Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 13, No. 2, 2025

Interpretable Olympic Medal Forecasting Using Hybrid GA-LSTM with SHAP Explanations

Hongzhe Zhao *

School of Science, Liaoning Technical University, Fuxin, Liaoning, China
*** Corresponding author Email:** 15247041570@163.com

Abstract: In recent years, more and more people have been paying attention to the Olympic Games and who will win the Olympic MEDALS. This study focuses on the prediction of Olympic medals and the analysis of related factors. It constructs a GA-LSTM model to forecast the number of medals each country will win at the 2028 Los Angeles Summer Olympics and employs the SHAP model to interpret the factors influencing medal counts. Helps to understand the development trend of sports competitiveness in various countries and provides a reference for better allocation of sports resources.

Keywords: GA-LSTM Medal Prediction Model; SHAP Model; Hypothesis Testing.

1. Introduction

In the context of today's globalization, the Olympic Games, as one of the most influential global sporting events, have attracted widespread attention from countries around the world. Accurately predicting the number of medals each country will win at the Olympics not only helps in understanding the development trends of national sports competitiveness but also provides a reference for the rational allocation of sports resources. Therefore, this paper conducts an in-depth exploration of these issues by constructing data-driven models.

2. Based on the GA-LSTM Medal Prediction Model

To predict the number of Olympic medals each country will win, including both gold and total medals, we developed a Long Short-Term Memory (LSTM) model. Its hyperparameters were optimized via the Tree-structured Parzen Estimator (TPE) algorithm. This model leverages LSTM's ability to capture complex temporal dependencies and non-linear feature interactions in historical data. Within a Bayesian optimization framework, TPE efficiently explores the hyperparameter space to enhance model performance. Using this approach, we forecast the distribution of future Olympic medals and conduct a quantitative analysis of the uncertainty in these predictions.

2.1. LSTM Model Construction

(1) Data preparation

Input characteristics: including country code (NOC), historical medal distribution, host country logo, number of sports participation, number of awards in each event, etc.

Output characteristics: Gold, Silver, Bronze, and total number of medals.

(2) Dataset division

In the previous section, the data has been preprocessed, and the feature X and target variable Y are extracted directly from data3, and the data is divided into training and test sets at a 7:3 ratio:

$$X_{\text{train}}, X_{\text{test}}, Y_{\text{train}}, Y_{\text{test}} = \text{split}(X, Y, \text{test\_size}=0.3) \tag{1}$$

The training set is used to learn the parameters of the model, and the test set is used to evaluate the predictive performance of the model.

(3) Encoder and decoder

Encoder: Maps the input sequence X to an implicit representation h of a fixed dimension.

$$h_t = f_{\text{enc}}(x_t, h_{t-1}) \tag{2}$$

where h_t represents the hidden state of the encoder at time step t, and f_{enc} is usually an LSTM or GRU unit.

Decoder: Based on the encoder's implicit representation h, the decoder generates the target sequence Y.

$$y_t = f_{\text{dec}}(y_{t-1}, h_{t-1}) \tag{3}$$

where f_{dec} is the nonlinear mapping function of the decoder, usually using LSTM or GRU units.

(4) Core components

The LSTM [4] model is mainly composed of the following components: forgetting gate, input gate, cell state, output gate, and cell

(5) Status updates

Cell Status Update: $C_t = f_t C_{t-1} + i_t \tilde{C}_t$

Hide Status Updates: $h_t = o_t \tanh(C_t)$

The specific optimization steps are shown in the following figure 1.

2.2. GA Hyperparameter Optimization

(1) Search space definition

Genetic algorithms give the evolution of populations to explore space, so define a broader search space within this range to find the optimal solution:

Learning rate α : $[0.0001, 0.1]$

Hides the number of layer elements $h:[32,512]$ Regularization coefficient $\lambda:[0.0001,0.1]$

Batch size $b:[16,128]$

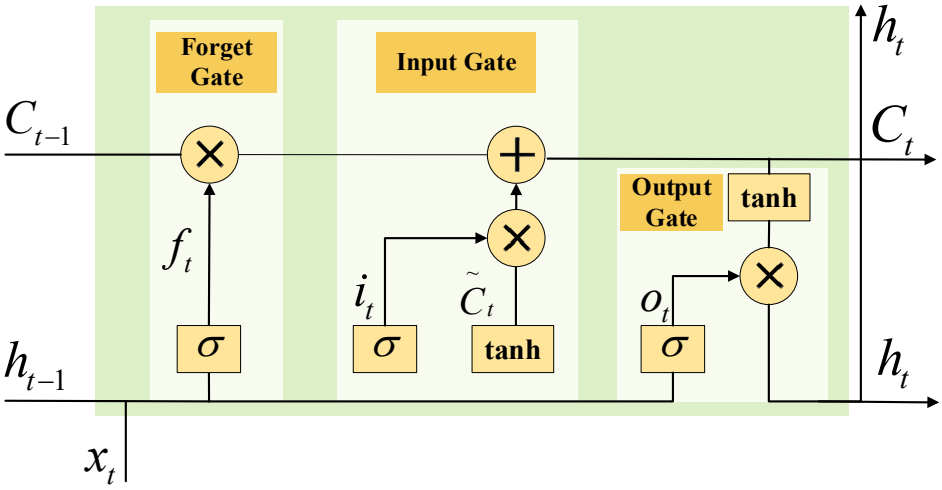


Fig 1. LSTM flowchart

(2) Objective function

The goal of the genetic algorithm is to maximize the fitness function, which corresponds to the minimization loss function in machine learning, by defining the fitness function as the negative value of the loss function on the validation set, so that the genetic algorithm can minimize the loss expressed as follows:

$$\text{Fitness}(\theta) = -\text{L}_{\text{val}}(X, Y; \theta) \tag{4}$$

(3) Genetic algorithm steps

The genetic algorithm is mainly divided into the following four steps: encoding, selection, crossover, mutation, and the specific process is shown in the following figure:

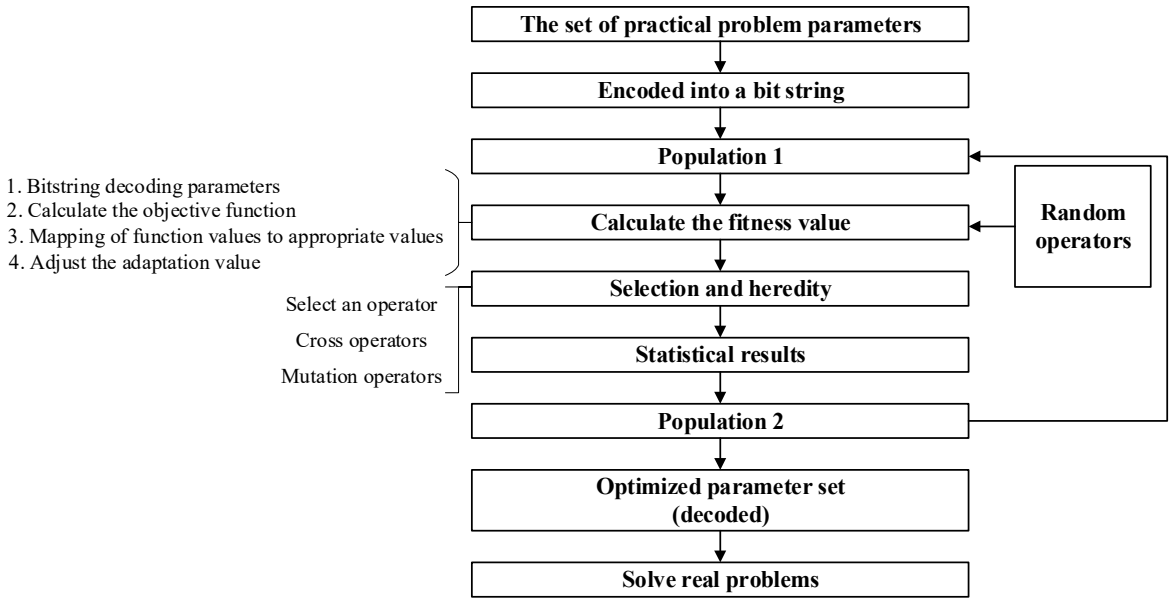


Fig 2. Genetic algorithm flow chart

2.3. Page Numbers Forecast Uncertainty Analysis

Monte Carlo simulations and confidence intervals were used to assess the uncertainty of the model's predictions. Using the trained LSTM model, the N sets of prediction values were generated by multiple sampling of random perturbation input data, and the mean and standard deviation of the prediction distribution were calculated.

$$\mu_{\hat{y}} = \frac{1}{N}\sum_{i=1}^{N}\hat{y}_i, \quad \sigma_{\hat{y}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - \mu_{\hat{y}})^2} \tag{5}$$

Assuming that the distribution of the predicted values satisfies the normal distribution, the confidence level $\alpha = 0.05$ is set, and the confidence interval of the predicted values is calculated:

$$\text{CI} = \left[\mu_{\hat{y}} - z_{\alpha/2}\cdot\sigma_{\hat{y}}, \mu_{\hat{y}} + z_{\alpha/2}\cdot\sigma_{\hat{y}}\right] \tag{6}$$

The stability of the predicted value is evaluated by the width of the confidence interval, and the narrower the width, the more reliable the prediction of the model. At the same time, the confidence interval is analyzed to verify the validity of the model's predictions.

3. Model Solving and Result Analysis

3.1. Hyperparameter Optimization

In hyperparameter optimization, a genetic algorithm is used to search for multiple hyperparameters, with the goal of reducing the loss function of the classification task. By defining the search space and multiple iterations, the loss function of the model on the validation set is gradually stabilized, and the optimal combination of hyperparameters is finally determined. The specific results are shown in the table below:

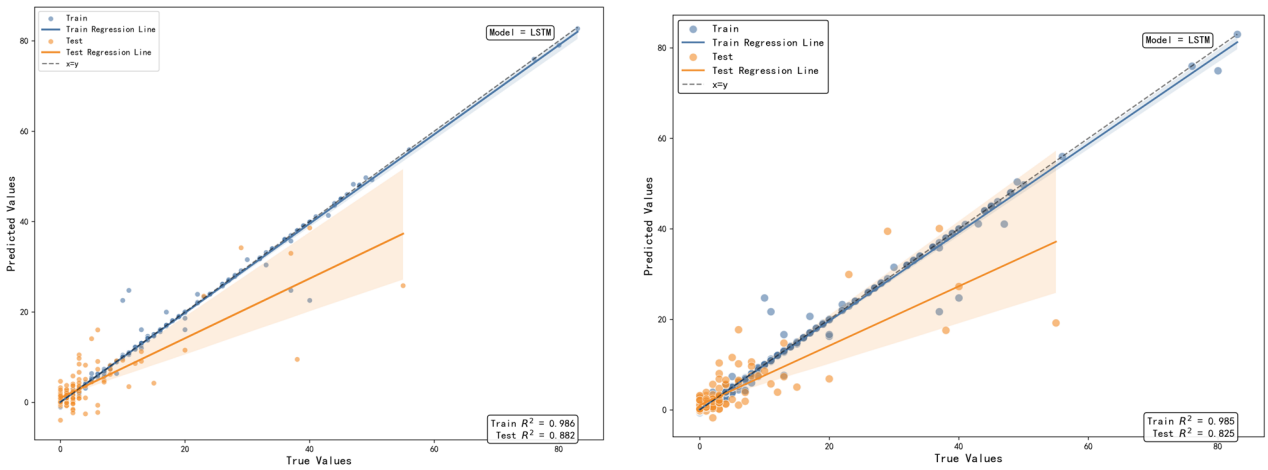
Table 1. Hyperparameter optimization results of GA-LSTM model

HYPERPARAMETERS	RANGE	BEST VALUE
Learning rate α	[0.0001,0.1]	0.01
Hidden layer h	[32,512]	256
Batch size b	[16,128]	64
Regularization coefficient	[0.0001,0.1]	0.0002
Layers of coding	[1,5]	2
Decoding Layers	[1,5]	2
Time steps	[2,15]	10
Activation function	[ReLU,Tanh]	ReLU

Through the optimization results, it can be seen that the learning rate, the number of hidden layer elements, and the batch size are the key hyperparameters that affect the performance of the model. The small learning rate supports the model to converge steadily, while the medium-sized number of hidden layer units and batch size balance the expressiveness and training efficiency of the model. The selection of regularization coefficients further inhibits the overfitting tendency of the model. Combined with the optimal number of time steps and the selection of activation functions for the model, the combination of these hyperparameters provides a solid foundation for improving the performance of the model.

3.2. Model Testing

In the initial training process of the LSTM model, under the default parameter setting, the R^2 of the model on the training set reached 0.985, while the R^2 of the test set was only 0.825, indicating that the model had an overfit. In order to alleviate this problem, the model hyperparameters were adjusted through GA [5] optimization, including reducing the number of hidden layer elements, increasing the regularization coefficient, and adjusting key parameters such as batch size and time steps. After the adjustment, the performance of the model was significantly improved, and the R^2 on the training set and the test set reached 0.986 and 0.882, respectively, effectively reducing the phenomenon of overfitting.



True vs Predicted Values　　True vs Predicted Values Model Performance

Fig 3. Comparison and Analysis of LSTM and GA-LSTM

Before the adjustment, the model fit the training set very well, with the point cloud densely distributed near the reference line, showing almost no deviation. However, on the test set, the point cloud distribution was more scattered, and the regression line was significantly different from the reference line, indicating weaker generalization ability. After optimization using the Genetic Algorithm (GA), the point cloud of the test set was notably contracted, with a distribution closer to the reference line, indicating improved prediction accuracy and stability. At the same time, the performance of the training set slightly decreased but remained at a high level, suggesting a more balanced overall performance of the model. The optimized LSTM model exhibits more consistent performance between the training set and the test set, effectively avoiding overfitting and enhancing its generalization ability.

3.3. Model Prediction

(1) Construct forecast data

In order to predict the number of Olympic medals in 2028, a new dataset of input features is first constructed. The dataset is based on existing data for the 2024 Olympic Games, with adjustments for the new events at Los Angeles 2028. The specific process is as follows:

1.Screening 2024 data

By selecting relevant records for 2024 from the original dataset, the underlying dataset X_{24} is constructed, from which the data for Russia is excluded due to the ban on Russia in 24.

$$X_{28} = X_{24}[(X_{24}[' year'] = 2024 \wedge X_{2024}[' NOC'] \neq 113] \tag{7}$$

2.Reset the index

Reindex the filtered data

3.New project forecast

According to the new sports approved by the IOC (cricket, squash, baseball and softball, tennis and flag football), the corresponding number of medals will be adjusted, and one gold medal will be added for each of the men's and women's categories according to the adjustment rules, and the corresponding formula will be adjusted as follows:

$$\begin{aligned} X_{28}['Baseball'] &= X_{28}['Baseball'] + 2 \\ X_{28}['Softball'] &= X_{28}['Softball'] + 2 \\ X_{28}['Cricket'] &= X_{28}['Cricket'] + 2 \\ X_{28}['Sixes'] &= X_{28}['Sixes'] + 2 \\ X_{28}['Squash'] &= X_{28}['Squash'] + 2 \\ X_{28}['Flagfootball'] &= X_{28}['Flagfootball'] + 2 \end{aligned} \tag{8}$$

4.years of renewal

Change the year information in the data to 2028

5.Host country logo

For the United States (NOC=147), as the host country of the 28 Olympic Games, it is marked as 1 and other countries as 0

6.Contestant assumptions

Assuming that the 28-year entrants and the '24 entrants remain the same, all variables are still 24-year data

Through the dataset of the above steps, the predictive analysis of the medal prediction model can be further carried out.

(2) Performance analysis

1.Los Angeles 2028 Summer Olympics medal table with confidence intervals

Based on the established medal prediction model, the number of medals in each country for the 2028 Summer Olympics in Los Angeles was predicted, and the uncertainty analysis was carried out to obtain the corresponding prediction intervals. Here are some of the predictions for some countries, including the number of gold, silver and bronze medals and their confidence intervals:

Table 2. Medal table (sorted by number of gold medals)

	Gold	Silver	Bronze	Total
United States	40	45	42	127
China	39	27	24	90
Japan	20	13	13	46
Australia	18	19	17	54
France	15	25	20	60
Great Britain	14	22	29	65
Netherlands	14	7	13	34

Table 3. Medal confidence interval

	Gold		Silver		Bronze	
	lower	upper	lower	upper	lower	upper
United States	37	41	33	46	37	42
China	36	40	26	27	20	25
Japan	13	20	11	14	10	13
Australia	12	18	8	19	12	18
France	12	16	19	25	16	20
Great Britain	9	15	18	23	22	30
Netherlands	8	14	6	8	8	14

The chart displays the predicted quantities of gold, silver, and bronze medals along with their confidence intervals. Curves of different colors represent the predicted values for each type of medal, with shaded areas indicating the corresponding confidence intervals. As the number of data points increases, the predicted values tend to stabilize, but there is significant fluctuation at the beginning, indicating higher uncertainty. The width of the confidence intervals reflects the precision of the predictions; wider intervals suggest higher uncertainty, while narrower intervals indicate more accurate predictions.

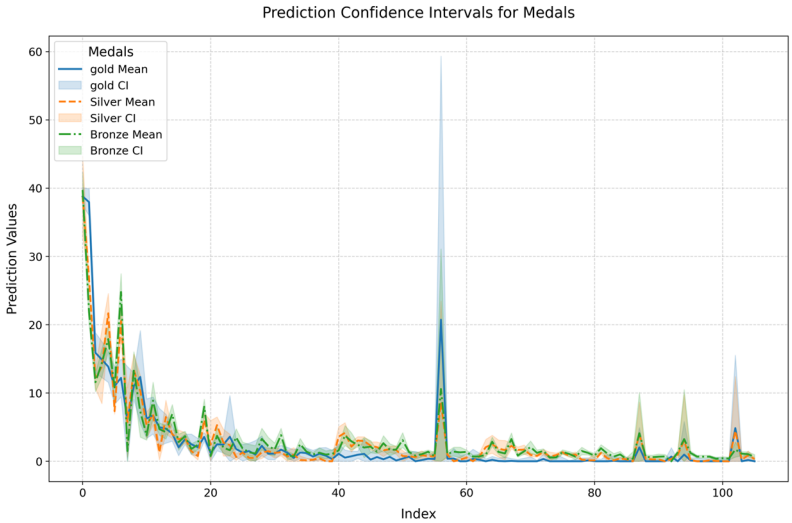


Fig 4. Prediction Confidence Intervals for Medals

By comparing the projected medal totals for 2024 and 2028, the following countries are expected to significantly improve their medal results:

Table 4. Countries that are expected to significantly improve their medal results

NOC	2024 Total	2028 Total	Improvement
ROC	0	17	17
Poland	7	13	6
Jordan	1	4	3
Mixed team	0	3	3
Greece	3	6	3
Thailand	4	7	3
Mexico	4	7	3
Morocco	1	3	2

The countries in the table above are projected to see significant growth in medals in 2028, especially ROC and Poland, both of which have significantly increased their totals, reflecting the rapid development of sports in these countries.

The following countries are expected to see a decline in performance in 2028 compared to the countries that have improved:

Table 5. Countries with projected declines in 2028

NOC	2024 Total	2028 Total	Improvement
South Korea	32	24	-8
Pakistan	8	2	-6
Iran	12	8	-4
France	64	60	-4
Ethiopia	8	5	-3
Portugal	8	5	-3
Indonesia	7	4	-3
Chinese Taipei	7	5	-2

The total number of medals in these countries is expected to decline in 2028, with South Korea and Pakistan in particular seeing a reduction of 8 and 6 medals respectively, reflecting the potential of their potential for future Olympic events.

2.Predict the country that won a medal for the first time

For countries that have not won medals in international sporting events, the following countries are likely to win medals in the 2028 Olympic Games through the prediction analysis of the model:

Table 6. Countries that won their first medal at the 2028 Olympic Games

	Gold	Silver	Bronze	mean
Andorra	0.5346	0.4769	0.6882	0.5666
Anguilla	0.5346	0.4886	0.6882	0.5705
Aruba	0.5262	0.4997	0.6823	0.5694
ROC	0.9734	0.9980	0.9991	0.9902
South Sudan	0.5203	0.5561	0.8110	0.6291
Syria	0.5269	0.5260	0.7764	0.6098
Vietnam	0.5273	0.5382	0.6852	0.5836
Zambia	0.5428	0.5297	0.7490	0.6072

It can be seen that ROC and South Sudan have the highest probability of winning a medal in 2028 at 0.9902 and 0.6291 respectively, indicating that they have a better chance of winning a medal in 2028.

4. SHAP Model Analyzes Impact of Olympic Events on Medal Performance

4.1. Establishment of the SHAP Model

To study the association between Olympic events and the number of medals won by countries, we used the SHAP (Shapley Additive Explanations) model to assess the specific impact of different Olympic events on the total number of medals won by countries.

The SHAP [6] model is a method for explaining machine learning model predictions, which draws on the Shapley value in cooperative game theory to evaluate the contribution of each feature to the model's output in a fair and transparent way.

For a given feature f and model output y ,the Shapley value $\phi(f)$ represents the contribution of feature f to the model prediction results, and is calculated as follows:

$$\phi(f)=\sum_{S\subseteq N\setminus\{f\}}\frac{|S|!(|N|-|S|-1)|!}{|N|!}[f(S\cup\{f\})-f(S)] \tag{9}$$

Where: S is the set of features; N is the set of all the features; $f(S)$ is the result of model prediction using only feature set S ; $f(S\cup\{f\})$ is the prediction result after adding feature f to feature set S ;$|S|$ and$|N|$ represent the size of sets S and N , respectively.

When using the SHAP model, we first use a trained model (e.g., the GA-LSTM model) to predict the number of medals

in the Olympic Games, which is based on a variety of input features. The model was then analyzed using the SHAP tool to assess the specific impact of each characteristic (e.g., number of sports, category, past medal statistics, etc.) on the total number of medals in different countries.

By using the SHAP model to delve into the relationship between Olympic events and the number of medals, we can provide valuable data for countries to formulate sports development strategies.

4.2. Solve the SHAP model

4.2.1. Analysis of the Impact of Olympic Events on the Number of Medals in Each Country

The results of the SHAP model show that the number and variety of Olympic sports have a significant impact on the gold, silver and bronze medals of each country, and these effects can be visualized through charts.

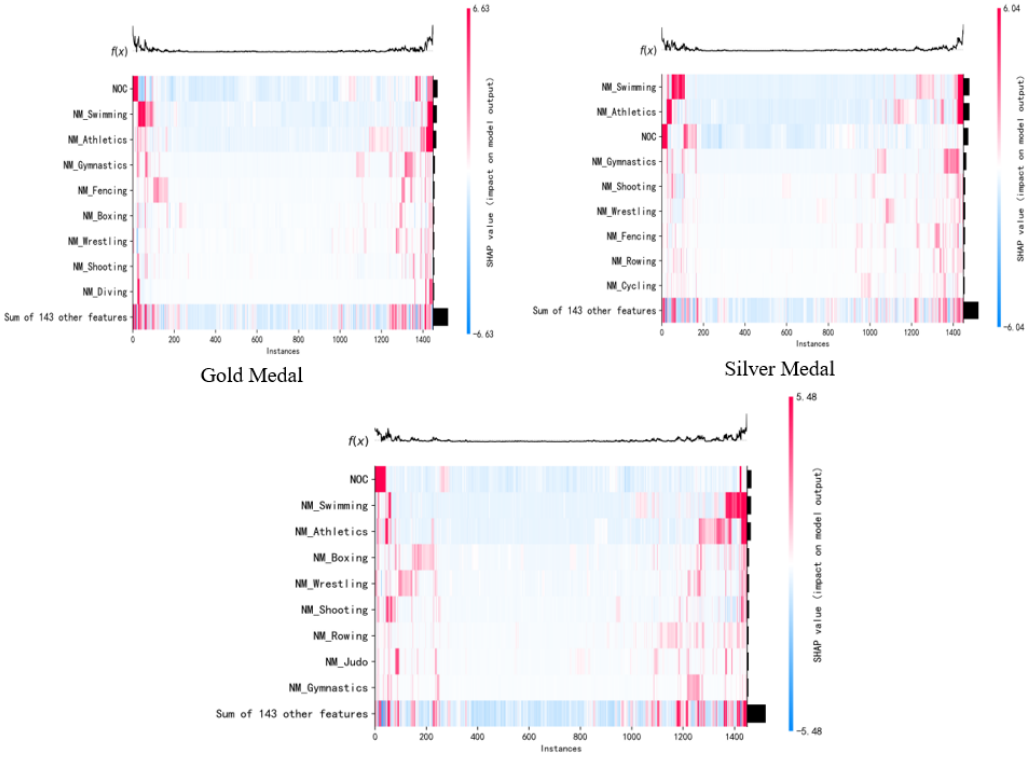


Fig 5. Impact of Olympic events on the number of medals by country

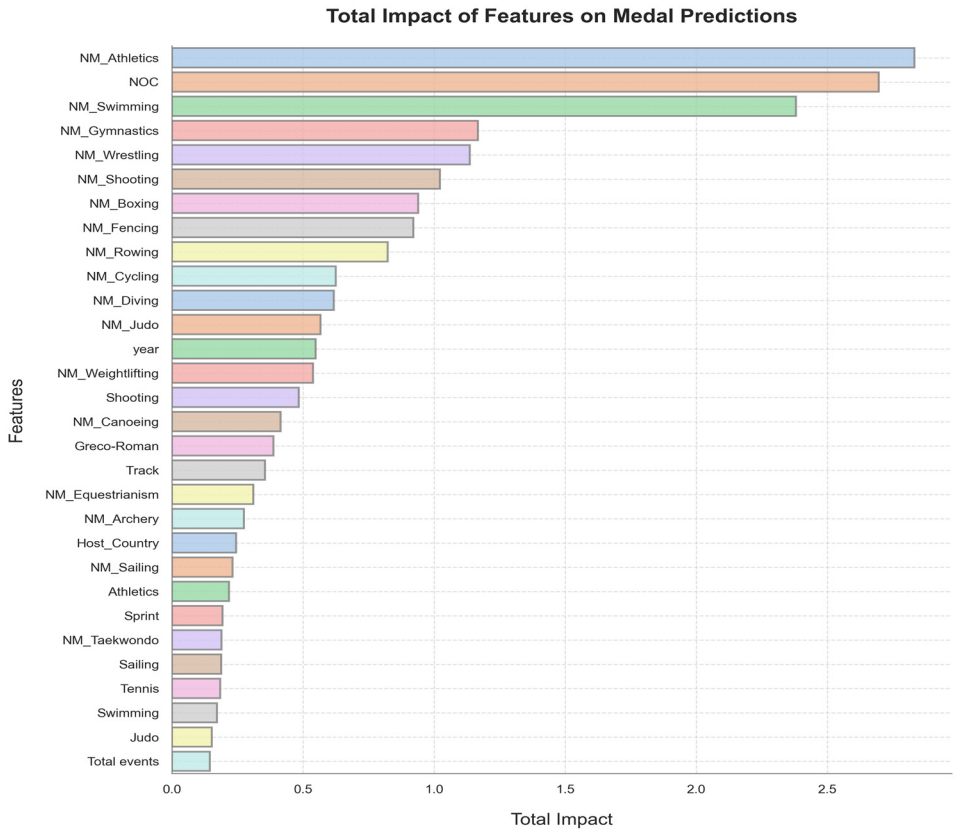


Fig 6. Total Impact of Features on Medal Predictions

The diagram shows that different Olympic events have different impacts on gold, silver and bronze medals. Athletics and swimming have a significant impact on gold and bronze medals, while fencing and gymnastics have a significant impact on silver medals. The number of events plays a major role in the overall medal prediction, while the type of event affects the details of medal distribution. The host nation may have an advantage in both gold and silver medals, especially in its strengths.

Based on the summary results of the SHAP values of the gold, silver, and bronze medals, it is possible to observe the magnitude of the influence of individual items and characteristics on the medal prediction in the model, and these effects have been ranked in descending order. The chart shows the contribution of each characteristic to the gold, silver and bronze medals, where "NM_" represents the number of events, "NOC" represents the country code, "Host_Country" indicates whether it is the host country, etc.

The analysis shows that athletics and swimming have the greatest impact on gold medals, while wrestling, gymnastics and shooting have the greatest impact on silver and bronze medals. The host country is likely to have more medals due to home advantage. Although events such as weightlifting and canoeing have a lesser impact, they are still an important source of medals in specific countries.

4.2.2. SHAP Value Analysis: Impact Analysis of the U.S. 2028 Olympic Medal Prediction

Based on the projections for the gold, silver, and bronze medals at the 2028 U.S. Olympic Games, the following graphs illustrate the impact of different characteristics on the number of medals projections through the contribution analysis of SHAP values.

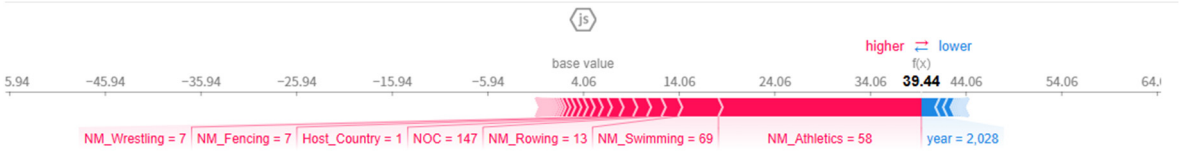


Fig 7. The effect of different characteristics on the number of gold medals

The chart shows that athletics and swimming have the greatest impact on the SHAP analysis of the number of gold medals in the United States, and the host country advantage also adds to the gold medal prediction.

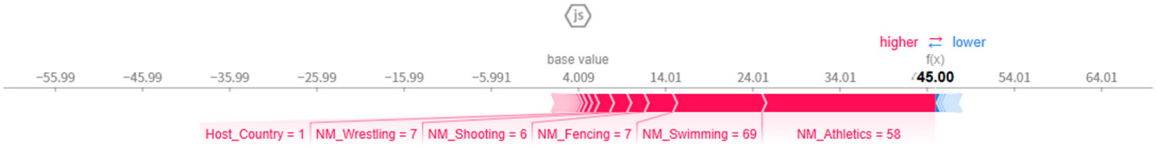


Fig 8. The impact of different characteristics on the number of silver medals

The analysis of the SHAP values in the graph shows that swimming has the greatest impact on the number of silver medals in the United States, and track and field is also a key factor, both of which contribute significantly to the total number of medals.

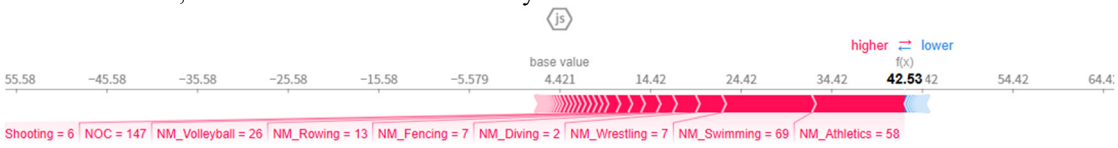


Fig 9. The effect of different characteristics on the number of bronze medals

The analysis shows that swimming and track and field contributed significantly to the U.S. bronze medal, while shooting and volleyball also had a significant impact on the total number of bronze medals.

The summary shows that track and field and swimming are the main contributors to the total number of Olympic medals in the United States, while events such as wrestling, fencing, shooting, and volleyball also have a significant impact on the medal count. These analyses help the U.S. Olympic team strategize, optimize resources, and focus on the most successful events.

5. Additional Insights on the Distribution of Olympic Medals

Based on the established SHAP model, GA-LSTM model and hypothesis testing model, the following original insights on the distribution of Olympic medals are revealed:

1. The impact of Olympic events on medals: The analysis of the SHAP model shows that different Olympic sports (such as track and field, swimming, fencing, etc.) have different degrees of influence on the medals of various countries, especially the gold medal and bronze medal. This insight provides NOCs with strategies for optimizing resources and concentrating their strengths, such as the U.S. in athletics and swimming, which can focus resources on competitiveness.

2. Host country advantage: SHAP analysis shows that host countries usually have a certain advantage in gold and silver medals due to home field advantage. This provides an important strategic basis for NOCs, especially before the Games, when it can predict the number of medals and make resource adjustments based on the host country's characteristics and historical performance.

3. Uncertainty in medal predictions: The GA-LSTM model helps NOCs understand the expected number of medals in different scenarios, especially in some specific events, by predicting the number of Olympic medals and providing confidence intervals, which helps the NOCs to understand the expected number of medals in different scenarios, especially in some specific events, and helps to identify potential challenges and opportunities in advance.

Through the analysis of these models, NOCs can adjust athlete training plans, funding allocations and participation strategies to improve the success rate of medals based on the performance potential of different events.

References

[1] Gong Fengjie, Zhou Conghua. Asthma missing value filling method based on improved [J]. Computer and Digital Engineering, 2024,52(08):2284-2288+2335.

[2] Siru W, Guoqi Q, John H. Integrated logistic ridge regression and random forest for phenotype-genotype association analysis in categorical genomic data containing non-ignorable missing values[J]. Applied Mathematical Modelling,2023,1231-22.

[3] Miao Changqing, Lv Yuekai, Wan Chunfeng. Prediction model of mechanical properties of corroded steel wire for bridge cable based on GA-LSTM[3][J/OL].Journal of Southeast University (Natural Science Edition),1-10[2025-01-27].http:// kns.cnki. net/ kcms/ detail/32. 1178.N.20241129.1639.012.html..

[4] ZHAO Nanyang, LIU Chao, DU Wenlong, et al. Health assessment of ship diesel engine based on LSTM prediction and cloud center of gravity evaluation[J/OL].China Ship Research, 1-8[2025-01-27].https:// doi.org/10.19693/j. issn. 1673-3185.04077.

[5] Xianyi L The optimization method of CNC lathe performance based on Morris sensitivity analysis and improved GA algorithm [J]. Journal of Vibroengineering,2024,26(2):438-454.{GA}.