World Journal of Information Technology
ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3034

A HYBRID SEQ2SEQ AND BAYESIAN OPTIMIZATION FRAMEWORK FOR PREDICTING OLYMPIC MEDAL DISTRIBUTION WITH UNCERTAINTY ANALYSIS

QiuLin Yao[1*], Feng Cheng[1], YanPeng Guo[1], KuiSong Wang[1], QiSheng Liu[1], Ning Ding[2]
[1]School of Mechanical Engineering, Jiamusi University, Jiamusi 154007, Heilongjiang, China.
[2]School of Materials Science and Engineering, Jiamusi University, Jiamusi 154007, Heilongjiang, China.
Corresponding Author: QiuLin Yao Email: wy524887227@163.com

Abstract: This paper proposes a robust and data-driven methodology to forecast Olympic medal outcomes with high accuracy and interpretability. A sequence-to-sequence (Seq2Seq) neural architecture is employed to learn temporal dependencies in national Olympic performance, while hyperparameter optimization is conducted using a Tree-structured Parzen Estimator (TPE) to enhance model generalization. To ensure data integrity, preprocessing steps include structured data cleansing and the use of a backpropagation neural network to address missing values. The model further integrates features such as national investment in sports, historical medal trends, and host country effects. In addition to deterministic predictions, uncertainty is quantified through Monte Carlo sampling and confidence intervals, providing probabilistic insights into future outcomes. Experimental results show that the proposed approach outperforms baseline models, achieving an R^2 improvement from 0.827 to 0.875 on the test dataset. The framework is applied to predict the medal distribution for the 2028 Los Angeles Olympics and highlights emerging medal-winning countries. These findings demonstrate the framework′s potential to assist national committees and policy makers in strategic planning for future Olympic participation.

Keywords: Olympic medal forecasting; Sequence-to-sequence neural network; Bayesian hyperparameter optimization; Prediction uncertainty quantification

1 INTRODUCTION

As the most influential sports event in the world, the number of Olympic medals is an important indicator to measure the strength of sports in different countries[1]. Accurate prediction of Olympic medals can not only satisfy the expectation of sports enthusiasts, but also serve as a key reference for the sports departments of different countries to formulate strategic planning and allocate resources. With the development of big data and artificial intelligence, the use of data-driven methods to predict the number of Olympic medals has become a research hotspot, bringing new opportunities for decision-making in the field of sports. Past studies have used a variety of methods in Olympic medal prediction Some studies have used traditional statistical models such as linear regression to make predictions by analyzing historical data and key information related factors, but such models are difficult to capture the complex nonlinear relationships in the data [2].

In the existing research on Olympic medal prediction, the traditional linear model is difficult to capture the complex nonlinear relationship due to structural limitations, the machine learning method is insufficient to mine the temporal dynamic features of the number of medals, and the neural network has problems such as lack of prediction uncertainty analysis and insufficient model depth. In view of these shortcomings, this study is improved in three aspects: constructing a hybrid architecture that fuses time series models and deep learning, and strengthening the modeling of nonlinear and time series features; A Bayesian deep learning framework is introduced to quantify the uncertainty of prediction results. Integrating multi-dimensional variables such as historical medal data, sports resource investment, and event characteristics, a high-precision special model is created to improve the accuracy of medal prediction for the Los Angeles 2028 Olympic Games, and deeply analyze the core factors affecting the medal performance of various countries, making up for the lack of model capabilities and analysis dimensions in existing research. In recent years, machine learning methods have been gradually applied to the field of decision counting, neural networks, etc. However, these models have limitations in dealing with the temporal and complex nature of the number of medals, and the accuracy needs to be improved, while the analysis of uncertainty in the prediction results is not deep enough. This study is only in overcoming the inadequacy of existing research to construct a high-precision prediction model for Olympic medals, and quantitatively analyze the uncertainty of prediction results. Specific objectives include predicting the number of medals for each country in the 2028 Olympic Games in Los Angeles and analyzing the factors affecting the performance of some countries. In terms of methodological innovation, this approach integrates time-series deep learning models with Bayesian frameworks, breaking through the limitations of traditional linear models in capturing nonlinear relationships, while also addressing the shortcomings of existing models in analyzing prediction uncertainty. In model construction, it incorporates multidimensional variables such as sports resource investment and event characteristics to create a highly accurate specialized model, deepening the systematic analysis of factors influencing medal outcomes. On the application level, it achieves precise predictions of medal counts for the 2028 Los Angeles

Olympics, and through variable attribution, it uncovers core influencing factors, providing deeper references for decision-making in sports departments, thus filling the gaps in existing research regarding model accuracy, uncertainty quantification, and multidimensional factor analysis.

2 MATERIALS AND METHODS

2.1 Data Acquisition and Pre-Processing

In this paper, this article collected information on the number of medals and athletes from each country from 1896-2024, which were obtained from the open source website https://www.contest.comap.com/undergraduate/contests/ Firstly, difference set analysis is used to supplement the countries that do not appear in the medal data, and then the item data is cleaned, including filling missing values, removing irrelevant fields, and handling special items [3]. Then the cleaned data is linked to the medal data by year, and the countries are inwardly linked and counted to win the awards in each program. Based on the error back propagation BP neural network, the missing values in the dataset are filled to ensure the integrity of the data, features are extracted, and the data is divided into a training set and a test set in the ratio of 7:3 [4]. The training set is used for learning the parameters of the model and the test set is used to evaluate the predictive performance of the model.

2.2 Methodology

The aim of this study is to predict the number of medals that each country will win at the Olympic Games and to estimate the uncertainty of the prediction results. First, data preprocessing and cleaning. After the data are organized, the features include the number of types of Olympic sports corresponding to each year, the number of medals, the country code, the distribution of medals in previous years, the logo of the host country, and the participation of each sport. Next, the number of gold, silver, and bronze medals won by each country each year is used as the dependent variable, and other features are used as independent variables to split the data into a training set and a test set. The task of predicting the number of medals can be handled by a variety of machine learning regression models. In order to improve the prediction performance and generalization ability of the model, this study uses a heuristic algorithm to optimize the hyperparameters of the model. The accuracy and stability of the model were ensured by adjusting the hyperparameters and using metrics such as R^2 for model evaluation. Regarding the uncertainty analysis of the prediction results, this study combines Monte Carlo simulation and confidence intervals to quantify the reliability of the model predictions. The uncertainty of the prediction results is comprehensively evaluated by introducing random perturbations to sample the input data multiple times, generating multiple sets of prediction results, and calculating the mean and standard deviation of the prediction distribution. Finally, after constructing the optimal model and completing the training, the prediction is made based on the relevant data of the 2028 Los Angeles Olympic Games, which provides scientific and accurate prediction and reference for the future Olympic medal distribution. Will add new items to adjust the relevant features, excluding Russia and other countries that do not participate, to get the medal list with its confidence interval, and then analyze the rise and fall of each country's performance); through the previous difference set analysis of the complementary non-winning countries, the model also predicts and evaluates which countries are likely to win medals for the first time in 2028, and provides the predicted probability, which can be nested into the activation function by the regression value.

In order to predict the number of Olympic medals (including the number of gold medals and the total number of medals) for each country, a sequence-to-sequence (Seq2Seq)-based deep learning model was developed and the hyperparameters of the model were optimized using a tree-structured Parzen Estimator (TPE) algorithm [5]. The model provides reliable predictions for future Olympic medal distributions by learning complex temporal and feature relationships in historical data, while the uncertainty of the prediction results is quantified and analyzed in detail.

The Seq2Seq model is a deep learning method for sequence prediction and consists of an encoder and decoder. The input features cover information such as country codes and the target variables are vectors about the distribution of medals [6]. The encoder transforms the input sequence into an implicit representation of fixed dimensions, and the decoder generates the target sequence based on this representation. The model is optimally trained by stochastic gradient descent using mean square error as loss function.

2.3 Model Evaluation Indicators

To assess the predictive ability of the model, R^2 was chosen as the main performance indicator:

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

Where y_i is the true value, $\hat{y}_i$ is the predicted value, and $\bar{y}$ is the mean value of the target variable. the closer the R^2 indicator is to 1, the stronger the explanatory power of the model on the target variable.

3 MODELING AND SOLVING

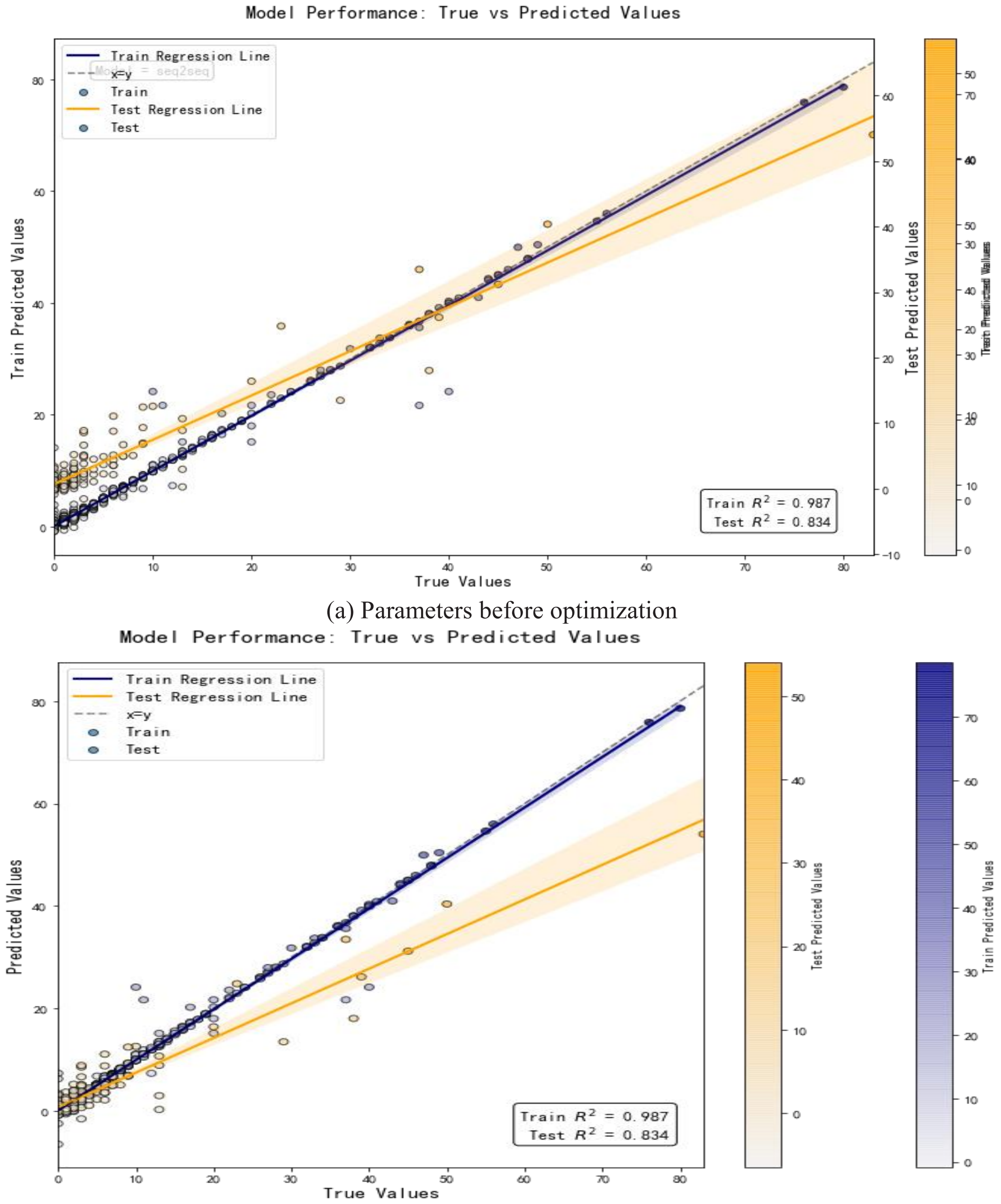
3.1 hyperparametric Optimization

The optimization results are as follows: the learning rate is 0.001, the number of hidden layer units h is 256, the batch size b is 64, the regularization coefficient λ is 0.0001, the number of encoder layers is 2, the number of decoder layers is 2, the time step T is 10, and the activation function is ReLU.

The above optimization results show that the learning rate α, the number of hidden layer units h and the batch size b are the key hyperparameters affecting the performance of the model. Among them, a small learning rate α ensures the stable convergence of the sub-model, while a medium-sized hidden layer unit number and batch size take into account the expressive ability and training efficiency of the model.

3.2 Model Checking

During the initial training of the Seq2Seq model, the R² of the model on the training set reaches 0.986 under the default parameter settings, while the R² of the model on the test set is only 0.827, which shows that the model is overfitting. To alleviate this problem, the hyperparameters of the model are adjusted through TPE optimization, including reducing the number of hidden layer units, increasing the regularization factor λ, and adjusting the key parameters such as the batch size and the number of time steps. After the adjustment, the performance of the model is significantly improved, and the R² on the training set and test set reaches 0.987 and 0.875, respectively, which effectively reduces the overfitting phenomenon, as shown in Figure. 1 below[7][8].



(a) Parameters before optimization

(b) TPE adjustment parameters

Figure 1 Model Performance: True vs Predicted Values

For the model under the recognized parameter settings before tuning, the fitting effect of the training set is very good, and the point cloud is densely distributed near the reference line with almost no deviation; however, in the test set, the dispersion of the point cloud is larger, and the regression line deviates from the reference line more obviously, indicating that the model's generalization ability is weaker.After the TPE adjusts the parameters, the point cloud of the prediction results in the test set shrinks significantly, and the distribution is closer to the reference line, indicating that the accuracy and stability of the prediction have been improved. This indicates that the accuracy and stability of the prediction have been improved. Meanwhile, the performance of the training set is slightly reduced, but still maintains a high R^2, indicating that the overall performance of the model tends to be balanced.

3.3 Model Prediction

3.3.1 Constructing the prediction data

To predict the number of Olympic medals in 2028, a new input feature dataset needs to be constructed first. This dataset is based on the existing data of the 2024 Olympic Games, and the relevant features are adjusted according to the new programs of the 2028 Los Angeles Olympic Games. The specific process is described below:

1) Screening of 2024 data

The base dataset X_2024 is constructed by selecting relevant records from the original dataset for the year 2024, from which the data for Russia (i.e., rows with a NOC value of 113) are excluded because Russia was banned in 2028:

$$X_{2028} = X_{2024}[X_{2024}['year'] == 2024 \wedge ['NOC'] \neq 113 \tag{2}$$

2) Re-indexing:

Index reset on filtered data to ensure that the data is neatly organized:

$$X_{2028}.reset\_index(inplace = True, drop = True) \tag{3}$$

3) Medal projections for new sports:

Based on the new sports that have been approved by the IOC (e.g., cricket, squash, baseball, softball, stickball and flag rugby), adjust the corresponding number of medals. The specific adjustment rules are as follows:

a) Baseball and softball: the new men's and women's events will have one gold medal each; thus adding 2 gold medals per event;

b) Cricket: one gold medal is created for each of the new men's and women's events, totaling 2 gold medals;

c) Stick tennis: creation of one gold medal for each of the new men's and women's disciplines, totaling 2 gold medals;

d)Squash: the creation of a men's and women's singles event with one gold medal each, totaling 2 gold medals;

e) Flag Rugby: create one gold medal for each of the new men's and women's events, for a total of 2 gold medals. The corresponding adjustment formula is:

$$X_{2028}['Baseball'] = X_{2028}['Baseball'] + 2 \tag{4}$$

$$X_{2028}['Softball'] = X_{2028}['Softball'] + 2 \tag{5}$$

$$X_{2028}['Cricket' = X_{2028}['Cricket'] + 2 \tag{6}$$

$$X_{2028}['Sixes'] = X_{2028}['Sixes'] + 2 \tag{7}$$

$$X_{2028}['Squash'] = X_{2028}['Squsah'] + 2 \tag{8}$$

$$X_{2028}['Flagfootball'] = X_{2028}['Flagfootball'] + 2 \tag{9}$$

4) Update the year:

Since 2028 is a future year for the Olympic Games, the year information in the data needs to be updated to 2028:

$$X_{2028}['YEAR'] = 2028 \tag{10}$$

5) Host country identification:

For the United States (NOC of 147) as the host country for the 2028 Olympic Games, it needs to be identified as 1 other countries remain at 0.

$$X_{2028}['Host'_{Country} = 0 \tag{11}$$

$$\mathrm{X}_{2028}.\mathrm{l}_{\mathrm{OC}}[\mathrm{X}_{2028}['\mathrm{NOC}'] == 147, '\mathrm{Host\_Country}'] = 1 \tag{12}$$

3.3.2 Performance Analysis

1) Medal Table and Confidence Intervals for the 2028 Los Angeles Summer Olympics

Based on the established medal prediction model, the number of medals for each country at the 2028 Los Angeles Summer Olympics was predicted, and an uncertainty analysis of these medal counts was conducted, resulting in the corresponding prediction intervals. Below are the predicted results for some countries, including the number of gold, silver, and bronze medals, as well as the corresponding confidence intervals, as shown in Table 1:

Table 1 Medal Count Projections and Confidence Intervals

NOC	gold	Silver	Bronze	gold-CI-lower		gold-CI-upper	Silver-CI-Iover	Silver-CI-upper	Bronze-CI-Iower
United States	39	45	43	36	40	36	45	37	43
China	40	27	24	37	40	24	27	21	24
Japan	20	13	12	14	20	11	14	9	12
Austraila	18	19	17	12	17	8	19	12	17
France	15	25	20	12	18	18	24	16	20
…	…	…	…	…	…	…	…	…	…
Samoa	0	0	0	0	1	0	1	0	1

Mixed team	1	1	2	0	1	1	2	1	2
Crylon	0	0	1	0	0	0	1	1	1
FR Yugoslavia	0	1	1	0	1	0	1	1	1
ROC	4	2	8	0	7	0	5	0	12

The following graph shows the predicted number of Gold, Silver and Bronze medals and the corresponding confidence intervals (CIs), where the predicted values are labeled by curves and scatters, and the confidence intervals are indicated by shaded areas. The horizontal coordinate indicates the index of the data point and the vertical coordinate indicates the number of medals predicted. The predicted values for gold, silver, and bronze medals are shown as dark purple, light purple, and rose curves, respectively, with each curve accompanied by its corresponding confidence interval (gold CI, silver CI, and bronze CI, shaded, respectively).

As can be seen from the figure, the predicted values tend to stabilize as the number of data points increases, whereas in some positions (e.g., near the first few data points), there are large fluctuations, indicating that the model has a high prediction uncertainty at these positions. This is further verified by the width of the confidence intervals, where wider regions represent greater uncertainty in the predictions, while narrower regions indicate more accurate predictions, as shown in Figure 2 below.

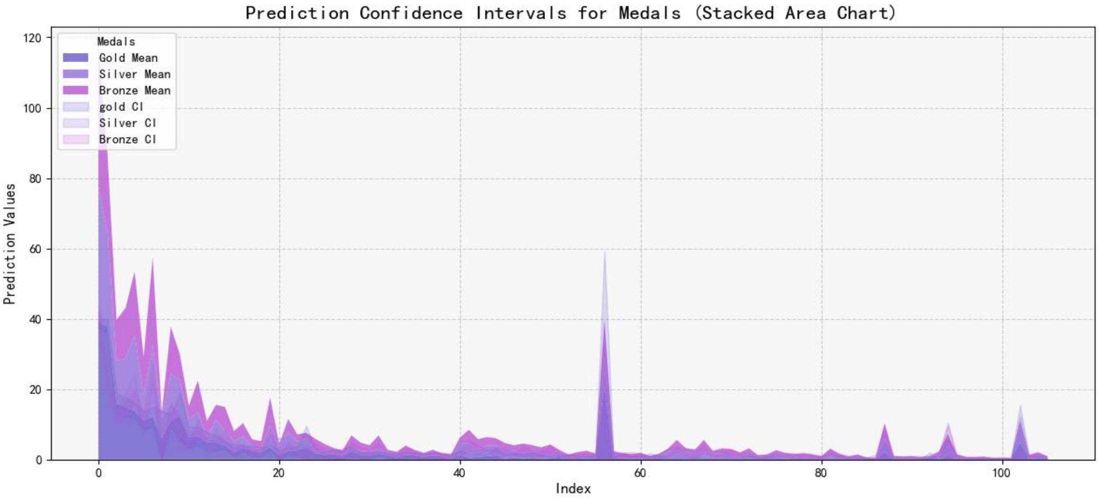


Figure 2 Prediction Confidence Intervals For Medal

By comparing the projected medal totals for 2024 and 2028, the following countries are expected to significantly improve their medal performance, as shown in Table 2 below:

Table 2 Achievement Gains in Selected Countries

NOC	2024Total	2028Total	Improvement
ROC	0	4	14
Mixed team	0	4	4
Monglia	1	4	3
Kazakhstan	7	9	2
South Africa	6	8	2
Serbia	5	7	2
Israel	7	9	2
FR Yugoslavia	0	2	2
Malaysia	2	4	2

These countries are projected to see a significant increase in performance in 2028, especially ROC and Mixed team, which have both seen significant increases in their medal totals, reflecting the strong momentum in these countries and regions.

The following countries are projected to see a decrease in performance in 2028 compared to the countries that have improved Table 3 below:

Table 3 Declining Performance in Selected Countries

NOC	2024Total	2028Total	Improvement
South Korea	32	23	-9
France	64	60	-4
Turkey	8	5	-3
Great Britain	65	63	-2
North Korea	6	4	-2
Greece	8	6	-2
India	6	4	-2

Iran	12	10	-2
Denmark	9	8	-1
Belgium	10	9	-1

The total number of medals for these countries is projected to decline in 2028, with South Korea and Frances in particular experiencing a reduction of 9 and 4 medals respectively, reflecting a possible limitation of their potential in future Olympic events.

2) Predicting countries that will win medals for the first time

For countries that have not yet won a medal, the model also makes a prediction and assesses which countries are likely to win their first ever

medals. By filtering out countries that did not win a medal in 2024 but are expected to win a medal in 2028, the following list of countries was obtained as shown in Table 4 below:

Table 4 National Medal Projections

NOC	Mean Probability of Winning
Zambia	0.602431
Independent Olympic Athletes	0.65913
Virgin Islands	0.594907
British West Indies	0.558821
Independent Olympic Participants	0.648978
Mixed team	0.752325
Ceylon	0.564092
FR Yugoslavia	0.602568
ROC	0.964915

It can be seen that ROC and Mixed team have the highest probability of winning a medal in 2028, 0.964915 and 0.752325 respectively, indicating that they are more likely to win a medal in 2028.

However, if the question asks to extend the caliber of the analysis to historical data, these countries have won medals in their history, so their probability of "winning a medal for the first time" is zero.

By analyzing the medal predictions for the 2028 Summer Olympics in Los Angeles and combining them with the predicted performance intervals for each country, some valuable predictive information can be obtained for future events. This information can not only help countries to make corresponding preparation strategies, but also provide data support for the IOC and event organizers to help them better plan and prepare for the upcoming Olympic Games. At the same time, through the prediction of first-time medal-winning countries, it is possible to better understand which countries and regions have not yet fully realized their potential for sports development and have more room for improvement.

4 CONCLUSION

In this paper, the Olympic medal distribution was successfully predicted by constructing a model. The Olympic medal prediction model provides a scientific basis for medal prediction through data preprocessing, BP neural network missing value filling, TPE hyperparameter optimization and prediction uncertainty analysis. However, the limitation of data and the complexity of the model still need further improvement. Future research can consider introducing more data sources, such as athletes' personal training data and international competition results, to improve the prediction accuracy of the model. In addition, more efficient model structures and optimization algorithms can be explored to reduce training time and resource consumption.

Overall, the model in this paper provides new methods and ideas for understanding and predicting Olympic medal distributions as well as assessing coaching effects. Through continuous improvement and optimization of the model, it can provide more scientific and accurate decision support for countries' sports development strategies and coaching resource allocation.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Shi Huimin, Zhang Dongying, Zhang Yonghui. Can Olympic medals be predicted? Journal of Shanghai Sport University, 2024, 48(04): 26-36.

[2] Chen Zhanshou, Liang Yan, Wei Qiuyue. Examination of structural variation points in linear regression models with LMSV errors. System Science and Mathematics, 2025: 1-18.

[3] Xiong Zhongmin, Guo Huaiyu, Wu Yuexin. A review of research on missing data processing methods. Computer Engineering and Applications, 2021, 57(14): 27-38.

[4] Denicolò V, Polo M. Duplicative research, mergers and innovation. Economics Letters, 2018, 166: 56-59.

[5] L Zhang, B Ding, JY Deng, et al. Study on urban subsurface change and runoff coefficient response based on BP neural network. Journal of Changjiang Academy of Sciences, 2025: 1-7.

[6] LUO Min, YANG Jinfeng, YU Hui,et al. A short-term load forecasting method based on TPE optimization and integrated learning. Journal of Shanghai Jiao Tong University, 2023(5).

[7] Li W J, Wu LL, Wen SH, et al. Optimization of LSTM-Seq2seq model for runoff simulation based on attention mechanism. Glacial Permafrost, 2024, 46(3): 980-992.

[8] You Lan, Han Xuewei, He Zhengwei, et al. An Improved Seq2Seq-Based Model for Short-Term AIS Trajectory Sequence Prediction. Computer Science, 2020, 47(09): 169-174.