Research on the Momentum of Tennis Players Based on the XGBoost Model and GA-RF Model

Tao Lu[1,*], Luxian Wang[1], Jiaxi Li[2]

[1] School of Management, Nanjing University of Posts and Telecommunications, Nanjing, China, 210023

[2]School of Science, Nanjing University of Posts and Telecommunications, Nanjing, China, 210023

*Corresponding author: b22130714@njupt.edu.cn

Abstract. In this paper, to better represent an athlete's performance using momentum, this paper proposed an indicator system to assess a player's performance at different stages of a match. The article tested this system using both LightGBM and XGBoost classification models. By comparing the accuracy, recall, and precision of the training sets for both models, this paper found that LightGBM achieved 91.8% for all these metrics, while XGBoost achieved an impressive 98.4%. Therefore, this paper used the feature importance derived from the XGBoost model to determine the weight of these indicators. After obtaining the weighted indicators for the players, the essay used a comprehensive calculation formula to assess the momentum changes of a player during the match. The essay also applied a genetic algorithm-optimized random forest regression model (GA-RF) to predict real-time momentum changes in a match. Additionally, this paper utilized the SHAP model to interpret the predictions, providing specific recommendations for players to prepare for the match.

Keywords: Momentum, XGBoost Model, SHAP Model, GA-RF Model.

1. Introduction

In the field of sports competition, teams or athletes sometimes experience a force referred to as "momentum" or "momentum flow," which seems to influence the progression of a match. However, this concept of momentum is challenging to quantify, and its specific impact and fluctuations within a game are difficult to capture intuitively. To assist coaches in gaining a deeper understanding of the effects of momentum on the game and to provide athletes with more targeted preparation advice, it is necessary to conduct a quantitative analysis of athletes' performance or momentum performance. Through this analysis, we can explore whether momentum indeed plays a role in competitions and how its trend might be predicted. Such in-depth research can help to reveal the essence of momentum in sports competitions, providing a scientific basis for developing more precise strategies and training plans.

Many researchers have provided extensive literature reviews on momentum, and [1] proposed a new theoretical framework that offers a fresh perspective for discussing momentum. In the field of sports competition, an athlete is often influenced by psychological or physiological factors, which are usually related to momentum; therefore, momentum is closely linked to an athlete’s performance. However, there is currently a lack of specific methods to quantify the effects of momentum. This calls for the use of models to quantitatively describe momentum and to assess its impact on athletes based on momentum changes. By doing so, we can analyze and identify the key influencing factors and enable athletes to make tactical adjustments according to these main factors and the changes in momentum.

Also, many researchers have offered different perspectives on the application of momentum across various fields. Walid Briki and Ruud J. R. Den Hartigh examined the influence of dynamic systems on psychological momentum from a psychological perspective and proposed the concept of PM to measure its effects [2]. Matías José Gómez Seeber explored the effect of breaks during competitions, noting that such pauses can disrupt players' psychological momentum [3]. Additionally, Walid Briki proposed the concepts of psychological and behavioral momentum, describing PBM as a unified

phenomenon reflecting different manifestations of momentum through the integration of psychological, physiological, and behavioral structures [4].

On the other hand, Philippe Meier and Raphael Flepp argued that behavioral and strategic momentum are independent and that the momentum effect can be influenced by breaks [5]. Muhui Zhong and Zikang Liu introduced an integrated logistic regression model to predict momentum shifts during competitions and inform strategic decisions [6]. Elia Morgulev, Yangqing Zhao, and Hui Zhang likened momentum to the "hot hand" phenomenon, proposing the concept of "success breeds success." They explored the key psychological, physiological, and economic factors that drive success, using hit rates in archery and the advantage of winning two out of three sets to illustrate that early success often increases the likelihood of winning. Morgulev also explained how this "success breeds success" mechanism is evident across many fields [7-10]. However, a clear explanation of this effect remains elusive, and skepticism around the hot hand effect persists.

In many competitive sports, an athlete's previous performance is closely related to subsequent performance. Andrew E. Evans and Paul Crosby identified a "hot-hand" effect in golf, noting that men display more evident momentum than women, often resulting in better performance [11]. Additionally, after narrowly winning a match, the performance of female athletes tends to decline compared to before, while male athletes remain largely unaffected. This suggests that women may be more impacted by negative momentum effects [12]. In summary, this paper utilizes data from the 2023 Wimbledon Championships to investigate the presence of momentum in tennis, illustrating the concept of "success breeds success". The contributions of this paper are as follows:

We analyzed match data and extracted 15 influencing factors to quantify the effects of momentum. Using the XGBoost and LightGBM models, we calculated the impact weight of these 15 factors on players and compared the models' predictive accuracy. The XGBoost model achieved an accuracy rate of 98.4%. By identifying the key influencing factors through the model and using a comprehensive calculation formula, we derived a performance score for each player.

We developed a random forest regression prediction model optimized by a genetic algorithm, using momentum scores as the dependent variable to identify the primary factors influencing momentum fluctuations. This model also allowed us to predict momentum shifts during the match. Through data visualization, we observed that the model's predictions closely aligned with actual momentum fluctuations.

We also applied the SHAP model to interpret the predictions of the random forest model, identifying the main factors that drive momentum fluctuations. This approach provided insights into the primary influences on athletes' momentum shifts during matches. Moreover, we found that the effect of running distance on momentum is uncertain; it can either increase or decrease momentum, though there is a slight tendency for momentum to increase when the running distance is shorter. Additionally, when the opponent's serve speed is high, momentum slightly decreases; conversely, when the serve speed is lower, momentum tends to increase.

2. Establishment of momentum evaluation models and prediction models

2.1. Construction of Indicator System

We have established an evaluation system for assessing a player's scoring performance, and have set up some notations as shown in Table 1:

Table 1: Notations used in this paper

Symbol	Description
X0	This game scores
X1	The number of games won in the current set
X2	The lead in points for this set
X3	The lead in points for this game
X4	The previous point scored
X5	Is the server
X6	The player served an Ace
X7	The player hit an Ace
X8	Whether the player has committed two serve faults in this game
X9	Whether the player has made unforced errors in this game
X10	The ratio of successful net points to the total number of attempts at the net
X11	The ratio of opportunities for the player to win the game when the opponent serves to the actual victories in the game
X12	Whether the player missed an opportunity to win the game
X13	Serve speed
X14	Distance covered during this point
X15	Total distance covered in the last three points
M	The strength or momentum felt by players during a match/game

2.2. Establishment of the XGBoost model

The XGBoost classification model is designed for a training dataset containing Nsamples and M features $D = \{(x_i, y_i)\}, i = 1,2,\cdots,N, x_i \in R^M, y_i \in R$, The final predicted value of the XGBoost algorithm $\widehat{y_i}$ The ensemble model, composed of multiple classification and regression trees (CART), calculates the final result, which can be expressed as:

$$\widehat{y_i} = \sum_{k=1}^{K} f_k(x_i), f_k \in F \quad (1)$$

In the equation, K represents the number of decision trees; $f_k(x_i)$ in the term represents the predicted score of the $k-th$ CART for the $i-th$ sample in the dataset; F represents the function space formed by all the CART functions [13], $\widehat{y_i}$ represents the momentum prediction value for sample i after the iteration.

In the XGBoost algorithm, the objective function for model learning consists of two parts: the loss function and the regularization term. The regularization term is used to control model complexity and prevent overfitting. Its expression is as follows:

$$\omega(f_k) = \sum_{i=1}^{n} l(y_i, \widehat{y_i}) + \sum_{m=1}^{t} \Omega(f_m) \quad (2)$$

In the expression, the true value y_i represents the actual momentum score of sample i; The predicted value $\widehat{y_i}$ represents the momentum score predicted by the model for sample i; The loss function $l(y_i, \widehat{y_i})$ represents the difference between the predicted value and the true value, and it is used to describe the discrepancy between the model's predictions and the actual values; $\Omega(f_m)$ represents the regularization term, which is designed to prevent the model from overfitting by penalizing overly complex models.

2.3. A random forest regression prediction model based on genetic algorithm optimization

We treat the previously calculated player performance scores as momentum scores and the 15 key indicators defined earlier as variables that may influence momentum changes. By using these 15 indicators as independent variables and momentum scores as the dependent variable, we can analyze

the factors that have the greatest impact on momentum fluctuations and, based on this, predict the trend of momentum changes during the match.

Given that the relationship between momentum changes and various indicators exhibits non-linear characteristics, traditional analytical methods are often inadequate in revealing the exact connection between momentum and these indicators. Therefore, applying machine learning techniques to explore their relationships is more appropriate.

2.3.1 Random Forest Regression Model

The Random Forest (RF) algorithm is a popular ensemble learning method that consists of multiple decision trees, capable of building classification and regression models for different tasks. This algorithm handles classification and regression problems by finding the optimal splitting points, allowing for deep analysis of the data. Using the Bootstrap technique, we can randomly sample NNN instances with replacement from the original dataset, where each sample is of the same size as the original dataset, thus creating multiple training sets. Then, for each training set, a CART (Classification and Regression Tree) model is built independently.

In classification tasks, the prediction results from each tree are integrated using a majority voting mechanism to determine the final class label. In regression tasks, the average of the predictions from all trees is taken to obtain the final regression prediction.

2.3.2 Genetic Algorithm

Genetic Algorithm (GA) is a global optimization technique that simulates the biological evolution mechanism in nature. It draws inspiration from natural selection and gene recombination in biological genetics to solve complex optimization problems.

In the Random Forest regression model, the genetic algorithm is used to optimize feature selection and model parameters to enhance the model's predictive power. The optimization process involves the following steps:

(1) Fitness Function Setup: A target function is defined to evaluate the model's performance, such as Mean Squared Error (MSE), which serves as the fitness metric to assess the quality of the model.

(2) Initial Population Creation: A set of candidate solutions is randomly generated within the given parameter space. Each solution is represented by a vector (chromosome) containing multiple genes (decision variables).

(3) Fitness Evaluation: The fitness of each candidate solution is calculated, i.e., evaluating the corresponding objective function value to assess the quality of the solution.

(4) Selection Process: Candidate solutions are selected based on their fitness. Poor-performing solutions are eliminated, while better solutions are selected for reproduction, forming a new generation.

(5) Crossover and Mutation: Crossover (pairing and exchanging genes) and mutation (randomly altering certain genes) operations are performed to increase the genetic diversity of the population and explore a broader solution space.

(6) Iteration Process: The steps of selection, crossover, and mutation are repeated until a stopping condition is met, such as reaching a predetermined number of iterations or when improvements in fitness become negligible.

(7) Model Construction and Application: The best feature subset and parameters selected by the genetic algorithm are used to construct the Random Forest regression model, aiming to achieve better predictive performance and model stability.

2.3.3 Genetic Algorithm-Optimized Random Forest Regression Predictive Model

Random Forest improves prediction accuracy by aggregating the results of multiple decision trees. However, during model construction, challenges such as overfitting and model instability may arise.

To address these challenges, we can use the Genetic Algorithm (GA) to optimize the Random Forest model. GA is an optimization technique that simulates the process of biological evolution,

using natural selection and genetic variation to find the optimal solution to a problem. In Genetic Algorithm-optimized Random Forest (GA-RF), GA is used to select the best feature combinations and adjust the parameters of the decision trees. By utilizing genetic operators such as crossover and mutation, GA-RF can explore better parameter configurations, enhancing the model's generalization ability and reducing the risk of overfitting [14].

The specific algorithmic flow of the Genetic Algorithm-based Random Forest regression prediction model can be referenced in Figure 1.

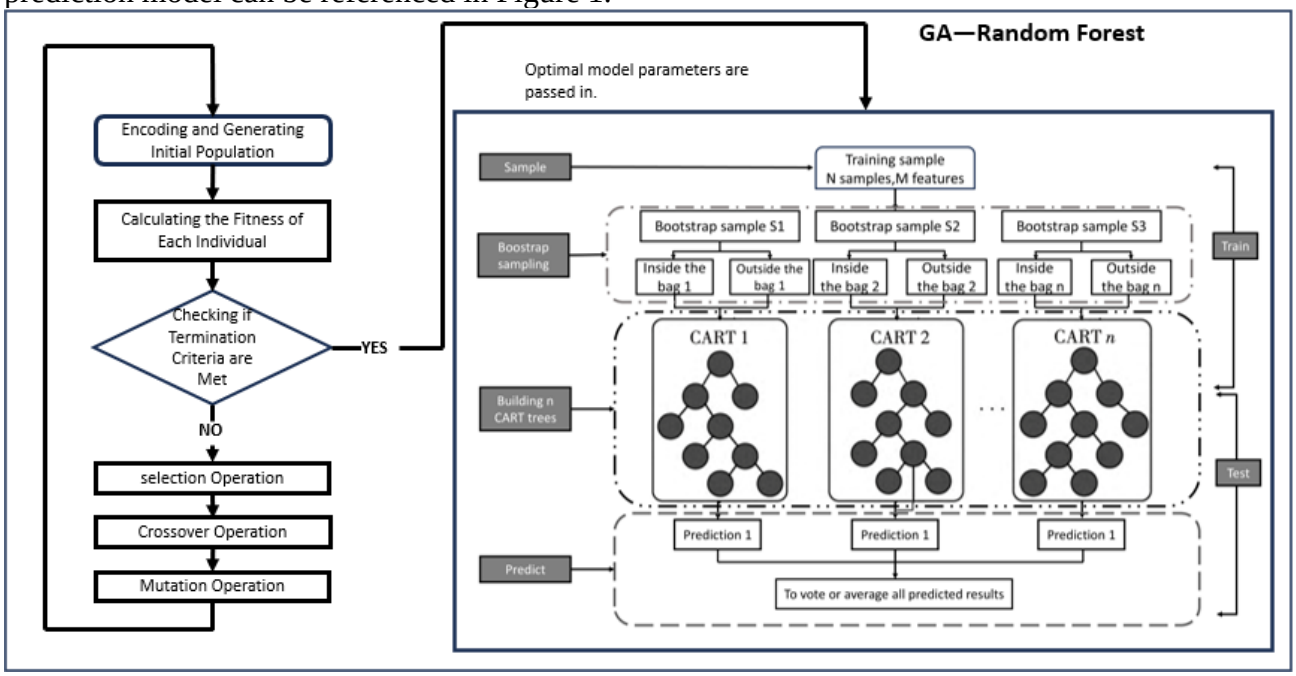


Figure 1 A-RF model schematic diagram

2.3.4 Analysis of Player Momentum Factors Based on GA-RF and SHAP Model

In previous applications, we effectively predicted momentum using the GA-RF model. Although Random Forest is widely used for its strong predictive power, its "black-box" nature makes the interpretability of prediction results a challenge. To improve the model's interpretability, we introduced the SHAP (SHapley Additive exPlanations) algorithm.

The SHAP algorithm, based on the Shapley value from game theory, is used to explain the predictions of machine learning models. It provides a comprehensive approach to interpret the outputs of any type of machine learning model. By applying SHAP, we can quantify the specific contribution of each feature to the prediction results, thus gaining a deeper understanding of the model's prediction logic.

In the Random Forest model, the SHAP algorithm helps us identify the key features that significantly impact the prediction outcomes. By analyzing the SHAP values for each feature, we can gain insights into how each feature influences the model's predictions, thereby revealing the underlying mechanisms of the model [15]. This approach not only enhances the transparency of the model but also provides valuable insights for improving and optimizing the model.

3. Result

3.1. XGBoost model results and analysis

After comparing the LightGBM and XGBoost classification models, we chose the one that performed better. Compared to XGBoost, LightGBM has advantages in training efficiency, memory usage, and prediction accuracy. It also supports distributed computing, handles large-scale datasets more efficiently, and helps reduce the risk of overfitting [16].

Key features of LightGBM include a histogram-based decision tree construction method for optimization, a Leaf-wise tree growth strategy with depth limitation, gradient-based one-side sampling (GOSS), and exclusive feature bundling (EFB) technology.

These features contribute to LightGBM's superior performance, particularly in large-scale machine learning tasks, and its ability to efficiently handle complex datasets.

We used the 15 indicators mentioned above as independent variables, with the player's point score (whether the point is won) as the dependent variable. The XGBoost model achieved an accuracy of 98.4%, and the LightGBM classification model had an accuracy of 91.8%. Based on these results, we selected the XGBoost model. The feature importance according to Figure 2 is as follows:

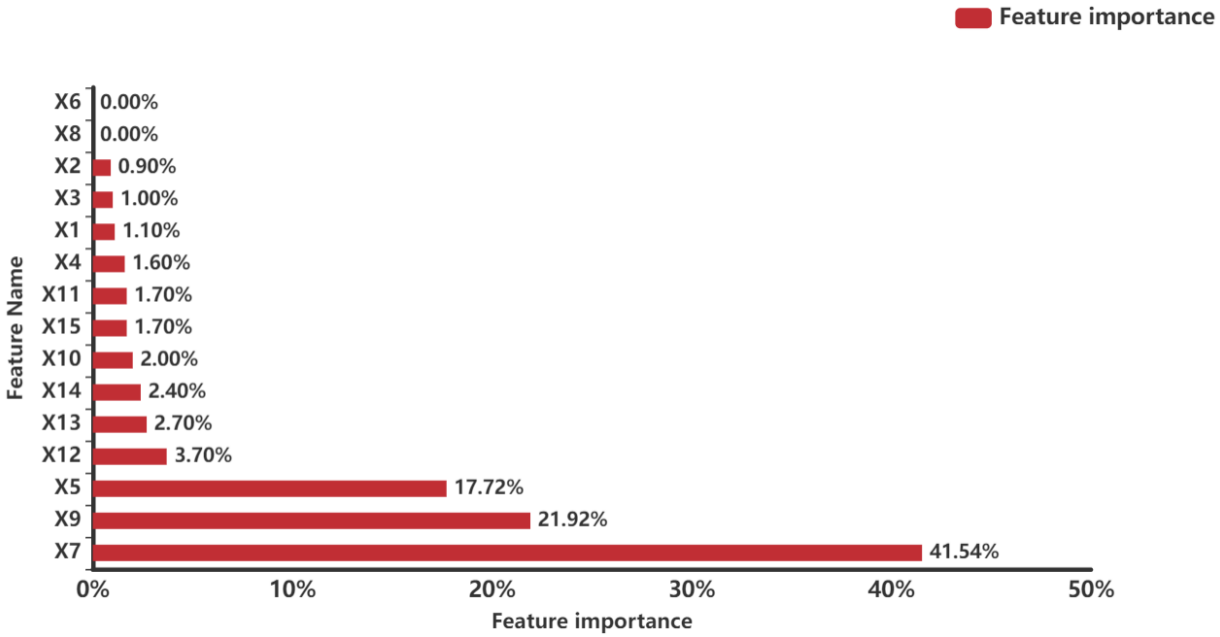


Figure 2 Weights of indicators

X7: 41.5%; X9: 21.9%; X5: 17.7%; X12: 3.7%,X13: 2.7%; X14: 2.4%; X10: 2%.The weights for X6 and X8 are both 0.

After a series of standardizations, the final weight results are shown in Table 2.

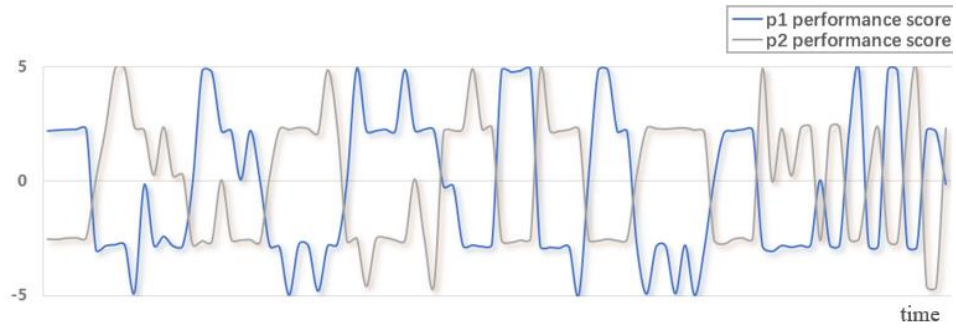
To avoid the impact of the scale of the data on the results, we applied Min-Max normalization to the indicator scores, scaling them within the range of [-5, 5]. Next, we multiplied each indicator by its corresponding weight to obtain the comprehensive performance score for the player.The calculation formula is as follows:

$$M = X7 * 45\% + X9 * 24\% + X5 * 19\% + X12 * 4\% + X13 * 3\% + 10 * 2.3\% \quad (3)$$

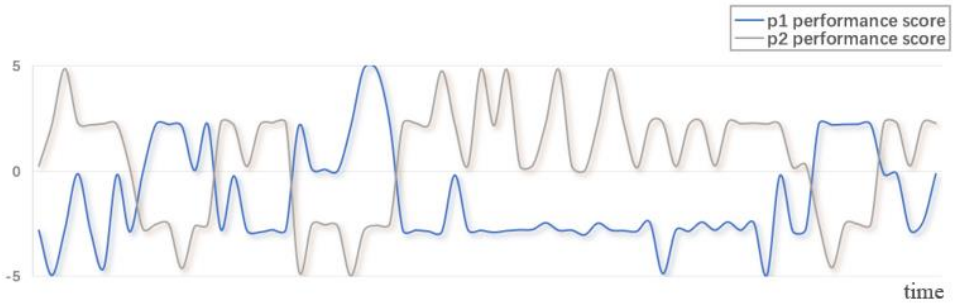
Table 2 Indicator Weights

Index	X7	X9	X5	X12	X13	X14	X10
Weight	45%	24%	19%	4%	3%	2.7%	2.3%

Based on this formula, we can calculate the performance score for the player at each moment. Taking the data from the match "2023-wimbledon-1701" as an example, we visualized the performance of the two players in the second and third sets, as shown in Figure 3(a) and (b). We observed a negative correlation between the performance scores of the two players: when one player performed better, the other player's performance was relatively weaker. This aligns with our expectations and indicates that the model we constructed is logically sound.



(a) Second set player performance score



(b) Third set player performance score

Figure 3 Player performance score

3.2. GA-RF Model Prediction Analysis

The optimal parameters of the RF model calculated by GA are shown in Table 3.

The 15 normalized indicators (x1 ~ x15) are used as input features, with the momentum score S as the target output. The optimal parameters R, calculated by the genetic algorithm, are applied to set the hyperparameters of the RF model. The model's prediction results are shown in Figure 4.

From the visualization, we can clearly see that the GA-RF model fits the fluctuations in momentum very well, indicating that the model has strong predictive capability.

Table 3 Optimal Parameters for the RF Model

Number of decision trees	459.94544513
Maximum depth	27.0666075
Minimum samples for split	4.96569574
Minimum samples for leaf node	1.38113491
Maximum features for split	0.84607177

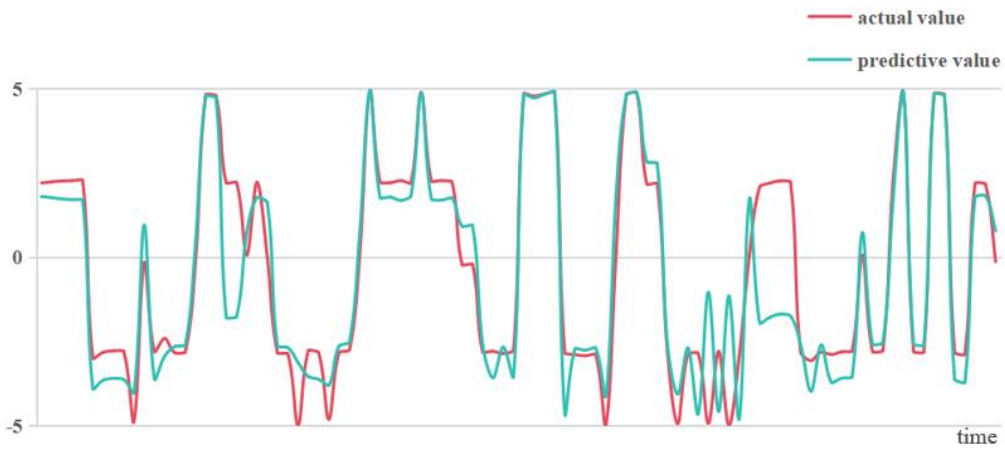


Figure 4 Prediction result fitting graph

As shown in Figure 5, the correlation between the number of iterations and the fitness value during the execution of the genetic algorithm is displayed. As the number of iterations increases, the fitness

value shows a downward trend, reflecting that the algorithm is gradually evolving towards a better solution.

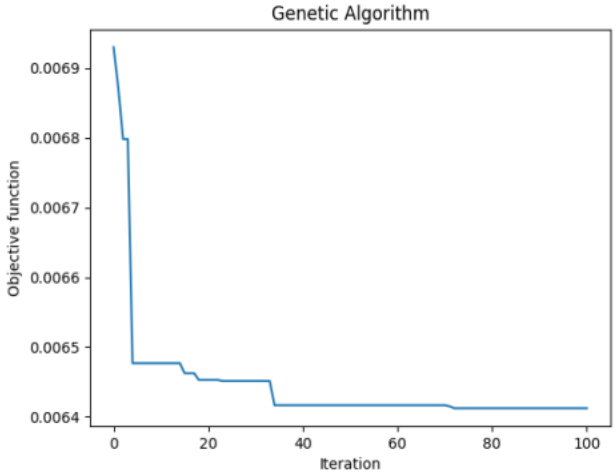
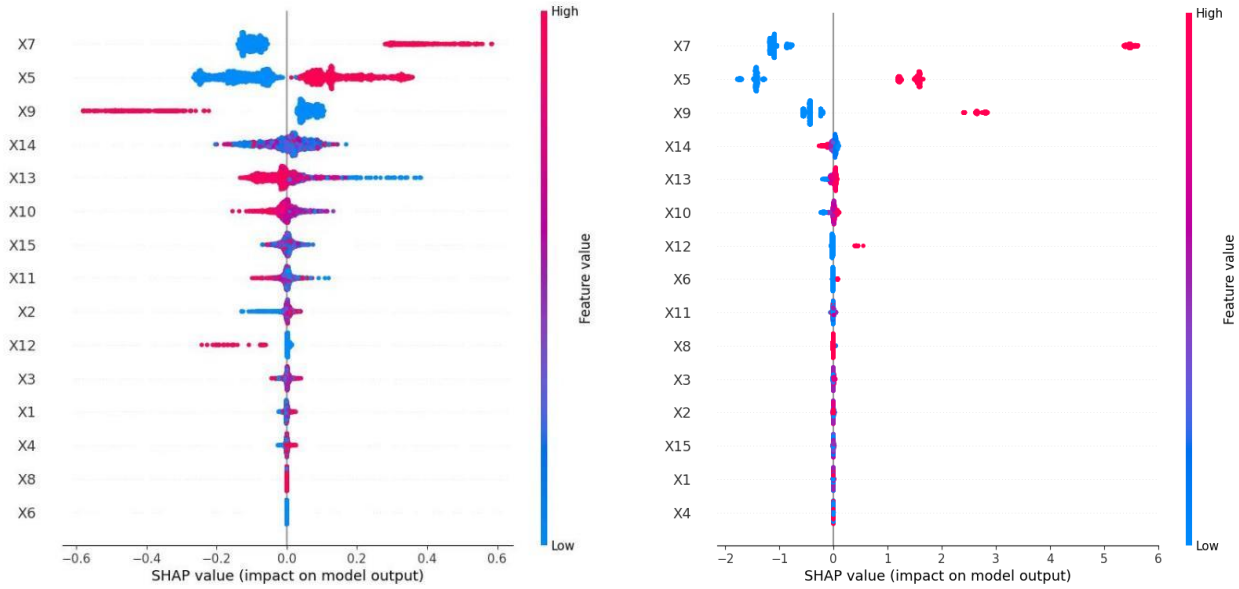


Figure 5 Relationship Between Iterations and Fitness

3.3. SHAP model results and analysis

We conducted SHAP algorithm calculations based on GA-RF using both all match data and the "2023-wimbledon-1701" match data as test data. The following are the SHAP calculation results. Figure 5 shows the SHAP plot based on all match data, while Figure 6 displays the SHAP plot based on the "2023-wimbledon-1701" match data.



(a)SHAP Plot Based on All Match Data

(b)SHAP Plot Based on "2023-Wimbledon-1701" Match Data

Figure 6 SHAP Plot

From Figure 6, we conclude that the main factors influencing momentum fluctuations include X7 (whether an Ace was served), X5 (whether the player is the server), X9 (whether there was an unforced error in the current game), X14 (the running distance during the point), and X13 (the serve speed). Among these, the impact of serving an Ace on momentum is the most significant.

4. Conclusions

This paper uses momentum to describe the changes in an athlete's performance during a match. First, the XGBoost model is used to evaluate the athlete's performance, effectively capturing the

momentum fluctuations during the game and making the concept of momentum more tangible. Then, the paper applies a genetic algorithm-optimized random forest model (GA-RF), which uses machine learning algorithms to analyze match data and predict the changes in a player's momentum. Through continuous iterations, the model refines its results, gradually converging toward the optimal outcome. Finally, SHAP is used to interpret the GA-RF model's predictions, enhancing the readability of the machine learning results and identifying the key factors that influence a player's momentum. The primary factors influencing an athlete's momentum, as interpreted through SHAP, can help athletes focus their training on improving these key areas. This targeted approach allows them to maintain better momentum during competitions, increasing their chances of winning. For researchers, quantifying momentum provides a clearer understanding of its fluctuations in matches and facilitates further studies on momentum. This quantification of momentum not only highlights its impact on athletes and individuals but also supports its application across various fields.

References

[1] Morgulev E, Azar O H, Bar-Eli M. Searching for momentum in NBA triplets of free throws[J]. Journal of Sports Sciences, 2020, 38(4): 390-398.

[2] Den Hartigh R J R, Van der Sluis J K, Zaal F T J M. Perceiving affordances in sports through a momentum lens[J]. Human Movement Science, 2018, 62: 124-133.

[3] Seeber M J G. Momentum-stop**: Effects on performance[J]. Sports Economics Review, 2024, 7: 100038.

[4] Briki W. Rethinking the relationship between momentum and sport performance: Toward an integrative perspective[J]. Psychology of Sport and Exercise, 2017, 30: 38-44.

[5] Meier P, Flepp R, Ruedisser M, et al. Separating psychological momentum from strategic momentum: Evidence from men's professional tennis[J]. Journal of economic psychology, 2020, 78: 102269.

[6] Zhong M, Liu Z, Liu P, et al. Searching for the Effects of Momentum in Tennis and its Applications[J]. Procedia Computer Science, 2024, 242: 192-199.

[7] Morgulev E. Success breeds success: Physiological, psychological, and economic perspectives of momentum (hot hand)[J]. Asian Journal of Sport and Exercise Psychology, 2023, 3(1): 3-7.

[8] Zhao Y, Zhang H. Does success breed success? An investigation of momentum in elite recurve archery[J]. Psychology of Sport and Exercise, 2023, 66: 102397.

[9] Morgulev E. Streakiness is not a theory: On "momentums"(hot hands) and their underlying mechanisms[J]. Journal of Economic Psychology, 2023, 96: 102627.

[10] Mago S D, Sheremeta R M, Yates A. Best-of-three contest experiments: Strategic versus psychological momentum[J]. International Journal of Industrial Organization, 2013, 31(3): 287-296.

[11] Evans A E, Crosby P, Shin S Y. Psychological momentum among non-experts: Evidence from club golfers[J]. Journal of Behavioral and Experimental Economics, 2023, 104: 102016.

[12] Lackner M, Weichselbaumer M. Can barely winning lead to losing? Evidence for a substantial gender gap in psychological momentum[J]. Evidence for a Substantial Gender Gap in Psychological Momentum (May 9, 2022), 2022.

[13] Hu X W, Ding R, Zhang Z Y, et al. A unified section-based restoring force model for reinforced concrete coupling beams based on the XGBoost model: Establishment, validation and application[J]. Journal of Building Engineering, 2023, 73: 106666.

[14] Han H, Wang W. A Hybrid BPNN-GARF-SVR PredictionModel Based on EEMD for Ship Motion[J]. CMES-Computer Modeling in Engineering & Sciences, 2023, 134(2).

[15] Fu Q, Wu Y, Zhu M, et al. Identifying cardiovascular disease risk in the US population using environmental volatile organic compounds exposure: A machine learning predictive model based on the SHAP methodology[J]. Ecotoxicology and Environmental Safety, 2024, 286: 117210.

[16] Ding Y. Construction of Financial Risk Analysis and Early Warning Model for Information Technology Enterprises Based on Complex Networks and LightGBM[J]. Procedia Computer Science, 2024, 243: 465-471.