



OPEN Quantifying momentum and influencing factors of tennis players using the XGBoost model

Donghong Wang^{1,2✉}, Xu Zhang², Yuneng Xie³ & Qinyan Zhu³

Momentum can directly or indirectly affect a tennis player's mentality and the trajectory of the game, thereby changing the outcome of the match. The article provides a clear quantitative description of the concept of momentum in tennis, analyzing the impact of momentum fluctuations on the trajectory of the game and the athlete's scoring; a secondary indicator system is established, and the weights of each indicator are determined through expert analysis method, CRITIC weighting method, and hierarchical analysis method. The final value of momentum is randomly tested with a random walk model. The results show that momentum is not random but influenced by specific factors, indicating that there is a certain correlation between the fluctuation of momentum and the success of the player. Due to the non-normal distribution of momentum, using the XGBoost model and Shap feature importance analysis can determine the significant influence of factors such as the distance run by athletes and the speed at which they hit the tennis ball. Moreover, the model demonstrated excellent performance in five randomly selected matches, with low error indicators and high fitting degrees, and an R^2 coefficient of up to 0.9814, showing the model's high precision and strong generalization capability.

Keywords Random walk, Shap feature analysis, Tennis momentum, XGBoost algorithm

As a highly competitive sport that emphasizes both technique and strategy, the performance of tennis players is influenced by various factors. Among them, momentum directly or indirectly affects the mentality of players and the trend of the game, thereby changing the outcome of the game, and is one of the important factors in the game situation and final result. In tennis matches, momentum not only includes psychological momentum¹, which is the psychological momentum generated by athletes due to continuous scoring or advantages in the game, but also includes physical momentum², which is the continuity of athletes' movement state and energy maintenance. Understanding and quantifying these two types of momentum is crucial for optimizing training and competition strategies.

Upon reviewing existing research, it can be found that competitions such as patent competitions, sports competitions, and lifestyle competitions typically rely on multiple small-scale "battles" as the final outcome³, which are commonly referred to as dynamic competitions⁴. In the literature, many scholars have investigated dynamic competitions, particularly addressing the question of whether momentum affects the results of such competitions⁵.

The research on momentum quantification technology and the application of psychological momentum in sports competitions has always been a focus in the field of sports psychology. Researchers mainly explore how psychological momentum affects athletes' performance and results in competitions. Early research has shown that psychological momentum can significantly affect the situation of a game⁶. Iso Ahola and Mobily (1980) proposed the concept of psychological momentum, stating that athletes develop a psychological advantage or disadvantage after consecutive successes or failures, which affects their subsequent performance. Richardson et al. (1988) validated the impact of psychological momentum on competitive performance through empirical research⁷. However, quantifying the impact of psychological momentum has always been a challenge due to its subjectivity and abstraction. Markman and Guenther (2007) proposed a simple quantification method based on continuous scoring and winning streak, but this method is too crude and fails to capture the complex characteristics of momentum⁸. To overcome this problem, researchers have begun to attempt to combine psychological momentum with physical momentum, using mathematical models and data analysis methods to quantify momentum. With the development of data science and machine learning technologies, more and

¹School of Public Administration, Zhongnan University of Economics and Law, Wuhan, China. ²School of Statistics and Mathematics, Hubei University of Economics, Wuhan, China. ³School of International Education, Hubei University of Economics, Wuhan, China. ✉email: wdhong@hbue.edu.cn

more research is adopting advanced analytical methods to quantify momentum. For example, Sarlis et al. (2020) conducted a study on basketball games⁹, using data science and machine learning to quantitatively analyze the performance of athletes and teams, and explored changes in momentum during the game. In addition, Sarlis et al. (2021) further investigated the impact of injuries on athlete performance and applied multivariate statistical models to explore the relationship between injuries and momentum changes¹⁰. Although these studies provide valuable references for momentum quantification, most still focus on single dimensional analysis.

In the field of tennis, Klaassen and Magnus (2001) analyzed the scoring patterns of tennis matches using a dynamic binary panel data model and found that scores were not independent and identically distributed, suggesting the existence of momentum effects¹¹. Their research provides a statistical basis for the quantification of momentum. Den Hartigh et al. (2014) used a dynamic system model to simulate momentum changes during a competition, considering the interaction between psychological and behavioral factors¹². Their model can capture the dynamic evolution of momentum during the game, but there are limitations when dealing with high-dimensional data and nonlinear relationships.

Regarding the application research of different machine learning algorithms in momentum analysis, with the development of machine learning technology, researchers have begun to apply advanced algorithms to quantify and predict momentum. Machine learning models are capable of processing large amounts of complex data, capturing non-linear and high-dimensional feature relationships, and are suitable for analyzing motion data¹³. XGBoost, as an efficient gradient boosting tree model, has attracted attention due to its fast speed and excellent performance¹⁴. The XGBoost model developed by Chen and Guestrin (2016) improves the performance and computational efficiency of the model by introducing regularization and parallel computing¹⁴.

In the momentum analysis of tennis matches, Gönültaş et al. (2020) applied XGBoost to predict match results with high accuracy, demonstrating the advantages of this model in capturing game dynamics and momentum changes¹⁵. Zhang and Jones (2019) used machine learning techniques to predict the results of tennis matches and identified key factors that affect the outcome of the game¹⁶.

In order to gain a deeper understanding of the mechanism of momentum formation, researchers use methods such as feature importance analysis to identify key influencing factors. SHAP (Shapley Additive exPlanations) values, as a method of explaining machine learning models, can quantify the contribution of each feature to the model output¹⁷. Lundberg and Lee (2017) proposed that SHAP values have consistency and essence, and are applicable to the interpretation of various models¹⁷. Li et al. (2020) applied the XGBoost model combined with SHAP value analysis in basketball games to identify key technical indicators that affect game results, providing new ideas for quantifying momentum¹⁸. However, there is still a lack of research on the combination of XGBoost model and SHAP value in tennis momentum analysis.

In summary, although some progress has been made in momentum quantification and model applications, there are still some shortcomings. Firstly, most studies only focus on a single aspect of psychological momentum or physical momentum, failing to fully integrate the two. Secondly, in terms of model selection, there is a lack of validation for model adaptability and generalization ability, and the stability of the model's performance on different datasets remains to be examined¹⁹. Therefore, in future research, especially in tennis, which is a long-term and high-pressure competition, it is necessary to explore in more detail the causes, performance, and long-term and short-term effects of momentum. In order to more accurately understand how momentum affects the game, research should comprehensively consider strategic (physical) momentum and psychological momentum, and develop a comprehensive and reasonable evaluation method to concretize the impact of momentum and make up for the shortcomings in previous studies.

Based on the analysis of the aforementioned literature, this study aims to create a novel definition of momentum that combines strategic (physical) momentum with psychological momentum. A new comprehensive momentum quantification method is proposed, using the XGBoost model and SHAP feature importance analysis to further explore how momentum is influenced by multiple influencing factors such as running distance, hitting speed, scoring patterns, etc. This innovative method provides a more comprehensive perspective for a deeper understanding of momentum fluctuations and precise support for optimizing competition strategies and athlete training in practical applications. Specifically, the main research questions of this study include: (1) Defining or quantifying momentum in tennis, which will comprehensively consider the physical, psychological, and skill states of athletes, making the concept of momentum more specific and objective. (2) Verify the validity of the new definition of momentum and explore the non randomness of momentum changes to confirm the existence of specific factors that affect momentum changes. (3) Using the XGBoost model combined with SHAP feature importance analysis, summarize and generalize the key factors that affect athletes' momentum changes. In addition, this study will use a wide range of datasets to validate the adaptability and generalizability of the proposed model, in order to ensure the practical application value of the research results. Through these research findings, coaches and athletes will be able to use momentum quantification results to optimize competition strategies, improve athlete performance, and achieve more effective competition preparation and training planning.

Data collection and processing

The data utilized in this study comes from the game data we collected from the game websites and some sports organizations, combined with the web crawler technology to crawl the relevant website data to collect a total of 7,284 pieces of data, covering 31 games. In the dataset, we found that there are missing values in the features of speed_mph, serve_width, serve_depth and return_depth. To ensure data integrity and usability, we will clean the raw data.

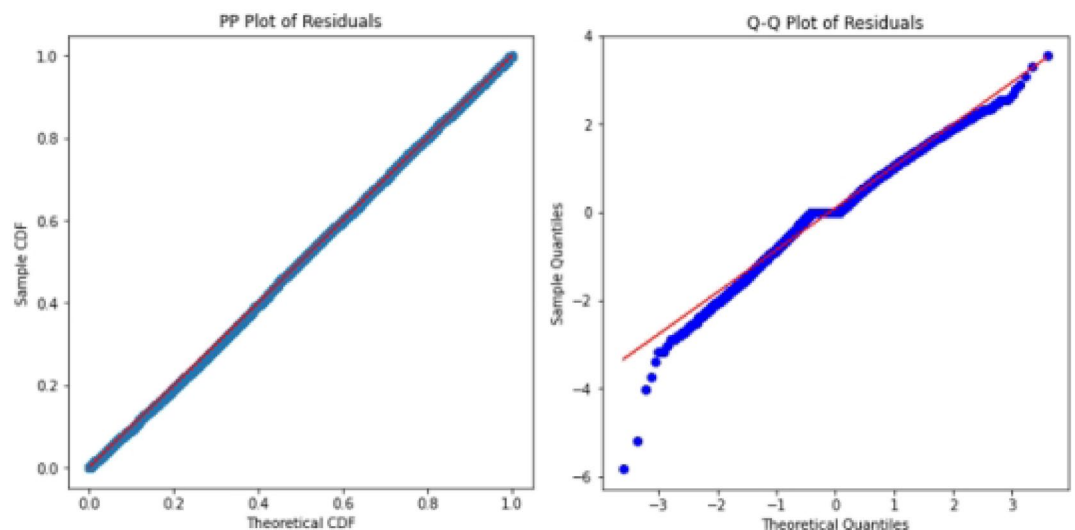


Fig. 1. Q-Q plot of normality test.

Variant	CV-value
p1_distance_run	0.358668
p2_distance_run	0.352120
rally_count	0.079587
speed_mph	0.304736

Table 1. Table of outlier test coefficients.

Missing value handling

For the processing of categorical data, this study will use mode filling for features with fewer missing values. For data with a large number of missing values, 'U' will be used to represent unknown values. Treating missing values as an independent category can maximize the preservation of data authenticity. To ensure the rationality of this filling method and its impact on subsequent analysis, we will conduct further validation in the subsequent model selection and performance evaluation stages. This study will compare the impact of different interpolation methods on prediction performance, such as KNN interpolation, mode interpolation, etc. Through these comparisons, we can further evaluate whether the "U" category imputation method has a significant impact on model performance and make appropriate adjustments based on the comparison results.

For the processing of quantitative data, this study mainly focuses on the handling of missing values in the column of speed (speed_cph). We face a crucial decision regarding this: how to effectively handle it? Considering the characteristics of the data and subsequent analysis, we choose to use random forest regression to fill in missing values. Tennis match data typically contains nonlinear relationships and complex interaction effects between multiple variables. As an ensemble learning method, random forest regression can capture these complex patterns by constructing multiple decision trees, which cannot be achieved by traditional methods such as linear interpolation or mean imputation. In practical applications, we also compared random forest regression with other common missing value handling methods such as mean imputation, median imputation, nearest neighbor interpolation, etc. The performance of random forest regression in cross validation is superior to these traditional methods, which further confirms its rationality as a missing value imputation method in this study. To further validate the effectiveness of the random forest regression filling method, we conducted normality tests and outlier detection on the filled data (as shown in Fig. 1). The results indicate that the filled dataset maintains statistical consistency with the original dataset, suggesting that random forest regression does not introduce any additional bias.

Outlier detection and handling

In this study, we first determine whether there are outliers through logical validation, for example, if player 1 takes points, the corresponding player 2 is not counted, etc., for which we can preliminarily conclude that the data are logical in accordance with the rules of the game. Subsequently, we calculated the coefficient of variation (CV value) of quantitative data in the data as shown in Table 1 below:

By observing the above values of coefficient of variation (CV), we find that the CV values of all the quantitative data except qualitative data are less than 0.5. Based on this observation, it is reasonable to assume that there are fewer outliers, or no outliers present in these characteristics. Therefore, it can be argued that the quality of the data in this study is better.

Construction and analysis of momentum quantification system

Evaluation system construction

This study firstly realizes the quantitative definition of momentum. Combined with the athlete data and provide a series of evaluation methods to achieve the quantification of momentum, the use of hierarchical analysis²⁰, CRITIC weighting method^{21,22}, from the professional skills, scoring performance, physical condition and psychological quality of the four aspects of the first level of the indicators for evaluation and analysis, which then combined with the characteristics of the data to determine each of the first level system under the second level of the system indicator selection. The evaluation system used in this study is shown in Fig. 2 below:

For the above evaluation quantitative system, this study utilizes the CRITIC weight method to assign the second-level weights to obtain the weights of the corresponding athletes in each game as $W_{i,j}$, where i represents the number of athletes; j represents the corresponding second-level index, and combines the hierarchical analysis method and expert evaluation method to determine the weights of the first-level index system as W_a , where a represents the number of corresponding first-level indexes. In this regard, the formula for calculating the athletes' momentum in each race is as follows:

$$S_p^b = M_0 + \sum_{a=1}^n \left(\sum_{j=1}^m W_{b,j}^b * X_{b,j}^b * W_a \right)$$

In order to facilitate the intuitive embodiment of the floating change of momentum, this study assumes that the initial momentum of each tennis player is M_0 , whose value is equal to 100, combined with the score of the player in the game and a variety of data situations will produce a corresponding increase in momentum, so as to realize the quantification of momentum.

where the above symbols are interpreted as shown in Table 2 below:

Quantitative momentum results and visualization

In order to explore the intuitive changes in momentum, this study screened the whole process of one of the most classic matches (Carlos Alcaraz vs. Novak Djokovic), and the corresponding indicator weights can be obtained through the above indicator system and the selected data as shown in Table 3 below:

Relying on the above system can be realized to quantify the momentum, in order to more intuitive representation, the whole game momentum trend change graph can be drawn as shown in Fig. 3 below:

We can visualize that as the score changes, the momentum changes regularly: as one athlete's momentum increases, the other athlete's momentum decreases. This confirms what happens more often than not in competitions: an opponent's score creates an invisible pressure on the athlete, which can weaken his or her performance. From the basic logic, it can be concluded that the momentum quantification is referable, and this study continues to carry out the randomness test of momentum to further verify the rationality of momentum quantification.

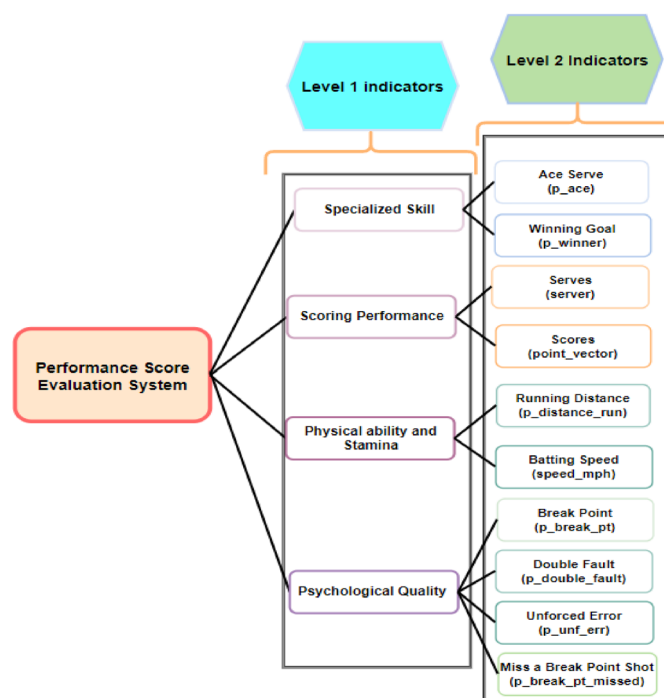


Fig. 2. Structure of the evaluation system.

Notation	Hidden meaning	Unit
b	Number of athletes per race	seat
n	Number of Tier 1 indicators	size
m	Number of secondary indicators under each a-level indicator(a akes the value[0, n])	size
M ₀	Initial momentum value	–
W _a	Tier 1 indicator weights	–
W _{b,j} ^b	Weights of secondary indicators for athlete b	–
X _{b,j} ^b	Competition value for athlete b	–

Table 2. Symbol meanings.

Level 1 indicators	Weights	Secondary indicators	Weights	
			Player 1	Player 2
Specialized skill	0.1584	Ace serve (p_ace)	0.2138	0.2135
		Winning goal (p_winner)	0.7862	0.7865
Scoring performance	0.2573	Serves (server)	0.4967	0.4857
		Scores (point_vector)	0.5033	0.5143
Physical ability and stamina	0.2647	Running distance (p_distance_run)	0.4585	0.4059
		Batting speed (speed_mph)	0.5415	0.5915
Psychological quality	0.3196	Break point (p_break_pt)	0.0792	0.0562
		Double fault (p_double_fault)	0.6228	0.6028
		Unforced error (p_unf_err)	0.1325	0.1169
		Miss a break point shot (p_break_pt_missed)	0.1655	0.2241

Table 3. Indicator system weighting values.

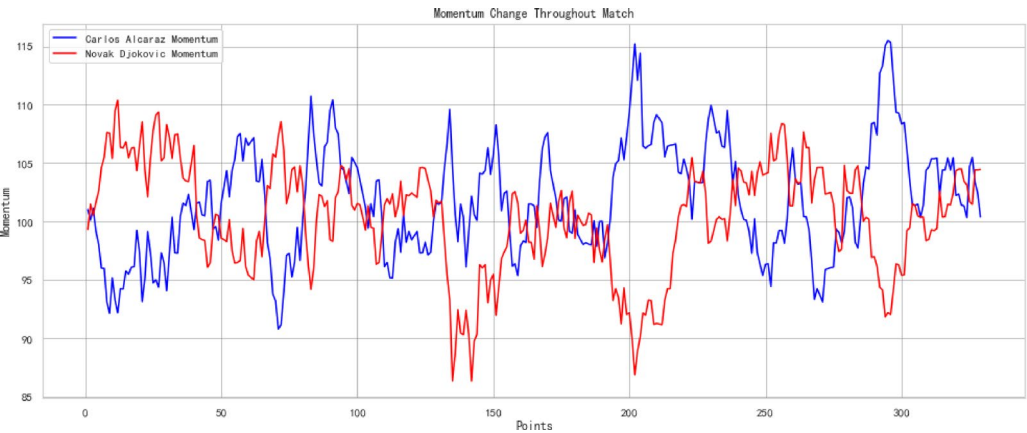


Fig. 3. Momentum trend change chart.

Non-randomness test for momentum
Introduction to the stochastic tour test model

The Run Model is a statistical model used to analyze patterns in sequential data, especially useful when studying random or non-random patterns. The run model evaluates the randomness of a sequence by identifying “runs” of consecutive identical elements in the data sequence. A “run” is the set of identical elements that occur consecutively in a sequence.

By performing the run test on the momentum feature, it is possible to assess whether its variations are stochastic or not, and to verify the feasibility and correctness of the momentum quantization²³.

The main steps in testing the stochastic tour model constructed in this study are as follows:

Step1: Counted Trips: records the total number of trips observed in a sequence of momentum features. This includes consecutively positive and consecutively negative trips.

Step2: Calculating the expected number of trips and variance: the expected number of trips and variance are calculated based on the length of the binary sequence and the distribution of positive and negative markers. The expected number of trips and variance can be calculated as follows:

Variant	Z-value	P-value
Player 1 Momentum	− 9.455	0.000***
Player 2 Momentum	− 11.55	0.000***
Winning player of each game	15.452	0.000***

Table 4. Randomness test coefficients per game.

Variant	Z-value	P-value
Player 1 Momentum	− 4.467	0.041*
Player 2 Momentum	− 3.59	0.023
Winning player of each set	− 2.058	0.040**

Table 5. Randomness test coefficients per set.

$$\mu_R = \frac{2n_1n_2}{N} + 1$$

$$\sigma_k^2 = \frac{2n_1n_2(2n_1n_2 - N)}{N^2(N - 1)}$$

where the number of positive markers is n_1 , the number of negative markers is n_2 , the total number of observations is $N = n_1 + n_2$, and the number of observed excursions is R .

Step3: Perform a hypothesis test: use the expected number and variance of trips to compute the statistic Z , which in turn determines whether the distribution of trips in the sequence meets the expectations of randomness. If the sequence is completely random, the value of R should be close to μ_R . where the statistic Z is calculated by the formula:

$$z = \frac{R - \mu_R}{\sqrt{\sigma_R^2}}$$

Step4: Judging the results: based on the Z - value and the selected significance level ($\alpha = 0.05$) determine whether the observed number of trips as R significantly deviates from the expected value of the random distribution μ_R . If the Z - value falls outside the critical value of the normal distribution, we reject the original hypothesis.

Based on the above modeling steps, we can analyze the momentum of the two athletes in each race to determine whether the change in momentum is random or not.

Changes in momentum are non-random in nature

A tennis match will have a number of sets, and each set will be played for a number of games, for which this study focuses on verifying the non-random nature of momentum from two perspectives: per set and per game.

1. Momentum test analysis at the end of each game.

Firstly, we screened to get the momentum state of each athlete at the end of each set at this moment, and then we calculated the statistic through the model for swim test to get the result as shown in Table 4 below:

Based on the results of the table above, we can find that the significance p-value is much less than 0.05, which presents significance at the level, and therefore has not been non-randomized data, so it is reasonable to suggest that the change in this momentum is related to some factors.

2. Momentum test analysis at the end of each disk.

Combined with the analysis of the ideas for each game, we can first filter out the momentum states of both athletes at the end of each set, and then calculate the results obtained by our swim test model as shown in Table 5 below:

Based on the results of the table above, we may find that there is also a non-randomness accompanying the change in the kinetic energy of the players at the end of each set, and similarly we can show that there are certain factors that can affect the change in momentum, and that the change in the athlete's momentum affects the performance of the match to a certain extent.

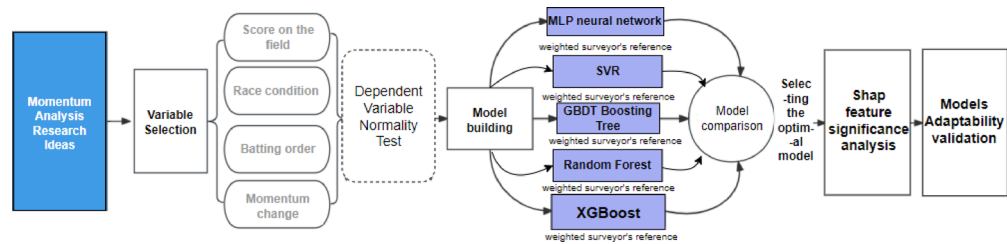


Fig. 4. Momentum Influencing Factors Research Idea Map.

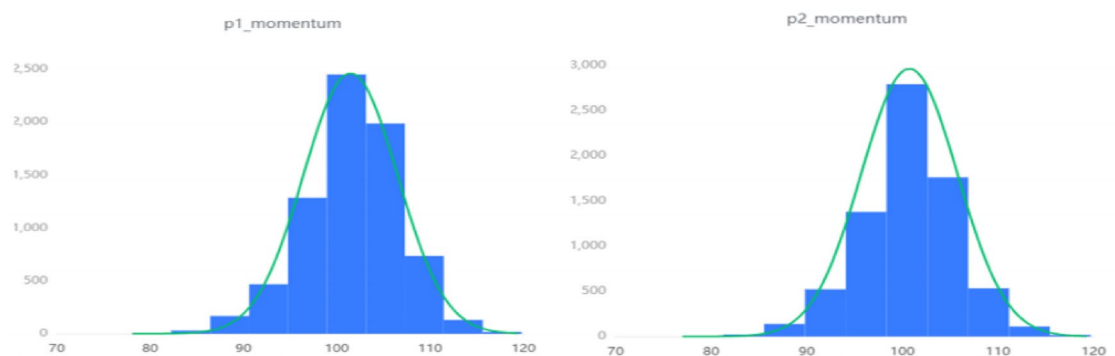


Fig. 5. Normality test chart.

Research on momentum influencing factors based on XGBoost model

Momentum influencing factors research ideas

The research line followed in this study to explore the study of momentum influencing factors is shown in Fig. 4 below:

In the selection of the independent variable, due to the quantitative properties of momentum considering the establishment of a regression model to explore the impact of momentum change factors, in this regard, we test the normality of the momentum distribution of the two athletes, you can draw a normal histogram as follows Fig. 5:

From the above normality test graph, we can find that the feature of momentum is not fully consistent with the absolute normal distribution, then we can consider using a nonlinear regression model. Subsequently combined with the data selected three models: MLP neural network, GBDT boosting tree and XGBoost for analysis, select the model optimal XGBoost model for regression analysis, followed by the use of Shap feature importance analysis intuitive analysis of factors affecting momentum, and finally the trained model nested in the other multiple matches to validate the adaptability of the model.

Model building

Taking into account the application of various machine learning algorithm models in regression analysis, we mainly used five algorithm models in the process of solving this problem:

The MLP neural network algorithm²⁴ model is a feedforward artificial neural network model that belongs to the category of deep learning models, especially performing well in dealing with complex nonlinear problems.

GBDT (Gradient Boosting Decision Tree) is a popular and powerful machine learning algorithm²⁵, belonging to the Boosting family of ensemble learning methods. It is mainly used for regression and classification problems, and improves prediction accuracy by combining multiple decision trees.

SVR (Support Vector Regression) is a supervised learning algorithm based on Support Vector Machines (SVM)²⁶, used to solve regression problems. SVR constructs an optimal hyperplane in a high-dimensional feature space to minimize the error between predicted and actual values while controlling the complexity of the model, making it particularly suitable for regression analysis of high-dimensional and nonlinear data²⁷.

Random Forest Regression is an ensemble learning method²⁸ that improves the accuracy and stability of predictions by constructing multiple decision trees and averaging their results. It can effectively handle high-dimensional data, reduce model variance, and prevent overfitting^{28,29}.

XGBoost (eXtreme Gradient Boosting) is an improved algorithm based on GBDT³⁰. It is widely used in fields such as spam detection, advertising placement, sales forecasting, and medical data analysis³¹.

This study first selected ten sets of data as our dataset and split it into training and testing sets. Then, we conducted regression analysis using the five model distributions mentioned above. Evaluate each model to obtain the absolute percentage error, average absolute error, r2 coefficient, confidence interval of test set error, cross validation R2, and normality test of residuals. The results are shown in Table 6 below:

Regression models	Train-MAPE	Test-MAPE	MAE	r2-score	95% confidence interval (MAPE) of the test set	95% confidence interval (MAE) of the test set	Cross validation R2	Cross validation MAE	Residual normality P-value
Multi-Layer Perceptron	25.24	29.85	1.85	- 0.04	[28.06, 30.32]	[1.77,2.01]	0.12	1.88	0.42
GBDT	41.56	42.59	1.62	0.24	[40.35, 43.22]	[1.53,1.72]	0.15	1.95	0.78
SVR	19.43	27.39	0.95	0.87	[17.91,29.12]	[0.93,0.97]	0.78	1.67	0.03
Random Forest	17.89	24.58	0.71	0.85	[13.55,32.49]	[0.67,0.75]	0.81	1.49	0.67
XGBoost	1.14	0.69	0.40	0.94	[0.67,0.79]	[0.37,0.43]	0.91	0.39	0.85

Table 6. Error coefficients for multi-model comparison.

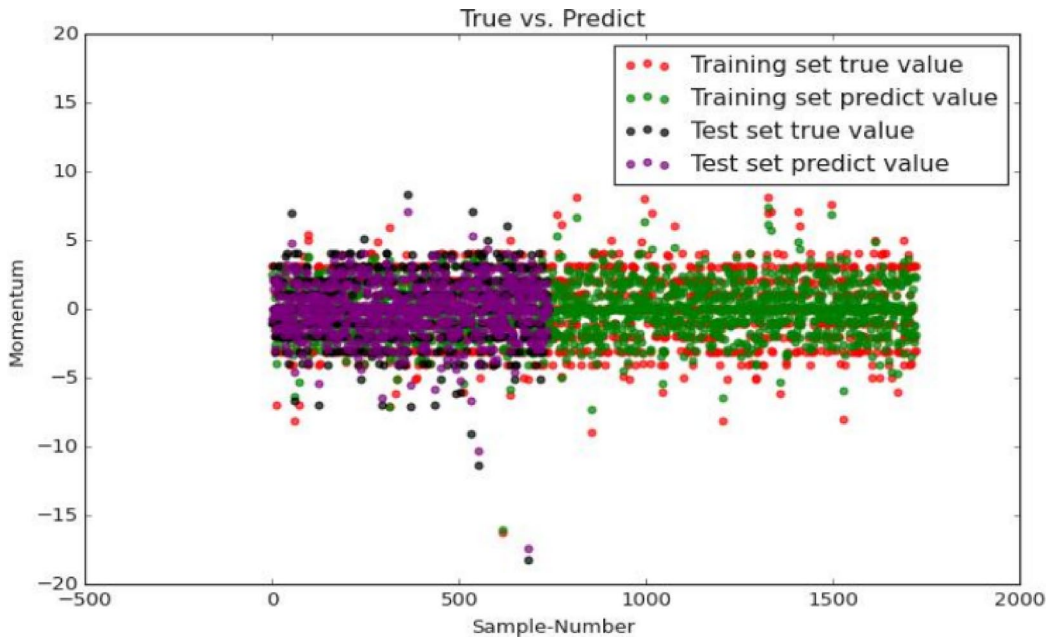


Fig. 6. Momentum scatterplot.

From the above analysis, it can be seen that the residuals generally follow a normal distribution (P value greater than the significance level), and the absolute percentage error and mean absolute error are both within the confidence interval, indicating that the model’s predicted error is within an acceptable range. Based on more statistical indicators, especially the evaluation of different interpolation methods through cross validation, we chose the XGBoost model with the best performance. The reason for choosing XGBoost is not only based on its performance indicators, but also on its advantages in handling complex nonlinear relationships, feature interactions, and large-scale datasets. The gradient boosting tree (GBDT) mechanism of XGBoost makes it particularly outstanding in predicting dynamic game data, effectively capturing complex patterns of momentum changes during the game.

To further validate the predictive ability of the model, we plotted a scatter plot (as shown in Fig. 6). From the figure, it can be seen that the XGBoost model can fit the data well, and the relationship between predicted values and true values shows a strong linear trend, indicating that the model has good regression performance. To improve the readability of the scatter plot, especially to help non-technical readers better understand the information in the chart, we have added clear labels and annotations to the scatter plot, and provided intuitive descriptions of the point correspondence between predicted and actual values. These annotations can help readers quickly grasp the core information of the chart without having to delve into the complex mathematical background.

Based on the exploration of missing value processing in this study, we evaluated various missing value processing methods using the xgboost model, and obtained the model evaluation coefficients as shown in Table 7:

According to the cross validation results, the KNN interpolation method performs the best, with a mean square error (MSE) of 0.23 and a coefficient of determination (R^2) of 0.88, indicating that this method can provide the most accurate prediction results and better explain the variability of the data. This indicates that KNN interpolation can more accurately capture nonlinear features in data, thereby improving the predictive performance of the model. In contrast, the performance of U-interpolation and mean interpolation is poor, with MSE of 0.26 and 0.29, and R^2 of 0.86 and 0.84, respectively. The prediction error is large, and the model fit is low,

Interpolation method	MSE (CV)	R ² (CV)
U interpolation	0.25987	0.86392
KNN interpolation	0.23095	0.88398
Mode interpolation	0.28712	0.84087

Table 7. Model evaluation coefficients for different interpolation methods.

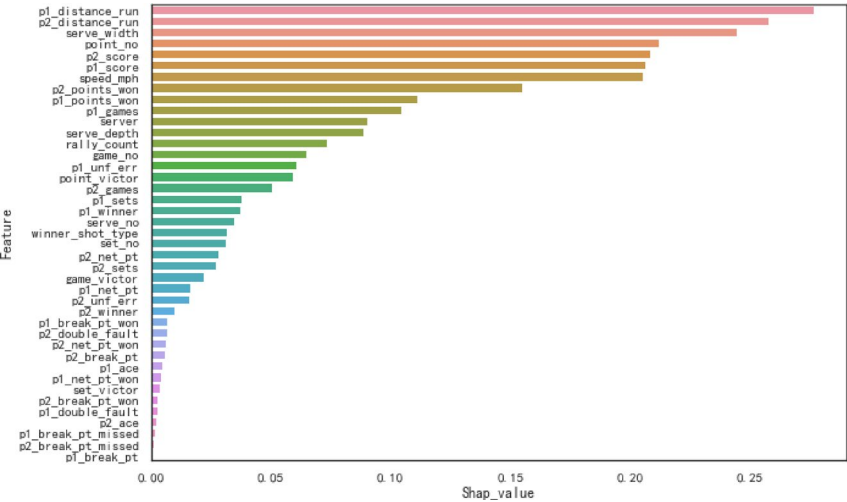


Fig. 7. Shap feature importance bar chart.

Feature	Shap_values
p1_distance_run	0.276618
p1_distance_run	0.257590
serve_width	0.244283
point_no	0.211779
p2_score	0.208419
p1_score	0.206171
speed_mph	0.205255
p2_points_won	0.154821
p1_points_won	0.111370
p1_games	0.104675

Table 8. Values for importance of characteristics.

indicating that these methods have certain limitations when dealing with complex data. Therefore, in this study, using KNN interpolation method will be more helpful in improving the accuracy and generalization ability of the model.

Analysis of influential factors based on the importance of Shap features

Based on the above research results, we can find that the XGBoost model has the best fitting ability. Next, we will analyze the influencing factors. Traditional feature importance calculation algorithms have certain limitations³². Therefore, this study uses Shap feature importance, which represents the importance of each feature by taking the average absolute value of its SHAP value. For ease of presentation, we present the importance of each feature by drawing a horizontal bar graph of the proportion of each feature (as shown in Fig. 7). It can be clearly seen from the graph that the key factors affecting momentum include the athlete’s running distance, hitting speed, scoring, scoring frequency, and serving direction.

It can be clearly seen from the graph that the key factors affecting momentum include the athlete’s running distance, hitting speed, scoring, scoring frequency, and serving direction. To further deepen our understanding of these factors, we have listed the top ranked features and their corresponding SHAP values in the table below, and specifically analyzed their contributions to momentum changes. We can analyze this in detail and obtain the SHAP values of the top ranked features as shown in Table 8:

Feature	sum_sq	mean_sq	F	PR(> F)
distance_run	1.456	2.08	10.53	0.0150
point_no	0.387	0.322	7.12	0.0293
speed_mph	0.287	0.302	6.89	0.0458
Residual	13.472	0.547		

Table 9. Covariance analysis statistical table.

Feature	Chi2	P-value	df	Cramer's V
serve_width	11.8773	0.0356	12	0.350
p1_score	37.8584	0.0447	39	0.431

Table 10. Chi square test analysis statistical value table.

Feature combination	SHAP interaction contribution
p1_distance_run & p2_distance_run	0.400
p1_distance_run & p1_score	0.350
speed_mph & serve_width	0.310
p2_score & point_no	0.290
p2_points_won & p1_points_won	0.270

Table 11. Composite contribution rate values.

From Table 8, it can be seen that p1-distance run and p2-distance run (athlete’s running distance) contribute the most to the impact on momentum, indicating that these two characteristics play a decisive role in the change of momentum. Following closely behind is the serve width, which reflects the significant impact of the direction of the serve on the change in momentum. Other features, such as point_no (game set points), p1_store, and p2_score, although slightly less important, still have an undeniable impact on momentum changes.

Based on the above analysis, we can delve into the causal relationship between high influencing factors and the final dependent variable momentum. We will conduct covariance analysis on quantitative data and divide momentum into four levels for categorical data, and perform chi square test analysis. The results are shown in Tables 8 9, and 10 below:

Through the dual exploration of covariance analysis and chi square test, we have revealed the complex correlation between player momentum and key features. Among quantitative features, running distance has the most significant impact on momentum, reflecting how physical exertion shapes a player’s psychological state, followed by scoring and hitting speed in each game. This indicates that scoring accumulation and offensive initiative play an important role in improving momentum. In the categorical features, the moderate correlation between serving width and momentum indicates that the choice of game strategy has a certain impact on the establishment of psychological advantage, while the player’s score reflects a stronger force, especially in the critical score stage where the momentum driving effect is particularly evident.

However, in addition to the influence of individual features, this study also needs to further explore the joint effects and interactions of multiple features on the contribution of momentum changes. For example, the combination of running distance and scoring patterns may have a greater impact on changes in momentum than running distance alone. This study will further validate these interaction effects through feature importance analysis of the XGBoost model, and combine them with the SHAP value combination contribution coefficient table (as shown in Table 11 below) to clarify which feature combinations have a greater impact on momentum changes, in order to provide more accurate strategic recommendations for coaches and athletes.

Model adaptation analysis

After our verification, the model is not only adapted to this game, but also has a better performance ability in many other games, in this regard, we will be trained XGBoost model for grid search tuning to analyze the data of five other randomly selected games, and draw the momentum distribution scatterplot results are shown in Fig. 8 below:

Subsequently, we calculated the MAPE value, MAE value, r2 score, and related confidence intervals of the model as shown in Table 12 below:

Combining the above visualization and error coefficients, the R2 coefficient is close to 1, indicating that the model fits well, and the mean absolute errors are small. In this regard, it is not difficult to find that the model also has a good prediction ability for other matches, which also proves the adaptability of the model.

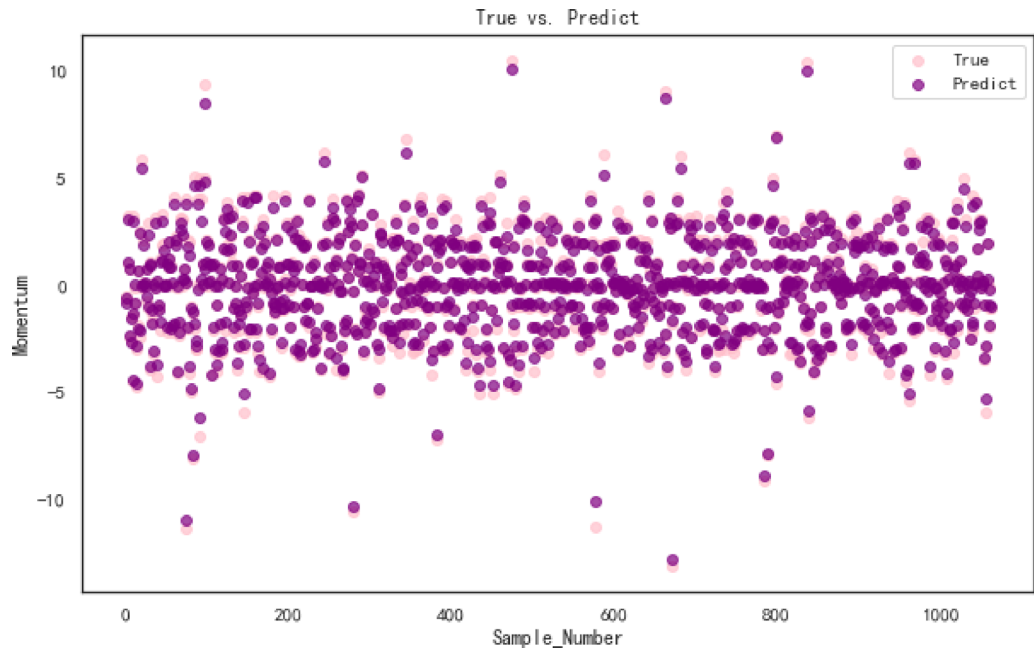


Fig. 8. Momentum prediction and scatterplot of true distribution.

Evaluate	Value
MAPE	0.00697
MAPE 95% confidence interval	[0.005, 0.009]
MAE	0.163534
MAE 95% confidence interval	[0.154, 0.233]
R2_score	0.981495

Table 12. Model error coefficients.

Summary and conclusions
Research conclusion

This study first conducted a detailed cleaning of the collected tennis match data, including using mode, KNN filling, random forest regression filling methods to handle missing values, and using CV value testing method to detect abnormal data. On this basis, this study successfully quantified the momentum in tennis matches and tested the non randomness of momentum changes from two levels, each set and each game, using a random run model. The P-values were significantly less than 0.05, confirming that momentum changes are not random events but are influenced by certain specific factors. In addition, this result also validates the rationality of the momentum quantification method in this study.

To further explore the specific factors affecting the momentum change, the study used XGBoost model combined with Shapley feature importance analysis. Through the visualization of feature importance, we identified seven key factors that have the most significant influence on the momentum change: (1) running distance: reflecting the athlete's physical fitness and speed, the longer the distance indicates better physical fitness, and the possibility of acquiring momentum is higher; (2) hitting speed: the faster the speed, the advantage in controlling the tempo and posture, and it is easier to acquire momentum; (3) individual scoring: the higher the scoring, the greater the possibility of continuous scoring, and the accumulation of momentum is accelerated; (4) total scoring: the existing scoring, the greater the possibility of accumulating momentum; and (5) the total scoring. (3) Individual score: the higher the score, the higher the likelihood of successive scores, accelerating the accumulation of momentum; (4) total score: scoring on the basis of the existing score difference, to obtain a decisive lead, can quickly form momentum; (5) changes in the width of the ball: tactical use of the better, more likely to create the opponents' mistakes, to gain momentum; (6) cumulative score: consecutive bowling bowling, which can be pushed up the individual and the accumulation of scores, to enhance the momentum; (7) the number of sets won: the number of sets reflects the momentum of winning, winning consecutively helps to gain the momentum; (8) the number of sets won. Winning momentum, winning streak is conducive to obtaining good momentum.

In order to verify the adaptability and generalizability of the model, the study also utilized data from multiple matches for nested model training and calculated the model R2 coefficient as high as 0.98, which indicated that

the model also showed good expressive ability in multiple other matches. This result further confirms the wide applicability and validity of the model in this study.

Exploration of practical applications

1. From the coach's perspective, the momentum quantification model in this study provides a scientific basis for developing competition strategies and training plans. On the one hand, coaches can use this model to monitor the momentum changes of athletes in real time during the game, combined with the changing trends of key features such as running distance and hitting speed, to accurately evaluate the physical exertion and psychological state of athletes. For example, when the running distance exceeds a certain threshold, the coach can adjust tactics appropriately to reduce the player's physical exertion or switch defensive strategies, avoiding momentum decline caused by excessive fatigue. On the other hand, the model results indicate a significant correlation between serving direction and player momentum, suggesting that coaches can enhance the psychological advantage of players by designing diverse serving strategies, such as using different serving widths and depths to disrupt the opponent's rhythm during critical score stages. In addition, the hitting speed during the game has been verified as an important factor affecting momentum, and coaches can help players more actively control the situation by improving their hitting power and rhythm control ability.
2. From the perspective of athletes, the momentum quantification model in this study provides a more intuitive basis for game feedback and self adjustment. Firstly, the significant relationship between running distance and momentum emphasizes the importance of physical fitness management. Athletes can control their running intensity reasonably during the competition based on the model prompts, avoiding the accumulation of psychological momentum caused by physical exhaustion. At the same time, the impact of hitting speed on momentum indicates that improving technical ability and controlling the pace of the game are key to enhancing momentum. Athletes can increase their hitting power and speed, enhance their offensive initiative in the game, and gain more momentum advantages. Secondly, the analysis of serving direction and player scoring patterns suggests that athletes can use serving strategies more flexibly in the game. For example, in the process of confrontation, changing the serving direction or adjusting the serving width can increase the opponent's difficulty in receiving the serve, thereby enhancing one's own psychological momentum. Thirdly, research on scoring patterns suggests that athletes should pay special attention to the competition for key scores during matches. For example, at critical moments such as 40–40 or break points, positive psychological suggestion and high-intensity focus can increase the success rate of scoring and steadily accumulate momentum.

Future research directions

This study reveals the quantification and influencing factors of momentum in tennis matches from both theoretical and practical perspectives, but there is still much research space worth further exploration.

Firstly, in the future, momentum quantification models can be applied to real-time data analysis in competitions, by combining sensor technology to achieve real-time feedback on player momentum, helping coaches develop precise tactical adjustment strategies at critical moments. The development of this real-time feedback system not only provides real-time decision support for competitions, but also further validates the applicability and robustness of the model.

Secondly, the adaptability and generalization of the model make it possible to study the momentum of other sports events. For example, sports with similar momentum characteristics such as badminton, table tennis, and squash can refer to the momentum quantification approach of this study to explore the impact mechanism of momentum on different types of sports. This cross project research will provide a broader empirical basis for the universality of momentum theory, and also bring new ideas for training and competition strategy optimization in other sports.

Thirdly, integrating environmental and psychological data will greatly enrich the connotation of momentum quantification. For example, the psychological state of athletes (such as anxiety, stress) and external environmental factors (such as weather, field conditions) may be closely related to changes in momentum. Through multimodal data analysis, we can combine these factors with momentum quantification to provide a more comprehensive explanation for changes in momentum. In addition, future research can explore how to combine sensor technology and physiological monitoring data (such as heart rate, brainwaves, etc.) with momentum quantification models to further enhance the accuracy and application value of momentum quantification.

Data availability

The data presented in this study are openly available in the following websites: (1) Wuhan Municipal People's Government Internet: <http://www.wuhan.gov.cn/hdjl/>. (2) Public Data open platform of Wuhan City: <http://data.wuhan.gov.cn>. (3) General Administration of Sport of China: <https://www.sport.gov.cn/>.

Received: 5 June 2024; Accepted: 13 May 2025

Published online: 19 May 2025

References

1. Iso-Ahola, S. E. & Mobily, K. Psychological momentum: A phenomenon and empirical validation of its influence in competitive sport tournaments. *Psychol. Rep.* **46**(2), 391–401 (1980).
2. Adler, P. & Adler, P. The gloried self: the aggrandizement and the constriction of self. *Social Psychol. Q.* **52**(4), 299–310 (1989).
3. Konrad, K. A. *Strategy and Dynamics in Contests* (Oxford University Press, 2009).
4. Jia, J. & Skaperdas, S. Technology and contests. *Econ. Theor.* **51**(2), 397–416 (2012).

5. Berger, J. & Pope, D. Can losing lead to winning? *Manage. Sci.* **57**(5), 817–827 (2011).
6. Crust, L. & Nesti, M. A review of psychological momentum in sports: why qualitative research is needed. *Athletic Insight*. **8**(1), 1–15 (2006).
7. Richardson, P. A., Adler, A. B. & Hanks, D. M. Psychological momentum in competitive athletics: A preliminary test. *Percept. Mot. Skills*. **66**(2), 531–537 (1988).
8. Markman, K. D. & Guenther, C. L. Psychological momentum: intuitive physics and Naive beliefs. *Pers. Soc. Psychol. Bull.* **33**(6), 800–812 (2007).
9. Sarlis, V. & Tjortjis, C. Sports analytics - Evaluation of basketball players and team performance. *Inform. Syst.* **93**, 101562 (2020).
10. Sarlis, V. et al. A data science approach analysing the impact of injuries on basketball player and team performance. *Inform. Syst.* **99**, 101750 (2021).
11. Klaassen, F. & Magnus, J. R. Are points in tennis independent and identically distributed? Evidence from a dynamic binary panel data model. *J. Am. Stat. Assoc.* **96**(454), 500–509 (2001).
12. Den Hartigh, R. J. R., Gernigon, C., Van Yperen, N. W., Marin, L. & Van Geert, P. L. C. How psychological and behavioral team states change during positive and negative momentum. *PLOS ONE*. **9**(5), e97887. (2014).
13. Bishop, C. M. *Pattern Recognition and Machine Learning* (Springer, 2006).
14. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794 (ACM, 2016).
15. Gönültaş, B., Kıyıcı, F. B. & Koyutürk, M. Predicting outcomes of tennis matches with machine learning: serving patterns and match statistics. *IEEE Access*. **8**, 113911–113920 (2020).
16. Zhang, X. & Jones, B. Machine learning in predicting tennis match outcomes. *Int. J. Comput. Sci. Sport*. **18**(3), 112–125 (2019).
17. Lundberg, S. M. & Lee, S. I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, 4765–4774 (2017).
18. Li, Y., Ma, J. & Deng, X. Application of XGBoost algorithm in basketball outcome prediction based on player tracking data. *IEEE Access*. **8**, 22307–22313 (2020).
19. Stone, M. Cross-validatory choice and assessment of statistical predictions. *J. Roy. Stat. Soc.: Ser. B (Methodol.)*. **36**(2), 111–147 (1974).
20. Zhou, J. et al. Research on the evaluation index system of high level tennis teams in ordinary universities in China. *J. Beijing Sport Univ.* **36**(02), 105–109. <https://doi.org/10.19582/j.cnki.11-3785/g8.2013.02.020> (2013).
21. Diakoulaki, D., Mavrotas, G. & Papayannakis, L. Determining objective weights in multiple criteria problems: the CRITIC method. *Comput. Ops Res.* **22**, (1995).
22. Zhang, Y. & Wei, H. Multi attribute decision combination weighting method based on CRITIC. *Stat. Decis. Mak.* **16**, 75–77 (2012).
23. Lan, J. & Yu, W. Run test for heteroscedasticity. *J. Sun Yat Sen Univ. (Natural Sci. Edition)*. **S1**, 9–11 (2004).
24. Zhang, C., Guo, Y. & Leon, L. Overview of the development and application of artificial neural network models. *Comput. Eng. Appl.* **57**(11), 57–69 (2021).
25. Cao, Y. Improved GDBT iterative decision tree classification algorithm and its application. *Tech. Vis.* (12), 105–149 (2017).
26. Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A. & Vapnik, V. Support vector regression machines. *Adv. Neural. Inf. Process. Syst.* **9**, 155–161 (1997).
27. Smola, A. J. & Schölkopf, B. A tutorial on support vector regression. *Stat. Comput.* **14**(3), 199–222 (2004).
28. Breiman, L. Random forests. *Mach. Learn.* **45**(1), 5–32 (2001).
29. Liaw, A. & Wiener, M. Classification and regression by random forest. *R News*. **2**(3), 18–22 (2002).
30. Zhang, S., Liu, C. & Miao, K. Etc research on quantitative evaluation and analysis of sports effects based on optimizing XGBoost. *J. Chengdu Sport Univ.* **48**(06), 85–92 (2022).
31. Ke, G. et al. Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural. Inf. Process. Syst.* 3146–3154 (2017).
32. Lundberg, S. M. & Lee, S. I. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768–4777 (2017).

Acknowledgements

Project of the Innovation and Development Center for Ideological and Political Work in Higher Education Institutions of the Ministry of Education (Shanghai JiaoTong University) in 2024 (XMT-GZ-2401).

Author contributions

Donghong WANG wrote the main manuscript text, Xu ZAHNG prepared Figs. 1, 2, 3, 4, 5, 6, 7 and 8; Tables 1, 2, 3, 4, 5, 6 and 7. Xie Yuneng and Zhu Qinyan did data analysis. All authors reviewed the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to D.W.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025