World Journal of Information Technology
ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3038

OLYMPIC MEDAL QUANTITY FORECASTING: A RANDOM FOREST ALGORITHM-BASED MODEL CONSTRUCTION

JunBo Zhu*, LinFeng Li
School of Mathematical and Physical Sciences, Chongqing University of Science and Technology, Chongqing 401331, China.
Corresponding Author: JunBo Zhu, Email:18983362087@163.com

Abstract: Against the backdrop of the unstoppable wave of globalization in sports, the competition for Olympic medals has shown an increasingly fierce trend. Countries have invested a lot of resources to improve their performance in the Olympic Games in order to be in a favorable position in the medal competition. In this study, a random forest model is developed to predict the number of gold medals and the total number of medals of each country in the 2028 Olympic Games. Firstly, the data were obtained from the official website of the Olympic Games and data preprocessing was carried out. After completing data cleaning and organizing, a series of key influence indicators such as whether it is the host country, the number of athletes, the total score and so on are introduced, and then a random forest model is built to predict the total number of medals and gold medals of each country. Finally, based on the prediction results, it was determined that in the 2028 Olympic Games, countries such as Cuba, Germany and Slovakia have the potential to achieve breakthroughs, while countries such as Belgium, Ecuador and Israel may experience a decline in the acquisition of medals. This study breaks through the limitations of linear assumptions in traditional econometric models, utilizes the nonlinear fitting ability of the Random Forest algorithm to capture complex variable interactions, and quantifies the dynamic impact of the 'host effect' on the distribution of medals, and reveals the role weights of the core factors such as historical performance and participation size through characteristic contribution analysis. Meanwhile, the prediction results can provide scientific basis for the National Olympic Committees to optimize resource allocation and formulate strategies, sports economic research and event public opinion prediction.
Keywords: Random forest model; Olympic medal prediction; Data preprocessing; Prediction accuracy

1 INTRODUCTION

With the advancement of sports globalization, the competition in Olympic events has become increasingly intense. Olympic medals, as symbols of a country's sports strength, have received extensive attention. Against this backdrop, predicting the distribution of medals in Olympic events has become a focal point in the sports community.

Currently, scholars at home and abroad have conducted research on the issue of Olympic medals from multiple perspectives. Scelles N, Andreff W, Bonnal L, et al. [1] developed Tobit and Hurdle models to predict the number of medals and verified the effectiveness of the econometric models. Bredtmann J, Crede C J, Otten S. [2] constructed a combined model of regression analysis and time - series to analyze the influencing mechanisms of Olympic medals. Andreff [3] established a Tobit estimation model to analyze the influencing factors of Winter Olympics medals and predict the performance of Russia and China. Vagenas G, Vlachokyriakou E. [4] used a log - linear regression model to study the relationship between economic indicators and Olympic medals, providing a basis for medal prediction. Cheng H R, Lü J, Yuan T G. [5] employed mathematical statistics and other methods to analyze China's track and field performance in the Olympics. Liu C Y, Wu M Q, Zhang A A, et al. [6] used a spatial analysis model to reveal the spatial - temporal distribution patterns and influencing factors of China's medals. Ding Weizhe[7] used data mining and statistical analysis models to analyze the influencing factors of the ranking on the Olympic medal table. Balmer N J, Nevill A M, Williams A M. [8] adopted a generalized linear interactive modeling approach and found that the home - field advantage was significant in event groups with subjective judgment or decision - making.

Previous studies mostly relied on traditional models such as the Tobit model and regression analysis. Constrained by linear assumptions, these models struggle to depict complex non - linear interactions, such as the combined effects of the host identity and the scale of athletes. Meanwhile, when faced with multi - dimensional data, traditional models are difficult to handle due to the problem of multicollinearity, resulting in insufficient fitting accuracy. In contrast, the random forest model, with its strong noise - resistance and high data - processing efficiency brought by decision - tree integration, demonstrates flexibility and prediction accuracy in predicting the number of medals in sports events.

At present, scholars both at home and abroad have carried out prediction - related research using random forest. Bao Y, Meng X, Ustin S, et al. [9] constructed a random forest model based on Vis - SWIR spectroscopy and CARS to predict soil organic matter content, providing an effective method for soil organic matter estimation. Hafezi M H, Liu L, Millward H. [10] established a random forest model combining CART and curvature search to analyze individual daily activity sequences, and proposed that the model had the best accuracy in simulating activity agendas and sequences. Zhang Y D, Senjyu T, Chakchai S, et al. [11] developed an RF - DBSCAN model to predict road travel time, with relatively high prediction accuracy. Shi H M, Zhang D Y, Zhang Y H. [12] used a random forest combined with an interpretable model to evaluate the predictability of medals in Olympic events and found that socio - economic factors had a significant impact. Sun J, Zou X K, Zhu X B, et al. [13] established a random forest model to solve the problem of

© By the Author(s) 2025, under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0).

medical registration, with an efficiency five times higher than that of traditional models. Yang Q F, Li T, Jia Z Q. [14] constructed a random forest model optimized by a genetic algorithm to predict retail consumption behavior, showing better accuracy. Gan M, Liu P F, Yue D B, et al. [15] constructed a random forest mineral - prospecting model based on geoelectrochemical data to explore lithium deposits, and the AUC value exceeded 80% after training.

Based on previous research achievements on Olympic medals, this study uses a random forest model to predict the number of gold medals and total medals of various countries in the 2028 Olympic Games. The research first collects historical competition data from the official Olympic website and preprocesses it, selects a series of key influencing indicators such as whether a country is the host, and then constructs a random forest regression model to complete the prediction, yielding the predicted values of the medal standings for participating countries in 2028. By comparing the predicted values with historical results, countries with potential for performance breakthroughs and those likely to decline are identified, and targeted recommendations are proposed accordingly.

2 METHODS

2.1 Data Preprocessing

This paper obtain the data from the official website of the Olympic Games. For data preprocessing, only records from 1950 onwards were retained. Prior to 1950, Olympic participation was limited, with fewer countries competing, which made the data insufficient to reflect nations' true athletic capabilities. Frequent wars between the early 20th century and the 1940s (e. g. , World War I and II) further disrupted regular participation, resulting in sparse and unreliable data that could skew research outcomes. Post-1950, improved international stability and increased global cooperation expanded participation, yielding more comprehensive and accurate data that better capture countries' sustained Olympic performance.

After filtering, missing data were addressed through a systematic approach. Countries with over 50% missing values in critical metrics (e. g. , medal counts, event participation) were excluded to avoid biasing predictions. For gaps between consecutive Olympic editions (e. g. , available data for Editions N and N+2 but missing for N+1), values were imputed using the mean of the adjacent editions' medal counts. For nations absent from recent Olympics due to political events or natural disasters but with historically relevant performance data, their most recent valid records were carried forward to maintain predictive continuity. These procedures enhanced data integrity, ensuring the dataset was robust for modeling and predictions aligned more closely with real-world outcomes.

2.2 Data Analysis Methods

2.2.1 Key metrics

This study analyzed publicly available data from past Olympic Games and extracted key metrics including host country status, number of athletes, total medals, proportion of medals per country, proportion of gold medals per country, number of events participated in, gold medal count, and score—with the latter serving as an indicator of a country's sporting strength based on Balmer et al. [8]scoring methodology: each gold medal is valued at 3 points, silver at 2 points, and bronze at 1 point, with the total score calculated as the sum of points from all medals won.

$$score = gold * 3 + silver * 2 + bronze * 1 \tag{1}$$

2.2.2 Standardization

When conducting an in-depth analysis of publicly available data from past Olympic Games, data standardization is a critical step. The indicators used in predictive models vary significantly in dimensions and value ranges; without standardization, subsequent data analysis and model building would be severely disrupted. Z-Score normalization, based on the mean and standard deviation of the data, transforms values into a standard normal distribution with a mean of 0 and a standard deviation of 1.

$$x_{new} = \frac{x - \mu}{\sigma} \tag{2}$$

Here, μ represents the mean of the dataset, and σ denotes the standard deviation.

2.2.3 Contribution of characteristic variables

The contribution of each feature variable to the model's prediction result is interpreted as "the impact of the variable (x) on the final prediction outcome (y) during the prediction process. " Based on this core definition, this study uses the additive feature attribution method and related formulas provided by Shi Huimin et al. [12] to analyze the association between feature variables and prediction results. The "total prediction contribution" of the predictive model can be expressed as:

$$g(x) = \varphi_0 + \sum_{i=1}^{M} \varphi_i(x)\mathbb{1}(x_i) \tag{3}$$

where $X = (x_1, \cdots, x_M)$ is an M-dimensional explanatory or feature variable. $\mathbb{1}(x_i) \in \{0, 1\}$ is a binary indicator variable, where $\mathbb{1}(x_i)$ means the i-th feature variable is used for prediction, and 0 means it is not. $g(x)$ represents the final prediction result, ϕ_0 represents the average value of predictions, and $\phi_i(x)$ represents the marginal contribution of the feature variable to the prediction result. The key issues measured in this study are addressed through

the additive feature - attribution method. Specifically for the research questions of this study, $g(x)$ represents "the logarithmic value of the number of awards/gold medals won by a certain team in a certain event in a certain year", ϕ_0 represents the average number of awards/gold medals won by all teams in that event, and x_i represents the value of the i-th feature variable. Through calculation, the impact of changes on the number of gold medals and awards can be obtained, so as to identify the features that contribute more significantly to predicting medal changes.

2.3 Random Forest Prediction Model

The Random Forest Prediction Model falls within the realm of Ensemble Learning. Ensemble Learning aims to construct a powerful learner by combining multiple weak learners, thereby overcoming the limitations of traditional single - prediction models. Traditional single - prediction models, such as simple linear regression models or single decision - tree models, often suffer from overfitting or underfitting when dealing with high - dimensional and complex - distributed data. This leads to unsatisfactory prediction accuracy and poor generalization ability of the models, meaning they perform decently on the training set but experience a significant drop in performance on new test data. The Random Forest Prediction Model, however, effectively mitigates these issues through its unique construction and integration strategies. Here is its specific operational process.

2.3.1 Bootstrap sampling to build training subsets

Bootstrap sampling method is applied to randomly draw a training subset of the same capacity of N from the original dataset of size N with putback. This type of putative sampling is characterized by the fact that the same sample may or may not be drawn multiple times in a single sampling. In this way, the training subset obtained from each extraction differs somewhat from the original dataset, preserving the main features of the original data while introducing randomness. This has the advantage of providing different but related training data for each decision tree constructed subsequently, allowing each tree to learn the features of the data from different perspectives and enhancing the diversity of the model. For example, in an original dataset containing 1000 samples, the training subset obtained by Bootstrap sampling may have some of the samples duplicated and others not sampled, although the sample size is also 1000.

2.3.2 Randomly selected subset of split features

When a data sample has M features, m (where m << M) features are randomly selected as the subset of splitting features for constructing a decision tree. Limiting m to be much smaller than M is to prevent one or a few features from dominating the decision - tree construction process, thus enabling the model to learn more complex relationships among features. For instance, when dealing with a dataset that has 50 features (M = 50), perhaps only 5 features (m = 5) are randomly chosen to build the splitting nodes of the decision tree. This random feature - selection approach increases the model's randomness and robustness, allowing different decision trees to grow based on different feature combinations and further enriching the model's diversity.

2.3.3 Decision tree growth

Each decision tree is allowed to grow fully during construction without pruning. During the growth of a decision tree, based on the selected feature subset, nodes are continuously split according to certain criteria (such as information gain, Gini coefficient, etc.) until stop conditions are met (for example, the number of samples in a node is less than a certain threshold, or all samples belong to the same category). Omitting the pruning operation is to enable each tree to learn the information in the training data to the greatest extent and uncover potential complex patterns in the data. However, this may also lead to overfitting of a single tree on the training set. Nevertheless, within the overall framework of the random forest, integration of multiple trees can effectively alleviate this problem.

2.3.4 Constructing a random forest

The above steps (1)–(3) are repeated to construct a large number of decision trees, which collectively form the random forest. The number of decision trees can be adjusted according to practical needs. Generally, a larger number of trees may improve the model’s stability and prediction accuracy, but it also increases computational costs and training time. For example, in some practical applications, dozens or even hundreds of decision trees may be constructed to form the random forest.

2.3.5 Generation of forecast results

For regression tasks, each decision tree in the random forest generates a numerical prediction for the input sample. The final prediction of the random forest model is obtained by averaging the predictions of all decision trees. This ensemble averaging reduces variance and enhances the robustness of the prediction, leveraging the diversity of individual tree outputs to produce a more stable and accurate result.

The algorithm workflow is shown in Figure 1:

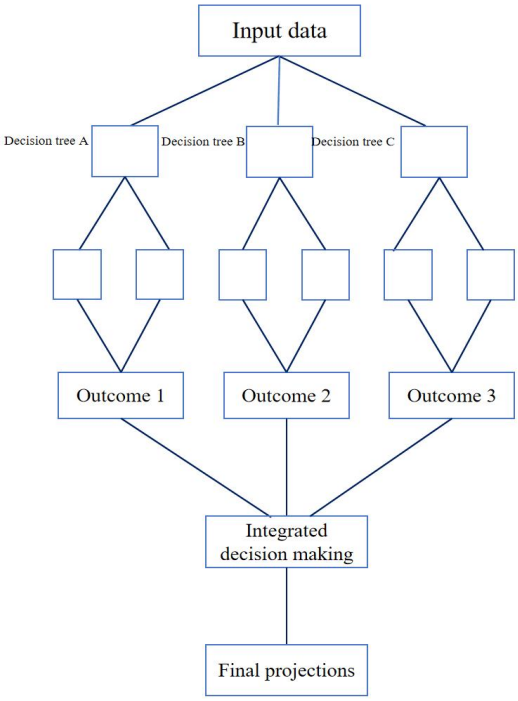

Figure 1 Flow Chart

The random forest prediction model aggregates multiple weak predictors (i. e. , individual decision trees) and integrates their forecasting results through a mean-based strategy. This ensemble approach significantly reduces variance in prediction tasks, endowing the model with excellent predictive accuracy and robust generalization capabilities, enabling it to handle various complex real-world prediction scenarios with ease. In this study, the random forest model is utilized to predict Olympic medal counts, facilitating event analysis and informed decision-making processes.

3 RESULTS AND DISCUSSION

3.1 Data Integration

After completing the preprocessing of the result data, to present the data more clearly and comprehensively, this study merged the data and counted the specific number of participating countries (regions), the detailed classification counts of sports events (both major and minor categories), and the number of athletes. The final results are shown in Table 1 below:

Table 1 Participating Countries/Regions, Sports/Disciplines, Athlete/Count

Type	Number
The number of unified states	206
Number of sport	50
Number of event	289
Total Athletes	11097

As shown in the table, the four-dimensional data of "large number of participating countries/regions, wide coverage of major sports, detailed classification of minor disciplines, and large athlete scale" intuitively demonstrates the characteristics of the Summer Olympics as "grand in scale and diverse in events", reflecting both the depth of global sports exchange and the integrity and complexity of the competitive sports system.

3.2 Analysis of the Contribution of Each Variable to the Medal Prediction Model

The contribution rates of each variable to the gold medal prediction model and the medal prediction model, calculated using the feature variable contribution rate formula, are shown in Table 2 and Table 3 . For the gold medal count prediction model in Table 2 , the most critical factors are historical gold medal count and the proportion of a country's gold medals in the total gold medals, with total medal count also playing a significant role. Host country status and the number of events participated in have less impact on gold medal count prediction, while athlete count and score have relatively minimal effects. For the Olympic medal count prediction model in Table 3 , the most critical factors are total medal count and the proportion of a country's medals in the total medals. Host country status and the number of events participated in also make important contributions. In contrast, athlete count, gold medal count, and score have relatively minor roles in predicting medal counts.

Table 2 Contribution of Each Variable to the Gold Medal Forecasting Model

Variate	Contribution rate to the gold medal model
Whether it is the host country	18. 6%
Number of athletes	6. 5%
medal tally	11. 2%
gold MEDALS in each country accounted	17. 4%
Number of events participated	8. 4%
Number of gold MEDALS	25. 8%
Score	12. 1%

Table 3 Contribution of Each Variable to the Medal Prediction Model

Variate	Contribution rate to prediction medal model
Whether it is the host country	17. 4%
Number of athletes	5. 6%
medal tally	15. 6%
gold MEDALS in each country accounted	16. 1%
Number of events participated	7. 3%
Number of gold MEDALS	19. 4%
Score	18. 6%

3.3 2028 Medal Count Forecast Analysis

When establishing the random forest model, this study adjusted the hyperparameters according to the data type and volume. The final configurations were set as follows: the number of decision trees was set to 200 to fully capture the relationships between data features and enhance the model's generalization ability; the maximum tree depth was limited to 15 to prevent overfitting by restricting the growth depth during training; and the minimum sample split number was set to 5 to ensure that decision tree nodes contain at least the specified number of samples when splitting, making the model more robust.

Meanwhile, historical Olympic medal data was divided at a 7:3 ratio, with 70% of the data used as the training set for model parameter learning and optimization, and 30% as the test set. This strategy ensures the model can sufficiently learn the characteristics of historical data while providing a reliable basis for testing its performance on unknown data. The study employed mean squared error (MSE), mean absolute error (MAE), and coefficient of determination (R^2) as evaluation metrics to assess model performance.

By constructing a random forest model to predict the number of medals for the 2028 Olympics, the following prediction results were obtained (only listing the top 10 countries in total medals and gold medals). Meanwhile, the model's mean squared error (MSE) is 8. 2 and mean absolute error (MAE) is 2. 3. These low values indicate small average discrepancies between predicted and actual values, suggesting high prediction accuracy. The coefficient of determination R^2 reaches 0. 89, close to 1, demonstrating a high degree of fit to historical data and the model's ability to explain 89% of the variation in medal counts.

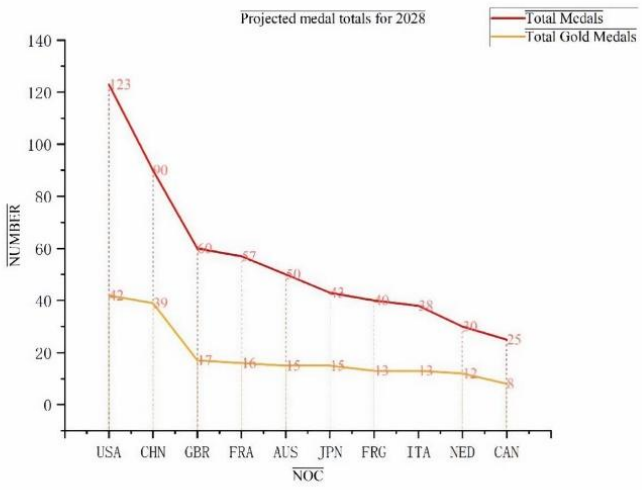

Figure 2 Comparison of the Projected Number of Medals for the 2028 Olympic Games

As shown in Figure 2 , the predicted total medal count for the United States is 123, far ahead of other countries, fully demonstrating its overall dominance in the sports arena. The U. S. boasts a mature sports talent development system, substantial financial investment, and a broad mass participation base, which enable it to maintain strong competitiveness across numerous sports disciplines and secure medals in both major and minor events.
The prediction results also reveal disparities in event-specific strengths among nations. Sporting powerhouses like the U. S. and China exhibit robust competitiveness across multiple sports. For example, Japan excels in combat sports such as judo and wrestling, while the Netherlands has achieved remarkable results in speed skating and cycling. These event-specific advantages help them gain outstanding performances in niche areas, though they also face the risk of over-reliance on single disciplines.

3.4 Medal Count Trend Analysis

The comparison between the predicted medal and gold medal counts for 2028 and the actual values for 2024 is presented in Table 4. Based on the table and an analysis of factors such as athlete reserves, the number of events participated in, and economic strength, Cuba, Germany, and Slovakia are expected to make progress in the upcoming Summer Olympics. In contrast, Belgium, Ecuador, and Israel may regress due to their respective circumstances, athlete reserves, economic strength, and other factors.

Table 4 Progress or Regression of Olympic-Related Countries

Noc	Situation
Cuba	Progressive country
Germany	Progressive country
Slovakia	Progressive country
Belgium	Backward country
Ecuador	Backward country
Israel	Backward country

Meanwhile, this study also focuses on the medal count trends of the top-ranked countries. Therefore, the predicted values for the top 5 countries in the 2028 Olympics are compared with their actual data from the 2024 Olympics, as shown in Table 5 below:

Table 5 Trends in the Number of Medals Won by Countries

Noc	The number of medals in 2024	The number of medals in 2028
USA	126	123
CHN	91	90
GBR	65	60
FRA	64	57
AUS	53	50

As indicated in the table, the medal counts of the United States, China, the United Kingdom, France, and Australia are projected to show a steady downward trend across the two Olympic Games. This finding suggests that although these countries are likely to remain among the top medal earners during the upcoming four-year Olympic cycle, they may face challenges in the development of competitive sports. For example, the decline in form or retirement of veteran athletes, coupled with delays in identifying and nurturing promising reserve talent, could directly lead to a reduction in medal tallies. In some disciplines, the lack of generational succession may make it difficult to sustain the exceptional performance achieved in 2024 into 2028. Therefore, over the next four years, these nations need to enhance their sports talent development frameworks to maintain their competitive advantages.

4 CONCLUSIONS

With the vigorous development of global sports, the competitive landscape on the Olympic stage has become increasingly intense, with a continuous increase in the number of participating countries and regions and the constant innovation of sports events. While promoting the improvement of competitive sports standards and cultural exchange, the uncertainty of medal distribution has significantly increased, posing challenges to event prediction and strategic planning, and adding variables to the development of sports economy and industry. Therefore, constructing a scientific Olympic medal count prediction model helps to grasp the sports competition situation and optimize resource allocation.
This study on medal predictions for the 2028 Olympics first highlights the outstanding performance of random forest models in complex data prediction scenarios. It then screens complete post-1950 data through data preprocessing, introduces key indicators such as host country status and athlete numbers, and performs standardization. Through feature contribution analysis, core influencing factors such as historical performance are identified. Finally, a random forest model is developed to predict gold medal counts and total medal counts. Based on the predictions, countries with growth potential such as Cuba and those at risk of performance decline such as Belgium are identified, providing a

scientific basis and practical reference for Olympic event analysis and decision-making by national Olympic committees.

COMPETING INTERESTS

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

REFERENCES

[1] Scelles N, Andreff W, Bonnal L, et al. Forecasting national medal totals at the Summer Olympic Games reconsidered. Social science quarterly, 2020, 101(2): 697-711.
[2] Bredtmann J, Crede C J, Otten S. Olympic medals: Does the past predict the future?. Significance, 2016, 13(3): 22-25.
[3] Andreff W. Economic development as major determinant of Olympic medal wins: predicting performances of Russian and Chinese teams at Sochi Games. International Journal of Economic Policy in Emerging Economies, 2013, 6(4): 314-340.
[4] Vagenas G, Vlachokyriakou E. Olympic medals and demo-economic factors: Novel predictors, the ex-host effect, the exact role of team size, and the "population-GDP" model revisited. Sport Management Review, 2012, 15(2): 211-217.
[5] Cheng H R, Lü J, Yuan T G. Prediction of China's track and field results in the Tokyo Olympic Games from the world top 20 national rankings of track and field events in 2018. Bulletin of Sports Science & Technology, 2020, 28(04): 4-8.
[6] Liu C Y, Wu M Q, Zhang A A, et al. Study on the temporal and spatial differentiation of Chinese Olympic medals from 1984 to 2016. Journal of Physical Education, 2019, 26(01): 75-82.
[7] Ding W Z. Data mining model of Olympic medals based on comprehensive national strength. Information Recording Materials, 2018, 19(03): 231-233.
[8] Balmer N J, Nevill A M, Williams A M. Modelling home advantage in the Summer Olympic Games. Journal of sports sciences, 2003, 21(6): 469-478.
[9] Bao Y, Meng X, Ustin S, et al. Vis-SWIR spectral prediction model for soil organic matter with different grouping strategies. Catena, 2020, 195: 104703.
[10] Hafezi M H, Liu L, Millward H. Learning daily activity sequences of population groups using random forest theory. Transportation research record, 2018, 2672(47): 194-207.
[11] Zhang Y D, Senjyu T, Chakchai S, et al. Smart Trends in Computing and Communications. Springer Singapore, 2022.
[12] Shi H M, Zhang D Y, Zhang Y H. Can Olympic medals be predicted? - From the perspective of interpretable machine learning. Journal of Shanghai University of Sport, 2024, 48(04): 26-36.
[13] Sun J, Zou X K, Zhu X B, et al. Research on random forest algorithm in the field of online scalper prediction. Computer Simulation, 2025: 1-6.
[14] Yang Q F, Li T, Jia Z Q. Consumption behavior prediction algorithm based on parameter-optimized random forest model. Computer & Digital Engineering, 2024, 52(07): 1959-1965.
[15] Gan M, Liu P F, Yue D B, et al. Prospecting prediction by geoelectrochemical technology in and around the Murong lithium mining area, western Sichuan based on random forest algorithm. Geology and Exploration, 2025, 61(02): 359-370.