A Study on Predicting the Medal Situation of the 2028 Olympic Games Based on Machine Learning and Predictive Modeling

Zhaohe Huo [#, *], Liren Zheng [#], Yubo Wang [#]

College of Mechanical and Vehicle Engineering, Chongqing University, Chongqing, China, 400044

* Corresponding Author Email: 13191138332@163.com

[#]These authors contributed equally.

Abstract. This paper predicts the medal situation of each country in the 2028 Olympic Games by analyzing data from the past 10 Olympic Games. It first selects 11 features (such as the total number of participants, total number of sports categories.) and 4 predictor variables (such as whether medals were won), and encodes the classification features. Random forest, LightGBM, and XGBoost models are then used to fit the data with five-fold cross-validation, establishing the relationship between features and predictor variables. A stacking model is used to integrate the prediction results of these three models, assigning different weights. The model achieved a score of 0.823 on the training set (80%) and 0.806 on the testing set (20%). Subsequently, ARIMA and grey prediction models are applied to predict the feature variables for the 2028 Olympics, and these are substituted into the medal prediction model to obtain the medal situation for each country in 2028. The results suggest that China and the United States are likely to perform better, while Japan and France may perform worse. Additionally, the paper explores the "Great Coach Effect," analyzing Lang Ping's impact as coach of the 2008 U.S. women's volleyball team and the 2016 Chinese women's volleyball team.

Keywords: Olympic Medal Prediction, Machine Learning, Stacking Model, Great Coach Effect, Lang Ping.

1. Introduction

Predicting Olympic medal counts is a vital area of research with far - reaching implications for countries' sports development and global sports standing.

Shi Huimin et al [1]. used explainable machine learning to explore medal predictability, a forward - looking approach. Wen Jing et al [2]. and Tian Hui et al [3]. concentrated on the Chinese team in the 2022 Beijing Winter Olympics. Wen Jing et al. applied multiple methods for prediction, while Tian Hui et al. stressed the home - field advantage. These studies offered valuable but narrow - focused insights.

Wang Jian [4] utilized traditional regression and ARIMA models. Their simplicity was a plus, but they couldn't fully account for the complex medal - winning factors. Yuan Junjie [5] explored a big - data - based model, yet it was mainly relevant to a specific sport. Schlembach et al [6]. developed a socioeconomic machine - learning model. However, its complexity made factor interpretation difficult. Li Haiwei [7] delved into track - and - field event predictions, providing in - depth sport - specific analysis. Paul Catumenn et al [8]. studied coaching changes in football, highlighting non - athlete factors. Long Jiayong and Wei Zhuohong [9] predicted the men's 100m gold - medal results, and Li Jiaqi [10] compared prediction methods for Chinese sprinting. These two studies were event - or sport - centered, offering targeted but limited knowledge.

However, current Olympic medal - count research has drawbacks, often focusing on single - element analysis. This paper aims to bridge the gap. It analyzes data from the past 10 Olympics, selects 11 features and 4 predictor variables, encoding classification features. Three machine - learning models (Random Forest, LightGBM, XGBoost) with five - fold cross - validation are used to fit the data, and a stacking model integrates their results. ARIMA and grey prediction models forecast 2028 feature variables for the medal - prediction model. The study predicts the 2028 medal situation, finding that China and the US may improve while Japan and France may decline. It also

explores the "Great Coach Effect" through Lang Ping's cases. This research offers a more comprehensive prediction method and valuable insights for countries' Olympic strategies.

2. Medal prediction model

2.1. Establishment of predictive models

The data used in this study is sourced from Problem C Data.zip on the website https://www.contest.comap.com/undergraduate/contests/mcm/contests/2025/problems/.

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 core of the random forest algorithm is a set of B trees $\{T_1(X), T_2(X), T_3(X), \cdots, T_B(X)\}$, were,

$$\left\{Y_1 = T_1(X), Y_2 = T_2(X), \cdots, Y_B = T_B(X)\right\} \tag{1}$$

In the problem of predicting the number of Olympic medals, the predicted value representing the b-th tree is $\widehat{Y_b}$. Usually, $\widehat{Y}$ is the average of the predicted values of all base learners (i.e. decision trees). Assuming the training set sample data is $D = \{(X_1, Y_1), \cdot\cdot, (X_n, Y_n)\}$, 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.). In this way, the random forest model can be trained and predicted based on these feature variables, effectively predicting the number of Olympic medals.

Using bootstrap to randomly select samples from the sample set for model training. Selecting m feature variables from branch nodes to construct the regression tree. By repeating the steps of random sampling and building regression trees until all regression trees are trained.

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%.

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

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

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

This study adopts the GridSearch method to automatically adjust n_estimators, x_depth and bootstrap. 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, as 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

LightGBM has increased its running speed, reduced memory consumption, while maintained high accuracy. 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, as 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

The XGBoost algorithm continuously improves prediction accuracy by learning the residual between the predicted value and the true value of the model in each round. 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 to improve the accuracy and robustness of the model. Finally, using the optimized parameter combination to construct the model significantly improved the prediction accuracy, as shown in Table.3.

Table.3. XGBoost's training set and test set scores

argument	value	Training set score	Test set score
learning_rate	0.01	0.821	0.847
n_estimators	1000		
max_depth	8		
lambda	0.3		
alpha	0.3		

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. The stacking method can more accurately capture the relationship between the number of Olympic medals and various factors, thereby improving prediction accuracy and helping to develop more scientific medal prediction models.

From the training effect, it is found that the fusion model has a better effect than the strong learner alone. The fusion fitting effect of Random Forest, LightGBM and XGBoost and the three are shown in the 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. 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, we 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 uses the reciprocal of relative error method for model fusion, combining the advantages of each model and assigning different weights to each model based on their prediction accuracy, thereby further improving the overall accuracy of the model. Figure 1 is a display of the final predicted Total Participants results.

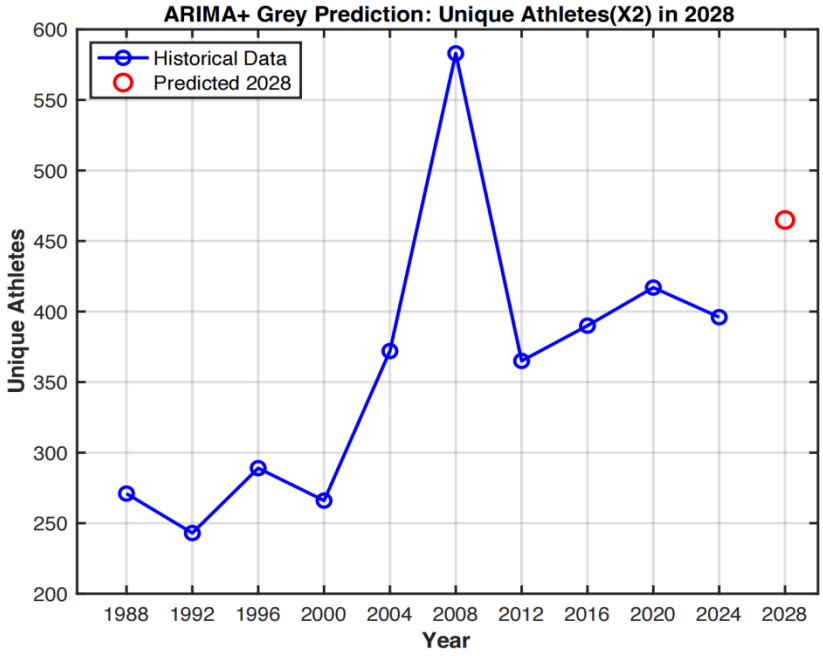

Figure 1: ARIMA+ Grey Prediction: Unique Athletes in 2028

2.3. Forecast of Medals in 2028

In this study, a model fusion approach was employed to predict China's medal count at the 2028 Olympic Games, leveraging the strengths of three advanced machine learning algorithms: Random Forest, XGBoost, and LightGBM. Each algorithm was carefully selected for its unique capabilities in handling complex data and generating accurate forecasts. By combining the predictive power of these models, a more robust and reliable prediction was achieved. The forecasted results indicate that China is expected to secure 44 gold medals, 31 silver medals, and 26 bronze medals, leading to a total of 101 medals. The results are visually represented in the Figure 2, demonstrating the model's forecasting capability and its potential to provide valuable insights for future sports analysis.

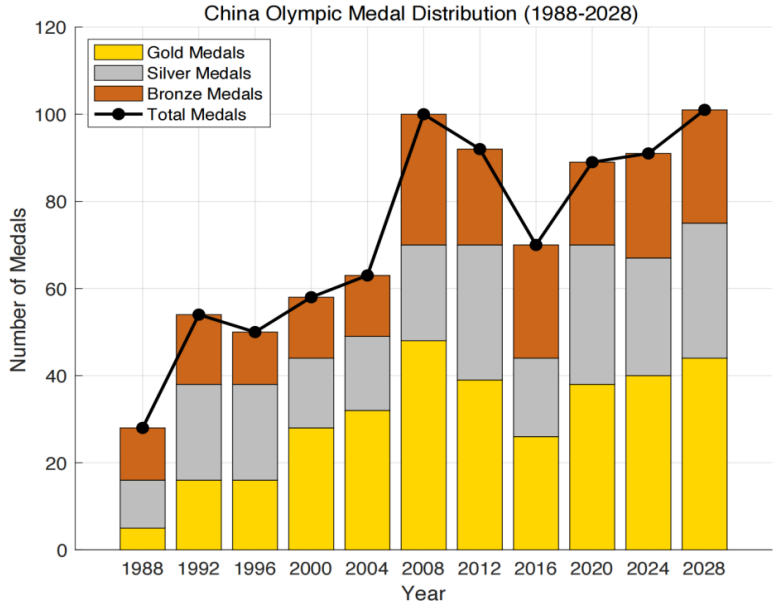

Figure 2: Forecast Results of Chinese Medals in 2028

2.4. Performance change analysis

To analyze which countries may perform better and which 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 prediction results are shown in the Figure 3.

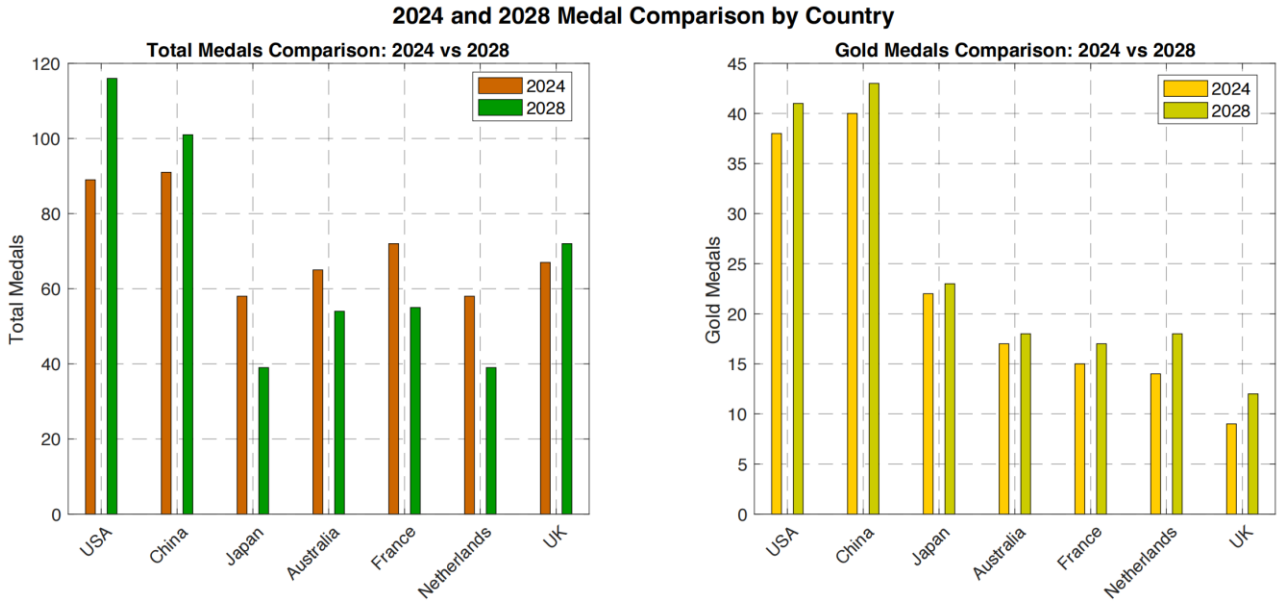

Figure 3: 2024 and 2028 Medal Comparison by Country

When analyzing the medal performance of the 2028 Olympic Games, it can be seen that the United States and China will still occupy the leading position in the medal table. With strong sports strength and deep athlete reserves, the two countries still have significant advantages in multiple events. However, other countries such as Japan, Australia, and France, although may have made breakthroughs in some projects, still face significant challenges in overall performance, especially in some traditional advantage projects. Although countries such as the Netherlands and the United Kingdom may experience 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 emerge in emerging projects, potentially breaking the monopoly of traditional strong countries and bringing about a new medal landscape. For instance, countries such as Monaco (MON), Bahamas (BAH), Bahrain (BRN), Saint Kitts and Nevis (SKN), and Djibouti (DJI) are projected to have their first-ever medal-winning performances with probabilities of 83%, 81%, 76%, 74%, and 72%, respectively. Therefore, the distribution of medals in the 2028 Olympic Games will become more diverse, 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.

2.5. Project Impact

The number of medals in each sports event will be counted, and the contribution ratio of each event to China's total medal count will be calculated. These contribution ratios can reflect the importance of different sports in China's Olympic performance, providing data support for subsequent analysis. Finally, presenting these contribution ratios through visualization tools helps us intuitively understand which projects have the greatest impact on China's Olympic performance. This analysis helps identify key projects and optimize future resource allocation and training strategies. Figure 4 is a bar chart used to visualize the contribution of each sports event, showing the proportion of each event in the total number of Chinese Olympic medals.

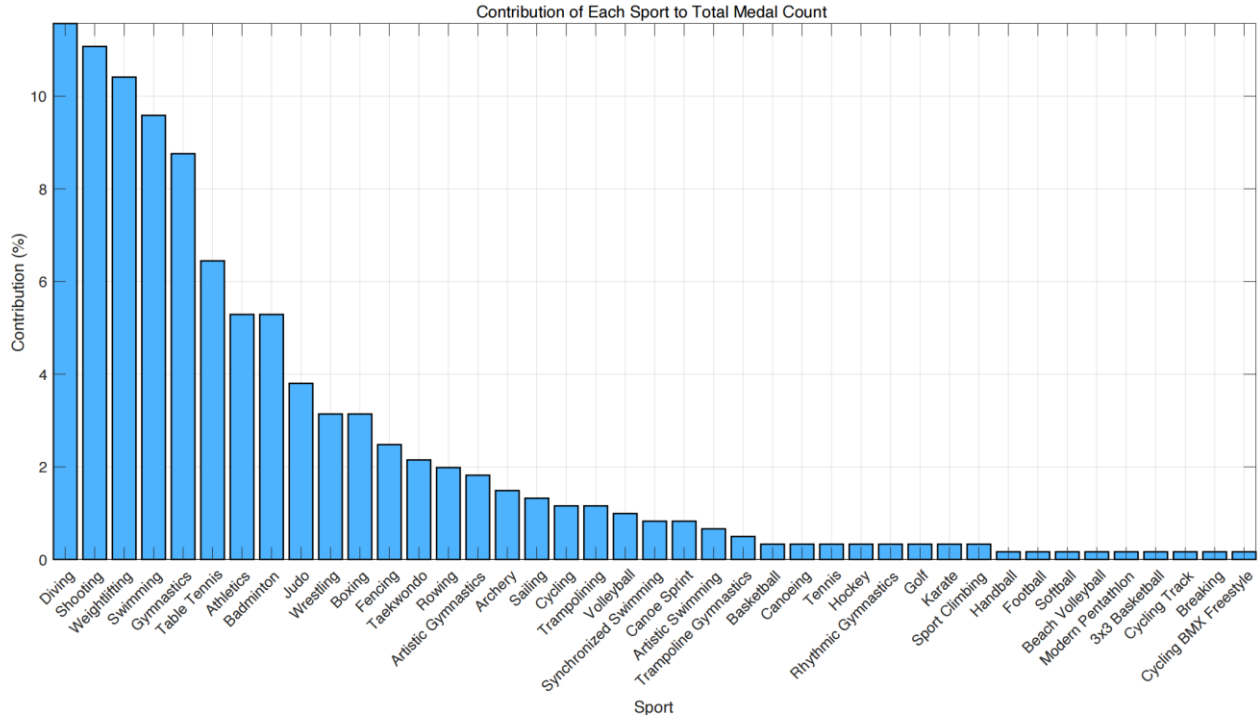

Figure 4: Contribution of Each Sport to Total Medal Count

Next, the medal data of the Chinese Olympic Games from 1988 to 2024 will be screened, and the number of medals will be counted by sports and the contribution will be calculated. According to the analysis of the contribution of various sports, China's long-term dominant events in the Olympic Games, such as diving, weightlifting, shooting, and gymnastics, have significantly contributed to its total medal count. Due to their high technical requirements, long training cycles, and competitive advantages, China's medal count in these fields is relatively stable. Compared to this, events such as table tennis and badminton also have a higher proportion of medals, further proving China's strong strength in individual technical sports. At the same time, with the addition of emerging projects such as breakdancing and 3x3 basketball, China has also shown potential in these emerging fields. Although their contributions are relatively low, they demonstrate China's ability to adapt and expand in diverse competitions.

3. The 'Great Coach' Effect

In order to study the "great coach" effect, this article analyzes the impact of Lang Ping's coaching on the medal scores of the Chinese and American women's volleyball teams. Firstly, we constructed a dataset that covers the medal performances of two teams from 1988 to 2024, and assigned specific score values for different types of medals. The specific scoring criteria are as follows: a gold medal is worth 1 point, a silver medal is worth 0.6 points, and a bronze medal is worth 0.4 points.

Next, by calculating the annual medal scores, we can quantify the fluctuations in medal performance for each year and focus on the years when Lang Ping coached to analyze whether her coaching had a significant impact on the overall performance of both teams. In order to visually display the trend of medal score changes, this article uses Figure 5 to help understand the score changes before and after Lang Ping's coaching, and provides a clear basis for evaluating her coaching effect.

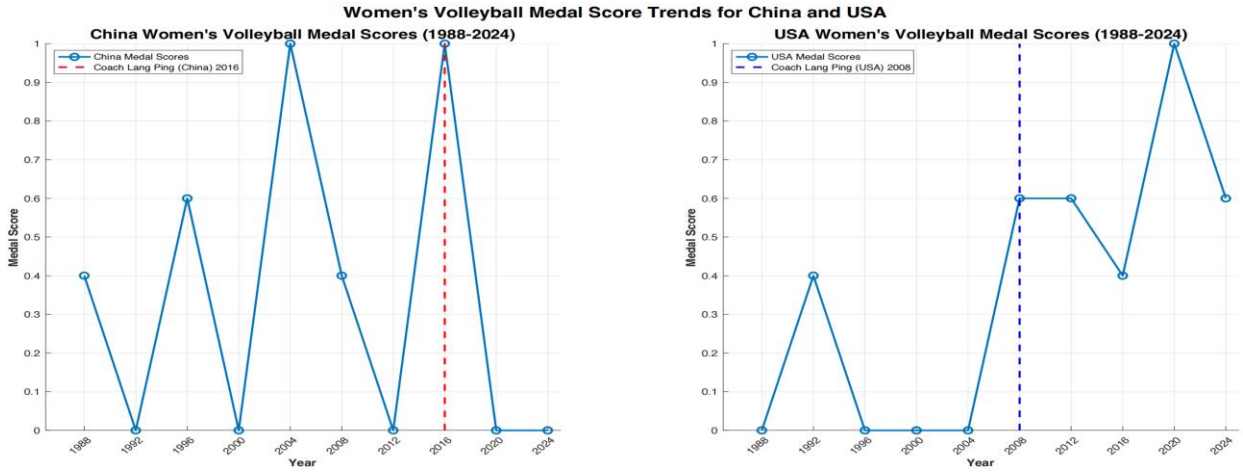

Figure 5: Women's Volleyball Medal Score Trends for China and USA

Furthermore, we explored the specific impact of Lang Ping's coaching on the medal scores of the Chinese and American women's volleyball teams. By calculating the total annual medal scores, we obtained the performance of each country in each year's Olympic Games. In order to evaluate the actual effectiveness of Lang Ping's coaching, this article calculated the average medal scores of the Chinese team before and after 2016 and the American team before and after 2008, and compared them with their performance before coaching, obtaining the percentage change in scores. Finally, a bar chart was used to display the comparison of medal scores between the two teams before and after Lang Ping's coaching, further revealing the potential impact of coach changes on team performance. Figure 6 shows Lang Ping's medal scoring before and after coaching the two teams.

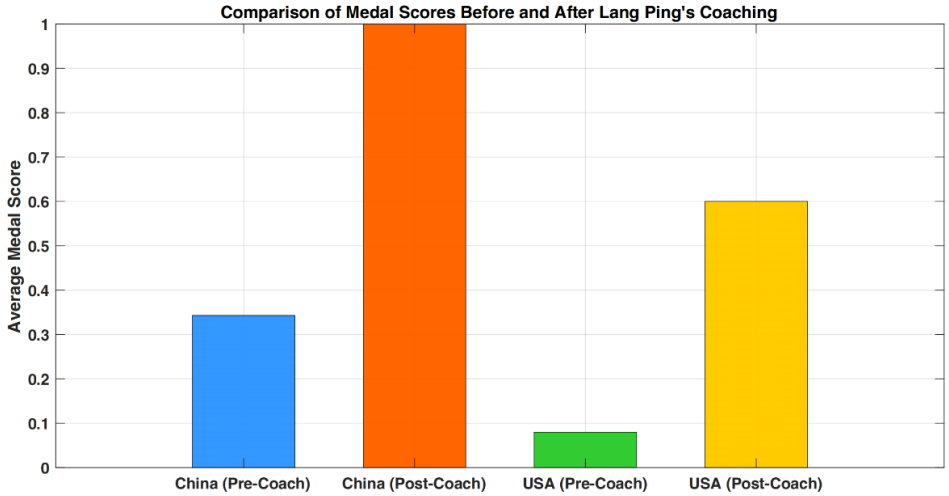

Figure 6: Comparison of Medal Scores Before and After Lang Ping's Coaching

From the above chart analysis, it can be concluded that Lang Ping's coaching has had a significant positive impact on both the Chinese and American women's volleyball teams. After her coaching, especially since 2016, the Chinese team has significantly improved their medal scores and performed outstandingly; After Lang Ping coached the US team in 2008, their medal scores increased significantly, demonstrating strong competitiveness.

In order to evaluate the impact of the "Great Coach" effect on medal count changes and provide coach introduction strategies for specific countries, this article first selects countries that have won medals in women's volleyball between 1988 and 2020 but failed to reach the podium in 2024. Subsequently, we identify the three teams with the highest medal scores during this period and based on this analysis, make a recommendation to invite 'great coaches' to provide guidance to these three teams that failed to win medals in 2024, in order to achieve future medal breakthroughs.

According to the analysis results, Brazil, China, and Cuba had high total and average medal scores between 1988 and 2020, indicating that these countries have strong volleyball capabilities. Although

they did not win a medal in 2024, their historical performance remains outstanding, whose stability and accumulated experience lay the foundation for future breakthroughs. Therefore, it is recommended that these countries further enhance their technical and tactical skills.

4. Conclusions

This study predicts the 2028 Olympics' medal situation per country. It analyzes 10 - Olympics' data, selects 11 features and 4 variables, and encodes classification features. Random Forest, LightGBM, and XGBoost models with 5 - fold cross - validation build feature - variable relationships. A stacking model combines their predictions. ARIMA and grey models predict 2028 features for medal forecasts. The "Great Coach Effect" is explored through Lang Ping's coaching. The research has high application feasibility for sports management and related aspects. For future development, incorporating more diverse data and further studying the "Great Coach Effect" are planned to improve the model's accuracy and understanding of sports success factors.

References

[1] Shi Huimin, Zhang Dongying, Zhang Yonghui. Can Olympic Medals Be Predicted? —— From the Perspective of Explainable Machine Learning [J]. Journal of Shanghai University of Sport, 2024, 48(04): 26 - 36.

[2] Wen Jing, Li Weiping, Lei Fumin. Research on the Prediction of Gold Medals and Medals Won by the Chinese Team at the Beijing Winter Olympics Using Multiple Methods [C]//Chinese Society of Sports Science. Abstracts of the 12th National Sports Science Congress - Symposium Reports (Sports Statistics Branch). School of Physical Education and Health, Hangzhou Normal University; Teaching and Research Section of Statistics, Xi'an Physical Education University, 2022: 3.

[3] Tian Hui, He Yiman, Wang Min, et al. Medal Prediction and Competing Strategies for Chinese Athletes at the 2022 Beijing Winter Olympics —— Based on the Analysis of the Home - Field Advantage Effect of the Olympics [J]. China Sport Science, 2021, 41(02): 3 - 13 + 22.

[4] Wang Jian. Research on the Prediction of the Number of Olympic Gold Medals Based on Regression and ARIMA Models [J]. Electronics World, 2018, (02): 46 - 47.

[5] Yuan Junjie. A Preliminary Exploration of the Olympic Gold Medal Prediction Model in the Big Data Era —— Taking the Results of the World Athletics Championships as an Example [J]. Bulletin of Sport Science & Technology, 2021, 29(06): 132 - 134.

[6] 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.

[7] Li Haiwei. Research on the Development Trend and Grey Markov Prediction of Track and Field Events in Previous Olympic Games [D]. Jiangxi Normal University, 2021.

[8] Paul Catumenn, Christopher Loch, Charlotte Kerchen, et al. The Impact of Coaching Changes on the Success of Professional Football Teams [J]. Journal of Beijing Sport University, 2019, 42(12): 61 - 76.

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

[10] Li Jiaqi. A Comparative Study on the Trend of Chinese Sprinting Results by Different Prediction Methods [D]. Tianjin University of Sport, 2023.