Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 11, No. 3, 2024

Study of Tennis Match Momentum based on Random Forest and BP Neural Network Models

Yifu Wang [1, *], Yiling Yang [2]

[1] School of Electrical Engineering, Northeast Electric Power University, Jilin, China
[2] School of Economics and Management, Chongqing Jiaotong University, Chongqing, China
*** Corresponding author:** Yifu Wang (Email: wangyifu0926@gmail.com)

Abstract: The aim of this study is to scientifically analyse the "momentum" of a tennis match by developing a mathematical model. Data preprocessing techniques, such as K-nearest neighbour interpolation, were used to deal with missing values in dynamic matches. By filtering key metrics through principal component analysis and constructing a random forest model, we are able to quantitatively assess player performance. Using BP neural network models and genetic algorithms to optimise weights and biases, we improve the accuracy of match trend prediction. The analyses showed the significant influence of match time and player status on match trends. These research results provide a basis for scientific and data-based tennis training, which can help coaches conduct post-match assessment and optimise training programmes to promote player performance improvement.

Keywords: Principal Component Analysis; Random Forest; Backpropagation Neural Networks.

1. Introduction

In today's competitive sports, sports performance assessment and training optimisation have become key factors for success. Tennis is a very challenging and competitive sport, and the performance of athletes is not only affected by their individual skill level, but also by a number of factors, including the momentum changes during the game. Based on the scientific and data-oriented method, this study explores the "momentum" in tennis. Through data pre-processing and mathematical modelling, we attempt to reveal the potential impact of momentum on match results, and provide coaches with more scientific post-match evaluation and training optimisation solutions. Through this study, we hope to provide more effective and forward-looking guidance for tennis players' training and competition, and promote the understanding and application of data-driven decision-making in the whole sports community [1].

2. Analysis and Modelling

2.1. Status of Player Performance during the Match

2.1.1. Missing Value Handling

In this paper, we divide the five columns of data into two groups, one group of numerical variables: rally_count and speed_mph, and one group of categorical variables: serve_width, serve_depth and return_depth. we adopt K-nearest neighbor interpolation (KNN) to deal with the numerical variables, and we adopt plurality interpolation to deal with the categorical variables. K-Nearest Neighbor Interpolation does not require specific assumptions about the distribution of the data and is therefore suitable for all types of complex distributions in real datasets. The method utilizes the local structure within the data to make predictions, and the interpolated values are often more reliable than the simple mean or median. The plurality interpolation method also makes no assumptions about the distribution of the data, and is especially advantageous when the data do not conform to a normal distribution. This method maintains data integrity and, by interpolating using values that already exist in the data set, it maintains the frequency distribution of the variable, a feature that is particularly applicable to the categorical variables covered in this paper [2].

The distributions of ally_count and speed_mph data before and after using K nearest neighbour interpolation were compared, as shown in Figures 1 and 2.

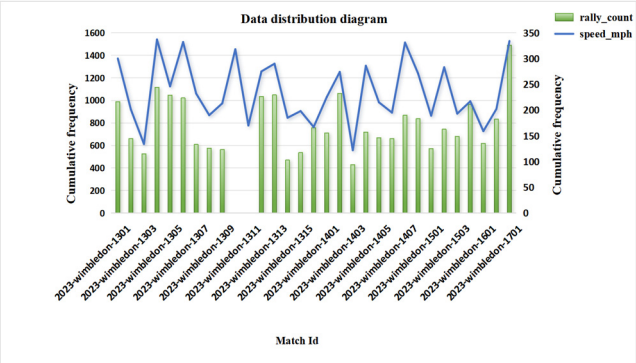

Figure 1. Before data cleansing

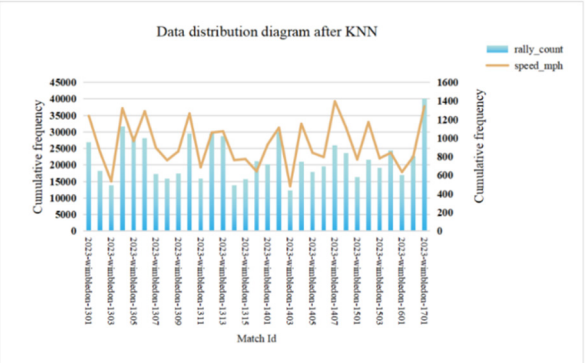

Figure 2. After data cleansing

In order to verify the reasonableness of the data after K-Nearest Neighbour interpolation, this paper uses residual analysis [3], which is a residual plot that checks whether any bias or outliers have been introduced by interpolating the data. This is shown in Figure 3 below:

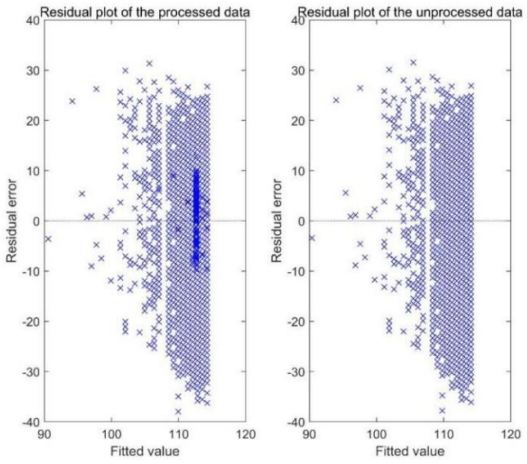

Figure 3. Residual plot

Comparing the two graphs, the residuals fluctuate up and down the horizontal line (the line with a residual of 0), and the distribution pattern is roughly the same. The residual values of the processed data fit are more concentrated in the 110 to 114 region, which indicates that interpolation improves the model fit in these regions. Data interpolation is reasonable.

Distribute serve_width, serve_depth, and return_depth data before and after using pattern interpolation, as shown in Figure 4 below:

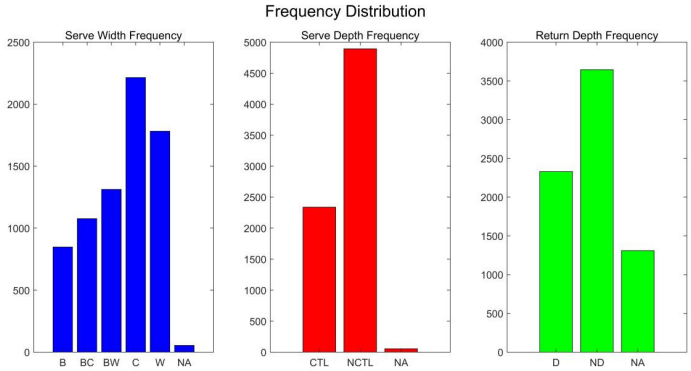

Figure 4. Before and after data distribution

In order to verify whether the data is reasonable after processing, the chi-square test is used, that is, whether there is a significant difference between the actual and expected frequency distributions. The SPSS software was used to prove that the data were not significantly different, so the method is reasonable to process the data [4].

2.1.2. Identify Relevant Metrics

There are many factors that reflect a player's performance on the playing field, such as the distance traveled on the court, the speed of hitting the ball, the depth of serve, etc. In this article, 15 characteristic indicators that reflect the performance of players are selected and specific definitions are given.

(1) Service Score Rate SPW (Service Points Won)

Definition: The service scoring rate is the percentage of the serving team in the service game.

$$SWP = \frac{A_{\text{ace}}+A_{\text{winner}}}{serve\_no} \quad (1)$$

Where A_{ace} represents 0 or 1 untouchable winning serve from player A in the authorities, A_{winner} means that player A has played one or 0 untouchable winning goals in the authorities; $serve\_no$ means the noth service game, and no can be equal to 1 or 2.

The server has an advantage in the match, and the score rate of the serve can be used as an indicator of whether the player is performing well in the service game based on the scoring rate of the serve.

(2) Unforced ErrorsRate (UER)

Definition: The frequency or proportion of unforced errors made by Player A during a match.

$$UER = \frac{A\_unf\_err}{A\_point\_won+B\_point\_won} \quad (2)$$

Where A_unf_err player A makes an unforced error, A_point_won denotes the number of points won by player A in the match, and B_point_won denotes the number of points won by player A's opponent player B in the match.

Through the unforced error rate, we can see the player's game mentality and environment, and the player's game mentality and environment will affect the player's performance, and the mentality and environment will directly affect the player's game performance.

(3)Average Serve Speed ASS

Definition: The average speed at which a player serves

$$ASS = \frac{sum(speed\_mph)}{serve\_no} \quad (3)$$

Where $speed\_mph$ represents the speed of the serve, and $serve\_no$ represents the first or second tee.

A fast serve has a speed advantage and is more likely to be a first-hand serve, which is a criterion for evaluating a player's performance.

(4)Physical endurance index SI

Definition: The ratio of the time spent in a game to the distance traveled by a player, assessing physical fitness and endurance.

$$SI = \frac{sum(A\_distance\_run)}{elapsed\_time} \quad (4)$$

Where $A\_distance\_run$ represents the distance that player A runs at the time of scoring, and $elapsed\_time$ represents the time from the start of the first score to the start of the current score.

Fitness and stamina affect the form of the players, which in turn affects the performance of the players.

(5) Serving Quality Index (SQI) (serve_quality_index)

Definition: The overall quality of an athlete's serve.

$$SQI = ASS \times SWP \quad (5)$$

where the average serving speed is denoted and the serving scoring rate is described.

The Serve Quality Index provides a visual indication of a player's performance in the game and can be used as an indicator to measure a player's game.

The remaining 10 characteristic indicators directly use the data score, victor, double_fault, break_pt_missed, return_depth, serve_depth, rally_count, point_no, serve_no, and point_victor in the original dataset.

These factors are linked, but not all of them have a significant impact on a player's performance. There are too many indicators of player characteristics, which will undoubtedly increase the difficulty and complexity of the analysis problem, and there may be a certain correlation between these multiple indicators. This is where principal component analysis can be used to identify salient indicators that reflect a player's performance.

The PCA calculation steps are as follows: Calculate the correlation coefficient matrix.

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ M & M & M & M \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix} \quad (6)$$

In the formula, $r_{ij}(\mathrm{i,j} = 1,2,\cdots,\mathrm{p})$ is the correlation coefficient between the original variable X_i and X_j , and its calculation formula is:

$$r_{ij} = \frac{\sum_{k-1}^{n}\left(x_{ki}-\bar{x_i}\right)\left(x_{kj}-\bar{x_j}\right)}{\sqrt{\sum_{k-1}^{n}\left(x_{ki}-\bar{x_i}\right)^2 \Sigma\left(x_{kj}-\bar{x_j}\right)^2}} \quad (7)$$

Because R is a real symmetric matrix (i.e.,$r_{ij} = r_{ji}$), only its upper or lower trigonometric elements need to be calculated.

Vegetarian is sufficient.

To normalize the original data, firstly, the original data is standardized, and the correlation coefficient matrix is calculated by the formula (4-7). Calculation of eigenvalues and eigenvectors.

First, the eigenvalue $|\lambda I - R| = 0$ is solved to find the eigenvalue $\lambda_i(\mathrm{i} = 1,2,\cdots,p)$ and arrange it in order of magnity, i.e., $\lambda_1 \geq \lambda_2 \geq \cdots, \geq \lambda_p \geq 0$, and then the eigenvector λ_1 corresponding to the eigenvalue $e_i(\mathrm{i} = 1,2,\cdots,p)$ is obtained respectively. The correlation coefficient matrix calculates the eigenvalues, as well as the contribution rate and cumulative contribution rate of each principal component. As can be seen from Table 1, the cumulative contribution rate of the first and second principal components is as high as 80.62%, so only the first and second principal components $\mathrm{z_1}$, $\mathrm{z_2}$ are required. Formula for calculating principal component contribution rate and cumulative contribution rate:

Principal component $\mathrm{z_i}$ contribution rate: $\mathrm{r_i}/\sum_{k=1}^{p}\gamma_k\,(\mathrm{i} = 1,2,\cdots,\mathrm{p})$ cumulative contribution rate: $\sum_{k}^{m}\gamma_k / \sum_{k}^{p}\gamma_k$.

Table 1. Eigenvalues and principal component contribution rates

Principal component	Eigenvalue	Variance contribution value	Cumulative contribution value
1	11.24	61.53	64.89
2	5.32	23.11	80.62
3	2.11	9.01	90.62
4	1.25	2.64	85.25
5	0.94	2.12	88.08
6	0.65	1.17	90.45
7	0.24	0.23	81.68
8	0.15	0.14	87.41
9	0.07	0.05	79.98
10	0.02	0.01	82.27

The contribution values and cumulative contribution values are visualised in Figures 5 and 6:

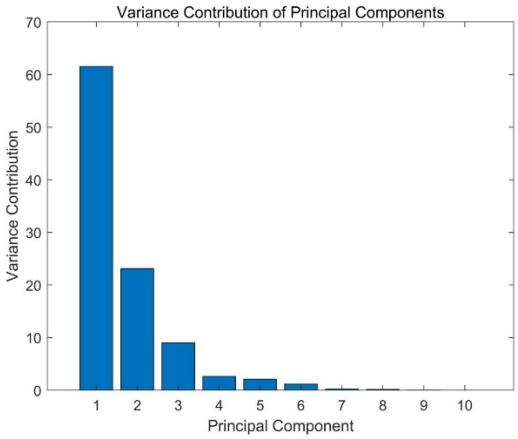

Figure 5. Variance contribution value

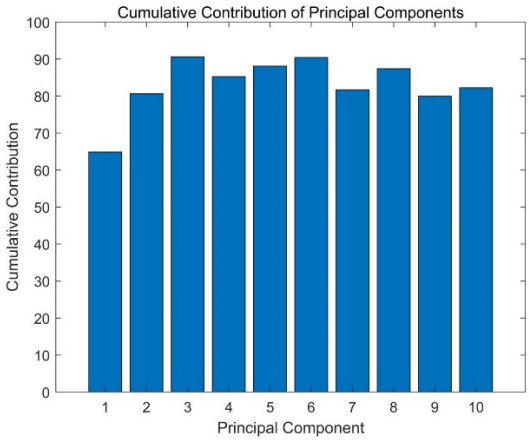

Figure 6. Cumulative contribution value

The variance contribution histogram shows that principal component 1 contains a lot of information about the characteristics of the performance players, and the cumulative contribution value histogram shows that most of them are greater than 80%, indicating that most of the principal components are sufficient to represent the characteristics of the players.

2.1.3. Random Forest Prediction Model based on Principal Component Analysis

Random forest model is an integrated learning algorithm based on decision trees [5]. Compared to single decision trees and traditional neural network algorithms, the random forest model has the advantages of strong interpretability, high computational efficiency and strong classification and regression performance. Random forests are inherently suited to handle high-dimensional datasets. Although PCA dimensionality reduction helps to reduce the number of features and mitigate the effects of dimensionality catastrophe, even after dimensionality reduction, the dataset may still retain complex structures and relationships. Random forests can effectively capture and model these complex nonlinear relationships, which are important for predicting a player's win rate per game.

After completing data preprocessing and feature extraction and CPA on the 15 indicator prediction datasets, the construction of the random forest model was started with the dimensionality reduced feature matrix $\mathrm{X_{n\times 10}}$ as the input data.

Step 1 Select MATLAB as the development model for predictive modeling.

Step 2 In the random forest model in this paper, the impurity weighting and $G(\mathrm{x_i}, \mathrm{v}_{ij})$ of each subnode are used to evaluate the quality of the sharding features and sharding points. The calculation formula is shown below.

$$G(x_i, v_{ij}) = \frac{n_{\text{left}}}{N_s} H(X_{\text{left}}) + \frac{n_{\text{right}}}{N_s} H(X_{\text{right}}) = \frac{1}{N_s}\left(\sum_{y_i X_{\text{left}}} (y_i - \bar{y}_{\text{left}})^2 + \sum_{y_i X_{\text{right}}} (y_i - \bar{y}_{\text{right}})^2\right) \quad (8)$$

Where X_i is the ith segmentation variable; is the segmentation value of the V_{ij} segmentation variable; n_{left}, n_{right}, and N_s represent the number of training samples of the left sub-node, the right sub-node, and the current node, respectively. $X_{\text{left}}, X_{\text{right}}$ represents the set of training samples of the left sub-node and the right sub-node, respectively. Represents the sample variable of the current node; where $\bar{y}_{\text{left}}$ represents the average value of the target variables of the samples of the left subnode and the right subnode, respectively.

Step 3 Perform random and repeated sampling of the row vectors of the dataset feature matrix $X_{n\times 10}$ to generate the sample matrix $X_{p\times 10}$, and the rest of the unselected data will be used as out-of-pocket data to increase the robustness and stability of the model.

Step 4 Among the Q column vectors in the sample matrix $X_{p\times 10}$, q column vectors are randomly selected as the optimal splitting node candidate ($q < Q$).

Solve for the q column vectors to minimize the impurity weighted and $\min G(\mathrm{x}, \mathrm{v})$, select the optimal segmentation variable x and the segmentation point v, generate two subnodes, and determine the corresponding output values.

Step 5 Repeat steps 2~4 to generate multiple single decision tree models until the maximum number of decision trees is reached, and then integrate all the generated single decision tree models to generate a random forest model.

Step 6 The prediction of the winning rate of each game of the player is a kind of regression problem, when the final output result is output, the model weights the prediction results of m decision trees on average, and then uses the obtained value as the model prediction result and outputs.

2.1.4. Solution of the Model

For a player's performance in a match, the measure is wins per set. The higher a player's win rate per set, the better his performance.

Individual differences in performance are large for different players. In this paper, we will start by analyzing two specific players. The following quotes are from the original dataset for Alejandro Davidovich Fokina and Holger Rune.

A player's winning percentage is defined as the number of points won by the player in the authority divided by the number of points won by both players in the game.

Figure 7 reflects a plot of the point in time at which each score was recorded versus the margin of victory in future matches.

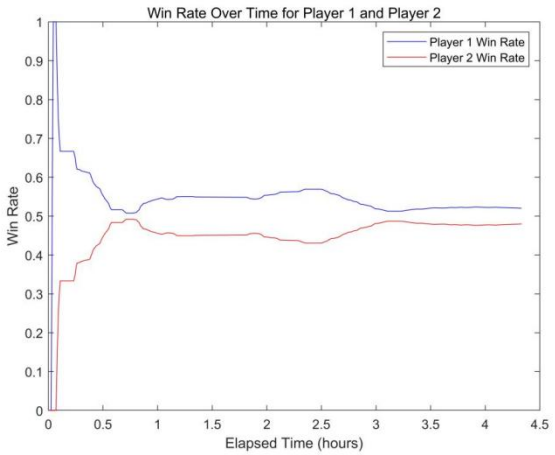

Figure 7. The relationship between the time of the score and the win rate

Player1 is Alejandro Davidovich Fokina, and Player2 is Holger Rune. By looking at the images, we can find that Player1's win rate peaks around 0.06(h), and then fluctuates around 0.5 as time increases. For Player2, his win rate rises first, then fluctuates around 0.5, peaking at around 0.7(h). Assessing the role of "momentum" in the game

Modeling of the Kruskal-Wallis H test

Before Kruskal-Wallis H test, it is necessary to carry out the homogeneity of variance test. The test process is as follows:

The hypothesis is made first, that is, the homogeneity of variance is satisfied. When the number of test repeats at each level is equal, that is:

$$m_1 = m_2 = \cdots = m_r = m \quad (9)$$

m is the number of repetitions at each level, which is the ratio of the maximum and minimum variances of r samples. Under the condition that the difference is equal, the H quantile of the H distribution can be obtained by random simulation. The distribution depends on the level number r and the degree of freedom $f = m - 1$ of the sample variance, so the distribution can be recorded as $H(\mathrm{r}, \mathrm{f})$. When H_0 is used immediately, there are:

$$\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_r^2 \tag{10}$$

The value of H should be close to 1, when the value of H is large, the difference between the parties will be large, the greater the value of H, the greater the difference between the parties, then reject the null hypothesis. It H_0 follows that for a given significance level, the rejection domain of test is H_0:

$$W_1 = \{\mathrm{H} > \mathrm{H}_{1-\alpha}(\mathrm{r},\mathrm{f})\} \tag{11}$$

The alternative hypothesis H_1 is that there is at least one pair of $\mu_i \neq \mu_j$.

Construct test statistics:

$$H = \frac{12}{N(\mathrm{N}+1)}\sum_{i=1}^{k} n_i(\bar{R}_i - \bar{R}) = \frac{12}{N(\mathrm{N}+1)}\sum_{i=1}^{k} n_i \bar{R}_i^2 - 3(\mathrm{N}+1) \sim \chi^2(\mathrm{k}-1) \tag{12}$$

2.2. Prediction of Game Changes and Establishment of Correlation Factors

2.2.1. GABP Prediction Model

When predicting changes in the flow of a tennis match, and in particular identifying potential momentum fluctuations, it is necessary to determine which factors are most relevant to these changes in the dynamics of the match. However, it is difficult to distinguish the real change of momentum from the random change in the game, and it is difficult for the simple BP neural network to accurately predict. Therefore, this paper optimizes the weight and bias of BP neural network by genetic algorithm, so as to improve the prediction accuracy of tennis match dynamics [6].

After optimizing the weight and bias of BP neural network, using genetic algorithm can significantly improve the performance of the network. By simulating the process of natural selection, genetic algorithms allow neural networks to explore a wider parameter space during training. This process not only reduces the risk of falling into local optimality, but also improves the network's ability to generalize to unseen data due to the global search nature of genetic algorithms. In addition, this method does not require gradient information and improves the optimization efficiency of the algorithm.

2.2.2. Correlated Factors are Established

The data entered the prediction model came from the characteristics of Roman Safiullin during the 2023-wimbledon-1503 tournament.

As shown in Figure 8, the comparison between the predicted value and the actual value of the model output.

By observing the predicted value and the actual value, the trend of the two curves is roughly the same, and the surface prediction effect is better. The following is the data of 2023-wimbledon-1306 to predict Tommy Paul's momentum swing, and to judge his on-court performance and final game results. Figure 9 is his momentum swing prediction chart.

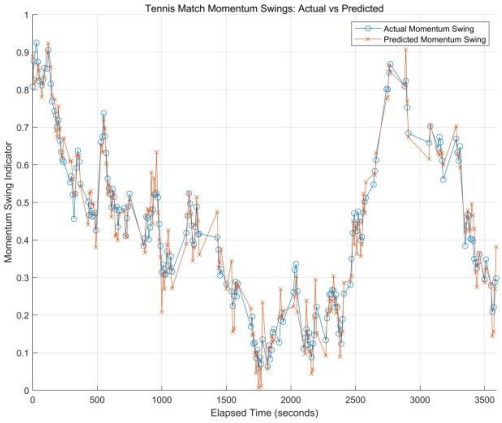

Figure 8. The model outputs a comparison of the predicted value with the actual value

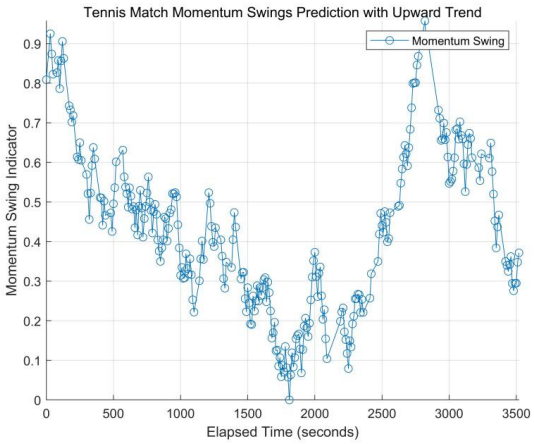

Figure 9. The model outputs a comparison of the predicted value with the actual value

By looking at the curve, the model predicts fluctuations in a player's performance, Like the player's performance on the field, and his final result was a loss, the result was consistent. As can be seen from the figure, his momentum swing indicator dropped rapidly in the late stage of the game, with large fluctuations, which had an impact on the result.

For the trained GABP neural network prediction model, the importance of input features can be evaluated by analyzing the weight of the hidden layer. A large weight means that the relevant features have a greater influence on the model output. This means that the greater the weight of the input features, the more relevant it is to momentum swing, which can be said to be the main factor to some extent. Table 2 is obtained from the weight analysis of the GABP neural network.

It can be seen from the above table that match_time has the largest weight, which is significantly different from the value of the second weight, so match_time can be considered as the main factor.

Table 2. Weight analysis of GABP neural networks

feature index	weight
match_time	13.01
score_condition	7.76
server	7.32
current_score	6.52
time_period	6.01
player_condition	9.50
score_gap	4.43
set_gap	8.36
service_game_count	5.99

According to the predicted player performance: Through weight analysis, the game time has the greatest impact on it, followed by the player's state. Therefore, it is suggested to improve the physical quality of players, reduce the sharp decline in physical state caused by long-term interaction with opponents, and pay attention to psychological construction during the game. Comprehensive analysis, maintaining a good mental and physical state is the key to winning.

3. Conclusion

Firstly, it scientifically analyses the "momentum" of tennis matches and its important influence on the match results. Secondly, the key indicators affecting players' performance were effectively screened, providing a more accurate assessment tool; the correlation between momentum and winning streak was explored in depth, improving the accuracy of match trend prediction.

Future research should expand the sample size, consider the effects of different venues and environments on momentum, and study the adaptability of different players to changes in momentum by combining actual game data and player characteristics. Optimizing the neural network model and algorithm can improve the prediction accuracy, deepen the understanding of the mechanism of momentum in tennis, and provide more scientific guidance for athlete training and game management.

This study is expected to advance the application of data analytics and mathematical modelling in sports science to support the improvement of athlete performance and competition.

References

[1] wang yongmin, li jing. Research on the influence of tennis players' psychological factors on tournament performance[J]. Sports World,2024, (04):121-123.

[2] Gao Qiang,Gao Jingyang, Zhao Di.GNNI U-net:Accurate segmentation network for MRI left ventricular contour based on group normalisation and nearest neighbour interpolation[J]. Computer Science,2020,47(08):213-220.

[3] JIANG Zhongqing, ZHANG Yue, ZHOU Yi,et al. Determination method of abnormal runoff volume by linking signal processing technique and residual analysis[J]. Jilin Water Resources, 2024, (04):37-41. DOI:10.15920/j.cnki.22-1179/tv.2024.04.006.

[4] P. Tu, Y. Xiong, Y. Wu, et al. A filtering algorithm for suppressing direction finding anomalies based on chi-square test [J]. Electroacoustic Technology,2023,47(06):148-152. DOI: 10. 16311/ j. audioe.2023.06.043.

[5] Qingxin He, Chuanfa Chen, Yuhui Wang,et al. A fusion method for multi-source remote sensing daily precipitation data: a random forest model with consideration of spatial autocorrelation[J]. Journal of Geo-Information Science, 2024, 26 (06):1517-1530.

[6] Saghi H, Nezhad S R M ,Saghi R , et al. Comparison of artificial neural networks and genetic algorithms for prediction of liquid wakes parameters (in English)[J].Journal of Marine Science and Application,2024,23(02):292-301.