LGBM-Based Real-Time Analysis of Tennis Player Performance

Jiatian Shang [1, *], Ziyao Zhou [2], Yumeng Sheng [1], Yiran Wang [3]

[1] School of Mathematics and Statistics, Ningbo University, Ningbo, China, 315211

[2] School of Finance (Chongqing University of Finance), Chongqing Technology and Business University, Chongqing, China, 400067

[3] College of Digital Technology and Engineering, Ningbo Univeasity of Finance & Economics, Ningbo, China, 315175

* Corresponding Author Email: 15127010903@163.com

Abstract. This study aims to develop a real-time performance evaluation model for tennis players, enabling the quantification of match dynamics and the identification of key factors influencing scoring outcomes. Utilizing data from the Wimbledon Open dataset, the research applies preprocessing techniques such as outlier removal to construct a two-level indicator system encompassing psychological, physiological, and service-related variables. A binary logistic regression analysis is conducted to test indicator significance, followed by a comparative evaluation of five machine learning algorithms: LGBM, XGB, SVC, MLP, and LR. Results demonstrate that the LGBM algorithm outperforms others, achieving an accuracy of 0.69 and an AUC of 0.77. The model effectively captures real-time player performance and provides actionable data support for tennis training and tactical decision-making. Future improvements may focus on refining the indicator system to enhance the model's generalizability and predictive robustness.

Keywords: Tennis Players, Real-Time Performance, LGBM Algorithm, Machine Learning, Indicator system.

1. Introduction

In modern sports, quantitative analysis of athletes' real-time performance is crucial for optimizing training strategies and tactical decisions. As a high-intensity and highly competitive sport, tennis involves multiple interacting factors—such as physiological state, psychological fluctuations, and serving advantage—that together create uncertainty in match outcomes. The 2023 Wimbledon Men's Singles Final between Carlos Alcaraz and Novak Djokovic highlighted frequent momentum shifts and raised a critical question: can data-driven models capture real-time turning points and explain the underlying factors affecting point outcomes.

Research on psychological momentum in tennis has progressed from traditional statistical modeling to multidimensional evaluation frameworks and, more recently, to machine learning-driven prediction. Ge first applied factor analysis to construct quantitative indices of momentum based on match dynamics, providing a foundation for early statistical modeling [1]. Building on this, Wu introduced an AHP–Entropy Weight Method (AHP-EWM) framework to quantify momentum characteristics and evaluate their impact on scoring probability [2], while Lv et al. further enhanced model robustness by developing an entropy-weighted momentum calculation model, improving performance under competitive pressure [3].

As the field evolved, researchers integrated broader performance indicators to better capture the multifaceted nature of momentum. He et al. developed a comprehensive evaluation model combining psychological resilience, technical proficiency, and physiological responses to provide deeper insights into player performance [4]. Wang et al. proposed a multidimensional momentum chain model based on differential equations, enabling dynamic representation of momentum changes within matches [5]. Most recently, Lv et al. applied CatBoost regression and random forest algorithms to predict momentum transitions, underscoring the potential of machine learning in modeling complex match dynamics [6].

Despite these advances, existing studies remain limited in terms of real-time applicability. Early statistical models [1-3] largely depend on post-match data and cannot reflect real-time fluctuations, while dynamic frameworks [4-5] still lack responsiveness during live play. Although machine learning approaches [6] show promise for predictive analytics, practical in-match tracking of momentum remains underexplored, highlighting the need for dynamic, data-driven models capable of capturing and forecasting momentum shifts as they occur.

Therefore, this study develops a real-time performance evaluation model based on data from the 2023 Wimbledon Men's Singles Final. The objectives are threefold: (1) identify key factors influencing player scoring, including serving advantage, physiological state, and psychological condition; (2) compare and optimize machine learning algorithms to achieve accurate, real-time assessment of performance and momentum shifts; and (3) visualize the model's application in critical match scenarios, providing a scientific basis for training optimization and tactical decision-making.

2. Research Methodology

2.1. Date Processing

The data utilized in this study comes from the dataset of the second-round men's singles match of the 2023 Wimbledon Championships. This dataset comprehensively records the match process over time across various scenarios, including scoring situations, serving conditions, and instances of broken serves, among others.

This study pre-processed this dataset by removing missing values and outliers, and normalizing the data as follows:

(1) Delete the rows with missing data in the "speed_mph" column.

Since the missing data in this column is not conducive to the next step in the data analysis, this study chose to delete them and all the data in the row they are in.

(2) Delete the columns "serve_width", "serve_depth" and "return_depth".

Since the data in these three columns have more missing values and the data in these three columns are not very useful for model building and analysis, this study choose to delete them for the sake of normal data analysis.

(3) Remove outliers in the "point_no" and "speed_mph" columns.

In the tennis rules, the score requirement is only "15", "30", "40", etc., and there is no "1", "2", "3", etc. ", "2", "3" and so on, so the error data that does not match the normal situation is deleted.

(4) Remove outliers in the "point_no" and "speed_mph" columns.

(5) This study choose the "min-max normalization" method to normalize the data, for each attribute $A_n(n = 1,2,\cdots 15,16)$, set $minA_n$ and $maxA_n$ as the minimum and maximum values of attribute A_n, respectively, and normalize the original value of A_n, x, by mapping the value of x' to the value of x' in the interval [0,1], as follows:

$$x' = \frac{x - minA_n}{maxA_n - minA_n} \ , n = 1,2,\cdots,15,16 \tag{1}$$

2.2. Binary logistic regression

Binary logistic regression is a statistical modeling technique widely employed for binary classification problems, aiming to predict the probability of a dependent variable taking a particular class (typically coded as 0 or 1) based on a set of independent variables (features). Unlike linear regression, logistic regression utilizes the logit function to establish a nonlinear mapping, thereby constraining the predicted values within the [0,1] interval.

In this study, binary logistic regression is applied to analyze the binary outcomes of players scoring or not scoring in a given rally, enabling the assessment of the significance of multiple indicators (e.g., psychological momentum, physiological momentum) on the scoring results. This approach facilitates the quantification of the relationship between each indicator and the scoring probability, thereby identifying key influencing factors and providing a theoretical basis for feature selection and

optimization in subsequent machine learning models. A notable advantage of this method lies in its high interpretability, as the sign and magnitude of the coefficients directly indicate the direction and significance of each feature's impact on the target variable.

Since a player's "score" and "no score" constitute a dichotomous variable, a binary logistic regression model is used to predict whether a player will score points or not, and thus to judge the player's performance.

The dichotomous variable of "score or not" was used as the dependent variable: "1" for score and "0" for no score. At the same time, the 16 momentum calculation related indicators in Figure 2 are used as independent variables, labeled as $x_1, x_2, x_3, \cdots, x_{16}$. Assuming that p is the probability of a player scoring a point in the game, then 1-p is the probability of not scoring a point, which is what this study get:

$$f(x) = B_0 + B_1x_1 + B_2x_2 + B_3x_3 + \cdots + B_{16}x_{16} \tag{2}$$

where B_n is the coefficient of each indicator and has no real meaning.

Also since the probability of scoring is defined in the model as $p = \frac{e^{f(x)}}{1+e^{f(x)}}$,

$$p = \frac{e^{B_0+B_1x_1+B_2x_2+B_3x_3+\cdots+B_{16}x_{16}}}{1 + e^{B_0+B_1x_1+B_2x_2+B_3x_3+\cdots+B_{16}x_{16}}} \tag{3}$$

Thus the probability of not scoring is obtained:

$$1 - p = \frac{1}{1 + e^{B_0+B_1x_1+B_2x_2+B_3x_3+\cdots+B_{16}x_{16}}} \tag{4}$$

The formula that defines the odds of winning is: $\text{odds} = \frac{p}{1-p}$, thus:

$$odds = \frac{p}{1-p} = e^{B_0+B_1x_1+B_2x_2+B_3x_3+\cdots+B_{16}x_{16}} \tag{5}$$

This in turn gives the regression model as:

$$\ln\left(\frac{p}{1-p}\right) = f(x) = B_0 + B_1x_1 + B_2x_2 + B_3x_3 + \cdots + B_{16}x_{16} \tag{6}$$

2.3. LGBM-algorithm

In addition to binary logistic regression, this study explores five classical machine learning algorithms commonly used for classification tasks, namely Light Gradient Boosting Machine (LGBM), XGBoost (XGB), Support Vector Classifier (SVC), Multilayer Perceptron (MLP), and Logistic Regression (LR). Among these, the LGBM algorithm was selected as the primary modeling approach due to its superior performance and ability to handle complex non-linear relationships efficiently. The detailed comparison and evaluation process among the five algorithms will be presented in a later section, while this part focuses on introducing the LGBM algorithm and its methodological advantages.

LGBM is an efficient machine learning algorithm based on the gradient boosting framework. It integrates multiple weak learners, typically decision trees, to form a strong classifier and enhance predictive performance. Compared to traditional Gradient Boosting Decision Trees (GBDT), LGBM demonstrates advantages in training speed, memory efficiency, and the ability to handle large-scale data, employing optimizations such as histogram-based splitting and feature parallelism.

LGBM iteratively trains models by using newly generated decision trees to fit the residuals of previous iterations, thereby reducing prediction errors. In this study, LGBM is applied in conjunction with 16 momentum-related indicators. By leveraging first- and second-order gradient information to evaluate feature importance and adopting a leaf-wise tree growth strategy to maximize information gain, LGBM effectively captures the complex nonlinear relationships inherent in players' scoring processes and accurately models momentum variations during a match.

Within this research, LGBM is employed to predict players' scoring probabilities and real-time performance, enabling the quantification of momentum shifts and supporting subsequent visualization analyses. Through comparison with other machine learning models (XGB, SVC, MLP, and LR), LGBM demonstrated superior overall performance, making it the primary modeling method for reliable dynamic match analysis.

2.4. Model Evaluation Metrics

To evaluate the predictive performance and reliability of the constructed models, several standard evaluation metrics are employed. These metrics enable a clear comparison of different algorithms and validate the effectiveness of the selected LGBM model.

First, based on all the confusion matrices of the above species models, this study evaluate them using five metrics, which are:

(1) Accuracy: The total number of correct predictions by the model as a proportion of the total number of predictions. The formula is expressed as:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{7}$$

(2) Precision: Precision is the ratio of correctly predicted positive observations to the total predicted positives. It focuses on the accuracy of the model when it predicts the positive class and is useful in situations where false positives are critical. The formula is expressed as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{8}$$

(3) Recall: Recall, also known as sensitivity or true positive rate, is the ratio of correctly predicted positive observations to all actual positives. It assesses the model's ability to capture all positive instances and is relevant in situations where false negatives are important. The formula is expressed as:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{9}$$

(4) F1 Score: The F1 Score is the harmonic mean of precision and recall. It pro-vides a balance between precision and recall, making it useful for evaluating models in scenarios with imbalanced class distributions. The formula is ex-pressed as:

$$\text{F1 Score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2 \cdot \text{precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10}$$

(5) Area Under the Receiver Operating Characteristic Curve (AUC)AUC-ROC evaluates the performance of a binary classifier by plotting the Re-ceiver Operating Characteristic (ROC) curve and calculating the area under it. It measures the trade-off between true positive rate and false positive rate at different thresholds.

In the above equation, TP, FP, TN, and FN are True Positive, False Positive, True Nega-tive, and False Negative, respectively.

3. Model Construction and Solution

3.1. Constructing a Player Performance Indicator System

First, this study classify momentum indicators into three categories: psychological, physiological, and combined psychological–physiological factors. This is because part of the factors increase or decrease one's per-ceived probability of success by modifying his and his opponent's perceptions and impres-sions of each other, such as whether or not he has made a service error, his current progress towards the lead, and so on; part of the factors are used to compute momentum by measuring his level of physical fatigue, and his personal technical ability, such as mileage of the run in a given period of time, and the ratio of the number of serves to the number of points scored from the net, and so on; and finally the the factor of whether or not to serve, which is sepa-rated out because it is related to both psychological and physiological momentum; serving itself is a natural advantage and usually has a higher probability of scoring, benefiting both psychologically and physiologically.

Secondly, this study perform a secondary categorization based on the above categorization, which is as shown in Figure 1.

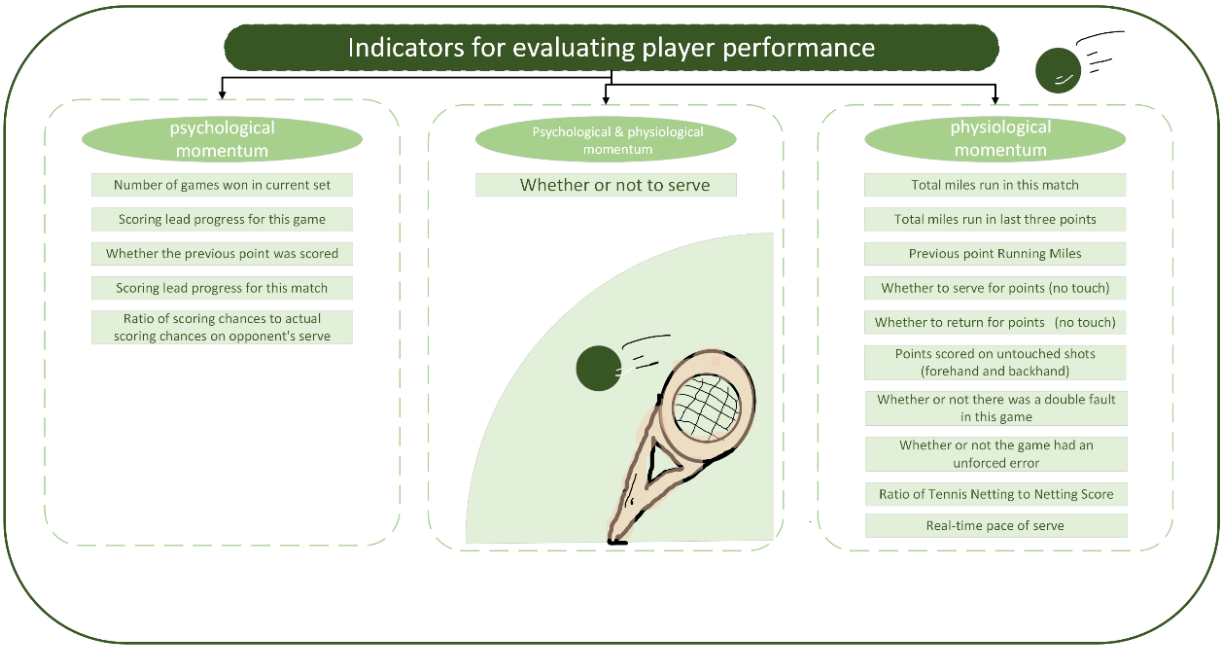


Figure 1. Indicators for evaluating player performance

3.2. Model Construction and Solution Based on Binary Logistic Regression

In this study, binary logistic regression was performed on the standardized data using SPSS, and the results are shown in Table 1.

Table 1. Classification performance of logistic regression

	Predicted : 0	Predicted : 1
Actual : 0	1944	1082
Actual : 1	977	2221

The data from SPSS shows that the LOSISTIC regression yielded an accuracy of 66.9%, and the predictive accuracy of the actual win was 64.2, and the probability of the predictive accuracy of the actual negative was 69.4%, although to some extent the value of these two probabilities is not very high, but the purpose of the regression analysis is to test whether the model's indicator system is able to have a significant effect on the player's performance and the winners and losers of the game.The results of the regression test are shown in Table 2.

Table 2. Results of logistic regression

	B	Standard Error	Wald	Degrees of Freedom	Significance	Exp (B)
x_1	-0.229	0.244	0.882	1	0.348	0.795
x_2	0.955	0.809	1.395	1	0.238	2.599
x_3	-7.161	3.538	4.098	1	0.043	0.001
x_4	1.422	0.603	5.567	1	0.018	4.145
x_5	0.102	0.280	0.133	1	0.715	1.108
x_6	0.219	0.193	1.279	1	0.258	1.244
x_7	0.455	0.169	7.201	1	0.007	1.576
x_8	0.472	0.153	9.491	1	0.002	1.603
x_9	0.064	0.146	0.195	1	0.659	1.066
x_{10}	0.292	0.170	2.942	1	0.086	1.339
x_{11}	2.155	0.493	19.129	1	0.000	8.628
x_{12}	0.774	0.412	3.532	1	0.060	2.168
x_{13}	-2.533	0.991	6.527	1	0.011	0.079
x_{14}	0.411	1.374	0.090	1	0.765	1.508
x_{15}	-1.455	1.326	1.204	1	0.272	0.233
x_{16}	8.719	4.497	3.758	1	0.053	6115.731
Constant	-0.925	0.580	2.546	1	0.111	0.397

From the Table 2, it can be learned that nearly half of the indicators have a P-value less than 0.05, they are $x_3, x_4, x_7, x_8, x_{10}, x_{11}, x_{13}$, and since their P-values are less than 0.05, it can be concluded that they have a significant effect on the dependent variable y. So, it can be said that the player's physical state and the player's mental capacity as well as the psychological state will significantly affect the player's performance and the the result of the match.

3.3. Real-Time Performance Model Based on LGBM Algorithm

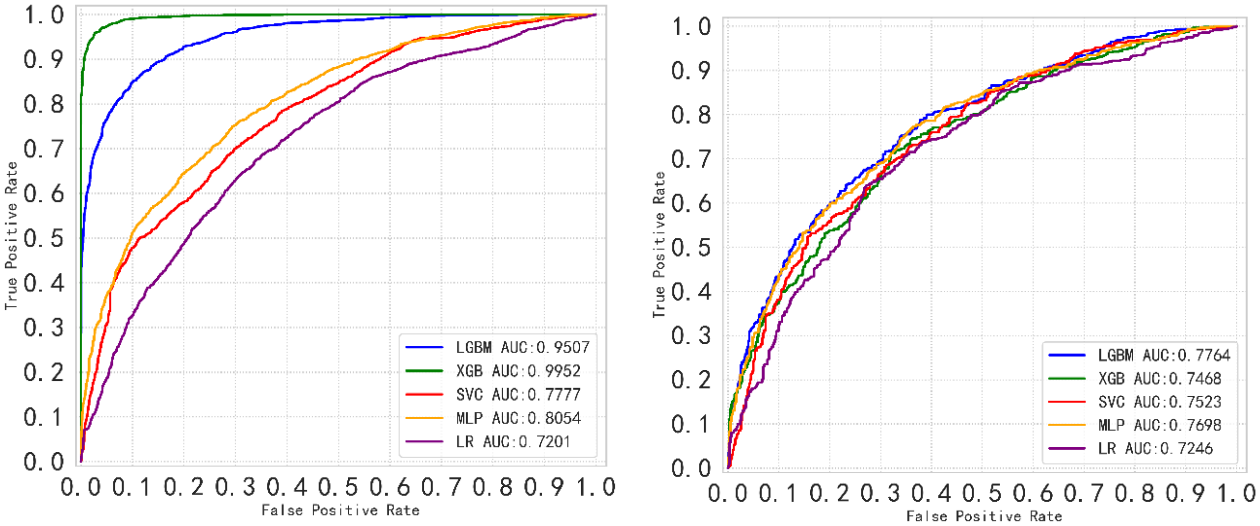
Firstly, five machine learning models—Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), eXtreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LGBM)—were selected along with corresponding evaluation indicators. The performance of each model was then assessed using 5-fold cross-validation [9] based on five evaluation metrics. The numerical results for each model across the selected indicators are presented in the Table 3.

Table 3. Results of the model evaluation

	ACC	Recall	Precision	F1	AUC
LGBM	0.69	0.69	0.7	0.69	0.77
XGB	0.67	0.68	0.68	0.68	0.75
SVC	0.67	0.75	0.7	0.67	0.75
MLP	0.69	0.66	0.71	0.68	0.76
LR	0.67	0.69	0.67	0.68	0.72

As shown in Table 3, the LGBM model demonstrates the best overall performance [10], with Accuracy, Recall, Precision, F1 Score, and AUC values of 0.69, 0.69, 0.70, 0.69, and 0.77, respectively. Its Accuracy value is the highest, indicating that the model is more accurate. This study focuses on the

F1 Score, as it better reflects the balance between Recall and Precision; a value closer to 1 suggests a better trade-off between the two. Moreover, the model achieves the largest AUC value, further confirming that the LGBM model performs best.



(a) ROC curve for training set (b) ROC curve for testing set

Figure 2. ROC curve

Next, the ROC curves of the five models were plotted to further validate the performance of the LGBM model, as illustrated in Figure 2. The ROC curves reflect the changes of TPR and FPR under different thresholds. The closer the ROC curve approaches the upper-left corner, the better the model's performance, as this reflects a higher TPR and a lower FPR. Obviously, the LGBM model performs the best. Where, $TPR = \frac{TP}{TP+FN}$, $FPR = \frac{FP}{FP+TN}$.

Finally, Python was used to obtain the contribution of each indicator, as shown in Figure 3.

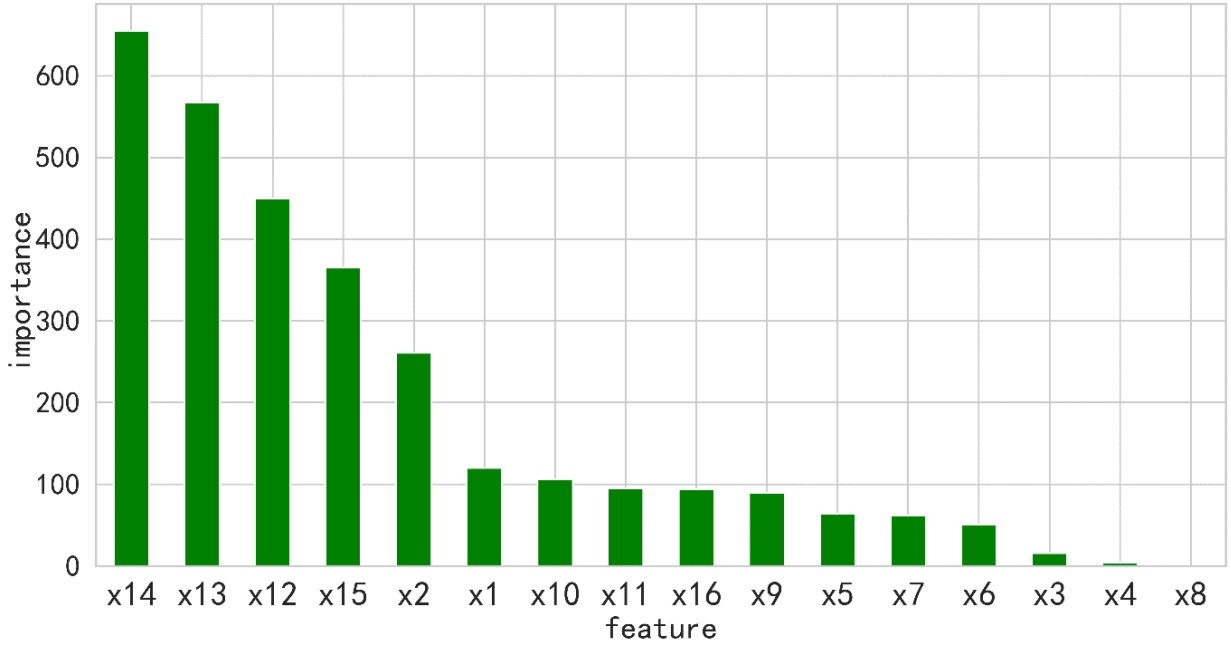


Figure 3. Contribution of indicators

Based on the aforementioned analysis, this study decided to retrain the model and select the best performing LGBM model. This study aimed to visualize in real time how this model would perform in the classic match between 20-year-old Spanish star Carlos Alcaraz and 36-year-old Novak Djokovic in the men's singles final at Wimbledon 2023.

As illustrated in Figure 4, this study selected Carlos Alcaraz for analysis and visualized the model's real-time predictions of the player's performance. This visualization provides an intuitive understanding of performance fluctuations during the match and offers insights into the accuracy of the model's predictions.

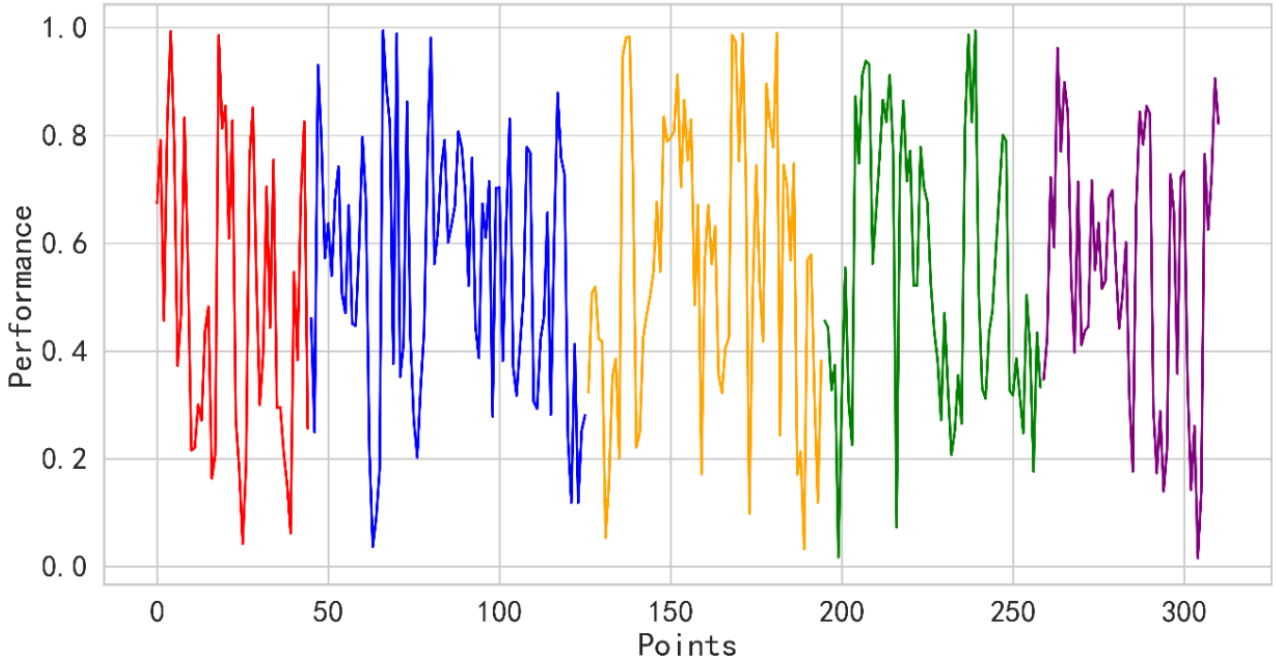


Figure 4. Real-time performance of the player

In fact, the accuracy of predicting scoring points has an upper bound, which aligns with intuitive reasoning. Because of the complexity of the factors that influence whether a player can score, the prediction of a player's score may be heavily influenced by noise when these complex factors are not taken into account. As a result, the model may only be able to achieve a certain level of accuracy in accurately determining how well a player ac-tually scores, at about 70%.

The analysis shows that the model's predictions are highly consistent with the players' actual scoring in this classic match. Specifically, in Carlos Alcaraz's victories, the model successfully captures his high-momentum state, whereas in defeats, it reflects a lower momentum. These findings indicate that the proposed model remains valid and effectively captures a player's scoring with considerable accuracy. Such an analysis helps to confirm the value of the model's application in real matches and to understand its sensitivity to match dynamics.

4. Conclusions

This study developed a real-time performance evaluation model for tennis players based on data from the 2023 Wimbledon Men's Singles Final. Through data preprocessing and the construction of a comprehensive indicator system, combined with binary logistic regression and machine learning techniques, the Light Gradient Boosting Machine (LGBM) algorithm was identified as the optimal model. The results indicate that physiological condition, psychological resilience, and serving advantage are key determinants of point outcomes. The LGBM model effectively quantified momentum shifts and achieved an accuracy of 0.69, offering technical support for targeted improvements in physical fitness, psychological regulation, and serving performance in tennis training.

Despite the model's promising performance, certain limitations remain. The historical indicator system did not fully capture player history or opponent strength, which may have impacted predictive generalizability. Future studies should incorporate more comprehensive historical data to refine the indicator framework and integrate deep learning approaches to enhance the model's capacity for interpreting complex match dynamics. These improvements will contribute to more personalized and data-driven development in professional tennis.

References

[1] Ge Y. Quantitative modelling of momentum in tennis based on factor analysis [J]. Transactions on Computer Science and Intelligent Systems Research, 2024, 5: 1204-1213.

[2] Wu W. Quantification of momentum in tennis matches and its impact: a study based on AHP-EWM method and data analysis [J]. Highlights in Science, Engineering and Technology, 2024, 100: 142-149.

[3] Lv Y, Yang Z, Zong T. Research on Tennis Players' Momentum Calculation Model Based on Entropy Weight Method [J]. Journal of Electronics and Information Science, 2024, 9 (3): 19-26.

[4] He J, Yang C, Cai M. Research on Predicting the Winning Rate of Momentum Tennis Athletes [J]. Highlights in Science, Engineering and Technology, 2024, 100: 661-669.

[5] Wang J, Guo S, Zhou Y. A multidimensional momentum chain model for tennis matches based on difference equations [J]. PLOS ONE, 2024, 19 (12): e0316542.

[6] Lv X, Gu D, Liu X, Dong J, Li Y. Momentum prediction models of tennis match based on CatBoost regression and random forest algorithms [J]. Scientific Reports, 2024, 14: 18834.

[7] Iso-Ahola S E, Mobily K E. Psychological momentum and performance in competitive sports [J]. Journal of Sport & Exercise Psychology, 1980, 2 (3): 211-222.

[8] Meier P, Flepp R, Ruedisser M, Franck E. Separating psychological momentum from strategic momentum: Evidence from men's professional tennis [J]. Journal of Economic Psychology, 2020, 78: 102269.

[9] Liu C, Yang J, Cui Y. Quantifying and predicting momentum in tennis match via machine learning approach [J]. International Journal of Racket Sports Science, 2025, 7 (1): 46-58.

[10] Liang R, Shen L, Zeng X, Zhong Y. Tennis player momentum analysis and match flow prediction model based on PCA [J]. Highlights in Science, Engineering and Technology, 2024, 115: 73-80.