Research on Tennis Match Result Analysis Based on Multi-model Integration

Xiaodi Shao, Zengqing Bai*, #, Zenghui Liu#

School of Engineering and Technology, China University of Geosciences, Beijing, China, 100083

*Corresponding author: bzengqing@gmail.com

#These authors contributed equally.

Abstract. Tennis is particularly prominent among the many highly competitive sports, and it is critical to provide scientific strategic guidance to players in modern tennis competitions. Based on data from the men's singles final of the 2023 Wimbledon Tennis Championships, this study synthesized match flow and scoring dynamics, quantified key metrics (e.g., serve dominance and player ability), and constructed the Momentum Evaluation Model to assess and predict trends in athlete momentum. By analyzing the main factors affecting momentum shifts, the model can predict the game's outcome more accurately. The study results show that momentum flows to the side of the more dominant player with different metrics. The model test shows that its prediction accuracy reaches more than 70%, which provides solid theoretical support for the tactical analysis of tennis matches and the optimization of athletes' on-court performances, as well as an effective guide for scientific, data-driven decision-making in tennis.

Keywords: Neural Network, Random Forest, Decision Tree Regression, Tennis.

1. Introduction

In modern sports competitions, data analytics has become an important tool to enhance athletes' performance and develop game strategies. Tennis is a highly competitive sport, every detail during the match may affect the result. With the development of big data technology, more and more researchers and coaches have begun to utilize data analytics to dig deeper into the key factors in the game to provide athletes with more scientific advice and guidance. Among them, the change of momentum is considered one of the most important factors affecting the trend of the game, which not only reflects the physical and technical state of the athletes in the game but is also closely related to the psychological state. However, how to quantify and evaluate the specific impact of momentum on the outcome of a match remains a complex and challenging problem.

This study applies the Random Forest algorithm to explore the effect of momentum on the outcome of a race. This algorithm, together with BP Neural Networks and Decision Tree Regression, constitute a variety of prediction techniques used in sports, where BP Neural Networks optimize the modeling of complex features through a hierarchical structure and Decision Tree Regression achieves an accurate association of features with outputs using recursive segmentation. Among the existing prediction studies, Claudino et al[1]used the BP Neural Network algorithms and Decision Tree algorithms to predict athlete ability and risk of sports injuries and validated the potential of AI technology in several sports. Goals et al[2], on the other hand, showed that the BP Neural Network algorithms outperformed the traditional regression model and performed more accurately in javelin score prediction. The NBA and the Australian Football League also use the BP Neural Network algorithms to optimize playing order and game strategy in terms of team performance and player selection to improve win rates[3][4]. The study by Xia et al[5] quantitatively analyzed the positive effects of physical activity on physical and mental health using the BP Neural Networks based on key exercise indicators. Wang et al in table tennis technical and tactical analysis pointed out that the BP Neural Networks can accurately fit the test dataset within the control error to ensure the objectivity and practicality of the evaluation[6].In summary, algorithms such as BP Neural Networks, the Random Forest, and Decision Tree Regression not only improve the accuracy of sports event prediction but

also show a wide range of application prospects in the areas of athletes' physical fitness, tactical selection, and injury risk assessment.

In this study, the data were preprocessed using the Random Forest Classification algorithm, and the Momentum Evaluation Model was developed for further quantifying momentum changes and match score changes. Secondly, by using the BP Neural Network algorithm with the Decision Tree Classification algorithm, this study calculates the weights of the factors affecting the change of momentum of athletes in tennis matches and predicts the results of the tennis matches by using the weights and analyses the accuracy.

2. Data Source and Preprocessing

Data for this article was obtained from https://www.comap.com/contests/mcm-icm,, there were multiple momentum shifts during the game.

2.1. Missing Data Handling

Since some of the data provided is missing, we use triple standard deviation to fill in the missing values, and at the same time calculate the mean and standard deviation separately for different features to fill in the missing values of the corresponding features. Filling the missing values with triple standard deviation can maintain the distributional characteristics of the data, and the filled values will not be heavily concentrated around the median value, which improves the accuracy of the data distribution.

2.2. Min-Max normalization

Transform the sequence$x_1, x_2, \ldots, x_n$:

$$y_i = \frac{x_i - \min\limits_{1\le i\le n}\{x_i\}}{\max\limits_{1\le i\le n}\{x_i\} - \min\limits_{1\le i\le n}\{x_i\}} \tag{1}$$

Then the new sequence $y_1, y_2, ..., y_n \in [0,1]$ is dimensionless. Advanced normalization can be considered for general data needs. The $Min-Max$ normalization is a linear transformation of the original data that maps values between $[0,1]$.

3. Factors Influencing Match Momentum and Their Weight

3.1. Random Forest Classification Algorithm

By checking the official ATP website and the websites of related sports events, the following technical data that significantly affect the momentum change were obtained[7]:

First-serve success rate, first-serve scoring rate, second-serve scoring rate, reception scoring rate, break rate, total scoring rate, ACE, double faults, unforced errors, and active scoring can reflect the technical and tactical use of both sides of the match and are easy to judge and record in the statistics . Therefore, this study summarizes and concludes the three main influences on the scoring of the match, and the pre-processed data will be analyzed and selected according to the rules and procedures of the match. This is illustrated in Figure 1.

Physical energy factor: Tennis matches are long and intense, requiring many swings and runs, therefore, tennis matches have extremely high endurance requirements. We analyze the data provided according to the rules of the game and judge that the running of each player in the game will consume physical energy, so the increase in running distance will hurt the change of momentum.

Skill factors: With the development of tennis, tennis players have higher skills and more comprehensive playing styles. Upon analysis, we determined that factors such as players hitting unreturnable balls, hitting the ball at a high speed, controlling the ball's landing position, and receiving

the ball as a non-server would have a positive effect on momentum. Since players Carlos Alcaraz and Novak Djokovic are high-level athletes who rarely make low-level errors such as double faults during matches, these factors are not considered.

Psychological factors: high-level athletes, not only have high physical ability and high skills but also have better psychological factors. After analyzing, we judged those factors such as whether there is a mandatory error, the total points obtained, and whether the ball struck is effective or not will have an impact on the momentum[8][9].

Figure 1: Factors affecting momentum

Following the selection of the data set, several factors were taken into consideration and the data underwent processing via a random forest classification algorithm. This enabled the assignment of weights to each factor affecting momentum, with the specific weights presented in Figure 2.

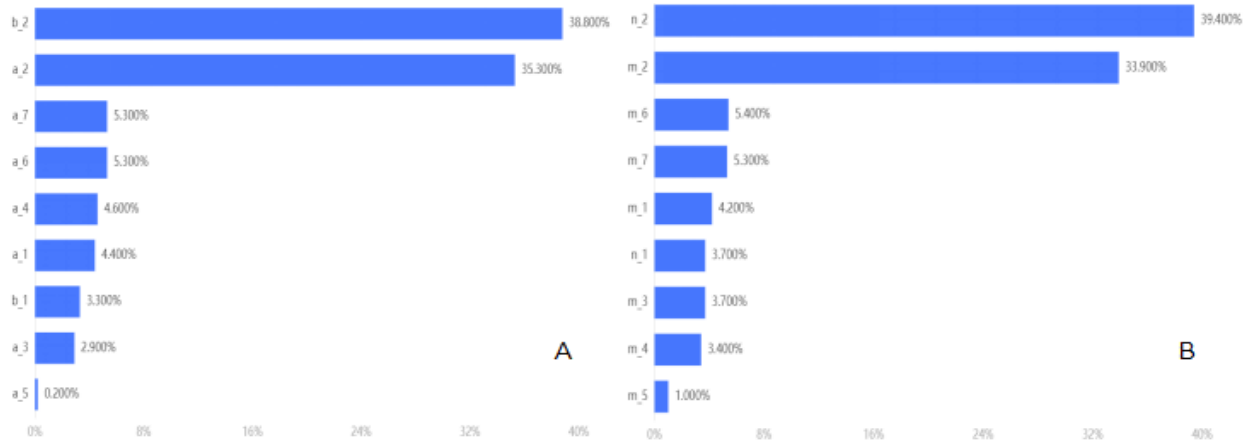

Figure 2: Quantification of the effect of factors on the momentum of different players. (A) Carlos Alcaraz. (B) Novak Djokovic.

The analysis revealed that the distance run and the total number of points scored after winning the point were the primary factors influencing Carlos Alcaraz and Novak Djokovic. The impact of being a scoring player or not winning the set on the serve or not, and hitting unreturnable shots or not on Carlos Alcaraz's and Novak Djokovic's respective momentum was found to be less significant.

3.2. Establishment of Model

3.2.1 Changes in Momentum

We categorize the factors into positive impacts (ω_i) and negative impacts (c_i) based on the different weights and directions of their impacts on momentum.

For player Carlos Alcaraz:

$$M_1 = \sum_{i=1}^{7} \omega_i a_i - \sum_{i=1}^{2} c_i b_i \tag{2}$$

For player Novak Djokovic:

$$M_2 = \sum_{i=1}^{7} \omega_i m_i - \sum_{i=1}^{2} c_i n_i \tag{3}$$

By quantifying the momentum, the performance of each player can be judged.The fluctuation of their momentum is shown in Figure 3: although the momentum of Carlos Alcaraz and Novak Djokovic often fluctuates, overall, the momentum of both Carlos Alcaraz and Novak Djokovic increases over time, and the momentum of Carlos Alcaraz fluctuates more compared to the momentum of Novak Djokovic.

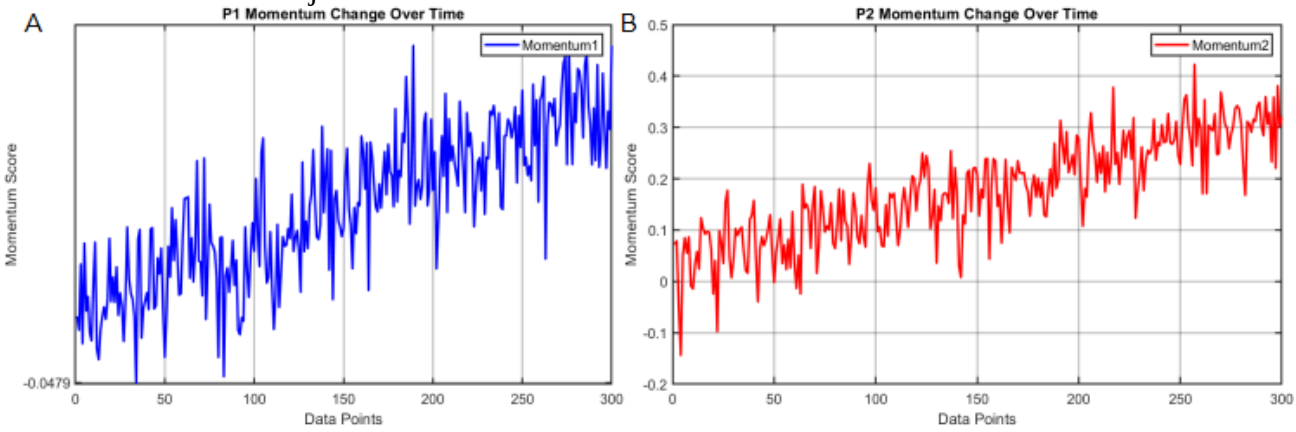

Figure 3: Visualization of match momentum over time. (A) Carlos Alcaraz. (B) Novak Djokovic.

3.2.2 Difference in Momentum

To get the difference between the change in momentum over time of Carlos Alcaraz and Novak Djokovic:

$$\Delta M = M_1 - M_2 \tag{4}$$

The difference in their momentum is shown in Figure 4:

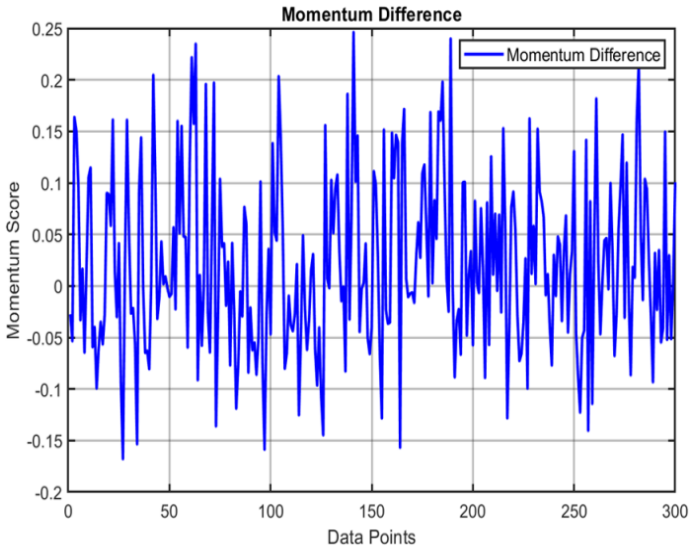

Figure 4: The momentum difference between Carlos Alcaraz and Novak Djokovic

3.3. Analysis of The Result

To determine which athlete performed better in a particular time, we divided the time into small segments of 50 innings, and divided ΔM into two parts, positive and negative, to make the following histogram, which are shown in Figure 5:

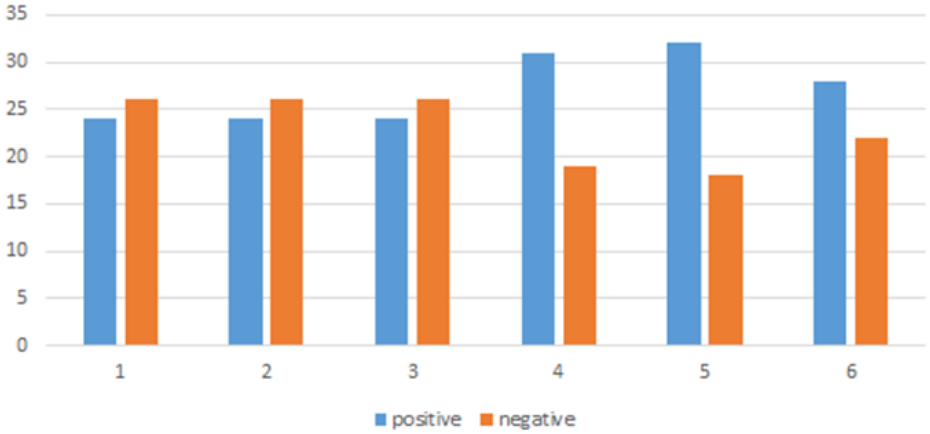

Figure 5: Comparison of the number of positive and negative values

In the first 150 games, Novak Djokovic's momentum was greater overall than Carlos Alcaraz's, so Novak Djokovic performed better compared to Carlos Alcaraz. And in the last 150 games, Carlos Alcaraz's momentum was greater overall than Novak Djokovic's, so Carlos Alcaraz performed better compared to Novak Djokovic.

4. Prediction and Analysis of Momentum

4.1. Establishment of Model

4.1.1 Back Propagation Neural Network

Back Propagation Neural Network is a common neural network structure used to solve classification problems. BP Neural Network is a supervised learning model that adjusts the weights in the network, through a backpropagation algorithm, to minimize the error between the predicted output and the actual samples. There are many common classification models in the BP Neural Network. These include a multilayer perceptron, which is a model that usually consists of an input layer, a hidden layer, and an output layer. Each neuron is connected to all the neurons in the previous layer and each connection has a weight. BP Neural Network computes the output by forward propagation and then uses a backpropagation algorithm to update the weights to minimize the prediction error[10].

By using a Back Propagation Neural Network classification model, we predicted the score of each sample point of Daniil Medvedev in the match between Daniil Medvedev and Marton Fucsovics at Wimbledon in 2023, based on the weight of each influencing factor on each sample point of Carlos Alcaraz's loss or win[11].The results of their predictions are shown below:

The prediction results are shown in Table 1 and the algorithm parameters are shown in Table 2:

Table 1. The results of their predictions

	training set	test set
Accuracy	0.723	0.665
Recall	0.723	0.665
Precision	0.735	0.679
F_1	0.72	0.665

Table 2. Parameter

Parameter name	Parameter Value
Training time	0.067
Data Slicing	0.7
Data shuffling	clogged
Cross-validation	clogged
Activation function	identify
Solver	lbfgs
Activation rate	0.1
L_2 regular terms	1
Number of iterations	1000
Number of hidden first-layer neurons	100

4.1.2 Decision Tree Regression Algorithm

The Decision Tree Regression algorithm is a machine learning algorithm commonly used to solve regression problems. Like the Decision Tree Classification algorithm, the Decision Tree Regression algorithm constructs a tree-like structure by recursively partitioning the data to make regression predictions on input features. The algorithm makes predictions by learning patterns in the dataset and traversing along the branches of the tree based on the values of the features, eventually reaching a leaf node to obtain a regression prediction. The study used the algorithm to calculate the importance of the different feature values (i.e., the various factors affecting changes in momentum), i.e., their weights, which are shown in Figure 6:

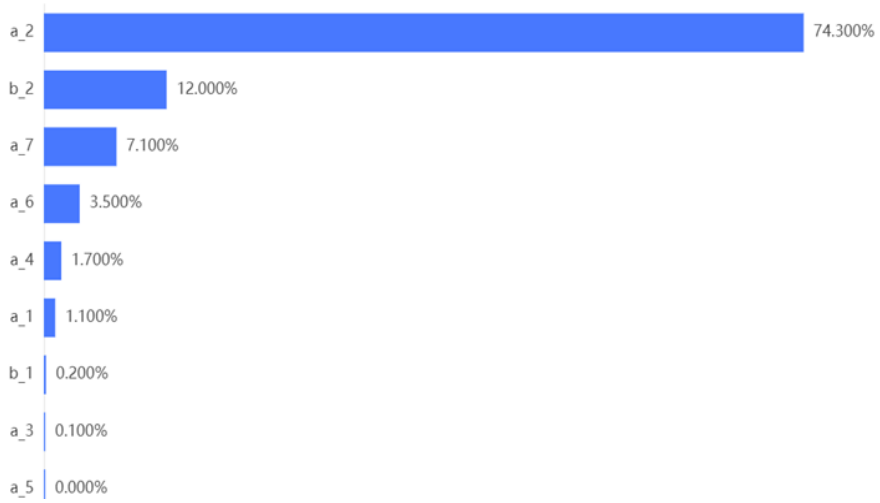

Figure 6: Impact of weighting of factors on momentum

It can be seen that the factors most relevant to the fluctuations in the game are the total number of points scored by the player per sample point (a_2), which has a weight of 74.3%, and the distance he or she runs (b_2), which is the second influencing factor, with a weight of 12%. There are other influencing factors, which are the ball's landing position (a_7), the speed of hitting the ball (a_6), and whether or not to hit an unreturnable ball (a_4), among others.

4.2. Analysis of The Result

We took the scores (denoted as x_1) obtained at 300 sample points in the Carlos Alcaraz tournament, defining a win as 1 and a loss as 0.

After analyzing the results, the predictions are shown in Table 3, where a portion of the data is extracted for interpretation:

Table 3. The results of their predictions

Forecast value	real value	Whether the prediction is correct
1	0	Wrong
0	0	Right
0	1	Wrong
0	0	Right
1	1	Right
0	0	Right
0	0	Right
1	1	Right
0	0	Right
0	0	Right
1	1	Right
0	1	Wrong

The essay extracted 12 data from 300 sample points. Among the 12 sample points, 9 sample points are correctly predicted and 3 sample points are incorrectly predicted, with an accuracy rate of 75%; among the 300 sample points, the accuracy rate is 66.5%. Since the number of 12 samples is very small, the accuracy rate is for reference only and the true accuracy rate should be based on 300 sample points.

4.3. Suggestions for the Competition

In the decision tree regression model, we get the main factors affecting the score of each sample point, which can be summarized as ability factor, psychological factor, and physical fitness factor. In the normal training of athletes, these three aspects can be emphasized.

In senior tournaments, the players' ability level is very comprehensive, but if their serves are not stable or aggressive, it is easy to put psychological pressure on themselves, leading to their errors. In a match, you can put psychological pressure on the opponent by increasing the speed of the ball and hitting the ball that the opponent cannot return. At the same time, when the game is anxious, stabilize their mentality, if they are on the serving side, stabilize their mentality when they are ahead, not anxious when they are behind, and prevent the appearance of double faults and other low-level error[12][13].

5. Conclusion

We analyzed the Wimbledon Open men's final and found that momentum played a key role in the outcome of the match. First, we built a momentum assessment model and processed the data by applying a random forest classification algorithm to determine the strengths and weaknesses of each player in the match, thus providing an effective means of visually displaying the momentum differences between players. Further, a decision tree regression algorithm was applied to calculate the weights of each factor leading to changes in scoring, and the analysis yielded that the most relevant factor to fluctuations in the game's momentum was the total number of points scored by the players per sample point, with a weight of 74.3%, and that their running distance was the second most influential factor, with a weight of 12%. Using the win/loss situation in the match between Carlos Alcaraz and Novak Djokovic as the training set and the win/loss situation in the match between Daniil Medvedev and Marton Fucsovics as the test set, the win/loss situation is predicted by the back-propagation neural network model and compared with the actual results, and the model accuracy is found to be more than 70%.

This paper provides a research idea and framework applied to the field related to sports analysis, and the model accuracy and generalizability of this study prove the feasibility of the method.

References

[1] Claudino J G, Capanema D O, de Souza T V, et al. Current approaches to the use of artificial intelligence for injury risk assessment and performance prediction in team sports: a systematic review[J]. Sports medicine-open, 2019, 5: 1-12.

[2] Maszczyk A, Gołaś A, Pietraszewski P, et al. Application of neural and regression models in sports results prediction[J]. Procedia-Social and Behavioral Sciences, 2014, 117: 482-487.

[3] Loeffelholz, Bernard, Bednar, Earl, Bauer, Kenneth W. Predicting NBA Games Using Neural Networks[J], Journal of Quantitative Analysis in Sports, 2009, 5(1): 7.

[4] McCullagh J. Data mining in sport: A neural network approach[J]. International Journal of Sports Science and Engineering, 2010, 4(3): 131-138.

[5] Xiangwei X, Lijuan M, Jinling H, et al. The Correlation between Physical Exercise and Comprehensive Health of Chinese Graduate Students: An Empirical Research Based on BP Neural Networks[J]. Global Education [2024-11-08].

[6] Jie W. The Resource-based Research-training-learning Platform Construction of Sports Techniques and Tactics[J]. Journal of Shanghai University of Sport, 2018.

[7] LV X C, GU D Y, LIU X H, et al. Momentum prediction models of tennis match based on CatBoost regression and random forest algorithms [J]. Scientific Reports, 2024, 14(1).

[8] MEIER P, FLEPP R, RUEDISSER M, et al. Separating psychological momentum from strategic momentum: Evidence from men's professional tennis [J]. Journal of Economic Psychology, 2020, 78.

[9] DEPKEN C A, GANDAR J M, SHAPIRO D A. Set-level strategic and psychological momentum in best-of-three-set professional tennis matches [J]. Journal of Sports Economics, 2022, 23(5): 598-623.

[10] LEI Y L, LIN A, CAO J N. Rhythms of Victory: Predicting Professional Tennis Matches Using Machine Learning [J]. Ieee Access, 2024, 12: 113608-17.

[11] YU H, JIA L Y, MIAO Y M, et al. CLSTM: A Dynamic Model for Predicting Winning Percentages in Tennis Matches; proceedings of the 5th International Conference on Computing, Networks and Internet of Things (CNIOT), Tokyo, JAPAN, F May 24-26, 2024 [C]. 2024.

[12] LI X H, LIU C, ZHANG Y, et al. Momentum Prediction Model: Catch the Flow and Swing; proceedings of the International Conference on Modeling, Natural Language Processing and Machine Learning (CMNM), Xian, PEOPLES R CHINA, F May 17-19, 2024 [C]. 2024.

[13] LEI Y L, LIN A, CAO J N. Rhythms of Victory: Predicting Professional Tennis Matches Using Machine Learning [J]. Ieee Access, 2024, 12: 113608-17.