# Olympic Medal Prediction Analysis Based on The Random Forest Model

Yuhui Li *, Yuxuan Zeng

School of pharmacy, Shenyang Pharmaceutical University, Benxi, China, 117004

* Corresponding Author Email: 15507090029@163.com

**Abstract.** During the 2024 Paris Olympics, both individual events and the overall medal tally have garnered significant attention. As a key indicator of a nation's competitive strength, the number of Olympic medals serves as a critical reference for countries preparing for the Games. This paper addresses three core questions: how to predict the medal rankings for the 2028 Los Angeles Olympics, the likelihood of countries that have never won medals securing their first, and the impact of Olympic events on medal rankings. By combining the Random Forest algorithm and the LSTM model to predict feature variables (such as gender ratio and number of participants), a high-precision medal prediction model was constructed, achieving an $R^2$ score of 0.929 and an MSE of 1.78098. For countries that have never won a medal, nine countries were selected, and their probability of winning a medal in 2028 was predicted. Additionally, the Random Forest Regressor model was used to analyze the importance of sports events, identify key events for each country, and reveal the strategic impact of host country event selection. The conclusions indicate that this model provides a scientific basis for countries to optimize their Olympic training strategies and resource allocation.

**Keywords:** Random Forest Model, LSTM, Olympic Medal Prediction.

## 1. Introduction

The Olympic Games, organized by the International Olympic Committee, are international games that include a variety of sports and are held every four years. It originated in Ancient Greece and was named after Olympia, where it was held. In 1894, the Frenchman Coubertin initiated the establishment of the International Olympic Committee, which started the modern Olympic Movement. The Olympic Games is not only a competitive event for global sports elites, but also a symbol of cultural exchanges and peace and friendship among countries, demonstrating mankind's pursuit of excellence, unity and fair competition, and attracting the attention of a large number of sports enthusiasts. In addition to watching the individual events on the Olympic schedule, fans also pay close attention to the overall medal standings for each country. The top of the standings are usually the ones that get all the attention, but the medal counts of other countries are just as significant - for example, there are countries that won their country's first medal at the Paris Olympics, and there are still more than 60 countries that have yet to win a medal for their participation. In addition, the ranking of the Olympic medal table is not static, it changes according to the adjustment of national sports strategies [1], the renewal of athletes, and the increase or decrease of competition events [2]. To assist countries in better preparing for the 2028 Los Angeles Olympics, this paper raises three core questions: how to predict the medal rankings for the 2028 Los Angeles Olympics, the likelihood of countries that have never won medals before securing their first medals, and the impact of Olympic events on medal rankings. Previously, most predictions of Olympic medals were based solely on time-series analysis, such as the predictions made by Cheng Hongren et al. for China's performance at the Tokyo Olympics [3]. However, Olympic medal rankings are influenced by numerous factors, and predictions based solely on time-series relationships have low reliability. This paper uses data from the International Olympic Committee's website covering all Summer Olympics from 1896 to 2024 as its dataset, selecting variables such as year, athlete gender ratio and number, host country, type and number of sports events as feature variables, and uses an LSTM model to predict the basic information of participating countries for the 2028 Olympics. Finally, a Random Forest model is employed to predict the 2028 Olympic medal standings and analyze the likelihood of countries that have never

won medals securing their first medals. Additionally, a Random Forest Regressor model is utilized to analyze the impact of Olympic events on medal rankings.

## 2. Forecasting the 2028 Los Angeles Olympics Medal Table

Since there is no official data available for the 2028 Los Angeles Olympics and only the venue is known, it is necessary to first predict the above feature variables and then use the predicted values of the feature variables to predict the medal table through the model.

This paper introduces the LSTM (Long Short-Term Memory Network) model to predict gender ratio, number of participants, number of events in each category, and other feature variables.

### 2.1. LSTM Model Description

LSTM is a deep learning model commonly used to process sequential data. Compared to traditional RNN (Recurrent Neural Network), LSTM introduces three gates (input gate, forget gate, output gate) and a cell state, which are mechanisms that allow LSTM to better deal with long-term dependencies in sequences [4]. The network structure is shown in Figure 1.

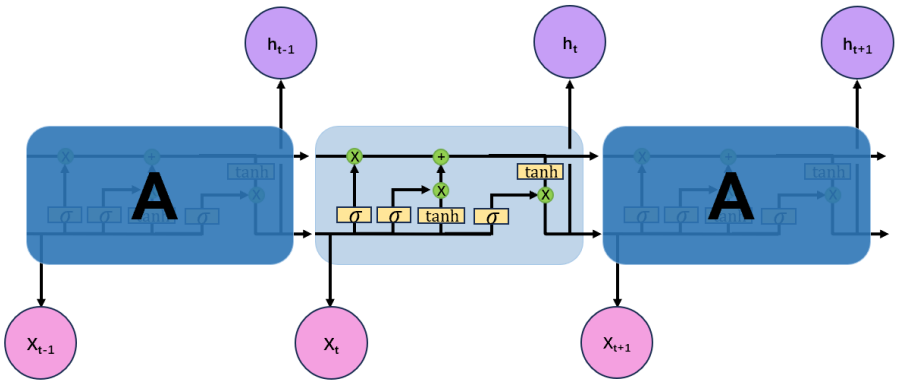


**Figure 1:** LSTM Interpretation Diagram

The core structure of LSTM is three gates, specifically, the input gate decides which new information will be stored into the memory cell, the forget gate decides which old information will be discarded, and the output gate controls which information in the memory cell will be output to the next section [5]. These gating units control the flow of information through sigmoid activation functions. The process of updating the specific memory state of LSTM is as follows:

(1) Decide what information to discard from the cell state:

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{1}$$

where:

$f_t$ denotes the output value of the forgetting gate, $\sigma$ denotes the sigmoid activation function, $W_f$ denotes the weight matrix of the forgetting gate, $h_{t-1}$denotes the hidden state of the previous time step, $b_f$ denotes the bias term of the forgetting gate, and $x_t$ denotes the input value of the current time step.

This section discards unneeded information.

(2) Determining what kind of new information is stored in the cellular state:

Here, it is necessary to get the new memory state by adding the previous information that needs to be left behind and the information that needs to be remembered now, that is, to get the new memory state, a process that requires the involvement of input gates. This part involves the formula for:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{3}$$

where:

$i_t$ denotes the output value of the input gate, $\tilde{C}_t$ denotes the candidate memory cell state, $W_i$ and $W_C$ denote the weight matrices of the input gate and the candidate memory cell, $b_i$ and $b_C$ denote the bias terms of the input gate and the candidate memory cell.

(3) Memory state update:

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \tag{4}$$

where:

$C_t$ denotes the state of the memory cell at the current time step, $C_{t-1}$ denotes the state of the memory cell at the previous time step.

In this way the memory of the cell state completes updating (removing old information and adding new information).

(4) Determine the output value:

Here, the output gate and filtered unit state are used, and the output value is awaited. The equations involved in this part are:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{5}$$

$$h_t = o_t \cdot \tanh(C_t) \tag{6}$$

where:

$o_t$ denotes the output of the output gate, $h_t$ denotes the hidden state of the current time step, $W_o$ denotes the weight matrix of the output gate, $b_o$ denotes the bias term of the output gate.

## 2.2. Predicting Characterization Variable Data

With the support of a wide range of Python libraries, this paper chose to use the tensorflow.keras library to build LSTM models.TensorFlow provides easy-to-use interfaces to build and train deep learning models, including LSTMs.Among them, Keras provides an LSTM layer, which is specifically designed to work with time series data. With the API provided by the tensorflow.keras library, LSTM models can be built quickly.

This paper used statistical data on characteristic variables (e.g., gender ratio) for each Olympic Games between 1896 and 2024 to input into the LSTM model, obtain predictions, and calculate the mean square error (MSE) for all data.

The MSE is calculated as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{7}$$

where:

$n$ denotes the number of samples, $y_i$ denotes the actual value of the ith sample, $\hat{y}_i$ denotes the predicted value of the ith sample.

Calculations are obtained:LSTM Mean Squared Error:0.7413834923111208

The mean square error between the predicted data and the original data is about 0.741383, indicating a small deviation between the two values. Therefore, it can be concluded that the LSTM model has good predictive performance for the feature variables of the 2028 Olympics for each country.

So, the predicted values of the feature variables can be used as the true values for 2028 to help predict the medal table.

## 2.3. Medal Predictions

The number of Olympic medal records changes with each passing tournament, and analyzing the changes in the data of these different countries can go some way to reflecting trends in the changing medal standings. By studying the historical data and trends, it is even possible to make informed

predictions about the performance of the participating countries in the future. In this case, this paper applied the Random Forest algorithm to machine learn the number of results provided and ended up with a prediction of the medal table for the 2028 Olympics. It should be noted that, for the sake of convenience in discussion, predictions for countries winning medals for the first time have not yet been included, i.e., countries that have never won medals in history are not included in this prediction.

### 2.3.1 Description of the Random Forest Algorithm

The Random Forest is a machine learning algorithm based on integrated learning [6], the core of which is by constructing multiple decision trees and integrating their predictions, it combines multiple decision trees together, each time the dataset is randomly have put back to the selection, while randomly selecting some of the features as the inputs, this model reduces the risk of overfitting and at the same time improves the model generalization ability [7]. The construction process is shown in Figure 2:

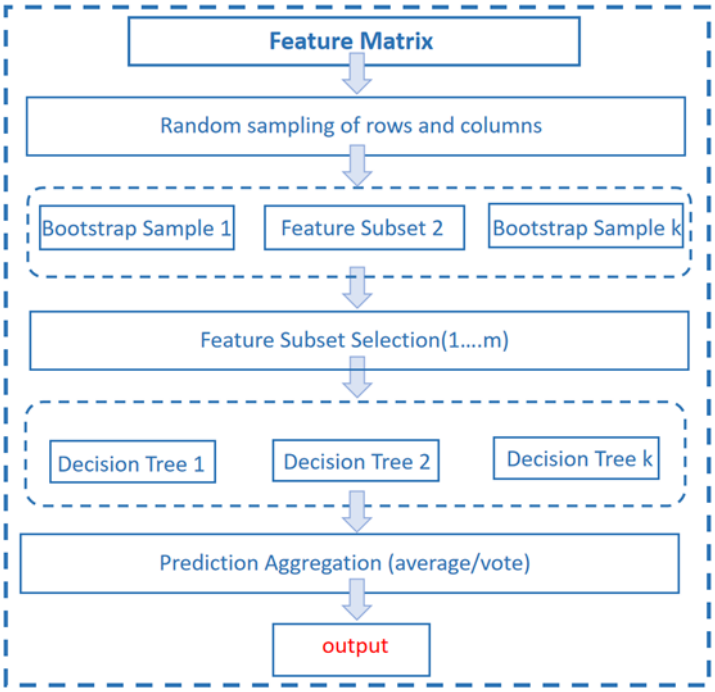


**Figure 2:** Random Forest Algorithm Construction Process

(1) Bootstrap Sampling:

N subsample sets (Bootstrap samples) are generated from the original training set via putative back sampling, each with the same capacity as the original dataset.

(2) Feature Selection:

Construct each decision tree by randomly selecting M features from the total feature set.

(3) Decision Number Construction Construct:

A decision tree for each Bootstrap sample using the selected features. Split the nodes by recursion until the stopping condition is reached.

(4) Aggregating Predictions:

For the classification task, the predictions of all trees are aggregated by Majority Voting; for the regression task, the average is used as the final output.

(5) The classification task:

For the input sample x final category for all the tree predictions for the plurality of results for:

$$\hat{y}(x) = \arg \max_{k} \sum_{t=1}^{T} I(h_t(x) = k) \quad (8)$$

where:

$h_t(x)$denotes the prediction result for the tth tree, I() denotes the indicator function, and T denotes the total number of counts.

(6) The regression task:

Uses mean aggregation and the final prediction is the mean of all tree predictions:

$$\hat{y}(x) = \frac{1}{T}\sum_{t=1}^{T} h_t(x) \tag{9}$$

### 2.3.2 Medal Prediction Results

Given the extensive support provided by Python libraries, the Random Forest model offered by the scikit-learn library was selected here [8]. It is a simple to operate and efficient tool for data mining and data analysis. scikit-learn supports a wide range of machine learning tasks, including Random Forest models.

This paper used 80% of the medal table data from every Olympic Games between 1896 and 2028 as a training set and the remaining 20% as a test set for the Random Forest model. The results of the Random Forest predictions (with fractional results rounded upwards), without considering the countries that won medals for the first time and showing only the top 20 total medal winners, are shown in Table 1:

**Table 1:** Projected medal table for 2028 (excluding first-time winners)

| Country | Gold | Silver | Bronze | Total |
|---|---|---|---|---|
| United States | 40 | 44 | 42 | 126 |
| China | 40 | 27 | 24 | 91 |
| Japan | 20 | 12 | 13 | 45 |
| Australia | 18 | 19 | 16 | 53 |
| France | 16 | 26 | 22 | 64 |
| Netherlands | 15 | 7 | 12 | 34 |
| Great Britain | 14 | 22 | 29 | 65 |
| South Korea | 13 | 9 | 10 | 32 |
| Italy | 12 | 13 | 15 | 40 |
| Germany | 12 | 13 | 8 | 33 |
| New Zealand | 10 | 7 | 3 | 20 |
| Canada | 9 | 7 | 11 | 27 |
| Uzbekistan | 8 | 2 | 3 | 13 |
| Hungary | 6 | 7 | 6 | 19 |
| Spain | 5 | 4 | 9 | 18 |
| Sweden | 4 | 4 | 3 | 11 |
| Kenya | 4 | 2 | 5 | 11 |
| Norway | 4 | 1 | 3 | 8 |
| Ireland | 4 | 0 | 3 | 7 |
| Brazil | 3 | 7 | 10 | 20 |

The predicted medal table visualization is shown in the Figure 3.

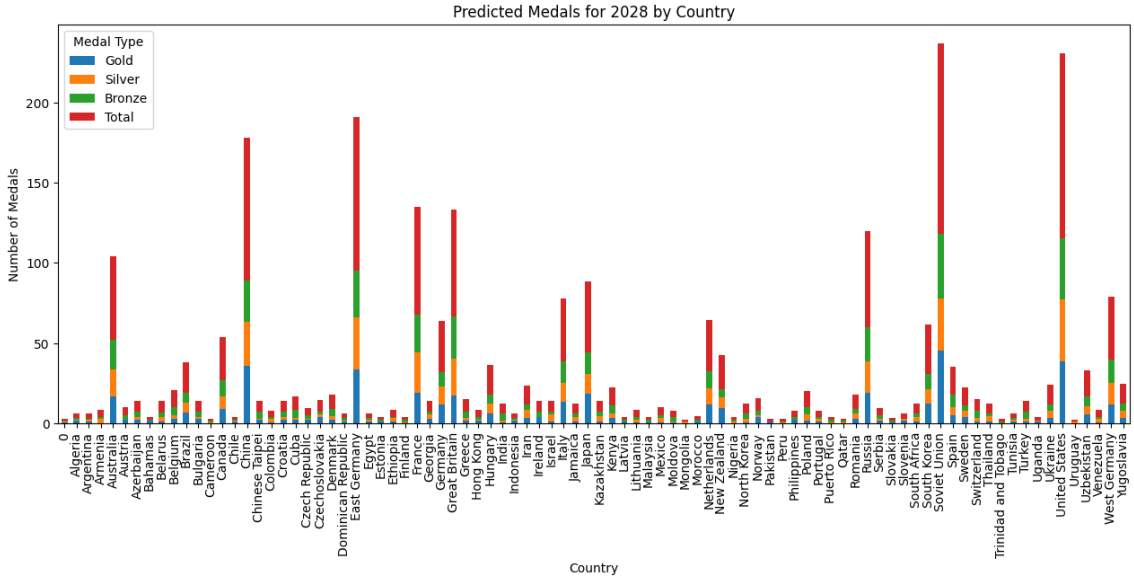


**Figure 3:** Projected medal table for 2028 (excluding first-time winners)

Calculate the (MSE)and the $R^2$ Score.

$R^2$ Score is calculated by:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{10}$$

where:

$n$ denotes the number of samples, $y_i$denotes the actual value of the ith sample, $\bar{y}_i$denotes the predicted value of the ith sample, $\bar{y}$ denotes the average of all actual values.

Calculated:Random Forest Mean Squared Error: 1.7809805555555558.

Random Forest $R^2$ Score: 0.929028671642823.

The mean square error between the predicted data and the original data is 1.78098, indicating a small deviation between the two values. Also, in statistics, the fit is considered to be better when the $R^2$ Score is close to 1. The above results are visualized as shown in Figure 4:

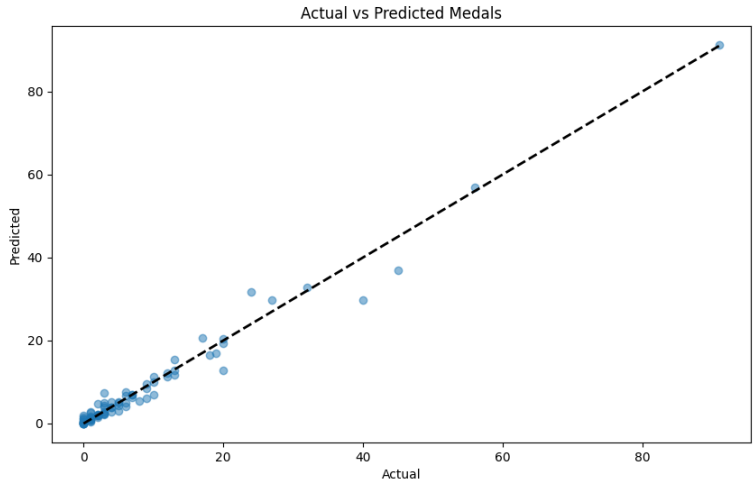


**Figure 4:** Visualization of the comparison between real and predicted values

Similarly, as can be seen in Figure 4 there is no significant difference between the predicted and true values.

Therefore, it can be concluded that the Random Forest model has good prediction performance for the number of medals of each country.

## 3. Random Forest Model with first-time award-winning countries

### 3.1. Data preparation for countries that have never won medals

In order to predict how many countries will win for the first time at the 2028 Olympics, this paper need datasets from countries that have never won a medal to be added to the dataset used by Model I for prediction. Since some countries that have never won a medal have participated in too few sessions of the Olympics, using their datasets would result in a smaller number of training and test sets for the model, which in turn would result in inaccurate model predictions, so this paper filtered and mined the total athlete dataset to find nine countries that participated in a large number of sessions and never won a medal, which are Libya, Palestine, Comoros, Zhide, Maldives, Benin, Somalia, Mali, Angola. Their data were selected and added to the dataset used for prediction. After that, perform a re-prediction.

### 3.2. Data Updating and Solving Random Forest Model

The process of modeling is the same as Random Forest Model. The LSTM model dataset needs to be updated first to get the new characteristic variables for the 2028 Olympic Games in each country. Then use the updated dataset's characteristic variables to make predictions. The medal list is obtained (medal predictions are rounded upwards if the result is a decimal).

The Mean Square Error, $R^2$ Score of the predicted data of the feature variables with respect to the original data indicates that the updated LSTM model also has a good predictive performance of the feature variables of the 2028 Olympic Games for each country. The mean square error of the predicted data for the number of medals versus the original data, $R^2$ Score, also indicates that the updated Random Forest model also has a good predictive ability for the number of medals for each country. The $R^2$ Score and Mean Square Error of the updated model are shown in Table 2.

**Table 2:** Tests of MSE and R2 Score for Updated Random Forest Models

| | LSTM | Random Forest |
|---|---|---|
| MSE | 0.7818822131456478 | 5.7666425 |
| $R^2$ Score | / | 0.9233346979740484 |

However, when compared to the above predictions without first-time winners it is clear to observe that the model does not predict countries that have never won a medal as well as those that have won a medal.

The results of the Forecast 2028 predicted medal table (with first time winners' countries) are visualized as shown in the Figure 5:

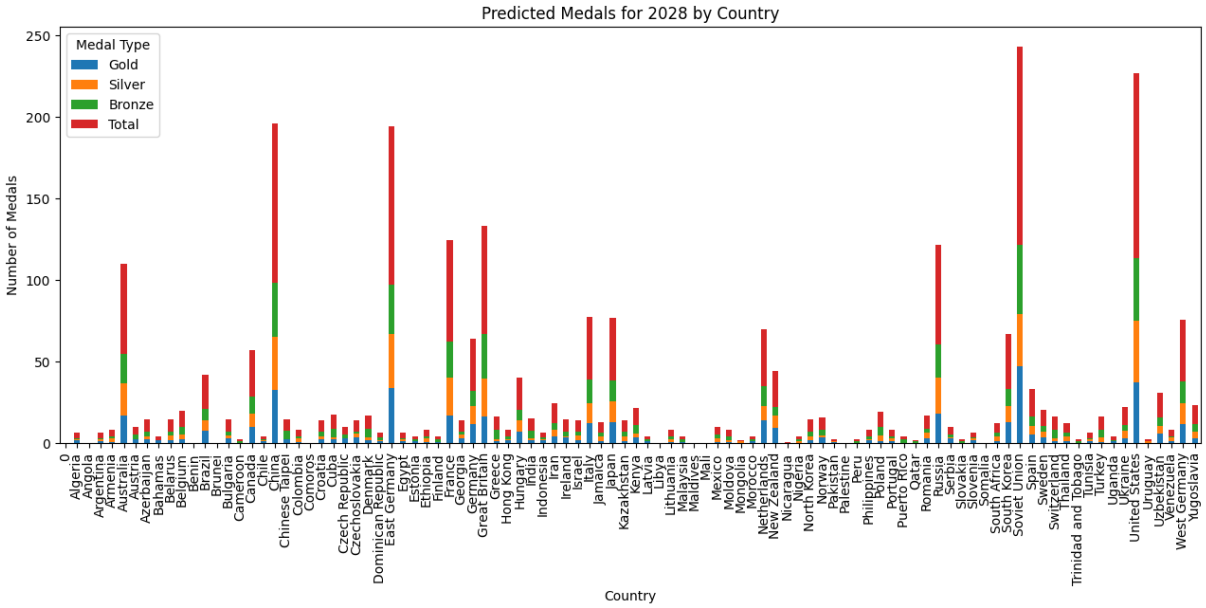


**Figure 5:** 2028 Projected Medal Table (with first-time winners)

### 3.3. Countries and probabilities of possible Olympic medals in 2028

Predicted medal table segments containing countries that have never won a medal before were selected, as shown in the Table 3.

**Table 3:** 2028 Olympic Medal Predictions for Non-Medalist Countries

| Country | Gold | Silver | Bronze | Total | P |
|---|---|---|---|---|---|
| Angola | 0.03 | 0.01 | 0.01 | 0.05 | 50.00% |
| Benin | 0 | 0.01 | 0.03 | 0.04 | 40.00% |
| Comoros | 0 | 0 | 0 | 0 | 0.00% |
| Libya | 0 | 0 | 0.02 | 0.02 | 20.00% |
| Maldives | 0.01 | 0.05 | 0.03 | 0.09 | 90.00% |
| Mali | 0 | 0 | 0.01 | 0.01 | 10.00% |
| Zhide | 0.01 | 0.05 | 0.04 | 0.1 | 100.00% |
| Palestine | 0 | 0.01 | 0 | 0.01 | 10.00% |
| Somalia | 0 | 0 | 0.01 | 0.01 | 10.00% |

According to the results, there are 8 countries that have never won a medal before may win in the next Olympic Games, and P in the table indicates the probability that the country will win in the next Olympic Games. In calculating the value of P, using the idea of fuzzy mathematics, the maximum value in the predicted value of the total number of medals is certain to win the award, and the minimum value is certain not to win the award, analogous to the relative affiliation to get the formula for P:

$$P = \frac{u_i - minu_i}{maxu_i - minu_i} \tag{11}$$

where:

$u_i$ denotes the total number of awards for $i$ countries($i = 1, 2, 3 \ldots 9$).

It is worth mentioning that according to the predicted results Zhide will surely take their first Olympic medal in the next Olympics, while Comoros still failed to complete their first win in the next Olympics.

# 4. Importance of the event program to countries

## 4.1. Construction of the Random Forest Regressor model

This part bases the regression model on the Random Forest algorithm for predicting continuous values by grouping the data by country and then performing a data row count check: check if the number of data rows for each country is less than 2. Since at least 2 data points are needed to perform a Random Forest regression analysis [9], if the number of rows is less than 2, the country is skipped from the analysis.

After that, extract the features and target variables: for countries with enough rows of data, extract the sports-related data from the grouped data as feature matrix X, and the total number of medals as target variable y [10]. Create a Random Forest Regressor model object.

## 4.2. Model solution results

Using the fit () method, the feature matrix X and the target variable y are passed into the model for training, allowing the model to learn the relationship between sport participation and the total number of medals.

Once the training is complete, the importance score for each sport is obtained via model. Feature importances, which indicates how much the sport influences the prediction of the total number of medals.

Because of space constraints, this papaer now shows the 20 most important sports in a given country with the largest importance scores, as shown in the Table 4:

**Table 4:** The 20 best sports in a country with the highest importance scores

| Country | Important sport | Ratio | Country | Important sport | Ratio |
|---|---|---|---|---|---|
| Norway | Shooting | 0.694551 | Australia | Aquatics | 0.416908 |
| Russia | Shooting | 0.591268 | Morocco | Shooting | 0.407951 |
| Austria | Shooting | 0.583483 | Netherlands | Athletics | 0.404427 |
| Algeria | Shooting | 0.565435 | Denmark | Shooting | 0.3934 |
| Belgium | Shooting | 0.565251 | UnitedStates | Fencing | 0.390629 |
| Estonia | Shooting | 0.541083 | Brazil | Athletics | 0.379895 |
| Spain | Athletics | 0.537014 | Sweden | Shooting | 0.379211 |
| Ethiopia | Weightlifting | 0.508884 | Latvia | Shooting | 0.377869 |
| Pakistan | Shooting | 0.419739 | Ivory Coast | Weightlifting | 0.370968 |
| China | Athletics | 0.418049 | Poland | Aquatics | 0.366045 |

These events are the most important events for the corresponding country in the table because they have the highest importance score among all sports in the corresponding country. That is, the events with the highest importance scores have the greatest positive impact on the number of Olympic medals won by the corresponding countries, and the higher the importance score, the greater the probability that the corresponding country will win in that event.

Therefore, when the host country selects Olympic events, if it chooses to increase the number of competitions in its most important events or other events with high importance scores, it will win more medals in the Olympics and increase its ranking in the medal table. At the same time, this selection method will favor countries that are close to the host country's expertise and disadvantage countries that are very different from the host country's expertise.

## 5. Conclusion

This paper focuses on predicting the Olympic medal tally by combining Random Forest algorithms and LSTM models to predict feature variables (such as gender ratio and number of participants), thereby constructing a high-precision medal prediction model. In the prediction of basic information about participating countries for the 2028 Olympics based on the LSTM model, after organizing and cleaning the data, variables such as year, athlete gender ratio and number, host country, event type and number were selected as the variables to be predicted, accurately predicting the basic information of participating countries for the 2028 Olympics, with good model accuracy. Subsequently, a Random Forest model was used to predict the medal rankings for the 2028 Olympics, and the likelihood of countries that have never won a medal securing their first medal was analyzed. Finally, a Random Forest regression model was employed to analyze the impact of Olympic event categories on medal rankings. The models constructed in this study exhibit high accuracy and precision, providing a scientific basis for optimizing training strategies and allocating resources for participating countries in the 2028 Olympics.

## References

[1] Luo Yubo, Cheng Yanfang, Li Mengyao, et al. Prediction of China's Medal Count and Overall Strength at the Beijing Winter Olympics: Based on the Host Country Effect and Grey Prediction Model [J]. Contemporary Sports Science and Technology, 2022, 12(21): 183-186.

[2] Lu Z, Li S, Sun J. Prediction of Olympic Medal Based on Multiple Linear Regression and Logistic Regression[J]. Frontiers in Computing and Intelligent Systems,2025,12(1):17-21.

[3] Cheng Hongren, Lyu Jie, and Yuan Tinggang. Predicting China's Track and Field Performance at the Tokyo Olympics Based on the 2018 World Top 20 Rankings for Track and Field Events [J]. Sports Science and Technology Literature Bulletin, 2020, 28(04): 4-8.

[4] Zhu X, Wang S, Li X, et al.A Study on Olympic Medal Table Prediction Based on LSTM and DBILSTM[J].Journal of Globe Scientific Reports,2025,7(2):249-258.

[5] Yan D. OLYMPIC MEDAL PREDICTION AND ANALYSIS BASED ON LSTM AND TOPSIS MODELS[J]. Journal of Computer Science and Electrical Engineering,2025,7(3):

[6] Zhang Yiming, Tang Yulei, Feng Junbo. Prediction and Analysis of Glaciers on the Qinghai-Tibet Plateau Based on a Random Forest Model [J/OL]. Arid Zone Geography, 1-14 [2025-06-12].

[7] Bai X, Zhang L, Feng Y, et al.Multivariate temperature prediction model based on CNN-BiLSTM and RandomForest[J].The Journal of Supercomputing,2024,81(1):162-162.

[8] Lemenkova P. Automation of image processing through ML algorithms of GRASS GIS using embedded Scikit-Learn library of Python[J]. Examples and Counterexamples,2025,7100180-100180.

[9] Lee J Y, Joo J M, Yu K H, et al.Random forest regressor for predicting sensory texture of emotional designed packaging films[J].Results in Engineering,2025,25104147-104147.

[10] Chowdhury S, Saha K A, Das K D. Hydroelectric Power Potentiality Analysis for the Future Aspect of Trends with R2 Score Estimation by XGBoost and Random Forest Regressor Time Series Models[J]. Procedia Computer Science,2025,252450-456.