Research on Medal Prediction for Events Based on LSTM and BP Neural Networks

Shihao Zhang*, Yang Zhou

Southampton Ocean Engineering Joint Institute, Harbin Engineering University, Harbin, China

*Corresponding author: 17669757689@163.com

Abstract. This study focuses on predicting medals in the Summer Olympics. It aims to analyze medal - distribution factors and predict future results through scientific modeling. A multi - input multi - output LSTM model is built to predict each country's medal wins in 2028. Considering athletes' strengths (quantified by past performance), participation arrangements, and historical awards, the model generates a 2028 predicted medal list. The top three gold - medal - winning countries are predicted to be the US, China, and Great Britain, consistent with historical trends. A BP neural network prediction model is also constructed to predict non - winning countries' performance in 2028. Based on the number of participants and athletes' international rankings, etc., the model shows high accuracy (F1 score of 0.93 with prediction errors around 0). It predicts AIN and LAT have a 0.466 and 0.415 probability of winning in 2028. This research offers valuable references for Olympic medal prediction and sports event analysis.

Keywords: Summer Olympics, Medal Prediction Model, LSTM Neural Network, BP Neural Network.

1. Introduction

The Summer Olympics, a global sports event, gathers top athletes worldwide, embodying highly - regarded sportsmanship and athletic values. The medal list not only reflects a country's sports strength but also symbolizes national honor and the effectiveness of sports strategies, attracting global sports enthusiasts' attention and triggering academic and sports - circle reflections on medal - winning factors.

Historically, Olympic medal distribution has changed. Initially, sports powerhouses dominated with their deep sports heritage and cultivation systems. However, emerging sports countries have since altered this pattern, driven by factors like athletes' improved strength, adjusted participation strategies, and changes in sports resource investment. For instance, some countries have increased investment in emerging sports, nurturing competitive athletes and achieving medal breakthroughs.

With the rapid development of big data and AI, scientific models for predicting Olympic medals have become both possible and valuable. Accurate predictions can help national sports departments formulate training and participation strategies, allocate resources rationally, meet sports enthusiasts' expectations, and provide decision - making references for related industries. Therefore, this study aims to use advanced machine - learning algorithms to build a scientific prediction model, explore key factors affecting Summer Olympics medals, and accurately predict the 2028 Olympics medals, contributing new theoretical and practical results to the sports field.

2. Medal table prediction

2.1. Establishment of prediction model based on LSTM long and short-term memory recurrent neural network

In this paper, the prediction of the awards for each country is made based on LSTM [1].

The memory of the human brain is persistent, and it can understand and learn the current knowledge through the accumulation of past knowledge. While the traditional neural network does not have persistence, each neuron can not be re-inferred learning through the learning results of the previous neuron, in order to solve this problem scientists proposed RNN Recurrent Neural Networks

(RNN), and to improve the proposed a special kind of recurrent neural network: the LSTM long and short-term memory network [2].

LSTM can solve the problem of long-term dependence that is common in general recurrent neural networks, effectively transfer and express the information in a long time sequence and will not lead to the useful information in the long time ago to be forgotten. At the same time, LSTM can also solve the gradient vanishing/exploding problem in RNN.

The LSTM also has the chain structure of an RNN, but the repeating module has a different structure that is based on four neural network layers interacting with information in a very specific way, as shown in Fig. 1.

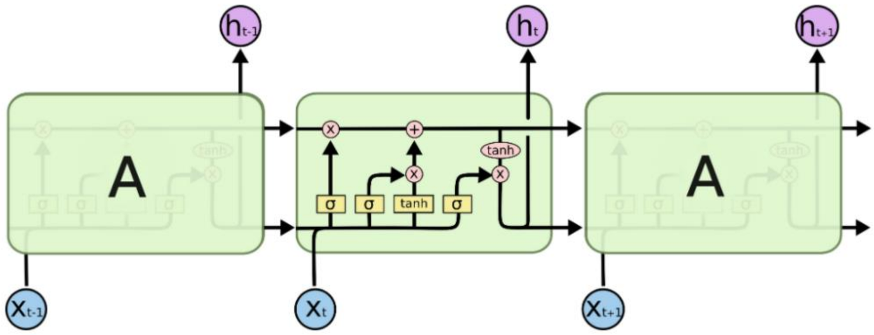

Fig. 1 LSTM structure

In Fig. 2, each line carries a complete vector from the output of one node to the inputs of other nodes; the pink circles represent point-by-point operations, such as adding and subtracting vectors; and the yellow boxes are the neural network layers used for learning.

The first layer is the “forgetting gate” layer, which decides what information we want to discard from the cell state. It utilizes h_{t-1} and x_t and outputs a number between 0 and 1 for each digit in the cell state C_{t-1}. An A of 1 means to keep the information, while an A of 0 means to forget it.

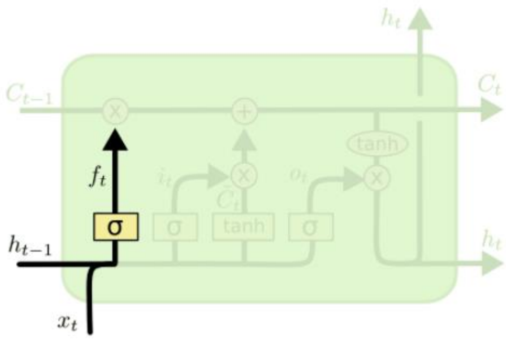

Fig. 2 Amnesia Gate

$$\boldsymbol{f_t = \sigma(W_f x_t + W_f h_{t-1} + b_f)} \tag{1}$$

The second layer is the “memory gate” layer, as shown in Fig. 3 which decides what new information we want to store in the cell state. First, a sigmoid layer called the “input gate layer” decides which values to update. Next, a tanh layer creates a vector of new candidates ($\tilde{C}_t$) that can be added to the state.

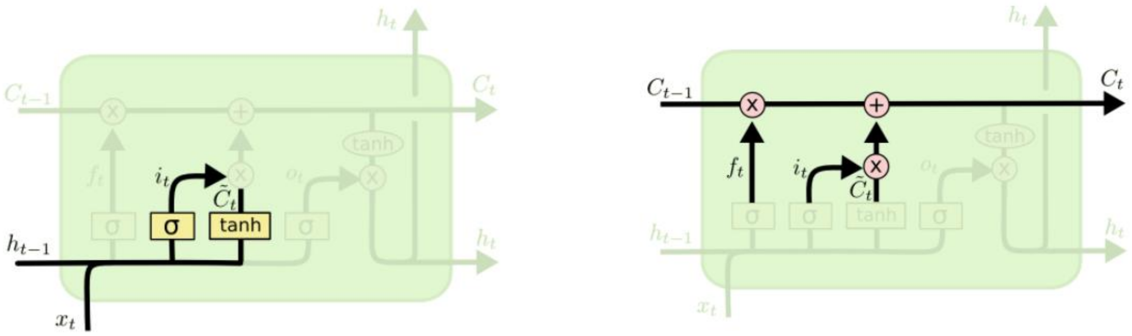

Fig. 3 Memory Gate

$$\boldsymbol{i_t = \sigma(W_i x_t + W_i h_{t-1} + b_i)} \tag{2}$$

$$\boldsymbol{\mathrm{C}_t = tanh(W_C x_t + W_C h_{t-1} + b_C)} \tag{3}$$

$$\boldsymbol{C_t = f_t * C_{t-1} + i_t * \mathrm{C}_t} \tag{4}$$

The third step is the "output gate" layer, as shown in Fig. 4. Which determines the information to be output. This output will be based on the cell state, but will be the result of filtering. First, a sigmoid layer is run to determine which parts of the cell state to output. The cell state is then passed through tanh (a value between -1 and 1) and multiplied by the output of the sigmoid layer so that only the information we decided to output, h_t, is output.

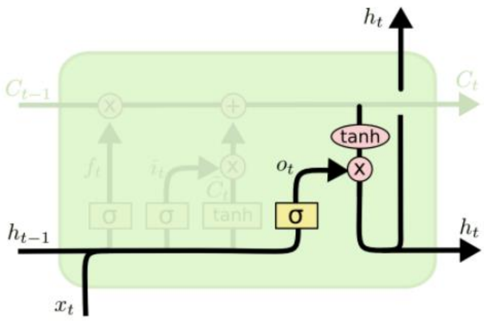

Fig. 4 Output gate

$$\boldsymbol{O_t = \sigma(W_o x_t + W_o h_{t-1} + b_o)} \tag{5}$$

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

In this paper, a multi-input medal prediction model is constructed in order to fully take into account the arrangement of athletes in the period to be predicted and the history of awards. Each feature is shown in Table 1 below:

Table 1. Input Characteristics of National Medal Count Prediction Models

Hallmark	Unit
Number of gold medals in the previous period	Lump
Number of silver medals in the previous period	Lump
Number of bronze medals in the previous period	Lump
Number of participants from the country during the period	People
Total number of entries by participants from that country during the period	Times
Average number of entries by participants from that country during the period	Times

The number of medals is the most intuitive measure of a country's level of sports competitiveness. The number of medals in the previous period reflects the country's sports performance and competitive strength in the past period, and is an important basis for predicting the number of medals in the future. By analyzing the historical medal data, it can be found that some countries have strong competitiveness in certain sports, which will be carried over to the future Olympic Games to a certain extent [3].

More participants means more opportunities for that country to show its strength at the Olympics. More participants means more potential medal winners. Therefore, the number of participants from that country in the current period is a good reflection of the number of medals won.

The total number of entries reflects the competition experience and athleticism of the participants. Experienced competitors are more likely to perform at their best in the Olympics and thus win medals. Competitors who have competed many times are more mature and stable both psychologically and technically, which helps increase the likelihood of winning medals.

The average number of competitions reflects the overall strength and athleticism of a country's sports teams. The higher the average number of entries, the more competitive the country's sports teams are in multiple sports.

2.2. Model predictions

Projections of the 2028 awards for each country based on the model described above. And evaluation based on historical awards for each country. The award projections for selected countries are shown Fig. 5 below:

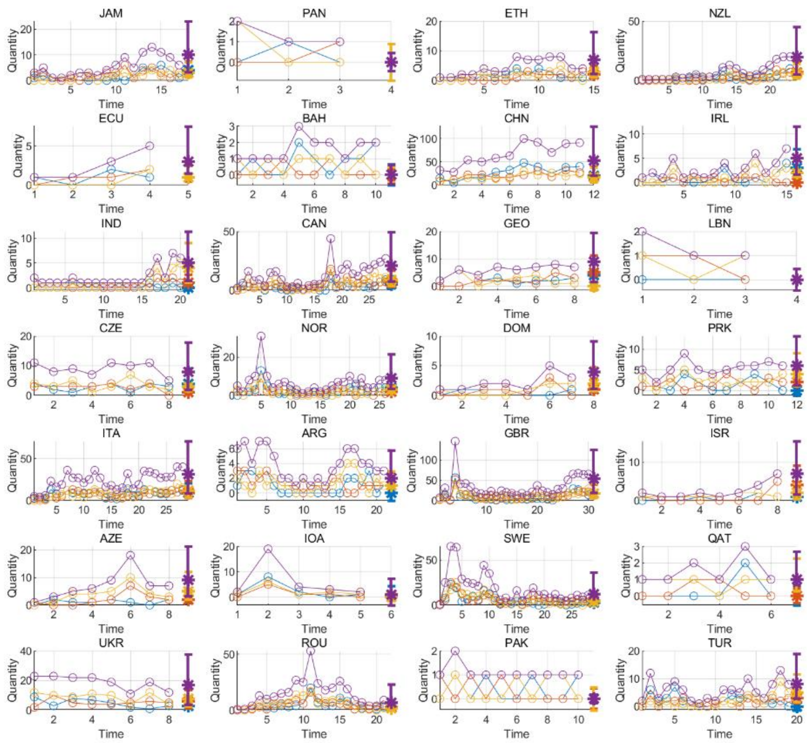

Fig. 5 Algorithmic Predictions for Selected Countries

In the graph, the hollow origin shows the actual awards for 2024 and earlier, and the asterisks show the projected awards for 2028. As can be seen from the above graph, the vast majority of the predicted values are consistent with the trend of historical awards. Some of the inconsistencies in the trends are due to the fact that the placement of entrants in 2028 will also affect the awards to some extent.

3. First Medal Prediction

3.1. Establishment of BP neural network

Assuming that the input from the ith neuron is X_i, and for each input signal, its connection weight for that neuron is not the same, the final actual input has to be multiplied by the corresponding weight ratio, i.e., ω_i, and the final summation of all the input signals is the full input, i.e., $\sum_{i=1}^{n} \omega_i x_i$. The final input signal is also processed once inside the neuron, i.e., the threshold θ of the neuron is subtracted, and the final result is the final input signal after processing, i.e., $\sum_{i=1}^{n} \omega_i x_i - \theta$.

For a specific neuron, it has a specific function which takes the input signal of the neuron as the independent variable and the output signal as the dependent variable, in this model, we choose the Sigmoid function as the transformation function of the neuron, and the final processed input value will be used as the independent variable of the function, and the obtained dependent variable will be the output signal, and the above process is the process of the neuron accepting a number of input signals and transforming them into output signals. The above process is the process in which the neuron receives a number of given input signals and transforms them into output signals. Therefore, the output signal, i.e., the dependent variable, is determined by the input signal, i.e., the independent variable [4,5].

For this model, because the transformation function is already determined, we need to do is to determine is the weight ratio ω_i and threshold θ, as long as to find the appropriate weight ratio and threshold, then the model can realize in the known input signal after the prediction of a reasonable output signal function, that is, in the prediction of all kinds of conditions are known to be able to predict the actual amount of goods. Here we introduce the mean square error:

$$\boldsymbol{E_k = \frac{1}{2}\sum_{j=1}^{l}(y_j^k - y_j^k)^2} \tag{7}$$

where $\frac{1}{2}$ is used for subsequent derivation and the input $\beta_j = \sum_{h=1}^{q} \omega_{hj} b_h$ of the jth output neuron and the threshold value of the jth output layer neuron is θ_j. The weight ratio and threshold value corresponding to the minimum mean square error are the ideal values, and also the parameters corresponding to the minimum error between the prediction result and the actual result. Therefore, the next step to be performed is to find the parameters corresponding to the minimum mean square error.

Here we adjust the parameters based on the gradient descent strategy (take the adjustment of ω as an example). After completing K rounds of iterations, ω_k can be obtained, and using Taylor's formula, expanding the error function E(ω) at ω_k, we have:

$$\boldsymbol{E(\omega) = E(\omega_k) + E'(\omega_k)(\omega - \omega_k)} \tag{8}$$

The above equation ignores the residual term R. After performing k+1 rounds of iterations, we let:

$$\boldsymbol{\omega = \omega_{k+1}} \tag{9}$$

Then there is:

$$\boldsymbol{E(\omega_{k+1}) = E(\omega_k) + E'(\omega_k)(\omega_{k+1} - \omega_k)} \tag{10}$$

In this paper, it is desired that the newly optimized parameters make the model's error smaller, then there is:

$$\boldsymbol{E(\omega_{k+1}) - E(\omega_k) = E'(\omega_k)(\omega_{k+1} - \omega_k) < 0} \tag{11}$$

Here we can find that the special solution $\omega_{k+1} - \omega_k = -E'(\omega_k)$ can be found to make the above equation hold, and its substitution into the above equation has:

$$\boldsymbol{E(\omega_{k+1}) - E(\omega_k) = -[E'(\omega_k)]^2 < 0} \tag{12}$$

On the other hand, the aforementioned proceeding Taylor expansion ignores the residual term R(ω), which is ignored on the assumption that ω_{k+1} is close to ω_k. In this case, a very small parameter η > 0 is introduced to reduce the degree of deviation, and η is referred to as the learning rate, which is then available:

$$\boldsymbol{\omega_{k+1} - \omega_k = -\eta E'(\omega_k)} \tag{13}$$

Therefore, there is an iterative formula:

$$\boldsymbol{\Delta\omega_{hj} = -\eta E'(\omega_{hj}) = -\eta \frac{\partial E_K}{\partial \omega_{hj}}} \tag{14}$$

By the law of chains:

$$\boldsymbol{\frac{\partial k}{\partial \omega_{hj}} = \frac{\partial E_k}{\partial y_j^k} \cdot \frac{\partial y_j^k}{\partial \beta_j} \cdot \frac{\partial \beta_j}{\partial \omega_{hj}}} \tag{15}$$

$$\boldsymbol{g_j = -\frac{\partial E_k}{\partial y_j^k} \cdot \frac{\partial y_j^k}{\partial \beta_j} = -(y_j^k - y_j^k)(y_j^k)'} \tag{16}$$

Since the activation function uses a Sigmoid function, i.e:

$$\boldsymbol{f(x) = \frac{1}{1-e^{-x}}} \tag{17}$$

Then there is:

$$\boldsymbol{(y_j^k)' = y_j^k(1 - y_j^k)} \tag{18}$$

Further:

$$\boldsymbol{\frac{\partial E_k}{\partial y_j^k} \cdot \frac{\partial y_j^k}{\partial \beta_j} = (y_j^k - y_j^k) y_j^k (1 - y_j^k) = g_j} \tag{19}$$

By the definition of β_j:

$$\boldsymbol{\frac{\partial \beta_j}{\partial \omega_{hj}} = b_h} \tag{20}$$

Associating the above equations, we get:

$$\boldsymbol{\Delta\omega_{hj} = \eta g_j b_h} \tag{21}$$

Similar there:

$$\boldsymbol{\Delta\theta_j = -\eta g_j} \tag{22}$$

Through this method, suitable parameters can be derived through continuous training and used for actual prediction, which is the solution process of the model.

In the first medal model constructed in this paper, the number of medals in period t-1, the number of participants in the last 3 sessions, the total number of participations of the current participants, and the average number of participations of the current participants are used as input variables. Compared with the medal prediction, the variable of the number of participants in the last 3 sessions is added in this section because if the number of participants in the last 3 Olympic Games of a country shows a steady increase, it usually means that the country's investment in sports is increasing, and the increase in the number of participants is often accompanied by an increase in the number of events.

3.2. Model Prediction Results

The convergence curve and error histogram of the training process of the BP neural network constructed in this paper are shown Fig. 6 and Fig. 7 below:

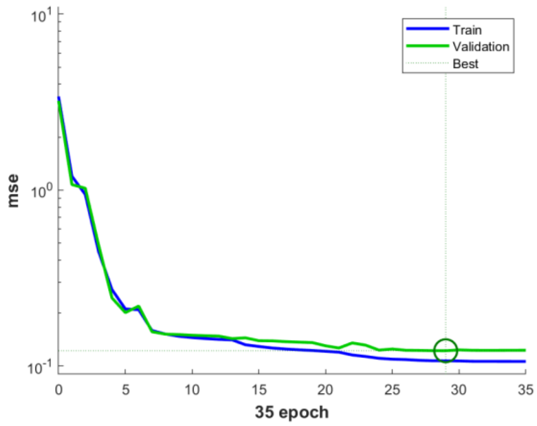

Fig. 6 Convergence Curve

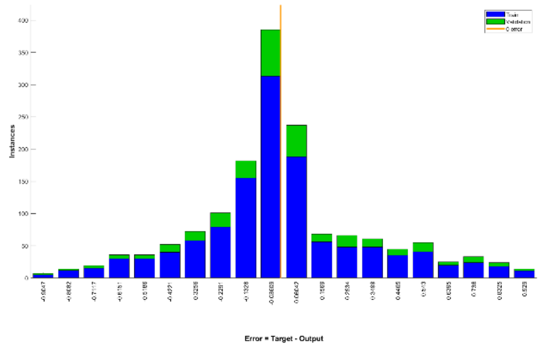

Fig. 7 Error Histogram

In order to quantify the predictive performance and reliability of the model, Accuracy, Precision, Recall and F1 Score are introduced to evaluate the performance of the model.

$$\text{Accuracy} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TN+FP+FN}} \tag{23}$$

where TP (True Positive) denotes the number of samples that the model correctly predicted as positive categories; TN (True Negative) denotes the number of samples that the model correctly predicted as negative categories; FP (False Positive) denotes the number of samples from negative categories that the model incorrectly predicted as positive categories; and FN (False Positive) denotes the the number of negative category samples that the model incorrectly predicted as positive.

Accuracy is the proportion of samples predicted by the model to be in the positive category that are actually in the positive category. It measures how accurately the model predicts a positive category. The formula is as follows:

$$\text{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}} \tag{24}$$

Recall is the proportion of all samples that are truly positive categories that the model correctly predicts as positive. It measures the model's ability to recognize positive category samples. The formula for calculating it is given below:

$$\text{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \tag{25}$$

The F1 score is the reconciled average of precision and recall, which combines the accuracy and recall of the classifier. The formula for the F1 score is as follows:

$$\boldsymbol{F1} = \boldsymbol{2} \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{26}$$

The results of the model test are shown in the Table 2 below:

Table 2. Model Performance

Norm	Value
Accuracy	0.906433
Precision	0.942857
Recall	0.908257
F1 Score	0.925234

As can be seen from the above metrics, the overall prediction performance is good. Accuracy is 0.906433, which means that more than 90% of all samples predicted are correct, indicating that the model has high reliability in distinguishing between countries that can win the first medal. Precision is 0.942857, which indicates that among the countries predicted to be able to win the first medal, a high percentage of them actually win the first medal, which means that the model's prediction is more accurate and the false alarm rate is relatively low. Recall is 0.908257, indicating that among all the countries that can actually win the first medal, the model successfully predicted a proportion of 90%, which reflects that the model is more capable of identifying the countries that can really win the first medal, with a low underreporting rate.

The F1 Score is 0.925234, which indicates that the model has achieved a good balance between precision and recall, and the overall prediction performance is more stable.

The top 10 countries predicted to be likely to win their first medal in 2028 are shown in the Table 3 below:

Table 3. 2028 Projections of First-Time Award Winning Countries

Country	2028 Probability of winning
AIN	0.466108585
LAT	0.414940378
SGP	0.408135654
ZAM	0.406546934
PAR	0.351874034
ANG	0.349894258
GUI	0.311911286
MLI	0.299912249
VIE	0.291186982
BOT	0.271151827

4. Conclusion

We constructed a multi - input multi - output LSTM medal prediction model to forecast the 2028 medal wins. The results show that the US, with its strong talent reserves and training systems in traditional sports like track and field and swimming, is likely to top the gold - medal count. China, making efforts in both emerging and traditional sports, maintaining strength in table tennis, diving and gymnastics while making breakthroughs in track and field and swimming, will rank second. The UK, with its advantages in cycling and rowing, will be in the top three. The prediction's consistency with historical trends verifies the model's reliability at the macro - level.

Based on the BP neural network model, AIN and LAT show potential in the 2028 Olympics, with winning probabilities of 0.466 and 0.415. The model has low prediction errors around 0 and an F1 score of 0.93, indicating high accuracy in predicting low - probability winning events.

Overall, this study offers data support for national sports departments to plan strategies, optimize resource allocation, and promote global sports development. It also provides references for sports event commercial operation and media communication, enabling industries to plan in advance for economic and social benefits.

References

[1] HOPE ZHANG, QIANKUN ZHU, XIANYU WANG, et al. Prediction of strain data for bridge monitoring based on VMD-KPCA-LSTM[J]. Journal of Applied Basic and Engineering Sciences,2025,33(01):76-86.DOI:10.16058/j.issn.1005-0930.2025.01.007.

[2] YANG Lei,HU Tonghao,WANG Wenxuan,et al. Prediction of process temperature attainment and stability in cigarette workshop based on RNN neural network[J]. HVAC,2024,54(S2):219-224.

[3] LU Yunxiang, TANG Ningxia, XU Zhenghui. Characterization of medal distribution of Chinese team in Paris Olympic Games[J]. Sports Science and Technology Literature Bulletin,2024,32(11):27-30.DOI:10.19379/j.cnki.issn.1005-0256.2024.11.007.

[4] Junyi Li,Xingxing Wang,Xiang Chen,et al. SOH prediction of lithium battery based on AO-AVOA-BP neural network model[J/OL]. Electronic Measurement Technology,1-10[2025-02-07].http://kns.cnki.net/kcms/detail/11.2175.tn.20250206.1445.050.html.

[5] G. H. Zhang, C. L. Wang, H. W. Wang, et al. Improved particle swarm optimization algorithm combined with BP neural network model for the prediction of total phosphorus concentration in water transmission spectra[J]. Spectroscopy and Spectral Analysis,2025,45(02):394-402.