Research on Olympic Performance Prediction System Based on Multi-source Data Fusion

Hongbo Dou *

Dalian University of Technology, Dalian, China

* Corresponding Author Email: cs_ai_eie_aut@yeah.net

Abstract. This paper proposes an Olympic performance prediction system based on multi-source data fusion. First, through data cleaning and feature engineering, we constructed a dataset containing features such as athlete scale, number of events, and host country identification, and introduced an advantage score to quantify national competitiveness in different events. Random forest algorithm was used to predict the medal distribution of the 2028 Olympics, and a linear regression model was employed to quantitatively analyze the impact of elite coaches on medal acquisition. Experimental results show that the prediction system achieved an R^2 value of 0.716 on the test set, outperforming other machine learning models. Meanwhile, the study found that the coaching factor has a significant contribution to improving competitive performance, providing important reference for countries to optimize their coaching configuration.

Keywords: random forest, advantage score, Olympic medal distribution, great coach effect.

1. Introduction

With the vigorous development of the Olympic Games, the prediction and analysis of athletic performance has increasingly become a focus of attention for national sports management departments [1-3]. Traditional prediction methods mainly rely on single data sources and simple statistical models, making it difficult to cope with the multi-dimensional information explosion in the era of sports big data [4-5]. This limitation is especially evident in quantifying the impact of elite coaching expertise, exemplified by the success of coaches like Lang Ping in volleyball and Béla Károlyi in gymnastics across different national programs. The increasing complexity of Olympic competitions, coupled with the growing availability of diverse data sources, necessitates a more sophisticated approach to performance prediction and analysis.

This study introduces an innovative prediction framework that leverages multi-source data fusion and machine learning techniques. Our approach addresses three key research gaps: (1) the integration of heterogeneous data sources, including historical performance metrics, athlete demographics, and event-specific characteristics; (2) the quantification of coaching impacts through intervention analysis; and (3) the development of an automated data processing pipeline for Olympic performance prediction. This comprehensive approach enables a more nuanced understanding of the factors influencing Olympic success.

The research employs Random Forest algorithms to process complex, interconnected datasets while maintaining interpretability. This methodology represents a significant advancement over traditional statistical approaches, offering enhanced predictive accuracy and practical applicability. The resulting framework provides national sports organizations with a comprehensive decision support tool, enabling data-driven strategy development and resource allocation optimization. The findings from this study have broad implications for sports management and performance optimization across various competitive domains.

2. Data Preprocessing and Feature Engineering

2.1. Data pre-processing

The data source coms from https://www.comap.com. The initial analysis of the dataset revealed several data quality issues, including extraneous spaces, special symbols, missing values, duplicate

entries, and outliers. Data cleaning was performed using Python's Pandas library to address the missing values and duplicate records, ensuring data completeness. Outliers were identified and removed using the 3-sigma principle to minimize their potential impact on the model's predictive performance.

Initial data examination revealed inconsistencies in country identification across datasets, with some using country codes while others using country names. A mapping set was constructed to establish correspondence between country codes and names, facilitating data harmonization and subsequent merging operations. All country identifiers were standardized to country codes.

Following the standardization process, feature engineering involved the calculation of two key metrics for each country: the total number of participants (Sum_Athlete) and the total number of events (Sum_Event) from previous Olympic Games. Additionally, a binary variable 'is_Host' was introduced to indicate host country status (1 for host countries, 0 for non-host countries). The analysis was restricted to data from 2000 onwards to ensure model prediction accuracy.

Subsequently, medal performance analysis was conducted by calculating the total medals won by each country in each Olympic Games. A binary variable 'HasMedal' was created to distinguish between medal-winning countries (1) and non-medal-winning countries (0), enabling the identification of countries that had never secured an Olympic medal.

Further analysis involved computing an AdvantageScore for each country's primary events using the following methodology.

First, for each item j, the total number of medals for that item across all countries and at time t is calculated and expressed as follows:

$$\text{Medals}_{i,j} = \sum_t \text{Medals}_{i,j,t} \tag{1}$$

Then, for each country i and major event j, the proportion of medals won by that country in event j to the global total, denoted $\text{Share}_{i,j}$, is calculated as follows.

$$Share_{i,j} = \frac{Medals_{i,j}}{Medals_{world,j}} \tag{2}$$

If $\text{Medals}_{world,j} = 0$, then $\text{Share}_{i,j} = 0$, to avoid dividing by zero, the next, calculate:

$$A_i = \sum_{j \in \vartheta} w_j * Share_{i,j} \tag{3}$$

Where w_j denotes the weight of item j, set by the importance of the item and the total number of historical medals.

Finally, in order to limit the AdvantageScore to the range between (0,1), which is convenient for training the model, we used the maximum-minimum normalization.

$$\text{AdvantageScore}_i = \frac{A_i - \min_i(A_i)}{\max_i(A_i) - \min_i(A_i)} \tag{4}$$

Finally, this study combines the above features according to the national code (NOC) to create a feature set. As shown in Table 1.

Table 1. Feature Set

Symbols	*Definition*
$Sum\quad Event_{i,t}$	the total number of events in which country i competed at the Olympic Games in year t
$Sum\quad Athlete_{i,t}$	the total number of athletes from country i competing in the year t Olympic Games
$is\quad Host_{i,t}$	whether country i is the host country for the Olympic Games in year t
$HasMedal_i$	whether the country i has won at least one medal at a previous Olympic Games
$AdvantageScore_i$	the dominance score of country i on the main item and Target Variable

$$Is\_Host_i, t = \begin{cases} 1, & \text{If country i is the host country in year t} \\ 0, & \text{Otherwise} \end{cases} \tag{5}$$

$$Has\ Medal\ i = \begin{cases} 1, & \text{If the country i has won at least one medal in its history} \\ 0, & \text{Otherwise} \end{cases} \quad (6)$$

And the Target Variable shown in Table 2:

Table 2. Target Variable

Symbols	*Definition*
$Gold_{i,t}$	the number of gold medals won by country i in year t of the Olympics
$Total_{i,t}$	the total number of medals won by country i in year t of the Olympics

2.2. Random Forest

A random forest regression model was implemented to predict both the number of gold medals (Gold) and total medals (Total). The model was configured with 100 decision trees (n_estimators=100) for training. Subsequently, predictions were performed on the test set, with root mean square error (RMSE) serving as the primary evaluation metric [6-8].

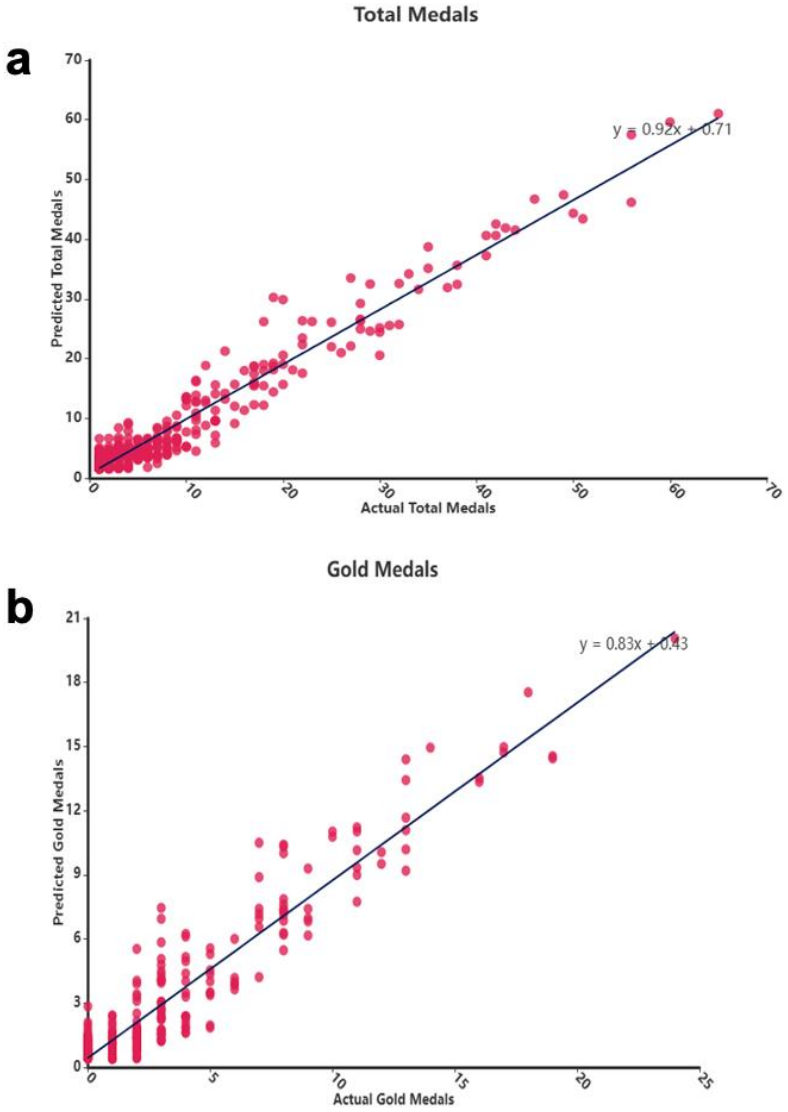

Figure 1. Scatter plot. (a) total medals; (b) gold medals.

This study used different nonlinear regression models to conduct the experiments, and the results are shown in Fig 2.

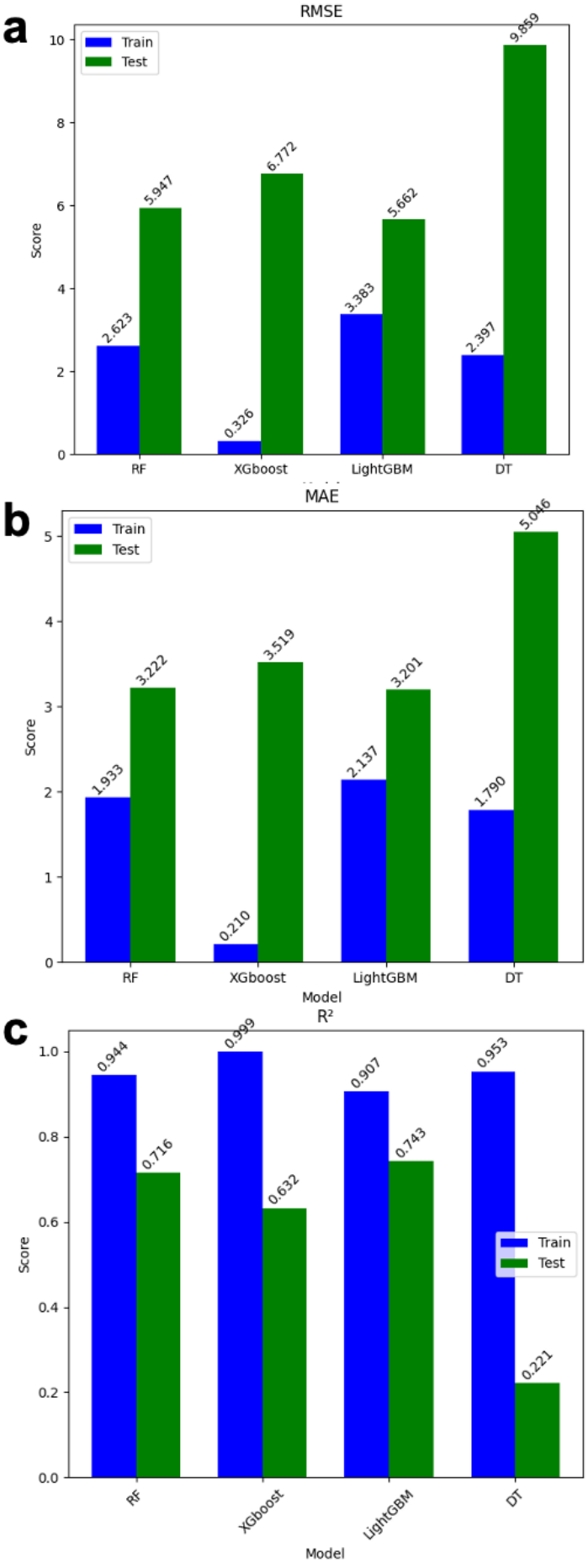

Figure 2. The results of the fitting of the different models. (a)RMSE. (b)MAE. (c)R^2

Analysis of model performance metrics revealed that Random Forest demonstrated superior performance with lower RMSE and MAE values, along with a higher test set R^2 of 0.716, indicating better capability in capturing complex feature relationships. While XGBoost exhibited exceptional performance on the training set with an R^2 of 0.999, it showed higher RMSE and MAE values, with significantly lower test set R^2 compared to Random Forest. Based on comprehensive evaluation of these metrics, LightGBM and Random Forest models demonstrated superior performance relative to Decision Trees and XGBoost. Given these comparative results, the Random Forest algorithm was selected as the final model for implementation.

3. Predictive Modeling of Olympic Medal Distribution

The random forest model was applied to predict both gold medal counts and total medal counts for participating countries in the 2028 Olympics. For input data preparation, information was sourced from official Olympic documentation, confirming Los Angeles, United States as the 2028 host city and identifying 49 confirmed sporting events. Given the uncertainty of other variables, three-year historical averages were utilized as proxy values for undetermined 2028 parameters.

To enhance the robustness of predictions beyond point estimates, interval predictions were generated using MAPIE with random forest ensemble methodology. The implementation involved wrapping the random forest model with MapieRegressor and training the wrapped model. Predictions were then generated on the test set with 95% confidence intervals (alpha = 0.05).

The resulting analysis produced projected gold medal and overall medal standings for the 2028 Los Angeles Olympics, complete with corresponding prediction intervals, as detailed in Table 3.

Table 3. Predicted Results

NOC	Pred_Gold	Gold_Cl_low	Gold_CI_high	Pred_Total_2028	Total_CI_low	Total_CI_high
USA	54.51	-26.99	162.01	153.44	33.61	373.26
CHN	51.87	26.13	77.601	135.68	36.64	176.61
GER	50.47	-8.33	109.27	106.03	13.65	198.40
FRA	34.43	-29.57	120.43	73.13	-7.99	214.25
AUS	21.48	0.09	66.86	62.36	16.28	182.43

Projections indicate that while the United States and China are expected to maintain their dominance in the gold medal standings, several smaller nations, including Albania, may emerge as contenders due to the addition of new events.

The analysis uses Total_diff as a key metric, calculated by subtracting each country's projected medal count for 2028 from their actual medal count in 2024. This metric serves as an indicator of a nation's Olympic trajectory: a positive Total_diff suggests progress, while a negative value indicates potential regression. Based on this analysis, 46 countries show potential for improvement, while 47 countries may experience a decline in performance.

When ranking countries by Total_diff in descending and ascending order, the data reveals the five nations with the most significant projected progress and the five with the most substantial projected decline in Olympic performance. As shown in Fig 3, 4.

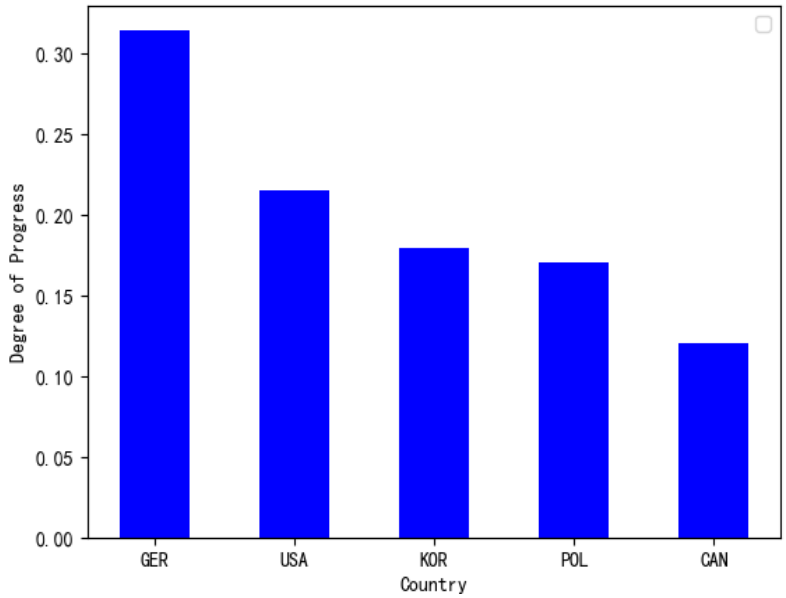

Figure 3. Comparison of the level of progress of the five countries that have made the most progress

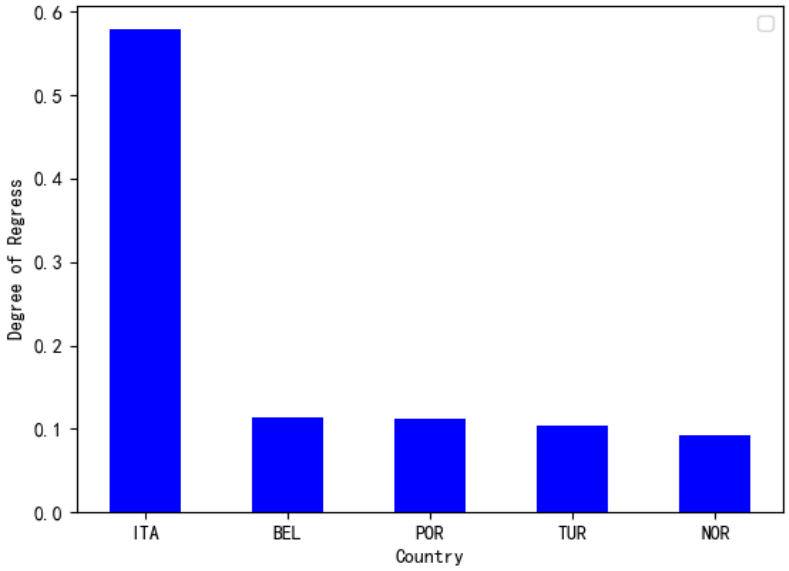

Figure 4. Comparison of the degree of regression of the five countries that have regressed the most

For nations that have never won an Olympic medal, a binary indicator HasMedal ∈ {0,1} is established to analyze their potential for achieving their first medal in 2028.

The probability of first-time medal acquisition is calculated for countries without previous Olympic medals. For instance, analysis indicates that Cape Verde has a 15% probability of securing its first Olympic medal in 2028.

The prediction methodology employs a Random Forest Classifier, which models Pit = P (HasMedali,t = 1|xit). This classifier enhances prediction accuracy by integrating multiple decision trees. In the Random Forest model, each decision tree is constructed independently and trained using randomly selected features and samples. The final classification is determined through a voting mechanism that combines the predictions of all individual trees.

The model training process utilizes historical data, where HasMedalit = (0/1) serves as the label while maintaining a consistent feature set. Through this approach, the probability of first-time medal acquisition in 2028 has been calculated for countries that have not previously won Olympic medals. As shown in Fig 5.

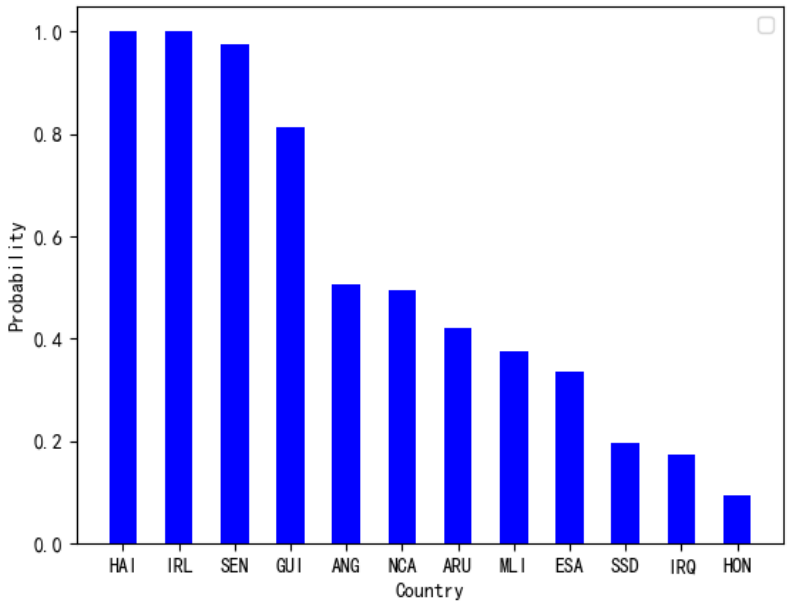

Figure 5. The probability that a country that has never won a prize will win it for the first time in 2028

4. Study on the Impact of Olympic Elite Coaches on Countries' Medal Counts

In a global arena such as the Olympic Games, the influence of coaches cannot be ig-nored. The ease with which coaches can move from country to country allows good coaches to utilize their expertise on a global scale, with far-reaching effects on the performance of sports in multiple countries.

This study will construct a mathematical model based on official data to quantify the contribution of "great coaches to the number of medals won by each country, and select three countries and their key sports to assess the potential impact of investing in "great coaches".

In order to find evidence of changes that may have occurred as a result of the "master-mind effect", we use the examples of two super-coaches, Lang Ping and Béla Károlyi.

This study first processed the tabular data to distill the volleyball medal trends in China and the United States. As shown in Fig 6.

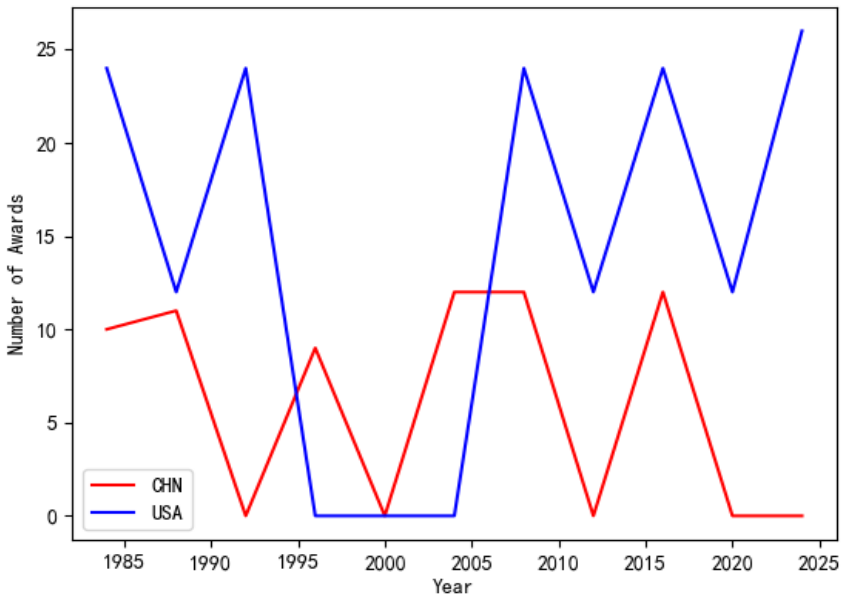

Figure 6. Comparison results

This study now analyzes the impact of gymnastics coach Bella Karolyi on the Romanian team and the U.S. women's gymnastics team. As shown in Fig 7.

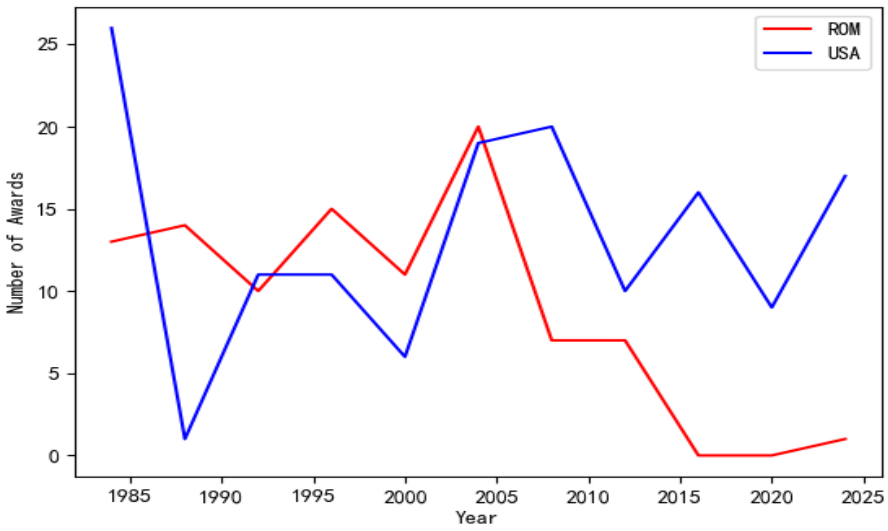

Figure 7. Comparison of awards in the women's gymnastics programs of Romania and the United States of America

The "Great Coach Effect is demonstrated through historical examples in Olympic sports. Bella's coaching tenure in Romanian gymnastics (1960s-1970s) and later with the U.S. Women's National Gymnastics Team (1981-2000s) shows significant medal count increases. Similarly, Lang Ping's leadership of both Chinese and U.S. volleyball teams illustrates this effect, particularly during her terms with the Chinese team (1995-2000 and 2013-present), culminating in Olympic gold.

However, the impact of elite coaches typically manifests gradually through skill development and team cohesion, rather than showing immediate results. This delayed effect often extends beyond their tenure, as evidenced by performance patterns in subsequent Olympic cycles. An OLS regression model was developed to quantify the relationship between elite coaches and medal acquisition.

Given the small amount of data and the need to quantify the impact of coaching, we chose a linear regression model with intervening variables.

A linear regression model with an intervening variable is a statistical model used to an-alyze the effect of an intervention (e.g., policy, event or condition) on an outcome variable. The intervening variable is usually a binary variable (0 or 1) indicating whether or not there is an intervention. The goal of the model is to quantify the independent effect of the inter-vention on the outcome by controlling for other variables. The strengths of the model are its simplicity, interpretability and good predictive power.

In this case, the intervening variable is "presence of a good coach (Coach) and the outcome variable is "total number of medals (Total_Medals). The model allows us to ana-lyze the specific contribution of good coaches to the number of medals. This study constructs a linear regression model as follows.

$$\text{Medals} = \beta_0 + \beta_1 \times \text{Year} + \beta_2 \times \text{coach\_dummy} + \epsilon \tag{7}$$

Where coach_dummy represents an intervening variable that indicates whether a "great coach has been hired (1 means that the coach has been hired, 0 means that the coach has not been hired).

$$\text{coach}_{\text{d}}\text{ummy} = \begin{cases} 1 & \text{if Year} \geq 2000 \\ 0 & \text{if Year} < 2000 \end{cases} \tag{8}$$

Let's assume that the "great coach started coaching in 2000.

Table 4. Symbols and Definitions

Symbols	*Definition*
Medals	the total number of medals won at the Olympics
Year	the year in which the Olympic Games were held
β_0	The intercept term
β_1	Coefficient of Year, indicating time trend
β_2	coefficients of coach_dummy
ϵ	the random error term

Through the regression model, this study can get the specific value of the coefficient of "great coach, which indicates the effect of the guidance of "great coach on the number of medals. Accordingly, this suty have drawn a bar chart of the influence of "great coach on the number of medals won by China, the United States and Romania. As shown in Fig 8.

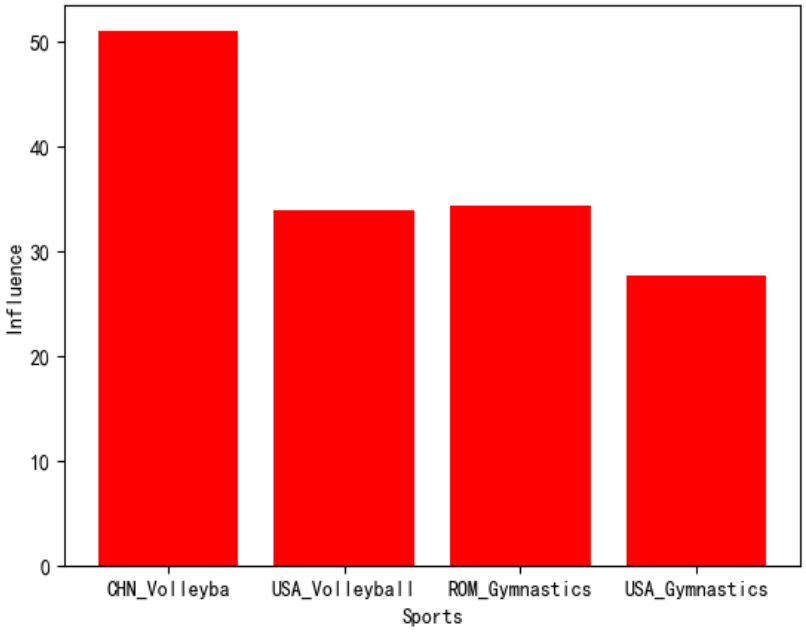

Figure 8. The extent to which the "great coach effect affects the number of medals won

It is easy to see that the "Great Coach effect has played a positive role in increasing the number of medals.

Analysis of elite coaching impacts, particularly in China's women's volleyball and U.S. women's gymnastics, reveals a distinct pattern. Nations typically seek elite coaches after experiencing extended periods of performance stagnation, despite previous success. These coaching appointments often follow a cycle where initial achievements are followed by prolonged periods of lower medal counts.

Research suggests that three consecutive years of stable medal performance in a particular sport indicates an optimal timing for elite coach recruitment to enhance competitive outcomes. Based on this criterion and historical data visualization, specific recommendations have been developed for three nations to optimize their athletic programs.

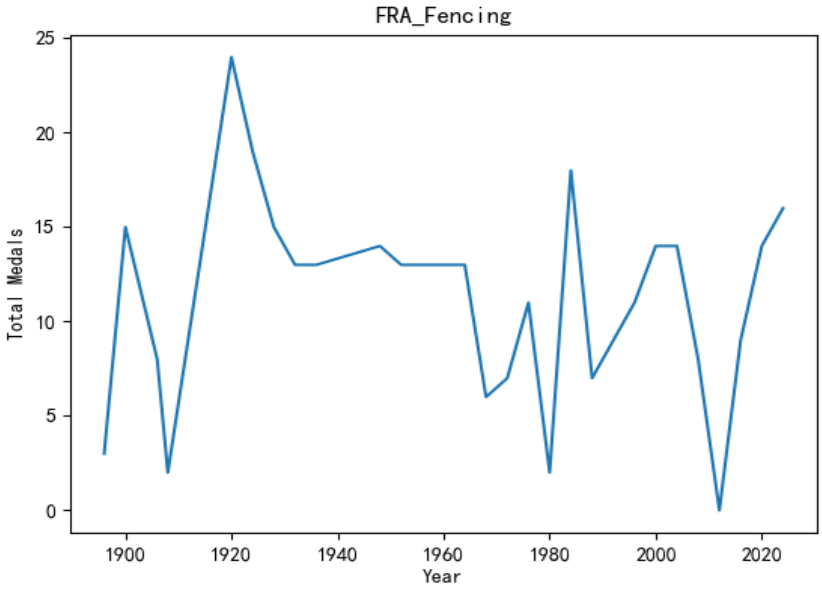

Figure 9. Trend of the number of prizes won by the French fencing program

As shown in Fig 9. France, with its rich fencing tradition, has seen declining and unstable performance in recent years [9]. Recruiting an elite coach could revitalize the program's competitive strength.

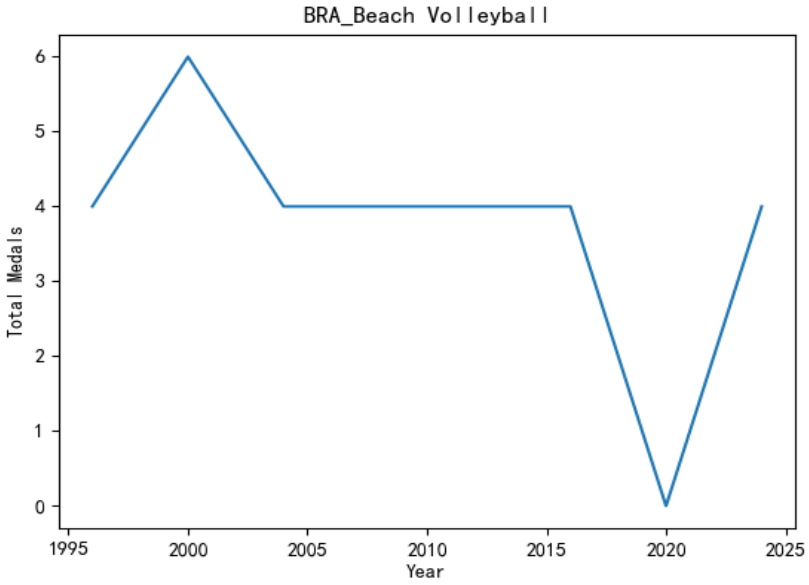

Figure 10. Brazilian beach volleyball awards chart

As shown in Fig 10. Brazil has displayed consistent strength in beach volleyball but faces increasing international competition and diminishing medal returns [10]. The addition of an elite coach could reinforce their dominance and enhance medal performance.

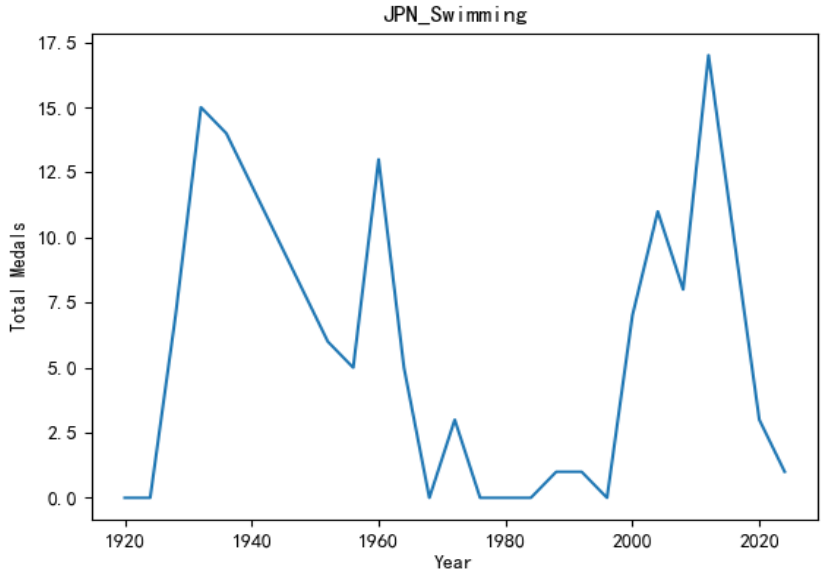

Figure 11. Trend of the number of prizes won by Japan's swimming program

As shown in Fig 11. Japan's swimming program shows potential but has experienced declining medal counts and struggles to develop elite athletes. An elite coach could strengthen talent development and improve medal performance.

Analysis of French fencing medal data from 1896 to 2024 using linear regression modeling demonstrates the significant positive impact of elite coaches on medal acquisition. The model results predict potential medal count increases following elite coach recruitment. This analytical framework can help identify optimal coaching investment opportunities for nations like France, Brazil, and Japan to enhance their Olympic performance through strategic coach recruitment. As shown in Fig 12.

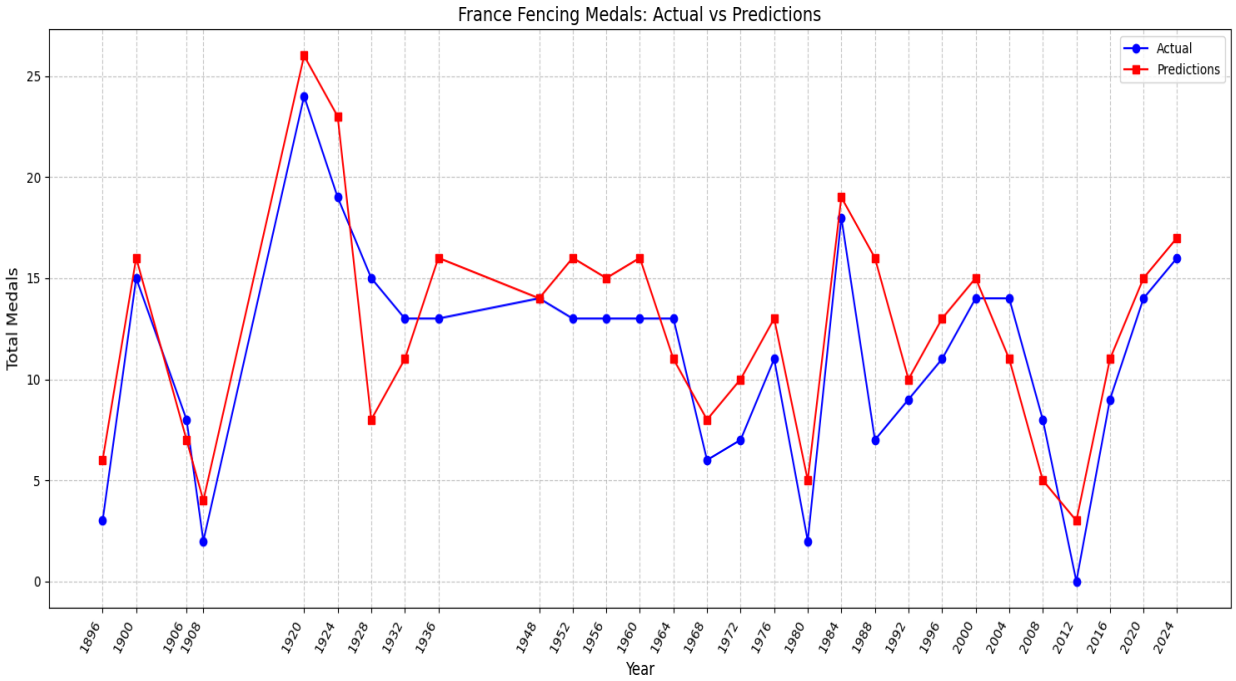

Figure 12. Number of awards in French fencing (coached vs. uncoached)

5. Conclusion

This study presents a novel approach to predictive modeling in sports performance analysis through the integration of machine learning techniques and historical data analysis. The research demonstrates the effectiveness of Random Forest models in capturing complex relationships between multiple

variables, achieving a robust test set R^2 of 0.716. Our feature engineering process, particularly the development of the AdvantageScore metric, provides an innovative framework for quantifying competitive advantages in multi-event competitions.

The methodology extends beyond traditional statistical analysis by incorporating confidence interval predictions through MAPIE implementation, enhancing the reliability and practical applicability of our predictions. Through systematic comparison of various machine learning models (including XGBoost, LightGBM, and Decision Trees), we established that Random Forest offers superior performance in handling the inherent complexity of sports performance data.

Furthermore, our research introduces a novel approach to quantifying the impact of expertise transfer through a linear regression model with intervention variables. This framework provides a statistical foundation for measuring the effectiveness of strategic personnel decisions in performance enhancement. The methodology developed in this study has broad applications beyond sports, potentially extending to any field where performance prediction and the impact of expertise need to be quantified.

Future research directions could explore the integration of deep learning techniques, the incorporation of additional contextual variables, and the development of more sophisticated intervention analysis frameworks to further enhance predictive accuracy and practical applicability.

References

[1] SAJADI S M, BAGHAIE S, REZAEI R. Optimizing sports development: Identifying and prioritizing key indicators for professional and competitive sports [J]. World Development, 2024, 180: 106651.

[2] MCCULLOUGH B P, ORR M, KELLISON T. Sport ecology: Conceptualizing an emerging subdiscipline within sport management [J]. Journal of Sport Management, 2020, 34 (6): 509 - 520.

[3] GREENWELL T C, DANZEY-BUSSELL L A, SHONK D J. Managing sport events [M]. Human Kinetics, 2024.

[4] EL-MAGHRABI Y, SHARIF M. Game Changers or Game Predictors? Big Data Analytics in Sports for Performance Enhancement and Fan Engagement [J]. Journal of Contemporary Healthcare Analytics, 2022, 6 (6): 19 - 39.

[5] ROSSI A. Predictive models in sport science: multi-dimensional analysis of football training and injury prediction [J]. 2017.

[6] HODSON T O. Root mean square error (RMSE) or mean absolute error (MAE): When to use them or not [J]. Geoscientific Model Development Discussions, 2022, 2022: 1 - 10.

[7] CHICCO D, WARRENS M J, JURMAN G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation [J]. Peerj computer science, 2021,7: e623.

[8] ROBESON S M, WILLMOTT C J. Decomposition of the mean absolute error (MAE) into systematic and unsystematic components [J]. PloS one, 2023, 18 (2): e279774.

[9] NASON A E. The Contribution of Domestic and International Conflict in Renaissance Italy to the Sport of Fencing [J]. 2023.

[10] MARQUES OLIVEIRA F, COSTA TELLES S, CARLOS NERY L, et al. Beach Volleyball Management in Brazil [J]. 2023.