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

Dynamic research of tennis match based on machine learning

Xiangyu Wu [1, *, #], Haoxuan Xu [2, #], Xuan Zeng [3, #]

[1] College of Information and Electrical Engineering, China Agricultural University, Beijing, China, 100083

[2] College of Electrical and Power Engineering, Taiyuan University of Technology, Taiyuan, China, 030002

[3] College of Life and Environmental Sciences, Minzu University of China, Beijing, China, 100081

* Corresponding author: jhb199702@gmail.com

[#]These authors contributed equally.

Abstract. The study employed the Adaboost algorithm to identify optimal indicators of momentum, demonstrating that momentum could significantly predict the likelihood of a player's subsequent wins or losses. This study offers a novel exploration into the dynamics of tennis matches, focusing on the concept of "momentum" and its impact on match outcomes. Through an analysis of the 2023 Wimbledon Men's Singles final, the research demonstrates how momentum significantly influences the likelihood of a player's success, challenging traditional perceptions of match progression as merely a sequence of random events. By employing advanced statistical methods, including the Adaboost algorithm for identifying key indicators of momentum and the Wald-Wolfowitz Runs Test to examine the randomness of winning streaks, the findings reveal a substantial effect of momentum on the game's flow. The analysis further utilizes the CatBoost algorithm to pinpoint factors critical in shifting momentum between players, such as set scores, serving dynamics, and return depths. These indicators provide concrete evidence against the randomness of match dynamics, suggesting that strategic adjustments based on these factors can potentially alter the course of a match. The implications of this research extend beyond academic interest, offering practical insights for players and coaches in strategizing to leverage momentum for competitive advantage. By debunking the myth of randomness in tennis match outcomes, this study underscores the importance of psychological and strategic elements in sports performance, providing a foundation for further research and practical applications in the field of sports science and coaching.

Keywords: AdaBoost; Catboost; Wald-Wolfowitz Runs Test; Tennis; Momentum.

1. Introduction

1.1. Background

The 2023 Wimbledon Gentlemen's final showcased the dynamic nature of tennis, where the rise of young talent like Carlos Alcaraz and the resilience of seasoned players like Novak Djokovic exemplify the sport's capacity for dramatic momentum shifts. These shifts not only captivate audiences but also underscore the complexity of competitive strategy in tennis. While the concept of momentum is often invoked in sports commentary, its tangible effects on match outcomes and the strategic elements that contribute to its ebb and flow are not yet fully understood.

In the realm of sports science, research has begun to delve into the nuances of player performance and the factors that influence the dynamics of a match. The study by Giles, Kovalchik, and Reid (2019) in the Journal of Sports Sciences represents a significant step in this direction. Their research utilized machine learning to analyze player tracking data from professional tennis matches, aiming to automatically detect and classify changes of direction (COD), which are critical elements in the sport's tactical and physical demands.

By examining the movement patterns of professional tennis players at the Australian Open Grand Slam, the study sought to develop an automated method for identifying and classifying COD

movements. The findings from this research offer valuable insights into the intensity and frequency of these movements, which are instrumental in understanding the momentum shifts that can occur within a match. The use of machine learning classifiers, such as the random forest algorithm, to accurately predict medium and high-intensity COD movements with an F1-score of 0.729, demonstrates the potential of advanced analytics in sports performance [1].

1.2. Research Purposes

This study is designed to achieve three primary goals that will deepen our comprehension of momentum in tennis. Firstly, it aims to construct a quantitative momentum model that delineates the fluctuating dynamics of a match. By pinpointing patterns in player performance, this model will translate the abstract concept of momentum into a measurable and analyzable framework.

Secondly, the research will empirically examine the influence of momentum on match outcomes. Through rigorous statistical analysis, it will determine whether perceived shifts in momentum have a substantial bearing on the results or if their impact is exaggerated. This will clarify the actual versus perceived effects of momentum on competitive advantage.

Lastly, the study will identify key indicators that affect the ebb and flow of a game, such as player endurance, mental fortitude, and tactical changes. By uncovering these factors, the research will provide actionable insights for strategic planning, enabling players and coaches to better manage and exploit momentum during matches.

By tackling these objectives, the research will deliver a nuanced understanding of momentum in tennis, offering strategic insights and actionable advice for enhancing competitive performance against a range of opponents.

2. Reasearch Methodology

2.1. Data Acquisition and Preprocessing

The data originates from the open-source website "2024 MCM/ICM Problems (immchallenge.org)".

The dataset includes several key features shown in Table 1:

Table 1. Key features

Key features	Description
match_id	The identifier for the match, such as "2023-wimbledon-1701" (where "7" represents the round of the match, and "01" represents the match number within that round).
player1	The first player's name and surname, e.g., Carlos Alcaraz
player2	The second player's name and surname, e.g., Novak Djokovic.
elapsed_time	The elapsed time from the start of the first point to the start of the current point (hours: minutes: seconds), e.g., 0:10:27.
set_no	The set number in the match, which can be 1, 2, 3, 4, or 5.
game_no	The game number within a set, which can be 1, 2, ...7, etc.
point_no	The point number within a game, which can be 1, 2, 3, etc.
p1_sets	The number of sets won by the first player, which can be 0, 1, or 2.
p2_sets	The number of sets won by the second player, which can be 0, 1, or 2.
p1_games	The number of games won by the first player in the current set, which can be 0, 1, ...6.
p2_games	The number of games won by the second player in the current set, which can be 0, 1, ...6.
p1_score	The current game score of the first player, which can be 0 (love), 15, 30, 40, AD (advantage).
p2_score	The current game score of the second player, which can be 0 (love), 15, 30, 40, AD (advantage).
server	Indicates who is serving; 1 for the first player, and 2 for the second player.
p2_double_fault	Indicates whether the second player lost a point due to a double fault; 1 for yes, 0 for no.
p1_unforced_error	Indicates whether the first player made an unforced error; 1 for yes, 0 for no.
p2_unforced_error	Indicates whether the second player made an unforced error; 1 for yes, 0 for no.
p1_net_pt	Indicates whether the first player approached the net; 1 for yes, 0 for no.
p2_net_pt	Indicates whether the second player approached the net; 1 for yes, 0 for no.
p1_net_pt_won	Indicates whether the first player won the point at the net; 1 for yes, 0 for no.

The dataset is partly shown in Table 2.

Table 2. Part section of Wimbledon_featured_matches.csv

Match_id	Player 1	Player 2	Elapsed_time	Set_no	Game_no	Point_no	P1_sets	P2_sets	P1_games
2023-wimbledon-1301	Carlos Alcaraz	Nicolas Jarry	0:00:00	1	1	1	0	0	0
2023-wimbledon-1301	Carlos Alcaraz	Nicolas Jarry	0:00:38	1	1	2	0	0	0
2023-wimbledon-1301	Carlos Alcaraz	Nicolas Jarry	0:01:01	1	1	3	0	0	0
2023-wimbledon-1301	Carlos Alcaraz	Nicolas Jarry	0:01:31	1	1	4	0	0	0
2023-wimbledon-1301	Carlos Alcaraz	Nicolas Jarry	0:02:21	1	1	5	0	0	0
2023-wimbledon-1301	Carlos Alcaraz	Nicolas Jarry	0:02:50	1	1	6	0	0	0
2023-wimbledon-1301	Carlos Alcaraz	Nicolas Jarry	0:03:33	1	1	7	0	0	0

Data Preprocessing Steps:

Encoding: The dataset conAdaboostategorical variables like "D", "F", "NCTL", etc., which were encoded to facilitate modeling.

Outlier Handling: Continuous variables such as p_distance_run, p2_distance_run, and speed_mph was subjected to outlier processing using SPSSPro to mitigate the influence of extreme values.

Feature Selection: Relevant features were selected for analysis, focusing on capturing score dynamics and identifying optimal momentum indicators using the Adaboost machine learning method.

Predictive Modeling: The Adaboost algorithm was applied to the preprocessed dataset to train a predictive model capable of forecasting momentum shifts and predicting match outcomes based on player performance trajectories.

2.2. Methodology

Adaboost: short for Adaptive Boosting, is a powerful ensemble learning technique used in machine learning for classification and regression tasks. It was proposed by Yoav Freund and Robert Schapire in 1996. AdaBoost is particularly renowned for its ability to improve the performance of weak learners, such as decision trees, by combining them into a strong learner. At its core, AdaBoost operates by iteratively training a sequence of weak learners, each focusing on the instances that the previous models misclassified. It assigns weights to the training instances and adjusts them at each iteration to emphasize the difficult instances. The final model is a weighted sum of the weak learners, where the weights are determined by their individual performance during training [2, 3].

Let's define some key components of AdaBoost:

Represents the weight assigned to the i^{th}training instance at iteration$t$.

$D_t(i)$is the weak learner or base classifier at iteration t.

α_t denotes the contribution weight of $h_t(x)$ to the final classifier

The AdaBoost algorithm minimizes the exponential loss function:

$$L(f) = \sum_{i=1}^{N} e^{-y_i f(x_i)} \quad (1)$$

Where:

N is the total number of training instances.

x_i is the i^{th} training instance.

y_i is the corresponding label (-1 or 1).

$f(x_i)$ is the output of the final classifier for the instance x_i.

The AdaBoost algorithm proceeds as follows:

(1) Initialize the weights $D_1(i) = \frac{1}{N}$, where N is the number of training instances.

(2) For $t = 1,2,\ldots,T$, do:

Train a weak learner $h_t(x)$ using the weighted training data.

Compute the error ϵ_t of $h_t(x)$

Compute $\alpha_t = \frac{1}{2} ln(\frac{1-\epsilon_t}{\epsilon_t})$, where ϵ_t is the error rate.

Update the weights:

$$D_{t+1}(i) = \frac{D_t(i) \cdot exp(-\alpha_t y_i h_t(x_i))}{Z_t} \quad (2)$$

Where Z_t is a normalization factor.

(3) Construct the final classifier:

$$(F(x) = \sum_{t=1}^{T} \alpha_t h_t(x) \quad (3)$$

Wald–Wolfowitz Runs Test: The Wald–Wolfowitz runs test (or simply runs test), named after statisticians Abraham Wald and Jacob Wolfowitz, is a non-parametric statistical test that checks a randomness hypothesis for a two-valued data sequence. More precisely, it can be used to test the hypothesis that the elements of the sequence are mutually independent [4].

A run of a sequence is a maximal non-empty segment of the sequence consisting of adjacent equal elements.

Under the null hypothesis, the number of runs in a sequence of N elements [note 1] is a random variable whose conditional distribution given the observation of N_+ positive values [note 2] and N_- negative values $N = N_+ + N_-$ is approximately normal, with:

$$mean: \mu = \frac{2N_+N_-}{N} + 1, variance: \sigma^2 = \frac{2N_+N_-(2N_+N_- - N)}{N^2(N-1)} = \frac{(\mu-1)(\mu-2)}{N-1} \quad (4)$$

Equivalently, the number of runs is

$$R = \frac{1}{2}(N_+ + N_- + 1 - \sum_{i=1}^{N-1} x_i x_{i+1}) \quad (5)$$

These parameters do not assume that the positive and negative elements have equal probabilities of occurring, but only assume that the elements are independent and identically distributed. If the number of runs is significantly higher or lower than expected, the hypothesis of statistical independence of the elements may be rejected.

CatBoost: an efficient machine learning algorithm for predicting categorical features, is a gradient boosting implementation that utilizes binary decision trees as base predictors. Proposed as an ensemble learning technique, CatBoost demonstrates proficiency in classification tasks and is particularly effective in handling categorical data.

The dataset is denoted as $D = \{(X_j, y_j)\}_{j=1}^{m}$, where X_j represents a vector of n features and $y_j \in \mathbb{R}$ is the response feature. y_j can be binary (e.g., yes or no) or encoded as a numerical feature (0 or 1). The samples (X_i, y_i) are independently and identically distributed according to some unknown distribution $p(\cdot)$. The learning task aims to train a function $H: \mathbb{R}^n \to \mathbb{R}$ that minimizes the expected loss given in Equation [5].

$$\mathcal{L}(H) := \mathbb{E}L(y, H(X)) \quad (6)$$

CatBoost operates similarly to other gradient boosting algorithms but excels in handling categorical variables. It employs binary decision trees as base learners, adapting them to enhance model performance iteratively.

Let's define some key components of CatBoost:

X represents the feature vector.

y is the response feature.

$h_t(x)$ is the response feature.

The CatBoost algorithm minimizes a suitable loss function, adjusting the weights and combining the weak learners to form a robust final model.

Initialize the model and set parameters.

For each iteration t, do:

Train a weak learner $h_t(x)$using the weighted training data.

Update model parameters based on the loss function.

Construct the final classifier:

$$F(x) = \sum_{t=1}^{T} h_t(x) \quad (7)$$

2.3. Model Evaluation Metrics

The evaluation results of the model are presented in Table 3, 4, illustrating the performance metrics for the cross-validation set, training set, and testing set. These metrics serve as quantitative indicators to assess the predictive capabilities of the AdaBoost and CatBoost model. The accuracy rates, recall rates, precision rates, and F1 scores are computed for both the training and testing sets. As observed from the data in the table, the AdaBoost and CatBoost model demonstrates high accuracy on the training set, achieving an accuracy rate, recall rate, precision rate, and F1 score.

Table 3. AdaBoost Model Evaluation Results

	Accuracy Rate	Recall Rate	Precision Rate	F1
Training Set	0.994	0.994	0.994	0.994
Testing Set	0.757	0.757	0.758	0.757

Table 4. CatBoost Model Evaluation Results

	Accuracy Rate	Recall Rate	Precision Rate	F1
Training Set	1	1	1	1
Testing Set	0.703	0.703	0.722	0.692

3. Model Construction and Analysis

3.1. The establishment of Momentum model

After running two models, XGBoost and AdaBoost, it was observed that XGBoost exhibited overfitting. This issue may be attributed to the model's sensitivity to the imbalanced nature of the dataset. Nur Huda Nabihan Binti Md Shahri and Sharmeen Binti Syazwan Lai supports this observation, as it states that, for simulated imbalanced datasets, XGBoost is identified as the best method for a 10% minority dataset. In our case, given that we encountered overfitting with XGBoost, it aligns with the literature's indication that the performance of XGBoost can vary based on the degree of class imbalance [6].

To address this challenge and improve generalization performance on imbalanced datasets, we have decided to switch to AdaBoost. The literature suggests that AdaBoost is particularly effective for datasets with a 25% minority class, indicating its robustness in handling moderately imbalanced scenarios. By leveraging the strengths of AdaBoost, which adapts by assigning more weight to misclassified instances, we aim to achieve a more balanced and accurate classification, mitigating the overfitting issues observed with XGBoost.

Additionally, these table 5 provide detailed information on the AdaBoost model parameters and feature importance, contributing to a comprehensive understanding of the model's performance in capturing changes in momentum during matches.

Table 5. AdaBoost Feature Importance

Feature Name	Weight of Importance
p2_distance	17%
p1_score	13%
p2_score	13%
p1_distance	13%
speed_mph	13%
rally_count	12%
game_victor	4%
P1_winner	2%
P1_net_pt	2%
Serve_width	2%
Return_depth	2%

3.2. Momentum Analysis and Run Test Evaluation

In accordance with the Momentum model [7, 8], we computed the momentum for both Player 1 and Player 2. The momentum scores were binary, represented by 1 and 0, indicating the high or low

momentum of each player. We conducted a run test to examine whether winning streaks were occurring randomly. The results of the run test are presented in Table 6.

Table 6. Wald–Wolfowitz Run Test

Name	Sample size	Z	P
Turning-point	299	2.393	0.017

3.3. The establishment of Swing model

Definition of Turning Points: Crucial moments where the outcome of the game may undergo a significant change. In this paper, a turning point is defined as a transition from a consecutive losing streak to a consecutive winning streak or vice versa. To identify indicators of momentum shifts from one player to another, this study, based on the data processing regarding turning points in the second question, utilizes a machine learning classification model. Due to the significant improvements and advantages of the CatBoost algorithm in handling categorical features, boosting techniques, and decision tree growth scores, this paper adopts the CatBoost algorithm to classify indicators related to momentum [9,10].

Moreover, the feature importance for the CatBoost model is summarized in Table 7:

Table 7. CatBoost Feature Importance

Feature Name	Feature Name
P2_score	14.8%
serve_width	8.9%
P1_score	8.1%
P2_distance	7.9%
set_no	6.7%
P1_distance	5.7%
Last_mom_diff	4%
P1_games	4.1%
return_depth	4.1%
game_no	3.7%
P2_games	3.2%
Rally_counts	2.9%
P2_points_won	2.8%
P1_points_won	1.8%
Point_no	1.5%
Game_victors	1.5%
P2_sets	1.3%

4. Conclusion

Our analysis of the 2023 Wimbledon final, coupled with machine learning techniques like Adaboost and Catboost, revealed the pivotal role of momentum in tennis matches. We disproved the idea of random swings through rigorous statistical tests and identified key indicators for predicting momentum shifts, aiding coaches in strategic preparations. Testing the model on various matches demonstrated its robustness, though acknowledging limitations in generalization. Overall, our research confirms the significance of momentum in tennis, providing coaches with a data-driven tool to enhance player adaptability and performance.

References

[1] Giles B, Kovalchik S, Reid M. A machine learning approach for automatic detection and classification of changes of direction from player tracking data in professional tennis [J]. Journal of sports sciences, 2020, 38 (1): 106 - 113.

[2] Asra T, Setiadi A, Safudin M, et al. Implementation of AdaBoost Algorithm in Prediction of Chronic Kidney Disease[C]//2021 7th International Conference on Engineering, Applied Sciences and Technology (ICEAST). IEEE, 2021: 264 - 268.

[3] Ke J, Zhengxuan Z, Zhe Y, et al. Intelligent islanding detection method for photovoltaic power system based on Adaboost algorithm[J]. IET Generation, Transmission & Distribution, 2020, 14 (18): 3630 - 3640.

[4] De La Rubia J M. Calculation of Two-Tailed Exact Probability in the Wald-Wolfowitz One-Sample Runs Test [J]. Journal of Data Analysis and Information Processing, 2024, 12 (1): 89 - 114.

[5] Abdullahi A. Ibrahim, Raheem L. Ridwan, Muhammed M. Muhammed, Rabiat O. Abdulaziz and Ganiyu A. Saheed, "Comparison of the CatBoost Classifier with other Machine Learning Methods" International Journal of Advanced Computer Science and Applications (IJACSA), 11 (11), 2020.

[6] Shahri N, Lai S B S, Mohamad M B, et al. Comparing the performance of AdaBoost, XGBoost, and logistic regression for imbalanced data [J]. Math Stat, 2021, 9: 379 - 85.

[7] Goyal A, Simonoff J S. Hot Racquet or Not an Exploration of Momentum in Grand Slam Tennis Matches [J]. arXiv preprint arXiv:2009.05830, 2020.

[8] Rusdiana A, Syahid A M, Kurniawan T. Momentum transfer analysis of upper body during backhand drive stroke in tennis [C]//AIP Conference Proceedings. AIP Publishing, 2023, 2646 (1).

[9] Depken C A, Gandar J M, Shapiro D A. Set-level strategic and psychological momentum in best-of-three-set professional tennis matches[J]. Journal of Sports Economics, 2022, 23 (5): 598 - 623.

[10] Ma K. A real time artificial intelligent system for tennis swing classification[C]//2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI). IEEE, 2021: 000021 - 000026.