A Study on Tennis Momentum Prediction Based on Random Forest Algorithm

Ying Li *, Junye Tan, Xiaokang Yang

School of Economics, Xinjiang University of Finance&Economics, Urumqi, China, 830012

* Corresponding Author Email: 19953829859@163.com

Abstract. As a strategic sport, tennis plays a key role in the study of its momentum from the training of players and the decision support of coaches. Firstly, this article obtains open source data of the 2023 Wimbledon 1701 competition, quantifies momentum through whether it is a serving party in the competition and personal skills, and determines 16 indicators. Secondly, a random forest model is built to predict the momentum changes of two players in this game from multiple aspects such as accuracy and AUC. The results show that although Player 1 has shown great momentum many times on the field, Player 2 has a greater momentum, especially within the first 30 minutes. Then, the fluctuations in performance in the competition and the randomness of success were evaluated. The KW test was introduced to test the "momentum" in the game, and the result showed that p-value=0.000<a=0.05, Ho was rejected, which means that the player's performance is positively correlated with the score. Finally, the assessment of the advantages and disadvantages of the model is made and some suggestions are given to provide feedback for coaches and players.

Keywords: Random Forest, Kruskal-Wallis H, Momentum.

1. Introduction

Tennis, as an aerobic sport, can improve one's cardiorespiratory function and endurance, exercise one's coordination and reaction ability, improve hand-eye coordination, and help people release pressure [1]. Tennis is a strategic sport that requires players to have good skills, physical strength, reaction and judgment. Players need to master the correct hitting skills, footwork and tactics in order to gain an advantage in the game. Players also need to be physically and mentally fit to cope with the pressures and challenges of the game. At the 2023 Wimbledon Gentlemen's final, unexpectedly, 20-year-old Spanish star Carlos Alcaraz defeated one of the greatest players in Grand Slam history, Novak Djokovic. In this tournament Djokovic should have won easily, but failed to do so. In this match, Djokovic should have won easily, but failed, and the dominant side had incredible swings, sometimes multi-point or even multi-game swings, which are usually attributed to "momentum"[2].

At present, scholars‘ research on tennis tactics is more extensive, Shi et al[3] quantitatively abstract the tennis ball landing point according to the regional distribution, and apply a variety of machine learning training models to construct a decision model for the optimal return landing point under multi-shot hold; Zhao et al[4] explore the predictive role of tennis players’ emotional intelligence on sports performance by testing 169 professional tennis players and using correlation and hierarchical regression analysis; Pei[5] proposed a prediction system for the best hitting point of tennis serve based on Kalman filtering to reduce the influence of signal interference and improve the prediction accuracy; Wang et al[6] took the technical and tactical characteristics of Djokovic's key points as the object of study, and used video observation and logical analysis to conduct strategy analysis.

The research of the above scholars mainly focuses on the static areas such as the ball return drop and the personal quality of the athletes, while there are few studies on the impact of the dynamic feature of the shift of the momentum of the athletes on the field on the athletes. Athletes ‘fatigue level [7], psychological quality [8] and technical level [9] all have an impact on athletes’ momentum, so analyzing momentum can provide guidance for players and coaches to make tactical shifts in the course of a match. Therefore, this paper collects the open source data of 2023 Wimbledon and extracts 18 features from it, constructs five machine learning regression models, and finally selects the Random Forest model for the prediction of match momentum through comparison.

2. Data sources

Math Works, Inc. (no specific author) (2024) Question C of the 2024 American Championship The datasets downloaded from the official Math Works website.

2.1. Modeling of Athlete Performance

The above open-source data was processed using missing and outlier values, and finally standardized. This article extracts 16 indicators from open-source data to quantify athletes' momentum performance. The following figures 1 show the selection of indicators in this article:

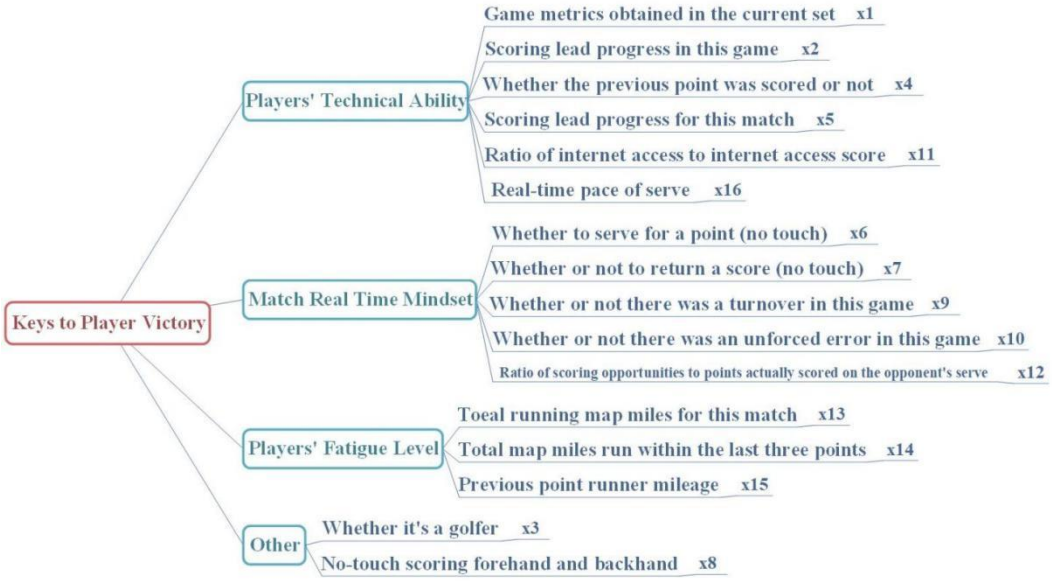


Figure 1. Quantitative indicators.

2.2. Model Selection

Various models are available within the prediction category. To select the optimal model for analyzing momentum, this study compared the five models depicted in Figure 2:

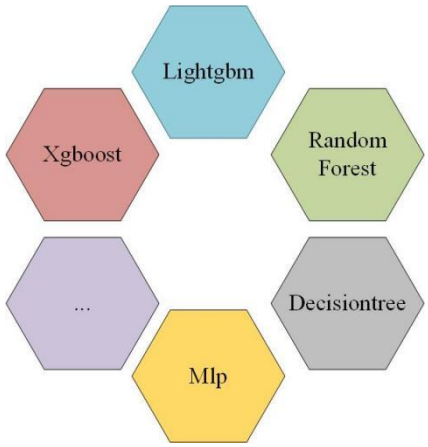


Figure 2. Model Selection.

The article evaluate the above models with the following main indicators:

1. Accuracy: an intuitive evaluation metric that indicates the proportion of all correctly predicted samples to the total samples.

2. Recall: reflects the accuracy of the model in predicting positive cases.

3. Precision: Precision rate indicates the proportion of all cases predicted as positive cases that are predicted accurately.

4. F1 Score: F1 Score is a weighted harmonic average of Precision and Recall, designed to comprehensively evaluate the performance of the model.

5. AUC: AUC is a commonly used metric for assessing the performance of classification models, indicating the degree to which the model distinguishes between positive and negative cases.

The closer the above indicators are to 1, the more accurate the prediction results are.

As shown in Table 1, by comparing the five models using Matlab, the random forest model can be used to make more accurate predictions on the data.

Table 1. Model comparison.

	Accuracy	Recall	Precision	F1	Auc
LGBM	0.68	0.68	0.69	0.69	0.76
XGB	0.67	0.69	0.68	0.68	0.74
Random Forest	0.68	0.68	0.69	0.69	0.79
Decision tree	0.61	0.62	0.63	0.62	0.61
MLP	0.67	0.68	0.69	0.68	0.75

2.3. Random Forest model

Random forest is an into machine learning method that can effectively learn to capture the complex nonlinear patterns of massive data by combining multiple decision tree combinations [10]. The random forest model demonstrates a wide range of applicability and becomes a powerful tool for solving complex problems. The risk of over fitting of a single tree is reduced by integrating multiple trees and is relatively less affected by outliers [11]. As shown in Figure 3 below:

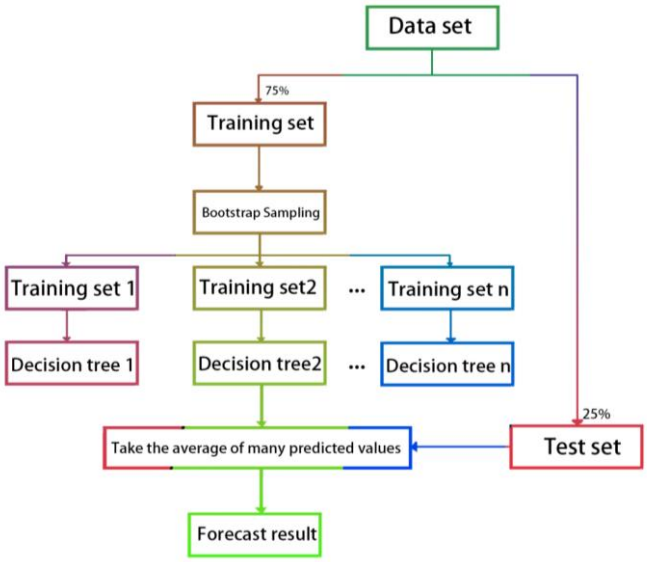


Figure 3. Random Forest Processes.

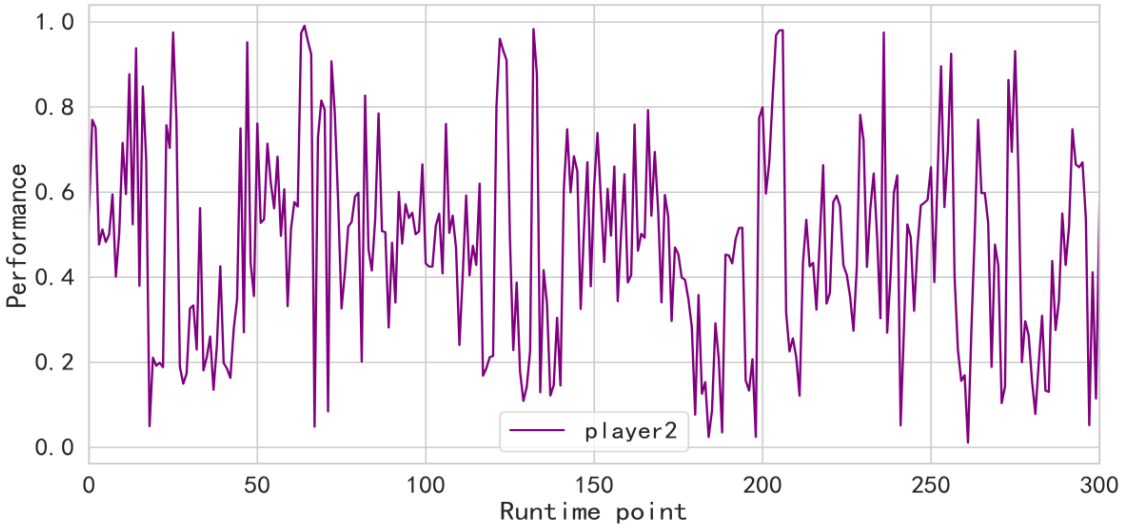
First, the article choose the standardized processed data as input, to ensure the quality of the data and model accuracy, then, the article divided the data set into training set and test set, the training set is used to train the model, the test level is used to assess the performance of the model, the training set accounts for 80% of the total datasetss, the test set accounts for the remaining 20% of the total datasetss, the use of the Random Forest algorithm for the training of the model, the completion of the training, the article use the five indicators mentioned above with the test level to assess the performance[12].

2.4. Performance evaluation of players

Through the Random Forest model, 16 relevant factors of Player one in the 2023-wimbledon-1701 game from open-source data are analyzed, and each running time point is assigned a value from 1-300. The comprehensive performance of Player one at each running time point is obtained as follows in Figure 4 (a) (b) below:



(a) Player1 performance per Runtime point.



(b) Player2 performance per Runtime point.

Figure 4. Player1&2 performance per Runtime point.

As the game progresses intensely, you can see that the real-time scores of Player one and Player two are constantly changing. The article believe that a real-time score in the interval [0.8, 1.0] indicates a better performance of the player, and the closer to 1.0 the better the player's performance. Player one and Player two have multiple real-time scores within the interval [0.8, 1.0]. In groups of six for every fifty running time points, the frequency of real-time scores within the interval [0.8, 1.0] is counted:

Take every fifty running time points as a group, a total of six groups, and count the frequency of real-time scores within the interval [0.8,1.0], as shown in Table 2 and Table 3.

Table 2. Frequency of player one real-time scoring in [0.8, 1.0].

	[0,50]	[50,100]	[100,150]	[150,200]	[200,250]	[250,300]	Total
Frequency	7	1	3	2	4	3	20
Frequency closer to 1	1	1	1	0	2	0	5

Table 3. Frequency of player two real-time scoring in [0.8, 1.0].

	[0,50]	[50,100]	[100,150]	[150,200]	[200,250]	[250,300]	Total
Frequency	5	4	2	0	2	4	17
Frequency closer to 1	3	1	2	0	2	2	10

Overall, Player one's frequency within [0.8,1.0] is greater than Player two's, but Player one's frequency of real-time scores closer to 1.0 is lower than Player two's, which suggests that Player two is performing better, especially within the running time point of [0,50], i.e., within the first thirty minutes of the game. It is also the case that Player two has more momentum and eventually wins the game.

3. Models to Quantify Momentum Analysis

To examine whether momentum is a key factor in the competition and to offer guidance for the same. This study encodes the "point_ Victor" column, designating victories as 1 and defeats as 0, and conducts normality and variance ratio tests on the initial question regarding momentum score outcomes. The data distribution does not meet the criteria for normality and variance ratio, hence the kWh test is applied to assess momentum and scoring [13]. Detailed procedures are as follows.

3.1. Normality Test

$$W = \{\sum_{i=1}^{[n/2]} a_i(W)[X_{n+1-i} - X_{(i)}]\}^2 / \sum_{i=1}^{n}[X_{(i)} - \overline{X}]^2 \quad (1)$$

First the article use the normality test to make a test on whether the data obeys a normal distribution. Normal distribution is a common probability distribution which is used in many natural phenomena and social science research [14].

Step 1: Arrange the n sample observations in descending order: $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(i)} \leq \ldots \leq x_{(n)}$

Step 2:The W-test statistic is:

Where, [n/2] denotes the integer part of n/2; the value of $a_i(W)$, given by the table of coefficients of the statistic *W*. The *W* statistic can be viewed as the ratio of the optimal estimate of the variance based on the squares of some linear combination of the order statistics to the sample variance of the data;

Step 3: Based on the given significance level α and sample size n, the p-quantile table of the statistic *W* is checked to determine the value of the α-quantile W_a. Its quantile table is generated by computer simulation;

Step 4: Calculate the value of the statistic *W*. If $W < W_a$, reject H0 and consider the data as not obeying a normal distribution; conversely, do not reject H_0. Since the distribution of the *W* statistic has a large skewness, values of *W* close to 1 may also lead to the rejection of the original hypothesis of normality.

Table 4. Ormality test.

Whether it scores or not	Kolmogorov Smirnov			Shapiro Wilke		
	statisticians	df	P-value	statisticians	df	P-value
0	0.071	2236	0.000	0.960	2236	0.000
1	0.045	5048	0.000	-	-	-

As shown in Table 4, from the result of normality test, P-value=0.000<α=0.05, so the original hypothesis H_0 is rejected and the data do not obey normal distribution.

3.2. Variance Chi-square Test

The variance chi-square test is used as a method in mathematical statistics to check whether the overall variance of different samples is the same. The basic principle is to make some assumptions about the characteristics of the population, and then make inferences about whether this assumption should be rejected or accepted through statistical inference in sampling studies [15]. Before the Kruskal-WallisH test, a test of chi-squareness is required, which is performed as follows:

Hypothesis formulated: H_0: Data satisfy variance chi-square

H_1: Data do not satisfy variance chi-square

When the number of replications of the test is equal under each level, i.e.

$$m_1 = m_2 = \cdots = m_r = m \tag{2}$$

m is the number of replications at each level, which is the ratio of the maximum to the minimum of the r sample variances. Under the condition that all variances are equal, the quantile of the *H* distribution can be obtained by stochastic simulation method, which depends on the number of levels *r* and the degrees of freedom of the sample variances *f=m-1*,so that the distribution can be denoted as *H(r, f)*. When *H*, holds, then we have.

$$\partial_1^2 = \partial_1^2 = \ldots = \partial_r^2 \tag{3}$$

The value of *H* should be close to 1. When the value of *H* is large, the difference between the variances is large, and the larger *H* is, the larger the difference between the variances is, at which point the original hypothesis *H0* is rejected. It can be seen that, for a given level of significance ɑ, the rejection domain of the test H_0 is

$$W_1 = \{H > H_{1-\partial}(r, f)\} \tag{4}$$

Where $H_{1-a}(r, f)$ is the *1-a* quantile of the *H* distribution. If the data in the unordered variable neither obey normal distribution nor satisfy variance chi-square, the Kruskal Wallis H-test can be considered. The results of the variance chi-square test are as follows Table 5.

Table 5. Variance Chi-square test.

	Levin statistics	df 1	df 2	significance
Based on average values	16.739	1	7282	.000
Based on median	12.778	1	7282	.000
Based on median with adjusted degrees of freedom	12.778	1	7246.502	.000
Based on post clipping average	15.749	1	7282	.000

From the results of variance chi-square test, it can be seen that P-value=0.000<ɑ=0.05 for the four categories, so the original hypothesis was rejected and the data did not meet variance chi-square.

3.3. Kruskal Wallis H

The correlation analysis of momentum and score requires analyzing the correlation between continuous and categorical variables, and as shown by the results of the variance chi-square test, the data satisfies neither normal distribution nor variance chi-square, so the Kruskal-Wallis H test was used. The Kruskal-Wallis H test is a nonparametric test used to test whether multiple groups of continuous or ordered categorical variables for the presence of differences.

3.3.1. Model Building

The principle is to sort all data from smallest to largest and then calculate the rank of each data. Where R_i is the rank sum of each group and n_i is the number of samples in each group. When the number of observations in each group of samples is five or more, the distribution of the sample statistic KW is very close to the chi-square distribution with k-1 degrees of freedom. Therefore, the KW statistic can be tested using the chi-square distribution. The KW test statistic for differences between multiple samples is:

$$H = \frac{S_{SST}}{S_{MST}} = \frac{12}{n(N+1)} \sum_{j=1}^{K} \frac{R^2}{n_j} - 3(n+1) \sim \chi^2(k-1) \tag{5}$$

S_{SST} is the sample inter-treatment sum of squares, S_{MST} is the total mean square, *k* is the number of samples; n_j is the jth sample capacity j=1,2,...,k; is the overall sample capacity; R_j is the rank sum of the jth sample.

According to equation (5) the KW test statistic can be calculated as its corresponding P-value, (α $H > \chi_{\alpha}^{2}(k-1) \text{or} P < \alpha$ is the level of significance, usually 0.05), indicating a significant difference between the samples.

3.3.2. Model Solution

The results of the KW test for the presence of differences in the data are as follows Table 6.

Table 6. Inspection Statisticsa, b.

	player1
Kruskal Wallis H	132.333
df	1
Asymptotic significance	.000

Note: a. Kruskal-Wallis test b. Grouping variables: whether or not to score

From the test results, P-value=0.000<α=0.05, the original hypothesis is rejected, i.e., there is a significant difference between the samples, and it can be concluded that there is a relationship between players' performance and scoring, and momentum plays an important role in the game.

4. Conclusions

For strategic sports like tennis, the impact of "momentum" on the game has been deeply and closely studied from multiple angles and levels through the construction of models such as random forests and Kruskal Wallis H. Here are the conclusions:

The data of tennis match has the characteristics of high dimension and large sample, so the application of random forest can be applied to the excellent performance of processing high dimension and large amount of data, without manual data cleaning, feature combination, unbalanced sample processing and even vacancy filling; At the same time, during the training, the random forest does not need to be specially tuned, but only needs to increase and decrease the appropriate trees, and its effect is obvious, that is, it will improve the accuracy, but it will also increase the risk of over-fitting, and there is no need to formulate a special loss function, which embodies the advantages of simple training and strong anti-interference.

Whether in finance, medical care or marketing, random forest can provide effective solutions to support the decision-making process.

The application scope is wide, and it can be applied to the analysis of non-normal distribution data. Tennis situation data does not always meet the assumption of normal distribution, and some indicators are affected by many factors, and their distribution patterns are complex. However, Kruskal-Wallis H test does not strictly require the shape of the overall distribution, and is suitable for all kinds of unknown distribution data, which has great advantages in the face of such non-normal data. The future research direction can promote the application of Kruskal-Wallis H test in interdisciplinary research, and combine with the methods and technologies of other disciplines to solve interdisciplinary problems. For example, the tennis momentum analysis and socio-economic factors cross-analysis.

References

[1] Liu Y, Qi CZ, He MY, etc Visual motion tracking training: the effect and mechanism of improving tennis players' perceptual prediction ability [J] China Sports Science and Technology, 2021, 57 (10): 31-37

[2] Zhao Y, Cui YX, Liu HY, etc Analysis of Winning Factors for Professional Male Tennis Players Based on their Performance in Grand Slam Events [J] Journal of Beijing Sport University, 2022, 45 (04): 78-90

[3] Shi GB, Ren YQQ, Shi L, et al. Decision-making model for optimal return landing point in tennis based on machine learning[J]. Sports Science, 2023, 43(06):53-60.

[4] Zhao MN, Qi B, Ge CL. The predictive role of emotional intelligence on sports performance in tennis players [J]. Journal of Xi'an Institute of Physical Education, 2016, 33(01):119-123.

[5] Pei CY. Prediction system of optimal hitting point of tennis serve based on Kalman filter [J]. Modern Electronic Technology, 2018, 41(11):162-165+170.

[6] Wang W, Zhou S, Zhang CH, et al. Study on Djokovic's technical and tactical use of key points in the 2016 Australian Open - based on angle, line and area perspectives [J]. Journal of Henan Normal University (Natural Science Edition), 2016, 44(05):182-188.

[7] Li JJ, Ding W, Wang BW, etc A method for evaluating the impact of folk sports on physical indicators based on random forests [J] Journal of Nanjing University (Natural Sciences), 2021, 57 (01): 59-67

[8] Zhao MN, Jiang HQ, Qi B, etc Construction of Emotional Intelligence Theory Model and Scale Development for Chinese Tennis Players [J] Psychological Research, 2024, 17 (01): 42-53

[9] Liu C, Hou HS, Shen DH, etc Empirical Study on Training and Competition Load of High level Tennis Athletes [J] Journal of Tianjin Sport University, 2024, 39 (02): 241-248

[10] Li WY, Liang H, Cheng H, etc Research on Identification Method of Chinese Herbal Flavor Based on Raman Spectroscopy [J/OL] World Traditional Chinese Medicine, 1-7 [2024-08-03]

[11] Zhang H, Gao SF, Zhao X, etc Construction and evaluation of corneal reshaping mirror myopia control effect and prediction model based on random forest and logistic regression [J] New Advances in Ophthalmology, 2024, 44 (08): 637-642

[12] Li JX, Liu JT, Wang ZT, etc Fine drawing and spatiotemporal evolution analysis of winter wheat in the Yellow River Delta using collaborative Sentinel-2 and random forest algorithm [J/OL] Journal of Agricultural Resources and Environment, 1-15 [2024-08-03]

[13] Chen JJ, Yu QZ, Tang QX, etc Analysis of Spectral Characteristics Differences between New and Old Leaves of Common Tree Species in Changbai Mountain [J] Spectroscopy and Spectral Analysis, 2024, 44 (08): 2372-2380

[14] Hu B, Hu ZL, Li G, etc Research on the Evaluation Method of Static Explosion Shock Wave Data Processing [J] Journal of Weapon Equipment Engineering, 2024, 45 (03): 114-120

[15] Ding Y, Wu F Research on Statistical Inference Method Based on Table Tennis Competition Techniques and Tactics [J] Journal of Guangzhou Sport University, 2021, 41 (03): 78-82+86