Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 12, No. 1, 2024

A Comprehensive Study on the Role of Momentum in the Sports Games

Yaqi Fan #, Shilin Fan #, Kaiyu Zhang *

College of Automotive Engineering, Jilin University, Changchun, 130022, China

*** Corresponding author:** Kaiyu Zhang (Email: karrychueng@gmail.com)

#These authors contributed equally

Abstract: This study aims to explore the effects of momentum and swing on tennis match results. The paper believes that momentum may be a key factor affecting player performance and match results. The research methods include data preprocessing, PCA dimensionality reduction, Pearson correlation coefficient analysis, logistic regression model and random forest model. Data preprocessing includes outlier treatment, one-hot encoding and normalization. PCA is used to reduce the dimension of match indicators and extract the main information. Pearson correlation coefficient analysis is used to test the correlation between match results and momentum. Logistic regression model is used to predict the probability of a player winning a point, and random forest model is used to predict match fluctuations. The study uses data from the 2023 US Open to validate the model. Sensitivity analysis is performed by changing model parameters to ensure model accuracy. The conclusion shows that momentum does affect match results, and our model can predict match performance and fluctuations with high accuracy. This is of great significance for coaches and athletes to formulate strategies and improve match performance. This study is of great value in understanding the momentum effect in tennis and how to use this knowledge to improve athlete performance.

Keywords: Random Forests; Momentum Forecast; Logistic Regression Pearson; Match Swings; Correlation Coefficient.

1. Introduction

The world of men's professional tennis is growing rapidly, with players improving their performances in the game and becoming increasingly competitive. The exploration of the characteristics and laws of the professional tennis game has received increasing attention. Among them, the influencing factors of winning and losing in high-level professional tennis matches is an issue that has always been paid attention to. Most of the previous studies have focused on the exploration of single factors, such as the predictive effect of ranking on the outcome of a match, the effect of a player's match experience on the outcome of a match, and the effect of a player's height on the serve, etc., and there have not been many studies that have provided a comprehensive answer to the question. Due to the complexity of the tennis program, the analysis of a single influencing factor is difficult to effectively explain the results of the game and requires a more comprehensive and integrated assessment.

Momentum, a term widely used in sports competitions, refers to the power and momentum of a team or individual in a game. In sports, momentum plays an important role in the trend and outcome of the game. The concept of momentum was made even more mysterious when Spanish rookie Alcaraz unexpectedly defeated Grand Slam legend Djokovic in the 2023 Wimbledon men's singles final. What role does momentum play in a game, making it was able to defeat Djokovic. How does it work, when does it work, and what factors are related to it? A series of questions have been raised. Therefore, in order to better understand the concept of momentum, there is a great need to devise some models to address these questions. Considering the background information and restricted conditions, firstly, the research want to develop a model capable of analyzing changes in scoring during a match that evaluates whether a player is dominant in a match and the degree of his strengths or weaknesses. Meanwhile, considering the serve advantage factor, the model can visualize the flow of the match. Secondly, it wants to build a model to evaluate whether player momentum and game outcomes are random. Thirdly, it needs a model to predict the moments in the match when the situation changes and identify the factors that are most relevant to the change in the situation of the match and make recommendations for the players based on the predictions. At last, testing the model with data from other matches, evaluating how well the model predicts fluctuations and analyzing the applicability of the model to other matches.

Based on the above research objectives, our workflow is as follows: Firstly, the study pre-processed the raw data, including outlier handling, data normalization and principal component analysis. Through principal component analysis, the paper obtained some factors that have a greater impact. Then, based on these factors, it analyzed the players' performance using a logistic regression model and showed the flow of the competition. Secondly, it used the Pearson correlation coefficient and the Spearman correlation coefficient to verify whether momentum can affect the outcome of the game. Correlation analysis is established between the momentum sequence and the competition results. Then the study used a random forest model to analyze the impact of momentum on the game and made recommendations to players and coaches on how to win the game based on the results of the analysis. Finally, it collected some data from other competitions in US Open's match data for the year 2023. The result shows that the model is suitable for other competitions.

2. Player Performance Analysis

2.1. Data Preprocessing

Using one-hot encode to process classified data, and

convert them into 0, 1 encode according to the classification of the S_{ew} column, the S_{ed} column and the R_{ed} column. Convert the hours in the E_{lt} column to minutes. By using python programming, it converts the p1_score into 0 and 1 according to whether the score is got. 0 means no score, 1 means score, and AD is judged by the score of the latter digit. The data of Table 2 on one-hot encode processing is as follows:

Table 1. one-hot encode

S_{ewB}	1	0	0	0	0
S_{ewBC}	0	1	0	0	0
S_{ewBW}	0	0	1	0	0
S_{ewC}	0	0	0	1	0
S_{ewW}	0	0	0	0	1
S_{edCTL}	0	1			
S_{edNCTL}	1	0			
R_{edD}	0	1			
R_{edND}	1	0			

2.1.1. Principal Component Analysis.

Principal component analysis is a multivariate statistical method that uses the idea of dimensionality reduction to transform multiple original indicators into several comprehensive indicators while ensuring that the loss of information is as small as possible. Because high-dimensional data is mapped to a low-dimensional space through linear projection based on the principle of maximum variance, using this method for feature dimensionality reduction can well solve the problem of feature redundancy in the original feature matrix.

The study splits the various indicators of athletes in the data table. Considering that the values of some indicators are too large, it uses the range transformation method to normalize them. The formula of the range transformation method is shown in equation (1). Use python coding to process the obtained data, and then obtain the principal components F1~F9 and the component matrix, as shown in Table 1.

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

Table 2. Principal component contribution rate

Principal component	Contribution rate %	Cumulative contribution rate %
1	16.765	16.765
2	16.093	32.858
3	10.396	43.254
4	9.234	52.488
5	8.070	60.588
6	6.886	67.444
7	5.566	73.010
8	5.125	78.135
9	4.934	83.070

The contribution rate of each component is shown in Table 1. By analyzing Table 2, it founds that the cumulative contribution rate of the 9th principal component has exceeded 80%, so the first 9 principal components are extracted to reflect the original high-dimensional features.

Because the contribution rate of the first five principal components is high, the study extracted the first five principal components as the main influential factors, as shown in Table 3. By analyzing of Table 2, it founds that the competition situation is mainly affected by S_{ed} and R_{ed}, and is also related to S_{ew}, S_{en}, S_e, P_{1n}, W_{is}. The actual competition situation is also related to depth of serve, depth of return, direction of serve, first or second Serve and other factors are related, indicating that principal component analysis is more accurate.

Table 3. The main influencing factors of the principal component

	Principal component				
The main influencing factors	1	2	3	4	5
1	S_{edNCTL}	R_{edND}	S_{ewC}	S_e	W_{is}
2	S_{edCTL}	R_{edD}	S_{en}	P_{1n}	S_e

2.2. Logistic Regression Model

The score changes in tennis matches reflect the strength and status of both parties in the match, as well as the intensity and trend of the match. Analyzing the score changes during the game can help us evaluate which tennis player has the advantage in the game, as well as the size and stability of the advantage. At the same time, in tennis matches, the server usually has a certain advantage because the server can actively control the rhythm and direction of the game, while the receiver needs to respond based on the quality and speed of the server's serve. Therefore, considering the server's advantage factors can allow us to more accurately analyze the score changes in the game.

Our goal is to develop a model that can analyze score changes in the game. The model should be able to analyze various factors in the game, such as the player's serving situation, return situation, movement distance, movement speed, and the server's advantage factors. It can predict a player's probability of winning to assess their advantage in a match. The probability of winning is higher for the first serve in tennis, the study analyzes the extracted principal components and weight the contribution of each index in the principal components, considering the first serve advantage, sever is higher in the 4th principal component. According to [11], the model appropriately increases the weight of the 4th principal component. The final loss function is as (2).

$$L(w) = \min(-\frac{1}{n}\sum_{i=1}^{n} y_i(w^T x_i + b) - \ln(e^{w^T x_i + b} + 1)) \tag{2}$$

Set the algorithm to output w and b when the loss function is less than the threshold, and predict whether the player scores or not based on the player's serving situation, return situation, movement distance, movement speed and other characteristics during the game. The 9-dimensional vector obtained by principal component analysis is input into the logistic regression model as x, and the processed score is used as for logistic regression. The results are shown in Table 4. After analysis, it was found that the accuracy of the model

in predicting non-scoring was 72%, and the accuracy of scoring was 68%, indicating that the model can reflect the player's scoring situation.

Table 4. Classification Report (Balanced)(player1)

	precision	recall	f1-score	support
0	0.72	0.56	0.62	719
1	0.68	0.62	0.64	474
Accuracy			0.58	1193
Macro avg	0.70	0.59	0.58	1193
Weighted avg	0.71	0.58	0.59	1193

The ROC curve (Receiver Operating Characteristic) is shown in Figure 1. The study found that the closer the image is to the (0,1) point, the stronger the model's ability to distinguish positive and negative samples. It can be seen that our model can reflect the scoring situation based on player characteristics.

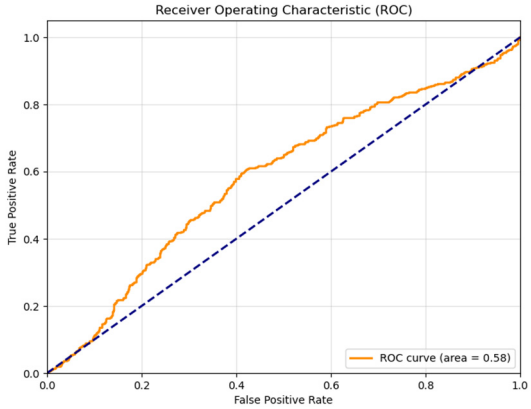

Figure 1. The ROC curve

Precision refers to the prediction accuracy of positive samples. For example: if it regards a player's win as a positive sample, it can know the model's prediction accuracy for the player's win. In our prediction results, there are two kinds of positive examples: it turns out that the positive example is predicted to be a positive example, and it turns out that the false example is predicted to be a positive example. Therefore, the model wants to know the proportion of real positive examples to all the positive examples in the prediction result. That is: the accuracy rate of a player's victory.

The formula of precision for a player to win is as the equation (3):

$$P = \frac{TP}{TP + FP} \tag{3}$$

Recall refers to the proportion of samples predicted as true examples to all true positive samples. For example: If the study regards player wins as positive samples, it wants to know whether the model can predict all player wins.

The formula of recall for a player to win is as (4):

$$R = \frac{TP}{TP + FN} \tag{4}$$

It can be seen that the logistic regression model predicts player wins with an accuracy of 72% and a recall of 68%. The study performed a comprehensive evaluation of the model using F_{1s} as the equation (5):

$$F1 = \frac{2TP}{2TP + FN + FP} = \frac{2 \bullet Precision \bullet Recall}{Precision + Recall} \tag{5}$$

Among them, TP, TN, FP, and FN respectively represent the number of true examples, true negative examples, false positive examples, and false negative examples. The mean value of F_{1s} obtained by this logistic regression is 0.63, so the prediction of this model is relatively accurate. The study extracts the main influencing factors from the principal components with large contributions, and add these factors according to their weights to obtain P_{1p}, which represents the trend of the game at a certain moment.

3. Results

3.1. The Influence of Momentum on the Result of Competition

3.1.1. Analysis of Momentum Influence Mechanism

From the principal component analysis, it can be seen that the player's score is related to many indicators such as S_{ed}, R_{ed}, S_{ew}, S_{en}, S_e, P_{1n} and W_{is}. In order to quantify momentum, the study detects athlete scores and use bonus points, subtraction points, continuous bonus points, and continuous subtraction points to quantify momentum. If score = 1, the momentum value + 1, if score = 0, the momentum value - 0.5, if score = 1 for 2 consecutive times, the momentum value + 1.5; if score = 1 for 3 consecutive times, the momentum value + 2; score Four consecutive times = 1, momentum value + 2.5; score 2 consecutive times = 0, momentum value - 1. Through the python program, it gets the momentum value of an athlete in multiple games, and then introduce a new variable winnner to record the winning or losing results of each game. If the athlete wins, winner=1, if the athlete loses, winner=0. Then the study uses momentum and winnner to do pearson correlation analysis. Pearson correlation coefficient[6] is often used to calculate the correlation between 2 variables. In this paper, it calculates the Pearson correlation coefficient between the value of momentum and winnner to determine whether the value of momentum affects the athlete's ability to win. When the 2 sets of variables are $X = (x_1, x_2, \cdots, x_n)$ and $Y = (y_1, y_2, \cdots, y_n)$. The formula for variables X and Y Pearson's correlation coefficient [1] is as shown in (6):

$$r = \frac{\sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y)}{\sqrt{\sum_{i=1}^{n}(x_i - \mu_X)^2}\sqrt{\sum_{i=1}^{n}(y_i - \mu_Y)^2}} \tag{6}$$

Where r represents the Pearson correlation coefficient, X represents the value of the momentum, Z represents whether the athlete won the race or not, μ_X is the average value of variable X, μ_Y is the average value of variable Y, the greater the absolute value of r, the higher the correlation between momentum and player winning.

The Spearman correlation coefficient determines the correlation through the hierarchical relationship between two variables[8]. It is a parameterless rank correlation coefficient, that is, its value has nothing to do with the specific values of the two related variables, but is only related to the size

relationship between their values. It is not affected by the sample data distribution and sample size, and it is suitable for missing values and extreme values. The existence of value distribution has better adaptability. The formula of spearman's rank correlation coefficient (continuous variable) [2] is as (7):

$$r = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)} \tag{7}$$

Where r represents the Spearman correlation coefficient; N represents the sample size; d_i represents the position difference of the paired variables after the two variables are sorted respectively.

The study used the Pearson correlation coefficient and the Spearman correlation coefficient to verify whether momentum can affect the outcome of the game. Correlation analysis was established between the momentum sequence and the competition results. The results are shown in Table 5 and Table 6. It can be seen that the Pearson coefficient r = 0.714, the Spearman coefficient r = 0.674, and the significance test values are all less than 0.5, indicating that the confidence interval is 95 %, the results are credible, indicating that score changes are related to momentum, and scores are not random.

Table 5. Spearman correlation coefficient

		P_{1m}	M_{aw}
P_{1m}	Spearman correlation coefficient	1.0	0.583
	Sig.(2-tailed)		0.032
	N	31	31
M_{aw}	Spearman correlation coefficient	0.583	1.0
	Sig.(2-tailed)	0.032	
	N	31	31

Table 6. Pearman correlation coefficient

		P_{1m}	M_{aw}
P_{1m}	Pearman correlation coefficient	1.0	0.672
	Sig.(2-tailed)		0.028
	N	31	31
M_{aw}	Pearman correlation coefficient	0.672	1.0
	Sig.(2-tailed)	0.028	
	N	31	31

3.1.2. Momentum-based Decision Suggestions

The random forest algorithm is a relatively effective classification prediction method, which can handle fields such as nonlinear relationships, classification, regression, high-order correlation and evaluation of the importance of variables by combining multiple decision tree combinations. Its classification accuracy is high, it can effectively prevent the classification accuracy from being too low in the presence of noise and outliers, and its generalization ability is relatively strong. It demonstrates broad applicability and becomes a powerful tool for solving complex problems. Firstly, the study defines the fluctuation of the match p1_p2 as (8):

$$p1\_p2 = p1\_points\_won - p2\_points\_won \tag{8}$$

Among them, p1_points_won: number of points won by player 1 in match; p2_points_won: number of points won by player 2 in match

The paper draw p1_p2, and the point where the polyline changes greatly is the swing point:

swing=1, Fluctuations occur

swing=0, No fluctuations

In order to find factors related to the fluctuation points of the game, the study extracted various indicators in the game as features and normalized them using the range change method. Then, it uses python programming to input the features and fluctuations into the random forest for feature sorting. Among them, it founds that P_{ov}, P_{1dr}, S_{pm}, R_{ac}, S_{ew} and other factors are of strong importance. It shows that the player's running distance, running speed, width of serve and game results have a greater impact on the fluctuations of the game. The study uses the top five factors that have a greater impact on volatility to make our forecasts [3]. It uses random sampling method to randomly select 10 games from the data set, 8 games as the training set and 2 games as the test set. The training set is used to train the model, 0 represents no fluctuation, and 1 represents fluctuation. Because there are few fluctuations in the competition, in order to improve the classification accuracy, the paper increases the weight of fluctuations [4], uses random forest parameters to select the best, and adjust the n-estimators value, that is, the number of decision trees. It determined that the optimal value of N_e is 11 through parameter tuning, and the results are shown in Tables 7.

Table 7. Classification report

P_{1dr}	0.390	
S_{pm}	0.360	
R_{ac}	0.186	
S_{ewC}	0.033	
P_{ov}	0.031	
	precision	support
0	0.80	132
1	0.67	34
accuracy	0.80	

The study found that the importance of p1distancerun and speed accounted for a large proportion, it can suggest that p1distancerun and speed have a great influence on the fluctuation. Random forest model has 67% accuracy in predicting the fluctuation and the prediction result is more accurate. For performance evaluation, it used the accuracy index as shown in (9). The model's accuracy is 79.5%, the accuracy is high and proves the validity of the model.

$$accuracy = \frac{TP + TN}{TP + TN + FP + TN} \tag{9}$$

Based on the player's game data, classification predictions are made using the Random Forest model. Players as well as coaches can then learn about possible fluctuations in P_{ov}, P_{1dr} S_{pm}, R_{ac}, S_{ew}, etc. during a match. Players can then adjust their position to improve the quality of their serves

on key points. It is beneficial to increase the advantage. It can analyze the opponent's players according to the model, understand the opponent's game data, and adjust the game strategy according to the opponent's vulnerability and our own strengths, such as increasing the attack when the opponent's physical ability is declining.

3.2. Analysis of Experimental Results

3.2.1. Analysis of Results

To test the model, the study collected match data from the US Open website for the year 2023. US Open's match data for the year 2023 contains both men and women. It ranked the features of each metric, found that point Winner and S_{pm} consistently accounted for a larger portion , and in this test, P_{os}, P_{1ue} and P_{1np} emerged as the factors that had a greater impact on game fluctuations, illustrating the fact that factors such as player's running speed and the number of mistakes made during the game, also have a greater impact on game fluctuations.

The names of the top five indicators with high impact are shown in Table 8.

Table 8. The names of the top five indicators

	Feature	Importance
2	S_{pm}	0.654
1	P_{os}	0.144
4	P_{1v}	0.076
6	P_{1ue}	0.048
7	P_{1np}	0.036

The study found that point Winner and S_{pm} consistently accounted for a larger portion , and in this test, P_{os}, P_{1ue} and P_{1np} emerged as the factors that had a greater impact on game fluctuations, illustrating the fact that factors such as player's running speed and the number of mistakes made during the game, also have a greater impact on game fluctuations. The paper still uses these five factors that have a large impact on match fluctuations for prediction. A random sampling method was used to randomly select 10 matches in the collected data set, 8 matches were used as the training set and 2 matches were used as the test set, and the results were shown in Table 9.

Table 9. Classification report

S_{pm}	0.567	
P_{1ue}	0.185	
P_{os}	0.136	
P_{1v}	0.062	
P_{1np}	0.050	
	precision	support
0	0.81	126
1	0.20	30
accuracy	0.80	

The study found that the importance of speed is more dominant in this test, so it can indicate that speed has a great impact on fluctuation. The accuracy of random forest model in the prediction of volatility is 0.20, the prediction accuracy has decreased.

To analyze the reason, the data analyzed in this test is for the preliminary stage, not the final stage, the gap between the players' strength may be larger, so the number of fluctuations will be reduced, while the increase in the number of non-fluctuations will cause the problem of imbalance in the category, so the prediction accuracy decreases.

For performance evaluation, the accuracy of this test is 79%. The accuracy is relatively high so the model can be considered effective.

3.2.2. Analysis of Experimental Results

As mentioned earlier, parameter optimization for random forests generally adjusts the value of n-estimators, i.e., the number of decision trees. The study determined through parameter optimization that the optimal value of n-estimators is 11: when the number of decision trees is 11, the model has the highest accuracy, with a precision of 0.8 for 0, and 0.67 for 1. By changing the values of n-estimators, the study found that most of the values of n-estimators made the model's precision for 0 to be 0.8, which is unchanged compared to the best precision. The model's precision for 1 is 0.58, which is a decrease of 0.09 compared to the best precision. they are within the acceptable range, then the model is considered to have passed the sensitivity test.

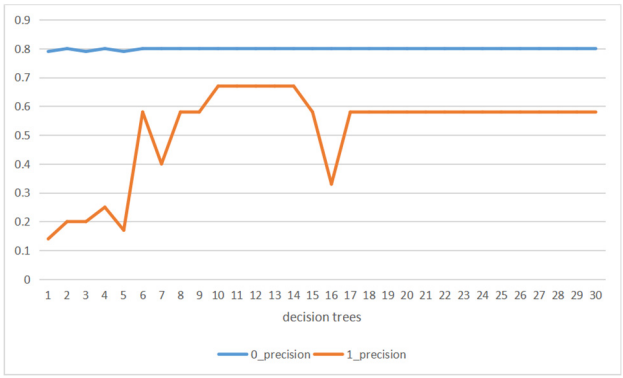

Figure 2. Sensitivity analysis of decision trees

4. Conclusion

Our study used pearson correlation analysis to analyze the correlation between the change of athletes' scores and momentum, and a random forest model to rank the influence of the indicators on the fluctuation of the game. The results showed that there was a significant correlation between the change of athlete's scores and momentum, and the scores were not random. It indicated that the player's distance ran during point and speed of serve were the most influential indicators, accounting for more than half of the importance. So, momentum does play a role in tennis matches and that the fluctuations in the game are not random. The results also identified the indicators that can help determine when the flow of play is about to change from favoring one player to the other. The paper uses the US Open's match data for the year 2023 to test the model, and the result shows that the model can be used for other competitions.

References

[1] Anbarci N, Lee J, Ulker A. Win at all costs or lose gracefully in high-stakes competition? Gender differences in professional tennis[J]. Journal of Sports Economics, 2016, 17(4): 323-353.

[2] Qiu H, Liu C, Zhang X. [Retracted] Intelligent Design of Tennis Player Training Schedule Based on Big Data of Complexity[J]. Complexity, 2021, 2021(1): 4759395.

[3] Wang Z H, Pan R C, Huang M R, et al. Effects of integrative neuromuscular training combined with regular tennis training program on sprint and change of direction of children[J]. Frontiers in physiology, 2022, 13: 831248.

[4] Coulon T, Barki H, Paré G. Conceptualizing project team momentum: a review of the sports literature[J]. International Journal of Managing Projects in Business, 2021, 14(2): 270-299.

[5] Russomanno T G, Lam H, Knopp M, et al. Within-Match Performance Dynamics–Momentary Strength in Handball[J]. Journal of Human Kinetics, 2021, 79(1): 211-219.

[6] Liu Y, Mu Y, Chen K, et al. Daily activity feature selection in smart homes based on pearson correlation coefficient[J]. Neural Processing Letters, 2020, 51: 1771-1787.

[7] GUO Liang, GUO Zixue, JIA Hongtao, et al. Residents electric larceny detection based on pearson correlation coefficient and SVM[J]; Journal of Hebei University (Natural Science Edition), 2023, 43(04):357-363.

[8] Ali Abd Al-Hameed K. Spearman's correlation coefficient in statistical analysis[J]. International Journal of Nonlinear Analysis and Applications, 2022, 13(1): 3249-3255.

[9] ZHANG Weifeng. Statistical Analysis of Spearman's footrule and Gini’s gamma[D]; Guangdong University of Technology, 2020.

[10] YAO Judeng, YANG Jing, ZHAN Xiaojuan. Feature selection algorithm based on random forest[J]; Journal of Jilin University (Engineering and Technology Edition), 2014, 44 (01):137-141.

[11] WANG Jiaqi, ZHU Junguo, YU Zhengtao. Low-Resource Machine Translation Based on Training Strategy with Changing Gradient Weigh[J/OL]. Journal of Frontiers of Computer Science and Technology:1-10.

[12] LoMonte F D, Suszan B, Dames P. Open and Shut? The Promise-And Problems-Of Government Open Data Portals in Meeting Community Information Needs[J]. UCLA JL & Tech., 2023, 28: 93.