

# Predicting and Analysing Match Momentum in Tennis Matches Using Entropy Weight Method-based TOPSIS and XGBoost Model

Jiyuan Zhang<sup>†,\*</sup>, Yuyang Lin<sup>†</sup>

School of Physics, Zhengzhou University, Zhengzhou, China, 450001

\*Corresponding author: william604734103@stu.zzu.edu.cn

<sup>†</sup> These authors contributed equally.

**Abstract.** Tennis players' round performance scores and the prediction and analysis of match momentum in tennis matches are important for players, coaches and the media. Accurate and appropriate predictions and analyses are important for players and coaches to adjust their tactics in time, for media to analyse and report, and for spectators to experience and interact. Firstly, in order to quantify the performance of players, a comprehensive evaluation model of TOPSIS based on entropy weight method was constructed based on the theory of TOPSIS model and the advantage of entropy weight method in reducing subjectivity in the evaluation process. Secondly, in order to accurately predict the match momentum of tennis matches, XGBoost (Extreme Gradient Boosting) model was constructed. The model not only has extremely high prediction accuracy, but also improves the generalisation ability of the model, reduces overfitting, and has the advantages of automation and robustness. Finally, the constructed machine learning model is reasonably interpreted using the shap model to provide a scientific basis for players and coaches to change tactics and adjust training methods in a timely manner.

**Keywords:** TOPSIS, Entropy Weight Method, XGBoost, Match Momentum.

## 1. Introduction

The sports industry is one of the most prosperous business sectors in the world today. The global sports market will be valued at approximately US\$ 512.14 billion by 2023, growing at a CAGR of 5.2% from US\$ 486.61 billion in 2022. This figure is expected to climb further by 2027, when it is expected to reach nearly \$623.63 billion. The growth of the sports industry has been growing at an ever-increasing rate as people's interest in physical activity grows day by day. As a result, more and more researchers are studying and analysing sports data, including scientific evaluation of athletes' performances in sports competitions as well as predictions of game scores and momentum, which is important for athletes and coaches, media bloggers and the sports betting industry.

With the booming of the sports industry, especially the booming of tennis, it becomes especially important to objectively evaluate the performance of players and predict the trend of the game. Tennis, as a globally popular sport, not only attracts many fans, but also becomes one of the key areas of scientific research and analysis. At the same time, machine learning has become an important tool for research in this field. In 2016, Kovalchik surveyed various academic papers addressing tennis prediction in 2016 and before [1], and summarised three types of methods, i.e., regression-based methods, point-based methods, and pair-wise comparison methods. In general, studies of regression-based methods use probit or logit estimators to predict the winner of a match based on the rankings of two players in the Association of Tennis Professionals (ATP). Papers belonging to the points-based approach focus on predicting the individual points of a win. Finally, the pairwise comparison method estimates the probability of winning based on a direct assessment of the potential ability of the two players. Examples of this approach are the dynamic pairwise comparison model proposed by Baker and McHale [2] and the Elo rating system. Many studies to date have targeted increasing sports betting odds revenues by improving the accuracy of tennis match outcome predictions. Examples of the same are [3-4]. In addition to the predictive models described above, several recent studies have used fairly simple machine learning methods to predict match outcomes by compiling, cleaning and

using the largest database of tennis match information to date. Their prediction accuracy was as high as 80% [5]. Some studies have also built on the standard Elo model by weighting the Elo model according to the score of the player's last match to obtain Weighted Elo (WElo), which showed that the WElo method, in terms of minimising the loss functions considered (i.e., Brier's Score (Brier, 1950) and the logarithmic loss function) as well as in terms of maximising the return on investment realised in the betting market, was able to make more accurate predictions [6]. Candila and Palazzo used a neural network framework to predict winning tennis players [7]. There have also been studies aimed at improving the quality of data analysis in tennis by combining wristband data generation and deep convolutional neural network (CNN) classification and using recent advances in the field of machine learning such as the Mish activation function and the Ranger optimiser [8]. Jack et al proposed a statistical method for predicting the outcome of Grand Slam tournament matches, in addition to applying Exploratory Data Analysis (EDA) to explore variables related to tournament outcomes, and the results showed that the prediction accuracy of their model was slightly higher than that of machine learning models [9]. Peter et al used a Bayesian hierarchical model to predict and analyse trends in tennis players' serves [10].

It is worth noting that all of the above articles focus on predicting who wins and who loses as a result of a match, ignoring the impact of player performance on match trends and the impact of the various influences in the match on the players that are mapped to the impact on match trends. Therefore, this article combines the evaluation model with the machine learning model and the machine learning explanatory model, and on the basis of training and obtaining a very high prediction accuracy, the results of the machine learning explanatory model are analysed to provide a scientific basis for the players and coaches to adjust the match strategy and training strategy in a timely manner, which also provides a new analysis and solution to similar problems and supplements the gap in the field.

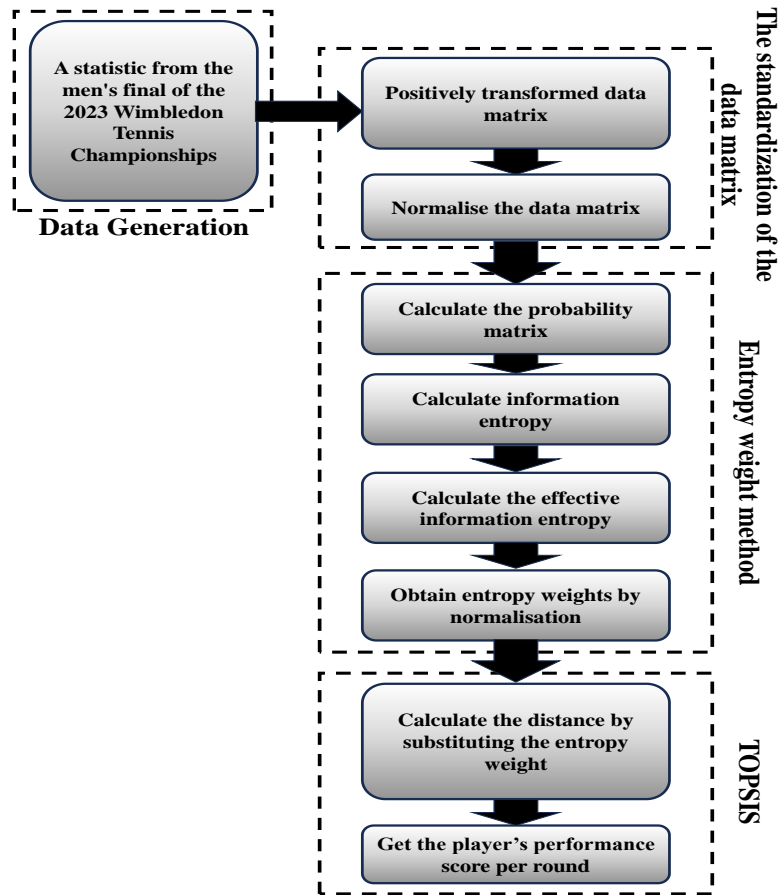
## 2. Model

### 2.1. Entropy Weight Method-based TOPSIS

Entropy Weight Method and Topsis Model (Technique for Order Preference by Similarity to Ideal Solution) are two models commonly used in multi-indicator decision analysis. Entropy Weight Method is a method for determining weights of multiple indicators, which is based on the concept of information entropy, in decision-making problems, if the entropy value of an indicator is larger, it means that the indicator contains more information, and has more influence on the decision-making process. Topsis Model is a method used for decision-making solution ordering. Using them together, first determining the weights by entropy weighting method and then applying them to Topsis model can make the decision-making process more objective and scientific. Such a combined model avoids the more subjective way of decision making and makes the decision results more convincing and credible.

The methodological process framework used in this study is shown in Figure 1. It consists of four parts:

- 1) Data matrix collection
- 2) The standardization of the data matrix: firstly, positively transformed data matrix, based on the judgement that only very small indicators existed in the data matrix that needed to be Positively transformed, so Equation 1 was used, and then the positively transformed data matrix was normalised using Equation 2.
- 3) Calculation of weights using the entropy weighting method: Calculate the proportion of each sample data under each indicator and consider it as the probability used to calculate the relative entropy Equation 3, calculate the information entropy Equation 4, calculate the information RMS value Equation 5, and then derive the weights of each indicator through the normalisation Equation 6.



**Figure 1.** Workflow diagram of entropy weighting method based on Topsis

4) Substitute the weights into Equation 7 to calculate the distance, and then Equation 8 to derive the performance scores of the two athletes in each round.

$$x = \max \{x\} - x, \tag{1}$$

Where x is the data for the indicators: p1\_serve\_no, p2\_serve\_no, p1\_double-fault, p2\_double-fault, p1\_unf\_err, p2\_unf\_err.

$$z_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^n x_{ij}^2}}, \tag{2}$$

Where x is each element in the data matrix, n is total number of rounds, z is the normalised data matrix.

$$p_{ij} = \frac{z_{ij}}{\sum_{i=1}^n z_{ij}}, \tag{3}$$

Where p is the probability matrix.

$$e_j = -\frac{1}{\ln n} \sum_{i=1}^n p_{ij} \ln(p_{ij}), \tag{4}$$

Where j represents the jth indicator,  $e_j$  is the information entropy of the jth indicator.

$$d_j = 1 - e_j, \tag{5}$$

Where  $d_j$  is the effective information entropy of the  $j$ th indicator.

$$w_j = \frac{d_j}{\sum_{j=1}^n d_j}, \quad (6)$$

Where  $w_j$  is the entropy weight of the  $j$ th indicator.

$$D_i^+ = \sqrt{\sum_{j=1}^n w_j (Z_j^+ - z_{ij})^2}, \quad D_i^- = \sqrt{\sum_{j=1}^n w_j (Z_j^- - z_{ij})^2}, \quad (7)$$

Where  $Z_j^+$  is the row vector consisting of the lower maximum value of each indicator,  $Z_j^-$  is the row vector consisting of the minimum value under each indicator.

$$S_i = \frac{D_i^-}{D_i^+ - D_i^-}, \quad (8)$$

Where  $S_i$  is the performance score of the  $i$ th round.

## 2.2. XGBoost

XGBoost is an algorithm based on Gradient Boosting Decision Trees (GBDT). It enhances the model by iteratively adding new decision trees, each constructed to correct the errors of the previous tree. In this article, we will select performance data such as player kills and missed balls as inputs and predict whether the winner of a given ball is more likely to be player one or player two by utilising the regression capabilities of the XGBoost model. The significant advantage of the XGBoost model over traditional gradient models such as GBDT is that it uses a Taylor expansion to optimise the loss function and introduces a regularisation term to prevent overfitting. Its objective function is derived as follows:

1) Defining the objective function: XGBoost's objective function is composed of a loss function and a regularisation term. The loss function measures the difference between model predictions and actual labels, while the regularisation term is used to control the complexity of the model.

$$L(\theta) = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k), \quad (9)$$

Where  $l(y_i, \hat{y}_i)$  is the loss function,  $\Omega(f_k)$  is the regular term.

2) Adding the structure of the tree: in the objective function, each tree  $f(x)$  is represented as a set of weights  $w$  of the leaf nodes and a function that maps instances to leaf nodes  $q(x)$ .

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i), \quad (10)$$

$$f_t(x) = \omega_{q(x)}, \quad (11)$$

Where  $\hat{y}_i^{(t)}$  is the predicted value of the prediction model after boosting from  $t$  trees.

3) Calculating the first and second order derivatives of the loss function: to optimise the objective function, XGBoost calculates the first and second order derivatives of the loss function. These derivatives are used in gradient descent and Newton's method to optimise the objective function.

4) Optimising the objective function: the XGBoost uses Taylor expansion to approximate the objective function and uses first and second order derivatives to optimise the weights  $w$  and structure  $q(x)$  of each tree.

$$L(\theta) \approx \sum_i \left[ l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t), \quad (12)$$

Where  $g_i$  and  $h_i$  are the first and second order derivatives of the loss function, respectively.

5) Calculation of splitting gain: XGBoost decides whether to split or not by comparing the objective function values before and after splitting. The splitting gain is calculated as:

$$GAIN = \frac{1}{2} \left( \frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right) - \gamma, \quad (13)$$

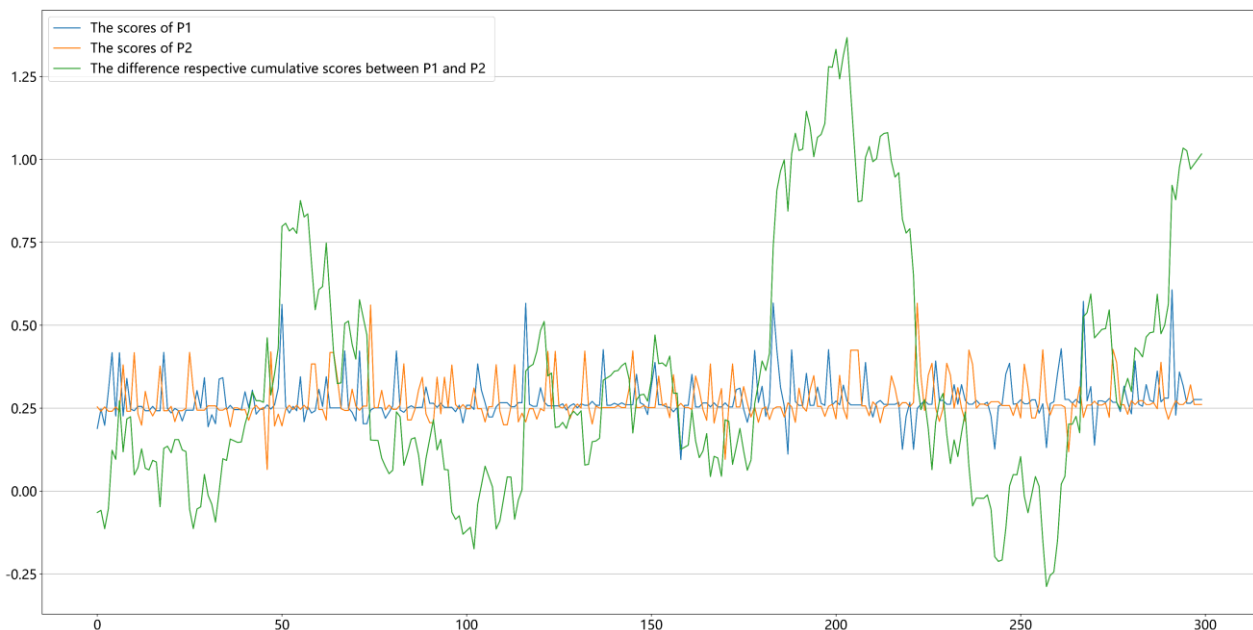
Where  $G_L$  and  $G_R$  are the sum of the second order derivatives of the left and right child nodes,  $\lambda$  is the L2 regularisation factor, and  $\gamma$  is the complexity parameter of the tree.

In summary, XGBoost improves the performance and efficiency of the model by introducing regularisation terms in the objective function, using second-order derivatives, and optimising the system implementation, making it perform well in various machine learning tasks.

### 3. Results

#### 3.1. Results of the Entropy Weighting Method-based TOPSIS

In this subsection, we use the Topsis model based on the entropy weighting approach to quantitatively rate the performance of two players, Carlos Alcaraz and Nicolas Jarry, in a match in the men's final of the 2023 Wimbledon Tennis Championships. to reveal the dynamics of the match and its evolution. By defining and studying the momentum of the game, it is possible to gain a deeper understanding of the nature of the game and the laws behind it, providing a more effective basis for analysing the results and formulating strategies. Here, we use the difference between the phenotypic situation scores of the two players to define the momentum of the match, and the results of the momentum of the match and the performance scores of the two players are shown in Fig. 2.



**Figure 2.** The graph of the momentum of the game and the evolution of the performance scores of the two players over the first three hundred overs

### 3.2. Analysis of experimental results

According to the match momentum evolution graph, the momentum in the match mostly hovers around 0.24, which shows that the average strength of the two players is not balanced. When a player gets a rare break chance or consecutive service points, there will be a clear peak in momentum, which means that one of the players is clearly overpowering the other, while when such a peak shows a clear decline, it means that the player who has the upper hand has made a serious mistake.

With the help of the relationship between momentum and force in physics i.e. Equation 14 and defining force as a factor that affects the trend of a match, i.e. differences in technical ability, mental state, fitness and endurance, tactics and strategy. Thus we summarise the two objective laws:

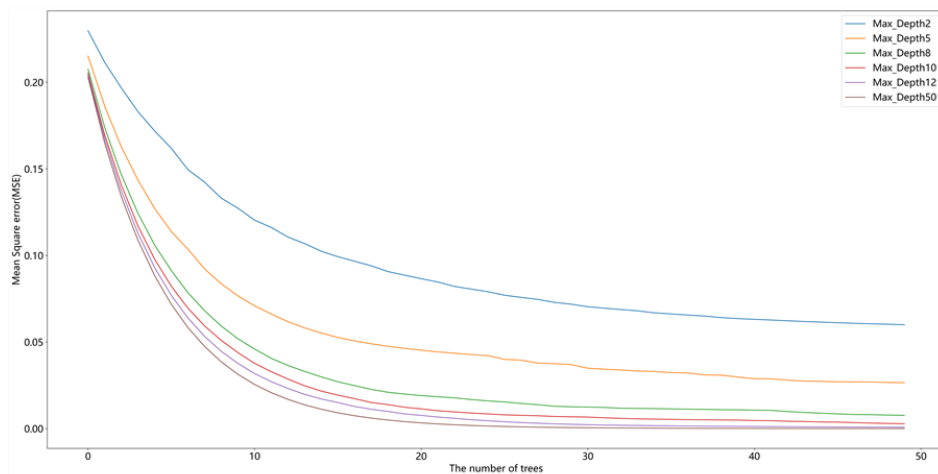
(1). When the absolute value of momentum is increasing, the factors influencing the trend of the match are biased in favour of the player who is in a position to gain the upper hand in the match, and this player will continue to accumulate advantages and have a stronger hold on the other player.

(2). When the absolute value of momentum is decreasing, the factors influencing the trend of the match favour the player who is in a disadvantageous position in the match, who continues to build up an advantage over the player who is in the ascendant position.

$$\text{Match Momentum} = y_1 - y_2, \tag{14}$$

Where  $y_1$  is Carlos Alcaraz’s performance score per round,  $y_2$  is Nicolas Jarry’s performance score per round.

### 3.3. Results of the XGBoost model



(a) Trend of mean squared error (MSE) with increasing number of trees at different maximum training depths.



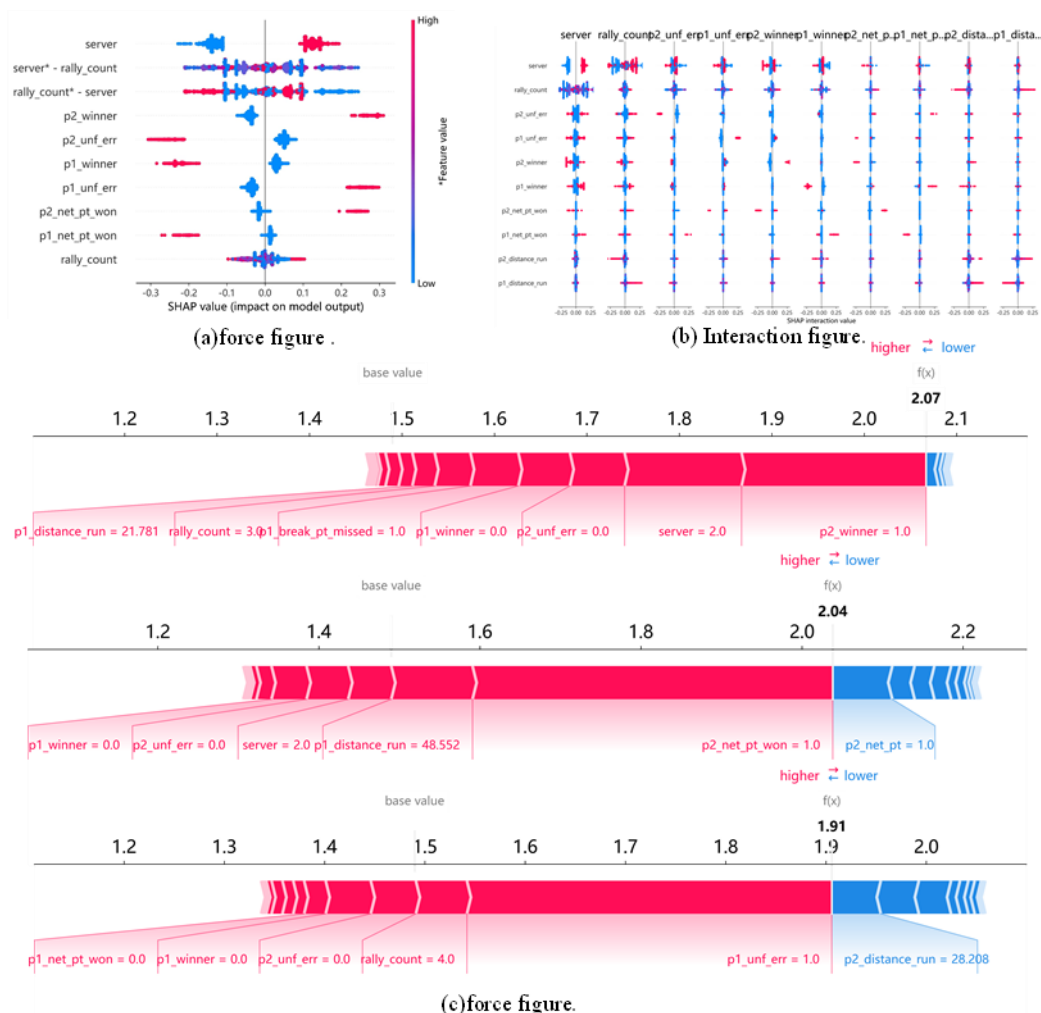
(b) The fit of the XGB model to the original data after training.

**Figure 3.** Training of the XGBoost model

In this section, we utilize an XGBoost model to perform regression training on the match momentum using various features from the dataset. The training process and results are shown in Figure 3, while the regression evaluation metrics are listed in Table 1. To better explain the impact of various factors on match momentum, we employed a model interpretation method: SHapley Additive exPlanations (SHAP). This method helps interpret the predictions of the machine learning model, allowing us to comprehensively and accurately assess the contributions of each feature as well as the interactions between features from both macro and micro perspectives. The results are illustrated in Figure 4.

**Table 1.** Regression evaluation metrics for the XGBoost model

Indicator scoring	Training set	Test Set
MSE	0.0008	0.0056
MAE	0.0117	0.0188
RMSE	0.0287	0.0746
MAPE	0.0087	0.0146
R <sup>2</sup>	0.9967	0.9777



**Figure 4.** Interpretation of match momentum at macro and micro level

### 3.4. Analysis of experimental results

Figure 3(a) illustrates the mean squared error (MSE) curves of the XGBoost model as the number of trees increases for different maximum tree depths. Two key observations can be made from this figure: First, for models with different maximum depths, their MSE tends to reach a minimum after training up to 30 trees. Additionally, as the maximum tree depth increases sequentially from 2 to 5,

8, 10, and 12, the MSE at the 50th tree becomes progressively smaller, approaching zero. However, when the maximum depth suddenly increases to 50, there is almost no reduction in MSE. Moreover, a maximum depth of 50 significantly increases computational costs compared to a depth of 12, leading to performance degradation and a higher risk of overfitting. Therefore, from Figure 3(a), we can deduce that the optimal model parameters are a maximum tree depth of 12 and 50 trees for training.

Figure 3(b) separately depicts the original data curve of the y1-y2 difference, the predicted data curve from the XGBoost model, and the loss value between the predictions and the target values. It can be visually observed that the predicted data curve almost overlaps with the original data curve, and the loss function fluctuates slightly around zero. This indicates that the XGBoost model performs well in training and has excellent trend prediction capabilities for this dataset.

Evaluation metrics based on the regressions shown in Table 1:

MSE measures the average squared difference between the predicted and true values of a model, with lower values indicating a better fit. MAE measures the mean absolute error between the predicted and true values of the model, which is the average of the absolute values of the individual errors. Similar to MSE, the lower the value, the better the model fit. The RMSE is the square root of the MSE, which has the same units as the target variable and is therefore easier to interpret. As with the MSE, lower values indicate a better fit to the model. MAPE measures the mean percentage error between the predicted value and the true value, which is the average of the ratios of the absolute values of the individual errors relative to the true value. Lower MAPE values indicate that the model's predictions are more accurate. The R-squared value measures how well the model fits the observed data, and ranges from 0 to 1, with values closer to 1 indicating a better fit to the data. All of the above metrics show that the XGBoost model has an excellent fit for the regression prediction of momentum in tennis matches.

Now let's analyse Figure 4(a), where each row represents an influencing factor and the horizontal coordinate is the Shap value. The features are sorted according to the average absolute value of the Shap value, and the top 10 most influential features are shown here, sorted from top to bottom according to their impact on the momentum of the game. A dot represents a sample, with redder colours indicating larger feature values and bluer colours indicating smaller feature values. Wide places indicate a large number of samples clustered together. It can be seen that servers have the greatest impact on match momentum, and samples with high server values have shap values greater than 0, indicating that they have a positive impact on match momentum, as do p2\_winner, which is ranked 4th in terms of the impact factor, p1\_unf\_err, which is ranked 7th in terms of the impact factor, and p2\_net\_pt\_won, which is ranked 8th in terms of the impact factor, which are all positive factors. On the contrary, p2\_unf\_err, ranked 5th in the impact factor, and p1\_winner, ranked 6th in the impact factor, as well as p1\_net\_pt\_won, ranked 9th in the impact factor, are all negative factors. In contrast, both the interaction term between server and rally\_count and rally\_count itself have a balanced impact on the momentum of the match. In Fig. 4 (c), we take the data of the 6th 151st 251st overs to represent the early, mid and late game respectively to show the impact of each factor on the momentum of the game at the micro level through force figure. It can be seen that p2\_winner at the beginning of the game, p2\_net\_pt\_won and p2\_net\_pt at the middle of the game, and p1\_unf\_err and p2\_distance\_run at the end of the game have the greatest impact on the momentum of the game. Where p2\_net\_pt and p2\_distance\_run are negative influences.

#### 4. Conclusions and outlooks

In this article, we first quantitatively score the performance of tennis players in each round of the match, then introduce the concept of match momentum to obtain the evolution of the match, and finally predict the match momentum and analyse the role of each influencing factor on the match momentum. In order to quantitatively rate the performance of the two players in each round of the match and to obtain the evolution of the match, a Topsis model based on the entropy weighting method was established and the momentum of the match was defined, and the following conclusions

were obtained by substituting the data of the men's final match of the Wimbledon Tennis Championships in 2023 into the model: when one of the players obtains a rare break of serve or consecutively receives the points on serve, a significant peak of the momentum occurs, which indicates that the momentum of the match is not as high as that of the other players. A significant peak indicates that one side is clearly overpowering the other, while when there is a significant drop in this peak, it indicates that the prevailing side has made a serious mistake. In order to predict the momentum of the game and analyse the role of each influencing factor on the momentum of the game, we built a XGBoost model and interpreted and analysed the machine learning model through a shap model, the  $R^2$  of the test set reached 0.9777, and through analysis we obtained that servers have the greatest influence on the momentum of the game, and the shap value of the samples with a high value of servers is greater than 0, indicating that servers have the greatest influence on the momentum of the game. It indicates that servers have a positive impact on match momentum, as do p2\_winner, which ranks 4th in the impact factor, p1\_unf\_err, which ranks 7th in the impact factor, and p2\_net\_pt\_won, which ranks 8th in the impact factor, all of which are positive factors. The model established in this paper is of great significance for players and coaches to adjust tactics and training plans in time, for media to analyse and report, and for spectators to experience and interact.

In the model we built, the XGBoost model may suffer from a local optimal solution problem without being able to find the global optimal solution. This may lead to poor model performance because it does not fully explore the hyperparameter space. We believe that more diverse search strategies can be introduced to better explore the hyperparameter space and reduce the risk of falling into local optimal solutions. For example, stochastic search or genetic algorithms can be combined to increase the diversity of the search.

## References

- [1] Kovalchik S A. Searching for the GOAT of tennis win prediction [J]. *Journal of Quantitative Analysis in Sports*, 2016, 12 (3): 127-138.
- [2] Baker R D, McHale I G. An empirical Bayes model for time-varying paired comparisons ratings: Who is the greatest women's tennis player? [J]. *European Journal of Operational Research*, 2017, 258 (1): 328-333.
- [3] Wilkens S. Sports prediction and betting models in the machine learning age: The case of tennis [J]. *Journal of Sports Analytics*, 2021, 7 (2): 99-117.
- [4] Vaughan Williams L, Liu C, Dixon L, et al. How well do Elo-based ratings predict professional tennis matches? [J]. *Journal of Quantitative Analysis in Sports*, 2021, 17 (2): 91-105.
- [5] Gao Z, Kowalczyk A. Random forest model identifies serve strength as a key predictor of tennis match outcome[J]. *Journal of Sports Analytics*, 2021, 7 (4): 255-262.
- [6] Angelini G, Candila V, De Angelis L. Weighted Elo rating for tennis match predictions [J]. *European Journal of Operational Research*, 2022, 297 (1): 120-132.
- [7] Candila V, Palazzo L. Neural networks and betting strategies for tennis[J]. *Risks*, 2020, 8 (3): 68.
- [8] Ganser A, Hollaus B, Stabinger S. Classification of tennis shots with a neural network approach [J]. *Sensors*, 2021, 21 (17): 5703.
- [9] 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): e0266838.
- [10] Tea P, Swartz T B. The analysis of serve decisions in tennis using Bayesian hierarchical models [J]. *Annals of Operations Research*, 2023, 325 (1): 633-648.