SJIF Impact Factor (2024): 8.675| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 9 | Issue: 4 | April 2024 - Peer Reviewed Journal

A RANDOM FOREST-BASED MODEL OF SCORE FLUCTUATIONS IN PROFESSIONAL TENNIS MATCHES

Yanqi Zhang[1], Jie Zhang[2], Mingxu Zhou[3], Liu Tao[4]
Dr. Hatem Hassanin[5]

[1] *Student at the College of Architecture and Civil Engineering, Xiamen University*
[2] *Student at the College of Information, Xiamen University*
[3] *Student at the College of Aeronautics and Astronautics, Xiamen University*
[4] *Student at the College of International, Hunan University of Arts and Sciences*
[5] *Professor, International College, Hunan University of Arts and Science*
[5] *orcid.org/0000-0001-7571-5519*

Article DOI: https://doi.org/10.36713/epra16361
DOI No: 10.36713/epra16361

ABSTRACT

In tennis matches, the victory and turning points of the game are often influenced by various factors. To explore the factors that affect match fluctuations (changes in the flow of scoring) and to provide suggestions for athletes' match strategies, this paper first identifies general indicators through a literature review and uses logistic regression to determine the effectiveness of the chosen model. Secondly, it employs the Fourier function fitting to identify turning points in the match. Considering the scarcity of turning points in the game, this paper uses the SMOTE method to expand the dataset. Subsequently, it tests with a random forest classification model, achieving an accuracy of 93.433%. To improve the model's accuracy, several indicators were added to the original model, resulting in a correct rate of 98.51%. Finally, to verify the model's results and applicability, the model was applied to other matches with good results. A sensitivity analysis was conducted, revealing good model stability. The model results indicate that the main factors affecting the appearance of turning points include the player's movement distance during the match, whether there are changes in the depth and width of the return, score differences, and the maximum number of consecutive wins. When tested in other types of matches, we found that the importance of these factors may change to some extent, but the results remain satisfactory.

KEYWORDS: *Volatility prediction, Random Forest, Logistic regression, Sensitivity analysis.*

1- INTRODUCTION

Tennis is known as the "second-largest global ball sport," a reputation earned for its intense competition and elegant techniques. It enjoys widespread popularity and a long history around the world. In recent years, the professionalization of tennis has accelerated, bringing professional tennis events into the broader public view. Studying the winning factors in professional tennis matches is of significant importance. Scholars have adopted a variety of perspectives and methods to analyze the winning patterns in professional tennis. Some researchers have summarized the winning patterns in competitive tennis by studying the characteristics of tennis matches and the training and competitive experience of athletes [1]; others have conducted comparative analyses on technical and tactical indicators by statistically analyzing open tennis match data, thereby highlighting the importance of technical and tactical factors in professional tennis matches [2]; and some have selected relevant indicators to establish models for analysis, thereby identifying the key factors that affect match outcomes [3].

A review of the research results indicates that while there are many studies on the winning patterns in professional tennis, models that reflect the flow of scoring and the dominance of athletes within smaller time spans in a match are relatively lacking. In fact, in the fierce competition of professional tennis matches, it is rare for an athlete to maintain an advantage throughout the end of the match. Instead, both athletes drive the match forward along a dynamic curve of alternating dominance. This dynamic curve can be referred to as the athlete's momentum curve, reflecting the real-time flow of scoring in a match. It is reasonable to hypothesize that

SJIF Impact Factor (2024): 8.675| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 9 | Issue: 4 | April 2024 - Peer Reviewed Journal

there are multiple turning points in a match, within which the higher probability of winning points shifts from one athlete to another over a small range of turning points. Research on indicators affecting the occurrence of turning points not only allows for reasonable prediction of point outcomes but also provides a scientific approach for athletes to gain an advantageous position in response to events that can affect the flow of scoring during the match.

After summarizing a large amount of literature, this study selects and analyzes indicators affecting turning points from three directions: psychological factors, physical factors, and strategic factors, as independent variables. It then applies the Fourier function fitting to the data of each point in the thirty-one matches following the second round of the men's singles at the 2023 Wimbledon Tennis Championships. The obtained turning points are used to expand the dataset and form the dependent variables, establishing a model to analyze and identify the key factors affecting match turning points. The Wimbledon Men's Tennis Open, as one of the world's top professional tennis events, provides a scientifically robust data source for studying professional tennis matches. Researching the factors that influence turning points also offers reference significance for coaches and athletes in deciding what strategies to adopt when facing events that can impact the flow of scoring during matches.

2.1 MODEL INTRODUCTION

2.1.1 FOURIER FUNCTION FITTING TO IDENTIFY TURNING POINTS

In sports matches, the accurate identification of turning points is crucial for understanding the rhythm and dynamics of the game. To capture these critical moments, we employed the Fourier function fitting method to analyze the scoring changes in the match. This approach reveals the periodic patterns of score differences over time, thereby identifying potential turning points.

We first constructed a time series model for the score difference in the match, which can be represented as:

$$S(t) = a_0 + \sum_{n=1}^{N} \left(a_n \cos\left(\frac{2\pi\, nt}{P}\right) + b_n \sin\left(\frac{2\pi nt}{P}\right)\right)$$

In the model, $S(t)$ represents the score difference at time point t, a_0 is the constant term, a_n and b_n are the Fourier coefficients, N is the order of the Fourier series, and P is the period, which reflects the repetitive pattern of score changes in the match.

To find the extremum points of the score difference curve, we need to solve the first-order derivative of $S(t)$ and find its zeros. The first-order derivative $S'(t)$ is expressed as:

$$S'(t) = \sum_{n=1}^{N} \left(-\frac{2\pi n a_n}{P} \sin\left(\frac{2\pi nt}{P}\right) + \frac{2\pi n b_n}{P} \cos\left(\frac{2\pi nt}{P}\right)\right)$$

Extremum points are the points where $S'(t)$ equals zero, that is, $S'^{(t)} = 0$. By solving this equation, we can locate the potential turning points in the match.

2.1.2 FEATURE INDICATOR EXTRACTION

To gain a comprehensive understanding of player performance and match dynamics, after reviewing literature [5][6][7], we have proposed a set of integrated indicators that cover three dimensions: psychological, physiological, and strategic. The psychological factor indicators include ranking points, the maximum number of consecutive victories, the number of ACEs, the number of winners, and the number of double faults. Physiological factor indicators, such as the distance covered at the net, serve as quantitative indicators of the player's running distance, which is related to recent physical exertion and its impact on scoring outcomes. Strategic factor indicators, including changes in serve width and depth, provide strategic insights into the variations in a player's serving

SJIF Impact Factor (2024): 8.675| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 9 | Issue: 4 | April 2024 - Peer Reviewed Journal

patterns, which affect the effectiveness of attacks and create opportunities based on the opponent's positioning. The selection and rationale for these indicators are displayed in Table 1.

Table no 1: Indicator selection and rationale

Variable	Explanation	Rationale
Point	The point number in the game	When a new game starts, both players have a higher frequency of consecutive balls, and the competition for the key points will greatly affect the future direction of the game.
Consecutive victories	Number of consecutive victories	Winning points within the first five goals, the value of which we believe has a large impact on a player's "momentum" and is important to the flow of scoring in a game.n
Aces	Number of aces in the previous five balls	Aces can be a great morale booster for your team. In the short period of time following an ace, a player's "momentum" increases and the probability of winning the match increases.
Winners	Number of winners in the previous five balls	Many "strikes" in the previous five balls. As with aces in the previous five goals, this indicator has a significant impact on increasing a player's "momentum";
Double fault	Number of double faults in the previous five ball	Number of "double faults" within the first five balls. The number of "double faults" will interrupt the player's rhythm and reduce the player's "momentum". We believe that "double faults" have a large impact on players' psychological factors.
Distance difference	Difference in running distance in previous points	The difference in distance run by players within the first five goals reflects the re- cent physical exertion of players, which is closely related to the current point winners and losers.
Change width & change depth	The width of the stroke and depth of the stroke changes	Changes in the width and depth of a server's serve reflect changes in a player's game strategy.

2.1.3 LOGISTIC REGRESSION TO VALIDATE THE EFFECTIVENESS OF INDICATOR SELECTION

In the process of building a predictive model, selecting the right feature indicators is crucial for enhancing the model's forecasting performance. To ensure that the indicators we have chosen can effectively reflect the turning points in a match, we have employed logistic regression analysis to verify their relevance and significance.

Logistic regression, as a widely used statistical method, is suitable for binary classification problems and helps us evaluate the impact of each feature indicator on the prediction outcome. The form of the model is as follows:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

In the model, p represents the probability that a sample is a turning point, β_0 is the intercept of the model, β_1, β_2, ..., β_n are the coefficients of the feature variables, and X_1, X_2, ..., X_n are the selected feature indicators.

We determine the coefficients \beta for each feature indicator through logistic regression analysis. These coefficients reflect the degree to which each feature variable affects the probability of a turning point occurring. Indicators with a significance level P-value less than 0.05 are considered to significantly influence the prediction of turning points and are retained in the final model.

Through this process, we not only validate the effectiveness of the selected indicators but also enhance the accuracy and reliability of the model in predicting match turning points. This lays a solid foundation for subsequent random forest classification predictions, ensuring that our model can capture key dynamic changes in the match.

2.1.3 USING SMOTE TO EXPAND THE DATASET

Considering that there are fewer fluctuation points and more non-fluctuation points in the dataset, leading to an imbalance in the data, the SMOTE method can balance the distribution of categories by adding samples from the minority class, reducing the risk of

SJIF Impact Factor (2024): 8.675| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 9 | Issue: 4 | April 2024 - Peer Reviewed Journal

overfitting while improving the model's generalization ability [4]. We use the SMOTE method to expand the dataset. The steps for SMOTE sample synthesis are as follows:

- Randomly select a positive sample from fluctuation points: $X_{positive}$.
- Use the KNN algorithm to find K nearest neighbors of the positive sample: $\{X_{neighbor,1}, X_{neighbor,2}, \dots, X_{neighbor,K}\}$.
- Randomly select a neighbor: $X_{selected}$.
- Synthesize a new sample using interpolation: $X_{synthetic} = X_{positive} + r(X_{selected} - X_{positive})$, where r is a random number betwee61n 0 and 1. By using the SMOTE method to expand the dataset, our dataset has expanded from 7285 to 17246, with the proportion of fluctuation points increasing from less than 1/60 of the original data to nearly 1/3.

2.2 RANDOM FOREST CLASSIFICATION PREDICTION

Random Forest builds on the Bagging ensemble constructed with decision trees as base learners, further introducing the random selection of attributes during the training process of decision trees. The Random Forest algorithm is simple, easy to implement, and has low computational overhead, showing strong performance in many practical tasks [8]. Here are the specific steps on how Random Forest works:

- Data Resampling

For the original dataset D, perform B times of bootstrap sampling, generating a new dataset D_b each time.

$$D\_b = \{(x\_i, y\_i) \mid x\_i \sim D, y\_i \sim Y, i = 1, \dots, n\}$$

where n is the number of samples, x_i is the feature vector, and y_i is the corresponding class label.

- Building Decision Trees

For each dataset D_b, construct a decision tree T_b. When selecting the splitting attribute at each node, choose the best splitting attribute from m randomly selected features.

The optimal splitting attribute A_{opt} can be selected by minimizing the Gini impurity loss:

$$A_{opt} = \arg\min_{A \in \{m\}} \Delta(A)$$

Where $\Delta(A)$ is the reduction in Gini impurity brought by attribute A.

- Collective decision-making:

For a new sample x, let it pass through each decision tree T_b to get B prediction results $\widehat{y_b}$.

The final classification $\hat{y}$ is determined by majority voting:

$$\hat{}\{y\} = "\{mode\}\{\{\hat{}\{y\}\}\_b \mid b = 1, \dots, B\}$$

Through this method, the Random Forest model can effectively predict turning points in tennis matches, providing valuable insights for coaches and analysts. The assessment of feature importance and model cross-validation further enhances the model's predictive power and generalizability.

3 MODEL RESULTS AND VALIDATION

3.1 MODEL PREPROCESSING RESULTS:

In the preprocessing stage, we first use a multivariate Fourier function to identify the extreme points in the game score, which symbolizes the transformation of the game momentum. The fitted Fourier function image shows the direction and turning point of the score(Figure 1)

SJIF Impact Factor (2024): 8.675| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 9 | Issue: 4 | April 2024 - Peer Reviewed Journal

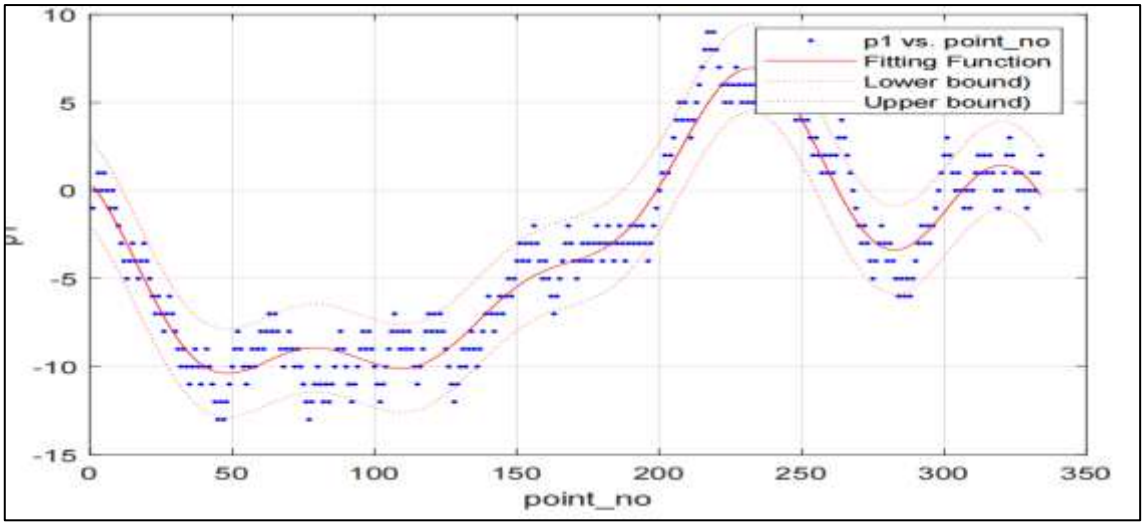

Figure 1: Fourier function fitting diagram

Furthermore, in order to address the issue of relatively scarce turning point data, we adopted the Synthetic Minority Class Over sampling technique (SMOTE). After SMOTE processing, the dataset was expanded from 7285 samples to 17246 samples, and the proportion of fluctuation points increased from less than 1/60 of the original data to nearly 1/3.

3.2 DISPLAY OF LOGISTIC REGRESSION RESULTS

Table 2 shows the evaluation indicators of the model, which can be used to evaluate the performance or validate the effectiveness of the model. The P-value is analyzed, and the value is less than 0.05, indicating that the model is effective. Table 3 shows that the significance P-value of each indicator is 0.000, showing significance at the horizontal level, rejecting the null hypothesis. Therefore, the indicators will have a significant impact on the volatility point. The selection of indicators is effective.

Table no 2: Biclassified logistic regression results

likelihood ratio chi-square value	P	AIC	BIC
18464.314	0.000	18482.314	18552.111

Table 3: Biclassified logistic regression results

Reg	StdErr	Wald	P	OR	Upper- lim	Lower- lim
Constant 2.376	0.069	1175.161	0.000	10.759	9.393	12.325
Point -0.128	0.008	280.252	0.000	0.88	0.867	0.893
Consecutive -0.369 victories	0.018	403.491	0.000	0.691	0.667	0.717
Aces -0.564	0.04	200.417	0.000	0.569	0.526	0.615
Winners -0.387	0.02	377.035	0.000	0.679	0.653	0.706
Double fault -1.117	0.064	307.259	0.000	0.327	0.289	0.371
Distance difference -0.011	0.002	39.023	0.000	0.989	0.986	0.993
Change width -0.885	0.039	523.046	0.000	0.413	0.382	0.445
Change depth -0.877	0.04	488.029	0.000	0.416	0.385	0.45

3.3 RANDOM FOREST PREDICTION RESULTS

The random forest algorithm was used for binary classification prediction, and the model demonstrated excellent predictive ability. On the dataset optimized and expanded by SMOTE, the classification accuracy of the random forest reached 91.39% (Figure

SJIF Impact Factor (2024): 8.675| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 9 | Issue: 4 | April 2024 - Peer Reviewed Journal

3). The confusion matrix of the model further proves its predictive performance, with a true positive rate of 93.43% and a false positive rate of 6.57% (Figure 4)

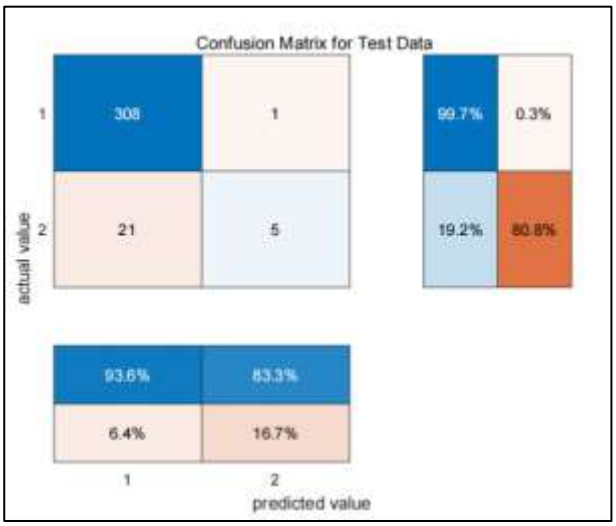

Figure no 2: Confusion matrix

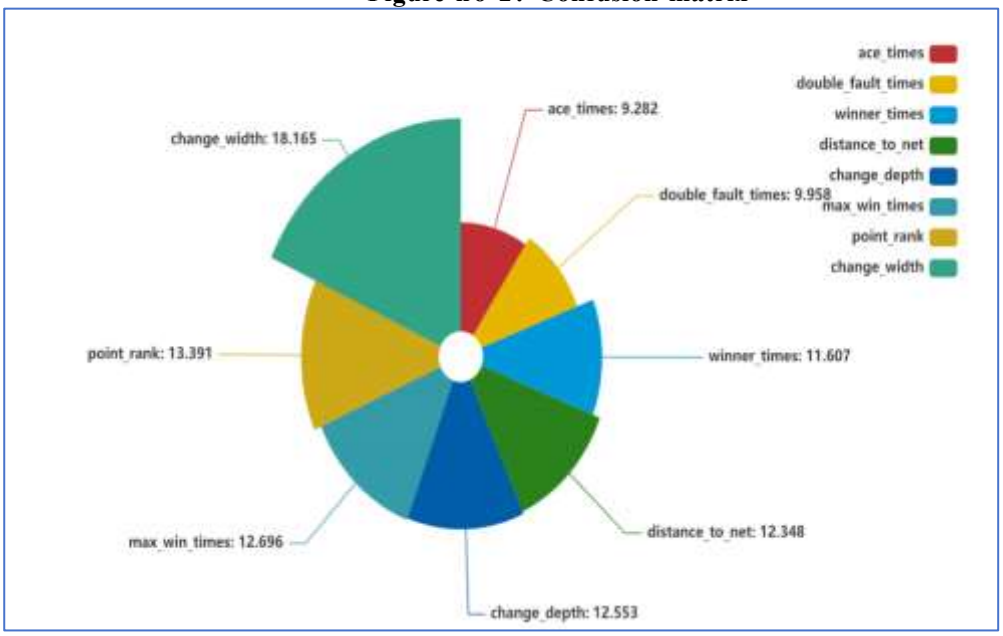

Figure 3: The weight of factors affecting state

We noticed turning point prediction inaccuracies. To refine the model, we strategically added key indicators like first-serve success rate, serve speed, and break rate. Considering the significance of the seventh set, we incorporated its outcome. Additionally, player metrics such as height and weight were included for a comprehensive approach. To maintain model specificity, we introduced three new indicators: "number of successful first five serves"、" last set break" and "tiebreak."

The turning points of the final data are also selected for prediction, and the RF classification accuracy of the optimized model reaches 98.51%, and the prediction effect is further improved. From the confusion matrix, it can be seen that the frequency of non-turning points predicted into turning points is greatly reduced, and the accuracy of the model is optimized, increasing its realistic practicality.

SJIF Impact Factor (2024): 8.675| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 9 | Issue: 4 | April 2024 - Peer Reviewed Journal

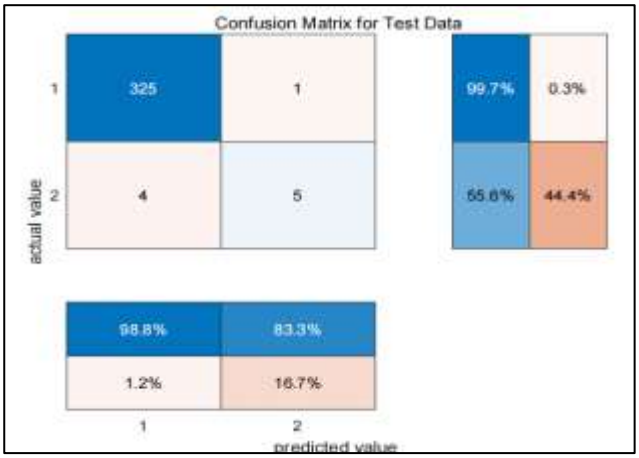

Figure 4: Confusion matrix

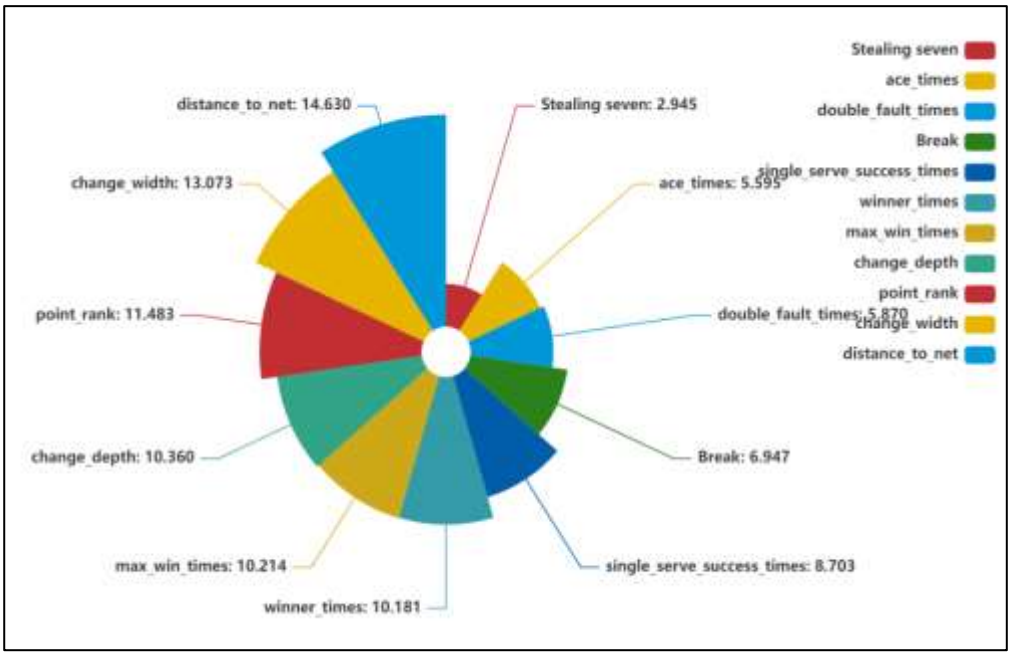

Figure 5 : The weight of factors affecting state

3.4 MODEL TESTING

Extensive testing:

The widespread applicability of the model was validated by applying it to data from the 2020 US Open women's singles tournament and the 2021 Wimbledon Championship. In these different competitions, the prediction accuracy of the model reached 90.1% (see Figure 6) and 97.4% (see Figure 7), respectively, demonstrating the model's good generalization ability in different competition environments.

SJIF Impact Factor (2024): 8.675| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 9 | Issue: 4 | April 2024 - Peer Reviewed Journal

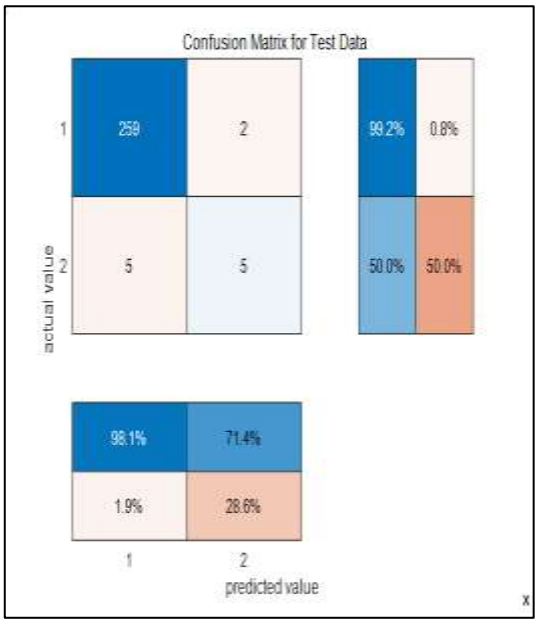

Figure 6: Confusion matrix

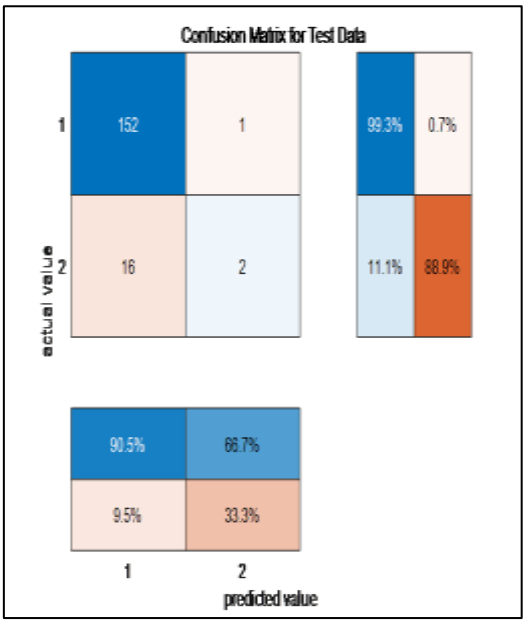

Figure 7 : The weight of factors affecting the state

Sensitivity testing

To verify the sensitivity of the model, we added 100 randomly generated perturbation data to the training set of the random forest classification model and re predicted the 23 year Wimbledon final. The results showed that even with the addition of perturbed data, the prediction accuracy of the model remained at 97.3% (see Figure 9), indicating that the model has high robustness to small changes in the data.

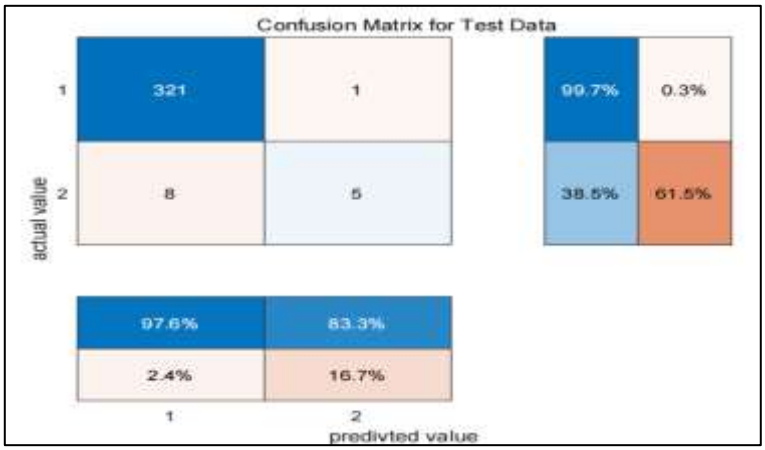

Figure 8: Confusion matrix after adding perturbed data

4. CONCLUSION

According to research, as the game progresses, factors such as the number of consecutive winning balls, ace balls, number of missed balls, difference in running distance, and changes in the depth and width of serve have a significant impact on game fluctuations in the first five goals. The changes in the width and depth of the serve have a significant impact on the turning point, which in turn drives the change in the direction of the game's score flow towards the target. There is a certain correlation between the flexibility of the strategy of changing the serving position and the high probability of the serving player winning the match point/match in tennis matches. Players can fully mobilize their opponents to attack their weaknesses and consume their physical energy by changing the serving position; You can also choose to play ace balls to enhance your attack and suppress opponents to maintain continuous scoring. Similarly, when serving the opponent, a strong and powerful serve should be played to gain the better initiative, and the overall strategy should be adjusted to consume reasonably, in order to avoid the opponent scoring continuously. Research has found that the number of double-serve errors in the top five psychological indicators has a significant impact on the progress and results

SJIF Impact Factor (2024): 8.675| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 9 | Issue: 4 | April 2024 - Peer Reviewed Journal

of the game. When athletes make double serve errors or other events during the game, they should strive to adjust their mentality and maintain a good competitive state.

The above analysis has reference significance for coach and athlete training. Excellent physical fitness is the foundation for formulating technical and tactical strategies, as well as an important factor affecting the progress and results of competitions. It is also the foundation for maintaining a good psychological state in competitions; Good psychological qualities can enable athletes to adopt appropriate strategies in events that affect the progress of the competition, such as double mistakes or consecutive scoring by the opponent.

5. REFERENCES

1. *Li Qingyou, Chen Zheng. Experience and winning factors of historic breakthroughs in China's women's tennis [J]. Chinese Sports Coaches, 2007 (01): 44-46*
2. *Jiang Hongwei. Analysis of Winning Factors for Jiangsu Provincial Men's Tennis Team in the 11th National Games [J]. Journal of Nanjing Institute of Sports (Social Science Edition), 2009,23 (04): 12-16. DOI: 10.15877/j.cnki. nsic.2009.04.024*
3. *Luo Weiquan, Zhang Lei. Research on the Model Construction of Winning Factors for Professional Tennis Athletes [J]. Journal of Guangzhou Institute of Sports, 2020,40 (03): 78-81. DOI: 10.13830/j.cnki.cn44-1129/G8.2020.030.21*
4. *Wang Xiaoxia, Li Leixiao, Lin Hao A Review of SMOTE Algorithm Research [J/OL]. Computer Science and Exploration: 1-29 [2024-03-30] http://kns.cnki.net/kcms/detail/11.5602.tp.20231108.1334.006.html.*
5. *Zhang Rong Prediction of Tennis Competition Results and Analysis of Players [D]. Yunnan University 2023. DOI: 10.27456/d. cnki. gyndu.2022.000807*
6. *Huang Zhiying Research on Tennis Match Score Prediction Model Based on BP Neural Network [D]. Fujian Normal University, 2022. DOI: 10.27019/dcnki. gfjsu.2022.000427*
7. *Shen Ye Key winning techniques in men's professional tennis based on logistic regression model* ***Empirical analysis of surgical indicators [J] Journal of Anhui Normal University (Natural Science Edition), 2017, 40 (6): 601-604***
8. *Lv Hongyan, Feng Qian. Review of Random Forest Algorithm Research [J]. Journal of Hebei Academy of Sciences, 2019,36 (03): 37-41. DOI: 10.16191/j.cnki. hbkx.2019.03.005*

Yanqi Zhang is currently studying at Xiamen University's School of Architecture and Civil Engineering, majoring in Civil Engineering.

Jie Zhang is currently studying at Xiamen University's School of Information Technology, majoring in Computer Science and Technology.

Liu Tao is currently studying at Hunan University's School of International, College of Information Management and Information Systems.

Mingxu Zhou is currently studying Electrical Engineering and Automation at Xiamen University College of Aeronautics and Astronautics.