Quantifying Momentum in Tennis: A Case Study of the 2023 Wimbledon Gentlemen’s Final

Xin Lyu
Faculty of Science and Technology, Beijing Normal University - Hong Kong Baptist University United International College, Zhuhai, Guangdong Province, 519000, China
r130026103@mail.uic.edu.cn

Abstract. In sports competitions, especially in famous sports like tennis, momentum has a very significant impact on the outcome of the competition. In the era of big data, research can predict the momentum changes in the game based on historical data, thereby providing players and coaches with reasonable strategies. This analysis can increase the player's winning rate. Explore the presence and changes of momentum in the game through a case study of the 2023 Wimbledon men's singles final. The research selected key moments in the game, including serve games, break games, and errors, to capture momentum through a series of quantifiable indicators, and applied random forest, support vector machine, and multi-layer perception models. Their accuracy rates are 0.6691, 0.6643, and 0.6479 respectively. The research aimed to predict scoring results and evaluate player performance. The research provides coaches and players with a new tool to quantify momentum in tennis matches, helping to better understand match flow, develop tactics, and improve match performance.

Keywords: Wimbledon Gentlemen’s Final, Support Vector Machine, Multi-layer Perception Model.

1. Introduction

As one of the most popular ball sports, tennis has a unique scoring system. In a tennis match, scores are calculated starting from 0 and passing through stages of 15, 30, and 40 points. To win the game, players need to ensure they are at least two points ahead of their opponents. During the game, players take turns serving and receiving the ball. In professional tennis, the server usually has a greater advantage. At each scoring point, players have two chances to serve the ball into the opponent's service area. If a player fails to serve twice in a row, it is deemed a "double fault" and the opponent receives that point. When the score of a certain set reaches a 6-6 tie, a tie-break will be held. At this time, one side needs to lead by at least two points to win the game. Usually, the player who wins by 7 points first wins. In addition, the game includes a break after the first inning and every two innings thereafter. Starting from the third game, a 90-second rest period is allowed for each defense change. In the tiebreaker, players switch sides every 6 points. After each set, players must rest for at least 2 minutes. Medical timeouts and a bathroom break are allowed [1]. The dictionary definition of momentum is "the power or force gained by motion or a series of events [2]." In tennis, the concept of momentum is a key factor in the outcome of the game, intertwining psychological and physiological factors. Philippe Meier decomposed motivation into strategic and psychological components and found that psychological motivation significantly improved player performance after winning key points and that this dynamic was moderated by game interruptions [3]. This nuanced understanding is consistent with the investigation by Helmut Dietl, which highlighted the dependence of momentum on race control [4]. They make it clear that while players demonstrate momentum by controlling the game, losing control can trigger counter-momentum, significantly reducing the likelihood of winning subsequent rounds. Silva integrated their findings on the influence of psychodynamics on gender, emphasizing its pervasive impact on tennis player performance and its potential impact on match outcomes [5]. These studies provide a solid basis for illustrating the critical role of momentum in tennis. They show the need to quantify momentum to understand its impact on game dynamics and player performance.

The research uses comprehensive data analysis and machine learning techniques to quantify motivation in tennis. Using a dataset from Wimbledon, the research preprocessed the data, handling missing values and normalizing scores for robust analysis. Then apply machine learning models, including random forest (RF), support vector machine (SVM), and multi-layer perception (MLP), to predict scoring outcomes, integrating features such as score variance and player performance metrics. This approach not only quantifies momentum but also provides insights into its impact on game dynamics. This study provides relevant researchers and scholars in the field with a novel method of quantifying motivational factors in tennis. By integrating machine learning technology with sports data analysis, and extending existing sports performance quantification methodologies, it provides a theoretical foundation and practical framework for further exploring similar dynamic factors in sports competitions. Through in-depth analysis of motivation factors, coaches can more specifically emphasize the importance of motivation in training and competition strategies, and players can more effectively adjust their playing methods and mental preparation to adapt to the flow of the game and their opponents. strategy.

2. Methodology

2.1. Data Collection and Description

This data set comes from the Mathematical Contest in Modeling (MCM). These competitions challenge teams to analyze, model, and solve open-ended problems, developing skills in research, analysis, and applied intelligence. It provides students with a platform to solve real-world problems, improve their teamwork, technical writing, and communication skills, and have the opportunity to earn international recognition and awards.

The data contains detailed 2023 Wimbledon tennis event data. The data records every important moment of the game, covering player information, game time, various stages of the game (including many sets, games, and scores), and player performance statistics. Each row represents a specific point in the match and contains the unique identifier of the match, the names of the two players, the time the match was played, the current set number, the number of games, the score, and the number of sets and games won by both players. In addition, detailed data such as the player's running distance, serving speed, serving direction and depth, and return depth are also recorded. The entire data has 7284 rows and 47 columns, providing a wealth of information for in-depth analysis of the game.

2.2. Data Pre-processing

In the preliminary stage of this study, the research performed a series of preprocessing steps on the "Wimbledon_featured_matches.csv" data set to ensure data integrity and facilitate subsequent analysis. The data preprocessing process mainly includes the following steps:

Normalized Score Representation: To maintain numerical consistency across the dataset, especially when dealing with score data, normalizing the "AD" symbol indicates an advantage to the number "55". This step helps simplify the numerical processing of the score, making it easier to perform quantitative analysis. The research uses the normalization method to ensure that the range of physical data is between [0, 1]. Standardization is the process of scaling data so that it falls into a specific range to eliminate dimensional differences between different variables. Using the following function for normalization

$$X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1}$$

Handle missing values: Identify the possible impact of missing values in the data set on the validity of the study, and use a median filling strategy to deal with missing values. This helps mitigate the impact of outliers and ensures the robustness of the dataset.

Extract and transform identifiers: The 'match_id' column in the dataset contains composite identifiers, and the numeric fields in it need to be extracted and transformed for analysis. By using regular expression processing, the research converted these numeric identifiers into integer format and created a new column 'match_id_numeric' as a key reference for subsequent analysis.

Data type conversion: To maintain consistency in quantitative analysis, the research converted the data types of the 'p1_score', 'p2_score', and 'match_id_numeric' columns to integer types. This step is key to ensuring data type consistency and numerical analysis accuracy.

Through these careful pre-processing steps, the research transformed the data set into a structured and reliable basis, laying a solid foundation for in-depth research on the quantification of momentum in tennis.

2.3. Feature Extraction

In the study, the research adopted a comprehensive approach to extract and construct player-specific performance metrics from the dataset "update.csv", which has undergone preliminary preprocessing. This extraction aims to create two distinct data frames, each encapsulating individual player performance in a tennis match, allowing for a granular analysis of each player's game dynamics.

For player 1, the research initialized a series of lists to obtain various performance metrics. 'match_id_numeric' provides a unique identifier for each match, ensuring the traceability of data points. The research calculates score differences between players, capturing the real-time competitive environment. Record player 1's score and serving status. To facilitate analysis, the serving status is converted into a binary indicator. Additionally, the research identified examples of aces and double faults, key events that can significantly impact a player's momentum and mental state. The distance Player 1 ran at each point was also recorded, providing insight into physical exertion and movement patterns. Finally, points victory status provides a direct measure of individual rally success.

At the same time, a similar extraction process is also reflected in Player 2, generating a second data frame with similar indicators, thus ensuring a symmetrical and comparative analysis framework. The research adjusted the score difference to reflect Player 2's perspective, and other parameters were adjusted accordingly.

2.4. Implementation of Momentum Quantification in Tennis

To rigorously analyze and quantify momentum in tennis, the study needed to standardize the various performance metrics that influence a player's momentum in the game. Including score advantage ('P1_scoreAD'/'P2_scoreAD'), current game score ('P1_score'/'P2_score'), serve status ('P1_serve'/'P2_serve'), score ('P1_ace'/'P2_ace'), double Fault ('P1_double_fault'/'P2_double_fault'), and the distance between players on both sides during the scoring process ('P1_distance_run'/'P2_distance_run'). The standardization process is key, converting each metric into a scale with a mean of zero and a standard deviation of 1. This uniform scaling is essential to mitigate the different sizes of the original data, thus facilitating balanced comparisons between different metrics and matching.

Based on these individual metrics, the research calculates a performance index for each player. For player 1, the "p1_performance" index is calculated as the sum of the normalized values in the relevant columns. This composite score encompasses multiple aspects of performance and serves as a proxy for a player's motivation at any given moment in a match. The “p1_performance” index is at the heart of the study of momentum. By aggregating standardized parameters, the index reflects the nuances of player performance, distilled into a single dynamic measure. Since motivation in sports is inherently a multi-factor structure, the "p1_performance" index captures the interplay between skill execution, physiological expenditure, and psychological factors, providing a quantitative perspective through which to assess changes in player momentum during the game

2.5. Visualization of Performance: A Case Study of Match ID 1301

To effectively illustrate the concept of momentum in tennis as quantified by the standardized performance index, the research presents a case study of the ID 1301 match. As shown in the Figure 1. The index has been reset to start from the first point of the match, providing an intuitive and sequential representation of the performance flow. The x-axis represents the chronological order of match points ("Match Instance"), while the y-axis represents the player's performance index. With Player 1's performance marked in blue and Player 2's performance in red, this line graph shows fluctuations in player performance levels throughout the game. Each point is represented by a marker, connected by lines to highlight the progression of the game. This visual analysis highlights the volatile nature of momentum. Spikes are moments when a player's cumulative performance in the metric is high, indicating a player is in a dominant position or experiencing a surge in momentum. Conversely, troughs represent a drop in performance, which may indicate a loss of momentum or a shift in game dynamics in favor of the opponent. Fig. 1 reveals not only the presence of momentum fluctuations but also their magnitude and frequency. For example, consecutive peaks for one player could mean a sustained period of dominance, while alternating peaks and troughs between players could mean a competitive match with frequent shifts in dynamics. Performance indices therefore provide a sophisticated means of quantifying and visualizing other qualitative concepts of momentum. By examining these metrics as shown by match ID 1301, the research can break down the match into performance strengths, neutrals, and weaknesses. This analysis not only enriches the understanding of match flow but also contributes to strategic insights, potentially guiding players and coaches to optimize performance and match preparation.

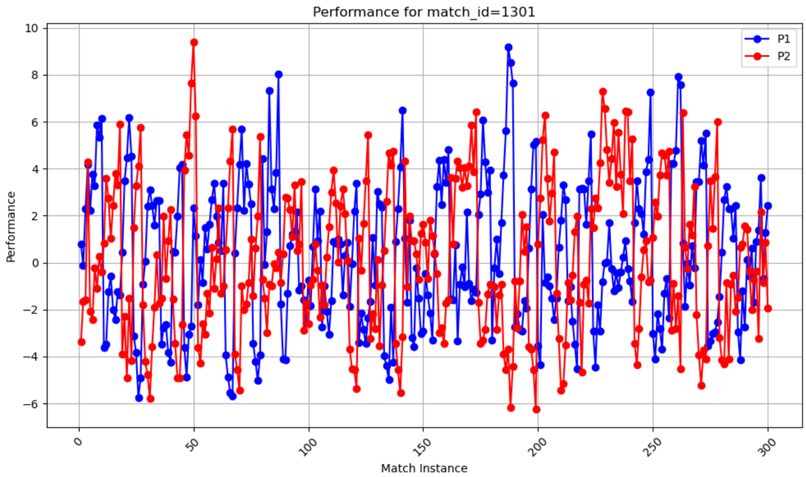

Fig. 1. The performance of the players in match number 1301 (Photo/Picture credit: Original).

Fig. 1 illustrates the performance fluctuations of two players over various instances of a tennis match identified by match_id=1301, showing their performance scores throughout the game.

2.6. Model Selection

Random Forest. RF is a method to improve the stability and accuracy of machine learning algorithms through ensemble decision-making. The RF algorithm enhances this approach by building a large number of decision trees at training time and outputting patterns of classes or average predictions for individual trees. Each tree in the forest is constructed by replacing drawn samples from the training set. Furthermore, when splitting nodes during tree construction, the best split is found from all input features or a random subset. This results in a forest of decorrelated trees [6].

Multilayer Perceptron. MLP is a basic model in the field of neural networks. This concept originated from the early perceptron theory proposed by Frank Rosenblatt [7]. The perceptron is a binary classifier that laid the foundation for later MLP. The way it works is that it takes a set of inputs, applies a weighted sum to them, and passes the result to a nonlinear activation function to produce an output for a classification task. MLP extends this idea by combining multiple layers of perceptrons, hence the name "multilayer," allowing the model to learn complex patterns through a technique called backpropagation. This learning process involves adjusting the weights of connections in the network in response to the error rate of the output compared to the desired outcome, thereby significantly enhancing the model's predictive power.

Support Vector Machines. SVM is a set of robust supervised learning methods used for classification, regression, and outlier detection. SVM is essential for high-dimensional data and is particularly good at performing complex classification tasks, distinguishing categories with clear separation ranges. By transforming data and using a technique called the kernel trick, SVM can perform linear classification in high-dimensional spaces, thereby enhancing their ability to handle nonlinear relationships [8]. The reliability and versatility of SVM make them the first choice for solving complex machine-learning challenges in a variety of fields.

3. Results

The research applies machine learning models to predict match outcomes using player performance data. And adopted SVM, MLP, and RF classifiers, and performed data preprocessing on SVM and MLP through standard scaling and principal component analysis (PCA). The accuracy of the model was evaluated, and SVM reached 66.9%, MLP reached 66.4%, and RF reached 64.8%, as shown in Table 1. These results highlight the potential of using advanced analytics to capture and quantify momentum in tennis matches, validating the effectiveness of these models as they play a key role in sports analytics.

Table 1. Accuracy of each model

Model	Accuracy
SVM	0.669183
MLP	0.664379
RF	0.647907

In an RF, the feature importance is evaluated during the construction of each tree, and a more robust feature importance evaluation is obtained by averaging the feature importance scores of all trees [6]. So the research used an RF classifier to evaluate the significance of various performance features in determining the point winner in a tennis match. The analysis shows that “P1_serve” is the most influential feature with an importance score of 0.528, indicating a strong correlation between a player’s serve and the momentum of the match. Followed closely by “P1_scoreAD” and “P1_ace” with scores of 0.182 and 0.121 respectively. In contrast, 'P1_double_fault' has the smallest impact, as shown in Fig. 2.

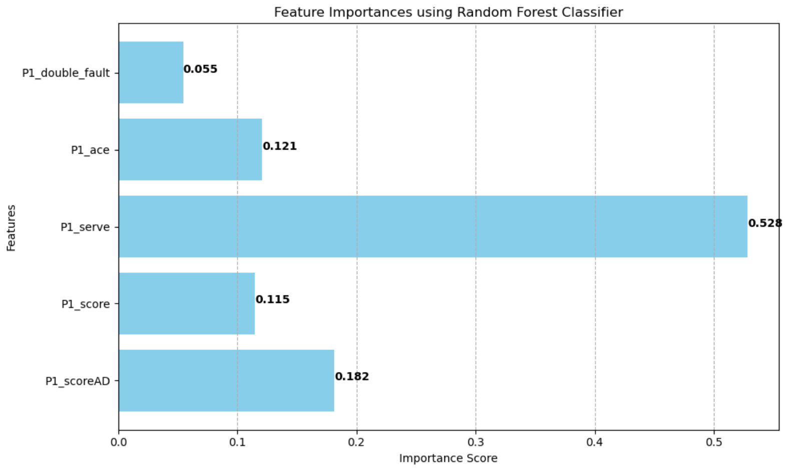

Fig. 2. The feature importance of each variable (Photo/Picture credit: Original).

Fig. 2 displays the feature importance as determined by an RF Classifier, highlighting 'P1_serve' as the most influential factor.

4. Discussion

This study explores the application of machine learning models to predict tennis match outcomes, highlighting a critical step forward in the field of sports analytics. Recent work by Anderson and Anderson highlights the transformative impact of artificial intelligence and machine learning on sports, particularly in data-driven athlete performance and strategic decision-making [9]. The study is consistent with these insights and demonstrates the effectiveness of machine learning in capturing the nuances of tennis movement.

The high accuracy of SVM, MLP, and RF classifiers in predicting match outcomes is like the findings of Zhao and van Buiten, who reported similar success in real-time strategy optimization in sports [10]. These technologies can fine-tune tactics in dynamic environments like tennis, where quick decision-making is crucial.

Furthermore, the critical role of the “P1_serve” function in determining race momentum confirms the recent computational analysis of Martin et al., emphasizing the strategic value of serve in tennis [11]. Together, these studies reinforce the concept that machine learning offers more than just predictive capabilities; it can also give players a deeper understanding of the strategic elements of a game. While the current research focuses on quantifying psychological dynamics, attention must also be paid to the broader implications of such analytical tools in training and development. Li et al. discussed the potential of data science in improving sports decision-making, suggesting that insights gleaned from such analysis can help tailor training programs and competition preparation to optimize athlete performance [12].

A model may be limited by its design and the quality of the training data. Since machine learning models are highly dependent on the quality and breadth of the data provided, future research should aim to collect more comprehensive data sets, possibly incorporating variables such as player fatigue and psychological state, by using a wider range of data sets and more complex models to overcome these limitations.

5. Conclusion

This study employs a machine learning model to quantify momentum in tennis matches and explore its impact on match outcomes. Through an in-depth analysis of Wimbledon 2023 data, this study demonstrates the significant role of momentum in matches and proposes new tools for predicting match dynamics. The SVM, MLP, and RF classifiers used in the study show that 'P1_serve' (serve) is the most important feature in determining the momentum of the game, with a feature importance score of 0.528, followed by They are 'P1_scoreAD' (advantage points) and 'P1_ace' (serve points). The findings indicate that machine learning methods can provide a more in-depth way to quantitatively analyze game dynamics than descriptive analysis of statistical data. The analysis pointed out the importance of serving the outcome of the game. Good serving ability can put pressure on the opponent and create opportunities to win the score. At the same time, this study also demonstrates the potential application of the research method in-game strategy decision-making, providing coaches and athletes with a new perspective to help them understand how psychological and strategic factors interact in the game process. Nonetheless, this study also has certain limitations. The dataset used is from only one specific Wimbledon match, which may limit the generalizability of the results. Additionally, due to limitations in model design and training data quality, future research could overcome these limitations by using broader datasets and more complex models. Future explorations could be expanded to broader data sets and incorporate additional variables to deepen understanding of the role of momentum in tennis and other sports. The research results not only promote the cross-integration of sports science and data science fields but also provide important decision support tools for athletes and coaches in practical applications, helping them better understand and utilize dynamic changes in the game and optimize tactics. and strategies to improve the winning rate of the game.

References

[1] Rivera, J.: Tennis scoring, explained: A guide to understanding the rules, terms & point system at Wimbledon. The Sporting News. Retrieved April 5, 2024, from https://www.sportingnews.com/us/tennis/news/tennis-scoring-explained-rules-system-points-terms/7uzp2evdhbd11obdd59p3p1cx. last accessed 2024/16/01.

[2] Momentum.: Merriam-Webster.com Dictionary, Merriam-Webster, https://www.merriam-webster.com/dictionary/momentum. last accessed 2024/07/04.

[3] Meier, P., Flepp, R., Ruedisser, M., Franck, E.: Separating psychological momentum from strategic momentum: Evidence from men’s professional tennis. Journal of Economic Psychology, 78, Article 102269 (2020).

[4] Dietl, H., Nesseler, C.: Momentum in tennis: Controlling the match. UZH Business Working Paper Series (No. 365). University of Zurich (2017).

[5] Silva, J. M., Hardy, C. J., Crace, R. K.: Analysis of psychological momentum in intercollegiate tennis. Journal of Sport & Exercise Psychology, 10, 346-354 (1988).

[6] Leo, B.: Bagging predictors, Machine Learning, 24.2: 123-140 (1996).

[7] Rosenblatt, F.: The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, 65(6), 386 (1958).

[8] Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning, 20(3), 273–297 (1995).

[9] Anderson, R., Anderson, L. F.: AI and machine learning in sports analytics: A New Era of data-driven athlete performance and match strategy. Journal of Sports Analytics, 7(1), 54-65 (2021).

[10] Zhao, Y., van Buiten, M.: Real-time strategy optimization using machine learning: A case study on tennis. International Journal of Performance Analysis in Sport, 22(2), 245-260 (2022).

[11] Martin, A. P., Jones, G., Ibrahim, K.: Implementing Machine Learning Techniques for Tactical Insights in Professional Tennis. Journal of Computational Science, 45, Article 101329 (2022).

[12] Lee, D., Kim, J. Y., Han, O.: The role of data science in sports decision making: current landscape and future prospects. Data Science and Management, 3(2), 1-12 (2020).