Research On the Performance Evaluation of Tennis Players Based on Stacking

Wenjie Yang *

College of Automation & College of Artificial Intelligence, Nanjing University of Posts and Telecommunications, Nanjing, China, 210023

* Corresponding Author Email: b21041616@njupt.edu.cn

Abstract. Evaluating the performance of tennis players in the competition not only improves the scientific and targeted training, but also realizes the accurate prediction of the match results, and promotes the fairness and transparency of competitive sports. At the same time, it enhances the spectator experience of the game, promotes the innovation and development of sports science and technology, and is of great significance to the individual growth of athletes and the overall progress of the sports industry. In order to evaluate the performance of athletes in the competition in real time, this paper establishes the PSPPM model, which combines the dynamic change and trend of time series, visualizes the competition process, and predicts the athlete's score in real time. The model uses a stacking approach, including logistic regression and XGBoost models. Stacking, as a multi-model fusion method, combines the interpretability of logistic regression, the visualization results, the feature importance assessment capability of XGBoost, and the ability to capture dynamic changes and trends in time series. This enhances the overall performance of the model, demonstrating better generalization capabilities when dealing with complex problems.

Keywords: Tennis, Stacking, Logistic Regression, XGBoost.

1. Introduction

In the 2023 Wimbledon Gentlemen's final, a dramatic match took place between rising star Carlos Alcaraz and established champion Novak Djokovic, featuring swings in momentum and game control. The concept of 'momentum' in sports is hard to measure and there's a debate about its significance.Therefore it is necessary to analyze Wimbledon match data, establish a highly generalizable model that can identify and predict momentum in matches, providing deeper insights, and evaluating the actual impact of momentum in tennis games.

Numerous studies have explored the application of machine learning algorithms in tennis matches. For instance, Jack C. Yue et al. [1] proposed using the Glicko rating model to assess players' strengths and employed EDA tools to identify key variables. Firas Bayram et al. [2] introduced advanced machine learning methods such as multi-output regression and privileged information learning to infer surface-specific and time-varying scores of professional tennis players. Sophie Chiang et al. [3] investigated the application of standard machine learning algorithms, including multi-output regression and random forest, for predictive purposes. However, traditional methods have limitations, such as over-reliance on past results and the risk of overfitting. To address these issues, this study employs logistic regression, XGBoost, linear regression, and random forest methods, combined with ensemble techniques, to create a more robust model capable of capturing complex patterns and enhancing generalization.

In conclusion, the accuracy of predicting the outcome of a tennis match depends on the quality and quantity of data used, the specific features chosen for the model, and the type of machine learning algorithm employed. The research focus of this paper is to explore and study these aspects in depth in order to improve the accuracy and practicability of prediction.

2. Point scoring probability prediction model

2.1. Data source

The data in this paper are from https://www.comap.com/undergraduate/contests/.

This paper filtered out missing values and outliers. This paper found missing values in the dataset, as shown in the table. Since the proportion of samples with missing values is 25%, we consider using the XGBoost model that can automatically handle missing values. And there are no missing values in the variables used in the logistic regression model. Considering that the data in this study are game record data, and after data auditing, no outliers were found.The data distribution is shown in Figure 1 below.

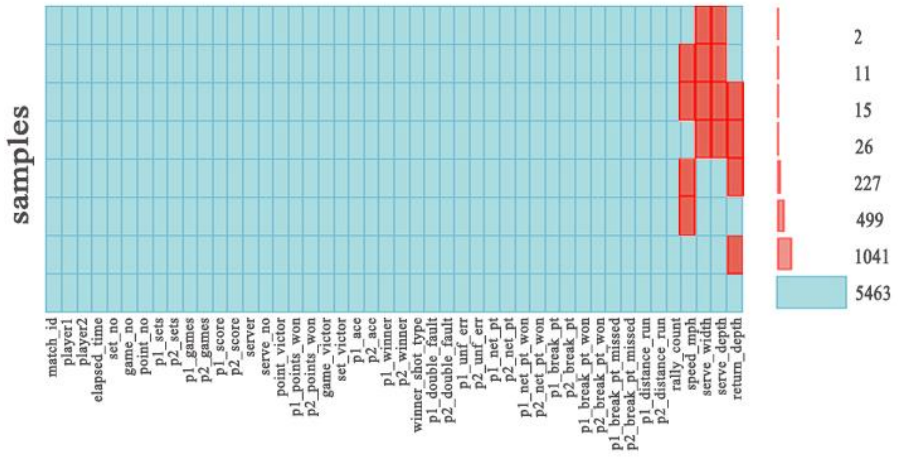


Figure 1. Data Distribution

To effectively eliminate the difference in dimensions between the data and facilitate model building, we used normalization to process some data.

2.2. The structure of model

In order to provide a visualized competition process, we applied logistic regression models separately for the performances of two players, Player1 and Player2, predicting their respective winning probabilities. It's worth noting that the winning probability models for Player1 and Player2 are constructed based on their individual features and strengths, without cross-usage. We chose logistic regression models primarily due to their excellent interpretability; the model's output results in probabilities, making it highly suitable for visual representation of outcomes. Additionally, logistic regression models can reveal the weights and impacts of different features. Moreover, logistic regression models exhibit outstanding computational efficiency and training speed, especially when dealing with small-scale datasets, demonstrating significant computational advantages in handling binary variables. Finally, since logistic regression models estimate parameters through maximum likelihood estimation, their susceptibility to the influence of outliers is relatively small. This forms the foundational layer of our predictive model.

Building upon this foundation, we further utilized the winning probability models for Player1 and Player2, combined with the common features Player1 and Player2, to apply the XGBoost model in constructing the winner prediction model. We chose XGBoost because, as a gradient boosting tree algorithm, it iteratively trains a series of decision trees, each correcting the prediction errors of the previous tree, gradually improving the model's performance. Additionally, XGBoost can handle missing values automatically, eliminating the need for additional data preprocessing. Most importantly, it provides an evaluation of feature importance, aiding in identifying which features have the greatest impact on the model's predictions. This constitutes the second foundational layer of our predictive model.

To capture the dynamic changes and trends in the time series, this paper ultimately chose a stacking model for the final prediction of the test data. The stacking model is a multi-model fusion method that effectively integrates the advantages of different models, enhancing the overall performance of the model and demonstrating better generalization capabilities when handling complex problems.The PSPPM structure is shown in Figure 2 below.

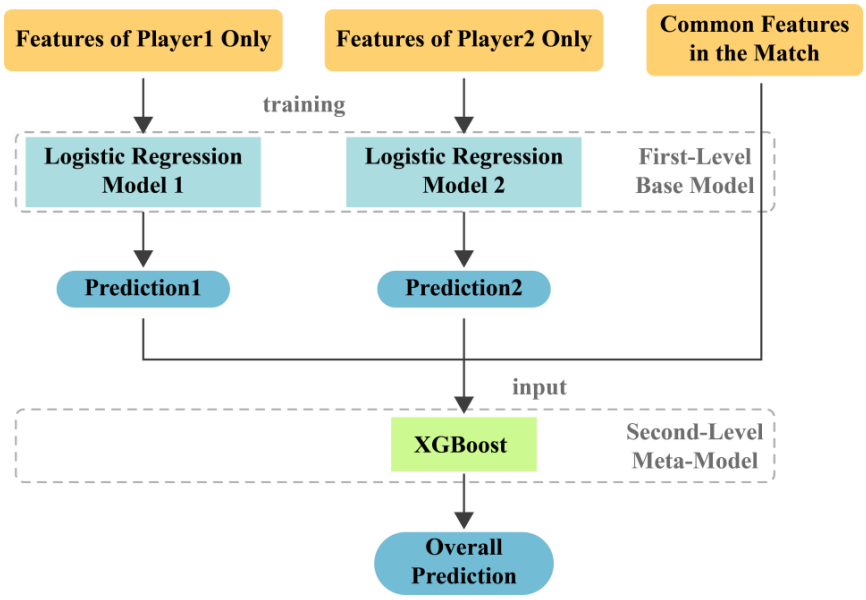


Figure 2. The Structure of PSPPM

2.3. Point Scoring Probability Prediction Model Based on Individual

2.3.1 Feature Selection

To incorporate the influence of momentum, besides measuring athlete's technical indicators, indicators reflecting historical sequence information were also selected, such as ($x_{1,p}$,$x_{3,p}$).

$$x = [x_{1,p}, x_{2,p}, x_{3,p}, \ldots, x_{17,p}] \tag{1}$$

In this feature selection, x_ {1, p} and x_ {3, p} are used to represent historical sequence information indicators, which can help measure the athlete's momentum and technical performance. These indicators assist in predicting the probability of the athlete scoring in the next match.

x_{1,p}: Whether the player scored in the last match. This feature can reflect the player's ability to score consecutively and the opponent's defensive situation. If a player scored in the last match, they may maintain confidence and motivation, increasing the likelihood of scoring in the next match.

x_{3,p}: Whether the player hit an ace in the last match. This feature is specific to serving games. If a player hit an ace in the last match, it indicates that their serving state may be good, which could influence the next serving game and increase the chance of scoring.

By considering these historical sequence information features, the model can better capture the athlete's state and momentum changes during the match, thereby improving the accuracy of predicting the probability of scoring in the next match.

2.3.2 Model Specification

The following is the theoretical framework of logistic regression:

For binary classification problems, considering that the value of the following function [4].

$$w^T x + b \tag{2}$$

is continuous, it cannot fit discrete variables. We can consider using it to fit conditional probability:

$$p(Y = 1 \mid x) \tag{3}$$

because the value of probability is also continuous.

The most ideal is the unit step function:

$$p(y=1 \mid x)=\begin{cases}0, & z<0\\ 0.5, & z=0,\ z=w^{T}x+b\\ 1, & z>0\end{cases} \tag{4}$$

but this step function is not differentiable, so the logistic function is a commonly used substitute function [5]:

$$y=\frac{1}{1+e^{-(w^{T}x+b)}} \tag{5}$$

The above formula can be transformed into another formula:

$$\ln\frac{y}{1-y}=w^{T}x+b \tag{6}$$

where y is viewed as the probability of x being a positive example, and 1-y is the probability of x being a negative example. The ratio of the two is called the odds. So, in logistic regression, the dependent variable value should be odds.

If y is viewed as the class posterior probability estimate, the formula can be rewritten as follows

$$w^{T}x+b=\ln\frac{P(Y=1|x)}{1-P(Y=1|x)} \tag{7}$$

$$P(Y=1 \mid x)=\frac{1}{1+e^{-(w^{T}x+b)}} \tag{8}$$

For multi-category logistic regression, spsspro defaults to comparing one category with the remaining categories as a binary classification problem [6-7]. N categories perform N-1 times classification to obtain N-1 binary classification models, calculate the probability corresponding to each binary classification, and the highest probability class is used as the prediction result of the new sample.

2.4. Point Scoring Probability Prediction Model Based on Both Side

$$Z=[x_{18},x_{19},x_{20},\ldots,x_{31},pred_{1},pred_{2}] \tag{9}$$

Here, pred_1 is the probability of player1 winning calculated using player1's features and the Point Scoring Probability Prediction Model Based on Individual. Similarly, pred_2 is the probability of player2 winning calculated using player2's features and the same prediction model.

2.4.1 Feature Selection

During the feature selection process, we chose indicators that reflect the dynamics of the match and historical sequence information to construct a more accurate prediction model.

Indicators reflecting the match dynamics: These indicators capture the dynamic changes and development of the match, significantly influencing the prediction of match outcomes. Examples include player scores, score differentials, and match stages, which reflect the progress and trends of the match.

Indicators reflecting historical sequence information: Historical sequence information provides essential clues about a player's past performance and event situations, guiding the prediction of future match outcomes. For instance, a player's past scores, technical indicators, and historical head-to-head records offer insights into a player's characteristics and potential match situations.

In the feature selection process, we considered these factors and selected a set of indicators that comprehensively reflect the match dynamics and historical sequence information. The feature set Z includes indicators derived from x_{18} to x_{29} both categories and the probabilities of winning pred_1 and $\mathrm{pred}_2$for both players calculated through the individual-based point scoring probability prediction model.

This feature selection approach integrates considerations of both match dynamics and historical sequence information, providing rich input features for the prediction model and enhancing the accuracy and reliability of predictions. By comparing players' historical performances and individual-based point scoring probability predictions, we can comprehensively assess players' strengths and match trends, offering a more reliable basis for predicting match outcomes.

2.4.2 Model Specification

The following is the theoretical framework of XGBoost:

XGBoost, short for "Extreme Gradient Boosting", is a synthetic algorithm formed by the combination of base functions and weights to achieve a good fit to data[8-9]. Due to its strong generalization ability, high scalability, and fast computation speed, XGBoost has been welcomed in the fields of statistics, data mining, and machine learning since its proposal in 2015.

For a dataset containing n records of m dimensions, the XGBoost model can be represented as:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(z_i),\ f_k \in F\ (i = 1,2,\dots n) \tag{10}$$

Were,

$$F = \{f(z) = w_{q(z)}\}(q: R^m \to \{1,2,\dots,T\}, w \in R^T) \tag{11}$$

is a collection of CART decision tree structures, q is the tree structure mapping samples to leaf nodes, T is the number of leaf nodes, and w is the real number score of leaf nodes [10]. When building the XGBoost model, the optimal parameters must be found according to the principle of minimizing the objective function to establish the optimal model. The objective function of the XGBoost model can be divided into the error function term L and the model complexity function term Ω. The objective function can be written as:

$$Obj = L + \Omega \tag{12}$$

$$L = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{13}$$

$$\Omega = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{14}$$

Where, γT is the L1 regular term, $\frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2$ is the L2 regular term.

When optimizing the model using training data, the original model needs to be retained and a new function f is added to the model to reduce the objective function as much as possible. The specific process is as follows:

$$\begin{aligned} \hat{y}_i^{(0)} &= 0 \\ \hat{y}_i^{(1)} &= \hat{y}_i^{(0)} + f_1(x_i) \\ \hat{y}_i^{(2)} &= \hat{y}_i^{(1)} + f_2(x_i) \\ &\cdots\cdots \\ \hat{y}_i^{(t)} &= \hat{y}_i^{(t-1)} + f_t(x_i) \end{aligned} \tag{15}$$

Where, $\hat{y}_i^{(t)}$ is the prediction value of the t-th model, $f_t(x_i)$is the new function added in the t-th time. At this time, the objective function is expressed as:

$$Obj^{(t)} = \sum_{i=1}^{n} \left(y_i - \left(\hat{y}_i^{(t-1)} + f_i(z_i)\right)\right)^2 + \Omega \tag{16}$$

In the XGBoost algorithm, to quickly find the parameters that minimize the objective function, the objective function has been expanded by Taylor's second order to obtain the approximate objective function.

$$Obj^{(t)} \approx \sum_{i=1}^{n} \left[\left(y_i - \hat{y}^{(t-1)}\right)^2 + 2\left(y_i - \hat{y}_i^{(t-1)}\right) f_t(z_i) - h_i f_t^2(z_i)\right] + \Omega \tag{17}$$

After removing constant terms, it is known that the objective function is only related to the first and second derivatives of the error function. At this time, the objective function is expressed as:

$$\begin{aligned} Obj^{(t)} &\approx \sum_{i=1}^{n}\left[g_i w_{q(x_i)} + \frac{1}{2}h_i w_{q(x_i)}^2\right] + \gamma T + \frac{1}{2}\sum_{j=1}^{T} w_j^2 \\ &= \sum_{j=1}^{T}\left[\left(\sum_{i\in I_j} g_i\right) w_j + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \lambda\right) w_j^2\right] + \gamma T \end{aligned} \tag{18}$$

If the structure part q of the tree is known, the objective function can be used to find the optimal Wj and obtain the optimal objective function value. Its essence can be attributed to a quadratic function minimum value solving problem. The solution is:

$$w_j^* = \frac{-\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda} \tag{19}$$

$$Obj = -\frac{1}{2}\sum_{j=1}^{T} \frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma T \tag{20}$$

Obj can be used as a scoring function to evaluate the model, the smaller the Obj value, the better the model effect. By recursively calling the above tree building method, a large number of regression tree structures can be obtained, and the Obj search is used to find the optimal tree structure, put it into the existing model, thereby establishing the optimal XGBoost model.

3. Results

3.1. Error Analysis

Table.1. Model evaluation result

AUC	MSE	MAE
0.972	0.0437	0.0277*

Forecast accuracy and mean square error are provided according to Table 1 above, which can initially evaluate the performance of the model in predicting the outcome of the match. Here's an analysis of the difference between the model's predictions and the actual results:

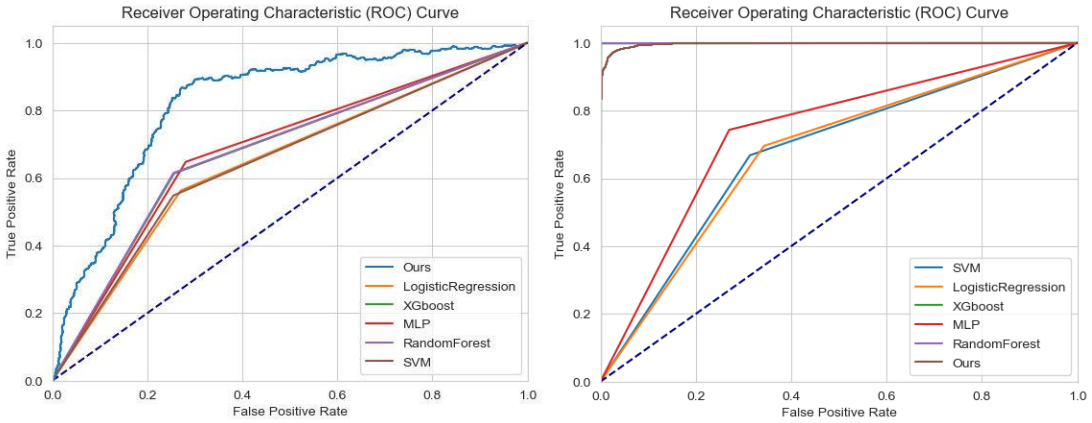
The prediction accuracy was 0.9723: This means that the model was able to accurately predict the outcome of the match 97.2 percent of the time. This is a very high accuracy rate, indicating that the model has been remarkably successful in capturing match dynamics and player performance.

The mean absolute error (MAE) is 0.0277: MAE is the mean absolute error between the predicted value and the actual value. A low MAE indicates a relatively small deviation between the model's predicted value and the actual value.

The mean square error (MSE) is 0.0437: Similar to MAE, MSE is a way to measure the difference between the predicted value and the actual value. In this case, the value of MSE is also relatively low, further indicating that the model's prediction effect is better.

In summary, based on the indicators provided, we can preliminarily believe that the model performs well in predicting the result of the competition, with high accuracy and low prediction error. This suggests that momentum may play an important role in a match, and that the model is able to effectively capture and use the influence of momentum to make accurate match outcome predictions. However, we still need more in-depth evaluation of the robustness and generalization ability of the model to ensure its validity and reliability in different competitions and scenarios.

3.2. Model Comparison Analysis



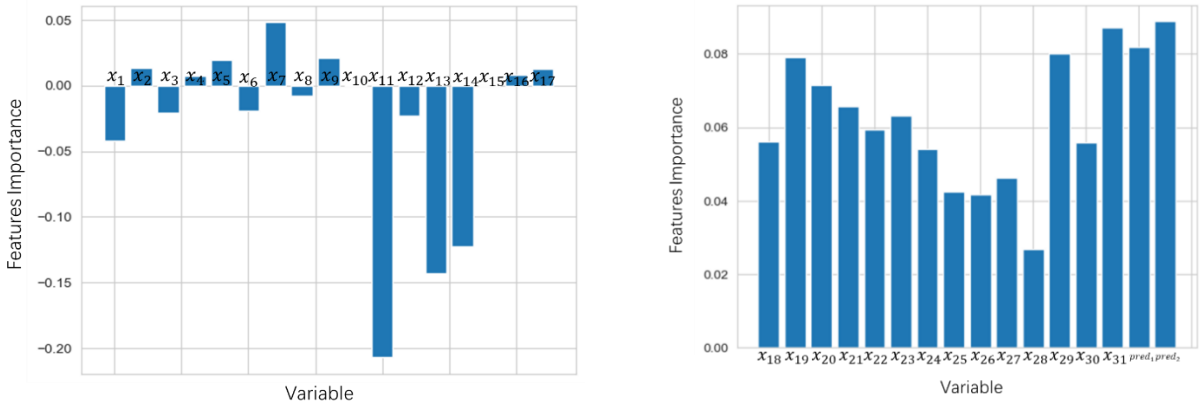
(a)ROC on Training Datasets (b)ROC on Test Datasets

Figure 3. ROC Curve

The ROC curve is a tool used to evaluate the performance of classification algorithms, by comparing the true positive rate (TPR) with the false positive rate (FPR). The Figure 3 shows the performance of five different machine learning models (Logistic Regression, XGBoost, MLP, Random Forest, and SVM). In this image, all models show good performance as their ROC curves are above the dashed line, which represents random guessing.

The area under the ROC curve (AUC) is a standard measure for assessing the efficacy of a classifier. It signifies the classifier's ability to differentiate between positive and negative instances. The closer the AUC value is to 1, the better the classifier's performance; the closer the AUC value is to 0, the worse the performance. In the given examples, our model outperforms others, with the highest AUC values. Other models have intermediate AUC values, indicating average performance.

3.3. Feature Importance Analysis



(a)PSPPM based on individual (b)PSPPM based on both sides

Figure 4. Feature Importance

The ranking of feature importance is shown in Figure 4 below, high Importance of Momentum-Related Indicators: In the analysis of feature importance, indicators related to momentum play a crucial role in the model. This indicates that momentum factors are a key consideration when predicting match outcomes.

Reliability of Predicted Results: Due to the high importance of momentum-related indicators, the model can more accurately capture and leverage the impact of momentum, thereby improving the reliability and accuracy of predicting match outcomes. This implies that the model can better take into account the athletes' conditions and the dynamic changes in the match when predicting match results.

4. Conclusions

This study proposes a machine learning model (PSPPM) aimed at enhancing the evaluation and prediction accuracy of athlete performance in sports competitions. The PSPPM model successfully visualizes the competition process by integrating the dynamic changes and trends of time series, and is capable of predicting athletes' scores in real-time. The model adopts a stacking method that combines logistic regression with XGBoost, preserving the interpretability of logistic regression while incorporating the feature importance assessment capabilities of XGBoost, thereby significantly improving the model's generalization ability in handling complex problems. The PSPPM model holds vast potential for application in the field of sports competitions. Its high-precision prediction capabilities and visualization of the competition process can not only provide scientific training guidance for coaching teams but also offer precise event planning and data analysis support for event organizers. Furthermore, with the continuous advancement of technology and the increasing abundance of data resources, this model is expected to further expand its application scope in the future, contributing more to the development of the sports industry.

References

[1] Yue J C, Chou E P, Hsieh M H, et al. A study of forecasting tennis matches via the Glicko model[J]. Plos one, 2022, 17(4): 100-108.

[2] Bayram F, Garbarino D, Barla A. Predicting tennis match outcomes with network analysis and machine learning[C]//47th International Conference on Current Trends in Theory and Practice of Computer Science, 2021: 505-518.

[3] Chiang S, Denes G. Supervised Learning for Table Tennis Match Prediction[J]. arxiv preprint arxiv:2023, 43(2):200-213.

[4] Zaidi A, Al Luhayb A S M. Two statistical approaches to justify the use of the logistic function in binary logistic regression[J]. Mathematical Problems in Engineering, 2023, 23(1): 552-560.

[5] Elkahwagy D M A S, Kiriacos C J, Mansour M. Logistic regression and other statistical tools in diagnostic biomarker studies[J]. Clinical and Translational Oncology, 2024, 26(9): 2172-2180.

[6] Gonaygunta H. Machine learning algorithms for detection of cyber threats using logistic regression[J]. Department of Information Technology, University of the Cumberlands, 2023, 34(10):1200-1208.

[7] Patel R K, Aggarwal E, Solanki K, et al. A Logistic Regression and Decision Tree Based Hybrid Approach to Predict Alzheimer's Disease[C]//2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES). IEEE, 2023: 722-726.

[8] Joshi A, Vishnu C, Mohan C K, et al. Application of XGBoost model for early prediction of earthquake magnitude from waveform data[J]. Journal of Earth System Science, 2023, 133(1): 512-520.

[9] Hong Z, Tao M, Liu L, et al. An intelligent approach for predicting overbreak in underground blasting operation based on an optimized XGBoost model[J]. Engineering Applications of Artificial Intelligence, 2023, 126(3): 107-115.

[10] Maleki A, Raahemi M, Nasiri H. Breast cancer diagnosis from histopathology images using deep neural network and XGBoost[J]. Biomedical Signal Processing and Control, 2023, 86(5): 105-112.