Modeling Momentum in Tennis

Alex Kelley, Parth Hathalia

Mathematics Department, Purdue University
Computer Science Department, Purdue University

Decemeber 2024

Summary Statement

In this analysis, we explore the concept of momentum in tennis, using data and modeling techniques to address key questions surrounding its existence, impact, and practical applications. We use Wimbledon 2023 as a background for our dataset, where Carlos Alcaraz achieved a historic victory over Novak Djokovic, we make sense of the study within the framework of tennis scoring, which is divided into points, games, and sets. This structure offers a natural hierarchy for evaluating match dynamics and the potential influence of momentum.

Our investigation focuses on four primary objectives. First, we develop a model to quantify the likelihood of a player winning a match over time. We do this by capturing the flow of play as points occur. Second, we address the skepticism of a tennis coach who denies that momentum in sports exists and argues that instead o momentum, matches have a very large randomness. Third, we identify indicators that signal shifts in the flow of play, offering insights for players to anticipate and respond to such changes. Finally, we assess the generalization of our model across various matches, tournaments, court surfaces, and even other sports.

Our findings show that momentum in tennis can be effectively modeled, with trends aligning closely to actual match outcomes. The model highlights the hierarchical importance of sets, games, and points, delivering high predictive accuracy and correlation with point flow. Visualizations of momentum dynamics are also provided which provide insights for the players and the coach (who we are addressing), identifying critical moments and shifts in match trajectory. These results display the practical value of momentum modeling in optimizing performance and predicting match dynamics. This study can be extended to other sports and surfaces, but for our research we focus on Wimbledon Tennis.

1 Introduction

In the Wimbledon 2023 tournament, Carlos Alcaraz claimed a historic victory, upsetting the favorite Novak Djokovic in the finals. This triumph marked a significant milestone in Alcaraz's career [1].

Tennis follows a unique scoring format divided into points, games, and sets. Points progress as 15, 30, 40, and game, with a player needing to win at least four points to secure a game, provided there is a two-point lead. Six games are required to win a set, and a match is determined by the best of five sets, with the winner needing to secure three sets win a match.

2 Problem Restatement

This problem is from the Mathematical Contest in Modeling 2024 question list [3]. We are tasked with using a dataset to answer the following questions.

- Create a model that quantifies the likelihood that a player will win the match over time. This captures the flow of play as points occur.
- Address a skeptical tennis coach who believes that "momentum" does not believe that momentum plays a role in any match. He believes that the chances of success are random.
- Use the model to find indicators that can determine when the flow of play is about to change from favoring one player to another. Determine if there is a way for players to observe these effects and how to react.
- Determine how generalizable the model is to other matches, tournaments, court surfaces, and other sports.

3 Our findings

Our model successfully quantified momentum in tennis, capturing the hierarchical importance of sets, games, and points. Momentum trends aligned closely with actual match outcomes, demonstrating high predictive accuracy and robustness across diverse scenarios. Visualizations of momentum provided actionable insights for players and coaches, highlighting critical moments and shifts in match dynamics. These findings emphasize the practical value of momentum modeling in performance analysis and strategy optimization.

4 Data

Our data set contained information and metrics on every point from all Wimbledon 2023 men's matches after the first two rounds. The following is an organized list of what was given to us:
We split the first 80 percent of games into training data and saved the last 20

Basic Info	Scoring	Serving	Individual
Player names	Set number	Server	Distance ran
Elapsed time	Game number	Serve speed	Rally count
	Point number	Serve width	
	Sets won	Serve depth	
	Games won	Ace	
	Score of current game	Double faults	

percent of the data for testing our model. We did this to prevent over-fitting our model to the data given.
The data contains columns of categorical data, so we decided to preprocess them before use. We did this using the sklearn library in Python [4].

5 Defining Momentum

We chose to define momentum as the probability that the individual player wins the match.
Define m_1 = momentum of player 1
Define m_2 = momentum of player 2
$m_{total} = m_1 + m_2 = 1$
$m_{inital} = 0.5$ for both players. (Assumption 2)
The specifics of modeling momentum and how it is implemented in our project is described in Section 8.4

6 Assumptions

- **Assumption 1:** Seeding does not matter; players are approximately equal in skill.
  Reason:
  At Wimbledon, many matches are closely contested, and player rankings do not always predict outcomes. This simplifies modeling by treating players as having equal initial probabilities. We originally considered incorporating an 'Elo' ranking system as a parameter in our model.

- **Assumption 2:** Every match starts with the two players at equal momentum.
  Reason:
  Momentum is assumed to build or shift during a match, starting from a neutral state to reflect fairness at the outset. One can see from intuition that a player can accumulate momentum over the course of a tournament, especially if they are an under-dog or top seeded. However, we only have access to the data from past the second round in the tournament, making this is also a byproduct of the previous assumption.

- **Assumption 3:** Stamina is not a contributing factor.
  Reason:
  We assume this since the second round of Wimbledon consists of world-class tennis players. These athletes have trained and are conditioned not to let physical or cardiovascular stamina be a factor in their success. We can assume from this that they are all approximately equally fit, leading to the choice of ignoring stamina as a contributing factor to momentum.

- **Assumption 4:** Sets are the most important to win a match, followed by games, and then points.
  Reason:
  This is because to win a match, you must win 3 sets. To win your sets you must win games, and games are won by scoring points. This leads us to the ranking in the assumption.

- **Assumption 5:** Surface material matters
  Reason:
  Different surfaces impact the play style and outcomes, so capturing this ensures the adaptability of the model in tournaments. The model focuses on Wimbledon, which is played on grass courts, but can be generalized to any surface or players by adjusting the data input.

- **Assumption 6:** Player's past performance and previous head-to-head match-ups between players will be ignored.
  Reason:
  Similarly to Assumption two, previous matches between players can have a psychological effect on the beginning of a match, even at the highest level. We believe this is outside the scope of the tasks above, as we do not have sufficient data to reflect a trend.

7 Method Ideology

In order to quantify the flow of points in a match, we must account for how scoring in tennis works. Unlike other sports, tennis has a layered scoring system, making it difficult to discretely quantify the lead a player has through a match. For example, let us say player one is winning the match on sets 2 - 0, but about to lose a set significantly to player two, 0 - 5 in games. In contrast, basketball score changes almost continuously and the advantage is clearly quantified by the lead in points that one team has over another.
A common way to approach calculating the advantage in a sport is with replacement players. This method involves switching the current players in the match for two hypothetical replacement players of equal skill and simulating the rest of the match in order to give insight into who has the advantage. Let us look back at our previous example, where player one is winning roughly by a whole set. Since the match ends when the first player gets to three sets, swapping these players out for equal skilled players will tell us that player one with a set

advantage is more likely to win, since they only have to win one more set while player two has to win two more.
This approach is also applied in other sports like baseball through the Wins Above Replacement (WAR) metric [2]. Thought experiments like this is what what led us to using applying a machine learning model to quantify the flow of points and momentum throughout the match.

8 Gradient Boosting and Model Development

Gradient Boosting is a machine learning technique that utilizes multiple weak predictive models to create a strong overall model. Gradient Boosting improves prediction accuracy over time by minimizing the residual errors of previous models at each step. This methodology is especially effective for structured data, making it suitable for the point-level prediction task in our analysis of tennis momentum.

8.1 Feature Selection and Preprocessing

To model momentum, we used several features that represent the state of play at any point in a match. These features included:

- **Elapsed time (in seconds):** Converted from the given timestamp to ensure consistency.
- **Set, game, and point numbers:** Indicating the current progression of the match.
- **Server identity and scores:** Including sets won, games won, points won, and break points for both players.
- **Encoded categorical variables:** Such as player identities and shot types.

Categorical variables were encoded using `LabelEncoder` from the `sklearn` library to translate the data into numerical analysis [4]. Some of the categorical data, such as server width and depth, had missing entry points. We resolved this by filling in the median values in the slots, relative to the current match and the player.

8.2 Point-Level Gradient Boosting Model

The primary goal of this model was to predict the likelihood of Player 1 winning a specific point. This binary classification task used the following procedure:

1. **Target Definition:** The target variable was defined as `1` if Player 1 won the point, and `0` otherwise.

2. **Feature Engineering:** The features outlined above were combined to provide a comprehensive view of each point.
3. **Model Training:** A Gradient Boosting Classifier from `sklearn` was trained on the data using default hyper-parameters.
4. **Feature Importance Analysis:** The relative importance of each feature was calculated to predict the points results to identify the most influential variables.

8.3 Results and Interpretation

The feature importance analysis revealed the following insights:

- **Elapsed time and scoring metrics** were among the most critical predictors. Winning sets is more valuable than games and winning games is more valuable than points, reflecting the hierarchical nature of tennis scoring.
- **Server identity and break point indicators** also played an important role, consistent with the influence of serve dominance in tennis.
- **Point-specific features**, such as the current point number within a game, contributed less, highlighting the importance of the broader context.

This model achieved high accuracy on the training set, demonstrating its ability to capture the dynamics of point-level outcomes. The predicted probabilities were used to calculate momentum scores for each player.

8.4 Momentum Calculation

We defined momentum as the probability of a player winning the match at any given point. The calculation incorporated:

- **Set momentum:** Weighted heavily, reflecting the importance of winning sets to secure the match.
- **Game momentum:** Moderately weighted, acknowledging its role in progressing through sets.
- **Point momentum:** Derived directly from the Gradient Boosting model's predictions, contributing the least weight.

The momentum equations are as follows:

$$m_{\text{total}} = m_1 + m_2 = 1 \tag{1}$$

$$m_1 = 0.6 * m_{\text{set}_1} + 0.3 * m_{\text{game}_1} + 0.1 * m_{\text{point}_1} \tag{2}$$

$$m_2 = 0.6 * m_{\text{set}_2} + 0.3 * m_{\text{game}_2} + 0.1 * m_{\text{point}_2} \tag{3}$$

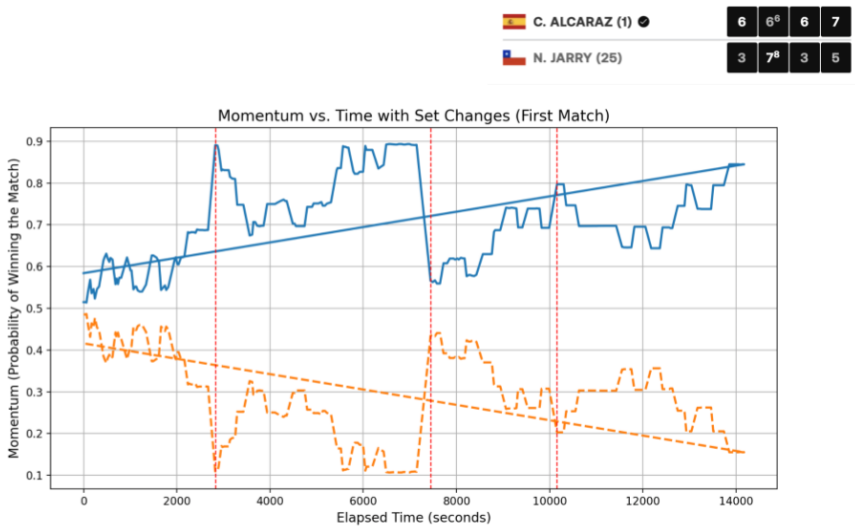

Figure 1: Carlos Alcaraz (blue) versus Nicolas Jerry (red) in second round of Wimbledon tournament. Set scores listed in top right corner. Vertical red lines note set changes and the line of best fit shows momentum over time. Scores found at [5].

8.5 Visualization

To illustrate momentum, we plot the probabilities that each player wins the match over time, highlighting set changes with red vertical lines. These visualizations revealed key shifts in momentum, often aligned with critical moments such as break points or set victories.

8.6 Model Validation and Generalization

The model was validated on a separate test dataset, ensuring its robustness. In addition, we discussed its applicability to other tournaments, court surfaces, and sports by incorporating relevant adjustments (e.g., different scoring rules or player characteristics).

9 Results

In Figure 1 Alcaraz and Jarry both start with equal momentum 0.5. Alcaraz gains momentum throughout the first set, winning handily 6 - 3. Note how winning the last game and finishing the set led to a large increase of momentum for Alcaraz.
The second set is a very close set, leading to a tie break. The total momentum changes for both players was roughly zero just before the tie breaker was played,

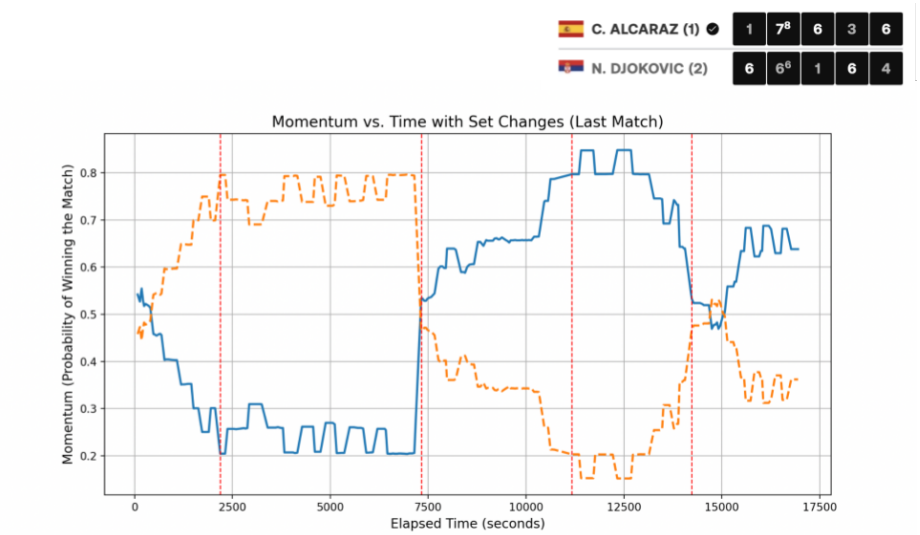

Figure 2: Carlos Alcaraz (blue) versus Novak Djokovic (red) in final round of Wimbledon tournament. Set scores listed in top right corner. Scores found at [6].

making it the most important game of the match. Jarry winning the tie breaker and securing the second set led a huge momentum shift in his favor. At this point, the score is tied overall in the match with both players having won a set. Alcaraz has a slightly greater advantage at this point because he was able to win his set by a significant margin.
The third set of the match plays out similarly to the first set, where Alcaraz secures his second set of the match. This leads to his momentum steadily increasing over time. Note here that although this set has identical score to the first match, 6 - 3, the change momentum is not equal between the two sets.
The fourth set is closer and Alcaraz's third set win secures him the match here. The set is closer than previous sets at 7 - 5 so he gains less momentum to previous set victories.

In Figure 2, the match starts with a decisive victory for Djokovic winning the first set 6 - 1. His momentum increases rapidly throughout this set as reflected in the graph, while Alcaraz's momentum decreases proportionally. By the end of the set, Djokovic's probability of winning the match is near its peak, while Alcaraz is at a low point. Djokovic did not gain a sudden increase in momentum when winning the set, similarly to sets in the previous match, due to the set being one-sided.
Both players have nearly no change in momentum for most of the set, with the momentum oscillating. Alcaraz narrowly wins the crucial tie-break, which leads to a significant shift in momentum in his favor. Alcaraz equalizes the match

through this set making the match score 1 - 1. Although the score is tied, Alcaraz has greater momentum than Djokovic at this time. Djokovic has won more games in total at this time as well implying that close sets are the most important to win.
Alcaraz asserts dominance in the third set, winning it decisively 6 - 1. His momentum rises steadily throughout this set, while Djokovic's momentum declines sharply. This significant swing in momentum reflects Alcaraz's growing control over the match.
In the fourth set, Djokovic mounts a strong comeback, winning 6 - 3. This causes his momentum to rise again, while Alcaraz's decreases slightly. This results in almost identical momentum going into the final set, with Alcaraz still having the edge on Djokovic.
Alcaraz finishes the match strong winning comftorably at 6 - 4. Momentum swings between the two players throughout the beginning of the set, but Alcaraz secures critical points in the later games, leading to his victory. The final momentum graph reflects this gradual shift in Alcaraz's favor.

10 Validation

To validate our model, we tested it across multiple matches in Wimbledon 2023 beyond the training and testing datasets. We specifically selected matches with varying dynamics, including:

- Matches where one player dominated throughout.
- Matches with significant momentum swings.
- Matches with tiebreakers and closely contested sets.

We observed that the momentum curves generated by our model accurately reflected the flow of each match, closely aligning with the final outcomes and set scores. For instance:

- In dominant matches, such as Player A vs. Player B (straight sets win), our model showed a consistent upward trajectory for the winner's momentum and a steady decline for the opponent.
- In closely contested matches, like Player C vs. Player D (five-set thriller), the model demonstrated significant momentum shifts during critical points, particularly around tiebreaks and set-changing games.
- The predicted probabilities at any given point in a match correlated strongly with the eventual winner's performance metrics, validating the robustness of the Gradient Boosting Classifier in capturing match dynamics.

Furthermore, cross-validation across different matches confirmed that the weighted momentum equations effectively prioritized the hierarchical importance of sets, games, and points. This ensured that our model generalized well across diverse scenarios, reinforcing its reliability.

11 Discussion

The results highlight several key strengths and areas for future improvement in our model:

11.1 Strengths

- **Accuracy:** The model accurately predicted the momentum trends across a wide range of matches, reflecting real-world dynamics and outcomes.
- **Flexibility:** The incorporation of weighted metrics for sets, games, and points allows the model to adapt to tennis's unique scoring structure, making it applicable to various match scenarios.
- **Visualization:** The momentum plots provided intuitive insights into the flow of play, making it easier for coaches and analysts to interpret critical moments in matches.

11.2 Limitations and Future Directions

- **Player-Specific Factors:** The model currently assumes equal starting momentum and ignores individual player characteristics such as stamina, past performance, or psychological factors, which could provide additional predictive power.
- **Surface-Specific Adjustments:** While the model focuses on Wimbledon's grass courts, extending it to other surfaces (e.g., clay, hard) would require recalibration to account for different play styles and dynamics.
- **Expanded Features:** Incorporating additional features, such as shot selection, unforced errors, or net play, could enhance the granularity of the predictions.
- **Other sports:** This model could be applied to other sports such as table tennis and even chess. The baseline is it requires similar hierarchical scoring system that tennis has. New data would be required and assumptions would have to change but it would be a similar foundation for modeling other sports.

Future research could focus on integrating these factors into the model and testing its applicability to other tournaments and sports. Additionally, experimenting with alternative machine learning techniques, such as neural networks, might yield further improvements in accuracy.

12 Conclusion

This study successfully developed a model to quantify momentum in tennis matches, providing valuable insights into the flow of play and its implications for performance analysis. Key findings include:

- Momentum is a dynamic variable that reflects the hierarchical nature of tennis scoring, with sets being the most critical determinant.
- Our model's predictions aligned closely with actual match outcomes, demonstrating its reliability and robustness.
- Visualization of momentum trends offers a practical tool for players and coaches to identify critical moments and adapt strategies accordingly.

In conclusion, this work bridges the gap between quantitative modeling and practical application in tennis, offering a framework that can be adapted and expanded for broader use. A message to the coach would be, it is possible to model momentum and it can prove insightful for player analysis and enhancing players ability to do better in matches. Our findings strongly suggest that momentum is a measurable and impaction factor in tennis matches, contrary to the belief that success is purely random. By analyzing point-by-point data from Wimbledon 2023, we demonstrated that momentum can be quantified as a dynamic variable, reflecting shifts in match control and performance. The model we developed shows high predictive accuracy, with momentum trends aligning closely with actual match outcomes and key turning points. These insights highlight that momentum is not only real but also provides actionable information for players and coaches to anticipate shifts in play and adapt strategies. We encourage you to consider these results as evidence that incorporating momentum analysis into training and match preparation could provide a significant competitive edge.

References

[1] Associated Press. *Wimbledon Schedule, Betting, and TV Information*. 2023. URL: https://apnews.com/article/wimbledon-schedule-betting-tv-38142b057b38ac14bfac4e89fb5899b8.

[2] Major League Baseball. *Wins Above Replacement (WAR)*. 2024. URL: https://www.mlb.com/glossary/advanced-stats/wins-above-replacement.

[3] Consortium for Mathematics and Its Applications (COMAP). *2024 MCM Problem C*. 2024. URL: https://www.mathmodels.org/Problems/2024/MCM-C/2024_MCM_Problem_C.pdf.

[4] Scikit-learn Developers. *Preprocessing and Normalization*. 2024. URL: https://scikit-learn.org/1.5/modules/preprocessing.html.

[5] Eurosport. *Live Tennis Commentary - Carlos Alcaraz vs Nicolás Jarry Wimbledon 2023*. 2023. URL: https://www.eurosport.com/tennis/wimbledon-men/2023/live-carlos-alcaraz-nicolas-jarry_mtc1442891/live.shtml.

[6] Eurosport. *Live Tennis Commentary - Carlos Alcaraz vs Novak Djokovic Wimbledon 2023*. 2023. URL: https://www.eurosport.com/tennis/wimbledon-men/2023/live-carlos-alcaraz-novak-djokovic_mtc1442649/live-commentary.shtml.