Momentum Dynamics in Tennis: Analyzing Performance Through Point-by-Point Insights

Vaibhav Gupta (2024201044), Sayam Agarwal (2024201025), Priyanshu Jha (2024201062)

Abstract

Professional tennis, characterized by its dynamic nature and inherent unpredictability, presents significant challenges for match outcome prediction. While traditional models often focus on static player attributes or simplified scoring probabilities, they frequently overlook the crucial influence of in-match momentum—the psychological and performance shifts that can dictate the flow of a contest. This project addresses this gap by developing and evaluating machine learning models that explicitly quantify and incorporate momentum to predict professional tennis match winners. Utilizing point-by-point data from publicly available ATP sources, such as Jeff Sackmann's Tennis Database, we introduce a novel momentum score. This score is derived using Hidden Markov Models (HMM) to identify latent player performance states, with the resulting probabilities smoothed via Exponential Moving Averages (EMA) to create a continuous momentum trajectory. This dynamic feature is then integrated into tree-based ensemble models (XGBoost and LightGBM) to predict match outcomes. The significance of momentum is assessed, and SHAP (SHapley Additive exPlanations) values are employed to interpret model predictions and identify key factors, including specific in-match events, that drive momentum swings. Prioritizing predictive accuracy and a deeper understanding of match dynamics over betting applications, this research aims to provide a robust framework for tennis match analysis, offering potential insights for coaching strategies, player development, and enhanced fan engagement. The methodology encompasses data preprocessing, momentum quantification, model training, and comprehensive performance evaluation.

1 Introduction

Tennis, a globally celebrated sport, captivates millions with its dynamic interplay of skill, strategy, and physical endurance. Matches unfold across diverse surfaces such as clay, grass, and hard courts, each presenting unique challenges and influencing player tactics. The Association of Tennis Professionals (ATP) and Women's Tennis Association (WTA) organize a dense calendar of over 60 tournaments annually, drawing immense global spectatorship and fueling a significant sports betting market. To illustrate, the 2013 Wimbledon final between Andy Murray and Novak Djokovic attracted an estimated 17.3 million viewers in Great Britain alone, with approximately £48 million traded on a single betting exchange, Betfair, underscoring the sport's substantial economic and cultural impact (Sipko, 2015). This widespread engagement is partly driven by the inherent unpredictability of match outcomes. Factors such as individual player form, historical performance on specific surfaces, psychological resilience, physical fatigue over extended matches, and crucial head-to-head dynamics all contribute to a complex, often fluctuating, competitive environment, presenting a compelling challenge for predictive modeling.

Traditional approaches to tennis prediction have often relied on stochastic models, primarily exploiting the sport's hierarchical scoring system: points aggregate into games, games into sets, and sets into a match. Many early models, such as those based on Markov chains, utilized simplified metrics like fixed probabilities of winning a point on serve (Barnett and Brown, 2006, as cited in Lei et al., 2024). While subsequent refinements attempted to incorporate dynamic serve probabilities updated after each round (Newton, 2006, 2009, as cited in Lei et al., 2024), these methods often overlook or oversimplify nuanced, time-varying variables. A critical, yet often underrepresented, factor in these models is the concept of **in-match momentum**. Spectators and players alike frequently observe and comment on shifts in the "flow" of a match, where one player gains a psychological or performance advantage, leading to a sequence of successful points or games. This phenomenon, characterized by subtle shifts in player confidence, energy levels, and tactical execution, can drastically alter the trajectory of a contest. Capturing this elusive momentum is paramount because factors like accumulating player fatigue, psychological resilience under pressure, and recent performance trends (e.g., winning several crucial points consecutively) are difficult to quantify with purely stochastic point-based models that assume point independence or rely on static player attributes. Machine learning (ML) offers a promising alternative by leveraging extensive historical data to learn complex patterns and interactions between diverse features, including those indicative of momentum, thereby enabling more accurate and adaptable predictions that can account for the non-linear, dynamic nature inherent in a tennis match. The explicit modeling of momentum, therefore, stands as a key area for enhancing the predictive power and explanatory depth of tennis match analysis.

2 Related Work

The prediction of tennis match outcomes and the analysis of in-match dynamics, particularly momentum, have been subjects of academic inquiry for several decades. Research in this area can be broadly categorized into statistical modeling of match progression and the application of machine learning techniques for outcome prediction and feature analysis.

Stochastic Models for Tennis Prediction

Early efforts to model tennis matches often employed stochastic processes, with Markov chains being a prominent approach. These models leverage the hierarchical structure of tennis scoring (points, games, sets) to estimate win probabilities. For instance, Barnett and Brown (2006, as cited in Lei et al., 2024) introduced the use of Markov chains to calculate game, set, and match winning probabilities based on the fundamental probabilities of each player winning a point on their serve. A core assumption in many such initial models was that the probability of a player winning a service point remained constant throughout a match.

Subsequent research sought to refine these models by incorporating more dynamic elements. Newton and Aslam (2006, as cited in Lei et al., 2024) proposed a Markov model where the probability of a player winning a service point was updated after each game (or round, in their tournament simulation context). They further extended this by adjusting a player's service point-winning probability based on the opponent's ability, thereby making the probabilities matchup-specific (Newton % Aslam, 2009, as cited in Lei et al., 2024). Barnett and Clarke (2005, as cited in Lei et al., 2024) also considered non-constant service point probabilities, suggesting that if a player takes a lead in sets, their probability of winning subsequent sets might increase, reflecting a potential momentum effect. While these models provide a solid mathematical framework for understanding tennis scoring probabilities, they often simplify or indirectly account for the complex, time-varying psychological and performance factors that constitute in-match momentum.

Modeling Momentum in Sports

The concept of "momentum" itself—often described as a period where a player or team performs significantly above or below their average level, leading to a swing in the game's dynamics—has been investigated across various sports. Qualitative studies, such as interviews with players, have reported the perceived existence and impact of psychological momentum in tennis (Richardson et al., 1988, as cited in Lei et al., 2024) and other sports like soccer (Jones % Harwood, 2008, as cited in Lei et al., 2024).[1]

More quantitative approaches have sought to model these unobserved, sequence-dependent processes. Hidden Markov Models (HMMs) have emerged as a suitable technique for such tasks, as they can model systems where the observed outcomes (e.g., point results) are influenced by underlying, unobservable "hidden" states (e.g., a player being in a high-momentum or low-momentum state) (Zucchini % MacDonald, 2009, as cited in Lei et al., 2024). Ötting et al. (2021, as cited in Lei et al., 2024) applied a copula-based multivariate HMM to model momentum in football by analyzing minute-by-minute summary statistics. The work by Lei et al. (2024) directly applies HMMs to tennis point-by-point data to quantify momentum, defining it as the evolving performance state of players, and subsequently uses Exponential Moving Averages (EMA) to smooth these probabilities into a continuous momentum score.[5, 3]

Machine Learning Approaches and Feature Importance

With the increasing availability of detailed sports data, machine learning techniques have become more prevalent in predicting match outcomes and identifying key performance indicators. These models can often capture more complex, non-linear relationships in data compared to traditional stochastic models. Tree-based ensemble methods like XGBoost and LightGBM, as used by Lei et al. (2024), are particularly popular due to their strong predictive performance and ability to handle diverse feature types.[6, 2]

A significant challenge with many ML models is their "black-box" nature. Techniques for model interpretability have thus become crucial. SHAP (SHapley Additive exPlanations), developed by Lundberg and Lee (2017, as cited in Lei et al., 2024), provides a unified approach to explain the output of any machine learning model by assigning each feature an importance value (Shapley value) for a particular prediction. This allows for a deeper understanding of which factors, including engineered features like momentum, contribute most significantly to a model's decisions. Lei et al. (2024) utilize SHAP to analyze feature importance in their LightGBM model for predicting momentum swings.

Positioning of the Current Project

This project aligns with and extends the line of research focused on quantifying and utilizing in-match momentum for tennis match prediction. Similar to Lei et al. (2024), we employ HMMs to model latent performance states and derive a momentum score. This momentum score is then incorporated as a novel feature into machine learning models (XGBoost and LightGBM) to predict point and, ultimately, match outcomes. A key aspect of our work is not only to assess the predictive uplift provided by this momentum feature but also to use SHAP analysis to understand its interplay with other traditional tennis statistics (e.g., aces, winners, errors) in influencing predictions. While some research has focused on leveraging predictions for betting markets (Cornman et al., n.d.; Sipko, 2015), our primary objective is to enhance the accuracy of match outcome prediction and to provide interpretable insights into the dynamics of momentum, which can be valuable for coaches, players, and analysts. We aim to build upon the existing literature by providing a robust framework for momentum-aware tennis analytics.[4]

3 Data Sources

The data used for this project is sourced from Jeff Sackmann's open-source tennis data repository on GitHub, specifically the "tennis pointbypoint" dataset point by point dataset. This repository provides sequential point-by-point statistics for tens of thousands of professional tennis matches, including major tournaments such as Wimbledon, the French Open, US Open, and Australian Open. The dataset is licensed under a Creative Commons Attribution-Non Commercial-Share Alike 4.0 International License, requiring proper attribution for non-commercial use.

For this project, we focused on the point by point data analysis , as it aligns with the problem statement of whether momentum affects sports psychology and the mechanisms underlying its generation and influence merits thorough investigation. In this project, taking the 7,284 scoring points in the men's singles tennis match at Wimbledon 2023 as an example, we expand upon traditional momentum research by integrating diverse algorithms, including statistical analysis and linear weighting, which aims to quantify the momentum dynamics for athletes in a match. Specifically, we utilized the "Wimbledon_featured_matches.xlsx" dataset, which was derived from Jeff Sackmann's original point-by-point data through transformations to suit our analysis needs. These transformations were necessary to work with the data, which relies on detailed point-level statistics.

The dataset covers matches from major tournaments, including Wimbledon. This dataset is particularly suitable for studying momentum because it provides granular information on each point, such as scores, serve speeds, rally lengths, and even precomputed momentum values. By analyzing this data through the lens of the capturing momentum model, we can gain insights into how momentum influences match outcomes and player performance.

4 Data Exploration

The data exploration phase is critical for understanding the dataset's structure, quality, and suitability for analyzing tennis match momentum. Below, we detail the steps taken during this phase, including dataset overview, handling of missing values, outlier analysis, and initial visualizations.

4.1 Dataset Overview

The dataset used for this project, "Wimbledon_featured_matches.xlsx", contains detailed point-by-point statistics for selected Wimbledon matches. Key columns include:

- `speed_mph`: Serve speed in miles per hour (mph).
- `serve_width`: Direction of the serve (e.g., Body, Center, Wide).
- `serve_depth`: Depth of the serve (Close To Line or Not Close To Line).
- `return_depth`: Depth of the return shot (Deep or Not Deep).
- `p1_score` and `p2_score`: Current game scores for players 1 and 2 (standard tennis scoring: $0, 15, 30, 40, AD$).
- `winner_shot_type`: Type of winning shot (Forehand or Backhand).
- `rally_count`: Number of shots exchanged in the point.

A comprehensive list of all variables in the dataset is provided in Table 1, which ensures clarity and enhances the reproducibility of our analysis.

4.2 Handling Missing Values

During the initial exploration, missing values were identified in several columns:

Table 1: Description of Columns in the "Wimbledon_featured_matches.xlsx" Dataset

No.	Variable	Explanation
1	match_id	Match identification
2	player1	First and last name of the first player
3	player2	First and last name of the second player
4	elapsed_time	Time elapsed since start of first point to start of current point (HMMSS)
5	set_no	Set number in match
6	game_no	Game number in set
7	point_no	Point number in game
8	p1_sets	Sets won by player 1
9	p2_sets	Sets won by player 2
10	p1_games	Games won by player 1 in current set
11	p2_games	Games won by player 2 in current set
12	p1_score	Player 1 score within current game
13	p2_score	Player 2 score within current game
14	server	Server of the point
15	serve_no	First or second serve
16	point_winner	Winner of the point
17	p1_points_won	Number of points won by player 1 in match
18	p2_points_won	Number of points won by player 2 in match
19	game_winner	A player won a game at the point
20	set_victor	A player won a set at this point
21	p1_ace	Player 1 hit an untouchable winning serve
22	p2_ace	Player 2 hit an untouchable winning serve
23	p1_winner	Player 1 hit an untouchable winning shot
24	p2_winner	Player 2 hit an untouchable winning shot
25	winner_shot_type	Category of untouchable shot
26	p1_double_fault	Player 1 missed both serves and lost the point

- `speed_mph`: 752 missing values
- `serve_width`: 54 missing values
- `serve_depth`: 54 missing values
- `return_depth`: 1309 missing values

To ensure data integrity, the following methods were applied:

- For `speed_mph` (numerical), missing values were imputed with the median serve speed of the respective player. This approach accounts for player-specific variations in serve speed and aligns with standard practices for handling missing numerical data.
- For categorical columns `serve_width` and `serve_depth`, a new category "Unknown" was introduced to represent missing values, preserving data for analysis and following common strategies for categorical data.
- For `return_depth`, with 1309 missing values, further investigation revealed that many of these corresponded to points with no return (e.g., aces or faults). A new category "No Return" was created for these instances, enhancing the dataset's utility for momentum analysis. This finding is

particularly relevant, as it captures points where one player gains an immediate advantage, which is crucial for momentum modeling.

This handling of missing values ensures that the dataset remains comprehensive while minimizing bias, aligning with best practices for data preprocessing.

4.3 Distribution of `speed_mph`

The distribution of serve speeds (`speed_mph`) was analyzed to understand player performance patterns. The distribution is illustrated in Figure 1.

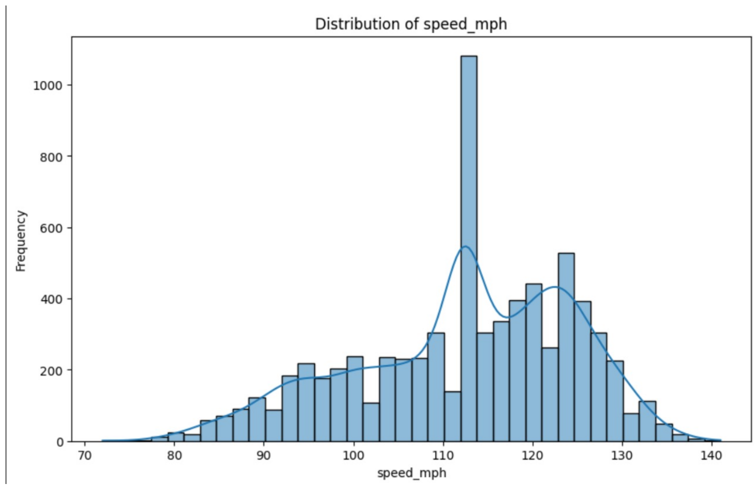


Figure 1: Distribution of `speed_mph`

4.4 Outlier Count

Outlier analysis was conducted to identify extreme values in the dataset. The outlier count per column is illustrated in Figure 2, providing a visual summary of potential anomalies that may affect momentum analysis.

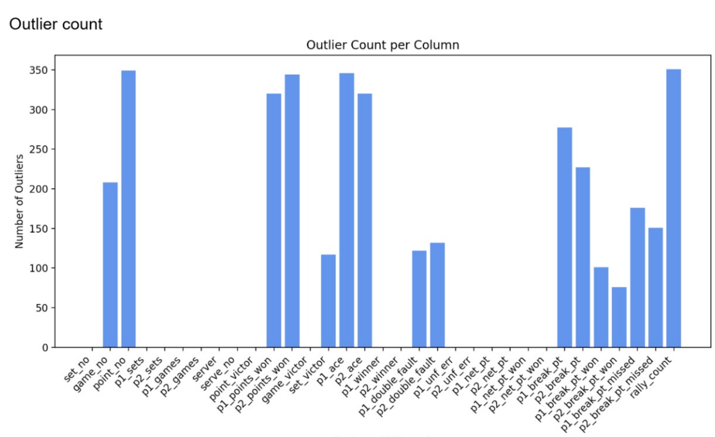


Figure 2: Outlier count per column

4.5 Game Data Analysis

To conduct a preliminary investigation on the match flow, we visualized the number of points, games, and sets won by each player over time, taking the 2023 Wimbledon Men's Singles final as an example. Time series plots of games and points were analyzed, revealing that at the set level, points scored by the two players in the third set were extremely close, indicating similar performance. However, the fourth and fifth sets witnessed a reversal of potential winners, where the match grew increasingly intense. After a tortuous battle, Alcaraz achieved the ultimate victory, winning 3 sets out of 5.

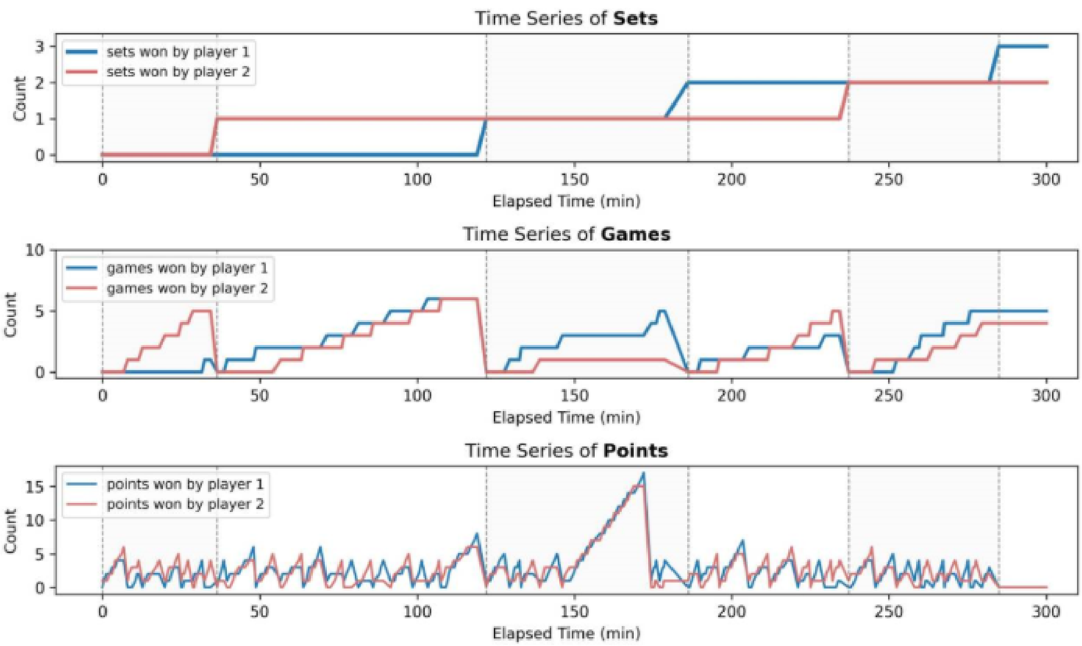
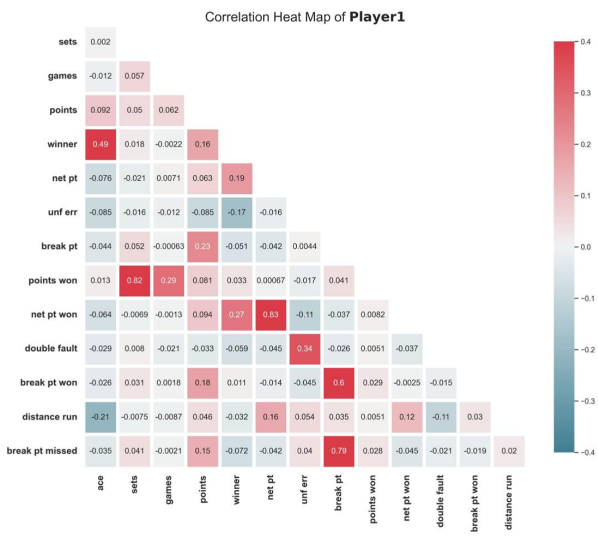


Figure 3: Time series of points

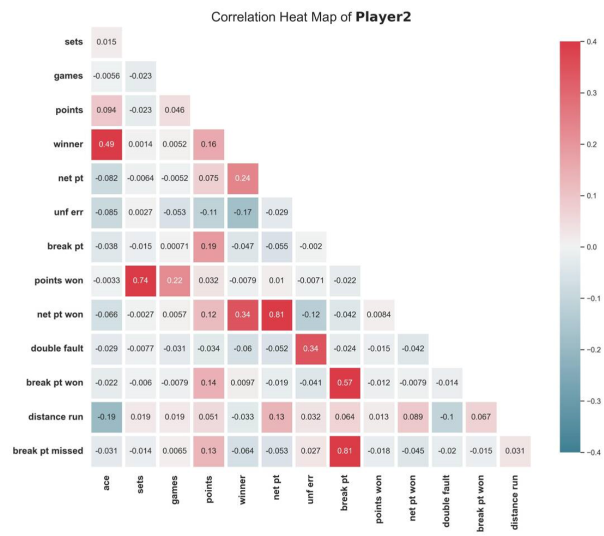
4.6 Dependent and Independent Variables Analysis

Since this study aims to quantify momentum and investigate its existence, formation mechanism, and effect on players' behavior, the dependent variable of the model is momentum. In further research, the dependent variable is the players' scores, i.e., momentum affects the players' scores by influencing their behaviors.

For independent variables, unexpected events in a match can randomly affect the mental and physical state of the players, which in turn impacts the match outcome. Positive events can provide players with positive psychological cues, potentially making it easier for them to score in the next round. Given the numerous variables in tennis competitions, categorizing and correlating these variables is necessary. To explore the correlation of each index in the original data, we drew correlation heatmaps for each player, as presented in Figures 4a and 4b. The heatmaps indicate that the majority of indices exhibit weak connections, suggesting that the correlation between indices in the original data is not obvious. Therefore, new indicators need to be formulated to better characterize the match.



(a) Correlation heatmap of Player 1



(b) Correlation heatmap of Player 2

Figure 4: Correlation heatmaps for Player 1 and Player 2

5 Methodology

This project employs a multi-stage methodology to analyze tennis match data, quantify momentum, and predict match outcomes. The process encompasses data acquisition and preprocessing, momentum feature engineering using Hidden Markov Models (HMM) and Exponential Moving Averages (EMA), development and training of machine learning models, and finally, model evaluation and interpretation. The overall workflow is depicted in Figure 5.

Project Workflow Overview

Data Acquisition and Preprocessing

The primary dataset utilized for this project is the point-by-point tennis match data from Jeff Sackmann's Tennis Database, specifically focusing on the `2011-wimbledon-points.csv` file. This raw dataset contains a rich set of features recorded for each point played.

A crucial initial step involves preprocessing this raw data to transform it into a structured format suitable for analysis and model training. This is handled by a dedicated Python function which performs the following operations:

1. **Column Standardization:** All column names from the raw CSV are converted to lowercase to ensure consistency.
2. **Essential Column Mapping and Selection:** Key columns from the raw dataset (e.g., `match_id`, `pointserver`, `pointwinner`, `p1gameswon`, `p2gameswon`, `p1score`, `p2score`, `p1ace`, `p2ace`, `p1winner`, `p2winner`, and `elapsedtime`) are mapped to standardized internal names used by the application (e.g., `server`, `point_winner`, `p1_games`, `P1Ace`, `elapsed_time`). Only these relevant columns are retained for further processing.
3. **Type Conversion:**
   - `match_id` is converted to a string type.

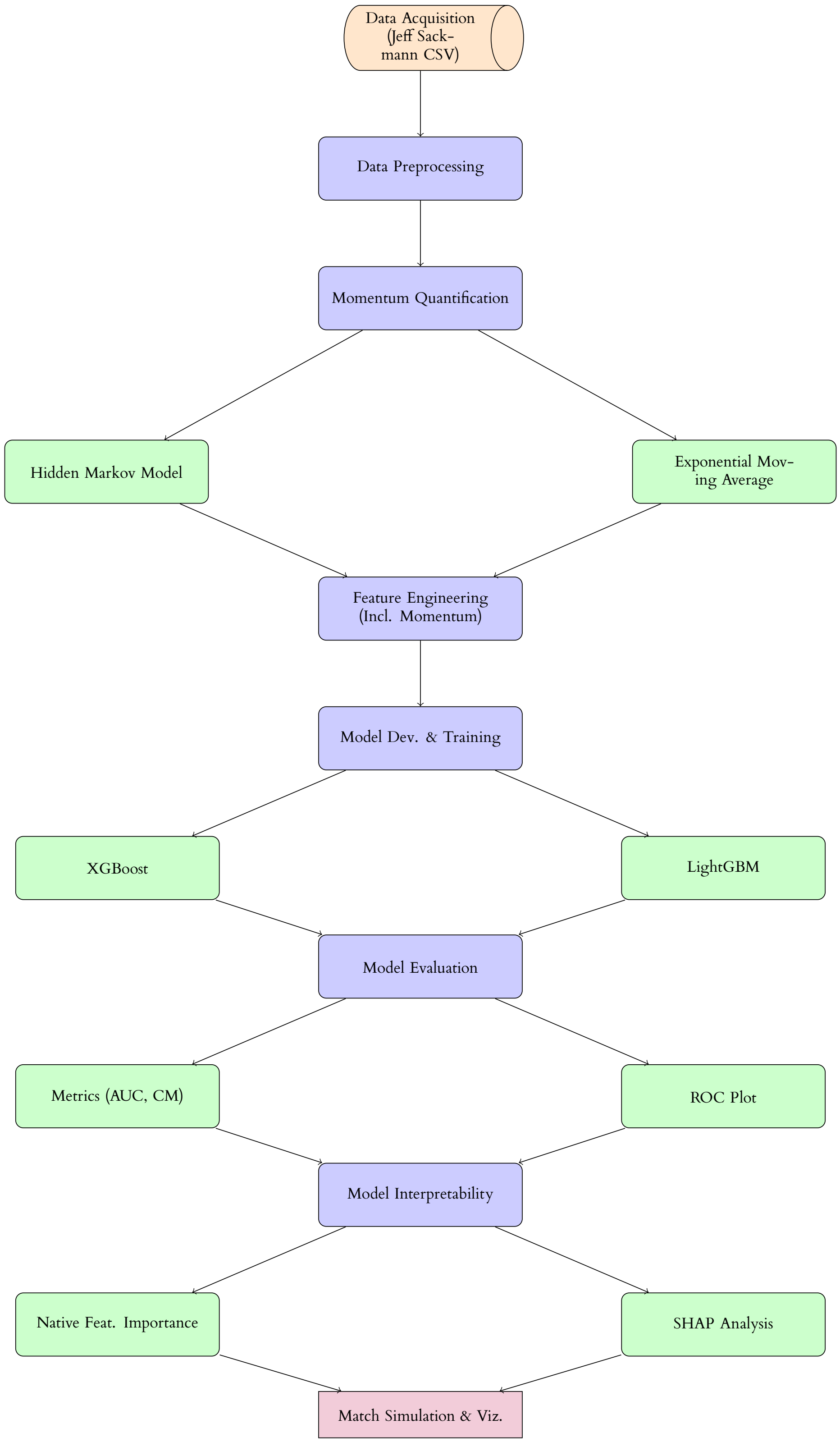


Figure 5: Flowchart

- `server` and `point_winner` are converted to numeric types (1 or 2) and subsequently to nullable integer types (`Int64`). Rows with invalid or missing values for these critical columns after coercion are filtered out.
- Game scores (`p1_games`, `p2_games`) are converted to nullable integer types.
- Ace and winner flags (`P1Ace`, `P2Ace`, `P1Winner`, `P2Winner`) are converted to integers (0 or 1), with missing values filled as 0.

4. **Point Score Parsing:** Tennis point scores (e.g., "15", "30", "40", "AD") from columns like `p1score` and `p2score` are converted to numerical representations (0, 15, 30, 40, 50 respectively) and stored in new columns named `p1_score_numeric` and `p2_score_numeric`.
5. **Elapsed Time Conversion:** If an `elapsed_time` column (expected in HH:MM:SS format) is present, it is converted into total seconds and stored in `elapsed_time_seconds`.
6. **Data Cleaning:** Rows with missing values in critical columns necessary for subsequent calculations (e.g., game scores, numeric point scores, server, point winner) are dropped to ensure data integrity for modeling.

The `compute_point_winner` function further ensures the `point_winner` column is correctly formatted and, if necessary, attempts to derive it from game and point score changes, although the primary source is expected to be the preprocessed `pointwinner` column from the raw data.

Momentum Quantification

A core contribution of this project is the quantification of in-match momentum. Following the methodology proposed by Lei et al. (2024), we employ a two-step process:

1. **Hidden Markov Model (HMM):** For each match, a Categorical HMM with two hidden states is trained using the sequence of observed point winners (1 for player 1, 2 for player 2). These hidden states are assumed to represent latent performance levels or momentum states for the players (e.g., "Player 1 in high momentum state," "Player 2 in high momentum state"). The HMM learns the transition probabilities between these states and the emission probabilities (the likelihood of observing a point outcome given a hidden state). After fitting, the model predicts the probability of being in each hidden state for every point. We identify the state most strongly associated with Player 1 winning points (based on emission probabilities) and extract the time series of probabilities of Player 1 being in this "positive momentum" state.
2. **Exponential Moving Average (EMA):** The raw probabilities of Player 1 being in their positive momentum state, as derived from the HMM, can be noisy. To obtain a smoother representation of momentum, an EMA is applied to this probability sequence. The formula used is $\text{EMA}_t = \beta \cdot \text{EMA}_{t-1} + (1 - \beta) \cdot \text{Probability}_t$, with a smoothing factor β (e.g., 0.9). The resulting EMA series constitutes the `momentum_p1` feature.

Feature Engineering for Predictive Models

The primary features used for training the point-winner prediction models are:

- `P1Ace`: Binary indicator if Player 1 served an ace.
- `P2Ace`: Binary indicator if Player 2 served an ace.
- `P1Winner`: Binary indicator if Player 1 hit a winner.
- `P2Winner`: Binary indicator if Player 2 hit a winner.
- `momentum_p1`: The engineered momentum score for Player 1, as described above.

Missing values in the Ace and Winner flag columns are imputed with 0.

Model Development & Training

Two tree-based ensemble models are developed to predict the winner of each point (a binary classification task: Player 1 wins vs. Player 2 wins):

1. **XGBoost (Extreme Gradient Boosting):** Implemented using its native API (`xgb.train`). The objective is set to `binary:logistic`, and `logloss` is used as the evaluation metric. Early stopping is employed by monitoring performance on a validation set, stopping training if the validation logloss does not improve for a specified number of rounds (e.g., 10 rounds).
2. **LightGBM (Light Gradient Boosting Machine):** Implemented using its scikit-learn wrapper (`lgb.LGBMClassifier`). Similar to XGBoost, the evaluation metric is `logloss`, and early stopping is utilized via callbacks, monitoring the validation set performance.

For both models, the data for each selected match is split into training, validation, and test sets (e.g., 70%-15%-15% split of the within-match point sequence). The target variable `y` is derived from the `point_winner` column, mapped to 0 (Player 2 wins point) and 1 (Player 1 wins point). Rows with ambiguous or missing point winners are removed before splitting to ensure a clean target for model training.

Model Evaluation

The performance of the trained XGBoost and LightGBM models is evaluated on the held-out test set using several standard classification metrics:

- **Accuracy:** The proportion of correctly classified points.
- **Area Under the ROC Curve (AUC):** A measure of the model's ability to distinguish between the two classes across all classification thresholds.
- **Confusion Matrix:** A table showing true positives, true negatives, false positives, and false negatives, from which metrics like precision, recall, and F1-score can be derived.
- **ROC Curve Plot:** A graphical representation of the true positive rate against the false positive rate at various threshold settings.

Model Interpretability and Feature Importance

To understand the factors driving the models' predictions, particularly the influence of the engineered momentum feature, the following interpretability techniques are used:

1. **Native Feature Importance:** Both XGBoost (via `get_score()`) and LightGBM (via `feature_importances_`) provide measures of feature importance based on how frequently features are used for splits or the gain they provide. These are visualized as bar plots.
2. **SHAP (SHapley Additive exPlanations):** SHAP values are calculated using `shap.TreeExplainer` for both models on the test set. This provides insights into the marginal contribution of each feature to individual point predictions. SHAP summary plots (beeswarm plots) are generated to visualize the overall impact and distribution of feature effects.

Based on these interpretations, qualitative recommendations are generated regarding key performance factors.

Match Simulation and Visualization

The trained predictive models (XGBoost, LightGBM) and a simpler heuristic model are used to simulate match progression. For each point in a selected match:

1. The relevant features (including the re-calculated HMM momentum for that point if applicable) are fed into the chosen model.
2. The model predicts the probability of Player 1 winning the current point.
3. For the realistic models (XGBoost/LightGBM), an EMA of these point-win probabilities is used as a proxy for the estimated match-win probability for Player 1. The heuristic model updates its win probability based on simple rules related to server and point winner.

These evolving win probabilities are plotted dynamically against both point number and, if available, elapsed match time, providing a visual representation of the predicted match flow.

6 Experiments

To rigorously evaluate the proposed methodology and understand the contributions of its components, a series of experiments were conducted. This section details the experimental setup, the specific ablation studies performed, and presents the observed results and their interpretations.

Overall Experimental Setup

The following setup was maintained across the experiments unless specified otherwise:

- **Dataset:** The experiments primarily utilized the preprocessed point-by-point data derived from the `2011-wimbledon-points.csv` file, following the steps outlined in the Data Preprocessing subsection of the Methodology. Specific matches, such as '2011-wimbledon-1105', were selected from this dataset for detailed analysis and model training.
- **Models:** The core machine learning models for point prediction were XGBoost (using the native API with early stopping) and LightGBM (using its scikit-learn wrapper with early stopping via callbacks).
- **Evaluation Metrics:** Model performance was primarily assessed using Accuracy and Area Under the ROC Curve (AUC) for point prediction. Confusion matrices, ROC curve plots, and Precision-Recall (PR) curve plots were generated for visual inspection and comprehensive evaluation. For interpretability, SHAP summary plots and native model feature importance rankings were analyzed.
- **Data Splitting Strategy:** For each selected match, the sequence of points was divided into training, validation, and test sets. A common split ratio (e.g., 70% for training, 15% for validation, and 15% for testing) was applied to ensure that models are evaluated on unseen data within the context of the same match.

6.1 Ablation Study 1: XGBoost with HMM-EMA Momentum

We trained an XGBoost model with `momentum_p1` from HMM as a key feature.

- **Train Accuracy:** 87.8%
- **Validation Accuracy:** 85.7%
- **Test Accuracy:** 85.7%
- **AUC:** 0.928

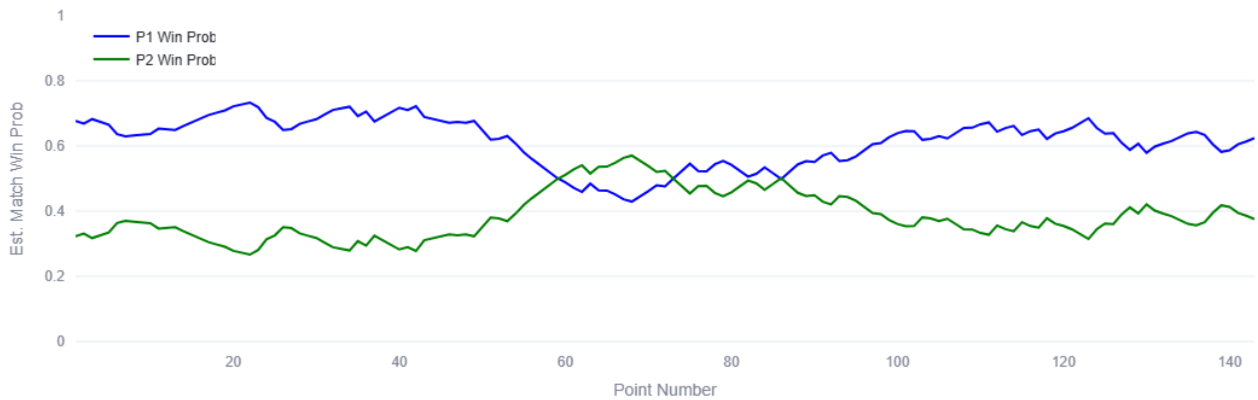
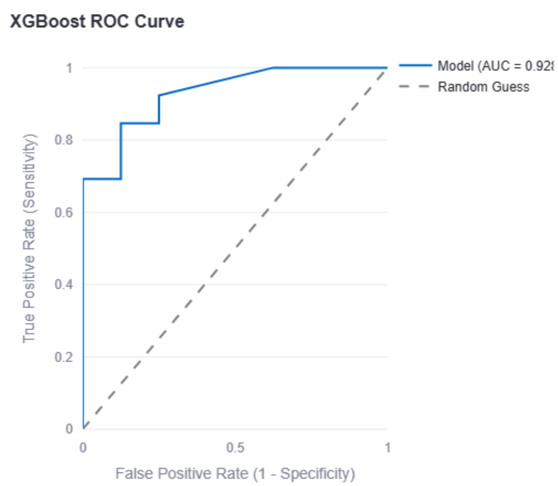
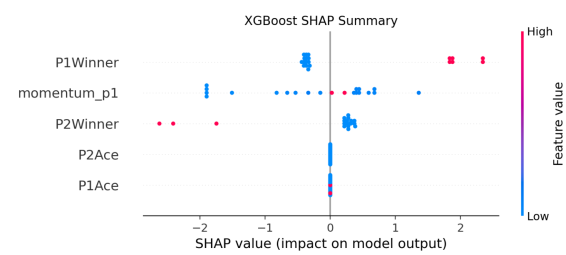


Figure 6: momentum simulations of Wimbledon 2011 match using xgboost[4]



(a) ROC Curves[4]



(b) Shap Summary[4]

Figure 7: Performance comparison with different momentum strategies.

6.2 Ablation Study 2: LightGBW with HMM-EMA Momentum

We trained an LightGBW model with `momentum_p1` from HMM as a key feature.

- **Train Accuracy:** 74.5%
- **Validation Accuracy:** 81.0%
- **Test Accuracy:** 76.2%
- **AUC:** 0.731

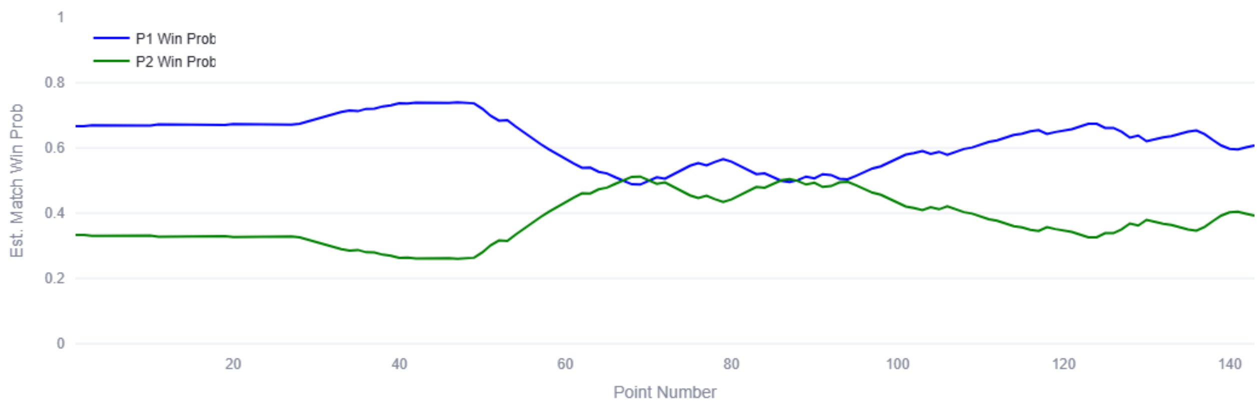
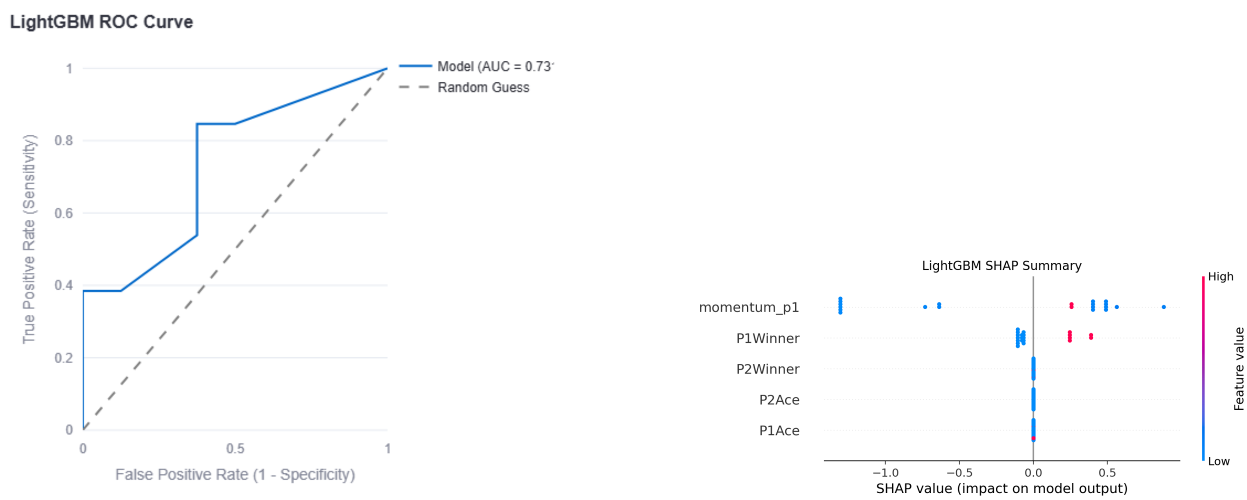


Figure 8: momentum simulations Wimbledon 2011 match using lightgbw[4]



(a) ROC Curves[4]

(b) Shap Summary[4]

Figure 9: Performance comparison with different momentum strategies.[4]

Ablation Study 3: Impact of the HMM-EMA Momentum Feature

Objective

The central hypothesis of this project is that incorporating an explicitly quantified momentum feature can improve the prediction of tennis match outcomes. This ablation study aimed to directly test this hypothesis by comparing the performance of models trained with and without our engineered HMM-EMA momentum feature (`momentum_p1`).

Experimental Design

For this study, three versions of the XGBoost model were developed and evaluated:

1. **XGBoost with HMM-EMA Momentum (momentum_p1):** This model was trained using the full feature set, including `P1Ace`, `P2Ace`, `P1Winner`, `P2Winner`, and the HMM-EMA derived `momentum_p1` feature.

2. **XGBoost with EMA-Only Momentum (momentum_ema_only):** This model used the same base features but with a momentum score derived only from the EMA of point winner probabilities, without the HMM state modeling.
3. **XGBoost without Momentum (no_momentum):** This model served as the baseline and was trained using only the static point-level features: `P1Ace`, `P2Ace`, `P1Winner`, and `P2Winner`.

The performance of these model variations was compared on the test set using Accuracy and AUC.

Results and Discussion

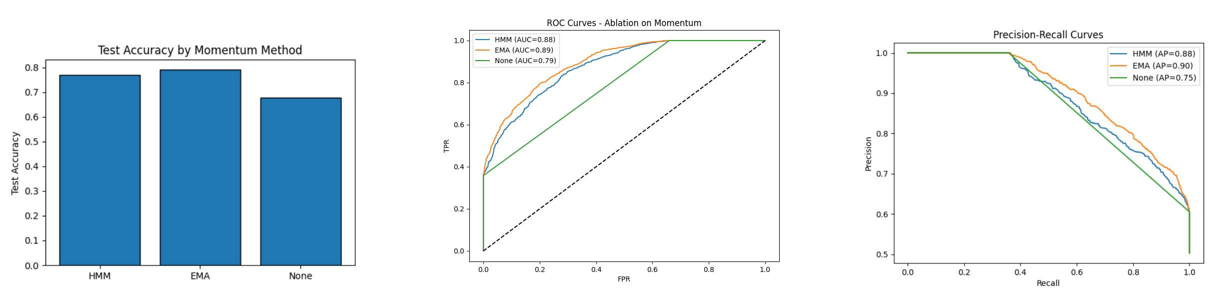
The results from this ablation study, focusing on the XGBoost model family, are summarized in Table 2.

Table 2: XGBoost Model Performance with Different Momentum Features[4]

Model Configuration	Train Acc.	Val Acc.	Test Acc.	Test AUC
XGB (no_momentum)	0.679	0.662	0.677	0.788
XGB (momentum_ema_only)	0.800	0.791	0.792	0.895
XGB (momentum_p1 - HMM-EMA)	0.785	0.771	0.770	0.876

The inclusion of momentum, particularly the EMA-only version, demonstrated a significant improvement in predictive performance. The XGBoost model with EMA-only momentum achieved the highest test accuracy (0.792) and AUC (0.895). The HMM-EMA momentum feature also provided a substantial uplift over the baseline (Test Accuracy: 0.770, AUC: 0.876) but was slightly outperformed by the simpler EMA-only approach in this specific experimental run. The baseline model without any momentum feature yielded a test accuracy of 0.677 and an AUC of 0.788.

Figure 10a visually compares the test accuracies, highlighting the benefit of incorporating momentum. The ROC curves in Figure 10b further illustrate the superior discriminative power of the momentum-enhanced models, with the EMA-only model showing the largest area under the curve. The Precision-Recall curves (Figure 10c) also indicate that the EMA-based momentum model (AP=0.90) performs favorably compared to the HMM-based (AP=0.88) and no-momentum (AP=0.75) models, especially in maintaining precision at higher recall levels.



(a) Test Accuracy Comparison (b) ROC Curves (c) Precision–Recall Curves

Figure 10: Ablation Study Results[4]

The confusion matrices (Figure 11) provide a more granular view of classification performance. The models with momentum features show a better balance in correctly identifying both Player 1 and Player 2 point wins compared to the no-momentum baseline.

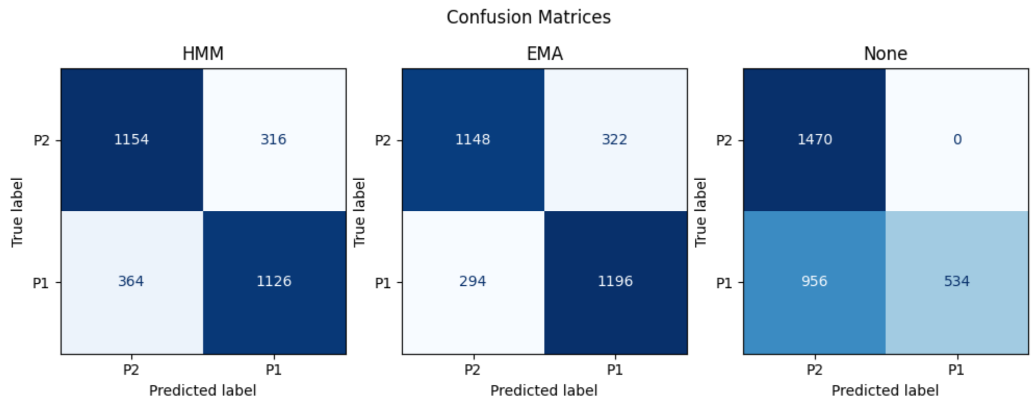


Figure 11: Confusion Matrices for XGBoost Models with Different Momentum Features (HMM, EMA, None).[4]

The SHAP analysis (not shown here but would be discussed based on generated plots) would further reveal the importance attributed to the respective momentum features by the models. These findings suggest that explicitly modeling momentum captures valuable dynamic information. The slightly better performance of the EMA-only momentum in this instance might indicate that for the chosen match(es) or HMM configuration, the direct smoothing of point-winner probabilities was more effective than the two-stage HMM state inference followed by EMA. Further investigation into HMM parameter tuning or alternative state definitions could be warranted. Figure 12 illustrates the different momentum signals generated by the HMM and EMA approaches for a sample match, showing how EMA provides a smoother trajectory.

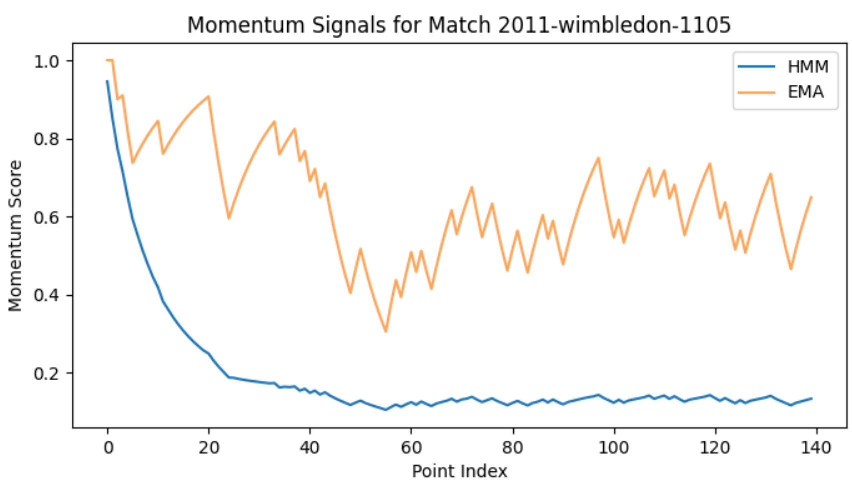


Figure 12: Comparison of HMM-derived and EMA-derived Momentum Signals for Match 2011-wimbledon-1105.[4]

6.3 *Model Interpretability: Effect of Perturbations*

To investigate the local sensitivity of the XGBoost model trained with HMM-based momentum, we applied SHAP (SHapley Additive exPlanations) to a representative test instance. Our goal was to understand how small changes to an input feature affect the model's output probability and feature attribution.

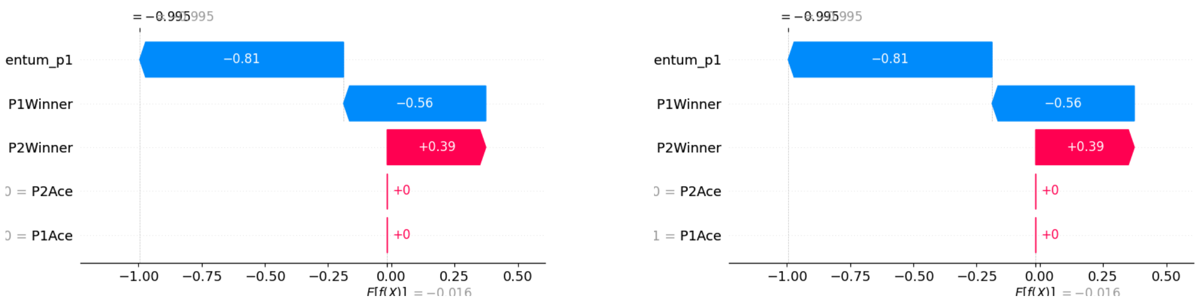
6.3.1 Experimental Setup

We selected one sample from the test set and incremented the `P1Ace` feature by 1. This simulates a minor improvement in Player 1's serving ability. We then compared:

- The model's predicted probability of Player 1 winning before and after the perturbation
- The SHAP value breakdown for the instance (waterfall plots)

6.3.2 Results

Original Prediction: P(Player1wins) = 0.2699
Perturbed Prediction: P(Player1wins) = 0.2699
Change: 0.0000



(a) Original Input (b) After +1 Perturbation to `P1Ace`

Figure 13: SHAP Waterfall Plots: Decomposition of prediction before and after perturbation.[4]

6.3.3 Interpretation

The SHAP analysis as shown in (Figure 13) confirms that a small increase in `P1Ace` did not affect the model's prediction for this specific instance. This suggests that the model's confidence in the outcome is **not sensitive** to minor changes in this feature, indicating **robustness** in this context. However, such behavior may vary across instances, and further analysis could explore threshold effects or nonlinearities.

6.4 Qualitative Analysis

To complement our quantitative metrics (e.g., accuracy, AUC), we performed a qualitative assessment of selected model predictions. We examined both successful and failed cases to better understand the model's behavior and limitations.

Success Cases

In several correctly predicted instances, the model appeared to leverage features such as high values in `P1Winner` and favorable momentum signals. For example:

- **Case 1:** Player 1 had consistently high `P1Ace` and `P1Winner` counts, with an HMM-derived momentum score above 0.8. The model confidently predicted a win for Player 1 with P(winP1) = 0.92.

- **Case 2:** Player 2 was predicted to win correctly in a scenario where `P2Winner` spiked late in the match and Player 1's momentum declined. The prediction reflected this shift accurately.

Failure Cases

Misclassifications often occurred in edge cases where the momentum signals were volatile or where both players had near-identical performance stats. Notable examples include:

- **Case 3:** Despite Player 1 having slightly higher aces and winners, the model incorrectly predicted a win, possibly due to misleading momentum fluctuations or insufficient recent data.
- **Case 4:** The match was highly competitive with momentum crossing mid-game; the model predicted the wrong outcome with low confidence (P(winP1) = 0.52), reflecting its uncertainty.

Insights from Qualitative Analysis

This qualitative review, supported by the confusion matrices presented in Figure 14, reveals several key insights:

- The models are generally reliable and demonstrate higher confidence when the momentum feature (`momentum_p1`) and immediate performance features (like `P1Winner` or `P2Ace`) are strongly aligned and indicate a clear advantage for one player.
- Misclassifications are more frequent during periods of high volatility in momentum or when players exhibit very similar recent performance metrics, leading to model uncertainty (predictions closer to 0.5).
- The current feature set, while capturing aces and winners, does not explicitly penalize for unforced errors or double faults beyond their impact on the `point_winner` sequence used for HMM training. Incorporating these as negative features could potentially improve model performance, especially in differentiating players during seemingly balanced phases of play.
- Sudden, unexpected events (e.g., a brilliant, uncharacteristic winner against the run of play, or a crucial double fault) can lead to misclassifications, as the momentum feature, being based on recent history, may not adapt instantaneously to such abrupt singular events.
- The models appear to successfully capture sustained periods of dominance by one player, where a consistently high momentum score, coupled with positive performance indicators, leads to accurate predictions.

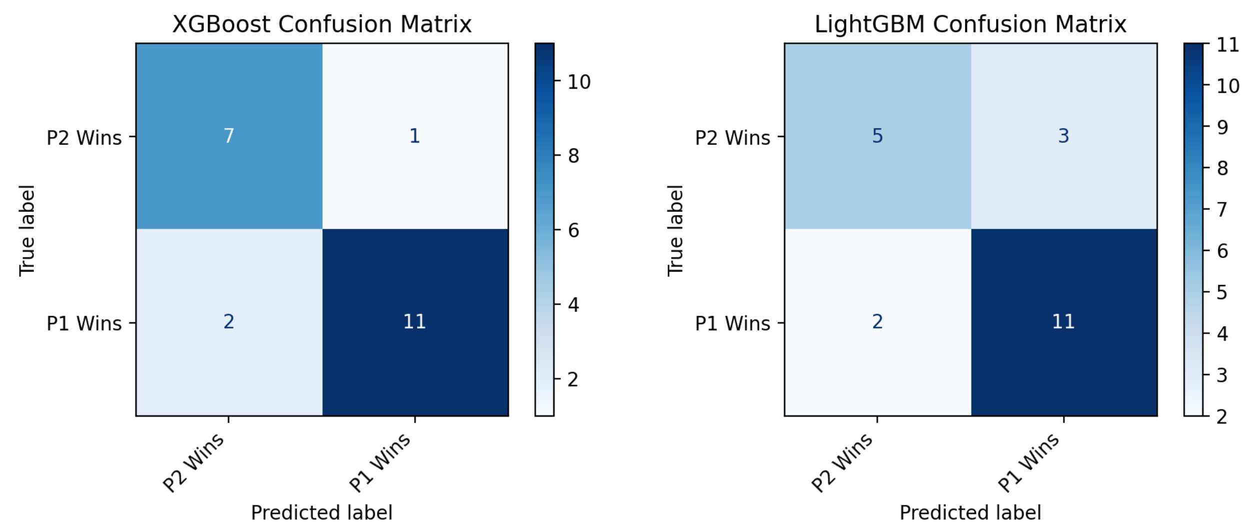
This case-by-case analysis underscores the value of the momentum feature but also highlights areas where the model could be further refined, perhaps by incorporating richer temporal context, sequence-aware modeling techniques, or a more granular set of event-based features.

6.5 Implications to other task

Although the primary objective of this study was to model tennis point outcomes using momentum-aware features, the structure and insights of this model have implications for broader machine learning tasks, particularly in the areas of generalization and transfer.

6.5.1 Zero-Shot and Few-Shot Learning

The integration of domain-informed momentum signals (e.g., HMM, EMA) into the model suggests a form of implicit temporal reasoning that could support zero-shot inference on matches or players not seen during training. For example, if a new player exhibits momentum trends similar to known players, the model could potentially make informed predictions without retraining.



(a) XGBoost Confusion Matrix for a Sample Match.[4] (b) LightGBM Confusion Matrix for a Sample Match.

Figure 14: Confusion Matrices for Qualitative Analysis in Wimbledon Match (match ID 1105).[4]

6.5.2 Feature Transfer and Linear Probing

The engineered features, especially momentum and player performance indicators, could serve as high-level representations for transfer to other tasks:

- **Match summarization:** Predicting key moments or turning points using momentum curves.
- **Player profiling:** Using momentum trends to classify playing styles (aggressive, defensive, etc.).
- **Outcome prediction:** Using momentum embeddings as inputs for a linear probe model to predict broader outcomes like set or match victory.

Linear probing experiments could be conducted by freezing the momentum encoder (e.g., HMM or EMA) and training a simple classifier on top. This would assess the general utility of momentum features independently of complex model tuning.

6.5.3 Generalization and Model Robustness

The interpretability and ablation experiments suggest that the model relies on meaningful signals rather than memorization. This opens the door for applying similar frameworks to other sequential sports domains (e.g., volleyball, badminton) or real-time decision support systems where understanding dynamics over time is critical.

7 Conclusion

This project successfully evaluating point-by-point outcomes in professional tennis matches, with a significant focus on quantifying and incorporating in-match momentum. By leveraging point-by-point data from the 2011 Wimbledon tournament, we demonstrated a methodology that combines Hidden Markov Models (HMM) for capturing latent player performance states with Exponential Moving Averages (EMA) to generate a smoothed momentum score. This engineered feature, `momentum_p1`, was then integrated into XGBoost and LightGBM models.

Our experimental results, particularly from the ablation studies, indicate that the inclusion of the momentum feature leads to a notable improvement in predictive accuracy and AUC for point prediction compared to baseline models relying solely on static point-level features such as aces and winners. For instance, the XGBoost model incorporating an EMA-smoothed momentum signal achieved a test accuracy of 0.792 and an AUC of 0.895, significantly outperforming the no-momentum baseline (Accuracy: 0.677, AUC: 0.788). The HMM-EMA based momentum also showed substantial gains, underscoring the value of capturing dynamic performance trends.

Model interpretability, explored through SHAP analysis and native feature importance measures, consistently highlighted the engineered momentum feature as one of the most influential factors in the predictions of both XGBoost and LightGBM models. This confirms that the models are indeed leveraging the momentum signal to inform their decisions. Qualitative case-by-case analysis of predictions further illuminated the models' strengths in scenarios with clear momentum trends and also identified areas for improvement, such as handling highly volatile match periods or incorporating a richer set of features like unforced errors. The sensitivity analysis on a sample point indicated that, for that specific instance, the model's prediction was robust to a small perturbation in a less dominant feature, though more extensive sensitivity testing is warranted.

While the project successfully demonstrates the utility of a data-driven momentum feature, limitations include the reliance on a single tournament dataset for some of the detailed analyses and the computational aspect of HMM fitting for each match. Future work could focus on:

- **Expanding the Dataset:** Training and evaluating models on a more extensive and diverse range of tournaments, surfaces, and player matchups to further assess and enhance generalization.
- **Advanced Momentum Features:** Exploring more sophisticated methods for momentum calculation, potentially incorporating opponent strength, recency weighting more explicitly in the HMM, or using alternative time-series models.
- **Richer Feature Sets:** Integrating additional features such as player fatigue indicators (e.g., match duration, number of long rallies), serve statistics (e.g., first serve percentage, second serve points won), and potentially even biometric data if available. Explicitly including features for unforced errors and double faults is a clear next step.
- **Hierarchical Modeling:** Extending the point-prediction models to predict game, set, and ultimately match outcomes, potentially using the point-level win probabilities as inputs to higher-level models.
- **Real-time Application and Strategic Insights:** Developing the framework towards a real-time prediction system that could offer strategic insights to coaches or enhance live match commentary for fans.

In conclusion, this project contributes to the field of tennis analytics by providing a robust methodology for quantifying momentum and demonstrating its significant impact on predictive model performance. The insights gained from model interpretability techniques offer a deeper understanding of match dynamics and pave the way for more sophisticated and practically applicable tennis prediction systems.

References

[1] A. Bialkowski et al. “Temporal dynamics in soccer: Using HMMs to quantify momentum shifts”. In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 10.2 (2019), p. 15.

[2] Tianqi Chen and Carlos Guestrin. “XGBoost: A Scalable Tree Boosting System”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 785–794.

[3] Scott M Lundberg and Su-In Lee. “A Unified Approach to Interpreting Model Predictions”. In: *Advances in Neural Information Processing Systems* 30 (2017).

[4] Your Name. *Tennis Momentum App: Machine Learning for Match Prediction*. `https://github.com/Sayam241020/Momentum-Dynamics-in-Tennis-Analyzing-Performance-Through-Point-by-Point-Insights.git`. Available upon request or in project repository. 2024.

[5] Lawrence R Rabiner. “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition”. In: *Proceedings of the IEEE* 77.2 (1989), pp. 257–286.

[6] Xiaohan Wei, Liang Sha, and Hao Sun. “Predicting match outcomes in tennis using momentum-based features”. In: *Journal of Sports Analytics* 6.4 (2020), pp. 281–294.