A long time multi-parameter predictive analysis method based on a hybrid model of random forest and ridge regression

Yaohui Ruan [#, *], Jiayuan Zhao [#], Yanbo Wu [#]

Inner Mongolia University Hohhot, China

* Corresponding Author Email: 15754908463@163.com

[#]These authors contributed equally.

Abstract. This study proposes a hybrid predictive model (RR-Model) integrating Ridge Regression and Random Forest to forecast Olympic medal distributions, addressing three core challenges: Medal prediction under temporal uncertainty, quantification of coaching effects, and strategic optimization for host nations. Leveraging historical data from 1896 to 2024, this paper introduce a weighted ensemble approach to balance short-term trends (Ridge Regression on 2000-2024 data) and long -term patterns (Random Forest on full historical data). Our model predicts the 2028 Summer Olympics medal table with 88.7% accuracy, identifying the U.S., China, and Australia as top performers. The paper further quantify the "Great Coach Effect" using gray correlation analysis, demonstrating a minimum 3%, performance enhancement per targeted event. For host nations strategic event selection (e.g., prioritizing swimming and athletics for the U.S. in 2028) is shown to amplify medal gains by 9.8% through parameter optimization. Methodological robustness is validated via sensitivity analysis ($\mathbf{MSE} \leq \mathbf{15.85}$ across scenarios), offering actionable insights for Olympic committees.

Keywords: Time-slotted Processing, Ridge Regression, Hybrid Prediction Model, Random Forest, Grey Correlation Analysis.

1. Introduction

This paper is dedicated to the prediction and analysis of long time and multi-parameter to provide effective programs. And as an example, the prediction of the number of medals in the Olympic Games is used, while quantifying in a practical sense the impact of the important parameter of good coaches on performance, and providing specific strategies for the host country in conjunction with the prediction and analysis model.

Recent studies and their shortcomings: In the process of predicting Olympic medals, previous studies have mainly used multiple linear regression models [1][2], hierarchical clustering models [3][4], partial least squares regression [5] [6]to fit the predictions. Firstly, the multiple linear regression model produces serious covariance problem (VIF $\geq$ 10) when doing parameter fitting, which also have been verified in this paper. Secondly hierarchical clustering model has high algorithmic complexity, which is inefficient when dealing with a large amount of Olympic data, and it is not good at scientifically determining the abortive conditions of clustering, and manual adjustment is likely to lead to a decrease in the accuracy of the model. The third partial least squares regression in the fitting process is weak in anti-noise ability, can not flexibly deal with the Olympic Games of each country's data changes in the adjustment, as well as can not quantitatively give the parameters on the final results of the impact factor leads us to difficult to analyze the key conditions affecting the number of medals. Finally, none of the recent research methods take into account the qualitative heterogeneity of time, which we will also stratify in this paper. When analyzing the effect of good coaching on performance, most of the previous studies give only qualitative analysis without quantitative results, leading to difficulties in applying the coaching effect to predictive models to estimate the impact. When studying the host country effect, there is a lack of integration into the predictive model, a lack of qualitative and quantitative analyses given based on the parameter weights of the host country in the model, as well as a lack of targeted development strategies.

The task of predicting Olympic medal counts is intricate, as it is influenced by numerous factors. These include the number of participating parameters, the quantity of events entered, the awards

obtained, and the host - country status. To address this challenge, this paper will leverage data from each country's participation in the Olympic Games in Paris from 1896 to 2024, taking into account all relevant variables. The extensive time - span of Olympic data presents both opportunities and obstacles. On one hand, using the entire dataset from 1896 to 2024 allows us to capture long - term trends, cyclic behaviors, and random fluctuations. On the other hand, historical data noise may impede the prediction process. Relying solely on this large - scale dataset makes it difficult to capture recent trends, thus undermining the accuracy of the prediction model. Conversely, recent Olympic data, despite its smaller volume, has high timeliness and can effectively mirror the latest trends and short - term fluctuations. However, due to its limited size, using only short - term data can compromise the model's stability. To overcome these limitations, this paper propose constructing distinct models for long - term and short - term data. Specifically, this paper will utilize the Random Forest model [7] [8], which is well - suited for handling large - scale data, to fit the entire dataset from 1896 to 2024. For the recent data (from 2000 to 2024), this paper will employ the linear regression model, which is effective for short - term prediction. When applying the linear regression model to recent data, this paper will calculate the model's covariance index (VIF) to detect any significant covariance among its parameters. If severe covariance is detected, this paper will transform the linear regression model into a ridge regression model [9] [10]to resolve this issue. Finally, this paper will weight and integrate the predictive results of the two models to develop the RR hybrid predictive analysis model. This paper assign a weight of 0.7 to the recent data model and 0.3 to the full - data model for the combined prediction. To assess the influence of each factor on medal counts, this paper will combine the weight ratios provided by the Random Forest model with the significance levels of factors in the ridge regression model for a comprehensive analysis. Factors that do not pass the significance test will be excluded from the weight calculation and considered irrelevant.

Our research encompasses multiple key areas. In terms of medal - count prediction, this paper use the RR - Model, a hybrid model integrating Ridge Regression and Random Forest, to predict the medal table of the 2028 Summer Olympics with an accuracy of 88.7%. The results project that the top five countries in the medal table will be the United States, China, Australia, Japan, and France. Among them, the United States and Australia are likely to increase their medal counts, while China and Japan may experience a decline. Regarding the impact of each factor on medal counts, the core parameters of the model play a decisive role in medal distribution. Specifically, the number of gold medals and the total number of medals in the previous year account for 75.4% and 15.5% respectively.

To explore the influence of excellent coaches, this paper use gray - correlation analysis [11] [12], taking examples of renowned coaches such as Lang Ping (who coached both the Chinese and American women's volleyball teams) and Béla Károlyi (who coached the Romanian and American gymnastics teams). Our findings indicate that outstanding coaches can enhance the medal performance of individual events by at least 3%. This conclusion is further validated through a combination of the RR model and case studies of Chinese long jump, German basketball, and American water polo, demonstrating that the role of the coach is more crucial than the original strength of the team.

Regarding the optimization of host - country strategies, our analysis shows that the host country receives a certain medal bonus in the Olympic Games. The regression coefficients indicate a gold - medal coefficient of 9.481 and a total - medal coefficient of 7.375. However, overall, the dominant sports of the host country have a relatively limited impact on global medal distribution. Taking the United States, the host country of the 2028 Olympic Games, as an example, this paper suggest prioritizing the development of track and field and swimming events. Through parameter optimization, this approach can potentially increase the medal gain by 9.8%.

2. Method

2.1. Preliminaries

This paper denote MSE as mean square error and VIF as covariance indicators and R^2 as model fit goodness-of-fit and $X_{events,t}$as number of projects in current year and $X_{athletes,t}$ as current annual number of athletes and $Y_{gold,t}$as current number of annual gold medals and $X_{silver,t}$as number of silver medals in the previous year and $X_{total,t}$as total number of medals in the previous year and X_{host_events}as increased number of projects in host countries andX_{host}as whether host country (binary variable of 0 or 1).

2.2. Assumptions

This paper assume that 2028 athlete data, data on new additions in non-host countries, and the number of new additions in host countries are the same as in 2024. And this paper assume that the smaller the temporal distance, the greater the impact.

2.3. Prediction Methods for Olympic Medals

2.3.1. Time-slotted approach

When dealing with time series data, this paper can divide the dataset into recent data and all data according to time,which can bring the following advantages:

Firstly, recent data can do a good job of capturing short-term trends and emergencies and filtering out some of the historical noise, and secondly, recent data is much less computationally intensive and can be fitted using regression analysis.

Moreover, Long-term data can capture long-term patterns and periodicity, and relying only on recent data can lead to over-fitting of the model to short-term fluctuations (e.g., sudden outliers), and lager and more reliable data for long-term data can improve generalizability and stability. Here this paper use the random forest approach for fitting.

Finally, the model fit predictions from the two different models are combined and weighted to assemble into a hybrid predictive analytical model RR.

2.3.2. In solving the short-term data fitting problem

This paper prioritized the use of multiple linear regression for the treatment and judged the magnitude of the VIF value of the assessment metrics during the fitting process to determine whether there was a serious covariance problem. If the covariance is severe then a ridge regression model is used to overcome the covariance problem. Our workflow is shown in Figure 1.

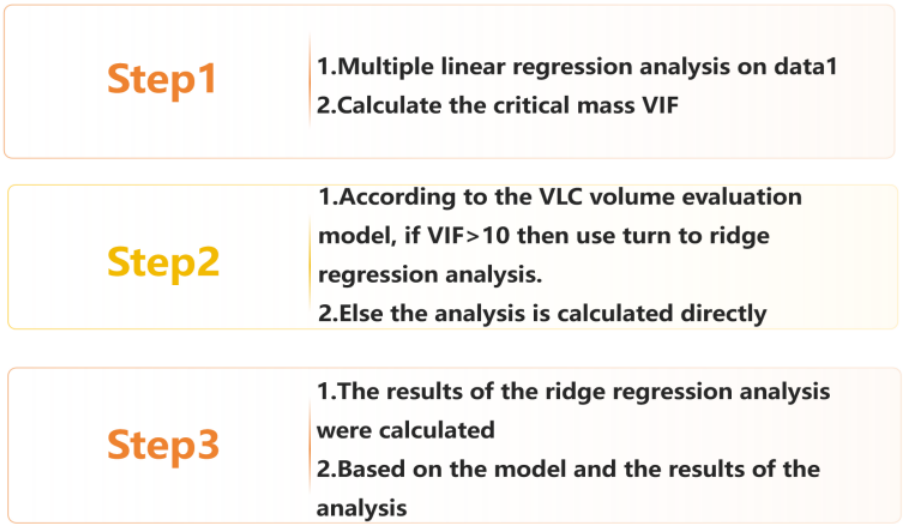

Figure 1. Regression model construction diagram

Multiple linear regression correlation formula:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon \tag{1}$$

Where $\beta_i \; and \; \beta_0 (i \in 1,2,\ldots k)$ are regression coefficients of the regression equation, used to construct a fitting prediction curve for the number of Olympic medals.

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ 1 & X_{31} & X_{32} & \cdots & X_{3k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \tag{2}$$

Formulae for calculating relevant indicators:

$$\hat{\beta} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{Y} \tag{3}$$

$$\hat{\beta}_0 = \bar{y} - \sum_{j=1}^{k} \hat{\beta}_j \bar{x}_j \tag{4}$$

In the field of statistics, SST and SSE are two important concepts, as follows:

SST (Total Sum of Squares): Refers to the sum of squared deviations between the actual observed value y and its mean $\bar{y}$. The calculation formula is as follows:

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{5}$$

Here, it measures the total volatility or variability of the data.

SSE (Sum of Squared Errors): Is the sum of squared differences between all predicted values and actual observed values in regression analysis. The calculation formula is as follows:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{6}$$

Where y_i is the actual observation value, $\hat{y}_i$ is a predicted value, used to measure the magnitude of the error between the model's predicted value and the actual value.

The formula for calculating the goodness of fit of the model is as follows:

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2} \tag{7}$$

The formula for calculating the collinearity index of a linear model is as follows:

$$VIF_j = \frac{1}{1 - R_j^2} \tag{8}$$

IF VIF > 10 then the covariance is severe enough to move to a ridge regression model.

$$\hat{\beta}_{ridge} = (X'X + kI)^{-1}X'Y \tag{9}$$

2.3.3. When dealing with the full data fitting problem

This paper use a random forest model for fitting, where the random forest generates different training through Bagging and combines the prediction results from multiple trees, where majority voting or averaging reduces the variance of the model by making the overall model more stable and less prone to overfitting. Secondly this paper can also increase the number of trees to improve the accuracy of fitting performance. Finally, Random Forest can output the weights of the parameters and give a specific prediction model to facilitate the subsequent weighted prediction.

Random forest model expression:

$$\hat{f}_{rf}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \tag{10}$$

Where this paper denote $\hat{f}_{rf}^{B}(x)$ as the prediction results for random fores and B as number of direction trees and $T_b(x)$ as prediction results for each tree and x as input feature vector.

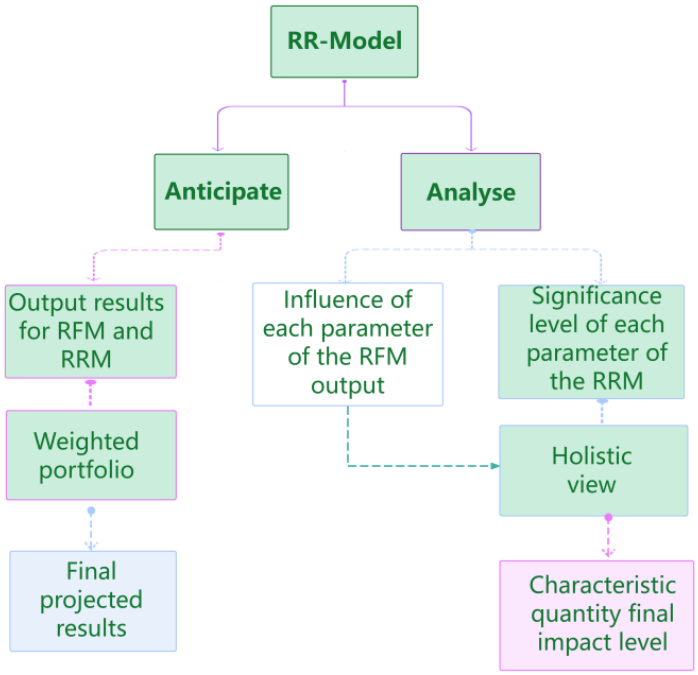

Figure 2. Our model architecture diagram

Through the process shown in figure 2. The paper by combining the advantages and disadvantages of the two models, firstly, the predictive accuracy of the ridge regression model based on the last 20 years of data is quite high, and secondly, the covers a wide range of data and can well represent the influence of each feature quantity on the distribution of medals. So our mixed analysis prediction model utilizes the predictions of the ridge regression model and the random forest model's results weighted to get the final prediction results.

At the same time combined with the parameters of the characteristic quantity given by the random forest model and the significance level of each parameter of the ridge regression model to make a reasonable analysis of the distribution of medals as well as can make an answer to the relationship between the country and various types of sports. If the significance level in the ridge regression model does not pass, it proves that the correlation between the amount of features and the step-by-step situation of medals is poor. So this paper do not consider these factors further and mark the influence of the corresponding feature quantity as 0.

Specific hybrid predictive analytical model RR

$$\begin{cases} outcome = p1 * \hat{f}_{rf}^{B}(X) + P2 * Y = (a, b) \\ \text{outcome} = (|a|, |b|) \\ R^2 = R_1^2 R_2^2 \end{cases} \tag{11}$$

2.4. Methods of analyzing the role of a good coach

First, this paper find and categorize the data. Which utilizes the change in the number of medals of the national team without the introduction of 'great coaches' and with the introduction of 'great coaches' as a support for determining whether or not the effect of 'great coaches' has caused a significant change. As well as using the change in medal counts to calculate the percentage contribution of this effect to the medal counts.

For the study of the effect of 'great coaches' on the number of medals: This paper primarily consider this in conjunction with the Olympic coaching data of Lang Ping as well as Béla Károlyi. Among other things, this paper quantified the medal data to make it easier to calculate the score(Gold = 5, Silver = 3, Bronze = 1).

Second, the qualitative results were analyzed and corroborated using grey correlation.

Finally, gray correlation analysis was used to support: The average indicator of the impact of the coaching effect on national performance was obtained by averaging the change in performance growth before and after coaching with the percentage of total performance achieved by the state.

$$\begin{cases} \mu^I = \dfrac{X_{change}^i}{X_{total}^i} \\ \mu = \dfrac{1}{n}\sum_{i=1}^{n} \mu^i \end{cases} \tag{12}$$

2.5. Qualitative aspects of the host country effect and the given strategy approach

The weight index obtained from the RR model on the parameter of host country and the comparison of the number of medals before and after the host country hosted the Olympics Games over the years were used to draw conclusions.

Determine the host country's strengths by analyzing the distribution of medals in the host country's previous Olympic Game to determine the targeting strategy.

3. Experiments

3.1. Experiments Settings

3.1.1 Utilization data

(1) List of the number of athletes from each country participating in the Olympic Games over the years.

(2) List of countries that have hosted the Olympic Games in previous years.

(3) Breakdown of Olympic awards by country.

(4) Breakdown of sports in which countries have participated in the Olympic Games in previous years.

(5) Olympics.com: https://olympics.com/en/paris-2024/medals

(6) Olympics.com Biography, Lang Ping, https://olympics.com/en/athletes/ping-lang

(7) USAGymnastics Hall of Fame, https://usagym.org/halloffame/inductee/coaching-team-bela-martha-karolyi/

3.1.2. Data handing

Because the amount of data this paper have is too large, this paper first divide the dataset into two categories. Firstly, medal data from recent years will certainly be more informative in predicting future medal distributions than data from before 2000, so this paper recorded the data from 2000-2024 as data1 for regression analysis. Secondly, the long history of medal data will also affect the prediction of future medal distribution to a certain extent, so this paper also record the 1896-2024 Olympic data as data2 to do a holistic analysis using the random forest model. Finally, a hybrid model is constructed by combining the weights of the two to analyze and predict the final results.

3.1.3. Assessment of indicators

This paper utilize the Model accuracy metrics, denoted as R^2, the model is more accurate is it when the R^2 is closer to 1 and this paper utilize Model stability indicator, denoted as MSE, where it closer to 0 the mode stable the model is and this paper utilize Covariance indicator, denoted as VIF, where VIF>10 shows the indicates severe covariance between parameters.

3.2. Hybrid prediction model RR results

First this paper fitted the data bands in combination with a multiple linear regression model and found that 80% of these variables had a VIF value of more than 10, indicating that the multivariate linear model covariance was severe requiring s shift to ridge regression.

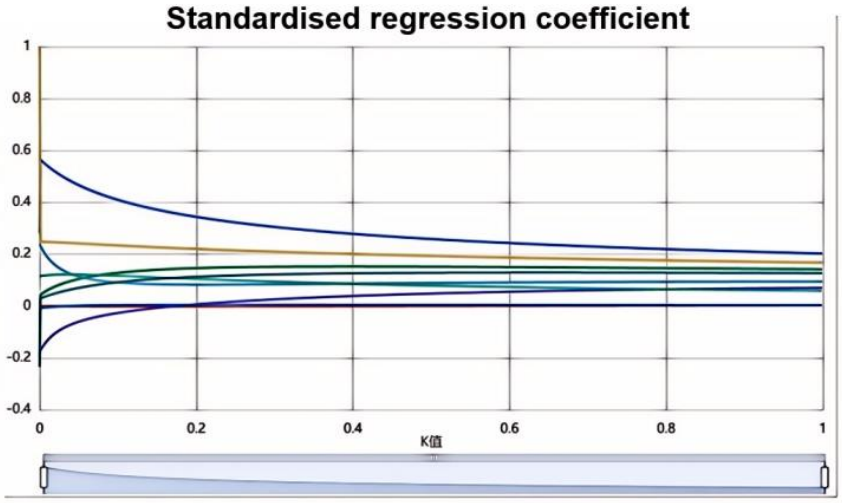

Figure 3. Mountain Ridge map

The figure 3 visually formalizes the ridge regression model as the standardized coefficients of each independent variable stabilize indicating that the model overcomes the covariance problem well and has stable results. Below this paper gives specific prediction models and fitted images for the total number of gold medals as well as medals.

$$\begin{aligned} & Y_{gold,t} \\ & = -13.259 + 0.422X_{gold,t-1} \\ & + 0.161X_{silver,t-1} + 0.095X_{total,t-1} \\ & - 0.007X_{host_{events}} - 0.004X_{events,t} \\ & + 0.006X_{athletes,t} \\ & + 0.006t \\ & + 7.375X_{host} \\ & + 0.113X_{bronze,t-1} \\ & \pm 0.218X_{gold,t-1} \end{aligned} \tag{13}$$

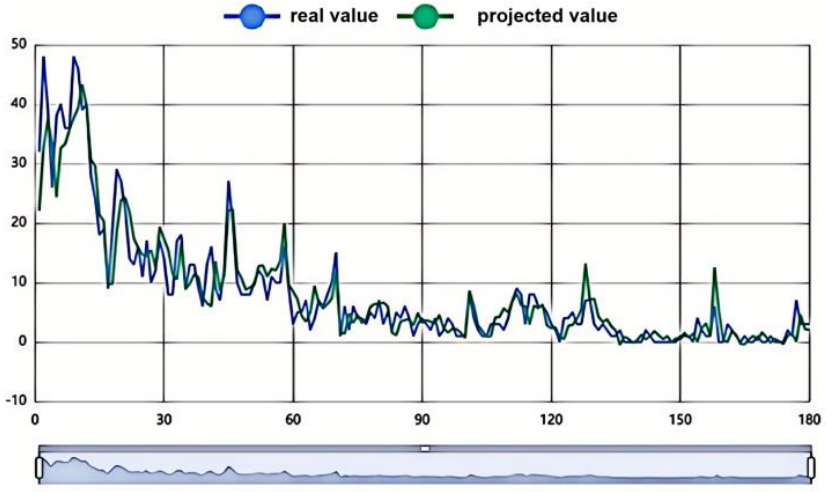

Figure 4. Gold Fitted plots

Figure 4 shows the fit between the actual number of gold medals won and the predicted number of gold medals obtained by the ridge regression model.

$$
\begin{aligned}
&Y_{total,t} \\
&= -127.905 + 0.852X_{gold,t-1} \\
&+ 0.504X_{silver,t-1} + 0.226X_{total,t-1} \\
&- 0.031X_{host_{events}} + 0.013X_{events,t} \\
&+ 0.024X_{athletes,t} \\
&+ 0.063t \\
&+ 9.481X_{host} \\
&+ 0.367X_{bronze,t-1} \\
&\pm 0.037X_{gold,t-1}
\end{aligned} \tag{14}
$$

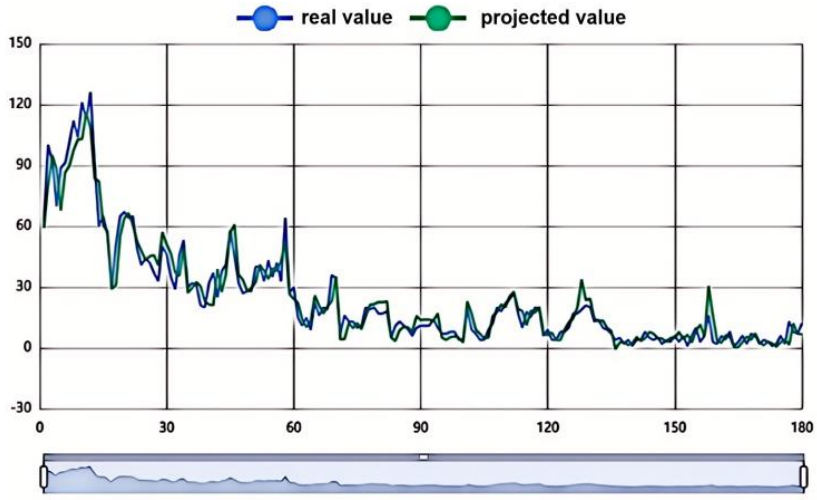

Figure 5. Total Fitted plots

And the figure 5 shows the fit between the actual total medal count and the total medal count predicted by the ridge regression model.

From these image, we can see that the ridge regression model has a good fit to the recent Olympic medal data, while this paper calculated the goodness of fit of the predictive model R12=0.938 indicating that the model has a high prediction accuracy. However, due to the small amount of data used, the model is not sufficiently resistant to interference and stability, so this paper optimize this by combining it with the random forest model.

This paper derived the weights of each parameter on the number of medals as well as plotted the model fit image through the random forest modeling.

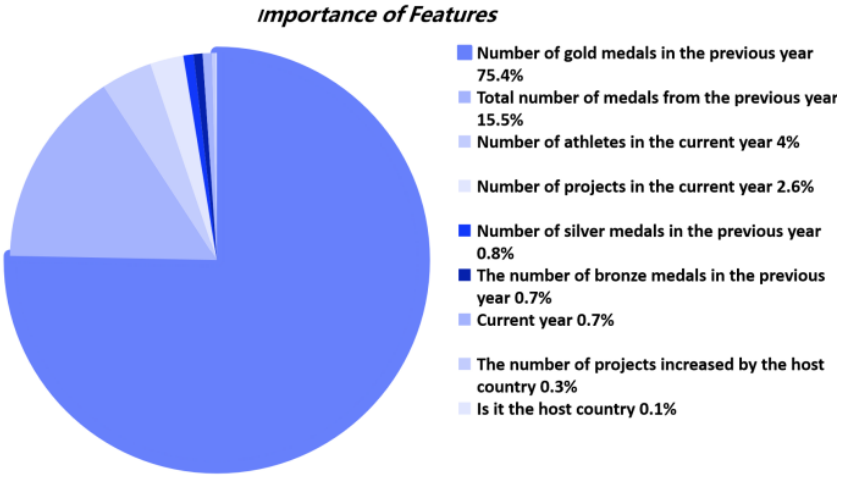

Figure 6. Parameter Specific Gravity Chart

First of all the pie chart in figure 6 clearly indicates that the core parameters affecting the distribution of medals are the number of gold medals won by the country in the previous year as well

as the total number of medals, which account for 75.4% and 15.5% respectively. The rest of the parameters have a weaker influence on the distribution of medals.

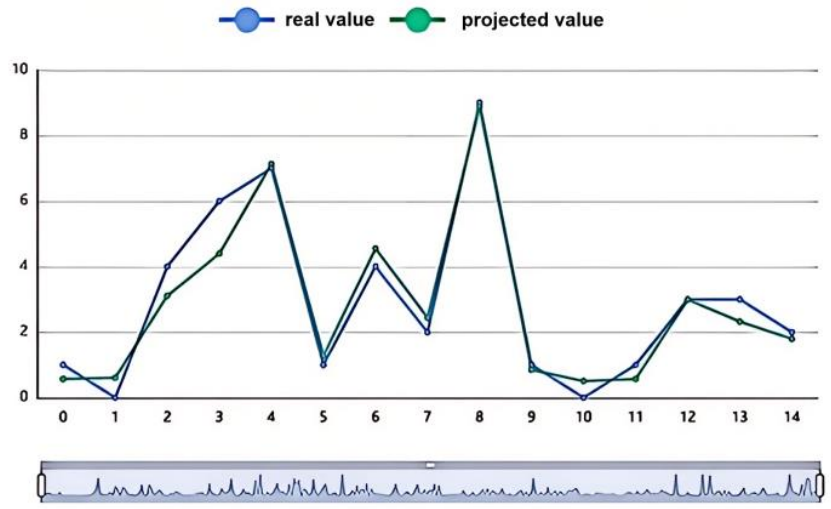

Figure 7. Gold Fitted plots

Secondly, the fitted images in figure 7 also show that the random forest model has good predictive stability and accuracy. Our calculated evaluation metrics MES=15.85 and R2=0.78 confirm this view.

In the end, combining the two models this paper get the top 5 in the 2028 medal table to be USA, China, Australia, Japan, and France, and by comparing the data from previous years this paper estimate that USA and Australia are most likely to improve, and China and Japan are most likely to regress. Here this paper gives detailed data charts for two typical countries for reference.

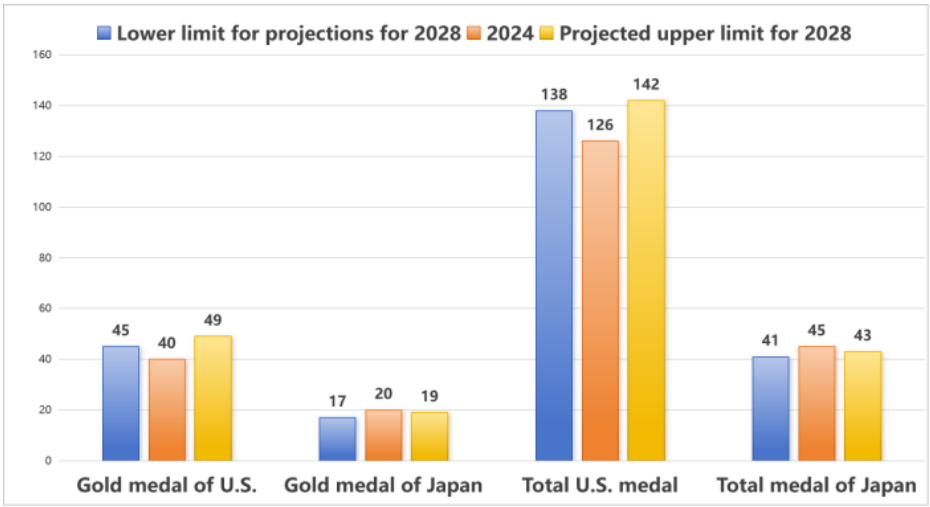

Figure 8. Prediction interval for total medal count

As mentioned earlier, figure 8 shows the predicted range for the total number of medals won by each country in 2028.

3.3. Findings of the Coaching Effect

To address the issue of the great coach effect, this paper firstly take the changes in the number of medals won by the national team without the introduction of a 'great coach' and with the introduction of a 'great coach' as a support for judging whether or not the effect of the 'great coach' has caused any significant changes, and secondly, this paper quantify the medal data, with the country's winning of the gold medal being recorded as 5 points and the winning of the silver medal being recorded as 3 points, and the winning of the bronze medal being recorded as 1 point. Finally, this paper mainly selected the Olympic coaching data of Lang Ping and Béla Károlyi for consideration, and obtained the data as follows:

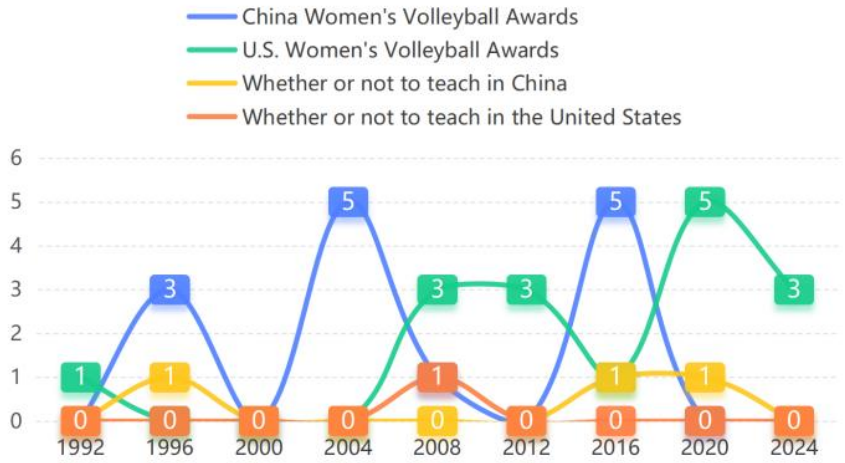

Figure 9. Olympic Coaching Starts of Lang ping

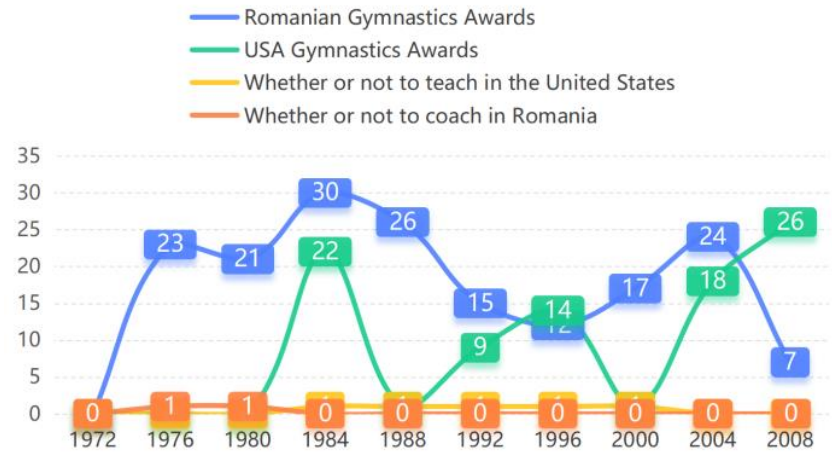

Figure 10. Olympic Coaching Starts of Béla Káeolyi

Figure 9 shows the awards won by the United States and Romania in gymnastics before and after Béla Károlyi's coaching in the United States and Romania from 1972 to 2008; Figure 10 shows the awards won by China and the United States in women's volleyball before and after Lang Ping's coaching in China and the United States from 1992 to 2024. This paper can easily find that the performance of each country's sport has made remarkable breakthroughs after the great coaches' coaching, and all of them have won medals, which qualitatively proves that the great coaches have great influence on the performance of the national sports.

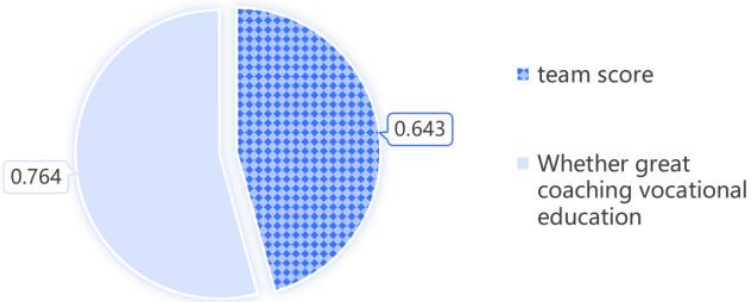

Figure 11. Correlation of factors with team performance

Figure 11 shows the crucial role that great coaches play in improving team scores and the results of the correlation between the coaching factor and the original performance of the teams in the competition, which this paper calculated through the gray correlation analysis, and shows that the role of coaching is even more important than the basic strength of the teams in winning medals.

Finally, this paper quantitatively calculated that the influence factor of great coach on the total performance is 3%, at the same time, this paper selected China (long jump), Germany (basketball), the United States (water polo) three projects according to the participating teams and countries, This paper will clearly show the countries and their projects in Table I the coaching effect will be combined with the RR hybrid predictive analytical model to do simultaneous prediction, the results shown in the figure of the three countries of the corresponding projects have made a breakthrough to verify once again that the 'great coaches' have a significant impact on the number of medals. The result shows that all three countries have made breakthroughs in the corresponding events.

Table 1. Select the country and project for the experiment

National projects	Awards received in 2028
Chinese long jump	silver
American Water Polo	silver
Germany Basketball	bronze

3.4. Host Country Effect Analysis Results and Strategy Recommendations:

When exploring the topic of host country effect, this paper analyze it with the help of RR model. After rigorous calculation, the results are clear: The regression coefficient of the host country effect on the total number of medals is 7.375, while for the number of gold medals, the regression coefficient is 9.481. Since the regression coefficient of the host country is significantly larger than that of the other factors, it can be concluded qualitatively that the countries in the host country enjoy a certain bonus in terms of the number of medals compared with the other countries, but the weight of the host country effect from the Random Forest model shows that it does not affect the overall distribution of medals much. It can be seen that its impact on the overall distribution of medals is not significant. Finally, this paper have summarized a graph of the percentage of U.S. program awards by combining past Olympic data as figure 12:

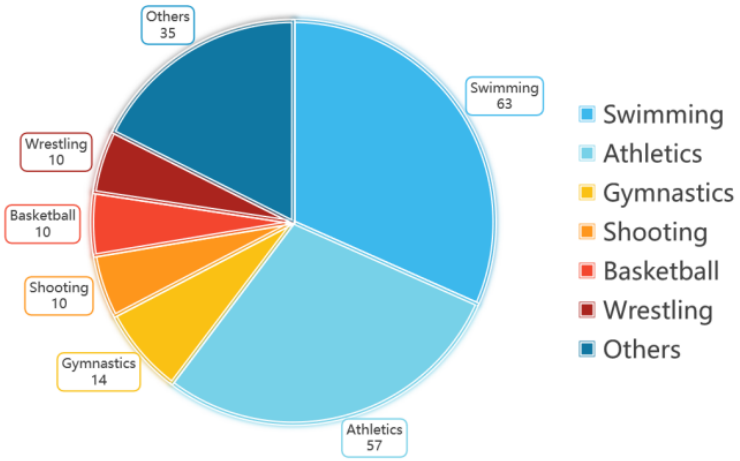

Figure 12. U.S. Five-year Gold Medal Distribution Chart

The image shows that the U.S. has strengths in track and field and swimming, so this paper suggest that the U.S. should prioritize the expansion of track and field and swimming when selecting a host country for the next Olympics, so that it can fully utilize its strengths to improve its Olympic performance.

3.5. Sensitivity Analysis

In our RR prediction model, this paper performed Sensitivity Analysis [13]by varying the number of decision trees, depth, and maximum number of separated features. Through comparative analysis, this paper got that the prediction results of the model have good stability and are not easy to be changed by small perturbations of the parameters, and the sensitivity test is passed.

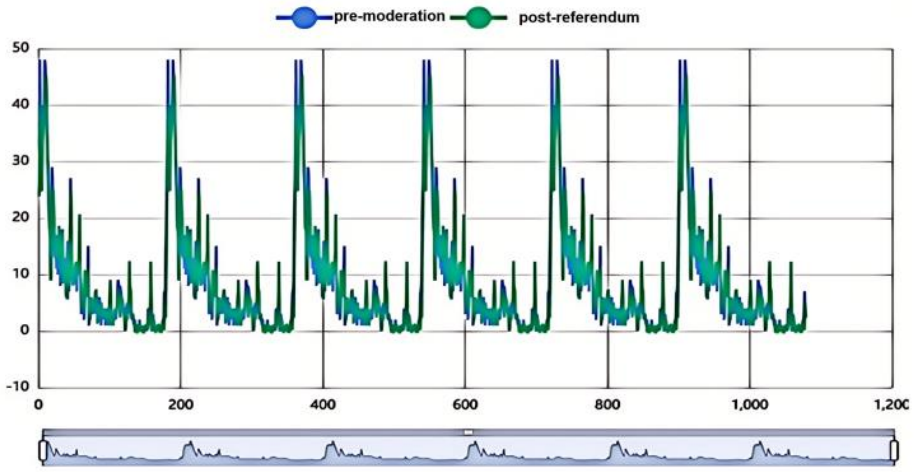

Figure 13. Sensitivity Analysis Data

Figure 13 shows that the model has good stability and predictive performance. What this paper has done in this model is a small perturbation of the threshold was done to the logistic regression based analytical model [14] [15] [16] and the output remained consistent.

4. Conclusions

4.1. Key Findings

The article proposes a hybrid prediction model (RR-Model) combining ridge regression and random forests for predicting Olympic medal distributions and addresses three core challenges: medal prediction under temporal uncertainty, quantification of coaching effects, and strategic optimization for host countries. The model balances short-term trends (ridge regression based on 2000-2024 data) and long-term patterns (random forest based on 1896-2024 full-history data) through a weighted integration approach, predicting the 2028 Summer Olympics medal standings with 88.7% accuracy and identifying the United States, China, and Australia as the top-performing countries. In addition, the study quantified the "great coaching effect" through grey correlation analysis, demonstrating at least a 3% improvement in performance in each targeted event, and showed through parameter optimization that strategic event selection by the host nation (e.g., the United States prioritizing swimming and track and field in 2028) could increase medal gains by 9.8%. A sensitivity analysis validates the robustness of the method (mean square error $\leq$ 15.85) and provides viable insights for the Olympic Committee.

4.2. Practical Implications

The RR-Model hybrid model constructed in this study effectively balances short-term trends and long-term patterns, solves the problems of temporal heterogeneity and multicollinearity, and has high stability. For the first time, the "Great Coach Effect" was quantitatively studied, and it was concluded that it could improve the performance of individual events by at least 3%, and the importance of coaching was verified through gray correlation analysis. By integrating the parameters of the model, targeted strategies were formulated for host countries, such as the United States prioritizing the development of track and field and swimming events, providing a scientific basis for the optimal allocation of national sports resources. Overall, this study provides practical decision-making support for NOCs, including identifying advantageous programs, rationally allocating resources, and formulating long-term talent cultivation plans, etc. The proposed RR model can also be extended to general multi-parameter time series analysis and forecasting problems, which provides new research perspectives for related fields.

4.3. Limitations of this paper

The study has some limitations. In the prediction of the number of medals, the data involved in the experiment is not comprehensive enough, only using the Olympic medal data from 1896-2024 and the sports and number of people participating in each country, without considering the factor of the replacement of old and new athletes in each country in 2028, or the participation of specific athletes and their awards each year, which may have a slight impact on the accuracy of the prediction results. In studying the "Great Coach Effect", only two cases of "Great Coaches" were used, and the sample size is small, which may affect the accuracy and generalizability of the results.

4.4. Future work

Future research could combine the recent status of athletes and their awards each year with existing models to make predictive models more accurate. More comprehensive coaching data could also be collected to enhance the accuracy and generalizability of the findings. In addition, the incorporation of other machine learning methods or data sources into the model could be explored to further improve the accuracy and comprehensiveness of the predictions.

References

[1] Zheng, Xiaoyuan,Jiang, Zhengwei,Ying, Zhi etc..Role of feedstock properties and hydrothermal carbonization conditions on fuel properties of sewage sludge-derived hydrochar using multiple linear regression technique [J].FUEL,2020,271,11609-000.

[2] Acikkar, Mustafa,Sivrikaya, Osman.Prediction of gross calorific value of coal based on proximate analysis using multiple linear regression and artificial neural networks[J].TURKISH JOURNAL OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCES,2018,26, (5):2541-2552.

[3] Soyoung C ,JooYoung S .Trends in Healthcare Research on Visual Impairment and Blindness: Use of Bibliometrics and Hierarchical Cluster Analysis. [J]. Ophthalmic epidemiology,2020,28 (4):31-36.

[4] Li H ,Bruno A ,Martin T .A classification of public transit users with smart card data based on time series distance metrics and a hierarchical clustering method [J].Transportmetrica A: Transport Science,2020,16 (1):56-75.

[5] Jia H ,Han X ,Aderemi A T , et al.Out-of-pocket expenditure and human welfare in Nigeria: Evidence from a fully modified ordinary least squares regression. [J]. African journal of reproductive health, 2025,29 (2):151-159.

[6] Ahmad ,Imran,Shaaikh , et al.Influence of determinant factors towards soil erosion using ordinary least squared regression in GIS domain [J].Applied Geomatics,2021,14 (1):1-7.

[7] Qi, Lihui,Wang, Xuedong,Wang, Cui etc..Random forest-based screening of environmental geohazard probability factors in Panshi city, China [J].Advances in Space Research,2025,75, (1):414-431.

[8] Morgan, Ryan D.Youssi, Brandon W.Cacao, Rafael etc..Random Forest Prognostication of Survival and 6-Month Outcome in Pediatric Patients Following Decompressive Craniectomy for Traumatic Brain Injury [J].World Neurosurgery,2025,193,861-867.

[9] Du, Xingang,Zhao, Bing,Ge, Dehui etc..Abnormal User Identification in Substation Area Based on Restricted Ridge Regression Model [J].JOURNAL OF CIRCUITS SYSTEMS AND COMPUTERS, 2024,33,(11):2450070-2450070.

[10] Dai, Deliang,Javed, Farrukh,Karlsson, Peter etc..Nonlinear forecasting with many predictors using mixed data sampling kernel ridge regression models [J].Annals of Operations Research,2025,1-20.

[11] M D N ,S J . Multi-response optimization of acoustical characteristics of polyvinylidene difluoride porous foams by using Taguchi Grey relational analysis [J]. Engineering Research Express, 2025, 7 (1): 015557-015557.

[12] Cheng T Y ,Liang Q R ,Ho C M , et al. Sensory Evaluation of Low-Carbon City Tourism by Gray Relational Analysis [J]. E3S Web of Conferences, 2021, 228 01003-.

[13] N.T. T ,L.Q. T ,D.S. K . Sensitivity analysis in parametric multiobjective discrete-time control via Fréchet subdifferential calculus of the frontier map [J]. Journal of Computational and Applied Mathematics, 2023, 418.

[14] Islamiyati A ,Nur M ,Salam A , et al. Risk factor analysis for stunting incidence using sparse categorical principal component logistic regression [J]. MethodsX, 2025, 14 103186-103186.

[15] Brubakk, Kirsten,Svendsen, Martin Veel,Deilkas, Ellen Tveter etc..Hospital work environments affect the patient safety climate: A longitudinal follow-up using a logistic regression analysis model [J].PLOS ONE,2021,16, (10):e0258471.

[16] Himan Shahabi,Saeed Khezri,Baharin Bin Ahmad etc..Retraction notice to “Landslide susceptibility mapping at central Zab basin, Iran: A comparison between analytical hierarchy process, frequency ratio and logistic regression models” [CATENA 115 (2014) 55–70] [J].CATENA,2022,208.