Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

A study of sports scores based on correlation analysis and logistic regression

Shaohang Chen [1, *], Jiacheng Chen [1], Ruizhe Jiang [2], Yian Xuan [1], Siming Liang [1]

[1] School of Mechanical Engineering, Shijiazhuang Tiedao University, Shijiazhuang, China, 050043

[2] School of Engineering Mechanics, Shijiazhuang Tiedao University, Shijiazhuang, China, 050043

* Corresponding author: chenshaohang1217@163.com

Abstract. This paper examines the relationship between "Momentum" and victory in the imbledon 2023 men's tournament, particularly between the semi-finals and the final. The definition and ormula of the Pearson's correlation coefficient are presented, and the Pearson's correlation oefficient is analysed based on the match data between Carlos Alcaraz and Novak Djokovic. The results show that there is a positive correlation between "Momentum" and winning. Then the logistic regression model was used to analyse the factors affecting the variation of "kinetic energy", such as onsecutive points, decisive points, key points, fatigue level and double faults. By training and testing the data set, a logistic regression model was developed for predicting the changes in "kinetic energy", and some suggestions based on the differences in "Momentum" were made to help players and coaches better cope with the changes in "Momentum" in the game. This paper also proposes some suggestions based on "Momentum" differences to help players and coaches better cope with "Momentum" changes in matches.

Keywords: Correlation Analysis; Logistic Regression; Momentum.

1. Introduction

As one of the top tennis events in the world, Wimbledon Championships attracts the attention of countless tennis fans all over the world. In such a competitive field, the players' form and performance are often one of the key factors determining the winners and losers. In this regard, "kinetic energy" is a concept that has attracted much attention in tennis as it may have a significant impact on the course of the match and victory. However, there is a lack of systematic quantitative analyses and research on the exact role of kinetic energy in matches and its relationship with victory [1].

This paper aims to explore the relationship between kinetic energy and victory in the context of the Wimbledon 2023 men's tournament, and proposes a corresponding analytical approach. Firstly, the concept and calculation method of Pearson's correlation coefficient and the application principle of logistic regression model are introduced. Then, through the quantitative analysis of the game data, the correlation between "kinetic energy" and victory was explored, and the key factors affecting the change of "kinetic energy" were analysed by the logistic regression model. Finally, based on the results of the analyses, recommendations for coaches and players are proposed, with a view to providing a reference basis for improving game performance and formulating strategies [2-3].

2. Proof of the Role of "Momentum"

2.1. Pearson Correlation Coefficient

The Pearson correlation coefficient is defined as follows: In statistics, the Pearson correlation coefficient, also known as PPMCC or PCCs (commonly represented as r or Pearson's r), is used to measure the degree of correlation (linear correlation) between two variables X and Y with values ranging between -1 and 1. In the field of natural sciences, this coefficient is widely used to measure the degree of correlation between two variables, primarily to assess their linear relationship [4].

The Pearson correlation coefficient between two variables is defined as the quotient of the covariance of the variables and the product of their standard deviations:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \tag{1}$$

The calculation formula for the Pearson correlation coefficient is:

$$r = \frac{N\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{N\sum x_i^2 - (\sum x_i)^2}\sqrt{N\sum y_i^2 - (\sum y_i)^2}} \tag{2}$$

In this formula, the closer the correlation coefficient is to 1 or -1, the greater the absolute value of the correlation coefficient and the stronger the correlation [5-6]. The closer the correlation coefficient is to 0, the weaker the correlation. Typically, the strength of correlation between variables is judged using the range of values in the following table 1.

Table 1. Levels of Correlation Strength

Value Range	Degree of Correlation
0.8-1.0	Very Strong Correlation
0.6-0.8	Strong Correlation
0.4-0.6	Moderate Correlation
0.2-0.4	Weak Correlation
0-0.2	Very Weak or No Correlation

In this section, we study the Pearson correlation coefficient between "momentum" and winning to demonstrate the relationship between the two.

The given data encompasses 31 matches, sourced from the Wimbledon 2023 men's matches after the first two rounds. We initially assume that all indicators present in the data are meaningful and start with an analysis of some of these indicators.

We analyze the Pearson correlation coefficient between "momentum" and the number of matches won in the final between Carlos Alcaraz and Novak Djokovic, resulting in Table 2.

Table 2. Pearson Correlation Coefficients

	Carlos Alcaraz	Novak Djokovic
Pearson Correlation Coefficient	0.614	0.597
Correlation Strength Level	Strong Correlation	Moderate Correlation

As indicated by the Table 2, in this match, the Pearson correlation coefficients for "momentum" and winning for both individuals are greater than 0.5, which substantiates that there is a strong correlation between "momentum" and winning [7].

2.2. Model Result Analysis

In this part, we analyze and compare the number of games won and the changes in "momentum" throughout the match for Carlos Alcaraz and Novak Djokovic during their head-to-head contest.

The number of games won and the changes in "momentum" for both players as the match progresses are illustrated in Figure 1.

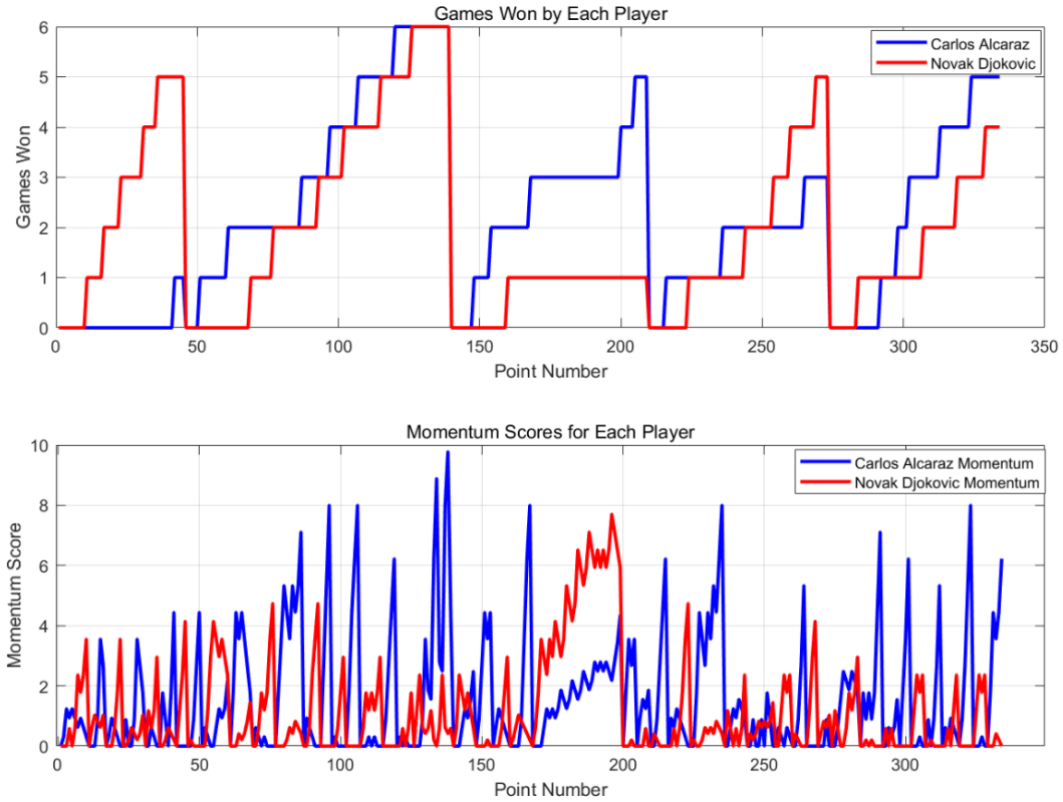


Figure 1. The Number of Games Won and The Momentum Score

From the figure 1, overall, Carlos Alcaraz's "momentum" was higher than Novak Djokovic's "momentum," and the result of Alcaraz's victory also proves that there is a certain degree of positive correlation between "momentum" and victory, thereby indirectly validating our model. The Pearson correlation coefficient, combined with the changes in the number of games won, suggests that an increase in a player's "momentum" at a given moment can to some extent lead to a higher probability of winning subsequent games. Thus, we consider the tennis coach's statement to be incorrect; "momentum" can play a role in a match to a certain extent.

3. The Transition of "Momentum" in Matches

In this section, we use the logistic regression model to focus on indicators that affect "momentum," tracking how they change over the course of the match to determine the timing and influence of "momentum" transitions [8].

3.1. Logistic Regression Model

3.1.1. Index definition.

To describe the transition of "momentum" during a match, we focus on 14 data indicators related to the match process and potential influences on the momentum change, which are categorized into 5 types.

Consecutive Scoring: Consecutive scoring can significantly boost a player's confidence and exert psychological pressure on the opponent. It can demonstrate dominance in the match and potentially cause the opponent to lose concentration or morale, thus shifting "momentum" towards the scoring player. The player's consecutive scoring situation can be judged through point_victor, p1_point_won, and p2_point_won.

Decisive Point: If a player scores a decisive point, thus winning the game, it will increase their "momentum." This can be determined through p1_ace, p2_ace, p1_break_pt_won, and p2_break_pt_won.

Critical Points: A high-quality shot demonstrates a player's skill, which can build confidence and disrupt the opponent's rhythm. Net play for scoring points indicates an aggressive style of play, which can also pressure the opposing player. The situation of critical points scored can be judged through p1_winner, p2_winner, p1_net_pt_won, and p2_net_pt_won.

Fatigue Level: A match consumes the energy of the participants, and effective management of physical stamina can actively adjust one's "momentum." High-intensity fatigue may make it more challenging to score, leading to a decrease in "momentum." The players' level of fatigue can be assessed through p1_distance_run, p2_distance_run, and rally.

Double Faults: Frequent double faults can diminish a player's "momentum," as they provide the opponent with easy points. The occurrence of double faults can be determined through p1_double_fault and p2_double_fault.

3.1.2. Construction of the Logistic Model.

Logistic regression is a generalized linear regression analysis model, widely applied in the field of data collection and result prediction. Thus, it is suitable for this problem. We select the 14 indicators as influencing factors to establish a logistic regression model for prediction [9-10].

Considering the independence of different rounds of matches, we first extract data that may influence the momentum indicators from the dataset as input values, denoted as f1, f2, ... f13, f14. The point difference (p1_point_won, p2_point_won) is used as the output value y. We define the moments where the score difference is greater than 3 as turning points in the game. We then split this dataset into 70% for training and 30% for testing.

The hypothesis function of logistic regression is as follows:

$$h_0(f_i) = \frac{1}{1+e^{-\theta^T f_i}} \tag{3}$$

Where fi are the input metrics, θ presents the parameters we aim to estimate.

We then begin to train our model on the training set, with the cost function $J(\theta)$ calculated as follows:

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} [y^{(i)} \ln (h_\theta(f^{(i)})) + (1 - y^{(i)}) \ln (1 - h_\theta(f^{(i)}))] \tag{4}$$

Where θ represents regression parameters, λ represents the regularization coefficient, which we set as 0.6. When the cost function $J(\theta)$ is the smallest, the regression effect is best, and the regression model can be considered the final model.

The logistic regression network diagram is shown in Figure 2:

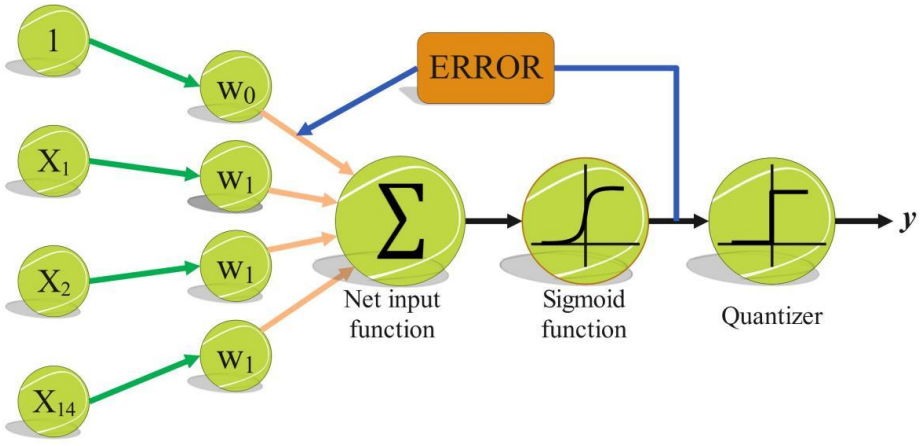


Figure 2. The Logistic Regression Network Diagram

3.2. Model Result Analysis

We first use the 2023_wimbledon_1701 dataset to test train the logistic regression model, with the testing results illustrated in Figure 3.

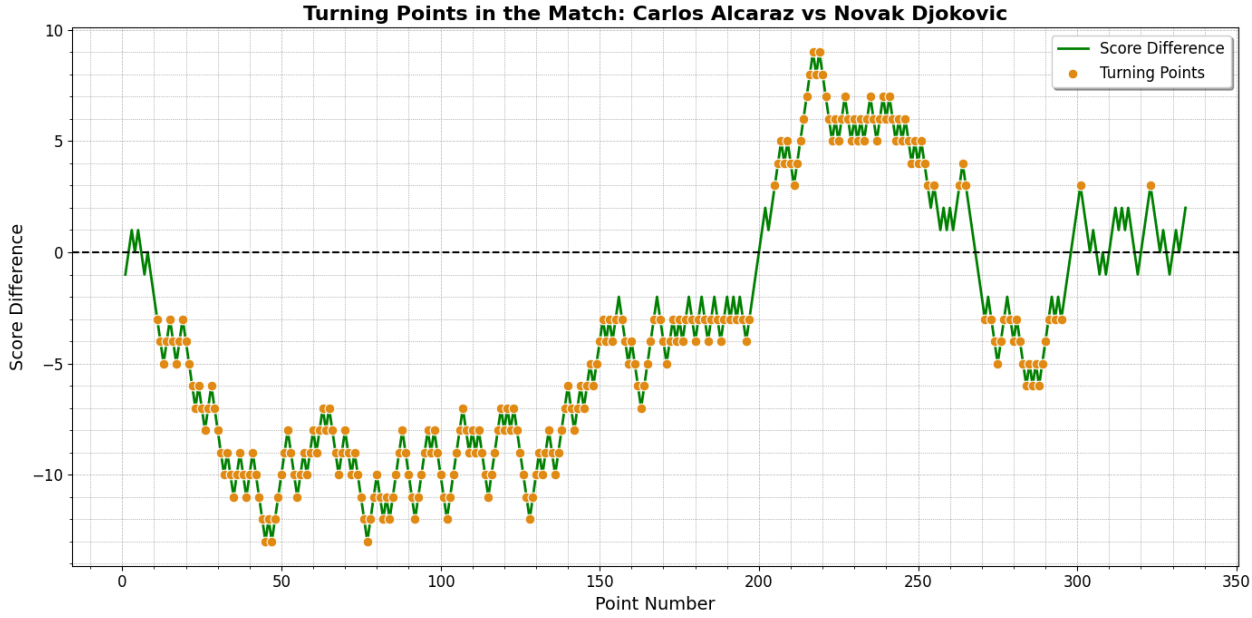


Figure 3. Training Results for Turning Points in Match Momentum

The predictive performance results of the logistic regression model are shown in Table 3:

Table 3. Logistic Regression Prediction

	precision	recall	f_1_score	support
0 (Benign)	1.00	0.62	0.76	13.00
1 (Malignant)	0.81	1.00	0.89	21.00
accuracy			0.85	34.00
macro avg	0.90	0.81	0.83	34.00
weighted avg	0.88	0.85	0.84	34.00

Based on Table 3, the regression analysis prediction results show high precision, recall, and f1_score, with the model's prediction accuracy being 0.85 > 0.8, indicating that our predicted results are quite good.

Furthermore, we obtained the percentage weights of each indicator on the prediction results as shown in Figure 4. Through logistic regression, we derived the percentage weights of each indicator, which reflect the relevance of the indicators to the predictive outcomes

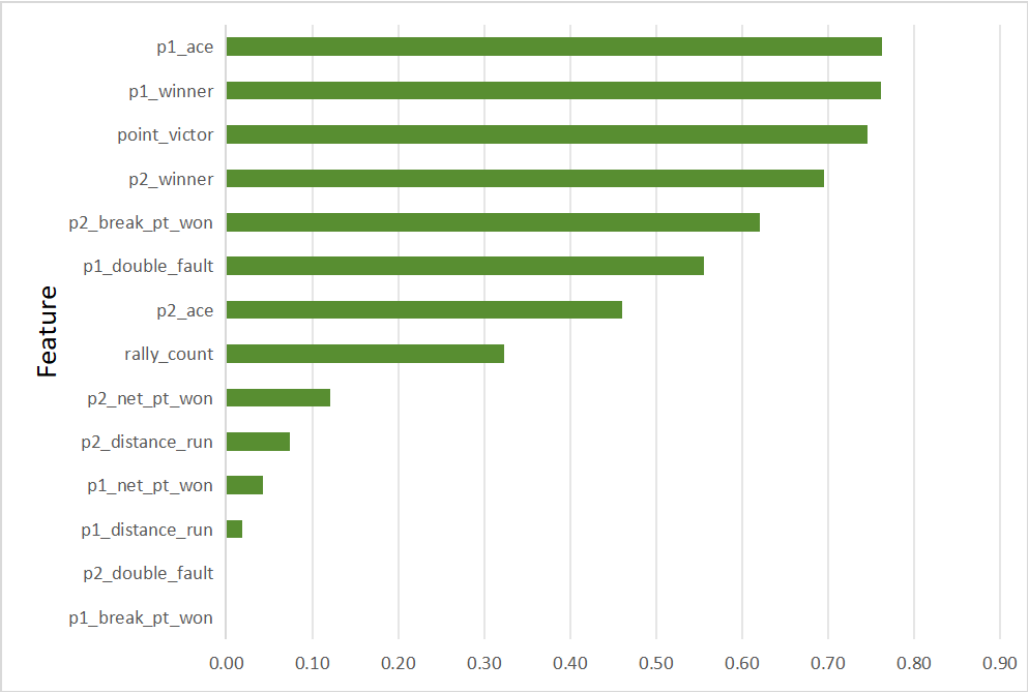
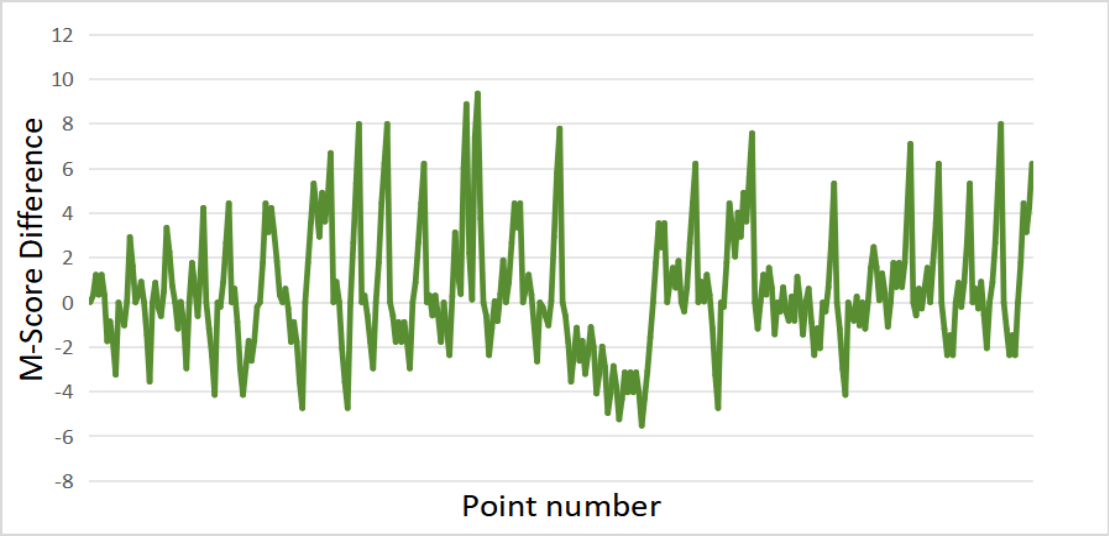


Figure 4. Importance coefficient of the feature

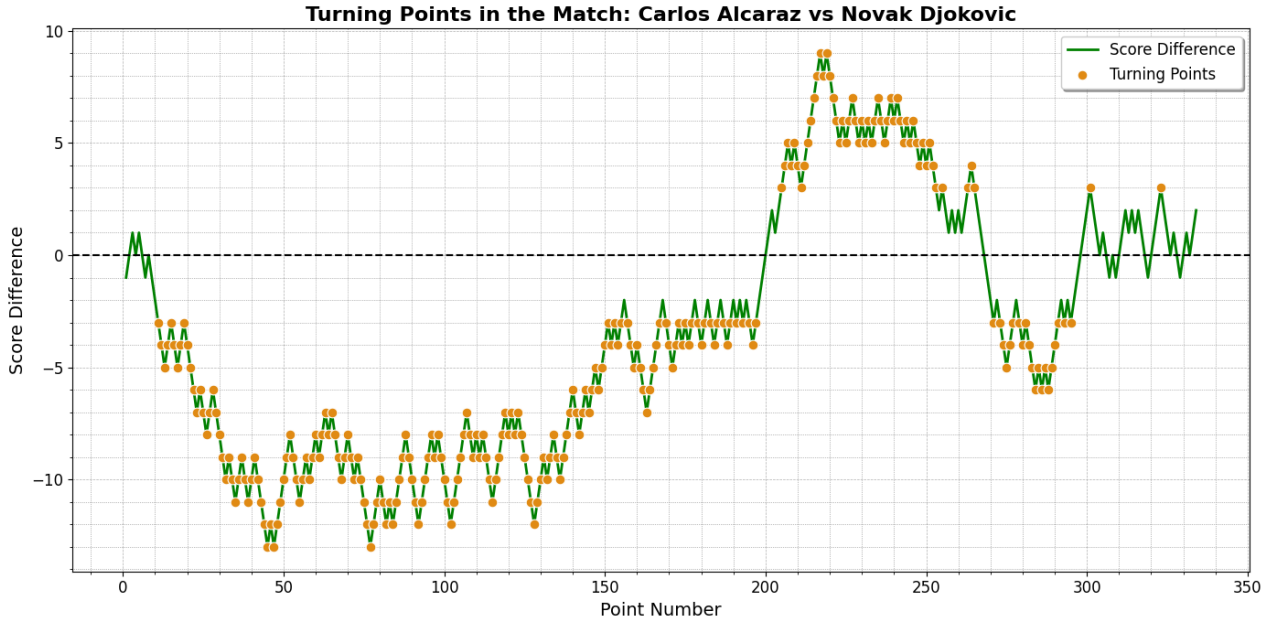
As shown in Figure 4, in this match, the importance coefficients for features such as p1_ace, p1_winner, point_victor, p2_winner, and p2_break_pt_won all exceed 0.6, having a significant impact on 'momentum' fluctuations. Meanwhile, the importance coefficients for features p1_break_pt_won and p2_double_fault are both 0, having the least impact on 'momentum' fluctuations.

3.3. Proposes Based on the "Momentum" Difference

We select the "momentum" scores quantified in Problem 1 as our dataset and plot a line graph as shown in Figure 5 of the "momentum" score differences between the two players to observe changes in "momentum."



(a)



(b)

Figure 5. M-Score Difference and Score Difference

The analysis results show that "momentum" turning points are densely packed where there are significant fluctuations in the "momentum" score difference. We provide the following recommendations for coaches of the players:

Pay attention to specific match turning point indicators, such as score difference and service utilization, as these factors may indicate imminent changes in match "momentum."

Train players to recognize and utilize turning points in the match, such as preparing mentally and adjusting tactics when scoring consecutively or facing critical points.

Analyze match data from different opponents to tailor strategies for varying match situations.

4. Conclusion

By quantitatively analyzing the relationship between kinetic energy and victory in the Wimbledon 2023 men's tournament, this paper has drawn some important conclusions. Firstly, this paper found a positive correlation between momentum and victory, with an increase in kinetic energy being predictive of victory, especially in key matches. Secondly, logistic regression modelling showed that key factors affecting changes in kinetic energy, such as consecutive points, decisive points and fatigue levels, had a significant impact on the outcome of the game. Finally, this paper provides some recommendations for coaches and players, including focusing on key points in the game, preparing in advance, and adjusting tactics, to better utilise the "kinetic energy" advantage and meet the challenges in the game.

These findings are not only important for understanding the relationship between "kinetic energy" and victory in matches, but also provide practical guidance for coaches and players to formulate strategies and improve their performance.

References

[1] WANG Wei, CHEN Xiaowei, KANG Shaojie, et al. Research on the performance prediction method of female throwers--taking the shot-put event as an example[C]// Chinese Society of Sports Science. Abstracts of the Thirteenth National Conference on Sports Science--Special Report (Sports Statistics Division). Chengdu Institute of Physical Education; Aba Normal College; 2023: 3.

[2] Lin Xianjian. Prediction of Physical Education Secondary School Examination Scores Based on BP Neural Networks[C]// Chinese Society of Sports Science. Abstracts Collection of the Thirteenth National Sports Science Conference - Written Communication (Sports Statistics Division). Yili Normal University; 2023: 2.

[3] J.Y. Long,Z.H. Ui. Research on the prediction of Olympic men's 100m gold medal performance based on GM (1,1) grey model [J]. Journal of Southwest Normal University (Natural Science Edition), 2023, 48 (07): 123 - 128.

[4] Li Ran. A study on the competitive performance of women's throwing events in the 24th-32nd Olympic Games and the prediction of the results of the Paris Olympics [D]. Qufu Normal University, 2023.

[5] Pei Zheng, Leng Qijian. Research on the development trend of women's event performance and grey GM (1,1) model prediction at the World Athletics Championships [J]. Sports Science and Technology Literature Bulletin, 2023, 31 (05): 37 - 39+64.

[6] Li Jiaqi. A comparative study of different prediction methods on the trend of sprint performance in China [D]. Tianjin Sports Institute, 2023.

[7] FAN Yingli, ZHENG Zhiqiang, ZHENG Wei. Prediction of college students' sports performance based on improved SSA-LSSVM model [J]. Computer Simulation, 2023, 40 (01): 283 - 289+483.

[8] Feng QM. Sports performance prediction based on time series analysis method [J]. Microcomputer Applications, 2022, 38 (12): 35 - 37+48.

[9] Huang ZY. Research on tennis performance prediction model based on BP neural network [D]. Fujian Normal University, 2022.

[10] Zhang M. A study on the prediction of competition time of China's outstanding speed skating 1500m athletes by 30sWingate test [D]. Jilin Institute of Physical Education and Sports, 2022.