Proceedings of Machine Learning Research vol 278:1–9, 2025 2025 2nd International Conference on Machine Learning and Intelligent Computing

Uncovering the Secrets of Momentum Hidden in the Game of Tennis

Haoqian Huo HUOHAOQIAN@STU.ZZU.EDU.CN
Zhengzhou University, Zhengzhou, China

Editors: Nianyin Zeng, Ram Bilas Pachori and Dongshu Wang

Abstract

Advances in sports technology have had a profound impact on the tennis game, not only improving the fairness and enjoyment of the game, but also changing the way players are trained and performance analyzed. This article builds a momentum evaluation model and deeply explores the impact of momentum on game results based on the game data set.

Before building the model, we cleaned and standardized the given data and classified it into four parts: fatigue level, psychological state, personal technical ability, and real-time conditions. Preliminary preparations were made for the construction and solution of the model.

We developed a comprehensive tennis player "momentum" evaluation model using Logistic-LGBM, employing point granularity and five-fold cross validation. This model dynamically assesses and captures real-time changes in player momentum. Our real-time visualization during the 2023 Wimbledon men's singles final revealed observable momentum trends. However, due to the complexity of factors affecting player scores, not accounting for them introduced significant noise and disrupted player scores. This insight serves as a foundation for refining the model.

Keywords: Logistic-LGBM, Point granularity, Five-fold cross validation method

1. Introduction

In various sports competitions, there are usually many factors involved in judging whether a player can win, and the higher the level of the competition and the higher the quality of the competition, the greater the impact of these different factors on whether the player can win.

Momentum in tennis has been a subject of research for decades, with studies analyzing its impact on match outcomes and player performance. Previous works have explored the psychological and physiological factors influencing momentum swings. For example, Kovacs and Ellenbecker (2011) discussed how fatigue impacts player momentum, while O' Donoghue (2013) investigated momentum shifts based on in-game statistics.

Sure, there're some meaningful advantages in this text: the research use a energetic Logistic-LGBM model to measure momentum as it happens, instead of just looking at things after the match is over. And then, while other studies usually focus on one factor at a time, this model brings together different elements like fatigue, technical skills, and the actual conditions during the game. This gives us a much clearer picture of how momentum changes.

As one of the top tennis events in the world, the two players in the 2023 Wimbledon men's singles final, Carlos Alcaraz and Novak Djokovic, have achieved every stage of the competition. Victory or failure is also closely related to the various factors mentioned previously.

This article conducts in-depth analysis and research on the information and data of the two players in the game to answer the questions mentioned below. In this article, we will solve the following problems:

- How does momentum influence match outcomes at different stages of a tennis match?

- What is the frequency and duration of momentum shifts during high-stakes matches?
- How do various factors such as fatigue level, technical ability, and psychological state contribute to momentum changes?
- Can a machine learning-based model accurately predict the impact of momentum on match results?

2. Assumptions and Justifications

In order to facilitate the establishment of the model, we made the following assumptions and simplifications based on the actual situation and tennis game rules. They are shown in Figure 1.

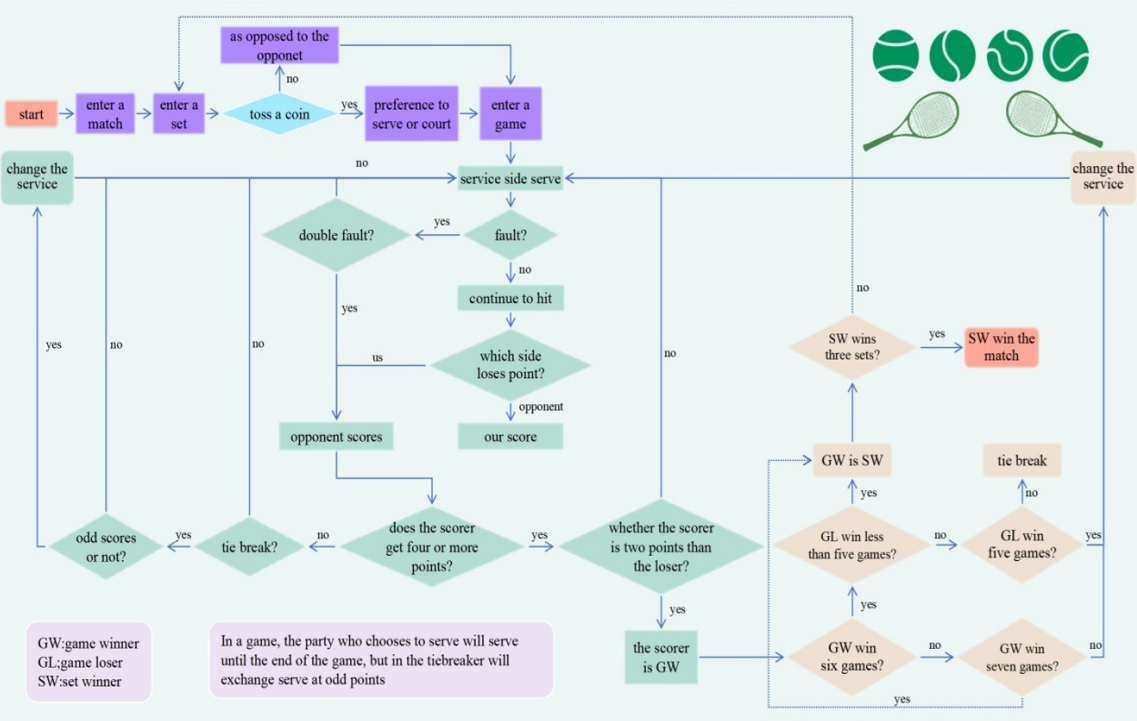

Figure 1: Schematic Diagram of Tennis Rules.

Assumption 1: Assume that the sample size of the attachment is large enough, the sample data is real, can reflect the specific situation, and eliminates erroneous data.

Assumption 2: Assume that the literature and conclusions cited in the article are correct and reliable.

Assumption 3: It is assumed that there will be no off-field factors that will affect the players during the game.

Assumption 4: It is assumed that a player's personal technical ability has a direct impact on his scoring performance.

Assumption 5: It is hypothesized that players may experience fatigue during the game to affect their performance later in the game.

Assumption 6: It is assumed that a player's success or failure in the game will affect his future performance.

Assumption 7: It is assumed that player scores are directly affected by personal skill level.

Assumption 8: In tennis matches, it is easier for the server to score than the receiver.

3. Notations

The key mathematical notations used in this paper are listed in Table 1.

Table 1: Notations used in this paper (Important variables are marked with an asterisk.)

Symbol	Description	Unit
gp*	The most influential factor, contributing significantly to momentum assessment	
mp	The set lead of this match	
sg	The number of games won in the current set	
s	Whether it is the server	
s1*	A key technical skill that heavily influences point outcomes	
s2*	Whether to score a counterattack in this game (no touch)	
s3	Whether the previous point was scored?	
v	Serve real-time pace	
sv	Is it an interaction term between serve and serve pace?	
f1*	Is there a double fault in this game?	
f2*	Are there any unforced errors in this game?	
d1	Total mileage in this match	m
d2	The total mileage in the last three points	m
d3*	A crucial fatigue-related factor affecting performance	m

4. Data Description

Before conducting data analysis, data availability must be ensured.

Step 1: In the data cleaning stage, we use the Pandas library in Python to read the file and convert its content into a DataFrame object. And use Python to check for missing values and outliers.

Step 2: Use the sklearn.impute.SimpleImputer class to fill the missing values in the selected numeric column with the median strategy. The specific performance is:

$$\begin{array}{ll} X_{i,j} & \text{if } X_{i,j} \text{ is not missing} \\ \text{median } (X_{i,j}) & \text{if } X_{i,j} \text{ is missing} \end{array} \tag{1}$$

Step 3: Apply sklearn.preprocessing.StandardScaler for feature scaling, converting each feature value into a standard normal distribution form of 0 and standard check as 1. The standardized formula is as follows:

$$\begin{aligned} b_{ij} &= \frac{a_{ij} - \bar{a}_j}{s_j}, i = 1, 2, \cdots m, j = 1, 2, \cdots, n \\ \bar{a}_j &= \frac{1}{m}\sum_{i=1}^{m} a_{ij}, s_j = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}(a_{ij} - \bar{a}_j)^2}, j = 1, 2, \cdots n \end{aligned} \tag{2}$$

5. Logistic-LGBM based "momentum" evaluation model of tennis players

When studying the key factors of players' scoring, we deeply explored factors such as fatigue, technical ability and game mentality, and constructed a comprehensive indicator system to more accurately evaluate players' scoring performance in the game. Using the statistical Logistic model (He and He, 2021) to perform label significance testing, we calculated various indicators on the tennis match data set, providing a solid data foundation for subsequent modeling. In the end, we chose the LGBM algorithm for modeling. Its advantage lies in evaluating player performance in real time and keenly capturing momentum changes in the game. This dynamic evaluation model improves the accurate prediction of player performance, provides a reliable basis for predicting game results, deepens the understanding of players' real-time tactical scores, and lays the foundation for further prediction and analysis.

5.1. Construction of indicator system

Index system of factors affecting tennis player scores are shown in Table 2. In order to evaluate personal technical ability, we can analyze past or real-time player scores; for fatigue, we can measure it through indicators such as the player's real-time running distance; for real-time mentality of the game, we can measure it through whether there are service errors and whether the player is in the game. Evaluated by indicators such as winning without overtime and whether points were scored when the opponent served. Based on these factors, we have constructed an evaluation system to comprehensively and accurately evaluate players' real-time performance in the game. This system helps to deeply understand and predict players' tactical scores, providing strong support for game analysis and prediction.

Table 2: Index system of factors affecting tennis player scores

Indicator type	Corresponding indicators
Fatigue level	d1,d2,d3,p1
mental state	f1,f2,sv,v
personal technical ability	gp,sg,mp,s1,s2,s3 ,p2
real time status	s

We select the 16 indicators defined above as influencing factors. Considering that each point score is a two-class situation, we establish a two-class Logistic regression model (Li et al., 2023) for prediction. Logistic regression has less impact on abnormal data, has certain anti-noise ability and robustness, and is highly robust; logistic regression is mainly used in linearly separable problems, but it also has certain generalization capabilities and can handle non-linearity data.

$$\begin{aligned} \ln \frac{p}{1-p} &= \beta_0 + \sum_{k=1}^{16} \beta_k x_k \\ p &= \frac{\exp\left(\beta_0 + \sum_{k=1}^{16} \beta_k x_k\right)}{1 + \exp\left(\beta_0 + \sum_{k=1}^{16} \beta_k x_k\right)} \end{aligned} \tag{3}$$

Use SPSS software to solve, and the results are as follows in Table 3(Constant is included in the model and the cut value is .500.):

From the SPSS statistical results, we can see that the accuracy of the logistic regression model is 65.5%. For those samples where the players did not score, the model's classification accuracy was only 34.4%. However, for those samples where the players actually scored, the model's classification accuracy was as high as 85.2%, which was much higher than the samples where no scores were scored. This shows that the current model prefers to classify the sample as the true score, that is, the case where index is yes.

Table 3: Forecast precision

			Predicted index		
			No	Yes	Percentage Correct
Step 1	Observed index	No	271	517	34.4
		Yes	184	1058	85.2
		Overall Percentage			65.5

However, the current logistic regression results are a forward-looking analysis, and the main purpose is to check whether the constructed indicators have a significant impact on the actual scores of the players. The logistic regression test results are shown in Table 4.

Analyzing the data in Table 4, we can see that half of the significant p-values of all independent variables are less than 0.05. When the significance p value is less than 0.05, we have reason to believe that the independent variable has a significant impact on the dependent variable. Therefore, in this indicator system, the independent variables gp, s1, s2, f1, f2, p1 and d3 can significantly affect the players' real-time scoring. Whether it is a player's fatigue, mental state or personal technical ability, they all significantly affect the player's real-time scoring.

5.2. Machine learning model inspection with point granularity

In the machine learning model construction part, we chose the currently excellent ensemble tree models LGBM (Nikhil and Nagalakshmi, 2023) and XGBOOST, and also selected classic machine learning algorithms, including support vector machines (Zhou et al., 2011), perceptron networks (Narayan, 2021) and logic Regression, as a comparison algorithm. We used the Accuracy, Percision, Recall and F1 scores based on the confusion matrix, as well as the AUC and ROC curves for the five-fold cross-validation method (Rojatkar et al., 2013).

The obtained confusion matrix results are shown in Figure 2, while the training set and test set results based on AUC and ROC curves are shown below respectively.

As can be seen from the above data, LGBM performs best in various indicators. Its Accuracy, Percision, Recall and F1 scores are 0.69, 0.69, 0.7 and 0.69 respectively, and the AUC is 0.77. Followed by the perceptron neural network, its Accuracy, Percision, Recall and F1 scores are 0.69, 0.66, 0.71 and 0.68 respectively, and the AUC is 0.76. The difference between various indicators is very small, which shows that the model does not show obvious discrimination preference for positive and negative samples. Therefore, it can be concluded that LGBM has the best effect. The training set ROC curve and test set ROC curve are shown in Figure 3.

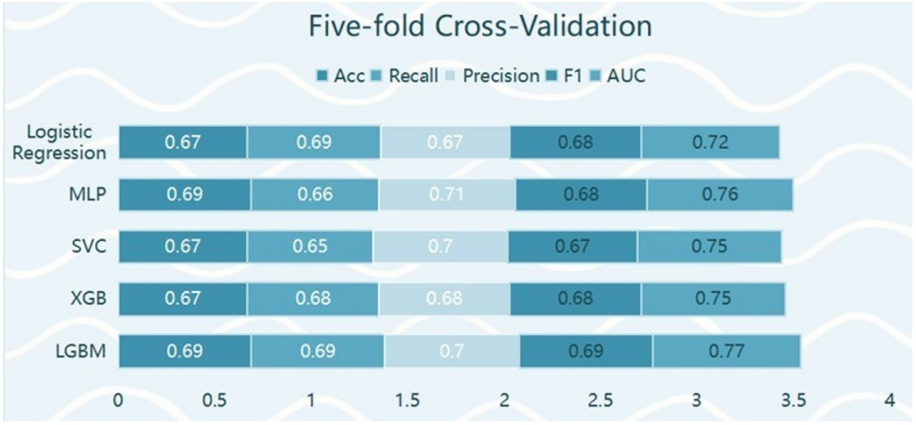

Figure 2: Five-fold cross validation algorithm results.

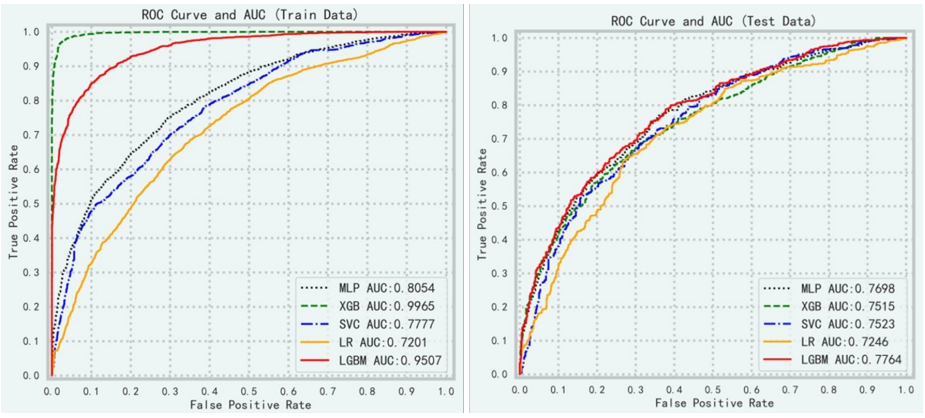

Figure 3: Training set ROC curve and test set ROC curve.

The ROC curve reflects the changes in precision and recall indicators under different thresholds. Through the ROC curve, we can also verify that LGBM is the best algorithm. The schematic diagram of LGBM principle is shown in Figure 4.

Based on the previous results, we decided to retrain using the best performing LGBM model. In the 2023 Wimbledon men's singles final, a 20-year-old Spanish star Carlos Al-caraz defeated the 36-year-old Novak Djokovic. We chose this classic match for real-time performance visualization. The final result is shown in Figure 5.

In fact, there is a limit to the accuracy of predicting score points, which is consistent with our intuition. Because the factors that affect whether a player's scores are too complex, the noise in a player's score is very loud when these complex factors are not taken into account. Therefore, for the model, it can only accurately predict about 70% of the players' true scores.

Through analysis, we found that the server and scorer in this classic matchup were consistent with the score of the players. When Carlos Alcaraz wins, the momentum is high; when he loses, the momentum is low. This shows that the model proposed in this paper is still valid. The indicator system for question is shown in Table 4.

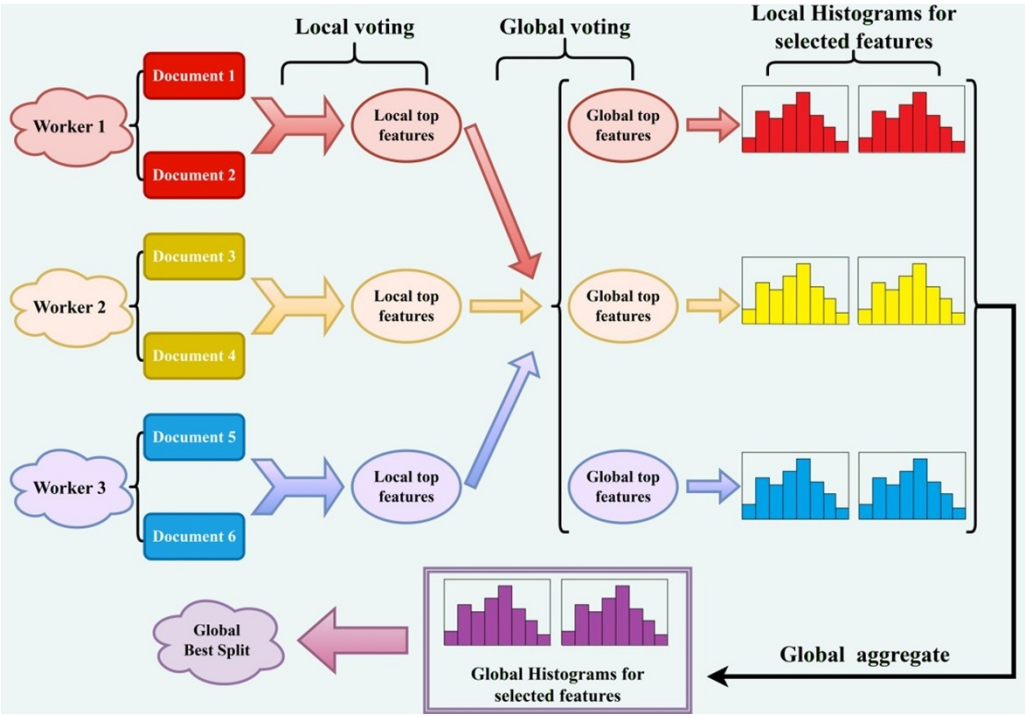

Figure 4: Schematic diagram of LGBM principle.

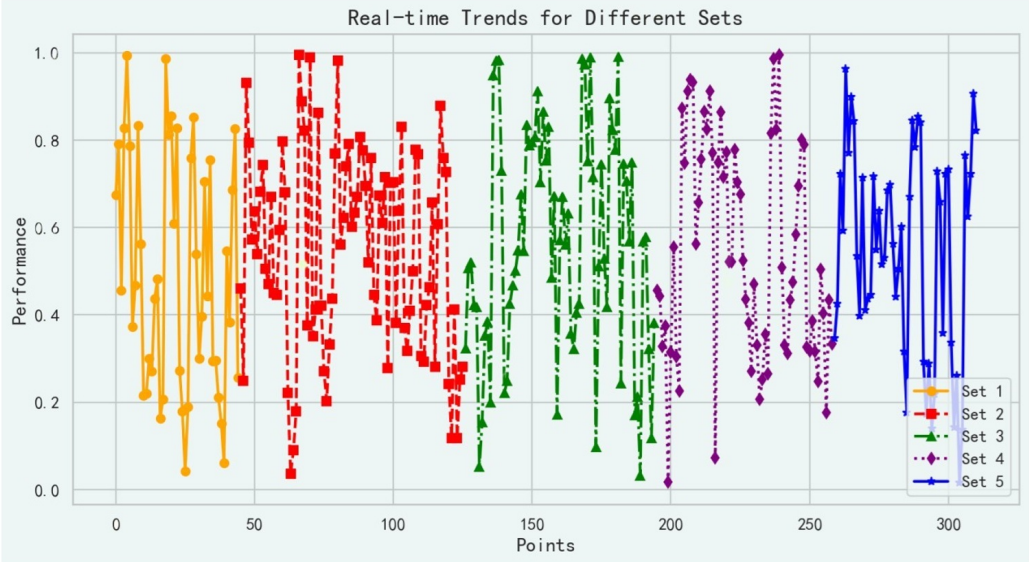

Figure 5: Final result of real-time trends for different sets.

References

Keren He and Cuiwei He. Housing price analysis using linear regression and logistic regression: a comprehensive explanation using melbourne real estate data. In *2021 IEEE International Con-*

Table 4: Variables in the Equation

		B	SE	Wald	df	Sig.	Exp(B)
Step 1a	d1	.120	.288	.174	1	.676	1.128
	d2	-.164	.482	.116	1	.734	.849
	d3	-1.445	.657	4.842	1	.028	.236
	p1	.914	.124	54.371	1	.000	2.493
	f1	.472	.153	9.491	1	.002	1.603
	f2	-.535	.109	24.287	1	.000	.586
	sv	-.876	1.693	.268	1	.605	.416
	v	.511	.626	.668	1	.414	1.667
	gp	-1.141	.363	9.881	1	.002	.319
	sg	.016	.176	.009	1	.926	1.107
	mp	.032	.203	.024	1	.876	1.032
	s1	.689	.127	29.400	1	.000	1.991
	s2	.456	.128	12.697	1	.000	1.577
	s3	.170	.166	1.048	1	.036	1.186
	p2	.073	.140	.273	1	.601	1.076
	s	.689	1.286	.287	1	.592	1.992
	const	-.072	.346	.043	1	.835	.931

ference on Computing (ICOCO), pages 241–246. IEEE, 2021. doi: 10.1109/ICOCO53166.2021.9673533.

M. Kovacs and T. Ellenbecker. An 8-stage model for evaluating the tennis serve: implications for performance enhancement and injury prevention. *Sports Health*, 3(6):504–513, 2011. doi: 10.1177/1941738111414175.

Wenqi Li, Yang Zhao, Mengli Dai, and Jiazhe Li. Prediction and classification of ancient glass types based on logistic regression models. In *2023 IEEE 3rd International Conference on Electronic Communications, Internet of Things and Big Data (ICEIB)*, pages 331–336. IEEE, 2023. doi: 10.1109/ICEIB57887.2023.10170035.

Yogendra Narayan. Analysis of mlp and dslvq classifiers for eeg signals based movements identification. In *2021 2nd Global Conference for Advancement in Technology (GCAT)*, pages 1–6. IEEE, 2021. doi: 10.1109/GCAT52182.2021.9587868.

K. Venkata Sai Nikhil and T.J. Nagalakshmi. Accuracy improvement in the attendance marking system using human face recognition with inception algorithm and comparing with light gradient boosting machine (lgbm) algorithm. In *2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)*, pages 1–6. IEEE, 2023. doi: 10.1109/ICONSTEM56934.2023.10142423.

Peter O' Donoghue. *Statistics for sport and exercise studies: An introduction*. Routledge, 2013. doi: 10.4324/9780203133507.

Dinesh V. Rojatkar, Krushna D. Chinchkhede, and G.G. Sarate. Handwritten devnagari consonants recognition using mlpnn with five fold cross validation. In *2013 International Conference on Circuits, Power and Computing Technologies (ICCPCT)*, pages 1222–1226. IEEE, 2013. doi: 10.1109/ICCPCT.2013.6528992.

Hao Zhou, Shaohong Li, and Jinping Sun. Unit model of binary svm with ds output and its application in multi-class svm. In *2011 Fourth International Symposium on Computational Intelligence and Design*, volume 1, pages 101–104. IEEE, 2011. doi: 10.1109/ISCID.2011.34.