

### Summary

Tennis is an incredibly thrilling competitive sport. In the recently concluded 2023 Wimbledon Men's Singles tournament, young sensation Carlos Alcaraz defeated Novak Djokovic to claim the championship title. Players' performances in the match can fluctuate due to various factors, and we define this ebb and flow as player momentum.

In task 1, we focus on tracking the match flow and differentiate the player's performance. Using three features extracted from the original data and the factor of server, **we employ K-means clustering along with computing the distance of each point to its corresponding centroid to measure the significance of each point**, thereby elucidating the flow of the match. **By evaluating the prevailing advantage based on player's winning rate**, we identify the concept of "momentum", thus discern which competitor is demonstrating superior performance. **The lack of directionality in distances within KMeans naturally eliminates the impact of serving**, as it is considered an external factor that does not inherently reflect the internal dynamics of momentum.

In task 2, our research endeavors to ascertain whether this intangible quality of momentum exerts a tangible influence on the whole match. We generate a trend graph that plots the discrepancy in score against the fluctuations in momentum throughout the duration of the match. The concordance observed in the patterns of these variables suggests a close correlation between the two. To delve deeper into the impact of momentum on the ultimate winner, **we employ a multiple linear regression model**, which confirms the presence of a significant effect.

In task 3, we initiate a predictive model to estimate the game swings, i.e. the differential in momentum. **We utilize the ARIMA model to segment the predictions according to whether the momentum trajectory exhibits positive or negative trends**. Remarkably, the predictions generated by this model demonstrate a high degree of accuracy. Through calculating the correlation between these game swings and various statistical features, we determine that faults emerge as the most influential factor affecting the swings. The analysis examines changes in momentum across matches to uncover strategic patterns, using Novak Djokovic as a case study to illustrate how insights can inform adjustments and improve athletic performance, forming the basis for targeted advice.

In task 4, We applied our method to more matches and analyzed the reasons leading to significant discrepancies between some results and reality. **We believe that players' playing strategies, weather and court conditions also affect player momentum**. We also modeled the Wimbledon finals in 2014 and 2015 and obtained similar conclusions.

Finally, starting from the role of player momentum in the match, we provided coaches with a reference guide. This included how to quantify and visualize player momentum during the match, how to identify key turning points in the match's rhythm, and strategies for players to respond accordingly.

**Keywords:** Player Momentum; K-Means; Linear Regression; ARIMA;

# Contents

<b>1</b>	<b>Introduction</b>	<b>4</b>
1.1	Backgrounds . . . . .	4
1.2	Restatement of the Problem . . . . .	4
1.3	Our Approaches . . . . .	4
<b>2</b>	<b>Assumptions and Notations</b>	<b>5</b>
2.1	Assumptions . . . . .	5
2.2	Notations . . . . .	5
<b>3</b>	<b>Task 1: The Momentum and Flow of Match</b>	<b>6</b>
3.1	Data Cleaning . . . . .	6
3.2	Feature Engineering for Each Point . . . . .	6
3.3	Importance of Points: Capture the Flow of Play . . . . .	6
3.4	Player Momentum: Which Player is Performing Better . . . . .	9
3.5	Dealing with the Impact of the Server Factor . . . . .	9
<b>4</b>	<b>Task 2: Validate the Effectiveness of Player Momentum</b>	<b>10</b>
4.1	Correlation Between Player Momentum and Scores Differences . . . . .	10
4.2	Fit LR model to Player Momentum and Outcomes of Each Set or Game . . . . .	11
<b>5</b>	<b>Task 3: Predict <i>Swing</i> - the Change in Player Momentum</b>	<b>12</b>
5.1	ARIMA . . . . .	12
5.1.1	Stationarity Check . . . . .	12
5.1.2	Determine $p, q$ . . . . .	12
5.1.3	Rolling Predictions . . . . .	13
5.2	Problem Restatement . . . . .	14
5.3	Predictions . . . . .	14
5.4	Factors Related and Sensitivity Analysis . . . . .	14
5.4.1	Critical Points (CP) and Swings . . . . .	14

5.4.2	Faults (F) and Swings . . . . .	15
5.4.3	Physical Exertion (S) and Swings . . . . .	15
5.4.4	Conclusion . . . . .	16
5.5	Advice . . . . .	16
<b>6</b>	<b>Task 4: Model Evaluation and Generalization</b>	<b>17</b>
6.1	Metrics of ARIMA for $sw_i$ prediction . . . . .	17
6.2	Partial Poor Performance and Reasons . . . . .	19
6.3	Model Generalization . . . . .	19
6.4	Memo for Coaches . . . . .	19
<b>7</b>	<b>Strengths and weaknesses</b>	<b>20</b>
7.1	Strengths . . . . .	20
7.2	Weakness . . . . .	20
	<b>Appendices</b>	<b>21</b>
	<b>Appendix A Feature extraction</b>	<b>21</b>
	<b>Appendix B Memo for Coaches</b>	<b>23</b>

# 1 Introduction

## 1.1 Backgrounds

Tennis is a racket sport played by individuals or pairs on a rectangular court divided by a net. The objective is to score points by striking a ball with a racket so that it lands within the opponent's court and cannot be legally returned. Each player has two services in a row, alternating between serving and receiving, and the first player to win four points in a set, with at least two points more than the opponent, wins the set. In case of a tie, a tie-break is played to determine the winner. The game's design fosters an environment that promotes cerebral and physical dexterity, demanding a harmonious blend of athleticism and psychological acumen. The dataset for this paper is derived from the men's singles matches played after the first two rounds of the 2023 Wimbledon Championship. This prestigious tennis tournament, held annually in London, England, is renowned for its grass courts and rich traditions. Analyzing the outcomes of these matches will provide insights into "momentum" and other factors influencing success or failure in the sport. The data will be used to identify patterns and trends, contributing to a deeper understanding of tennis competition and player performance.

## 1.2 Restatement of the Problem

Considering the background information and related conditions given in the title, we need to solve the following problems:

- Develop a model to track tennis match dynamics, differentiate player performance, and visualize match flow, considering serve advantage.
- Test the coach's claim about the randomness of "momentum" swings using the model to analyze tennis match patterns.
- First, create a predictive model to forecast match swings, identifying relevant factors for predicting shifts in match control. Then, provide strategic advice to a player based on historical match momentum swings for a new opponent.
- Assess the model's accuracy in predicting match swings across different matches, surfaces, and sports, identifying necessary improvements for broader applicability.

## 1.3 Our Approaches

In summary, the whole modeling process can be shown as in Figure 1: First, we extract the features of the data and cluster the different points accordingly. Then, we calculate "momentum" based on the results above. We then conduct a linear regression to test the correlation between "momentum" and the flow of the match, as well as the final result, and therefore evaluate the effectiveness of our model. Next, we differentiate the obtained "momentum" to derive the "swing" for each point and employ the ARIMA model to predict the trend of the "swing" value, thus forecasting the shifts of situation in the match. Subsequently, we attempt to apply the model under different contexts to test its generalizability. Finally, we explain the role of momentum to coaches and provide strategic recommendations for the match accordingly.

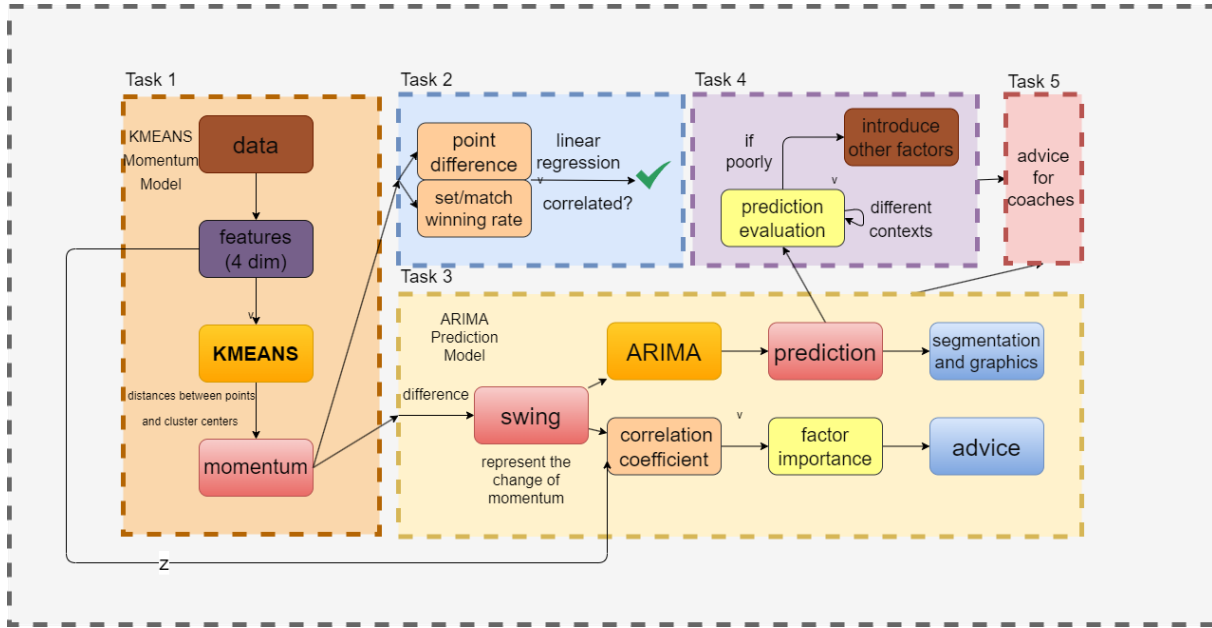


Figure 1: Roadmap

## 2 Assumptions and Notations

### 2.1 Assumptions

1. Key points in the match are few, with the majority of points being relatively ordinary.
2. Key points in the match are unique and can be considered as 'outliers' compared to most ordinary points.
3. If the extracted features are sufficiently discriminative, then the feature vectors of these key points are also outliers.

### 2.2 Notations

Notation	Description	Data Type
$Match_i$	the $i^{th}$ match in chronological order	object
$Final$	the final match, or $Match_{31}$	object
$n_i$	number of points in $Match_i$	int
$p_j$	the $j^{th}$ point in $Match_i$	1d-array
$CP_i$	critical points in $p_i$	float
$F_i$	faults in $p_i$	float
$S_i$	stamina in $p_i$	float
$Server_i$	the server in $p_i$	bool
$V_i$	the feature vector of $p_i$	$\mathbb{R}^4$
$Im_j$	the importance of $p_j$ in $Match_i$	float
$m_i$	the "momentum" of player at $p_i$	float

Table 1: Notations used to describe match-related data

### 3 Task 1: The Momentum and Flow of Match

#### 3.1 Data Cleaning

Firstly, for the sake of comparative analysis, we have made the following adjustments to the raw data: Non-numeric entries are converted to integers or floats for ease of feature engineering. For example, the score "px\_score," originally represented as "0" (love), "15," "30," "40," and "AD," are uniformly adjusted to "0," "1," "2," "3," and "4" respectively. Similarly, the server depths "server\_depth," "CTL," and "NCTL" are mapped to "0" and "1" accordingly. Simultaneously, certain missing values are standardized to either "0" or the median. For instance, missing values in "speed\_mph" are replaced with the median, while missing values in "return\_depth" are set to "0" (Deep).

#### 3.2 Feature Engineering for Each Point

Subsequently, we performed feature extraction for each point. Given a point  $p_i \in Match_i$ , the  $p_i$ -specific feature vector  $V_i$  can be represented as:

$$f_i : Match_i \rightarrow \mathbb{R}^4, \quad p_j \mapsto V_j := (CP_j, F_j, S_j, Server_j), \quad i = 1, 2, \dots, 31, \quad j = 1, 2, \dots, n_i.$$

- $CP$ , or critical points, represents the performance of both players in key scoring ability over a certain period. It is a weighted sum of metrics such as break points, saved break points, aces, and untouchable shots.
- $F$ , or faults, reflects the error rates of both players over a certain period. It is a weighted sum of indicators such as double faults, unforced errors, and net approaches.
- $S$ , or stamina, indicates the physical exertion of both players over a certain period. It is calculated from factors such as total distance covered, average distance covered per rally, average distance covered per point, return speed and return depth.
- $Server$  is a boolean variable specifically used to denote the serving side.

Detailed explanation and calculation methods for these four variables can be found in Appendix A. It is worth noting that the first three dimensions of  $V_i$  are not biased towards any particular player. The features only measures the relevant aspects of this particular point in this match. In the next section, we will compute features tailored to different players through alternative methods.

#### 3.3 Importance of Points: Capture the Flow of Play

We hold that the "flow of play" here refers to the criticality of the match progression. Although every point holds significance for the final outcome, only certain points serve as pivotal moments in the match. We attempt to capture these turning points or key points in the game. **Here, we employ**

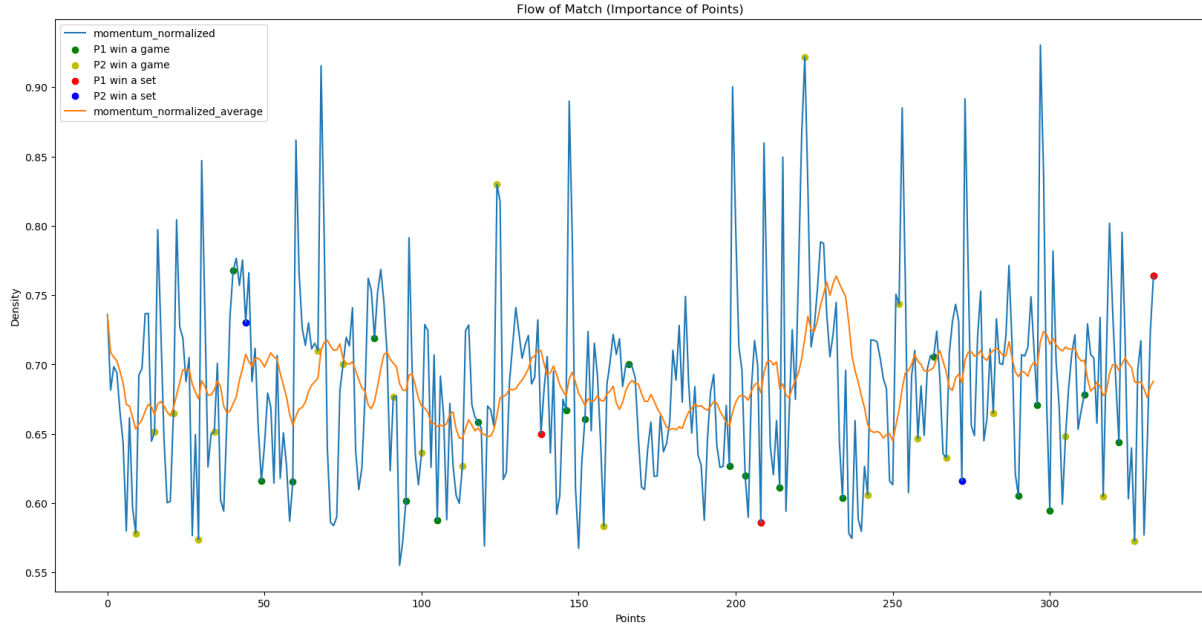


Figure 2: The Flow of Wimbledon 2023 Men's Final

**K-means clustering along with computing the distance of each point to its corresponding centroid to measure  $Im_i$  – the significance of point  $p_i$  in the match.**

K-means is an unsupervised learning algorithm that efficiently clusters data points with similar features together. It offers high flexibility, accommodating various types of features (binary, categorical, numerical, etc.), and can utilize any distance metric.

According to Assumption 2 & 3, key turning points in the match are few and outliers, as well as their feature vectors  $V_i$ . The natural clustering property of K-means will result in greater distances between outliers  $V_i$  and the centroid if there are few clusters. Thus, we can obtain a quantitative measure of  $Im_i$  by computing the distances from the k-means classification to the centroid points. The process can be summarized as follows[AI: 1]:

1. Start with  $k$  random center points (centroids)  $c_1, c_2, \dots, c_k$ . Without loss of generality, we can take  $V_1, V_2, \dots, V_k$ .
2. Group data points based on the closest centroid, forming  $k$  clusters.

$$C(i) = \arg \min_j d(V_i, c_j)$$

3. Update centroids by averaging the points in each cluster.

$$c_j = \frac{1}{|C_j|} \sum_{i \in C_j} V_i$$

4. Repeat steps 2 and 3 until the centroids do not change significantly or a maximum number of iterations is reached.

$$\text{while } \|c_j^{(t)} - c_j^{(t-1)}\| \geq \epsilon \text{ for any } j \quad \text{or} \quad \text{iters} \leq \text{max iters}$$

5. Measure  $Im_i$  by the distance from the point to its centroid and normalized.

$$Im_i = \sigma(d(V_i, c_{C(i)})), \quad \sigma(x) = 1/(1 + e^{-x})$$

Symbol	Description
$C(i)$	The cluster assigned to $V_i$ , the feature vector of point $p_i$
$d(V_i, c_j)$	The distance between $V_i$ and centroid $c_j$
$c_j^{(t)}$	The position of the j-th centroid at iteration t
$  \cdot  $	The Euclidean norm (distance)
$\epsilon$	A small positive number (the stopping criterion)
$ C_j $	The number of points in cluster j

Table 2: Notations used in K-means algorithm

Another issue here is how to determine  $k$ , the appropriate number of clusters. We utilized the elbow method to find the optimal number of clusters in k-means. It involves running k-means clustering on the dataset for a range of values of  $k$  (e.g., from 1 to 10), and for each value of  $k$ , calculate the sum of squared distances from each point to its assigned center (also known as within-cluster sum of squares, or WCSS). When you plot WCSS against the number of clusters, the curve will typically decrease towards zero in a steep slope and then flatten out, creating an "elbow"-like shape. The "elbow" (the point of inflection on the curve) is a good indication that the underlying model fits best at that point.

Taking the 2023 Wimbledon *Final* as an example. In Figure 4,  $k = 3$  is the "elbow" of this WCSS curve, and thus the optimal number of clusters for *Final* is 3.

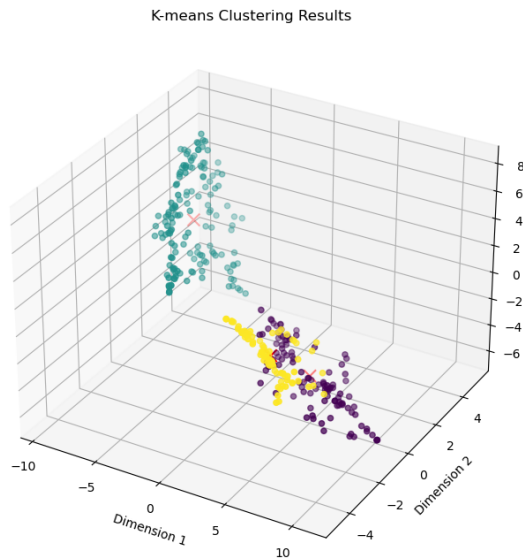


Figure 3: An Example of K-Means

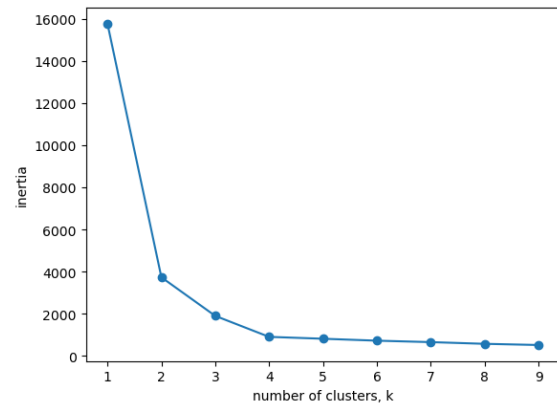


Figure 4: WCSS of the  $V_i$  in *Final*



### 3.4 Player Momentum: Which Player is Performing Better

Previously, we obtained  $Im_i$ , the significance of  $p_i$ , by calculating the distance between data points and cluster centers after K-means clustering. This  $Im_i$  value characterizes the degree of “particularity” of each point compared to the average, and thus represents how critical the point is in the match and how much it affects the match result. Now, we need to measure whether this “particularity” is advantageous or disadvantageous for a given player, therefore depict his “player momentum” during the match.

The basic idea is that for a certain point, we measure the winning rate of the player in a few points preceding and following it, and compare it with the player’s average winning rate throughout the entire match. If the player’s winning rate around this point is significantly higher than the average, then the “particularity” of this point is advantageous to the player. Otherwise, we consider it disadvantageous. So  $m_i$  can be represented as:

$$m_i = \begin{cases} Im_i, & (W_i > W - \delta) \\ -Im_i, & (W_i < W - \delta) \end{cases}$$

$W_i$  stands for the winning percentage of the last five points from the same server, and  $W$  stands for the winning percentage of points featuring the same server throughout an entire game.  $\delta$  is the threshold value.

In order to eliminate the impact of service, we only consider points with the same server when we are calculating the winning percentage. If this point’s “particularity” is advantageous for the player, then the momentum is taken as positive. Otherwise, it is negative. After normalization, we get the “player momentum” for a certain player. The higher the momentum value is, the more it indicates that, compared to the equilibrium, the situation on the court is more advantageous for that certain player. Taking the 2023 Wimbledon *Final* as an example, we provide  $m_i$ , the “player momentum” value for Carlos Alcaraz, the final champion.

In Figure 5, the blue curve represents Alcaraz’s “player momentum”, while the yellow curve denotes the trajectory of the blue curve. The horizontal line at 0.5 on the graph serves as the baseline, with values above indicating an advantage and those below indicating a disadvantage. The match begins with Djokovic dominating in the first set, reflected in Alcaraz’s consistently low momentum. Alcaraz then improves in the second set, with momentum fluctuating around the baseline, leading to a narrow victory. In the third set, Alcaraz gains the upper hand, as indicated by consistently high momentum, resulting in an easy win. However, Djokovic stages a comeback in the fourth set, capitalizing on Alcaraz’s declining momentum to secure an advantage. In the decisive fifth set, Alcaraz’s momentum rebounds, culminating in consecutive victories and ultimately clinching the championship.

The analysis demonstrates a strong correspondence between the computed momentum values for an individual player and the actual conditions of the match, demonstrating a high explanatory power.

### 3.5 Dealing with the Impact of the Server Factor

In tennis matches, the player who serves the ball often has a higher likelihood of winning the point. Therefore, when measuring players’ performances, we should remove the impact of service, involving the following two aspects: Firstly, we employ the K-Means clustering to calculate the “momentum”. One of the features used by K-Means for clustering is which player served the

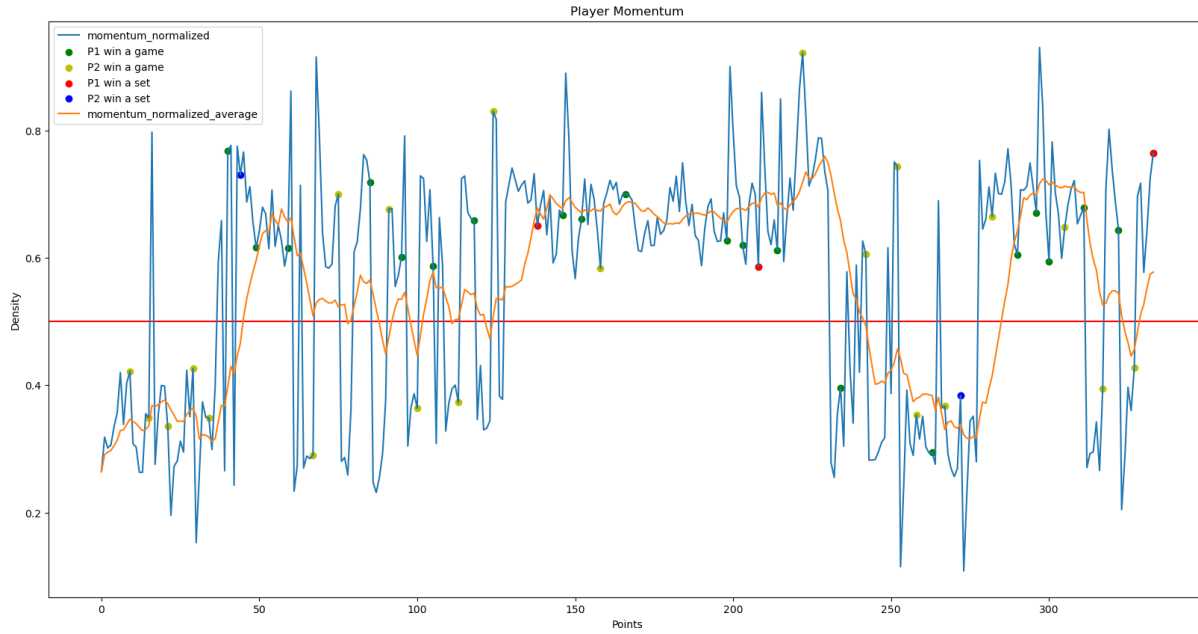


Figure 5: Player Momentum  $m_i$  in *Final*

ball. Consequently, points scored by different servers are allocated to different clusters, while those points with the same server are grouped in the same cluster. When calculating the distance of data points from the center of the cluster, since every point in a cluster has the same server, the results are independent of who the server was. Thus, the momentum we obtain merely measures the difference in performance between two players after removing the impact of service. Secondly, as mentioned, we also only consider scores with the same server while calculating “player momentum”, thus eliminate the effect of service. It can be observed that the momentum obtained for an individual player corresponds well with the actual conditions of the match, demonstrating high explanatory power.

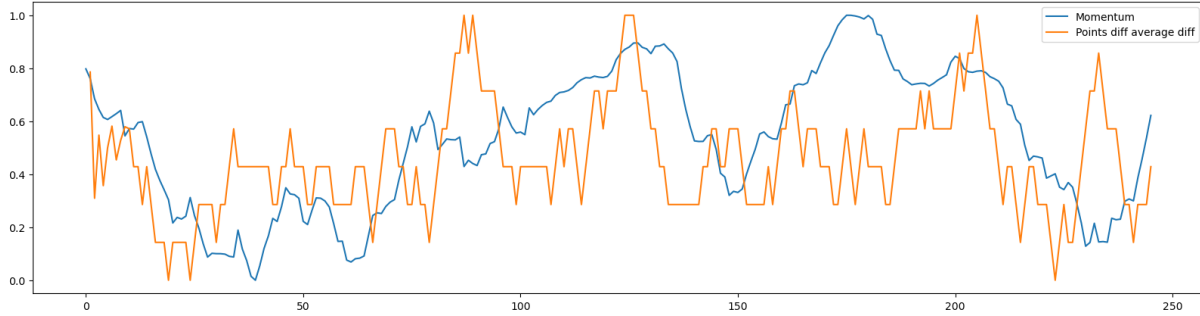
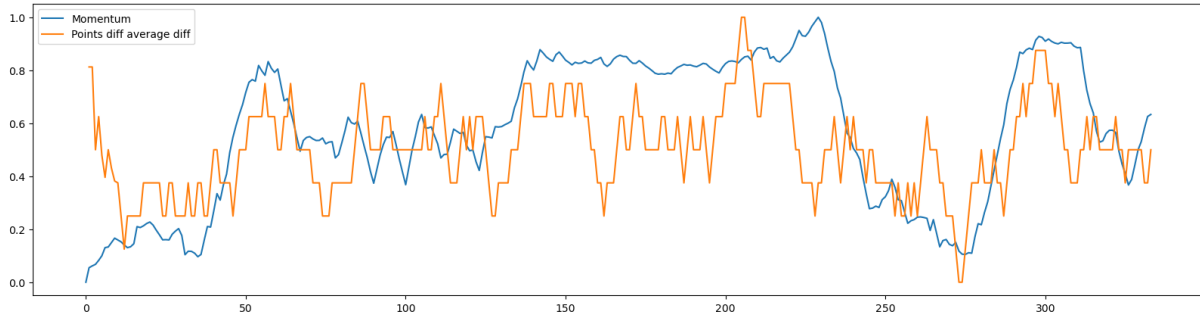
## 4 Task 2: Validate the Effectiveness of Player Momentum

Player momentum is a key representation of the match and greatly influences its trajectory. The progression of the match and the final outcome are not random. We will validate the correlation between player momentum and the match progression, and establish a linear regression model to predict match outcomes based on momentum. This will verify the effectiveness of the player momentum indicator.

### 4.1 Correlation Between Player Momentum and Scores Differences

One of the most direct ways to reflect the progression of the match is to observe changes in the score difference. We will use the data from the final match  $Match_{31}$  and  $Match_5$  as references. In Table 3, the high Pearson and Spearman correlation coefficients suggest a strong positive relationship between player momentum  $m_i$  and scores differences. The Figure 6 also visually reflects this intuitively.

Match	Pearson Correlation Coefficient	Spearman Correlation Coefficient
31	0.574	0.552
5	0.443	0.451

Table 3: Correlation Coefficients Between  $m_i$  and Scores Differences in Matches 31 and 5Figure 6: Visualized Comparison in  $Match_5$ Figure 7: Visualized Comparison in  $Match_{31}$ 

## 4.2 Fit LR model to Player Momentum and Outcomes of Each Set or Game

We used a linear regression model to fit and predict match results based on momentum. Specifically, we divided the momentum into different levels, including at the game and set levels. We categorized momentum into 10 histograms based on magnitude and counted the number of occurrences in each histogram. This allowed for dimensionality reduction of momentum, making the linear regression model representable as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_{10} x_{10} + \epsilon$$

And the predictions prove the effectiveness of player momentum  $m_i$  again, with metrics in Table 4

Level	Pearson Correlation Coefficient	Spearman Correlation Coefficient
set	0.315	0.298
game	0.443	0.451

Table 4: Correlation Coefficients Between  $m_i$  and Predicted Outcomes in Match 31

## 5 Task 3: Predict *Swing* - the Change in Player Momentum

In the preceding discussion, we have calculated "importance of points"  $Im_i$  and "player momentums"  $m_i$  as the indicators capable of grasping the flow of play. Coaches and players, however, are often more concerned with **predicting changes** in the trajectory of the game, as this enables them to adjust their tactics properly and formulate targeted responses. In this section, we aim to explore indicators that represent changes in the flow of play and utilize the ARIMA model to predict their future trends.

### 5.1 ARIMA

As previously mentioned, the "flow of play" in this context corresponds to the previously derived "player's momentum" indicator, which reflects the dynamics of advantage and disadvantage on the field. Considering that we are aiming at predicting changes in the flow of the match, we introduce the one order difference of "player's momentum", termed as "swing"  $sw_i$ .

$$sw_i = m_i - m_{i-1}, \quad i = 1, 2, \dots, n$$

$sw_i$  is a typical time-series, suitable for analysis using an ARIMA model.

#### 5.1.1 Stationarity Check

Stationarity is an important characteristic of time series data. A time series is said to be stationary if its statistical properties such as mean, variance, autocorrelation, etc. are all constant over time. Most of the time series models work on the assumption that the time series is stationary. Augmented Dickey-Fuller (ADF) test is a commonly used method for testing the stationarity of time series data. We still chose the data from the final match, and results are as follows in the Table 5. The p-value is sufficiently small, which means  $sw_i$  is stationary and does not require differencing.

Metric	Value
ADF Statistic	-5.317653
p-value	0.000005
Critical Value (1%)	-3.451
Critical Value (5%)	-2.871
Critical Value (10%)	-2.572

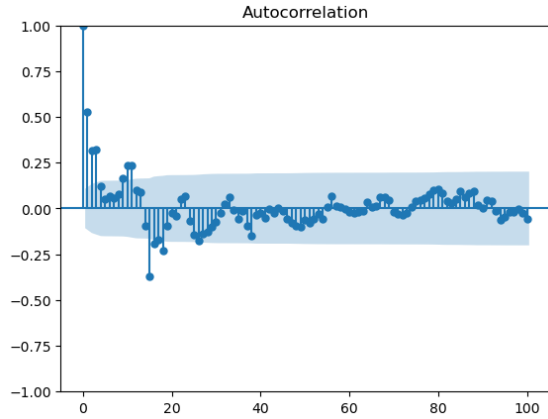
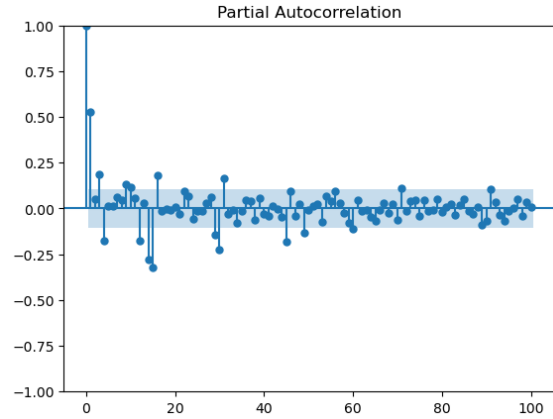
Table 5: ADF Test Results

#### 5.1.2 Determine $p, q$

The Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots are useful tools to identify the order of the Autoregressive (AR) or Moving Average (MA) terms in an ARIMA model.[AI: 2]

- Identifying the order of the AR term ( $p$ ): The PACF plot is used to identify the order of the AR term. The PACF plot displays the correlation between a series and its lags. The point where the PACF cuts off (drops to 0) for the first time gives the order of the AR term.

- Identifying the order of the MA term ( $q$ ): The ACF plot is used to identify the order of the MA term. The ACF plot displays the autocorrelation between a series and its lags. The point where the ACF cuts off (drops to 0) for the first time gives the order of the MA term.

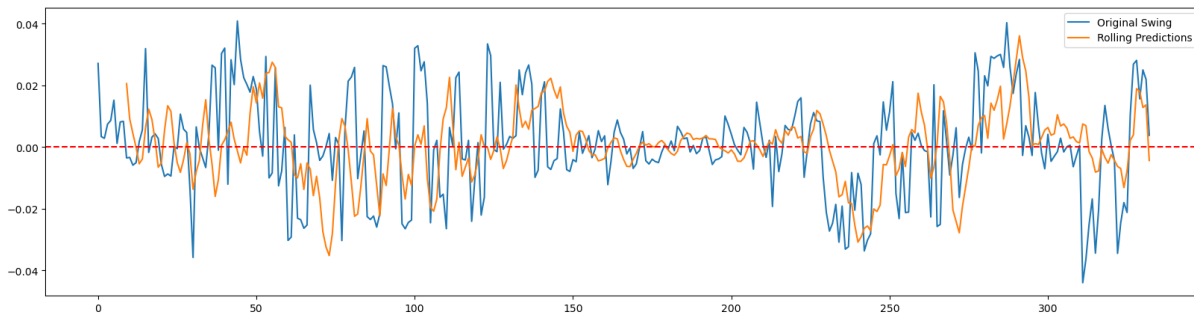
Figure 8: ACF for  $sw_i$  in *Final*Figure 9: PACF for  $sw_i$  in *Final*

Based on the features exhibited in Figure 8 & 9, we set  $p = 16, q = 4$ .

### 5.1.3 Rolling Predictions

To obtain more accurate forecasting results, we adopt a rolling prediction approach. This allows the model to adapt to changes in data over time and simulates how the model would be used in real tennis matches. This can be summarized as steps below and Figure 10:

1. Train Model:  $\text{Model} = f(X_{1:n})$
2. Predict:  $\hat{x}_{n+1} = \text{Model}(X_{1:n})$
3. Update Model:  $\text{Model} = f(X_{2:n+1})$ , then predict  $\hat{x}_{n+2}$ , and so on.

Figure 10: predictions by rolling ARIMA model for  $sw_i$  in *Final*

## 5.2 Problem Restatement

For a given player, a more positive “swing” indicates a faster increase in his “player’s momentum”, and conversely, a more negative “swing” indicates a faster decrease. What Players and coaches should pay attention to are **the zero points of  $sw_i$** , as they represent the critical turning points of the trend in the “players momentum”.

## 5.3 Predictions

We have already obtained the prediction of the player’s swing during the match, which refers to the one order difference of the “player’s momentum”. To display the prediction results distinctly, we have segmented the swing based on the prediction outcomes. The swing has been segmented based on the positive or negative characteristics of the curve. We have calculated the average swing values along the curve. if the mean value of a segment exceed a given threshold, we deem the swing values in the vicinity to be more positive, indicating that the athlete’s momentum is on the rise, and the current tactic is effective. Conversely, if the average value falls below another negative threshold, it suggests that the swing at this point is more inclined to be negative, signifying the athlete’s performance is sliding, and adjustments should be made promptly. If the swing consistently fluctuate around the baseline, it implies that the situation on the court remains unchanged, and the strategy may be considered based on the situation. In the Figure 11, we present the swing curve of Alcaraz in the 2023 Wimbledon Final and represent the segments of momentum rising, maintaining, and declining with red, green, and blue, respectively.

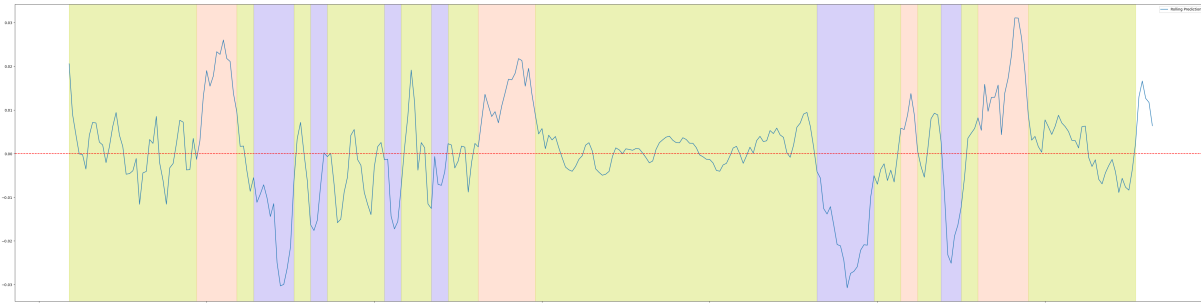


Figure 11: This is the swing curve of Alcaraz(Player 1) in the 2023 Wimbledon Final. The red, green and blue bolock here represent the segments of momentum rising, maintaining, and declining respectively.

## 5.4 Factors Related and Sensitivity Analysis

The analysis of the relationship between the variables critical points (CP), faults (F), and physical exertion (S) and the swings in tennis matches reveals several insights. Swings, representing the fluctuations in a player’s state over a short period, with positive values indicating an upward trend and negative values indicating a downward trend, are influenced by various factors.

### 5.4.1 Critical Points (CP) and Swings

The RMSE of 0.8069 for CP indicates a relatively high level of uncertainty in predicting swings based on critical points. The MAE of 0.6517 reflects a moderate average error in these predictions.

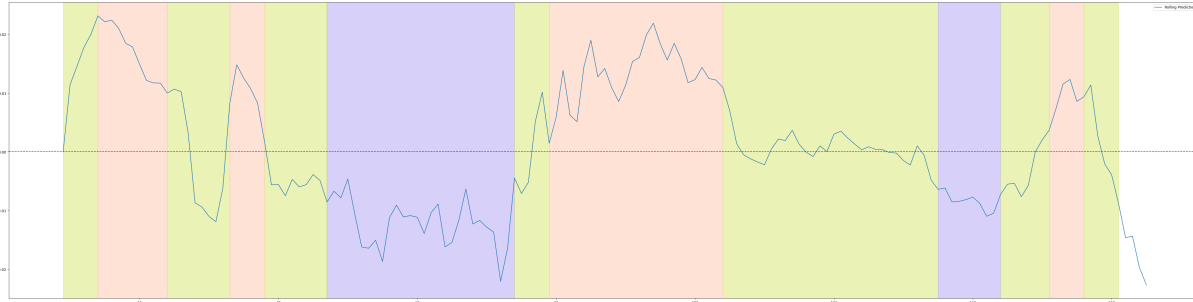


Figure 12: This is the swing curve of Player 1 in Match 16

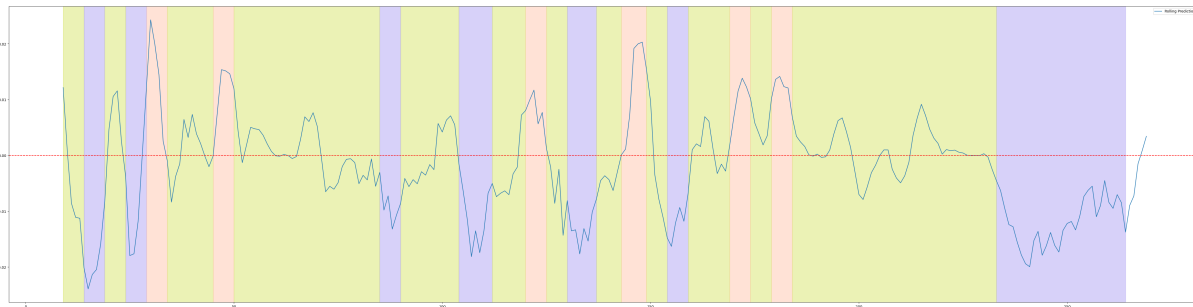


Figure 13: This is the swing curve of Player 1 in Match 24

The high MAPE of 0.9453 suggests a significant prediction bias, indicating that critical points may not accurately reflect the overall state fluctuations of a player. The R-square of 0.0747 indicates that critical points can only explain a small portion (7.47%) of the variance. These findings suggest that while critical points are important in evaluating a player's performance in key scoring situations, they may not be the primary indicator of state fluctuations. The correlation coefficients indicate a weak association between critical points and swings, indicating that other factors, such as tactical decisions and mental resilience, may have a more significant influence on a player's state swings.

#### 5.4.2 Faults (F) and Swings

The RMSE of 0.6027 for F indicates a relatively lower level of uncertainty in predicting swings based on faults. The MAE of 0.5398 reflects a smaller average error in these predictions. The MAPE of 0.6330 and the R-square of 0.0950 suggest a somewhat better prediction of swings based on faults. The Pearson correlation coefficient of -0.3082 indicates a weak negative correlation between faults and swings, while the Spearman correlation coefficient of -0.2996 suggests a weak monotonic negative relationship. The analysis implies that faults may have a moderate impact on state fluctuations. A higher error rate in terms of faults may lead to a decline in a player's state, indicating a potential negative correlation between the occurrence of faults and the upward momentum of a player's state. However, the strength of the correlation coefficients suggests that faults alone may not be a decisive factor in predicting swings.

#### 5.4.3 Physical Exertion (S) and Swings

The RMSE of 0.8166 for S indicates the highest level of uncertainty in predicting swings based on physical exertion. The MAE of 0.6359 reflects a relatively larger average error in these

predictions. The MAPE of 0.6839 and the R-square of 0.0549 suggest a significant prediction bias and a limited ability of physical exertion to explain swings. The Pearson correlation coefficient of 0.2342 and the Spearman correlation coefficient of 0.2469 indicate weak positive correlations between physical exertion and swings. These findings suggest that while physical exertion is an important aspect of a player's performance, it may not be a strong predictor of state fluctuations. The weak correlations indicate that other factors, such as technique, tactics, and mental fitness, may play more significant roles in determining a player's state swings.

<b>Metric</b>	<b>RMSE</b>	<b>MAE</b>	<b>MAPE</b>	<b>R-square</b>	<b>Pearson</b>	<b>Spearman</b>
Critical Point	0.81	0.65	0.95	0.07	0.27	0.26
Faults	0.60	0.54	0.63	0.09	-0.31	-0.30
Physical Exertion	0.82	0.64	0.68	0.05	0.23	0.25

Table 6: Metrics for Critical Point, Faults, and Physical Exertion

#### 5.4.4 Conclusion

In summary, Table 6 reveals that faults have a relatively stronger relationship with swings, followed by critical points and physical exertion. However, the correlation coefficients are generally weak, indicating that these variables alone may not fully capture the complexities of a player's state fluctuations. The findings suggest that a comprehensive analysis considering multiple factors, such as tactical strategies, mental aspects, and physiological conditions, is necessary to accurately predict and understand the swings in tennis matches.

### 5.5 Advice

In tennis as well as other sport game, formulating a strategic approach based on an opponent's momentum swings across multiple matches is a demonstration of strategic acumen. It is imperative to meticulously study the opponent's recent match footage, analyzing their playing rhythm, tactical habits, and fluctuations in momentum.

Now, suppose we possess the performance data of one player and his opponent (let's say Djokovic, as an example) in multiple matches, specifically, we understand their pace and momentum in each game. Figure 14 illustrates Djokovic's performance data from the 2023 Wimbledon Championship, enabling us to discern his momentum and swing traits with a modicum of approximation. Despite his lack of a consistent playing template, it is manifest that he exhibits superior form at the outset and conclusion of contests, albeit with some alternations in his level of play during the mid-section thereof.

Based on this and more comprehensive material found on internet, we can come up with some strategies to enhancing the probability of success.

1. Particularly in the early stages of the match, Djokovic is often in good form, so it's important to stay focused in the first few games and try to compete with him, possibly breaking his serve when both fitness and mental state are at their peak.
2. In the middle of the game, when the opponent might experience a decline in form, tactics should be adjusted accordingly, such as increasing serve speed and angle, adding spin to the



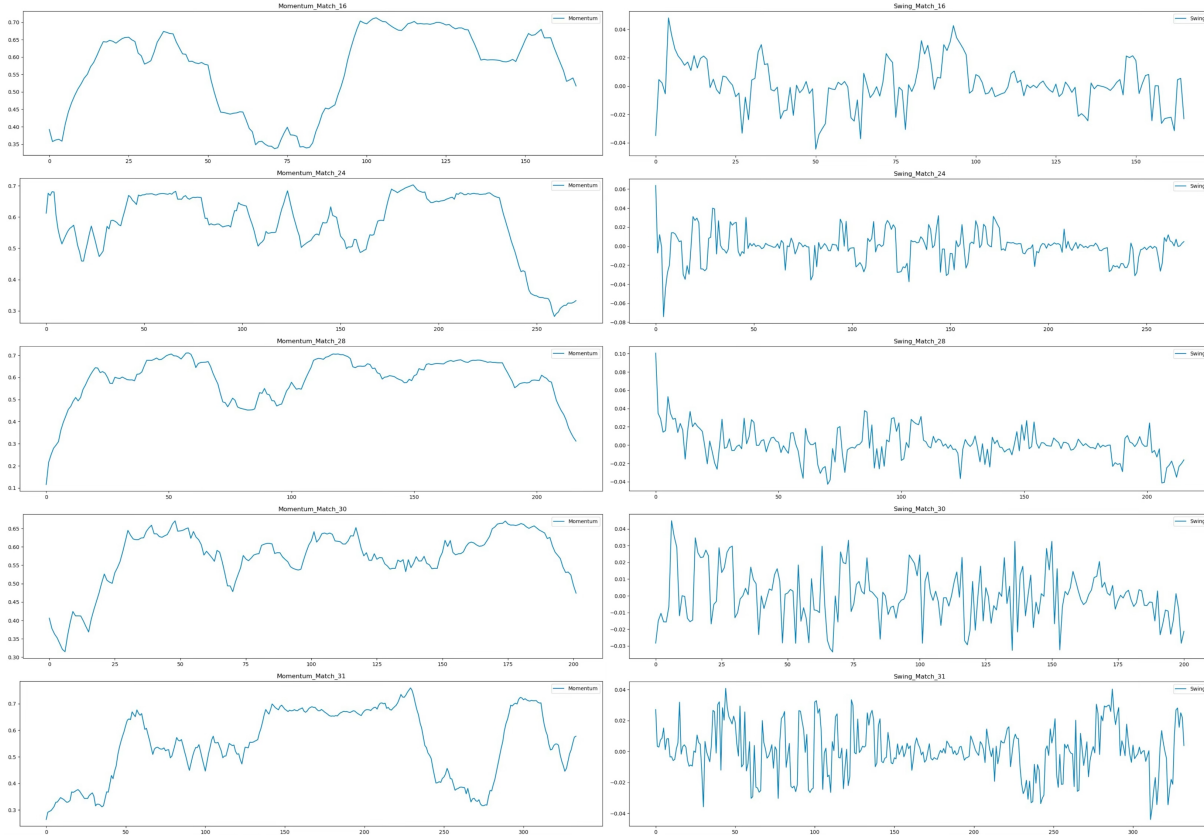


Figure 14: Navok Djokovic's performance in all matches at the 2023 Wimbledon Championships

ball, and changing the placement of shots. At the same time, maintaining a stable mindset and proper energy allocation is crucial for a long-lasting battle.

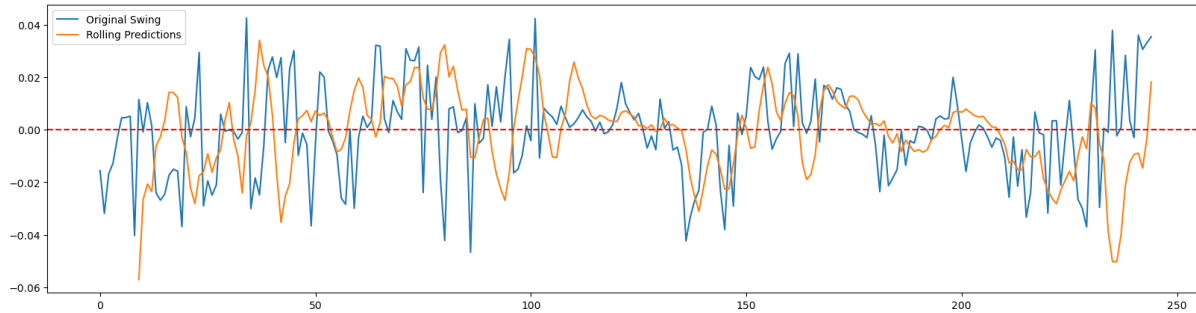
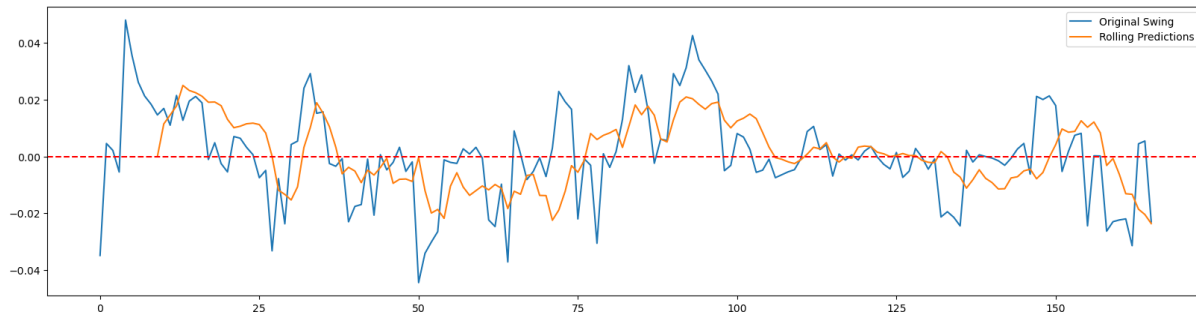
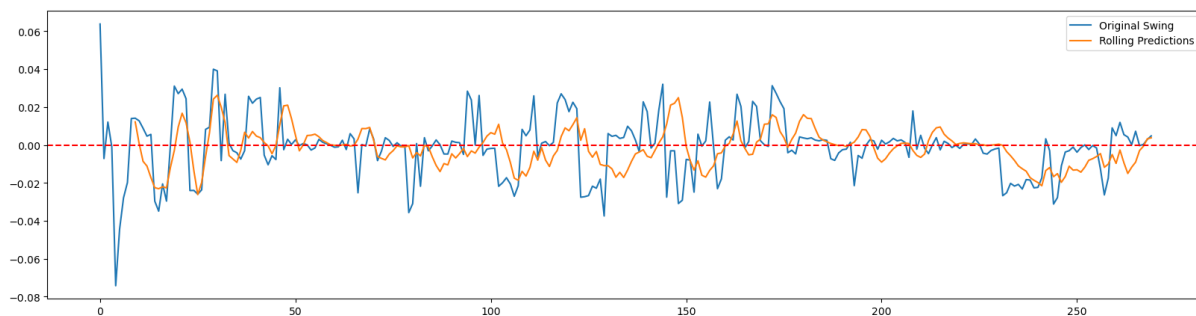
3. Give high importance to Djokovic's counterattacking ability and try to minimize unnecessary mistakes.
4. Even more to the point, maintaining flexibility and adaptability throughout the match is non-negotiable. It is advisable not to rigidly adhere to initial strategies but instead to make adjustments in response to the opponent's changing momentum. For example, not pandering to the opponent's expected ball path or altering strategies at critical moments can disrupt their rhythm. Or, when an opponent exhibits signs of gaining momentum, it is advisable for the coaching staff to promptly invoke a timeout in order to disrupt the opponent's rhythm and provide the player with a brief respite.

## 6 Task 4: Model Evaluation and Generalization

### 6.1 Metrics of ARIMA for $sw_i$ prediction

From Figure 10 and Table 7, we can see that the results of the rolling ARIMA predictions are relatively close to the ground truth. This indicates that the ARIMA model demonstrates good performance in forecasting swing sequences  $sw_i$  in *Final*.

Metric	Value
RMSE	0.171
MAE	0.130
MAPE	0.132
R2	0.096
Pearson Correlation Coefficient	0.311
Spearman Correlation Coefficient	0.270

Table 7: Metrics of Rolling ARIMA for predicting  $sw_i$  in *Final*Figure 15: predictions by rolling ARIMA model for  $sw_i$  in *Match<sub>5</sub>*Figure 16: predictions by rolling ARIMA model for  $sw_i$  in *Match<sub>16</sub>*Figure 17: predictions by rolling ARIMA model for  $sw_i$  in *Match<sub>24</sub>*

We also conducted the same modeling and evaluation for Matches 5, 6, and 7, and the data is shown in the Figure 15, 16, 17 and Table 8.

<b>Metric</b>	<i>Match</i> <sub>31</sub>	<i>Match</i> <sub>5</sub>	<i>Match</i> <sub>16</sub>	<i>Match</i> <sub>24</sub>
RMSE	0.171	0.223	0.748	0.159
MAE	0.130	0.172	0.713	0.113
MAPE	0.132	0.172	0.458	0.113
R2	0.096	0.072	0.218	0.169
Pearson Correlation Coefficient	0.311	0.085	0.467	0.411
Spearman Correlation Coefficient	0.270	0.153	0.447	0.336

Table 8: Metrics of Rolling ARIMA for predicting  $sw_i$  in *Match*<sub>5</sub>, *Match*<sub>16</sub>, *Match*<sub>24</sub>

## 6.2 Partial Poor Performance and Reasons

Considering that our performance model does not match the actual situation in certain stages of some matches(e.g, match 16), we believe the following reasons may contribute:

1. During the feature selection process, we did not take into account objective factors such as court conditions and weather, only focusing on the players' performance.
2. Different players have varying playing strategies, with some preferring baseline rallies while others prefer aggressive net play. These nuances are challenging to capture from the data.

## 6.3 Model Generalization

We obtained data for the Wimbledon Men's Singles from 2014 and 2015 online [2], and subsequently conducted data cleaning, feature extraction, KMeans clustering, and rolling ARIMA forecasting. The results are list in Table 9. We can see that the model still maintains good generalization ability.

<b>Metric</b>	<b>2014</b>	<b>2015</b>
RMSE	0.163	0.173
MAE	0.122	0.132
MAPE	0.121	0.132
R2	0.176	0.154
Pearson Correlation Coefficient	0.302	0.403
Spearman Correlation Coefficient	0.281	0.352

Table 9: Metrics of Rolling ARIMA for predicting *swing* in 2014 & 2015 Wimbledon Men's Final

## 6.4 Memo for Coaches

See Appendix B.

## 7 Strengths and weaknesses

### 7.1 Strengths

**KMEANS** In task 1, where we aimed to grasp the flow of the match, we employed the KMEANS algorithm to measure momentum, taking into account various factors such as critical points, faults, fatigue, and serving. This approach reflected the particularity of each point from multiple perspectives, thus faithfully depicted the player's competitive state and the overall situation on the court. Compared to other algorithms, our approach eliminated the impact of service by comparing and calculating within points featuring a same server. The results in task 2 also demonstrated that our measure of momentum accurately reflected the progression of the match and was highly correlated with the outcome of the game.

**ARIMA** In task 3, we employed the ARIMA model to predict the swing. The ARIMA model is adept at handling non-stationary time series data, rendering them stationary through differencing operations, which allows the model to capture long-term trends. For short-term predictions, the ARIMA model generally exhibits low forecasting errors. Additionally, the model ensures stability in the parameter estimation and forecasting process, mitigating issues of overfitting.

### 7.2 Weakness

In the features extraction, there is a lack of consideration for factors that are not provided in the data, such as weather conditions and players' tactics, etc. Long-term predictions may be suboptimal due to the lack of information. Our model is more focused on predicting and analysing the point's particularity and its impact during the competition, which may lead to biases when the match involves players with a significant disparity in level.

## References

- [1] Box, G. and Jenkins, G. (1970) Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco.
- [2] <https://github.com/ibm-datapalooza/wimbledon-datasets>

# Appendices

## Appendix A Feature extraction

---

```
# ...
weights_critical_points = {
    'p1_break_pt_won' : 10,
    'p2_break_pt_won' : -10,
    'p1_ace' : 3,
    'p2_ace' : -3,
    'p1_winner' : 2,
    'p2_winner' : -2,
    'winner_shot_type' : 0,
    'p1_break_pt_missed' : -2,
    'p2_break_pt_missed' : 2,
    'p1_net_pt_won' : 1,
    'p2_net_pt_won' : -1
}

weights_faults = {
    'p1_double_fault' : 3,
    'p2_double_fault' : -3,
    'serve_no' : 0,
    'p1_unf_err' : 1,
    'p2_unf_err' : -1
}

df_original['p1_ave_distance_run'] = df_original['p1_distance_run'] / (
    ↪ df_original['rally_count'] + 1)
df_original['p2_ave_distance_run'] = df_original['p2_distance_run'] / (
    ↪ df_original['rally_count'] + 1)
df_original['distance_diff_p1_p2'] = df_original['p1_ave_distance_run'] -
    ↪ df_original['p2_ave_distance_run']
df_original['distance_diff_p2_p1'] = df_original['p2_ave_distance_run'] -
    ↪ df_original['p1_ave_distance_run']
df_original['p1_distance_last_n'] = 0
df_original['p2_distance_last_n'] = 0
for i in range(last_n, len(df_original)):
    for j in range(1, last_n + 1):
        df_original.at[i, 'p1_distance_last_n'] += df_original.at[i-j, '
            ↪ p1_distance_run']
        df_original.at[i, 'p2_distance_last_n'] += df_original.at[i-j, '
            ↪ p2_distance_run']

df_original['p1_ave_distance_run'] = (df_original['p1_distance_run'] -
    ↪ df_original['p1_distance_run'].mean()) / df_original['p1_distance_run'
    ↪ ].std()
df_original['p2_ave_distance_run'] = (df_original['p2_distance_run'] -
    ↪ df_original['p2_distance_run'].mean()) / df_original['p2_distance_run'
    ↪ ].std()
df_original['p1_distance_last_n'] = (df_original['p1_distance_last_n'] -
    ↪ df_original['p1_distance_last_n'].mean()) / df_original['
```

```
    ↪ p1_distance_last_n'].std()
df_original['p2_distance_last_n'] = (df_original['p2_distance_last_n'] -
    ↪ df_original['p2_distance_last_n'].mean()) / df_original['
    ↪ p2_distance_last_n'].std()
df_original['speed_mph'] = (df_original['speed_mph'] - df_original['speed_mph'
    ↪ ].mean()) / df_original['speed_mph'].std()

weights_fatigue = {
    'p1_ave_distance_run' : 1,
    'p2_ave_distance_run' : -1,
    'distance_diff_p1_p2' : 1,
    'p1_distance_last_n' : 1,
    'p2_distance_last_n' : -1,
    # 'speed_mph' : 1,
}
```

---

## Appendix B Memo for Coaches

Firstly, by utilizing four detailed characteristics derived from the initial data, we apply the K-means clustering algorithm and calculate the distance of each data point from its cluster center to gauge the significance of that point, thus clarifying the game's flow. By assessing the dominant advantage based on the player's success rate, we establish the concept of "momentum" to determine which competitor is exhibiting superior performance.

Next, we plot a trend graph that displays the discrepancy in scoring against the changes in momentum throughout the match. The similarity in the patterns of these variables suggests a strong correlation between them. To further explore the impact of momentum on the eventual winner, we use a multiple linear regression model, which confirms a significant effect. Thus, we believe that the intangible indicator of momentum is actually closely related to players' performance on the field and the eventual outcome of the match.

Moreover, we develop a predictive model to estimate the game swings, i.e., the differential in momentum. We apply the ARIMA model to segment the predictions based on whether the momentum trajectory shows positive or negative trends. By calculating the correlation between these game swings and various statistical features, we identify "faults" as the most influential factor affecting the swings.

Subsequently, we examine the shifts in momentum across multiple matches to discern patterns in players' strategies. This empirical analysis forms the basis for offering targeted recommendations to improve performance. Using Novak Djokovic as a case study, we illustrate how these insights can be employed to inform strategic modifications and optimize the player's performance.

Finally, starting from the role of player momentum in the match, we provided coaches with a reference guide. This included how to quantify and visualize player momentum during the match, how to identify key turning points in the match's rhythm, and strategies for players to respond accordingly.

In summary, momentum deserves particular attention from coaches due to its comprehensive description of players' status and its high determinacy on the trajectory and eventual outcome of matches. To provide players with a competitive advantage during competitions, coaches should assess opponents' strategies by analyzing the trends of their player momentum and swing from recent matches. This includes understanding when opponents typically surge, when they are more likely to prone to stagnation, and whether they experience a severe fatigue decline in the latter stages of the game, among other insights. Based on these insights, coaches should purposefully adjust the match tempo of the athletes.

During matches, coaches can employ our model to calculate the on-court swing in real-time, thus grasping the dynamics of the game. By examining the segmentation outcomes of swing, coaches can determine the effectiveness of the current tactics. Coaches should pay special attention to changes in segmentation outcomes, as they reflect alterations in the comparative status between the opponents and uss, thereby formulating targeted strategies accordingly. If the match's swing consistently remains in the positive segment, which indicates our team is accumulating an advantage, coaches should maintain the tactical approach while controlling faults. If the swing persist in the negative zone, coaches may need to disrupt the opponents' rhythm through timeouts, breakaways, or alterations in return strategies to reverse the declining trend.

## Report on Use of AI

### 1. Github Copilot

Prompt: Give some math equations to depict the process of kmeans in latex. You can define any symbols you think appropriate and necessary.

Output: Sure, here are the LaTeX equations that depict the K-means algorithm:

Initialization: Choose  $k$  initial centroids. Let's denote them as  $c_1, c_2, \dots, c_k$ .

Assignment step: Assign each data point  $x_i$  to the nearest centroid.

This can be represented as:

$$C(i) = \arg \min_j d(x_i, c_j)$$

Update step: Recalculate the centroids by taking the mean of all data points in the cluster.

This can be represented as:

$$c_j = \frac{1}{|C_j|} \sum_{i \in C_j} x_i$$

Repeat steps 2 and 3 until the centroids do not change significantly or a maximum number of iterations is reached.

This can be represented as:

$$\text{while } \|c_j^{(t)} - c_j^{(t-1)}\| > \epsilon \text{ for any } j$$

$$\text{or } \text{iters} \leq \text{max iters}$$

### 2. Github Copilot

Prompt: How to determine the values of  $p$  and  $q$  in ARIMA based on the ACF and PACF?

Output: The Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots are useful tools to identify the order of the Autoregressive (AR) or Moving Average (MA) terms in an ARIMA model.

1. **\*\*Identifying the order of the AR term ( $p$ )\*\***: The PACF plot is used to identify the order of the AR term. The PACF plot displays the correlation between a series and its lags. The point where the PACF cuts off (drops to 0) for the first time gives the order of the AR term. For example, if the PACF cuts off after 2 lags, then we can use an AR term of order 2 (AR(2)).

2. **\*\*Identifying the order of the MA term ( $q$ )\*\***: The ACF plot is used to identify the order of the MA term. The ACF plot displays the autocorrelation between a series and its lags. The point where the ACF cuts off (drops to 0) for the first time gives the order of the MA term. For example, if the ACF cuts off after 1 lag, then we can use an MA term of order 1 (MA(1)).

Remember, these are just rules of thumb and the actual order of AR or MA terms can be higher than the point of cutoff. It's always a good idea to experiment with different orders and see which model gives the best fit.