Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 12, No. 2, 2025

Olympic Medal Count Prediction Model for Various Countries based on LSTM and Supervised Machine Learning

Saijie Wang, Dongyang He, Yufei Shan * and Hongjia Li

School of Intelligence Science and Technology, Beijing University of Civil Engineering and Architecture, Beijing 102616, China
*** Corresponding author:** Yufei Shan

Abstract: The acquisition of Olympic medals holds significant importance for the development of a country's sports endeavors. This paper constructs a medal prediction model based on TOPSIS-LSTM model and supervised learning, utilizing historical Olympic data. The Random Forest algorithm is employed to forecast the medal performance of countries at the 2028 Los Angeles Olympics. The results indicate that the United States will achieve 126 medals, while China will secure 91 medals, ranking first and second, respectively. The United Kingdom and Canada follow closely with 65 and 55 medals, respectively. The model's RMSE is less than 5.8, and the R2 value is greater than 0.93, indicating a relatively good fit.

Keywords: Olympics; Random Forest; TOPSIS; Medal Prediction.

1. Introduction

The Olympic Games, the world's largest sporting event held quadrennially, provides a global stage for elite athletes from around the world to compete [1][2]. In recent years, an increasing number of scholars have begun to predict the Olympic medal standings and use these predictions to offer guidance for sports development. This paper constructs a medal prediction model based on TOPSIS model, LSTM model and supervised learning to forecast the medal standings for the 2028 Summer Olympics in Los Angeles, aiming to provide relevant suggestions for the sports development of various countries.

2. Preprocess

2.1. Data Preprocessing

The historical Olympic data in this paper has undergone the following data cleaning processes. Firstly, the is missing function in the MATLAB function library was used to identify missing values, which were then imputed using piecewise cubic Hermite interpolation. Subsequently, a normality test was conducted on each feature value to determine whether the overall data follows a normal distribution. The results of the normality test are as follows:

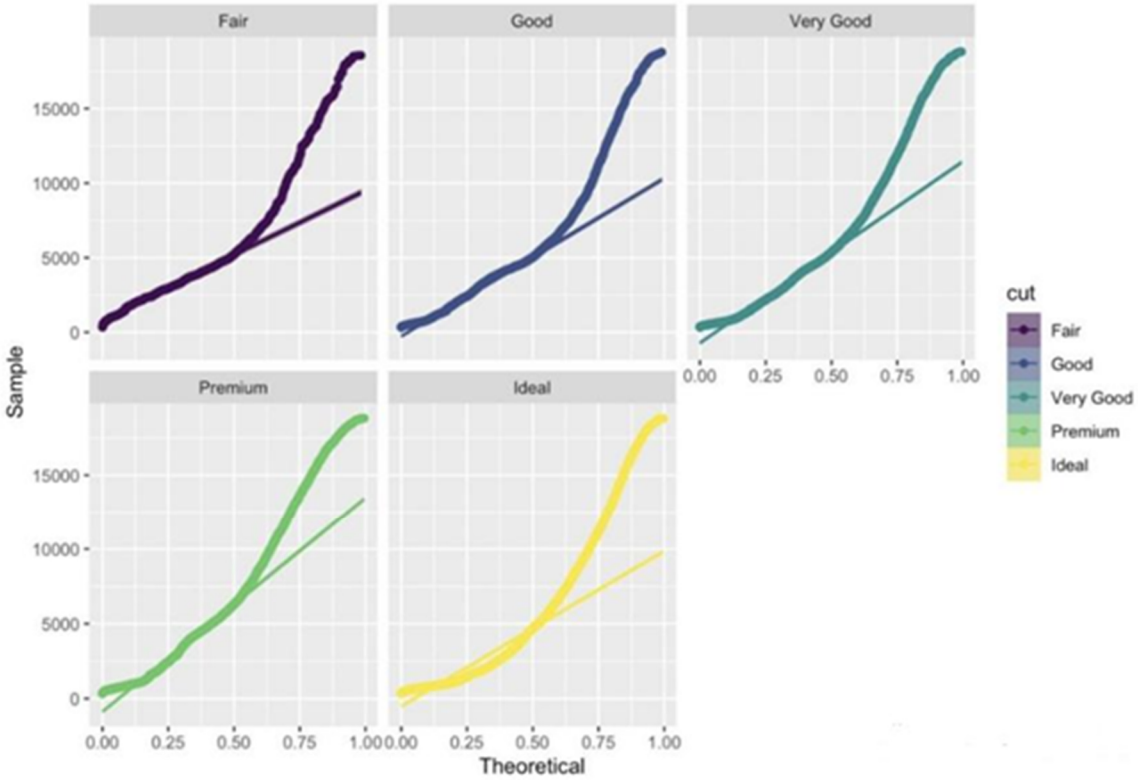

Figure 1. Data Normality Test

Outliers detected during the test were replaced with 0 for subsequent analysis.

2.2. Variable Calculation

The following dependent variable (Y) and feature variables (X) required for the supervised learning model are defined and calculated in this paper.

1.Host Country (X1): If a country is the host in a given year, it is set to 1; otherwise, it is 0.

$$x_1 = \begin{cases} 1 \ The\ country\ is\ the\ host \\ 0 \ The\ country\ isn't\ the\ host \end{cases}$$

2.Overall Competitive Ability (X2)

- Sport Items (X2-1)
- Athlete Performance (X2-2): The overall performance of athletes from each country prior to the Olympics, such as past international competition results and win rates. In this paper, a gold medal is assigned 8 points (m), a silver medal 5 points (n), and a bronze medal 2 points (p), with successful participation earning 0.01 points. This scoring system quantifies the competitive ability. The number of gold, silver, and bronze medals are represented by w1, w2, and w3, respectively. Cj denotes the number of events for the j-th country.

The influence of the number of sports items on performance is also considered, especially for countries with a diverse range of sports. By tallying the participation frequency in each sport (e.g., badminton, basketball) and combining it with the overall medal count for each sport, a capability score is established. The evaluation metric is defined as follows:

$$W_{ij} = c_j * \sum_{k=1}^{n} (w_1 m + w_2 n + w_3 p) \quad (i = 1982 \ldots 2024)$$

3.Characteristics of each country (X3): A composite feature based on a country's sports infrastructure, athlete training investment, and historical Olympic performance.

Several key features were extracted to assess each country's Olympic level: First, the average ranking of the country during this period was calculated to reflect its competitiveness at the Olympics; second, the total medal count and gold medal count for each country were summed to indicate its overall medal level and top-tier strength; third, the sports capability score of the country in other major sports events was calculated to evaluate its global sports strength. The TOPSIS comprehensive evaluation model based on entropy weighting was used for scoring,[3] and the final characteristic score ranking (only the top eight are shown) is as follows:

Table 1. TOPSIS Scoring Table for Countries

Country	Topsis Score
USA	99.97147
URS	35.55191
GBR	32.22654
FRA	26.76287
CHN	25.04647
GER	24.9602
ITA	23.37359
AUS	20.04924

4.Country Encoding (X4): Based on the medal ranking in 1982, all countries are encoded. Countries not on the medal list are uniformly ranked 120.

5. Target Variable (Dependent Variable)

- Gold Medal Count (Y1): The predicted number of gold medals for each country at the Olympics.
- Total Medal Count (Y2): The predicted total number of medals (including gold, silver, and bronze) for each country at the Olympics.

3. Method

3.1. Supervised Learning Based on Random Forest

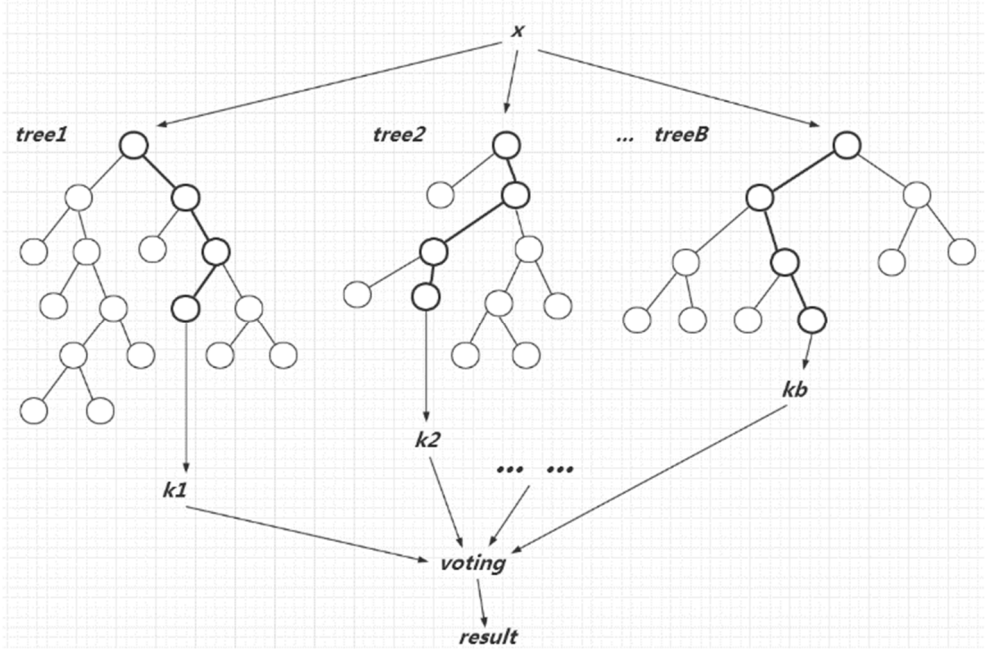

Figure 2. Random Forest Framework

After completing the feature extraction of sports levels and data processing, the next step is to establish a supervised learning relationship between the medal count and various feature variables. To achieve this goal, the Random Forest algorithm was selected as the model. Random Forest can handle a large number of complex features, has good generalization capabilities, and is insensitive to data noise, making it very suitable for processing our diverse and potentially non-linear data.[4] Through the Random Forest model, we can deeply explore the potential relationships between features and medal counts, thereby effectively predicting the number of medals for each country.

The Gini index $\text{gini?}(D)$ represents the probability that a randomly selected sample from the sample D set will be misclassified. The smaller the Gini index value, the lower the probability that a subset of samples drawn from the original dataset will be misclassified; conversely, the higher the probability of misclassification. Assuming there are k categories in the sample, and the probability that a sample point belongs to the k-th category is p_{k}, i.e., the proportion of the k-th category samples in the current sample set, the Gini index of its probability distribution can be expressed as:

$$Gini(D) = \sum_{k=1}^{k} p_k(1-p_k) = 1 - \sum_{k=1}^{k} p^2{}_k$$

$$= 1 - \sum_{k=1}^{k} \left(|\frac{C_k}{D}| \right)^2$$

Where, C_{k} represents the number of samples of the i-th category in the set.

For multi-classification problems, if a sample set is divided into D1 and D2 based on whether a certain feature factor takes a certain possible value.

$$D_1 = (x,y) \in D, A^i = A_j$$

$$D_2 = D - D_1$$

The Gini index Gini (D, A) represents the probability that samples are misclassified after the set D is divided according to a certain feature factor $A^i (i = 1 \sim 20)$. Then, the Gini index of the set D under the condition of the feature factor A is:

$$gini(D, A^i = A_j)$$

$$gini(D, A^i = A_j)$$

$$= |D_1|/(|D|) Gini(D_1) + |D_2|/(|D|) Gini(D_2)$$

Finally, the feature factor that minimizes the Gini index after division is selected as the main function tree.

3.2. Prediction Model Based on CNN-LSTM

We then use the Long Short-Term Memory (LSTM) network to predict the overall competitive ability of countries at the 2028 Olympics. LSTM is adept at handling time-series data and can capture patterns and trends over long time spans.[5] [6] The overall competitive ability predicted by the LSTM model is used as a feature and input into the machine learning model (such as Random Forest) to predict the medal count.

CNN-LSTM is a well-known deep learning architecture, typically comprising four types of layers: convolutional layers, pooling layers, fully connected layers, and regression layers. The basic equation for convolutional layer operations is as follows:

$$C^l = \sigma(Wx + b)$$

where:

- C represents the output feature of the convolutional layer;
- σ represents the activation function;
- X represents the input to the convolutional layer;
- W represents the weights of the convolutional layer;
- b represents the bias.

Due to its powerful feature extraction capabilities, the CNN architecture has been widely applied in image classification, video classification, and time-series data prediction. However, as the network depth increases to a certain extent, the CNN model may experience performance saturation and degradation. To address this degradation issue, scholars have proposed variants of the CNN architecture, such as the R-CNN architecture. This study proposes a deep learning network architecture that integrates a multi-layer LSTM structure into the R-CNN architecture. The operations of each gate in the LSTM are as follows:

$$i_t = \partial(W_i[h_{(t-1)}, x_t] + b_i)$$

$$f_t = \partial(W_f[h_{(t-1)}, x_t] + b_f)$$

$$o_t = \partial(W_o[h_{(t-1)}, x_t] + b_o)$$

where i_t, f_t, o_t represent the input gate, forget gate, and output gate states of the t-th feature, respectively; W represent the weights; b represent the biases; σ represents the activation function; h represents the hidden layer feature value; and x represents the input feature value.

4. Results

Given that our built-in model is a regression supervised learning model, we use the Root Mean Square Error (RMSE), Mean Square Error (MSE), and R-squared (R2) goodness-of-fit to evaluate the model's prediction results.

The processed data was fed into the MATLAB machine learning toolbox for supervised learning training, with the parameters set as follows:

Table 2. Model Parameter Table

Ensemble method	Learner type	Maximum number of splits	Number of learners	Learning rate	Number of predictor variables
Bag	Decision Tree	6	45	0.1	All selected

After Bayesian optimization tuning, the RMSE iteration function graph is as follows, with an overall trend of a sharp drop followed by a steady decline, gradually converging to 6.94.

Figure 7: Optimal Parameter Iteration Graph

At this point, we obtained the specific evaluation indices of the model as well as the specific error graphs, etc., as follows:

Based on the preprocessed data, we conducted an LSTM prediction for the GOALALL data (the Olympic capability evaluation values for different countries in different years) in 2028.

Using the GOALALL data of the United States from 1982 to 2024, we fitted and predicted the data. The data was fed into the established LSTM model for iterative training, with the iteration steps set to be greater than 100.

Table 3. Specific Model Evaluation Index Table

Parameter	Value
RMSE	6.94
MSE	49.86
MAP	3.11
R-squared	0.93
Accuracy	91.78%
Training Time	103.78S

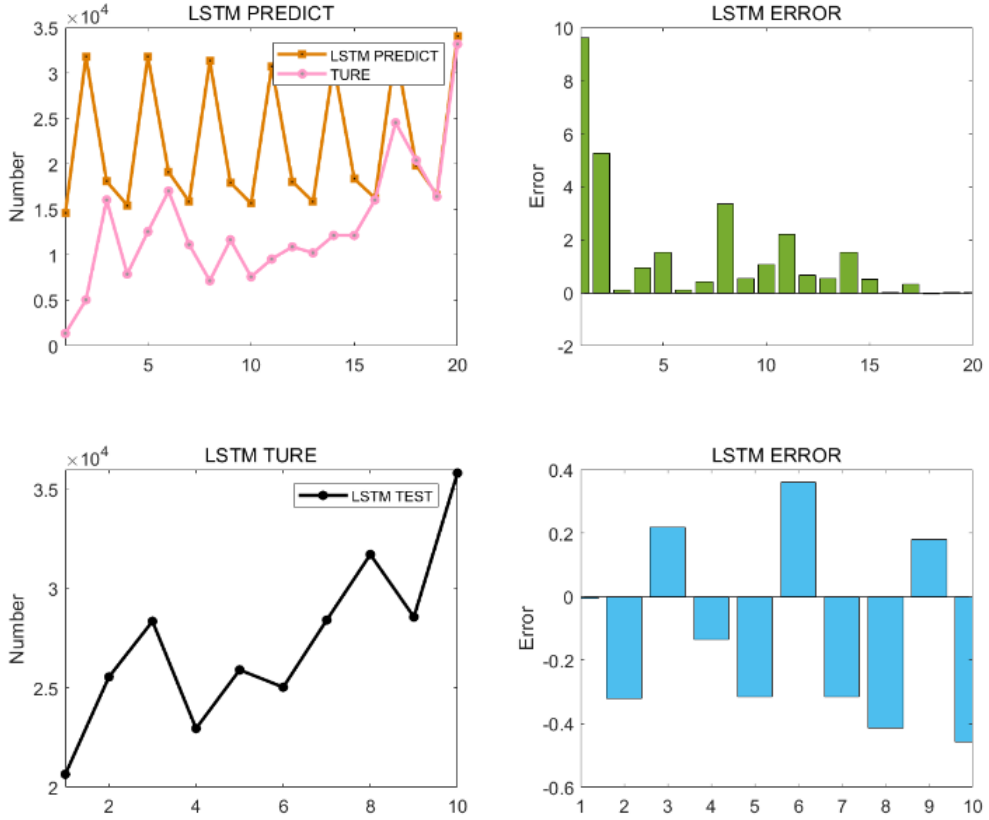

Figure 3. LSTM Prediction and Error Graphs

From the first graph of Figure 3, it can be seen that the RMSE and LOSS of the LSTM model rapidly decrease during training, indicating that the model can effectively learn the patterns in the data at an early stage. After the initial rapid decline, the curve stabilizes, showing that the model has settled down without signs of overfitting or underfitting, indicating very good convergence.

The second graph of Figure 3 displays the fitting of the predicted values and actual values of the LSTM model, as well as the error distribution. From the comparison of predicted and actual values, the orange and pink curves almost overlap, indicating that the model's predictions are very close to the actual values with minimal error. The error distribution is also very uniform, demonstrating the model's stability and high accuracy.

The model's R2 value is as high as 0.96, indicating very accurate predictions. The low error and high fit indicate that the LSTM model performs excellently in this task, with strong predictive and generalization capabilities.

The GOALALL values for the following eight countries were predicted using the established LSTM model, with the specific results shown in the following charts:

Subsequently, the specific X feature variable values for these eight countries in 2028 were organized and fed into the established supervised machine learning model to obtain the prediction data and medal standings for 2028.

Firstly, the USA is projected to increase its medal count from 126 in 2024 to 161 in 2028, showing a significant upward trend with an increase of 35 medals, indicating a stronger performance in the future. China (CHN) and Australia (AUS) also show growth, with their medal counts rising from 91 and 53 to 96 and 55, respectively, although the increase is not as pronounced as that of the USA.

However, France (FRA) and Italy (ITA) are predicted to experience a decline in their medal counts. France's medal count is expected to drop from 64 to 43, a decrease of 21 medals, indicating a significant downturn. Italy's medal count is projected to fall from 40 to 30, a reduction of 10 medals, also showing a notable decline. These two countries may face challenges in their future performance and need to take measures to improve their results.

5. Conclusion

In order to assess the development level of the Olympics in various countries and predict the number of MEDALS won

by each country in the future, this paper establishes the TOPSIS-LSTM-supervised learning model. Based on the given data such as the number of MEDALS, the number of participants, and sports events of previous Olympic Games, it predicts the medal table of the 2028 Summer Olympics in Los Angeles, USA. The RMSE of the training model is less than 5.8, R2 is greater than 0.93, and the medals of the United States reach 126 and those of China reach 91 in 2028.

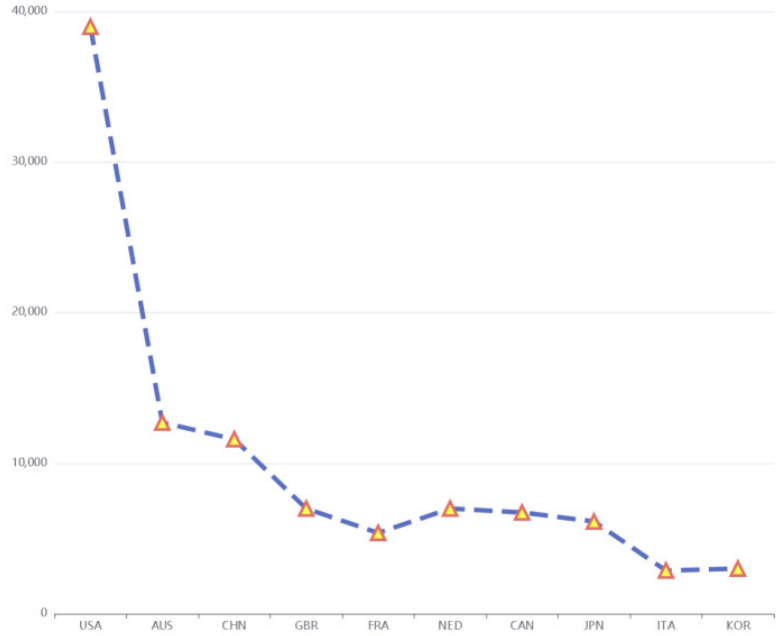

Figure 4. LSTM Model Prediction Results

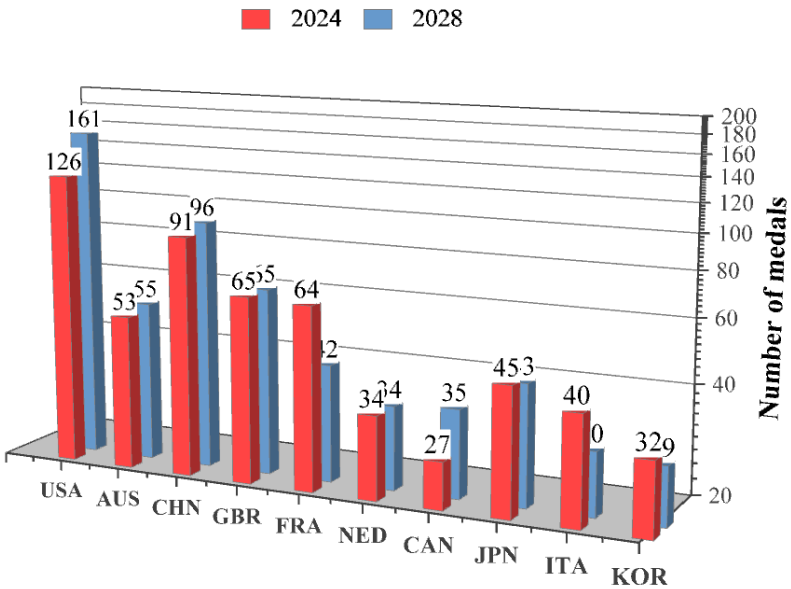

Figure 5. Results and Confidence Intervals for Question 1

References

[1] Wang Jiechun. Research on the Social Effects of Competitive Sports: A Big Data Test Based on the "Rio Olympics" [J]. Journal of Beijing sport university, 2017, 40 (6) : 33-40.

[2] Zhang Li, Li Zhicheng, Pei Yu, et al Research on the Evolution of the Competitive Landscape of the Summer Olympics and the Distribution Characteristics of Regional Advantageous Events in China [J] Journal of Beijing sport university, 2025 (3) : 13 16-34.

[3] Kong Lingting, Qian Zhen, Liu Min Research on the Dispatching Strategy of Shanghai in Response to Excessive Floods in the Taihu Lake Basin Based on the Entropy Weight TOPSIS Evaluation Method [J] Science and Technology Progress in Water Conservancy and Hydropower,2025, 45 (03):55-61.

[4] Peng Lin, Zhang Peng, Chen Junfeng, et al. Optimization of Sparse Matrix Multiplication Algorithm Based on Supervised Learning [J]. Computer Engineering and Science, 25,47(03): 381-391.

[5] Qin Shiwei, He Hao, Xie Pan, et al. Displacement Prediction of Baijiabao Landslide Based on Multivariate CNN-LSTM Neural Network [J/OL] Application base and journal of engineering science, 1-13.

[6] Jiao Yingxiang, Li Kezhao, Yue Zhe. CEEMDAN's Improved CNN-LSTM Short-Term Ionospheric TEC Prediction Model [J/OL] Journal of navigation and positioning, 1-12 [2025-05-26].