumulated profound technical and experiential advantages over a long period of time, making the Chinese delegation's medal achievements in this event particularly outstanding. Therefore, from a data perspective, the choice of host country and investment sources play a decisive role in the significant increase in the number of medals for specific events.

4. Conclusion

This article combines historical data and various machine learning models to propose a comprehensive and actionable Olympic medal prediction method for the 2028 Los Angeles Summer Olympics. An efficient and accurate medal prediction model was constructed through data preprocessing, feature selection, model training, and fusion. By utilizing random forest, LightGBM, and XGBoost models, and integrating the prediction results of multiple models through stacking methods, high prediction accuracy has been achieved on both the training and testing sets. In addition, by combining ARIMA and grey prediction models, a comprehensive prediction of the medal distribution for the 2028 Olympics was made, and it was found that the performance of China and the United States may be better in 2028, while the performance of Japan and France may decline. Finally, based on the constructed medal prediction model, this article provides five insights for the Olympic Committees of various countries, including identifying the projects that contribute the most to future medals, adjusting resource allocation, utilizing home advantage, tapping into the breakthrough potential of countries with fewer medals, and developing personalized preparation strategies. These insights can provide data support for the decision-making of the Olympic Committee, helping it optimize resource allocation, improve overall competitive level, and prepare for the 2028 Olympic

Games and future Olympic cycles. Random forest, LightGBM, and XGBoost can effectively handle high-dimensional and nonlinear data, accurately capture complex relationships between features, improve the accuracy of medal prediction, and meet the competition needs of future Olympic medal prediction. By integrating multiple base learners through a stacked model, bias and variance are reduced, significantly improving the robustness and generalization ability of predictions, and achieving high scores on both the training and testing sets. These models can efficiently process complex data involving multiple countries and sports events, quickly extract effective features, and meet the needs of large-scale Olympic data analysis in competition questions.

References

[1] English P, Fleischman D, Mulcahy R, et al. The Buzz of Brisbane 2032: Themes of Online and Social Media Olympic Sentiment [J]. Communication & Sport, 2024.

[2] Cetinkaya A, Peker S, Kuvvetli Ü. Analysis of countries' performances in individual Olympic Games using cluster analysis and decision trees: the case of Tokyo 2020 [J]. Sport, Business and Management: An International Journal, 2024, 14 (2): 209-229.

[3] Orchard J W, Luies N, Buckley R J, et al. Possible impact of national responses to the COVID pandemic on medal tallies at the Paris 2024 Olympics [J]. medRxiv, 2024.

[4] Breiman L. Random Forests [J]. Machine Learning, 2001, 45 (1): 5-32.

[5] Ke G, Meng Q, Finley T, et al. LightGBM: A highly efficient gradient boosting decision tree [C] // Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017: 3146-3154.

[6] Li H, Wang Y, Li Y, et al. Olympic Medal Prediction Through Multidimensional Features and Enhanced XGBoost Modeling [C] // Proceedings of the 2025 3rd International Conference on Machine Learning and Computer Science. 2025: 1-7.

[7] 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.

[8] Wang Z. Olympic Medal Prediction Based on Multiple Regression and Time Series Analysis [J]. International Journal of Housing Science and Its Applications, 2025, 46 (4): 551-562.

[9] Kumar A, Singh S, Singh P. Predicting Medal Counts in Olympics Using Machine Learning Algorithms: A Comparative Analysis [C] // 2024 3rd International Conference on Innovative Practices in Technology and Management (ICIPTM). IEEE, 2024: 1-6.

[10] 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 [C] // 2024 4th International Conference on Advanced Computing and Communication Systems (ICACCS). IEEE, 2024: 1-6.