Visualization of players' game performance based on random forest and particle swarm models

Yilu Tan[1, #], Junfeng Ni[2, *, #], Yitian Zhang[3, #], Jinyi Gao[4], Tianyi Yang[3]

[1]School of Geographical Sciences, University of Reading, UK

[2]School of Computer Science, Cyber Science and Engineering, Nanjing University of Information Science and Technology, Nanjing, China, 211800

[3]School of Mathematical, Physical and Computational Sciences, University of Reading, UK

[4]School of Mathematics and Statistics, Nanjing University of Information Science and Technology, Nanjing, China, 210044

* Corresponding Author Email: 202183290391@nuist.edu.cn

#These authors contributed equally.

Abstract. In this study, a model combining Particle Swarm Optimization (PSO) and Random Forest Algorithm (RF) is proposed for processing and selecting features in a sports competition dataset and optimizing the model parameters. First, the scale differences between different features were eliminated by data normalization, followed by optimizing the parameters of the random forest model using the PSO algorithm, which significantly improved the prediction accuracy of the model. The study also introduced the performance level scoring model, which assigns weights according to the importance of selected features, and the momentum scoring model, which explores the relationship between dynamic changes in the game and the outcome of the game. The validity of the models was verified through real-world data analysis, showing that this approach can accurately predict the outcome of a match. In addition, Wilcoxon rank sum test and Spearman correlation coefficient were applied to statistically validate the model, and the results showed that the selected features were significantly correlated with the match results. The application of these methods not only improves the explanatory ability of the model, but also provides a new research direction and methodological guidance for the field of sports data analysis. In summary, the integrated modeling framework developed in this study demonstrates high efficiency and accuracy in handling large-scale sports data, demonstrates its potential for application in competitive sports data analysis, and provides a solid foundation and new perspectives for subsequent research in related fields.

Keywords: Random Forest Model, Particle Swarm Optimization, Performance Visualization.

1. Introduction

In tennis matches, success is often determined by a multitude of factors. As one player starts scoring consecutively, they often undergo a remarkable transformation, continuing to score until victory. This phenomenon is termed "momentum," referring to when one player or team in a competition consecutively wins multiple points, games, or segments, thereby gaining a psychological or score advantage. This momentum can enhance the confidence and performance of the player or team, while potentially exerting pressure on the opponent, affecting their performance. Momentum is a dynamic, fluid state, triggered by various factors including outstanding individual performance, opponent errors, and unforeseen circumstances.

Traditionally, momentum is challenging to quantify. However, with the advancement of digital analysis techniques, it is now possible to attempt a quantitative approach to capturing and analyzing momentum by analyzing match data. Metrics such as consecutive scoring, break point conversion rates, and service game conversion rates can be used to quantify momentum and analyze its causes and impact on the game's progress and outcome.

Analyzing momentum allows players and coaches to adjust tactics timely, leveraging the current match momentum, such as intensifying the offense or strategically pausing when the opponent is

about to gain momentum. The advantage of analyzing momentum lies in providing players with a comprehensive perspective to observe and understand the match flow, enabling them to respond more strategically to various situations during the match, thereby enhancing performance and success rates. However, accurately defining and quantifying momentum and applying it in actual matches presents challenges.

Moreover, the regularity and predictability of turning points in matches remain debatable. It is unclear whether these are the result of systematic patterns or merely random events. The abstractness and variability of momentum make its analysis and application fraught with challenges. Discussing these issues is not only crucial for athletes and coaches but also significantly contributes to the study of sports science.

To this end, the study builds a model to capture the dynamics of scoring during matches while assessing and quantifying players' performance levels. The file Wimbledon_featured_matches.csv contains detailed match information for 31 matches. Therefore, it was necessary to pre-process the data before constructing the model. After confirming the absence of anomalies and other similar issues, the data was organized to derive potential features, including the streak status of the players at the beginning of each match. In order to identify features that are closely associated with score changes, a random forest algorithm optimized iteratively through particle swarm optimization was used [1]. This process fits each feature to the score, resulting in a regression coefficient for the feature that serves as the basis for assigning weights [2]. Subsequently, in order to quantitatively assess the performance level of the players, the study constructed a player performance scoring model using these important features to visualize the progress of the game and the performance of the players. Meanwhile, in order to explore whether there exists some relationship between momentum and the course of the game, the study established a momentum score difference evaluation model based on the satisfaction evaluation model[3], and the momentum score difference between the two sides of each game was obtained by assigning corresponding weights to each of the four important features that were screened out again, and the momentum score difference between the two sides of each game was calculated by using Support Vector Machine(SVM)[4]. And the relationship between the momentum score difference and the course of the game is captured by rank sum test [5]and Spearman correlation coefficient [6].

2. Model formulation and solving

2.1. Data Preprocessing

2.1.1. Data Normalization

When the numerical ranges of features in a dataset are inconsistent, features with higher numerical ranges may disproportionately influence distance calculations, thereby affecting model accuracy. Through normalization, the importance of all features is treated equally. For instance, the feature of running distance, whose numerical range significantly exceeds that of other features, requires data normalization.Normalization of Running Distance Data is shown in table 1.

Table 1. Normalization of Running Distance Data

Running Distance Data	0	17.894	114.549	148.723
Normalized Data	0	0.120317637	0.770217115	1

2.1.2. Feature Selection

Initially, after screening and organizing the Wimbledon_featured_matches.csv file data for outliers and missing values, the study initially selected eight features that were associated with scoring results, including “break point winners,” “serve winners “, ‘forced errors’ and ‘unforced errors’. In practical applications, the initially proposed Random Forest algorithm did not achieve the expected level of accuracy. Therefore, a Random Forest algorithm based on Particle Swarm Optimization was used, resulting in the PSO-RF model[7].

Particle Swarm Optimization[8] is an evolutionary computing technique that simulates social behavior and updates the positions of particles based on performance results. The rules for updating particle positions are as in equations (1) and (2).

$$P_2 = P_1 + v \tag{1}$$

The speed update rule is:

$$v_2 = \omega \cdot v_1 + c_1 \cdot r_1 \cdot (P_b - P_n) + c_2 \cdot r_2 \cdot (P_b' - P_n) \tag{2}$$

In this context, ω represents the inertia weight, c_1, c_2 represents the learning factor, and r_1, r_2 is a set of random numbers.

Through the Particle Swarm Optimization algorithm, as shown in table 2, we obtained the locally optimal parameters related to the random forest algorithm.

Table 2. Optimal Parameters of the Random Forest Algorithm

max_depth	n_estimators	random_state
36	91	42

Subsequently, random samples from the original dataset are bootstrap sampled using the Random Forest algorithm to train each tree. Once a sample is selected, it is eligible to be selected again, thus facilitating the generation of multiple training datasets. Each dataset is used to train a decision tree. No pruning is done in this process to allow each decision tree to reach its full potential. In the application of Random Forest Algorithm, we use the first 80% of the dataset for training and the remaining 20% for testing and the test results are shown in Table 3.

Table 3. Accuracy Comparison between the PSO-RF Model and the RF Model

PSO-RF model		precision	recall	f1-score
	accuracy			0.9
	macro avg	0.93	0.88	0.89
	weighted avg	0.92	0.9	0.9
RF model				
	accuracy			0.58
	macro avg	0.29	0.5	0.37
	weighted avg	0.34	0.58	0.43

This shows that the Random Forest algorithm with particle swarm optimization has significantly improved in terms of model accuracy. This lays a solid foundation for accurately identifying highly correlated features.

On this basis, the study obtained the regression coefficients of these 8 features, which indicate the degree of correlation between each feature and the win/loss results. And it is defined as “feature importance”, and the specific data are shown in fig.1.

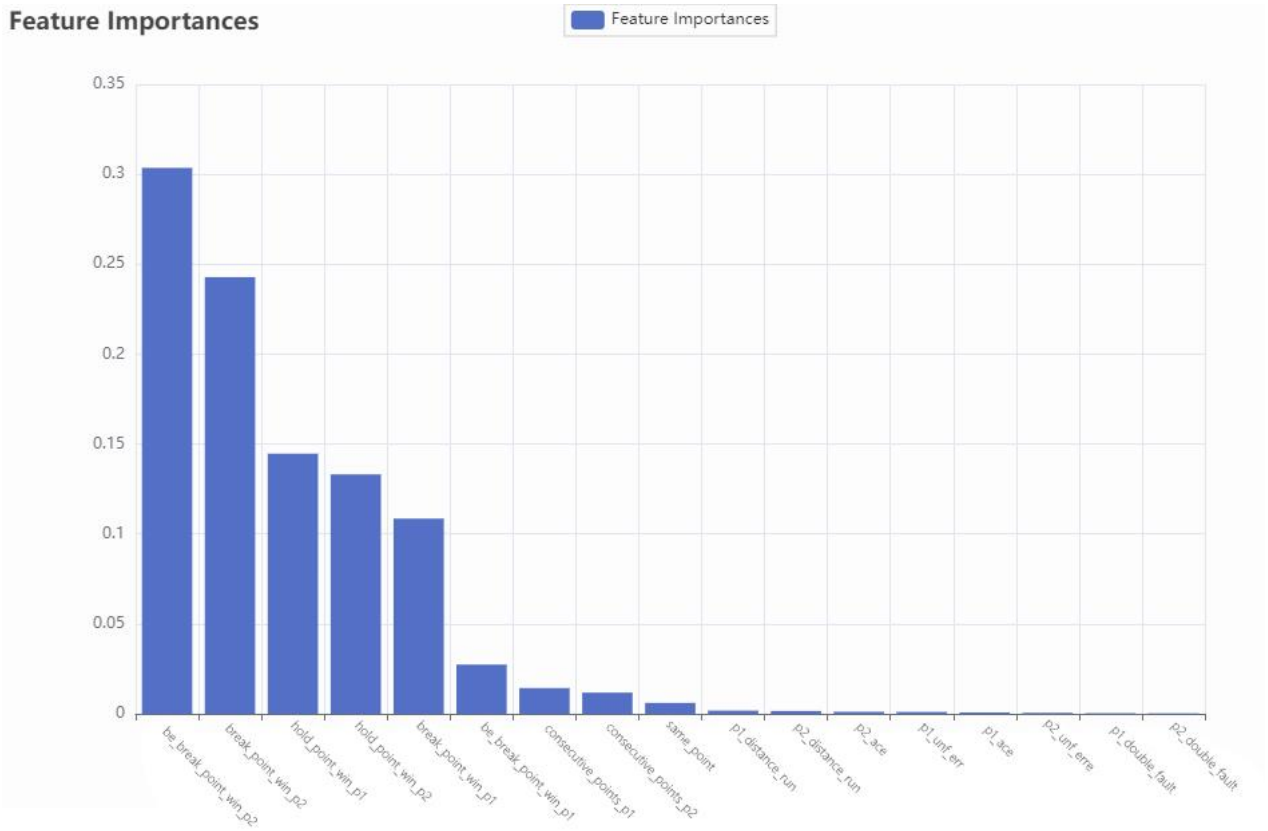


Fig. 1 Feature Importance

Since Player 1 and Player 2 were treated as separate training sets, to obtain an overall importance ranking of the features, the average importance value for each feature was calculated to represent its significance. The ranking results obtained are as Table 4.

Table 4. Top 5 Ranked Feature Importance

Feature Name	Break Points	Break Points Saved	Service Games Lost	Consecutive Points	Running Distance
Feature importance	0.308	0.399	0.209	0.062	0.009

The ranking results readily demonstrate that features related to serving, such as break points, break points saved, and service games lost, play significant roles in determining the match outcomes. Therefore, weights will be assigned according to the rank of top 5 feature importance in the subsequent performance scoring model.

2.2. Performance Level Score Model

Through feature selection, the top 5 features identified in terms of importance are "break points," "break points saved," "service games lost," "consecutive points" and "running distance." It is therefore posited that a player's on-court performance score is primarily determined by these five features. The weights assigned to features according to importance are 0.308, 0.399, -0.209, 0.062, 0.009 respectively.

The specific rules for calculating the player performance score are as follows.

In a single rally, the 5 identified features are considered as discrete actions, with a value of 1 indicating successful execution of the action, and a value of 0 indicating failure to complete the action. The weights assigned to these actions are their respective scores, and the sum of the scores of the 5 actions in a rally is the performance score of that rally. Based on this framework, the study constructed a performance score evaluation model, the derivation of which is summarized below:

In every point of the match, the study first obtain the scoring situation for each action:

$$p_{1j} = 0.308 \cdot \alpha_{1j} \tag{3}$$

$$p_{2j} = 0.399 \cdot \alpha_{2j} \tag{4}$$

$$p_{3j} = 0.209 \cdot \alpha_{3j} \tag{5}$$

$$p_{4j} = 0.062 \cdot ct_j \tag{6}$$

$$p_{5j} = 0.009 \cdot \alpha_{5j} \tag{7}$$

Where $\alpha(1,0)$ indicates whether the action has been completed, ct indicates the consecutive scoring streak of the player prior to the start of this point in the match. Thus, the performance score for each point in the match can be represented as:

$$p_j = p_{1j} + p_{2j} + p_{3j} + p_{4j} + p_{5j} \tag{8}$$

Subsequently, the study accumulate the total score of each player for every point in the match, obtaining the cumulative performance score of the player at any given moment. At this point, the cumulative performance score can be represented as:

$$P_j = \sum_{j=1}^{i} (p_{1j} + p_{2j} + p_{3j} + p_{4j} + p_{5j}) \tag{9}$$

2.3. Performance Score Visualization

The study will demonstrate the progression of matches using the previously established performance scoring model[9].

As shown in table 5, Taking the match identified by the code "2023-wimbledon-1701" mentioned in the query as an example. Through performance scoring model, the study first obtain the performance score of each player for every point in the match (partial data from the match):

Table 5. The performance score situation for each point in the match

break_point_win_p1	consecutive_points_p1	hold_point_win_p1	p1_distance_run	be_break_point_win_p1	perform_points
0	1	0	0.0380640	0	0.124561
1	2	0	0.1244057	0	0.352898
1	3	0	0.0321806	0	0.477448
0	1	0	0.0329942	0	0.124560
1	2	0	0.0866106	0	0.352895
0	1	0	0.0490374	0	0.124561
0	1	0	0.1565527	0	0.124570
1	2	0	0.1558534	0	0.352900
0	1	0	0.1957666	0	0.124573
0	1	0	0.0601184	0	0.922562

Subsequently, aggregate performance scores for each point, resulting in a line graph that displays the cumulative performance scores of both players at various time points throughout the match.

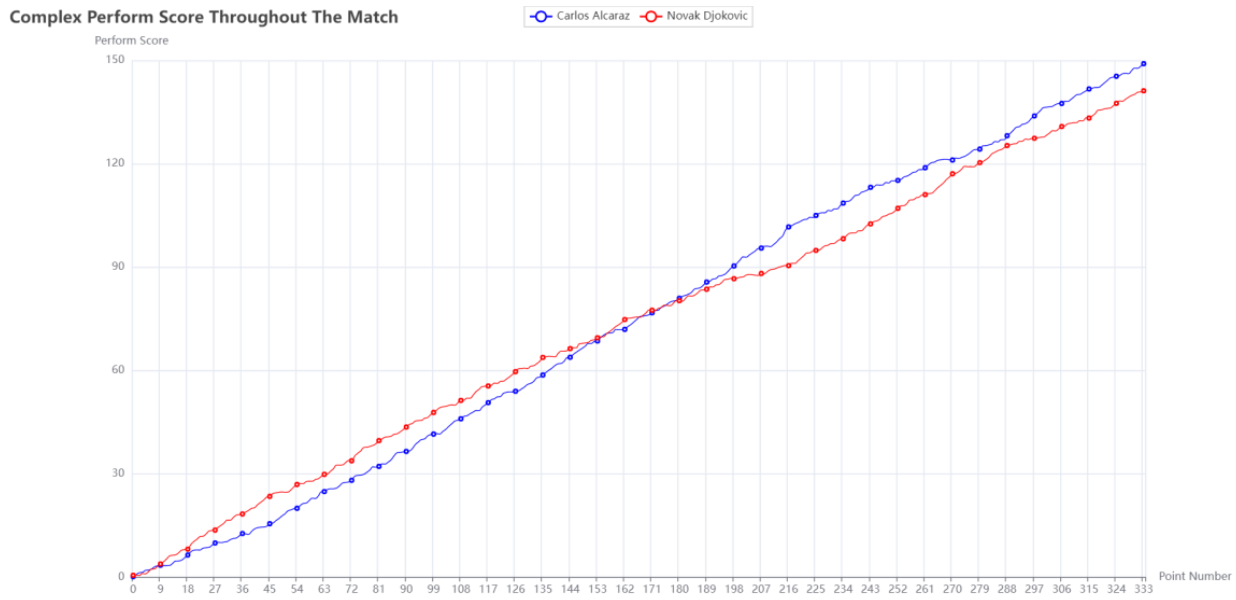


Fig. 2 Trend of performance scores of both sides in 1701 matches

From the fig.2, it is evident that Novak Djokovic's performance score held a comprehensive advantage at the start of match, but was soon overtaken by Carlos Alcaraz, who then maintained the lead until the end of the match. This closely mirrors the actual sequence of events described in the query, suggesting that the performance scoring model the study have developed is capable of effectively illustrating the progression of matches and clearly identifying which player exhibits superior performance at certain stages.

2.4. Further Feature Selection

With this in mind, the study planned to develop a model on Momentum to investigate whether there is a direct correlation between Momentum and the course of the match. Initially, based on the five characteristics identified in Question 1 - “break points”, “break points saved”, “service games lost”, “consecutive points” and ‘distance run’-were further filtered. It has been proposed that momentum may be primarily related to the number of successful breaks, the number of successful holds, the maximum sequence of consecutive points scored, and the frequency of breaks within a service game. To test this hypothesis, a Support Vector Machine (SVM) model was used to assess the contribution of these four features in determining the outcome of the match.

The objective of the Support Vector Machine is to identify a hyperplane $w \cdot x + b = 0$,that maximizes the margin between positive and negative samples. This can be achieved by solving the following optimization problem:

$$\min_{w,b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{N} \xi_i \tag{10}$$

The constraint condition is:

$$y_i (w \cdot x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \forall_i \tag{11}$$

Among them, C is the regularization parameter, ξ_i is the slack variable. Similarly, we divided the dataset into an 80% training set and a 20% test set yielding a model accuracy of 0.9705882352941176. In this context, the derived contribution values for the 4 features are deemed to possess academic relevance, with respective contributions outlined as Table 6.

Table 6. Feature Contributions

Feature Name	break_times	hold_times	be_break_times	max_consecutive_points
contribution	0.6669862	1.667298658	-1.167252241	0.333254649

2.5. Momentum Score Evaluation Model

To investigate whether there is a direct correlation between momentum and the outcome of a match, it is necessary to examine the relationship between quantified momentum scores and winning or losing. Herein, the study initially employ the Wilcoxon rank-sum test, with the following steps:

First, the study established the null hypothesis(H_0):the median of two sample datasets is equal, simultaneously established the alternative hypothesis(H_1):themedians of the two sample datasets are not equal. Subsequently, the two samples are combined into one dataset, disregarding the individual labels of the samples, and the combined dataset is then ranked. Next, the rank sums for each of the two samples are calculated, denoted as R_1 *and* R_2.

Subsequently, the study calculated the Wilcoxon rank-sum test statistic using rank sums $R_1$and R_2. Here employs the Mann-Whitney U statUistic[10]. The formula for calculating the U statistic is as follows:

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2} \tag{12}$$

$$U_2 = R_2 - \frac{n_2(n_2+1)}{2} \tag{13}$$

Among them,n_1, n_2 are the sizes of two samples respectively. Finally, based on sample size and the corresponding U-value, the statistic is calculated to get p value. Of which, the statistics: 340339.0, p value: 0.30149513040022646

Through$p \geqslant 0.05$, it can be concluded that the null hypothesis is valid, i.e., there is no significant difference between the two samples, indicating that there is a link between momentum differences and match won.

2.6. Spearman's correlation coefficient

The Wilcoxon rank-sum test preliminarily indicates that there is no significant difference between the momentum score differences and the outcomes of matches in terms of winning or losing, but it does not demonstrate their interconnection. To further substantiate the impact of momentum score differences on the match trajectory, specifically on match victories, we calculated the correlation coefficient between momentum score differences and match outcomes.

To explore the correlation between the two datasets, the study first tested whether the sample data follow a normal distribution to select the most appropriate method.

Similarly, for the momentum score difference data, the study formulated the null hypothesis: assuming it follows a normal distribution, as shown in table 7, calculated using the Shapiro-Wilk statistic to obtain the p value:

Table 7. Normality Test Results

	Shapiro-Wilk Test Statistics	p-value
Momentum Score Difference	0.985577345	1.85E-09

It is evident that the p value is significantly lower than the significance level (0.05), thereby refuting the null hypothesis, which implies that the momentum score difference data do not follow a normal distribution. Consequently, the study propose the use of Spearman's rank correlation coefficient to analyze this set of non-normally distributed data, with the calculation formula as follows:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{14}$$

Among them d_i denotes the difference in rank values for the i_{th} pair of data, where n represents the total number of observed samples.

To visualize the results, we plotted scatter diagrams depicting the distribution of momentum score differences under conditions of loss or victory, and calculated the Spearman's rank correlation coefficient between the momentum score differences and the outcomes of win or loss, as fig.3.

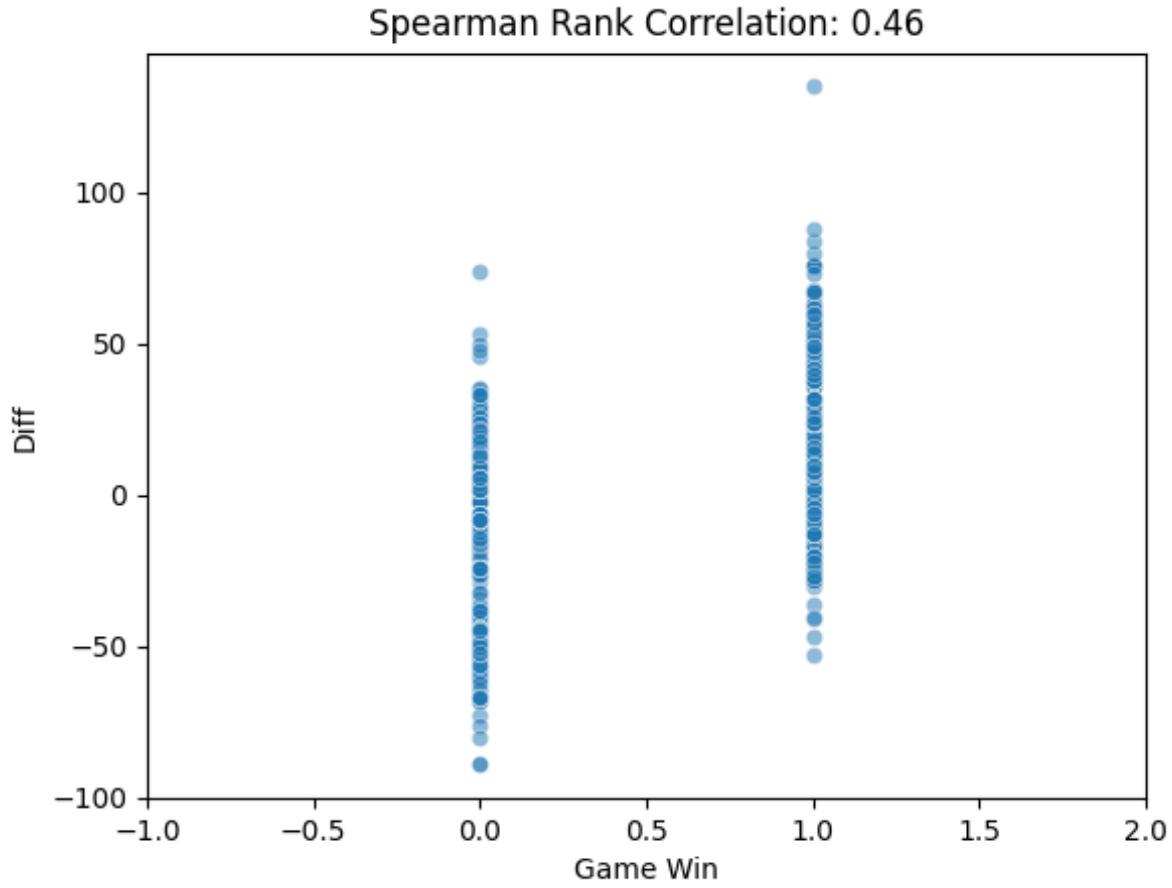


Fig. 3 Scatterplot of the difference in momentum scores

Spearman's correlation coefficient is 0.46, according to the variable correlation strength table 8:

Table 8. Variable Correlation Strength Table

Degrees of Relevance	Highly correlated	Significantly correlated	Moderately correlated	Marginally correlated	Negligibly or not correlated
Absolute value of correlation coefficient	0.8-1	0.6-0.8	0.4-0.6	0.2-0.4	0-0.2

This suggests that momentum score differences are positively and equivalently correlated with victory.

From the graph, it is evident that the x-axis represents 0.0 and 1.0, corresponding to Player 2 and Player 1 winning the match, respectively, with a noticeable vertical shift in the graph. Under the given sample size, when Player 2 wins, the momentum score differences are predominantly on the negative axis; conversely, when Player 1 wins, the momentum score differences tend to be on the positive axis. This further demonstrates that momentum score differences have a direct impact on the outcomes of matches.

To quantify the specific distribution of momentum score differences under conditions of loss or victory, we also analyzed the proportion of momentum score differences across different score intervals and utilized Kernel Density Estimate (KDE) to plot the curves, yielding the following fig.4:

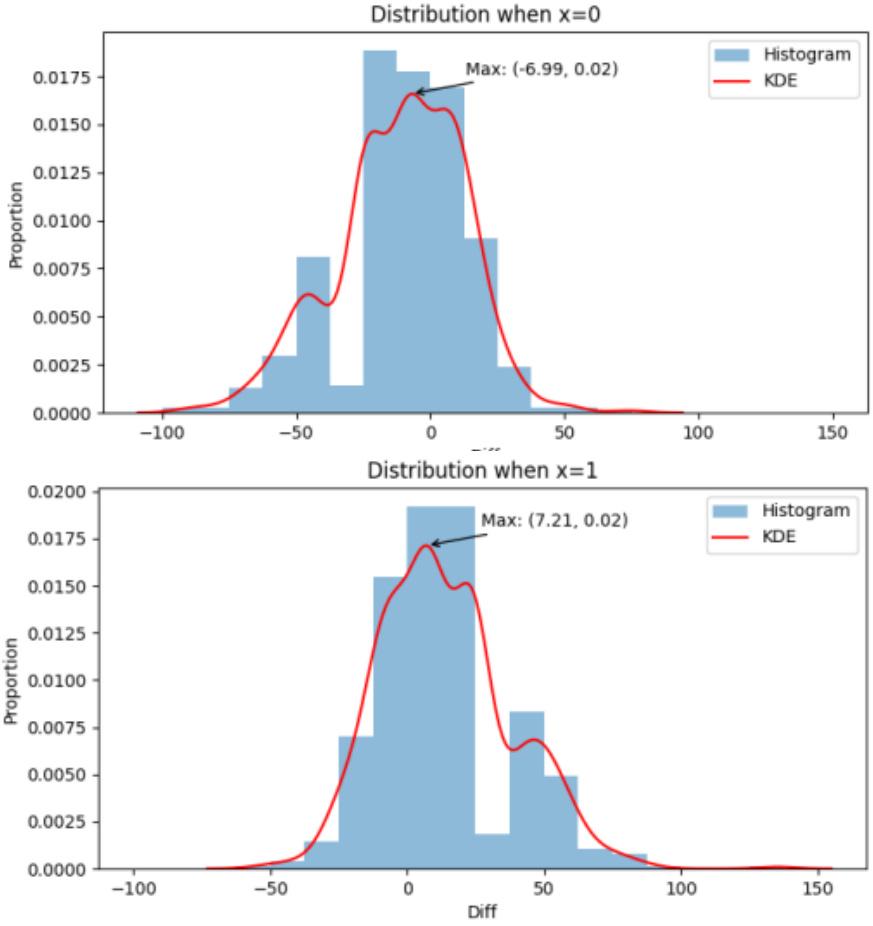


Fig. 4 Momentum Score Difference by Score Band

The histogram represents the relationship between momentum score differences and the proportion of such differences when a player wins, while the curve graph is the result of a linear fit to the histogram. Observing the KDE curve, it is noted that the largest proportion of momentum score difference for Player 1 winning the match is (7.21, 0.02), and for Player 2 winning, the momentum score difference proportion is (-6.99, 0.02). From the overall distribution, it is evident that there is a relatively close relationship between momentum score differences and the outcomes of matches. Therefore, aspects of the match progression, including sequences of consecutive wins and turning points, have a certain connection with momentum.

3. Conclusion

This study employed a combination of Particle Swarm Optimization and Random Forest algorithms to process and select features from a dataset, optimizing the Random Forest model parameters through Particle Swarm Optimization. This approach not only significantly enhanced model accuracy but also established a ranking of feature importance, thereby validating the model's effectiveness and applicational value.

The performance level scoring model and momentum scoring model developed in this research, based on weighted calculations of selected features, demonstrated high predictive accuracy and practicality in real-world applications. Particularly, the momentum scoring model revealed potential connections between consecutive winning sequences and match outcomes, offering new insights into strategic shifts during games.

Further explorations of feature interactions and their impact on match results were conducted using the Wilcoxon rank-sum test and Spearman's correlation coefficient. These statistical methods underscored the importance of advanced data analysis techniques in handling complex datasets,

enhancing the explanatory power of the models and providing methodological guidance for future research.

Overall, the comprehensive model frameworks and methodologies proposed in this paper proved to be efficient and accurate in handling large datasets, demonstrating their potential application and value in the field of competitive sports data analysis. These findings not only advance the level of data processing and analysis techniques but also equip researchers in the field with robust tools and new directions for further investigation.

References

[1] PRAKASHRAO KAULGUD S, HULIPALLED V, SOMANAGOUDA PATIL S, et al. Effective Prostate Cancer Detection using Enhanced Particle Swarm Optimization Algorithm with Random Forest on the Microarray Data[J/OL]. International journal of electrical and computer engineering systems, 2023: 251-258.

[2] AWADALLA M, NOUPOUE Y Y Y, TANDOGDU Y, et al. Regression Coefficient Derivation via Fractional Calculus Framework[J/OL]. Journal of Mathematics, 2022, 2022: 1-9.

[3] ZHANG M, QIN X. Research on Satisfaction Evaluation Model Based on Ordered Logistic Regression[C/OL]//Proceedings of the 2022 5th International Conference on E-Business, Information Management and Computer Science. 2022.

[4] THARWAT A, HASSANIEN A E. Quantum-Behaved Particle Swarm Optimization for Parameter Optimization of Support Vector Machine[J/OL]. Journal of Classification, 2019, 36(3): 576-598.

[5] DAO P B. On Wilcoxon rank sum test for condition monitoring and fault detection of wind turbines[J/OL]. Applied Energy, 2022: 119209.

[6] WANG S, ZHENG J. MonoSKD: General Distillation Framework for Monocular 3D Object Detection via Spearman Correlation Coefficient[J]. 2023.

[7] Mohamed G. El-Shafiey, El-Sayed A. El-Dahshan, Ahmed Hagag, Manal A. Ismail. A hybrid GA and PSO optimized approach for heart-disease prediction based on random forest [J]. Multimedia Tools and Applications, 2022, 3(8).

[8] ASRI A A, MALLAH G A, HASSAN N U, REHMAN N U. An Improved Particle Swarm Optimization Algorithm for Data Classification[J/OL]. Applied Sciences, 2023, 13(1).

[9] Marina Evers, Adrian Derstroff, Simon Leistikow, Tom Schneider, Larissa Mallepree, Jan Stampke, Moritz Leisgang, Sebastian Sprafke, Melina Schuhl, Niklas Krefft, Felix Droese, Lars Linsen. Visual analytics of soccer player performance using objective ratings[J]. Information Visualization, 2024, 1(12).

[10] XIE J, LI L. Comments on the utilization of Mann-Whitney U test and Kaplan-Meier method[J/OL]. Journal of Gynecologic Oncology, 2021, 32(3).