A Study of Athlete Performance Prediction Based on TOPSIS and Logistic Regression Models

Qiyu Lei[1, #], Qizhuo Lei[2, *, #], Xuan Huang[3, #], Jinfan Wang[4, #]

[1]Guangdong University of Technology School of International Education, Guangdong University of Technology, Guangzhou, China, 510006

[2]Aberdeen Institute of Data Science and Artificial Intelligence, South China Normal University, Foshan, China, 528225

[3]School of Chemical Engineering and Light Industry, Guangdong University of Technology, Guangzhou, China, 510006

[4]Institute of Mechanical and Electrical Engineering, Guangzhou Huali College, Jiangmen, China, 529100

* Corresponding Author Email: 18520046845@163.com

#These authors contributed equally.

Abstract. Tennis is a popular sport characterized by intense competition and demanding physical demands. To assess the performance of athletes in matches and to explore the key factors affecting their momentum and real-time scores, this paper selected eight indicators that have a significant impact on athletes' performance and processed the data using the entropy weighting method to derive the impact weights of each indicator on the athletes. The combined use of the TOPSIS method enabled the evaluation of the athletes' performance at each stage, while the real-time performance was demonstrated through visual analysis. To establish a real-time prediction model, 17 player indicators with a significant influence on training were selected, and a logistic regression model was established. The final model demonstrated an accuracy of 92%. The analysis revealed that serve percentage and serve speed exerted a significant positive influence on the player's momentum, whereas the unforced error rate exerted a significant negative influence on his performance. The prediction curve analysis yielded 2911 turning points in the players' status, providing a valuable reference for athletes' real-time performance.

Keywords: TOPSIS, Logistic Regression Modelling, Athlete Performance Prediction.

1. Introduction

Tennis is a sport that is widely popular, characterized by intense competition and demanding physical fitness requirements. In tennis matches, players who gain an advantage often exhibit superior performance. This trend tends to impact the players' technical abilities and mental outlook. In Utility Analysis of Different hitting areas for World elite male tennis players: A case study of Djokovic and Federer [1], this advantage is commonly referred to as "momentum". However, this phenomenon is not exclusive to tennis; it is observed in many competitive sports. According to Research on the influence of Tennis players' Psychological Factors on competition results [2], Chinese tennis player emotional intelligence theory model building and scaling [3] and Special physical characteristics and competitive tennis training strategy research [4], momentum is linked to the match situation, players' psychology, tactical adjustments and so on. Also, for the reach of Momentum in tennis: Controlling the match [5] and Analysis of Nadal's Serve Receiving technique -- Taking the 2022 Australian Open as an example [6] an understanding of how players respond to "momentum," the impact of court changes on "momentum," is not only advantageous for players in enhancing their lead but also crucial for timely adjustments during pivotal moments in a match. Therefore, research into the factors influencing this "momentum" and its effects on match scores is of paramount importance.

The objective of this study is to evaluate the performance of athletes in competition and to identify the key factors influencing their poise and real-time scoring. In order to achieve this objective, this text have been working on Analysis of kinematic characteristics affecting service speed of Chinese elite female tennis players [7] and Professional tennis players' serve correlation between segmental angular momentums and ball velocity [8], the problem will be divided into two distinct parts. The first part of the study will involve the identification of athletes' performance. The second part will assess the evaluation of the prediction of players' volatility. The data used in this paper was obtained from the website https://www.comap.com/contests/mcm-icm.

2. Identification of player performance

2.1. Model Construction

The performance of a player on the field is considered to be a comprehensive score of various data on the field. In response to the question, it can be posited that the server is more likely to accrue points. The evaluation index factors of serve score are selected as direct serve score (ACE), serve score (serve_point_victor), and the actual number of successful breaks (p1_break_pt_won, p2_break_pt_won). according to The role of psychological Quality in Tennis Match[9]**,** the psychological state of the player is altered when their opponent scores consecutively. Therefore, the continuous score (point_victor) is employed as an internal evaluation index, with its objective weight determined through the entropy weight method. The TOPSIS algorithm is employed in conjunction with the scoring of each athlete in each time period. This enables the determination of which athlete is closest to the ideal performance in a given time period, thus allowing for the optimal score to be obtained by integrating various indicators. The flowchat of player performant identification is shown in Figure 1.

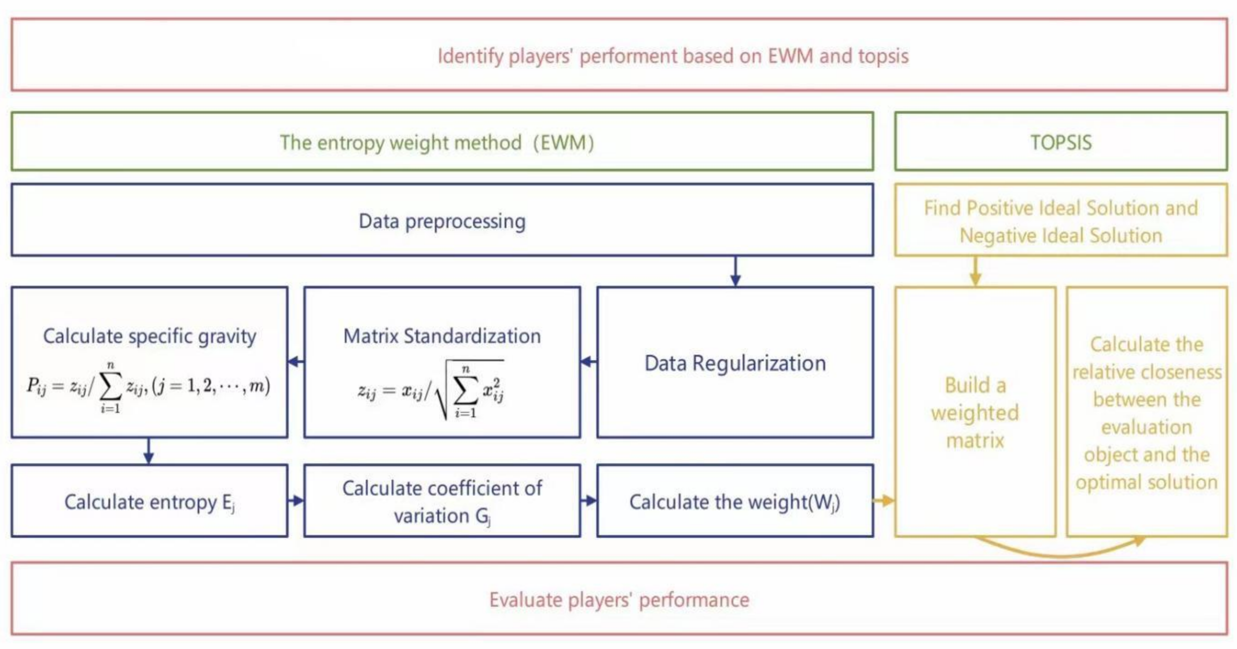

Figure 1. Flowchat of player performant identification

2.2. Actual steps for modeling

2.2.1 Regularization processing of original matrix

According to the characteristics of the data, we will use appropriate methods to forward the data, as shown in Table.1.

Table 1. Regularization processing method table

Data name	Indicator type	Normalization methods
server	Intermediate indicators(Approaching 1)	Method one
P1_ace	Intermediate indicators(Approaching 1)	Method one
P1_break_pt_won	Intermediate indicators(Approaching 1)	Method one
Point victor	Intermediate indicators(Approaching 1)	Method one
P2_ace	Smallest indicator	Method two
P2_break_pt_won	Smallest indicator	Method two
P2_points_won	Smallest indicator	Method two
Consist of win	Largest indicator	Keep original value

Method one:

$$M = max\{|x_i - x_{best}|\}, \tilde{x}_i = 1 - \frac{|x_i - x_{best}|}{M} \tag{1}$$

Method two:

$$\mathrm{M} =^{\mathrm{max-x}} \tag{2}$$

Matrix standardization to eliminate the influence of different indicator dimensions, assuming there is n (n is determined based on the total score of point no in each game, point no is the serial number of each scoring point in the game, providing each point of the game Unique identification) evaluation object, m evaluation indicators (four evaluation indicators), and the forwarding matrix X is:

$$\mathrm{X} = \begin{bmatrix} \mathrm{x}_{11} & \mathrm{x}_{12} & \cdots & \mathrm{x}_{1\mathrm{m}} \\ \mathrm{x}_{21} & \mathrm{x}_{22} & \cdots & \mathrm{x}_{2\mathrm{m}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{x}_{\mathrm{n}1} & \mathrm{x}_{\mathrm{n}2} & \cdots & \mathrm{x}_{\mathrm{nm}} \end{bmatrix} \tag{3}$$

For the corresponding normalized matrix Z, the normalized formula for each element in the matrix is:

$$z_{ij} = x_{ij} / \sqrt{\sum_{i=1}^{n} x_{ij}^2} \tag{4}$$

Then the normalized matrix Z is:

$$Z = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1m} \\ z_{21} & z_{22} & \cdots & z_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{nm} \end{bmatrix} \tag{5}$$

Calculate the value of the i-th solution under the j-th indicator (the i-th point, the j-th performance indicator of the player) as a proportion of this indicator Pij is:

$$P_{ij} = z_{ij} / \sum_{i=1}^{n} z_{ij}, (j = 1,2, \cdots, m) \tag{6}$$

Calculate the jth performance index entropy value Ej. When Pij=0, Pij=0:

$$E_j = -\frac{1}{\ln n} \sum_{i=1}^{n} P_{ij} \ln P_{ij}, (j = 1,2, \cdots, m) \tag{7}$$

Calculate the difference coefficient Gj of the jth indicator:

$$G_j = 1 - E_j \tag{8}$$

Calculate the weight Wj of the jth indicator:

$$W_j = \frac{G_j}{\sum_{j=1}^{m} G_j} \tag{9}$$

2.2.2 Topsis

Based on the standardized matrix, determine the positive ideal solution and the negative ideal solution. The positive ideal solution consists of the maximum value of the elements in each column of the matrix, and the negative ideal solution consists of the minimum value of the elements in each column. At this time, the standardized matrix Z is (n represents n games played, m represents four evaluation indicators):

$$Z = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1m} \\ z_{21} & z_{22} & \cdots & z_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{nm} \end{bmatrix} \tag{10}$$

positive ideal solution:

$$\begin{aligned} Z^{+} &= (Z_1^{+}, Z_2^{+}, \cdots, Z_m^{+}) \\ &= (max\{z_{11}, z_{21}, \cdots, z_{n1}\}, max\{z_{12}, z_{22}, \cdots, z_{n2}\}, \cdots, max\{z_{1m}, z_{2m}, \cdots, z_{nm}\}) \end{aligned} \tag{11}$$

negative ideal solution:

$$\begin{aligned} Z^{-} &= (Z_1^{-}, Z_2^{-}, \cdots, Z_m^{-}) \\ &= (min\{z_{11}, z_{21}, \cdots, z_{n1}\}, min\{z_{12}, z_{22}, \cdots, z_{n2}\}, \cdots, min\{z_{1m}, z_{2m}, \cdots, z_{nm}\}) \end{aligned} \tag{12}$$

Construct a weighting matrix, and multiply each column of the standardized matrix Z by the corresponding weight:

$$Z = \begin{bmatrix} \omega_1 * z_{11} & \omega_2 * z_{12} & \cdots & \omega_m * z_{1m} \\ \omega_1 * z_{21} & \omega_2 * z_{22} & \cdots & \omega_m * z_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \omega_1 * z_{n1} & \omega_m * z_{n2} & \cdots & \omega_m * z_{nm} \end{bmatrix} \tag{13}$$

Use the negative ideal solution distance to calculate the relative closeness of the evaluation object to the optimal solution. The score is between 0-1. The higher the score, the better. The higher the score Si of the i-th evaluation object:

$$S_i = \frac{D_i^{-}}{D_i^{+} + D_i^{-}} \tag{14}$$

Finally, normalize the scores:

$$S_i = S_i / \sum_i^n S_i \tag{15}$$

2.3. Results analysis

After processing by the entropy weight method and the topsis model, we use the smallest unit of tennis, points, as the x-axis. Use the score Si as the y-axis and make a line chart for visual analysis. Take the first and 14th games as examples, The First game player performance line chart is shown in Figure 2.

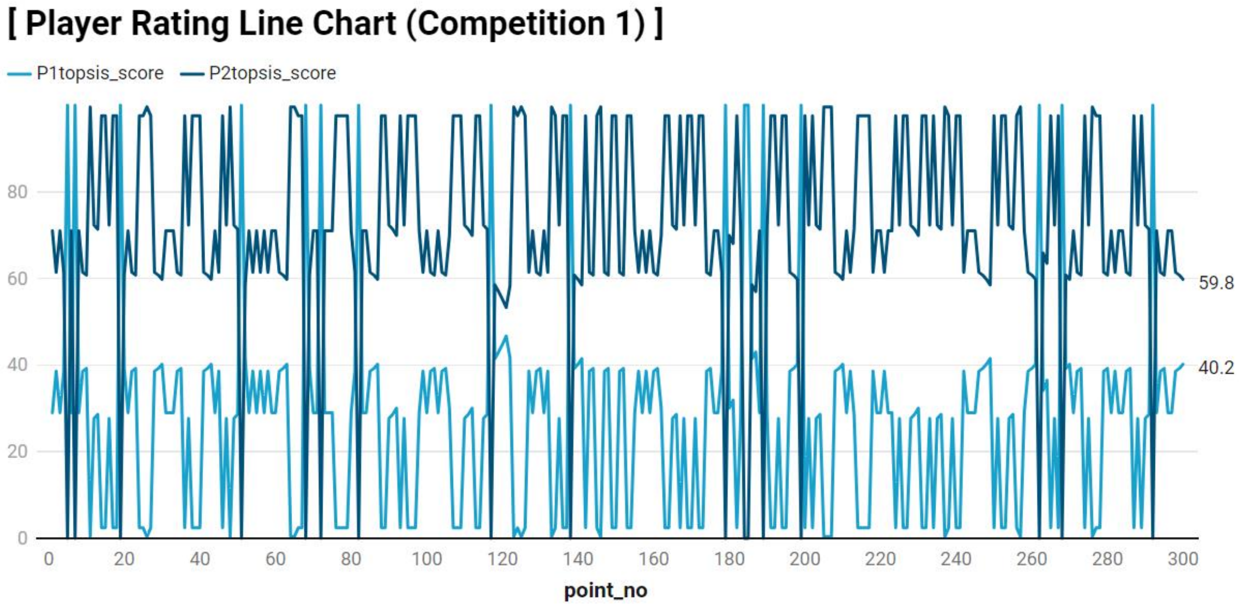

Figure 2. First game player performance line chart

The first game pitted Carlos Alcaraz (P1) against Nicolas Jarry (P2). From the picture we can see that Nicolas Jarry had the upper hand throughout the game. Among them, Carlos Alcaraz's performance took a turn throughout the game, but he failed to maintain it. Therefore, Nicolas Jarry performed well in all periods of the game when facing Carlos Alcaraz. However, Carlos Alcaraz was able to create a turning point in the overall situation, but he has never been able to. Keep up the good work.

3. Prediction of player swings

3.1. Logistic Regression Model

The fluctuations of players in the game can be seen as a binary classification problem, where there are only two results for the player's next goal score (score or loss) and obtaining the correlation between various factors is also one of the purposes of establishing a model. The logistic regression model is fast and suitable for binary classification problems. It can directly see the weights of each feature and continuously update the model by absorbing new data. Therefore, we choose to use logistic regression models for prediction. The flowchat of player swing prediction is shown in Figure 3.

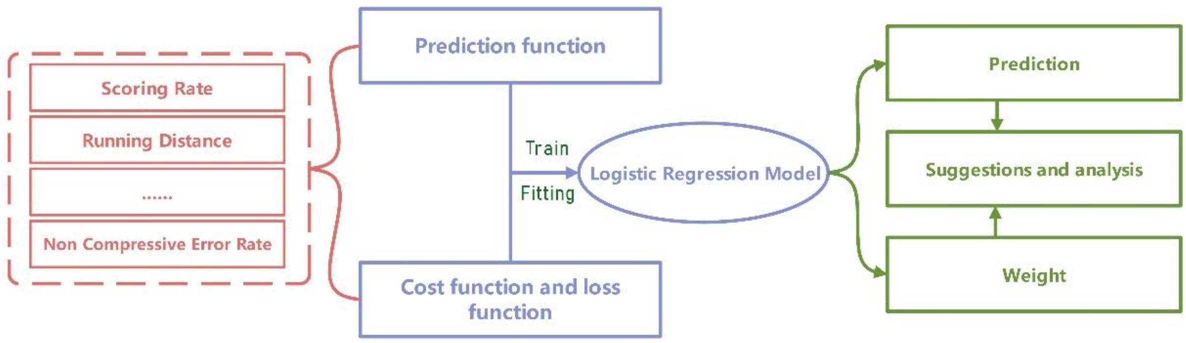

Figure 3. Flowchat of player swing prediction

3.1.1 Select characteristic indicators.

To predict the timing of player state fluctuations, we selected some game data that will affect players. In the reach of Influence of endurance Running Training on the stability of tennis stroke **[10]**, the running distance and rally count will affect their physical strength. The number of consecutive points scored by a player, the unforced error rate, and the actual score difference can affect the player's psychology. The score of a player's serve and the number of ace balls is factors that affects the player's score in the serving game. The success rate of net play, the success rate of breaking serve, and the speed of scoring serve are related to the player's ability. Therefore, we selected these indicators to train the model and predict.

3.1.2 Construct prediction function

For binary classification problems, if the probability of an event occurring (which we consider the probability of Player 1 winning) is p, then the probability of the event not occurring is 1-p. The probability of this event is defined as the ratio of the probability of occurrence to the probability of non-occurrence:

$$odds(p) = \frac{p}{(1-p)} \tag{16}$$

By taking the logarithm of the probability, we can obtain the logarithmic probability (log adds), which is the logit function. The range of logarithmic probability is the entire real number field, from negative infinity to positive infinity.When we choose the binomial distribution and logarithmic probability linking function, we obtain logistic regression, written in the following logarithmic probability form:

$$\ln\left(\frac{p}{1-p}\right) = w^T \cdot x + b \tag{17}$$

Our linear predictor is equal to the logarithmic probability of the sample. By transforming the above equation, we can solve for p, which is the probability that the sample belongs to a positive class. The formula is:

$$y = p = \frac{e^{w^T \cdot x+b}}{1+e^{w^T \cdot x+b}} \tag{18}$$

3.1.3 Construct cost function and loss function

Based on maximum likelihood estimation, the cost function and loss function functions are derived as follows:

$$\mathrm{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) \mathrm{if} y = 1 \\ -\log(1-h_\theta(x)) \mathrm{if} y = 0 \end{cases} \tag{19}$$

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \mathrm{Cost}(h_\theta(x_i), y_i) = -\frac{1}{m}\left[\sum_{i=1}^{m}\left(y_i \log h_\theta(x_i) + (1-y_i)\log\left(1-h_\theta(x_i)\right)\right)\right] \tag{20}$$
$$\mathrm{feature} X \in 1 \ldots 17, m \in 1 \ldots 14568$$

3.1.4 Training model and results

According to the prediction function and loss function, input numerical values for training. In each iteration, the algorithm updates the model parameters until the stopping condition is met. After training, the accuracy of the model data we obtained reached 92%, and the fitting degree graph of the first 150 points before the first game was shown in Figure 4.

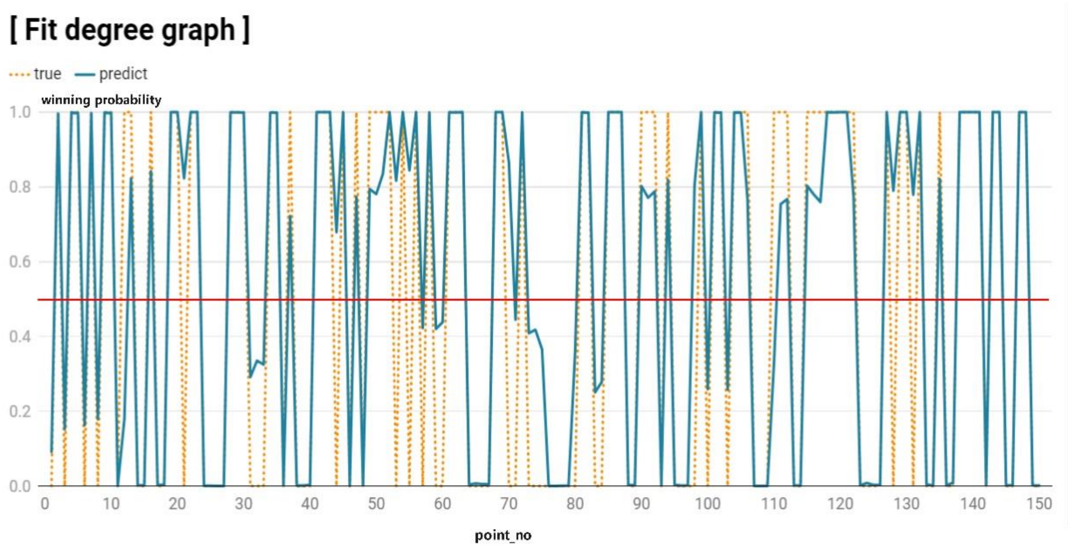

Figure 4. First fitting degree graph

At the same time, we can obtain the weight of players' fluctuations in various game indicators through the model. Create a heatmap of the weight data that positively and negatively affect player performance. The heatmap is shown in Figure 5.

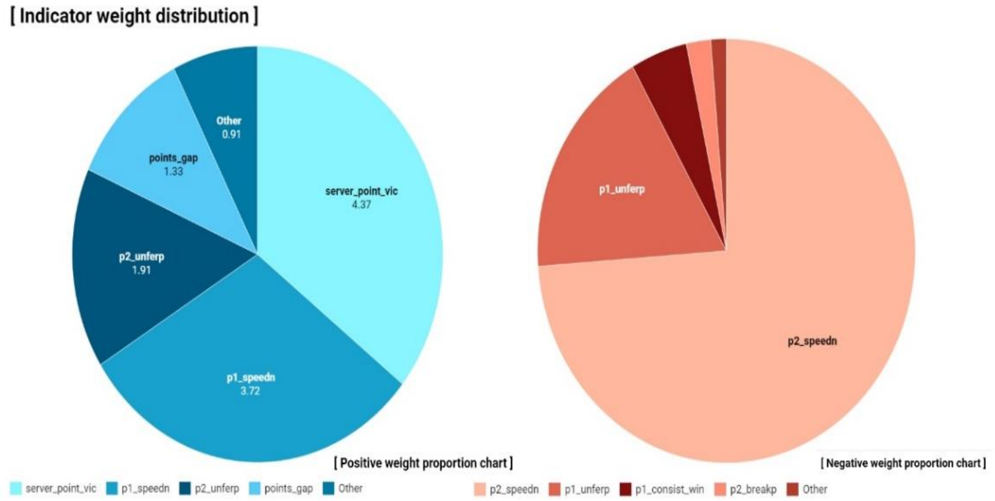

Figure 5. Heat map of weight for each indicator

From the positive weight chart, we can see that the player's scoring rate, serving speed, difference in points, and the opponent's unforced error rate will all have a positive impact on the player. Among them, the weight of scoring rate and serving speed in the serve is the highest.

From the negative weight chart, the number of unforced error rate made by a player and the opponent's serving speed can have a negative impact on the player's state, especially when the opponent's serving speed is fast.

4. Conclusions

In this study, this text established a comprehensive evaluation model based on various data of players. In the process of evaluating the model, this text first preprocesses the data and use the entropy weight method to obtain the weight of the influence of each player index on the player's score. Then combined with topsis method, the performance of each stage of the player is scored, and the real-time status of the player is intuitively seen. To understand the key factors that influence a player's "momentum" and to predict a player's scoring outcome in real time from the data, this text built a logistic regression model. Firstly, text selected 17 influential player indicators and substituted them for training. The final model achieved an accuracy of 92%. Through the analysis of the weight of each index by logistic regression model, this text draws the conclusion that the player's service score rate and service speed have a great positive impact on the player's "momentum" and the player's non-pressure error rate has a great negative impact on the player References.

References

[1] Li Meng, Li Weifeng. Utility Analysis of Different hitting areas for World elite male tennis players: A case study of Djokovic and Federer [J]. Sports science and technology literature bulletin, 2024, 32 (4): 59-67.

[2] Wang Yongmin, Li Jing. Research on the influence of Tennis players' Psychological Factors on competition results [J]. World of Sports, 2024, (04): 121-123.

[3] ZHAO Mingnan, Jiang Haoqi, Qi Bing, et al. Chinese tennis player emotional intelligence theory model building and scaling [J]. Journal of psychological research, 2024 (01): 42-53.

[4] Xia Lei. Special physical characteristics and competitive tennis training strategy research [J]. Journal of sport science and technology literature bulletin, 2024, 32 (4): 91-94 + 201.

[5] Laurent D, Carien R, Robin N. Influence of Motor Imagery Modality on First-Serve Performance in Tennis Players. [J]. Motor control, 2024, 11-14.

[6] Yang Xiaopeng. Analysis of Nadal's Serve Receiving technique -- Taking the 2022 Australian Open as an example [J]. Sports Vision, 2024, (02): 95-97.

[7] Xiao Yunfei. Analysis of kinematic characteristics affecting service speed of Chinese elite female tennis players [D]. Beijing sport university, 2020.

[8] Caroline M, Richard K, Paul D, et al. Professional tennis players' serve correlation between segmental angular momentums and ball velocity. [J]. Sports biomechanics, 2013, 12(1): 2-14.

[9] Yue Feng. The role of psychological Quality in Tennis Match [J]. Tennis World, 2024, (01): 58-60.

[10] Dong Weijia, Zhang Qi. Influence of endurance Running Training on the stability of tennis stroke [J]. Boxing & Fighting, 2023, (12): 37-39.