Problem Chosen	2024 MCM/ICM Summary Sheet	Team Control Number
C		2423183

SVM-based Adaptive Markov Chain for Tennis Match Prediction: Updatable Probabilities and Momentum Analysis

Summary

As a global sporting event, tennis competitions attract attention for their intense competition. This paper aims to establish a model predicting match volatility and forecasting event outcomes. A **Probability-Updatable Markov Chain (PUMC)** model based on Support Vector Machines (SVM) is proposed to predict **dynamic winning probabilities** of each point and **conditional winning probabilities** of the match at each point.

For TASK 1, an **Adaptive Multifaceted Integration Model** is designed to assess player on-court performances under the impact of historical match results. After pre-processing and **Exploratory Data Analysis(EDA)**, this paper constructs 15 **static indicators** and 6 **dynamic indicators** based on different characteristics. Static indicators measure their own average levels of players, and **Factor Analysis** is used to extract out four crucial factors. On-court performance coefficient is generated by dynamic indicators that use **Grey Relational Analysis** to determine correlations and weights. Applying the **Exponential Moving Average** equation to adapt the performance coefficient to the average levels, on-court performance scores for any player at any point can be easily obtained. Comparing the model with real matches, it has an average accuracy of **76.5%**, with a maximum of **87%**, indicating effective evaluation of player performances closely aligned with real-world scenarios.

For TASK 2, by quantifying momentum from **psychological and behavioral** perspectives, the correlation between momentum and player swings in the match is validated. Subsequently, it performed **The Granger Causality Test** and **Causality and Correlation Analysis**. According to the Akaike Information Criterion (AIC), the Granger Causality Test revealed mutual influence between momentum and performance swings in 22 out of 31 matches. Further analysis, Correlation Analysis shows a strong positive correlation (FCC = 0.76) with momentum preceding swings.

For TASK 3, it proposes a **Probability-Updatable Markov Chain (PUMC)** model based on **SVM** to predict match swings, including **dynamic winning probabilities** for each point and **conditional winning probabilities** for the entire match. To interpret the model further, it then analyzes the accuracy and rationality of the model's predictions both **horizontally (between matches)** and **vertically (within games)**. The model achieves an average accuracy of **77.3%**, with a maximum of **92.1%**, and the highest recall is **95.4%**. Using **Shapley Additive exPlanations (SHAP)**, it identifies the four factors with the greatest impact on swings: SSER, SBR, FSR, and SSCR. Finally, based on the above analysis, it offers practical recommendations for players entering new matches.

For TASK 4, when applying the prediction model to three-set, two-win women's matches, the accuracy rate is **71.1%**. Therefore, it refines and improves the model by introducing the **serve correction factor** e and the **court type factor** p_α. Considering the similarity of racquet sports, a slight modification to the model allows for predictions in different types of matches.

Finally, this paper conducts sensitivity analysis and robustness testing on the model, revealing its good sensitivity to the **smoothing coefficient** and strong stability regarding **dynamic factor**.

Keywords: PUMC; Winning Probabilities; Exponential Moving Average; SVM; SHAP

Contents

1 Introduction

1.1 Problem Background

With the conclusion of the 2023 Wimbledon Gentlemen's Final, an increasing number of researchers are turning to modern technologies and methodologies to analyze the determinants of victory in sports competitions. "Momentum" emerges as a pivotal element among these factors, yet its direct analysis remains elusive. Consequently, the establishment of a scientific and generalizable model to investigate this phenomenon is both necessary and urgent.

1.2 Restatement of the Problem

Upon assimilating the pertinent background information, the team is tasked with executing the subsequent actions:

***Task* 1:** Capture and exemplify the match flow when points occur, assess player performance within designated timeframes and visualize the flow of the match.

***Task* 2:** Validate the model's efficacy by demonstrating the role of 'momentum' in the dynamics of the match.

***Task* 3:** Identify indicators for measuring shifts in 'momentum' and predict the correlation between these indicators, providing match-specific recommendations.

***Task* 4:** Evaluate the effectiveness of the model and propose potential refinements, assessing the model's generalizability.

***Task* 5:** Draft a two-page memorandum to communicate strategies, modeling, and outcomes with coaches and players.

1.3 Overview of Our Work

The work we have done in this problem is mainly shown in the following Figure 1.

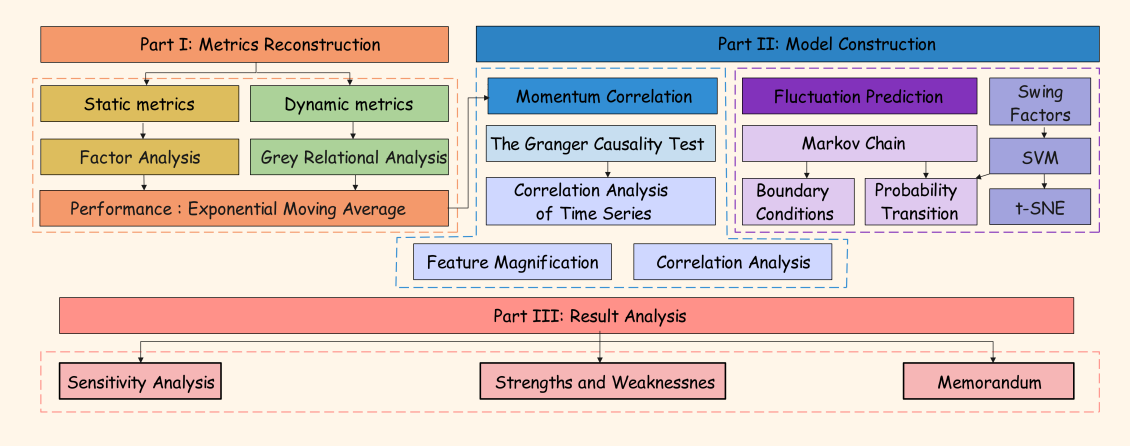


Figure 1: Flow of Our Work

2 Assumptions and Justifications

Given the multifaceted complexities inherent in the practical scenario, this study posits a set of rational assumptions to distill the issues at hand. Each assumption is meticulously corroborated with its respective rationale:

- **Assumption 1: Exclusion of external environmental disturbances, such as audience chatter and movement.**
  Justification: Evidence suggests that noise and other environmental factors can affect players' performance. In official competitions, such behaviors should be avoided.
- **Assumption 2: Factors not considered in this paper do not affect the match outcome.**
  Justification: There are numerous factors that can influence a match, such as weather conditions and audience presence. However, most of these impacts are minimal; thus, it is reasonable for the model to disregard these minor factors.
- **Assumption 3: The provided data accurately reflects the average level of the players.**
  Justification: A match consists of sets and games, leading us to believe that the provided data is sufficient to assess the average individual level of the players.

3 Notations

The key mathematical notations used in this paper are listed in the following Table 1.

Table 1: Notations used in this paper

Symbol	Description
S_j	the average individual level of the j-th player
$D_{i,j}$	the composite dynamic measure for the j-th player at the i-th point
$B_{i,j}$	the performance of the j-th player at the i-th point
L_i	the probability of winning changes with scoring or concedingpoints
m_i	the value of momentum at the i-th point
SF_s	the s-th swing factor
α	the smoothing coefficient
β	the dynamic factor

4 Model Preparation

- Data Transformation
  We performed data transformation regarding scoring principles related to the competition rules. For instance, for certain scoring rules involving data such as AD and score, we replaced them to facilitate subsequent model calculations.
- Data Standardization
  To enhance model performance during the subsequent modeling process, we employed the Z-score method for data standardization. This method involves scaling each feature by its standard deviation, transforming the data into a standard normal distribution with a mean of 0 and a standard deviation of 1. This aids the model in handling scale differences among different features, thereby improving training effectiveness.

5 Task 1: Flow of Matches With Performance Evaluation

5.1 Metrics Reconstruction

Observing the dataset "data_dictionary.csv", we find that it comprises 46 variables. Given the large number of variables, it is essential to categorize and combine them for more efficient analysis.

In this regard, we have considered the characteristics of each variable, constructing new metrics from both static and dynamic perspectives, and have further classified these metrics on this basis.

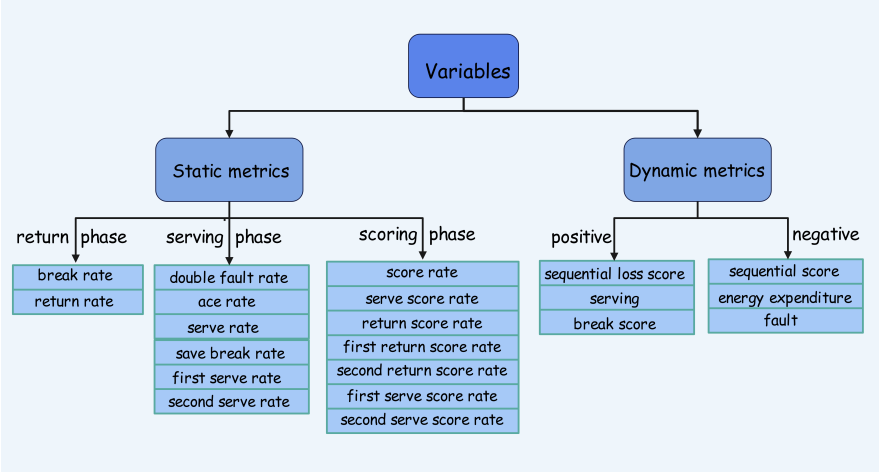


Figure 2: Refactoring metrics from static and dynamic perspective

In Figure 2, static metrics reflect the player's average level of play, considered to be the ability accumulated through long-term training, and are characterized by stability. Dynamic metrics, on the other hand, reflect the player's temporary state, which changes throughout a match. Explanations for each reconstructed metric can be found in the Appendices (Figure 19).

In the following two sections, we will explore different methods for extracting metrics, taking into account the unique characteristics of both types of metrics.

5.2 Factor Analysis of Static Metrics

Given the abundance of static metrics (see those metrics in Appendices, Figure 19), factor analysis can be employed for dimensionality reduction and simplification of metrics, using several factors to describe the relationships between metrics and to extract the main influencing information.

***Step* 1: Positive direction adjustment of static metrics**

***Step* 2: Statistical analysis**

We apply the Kaiser-Meyer-Olkin (KMO) test and Bartlett's test of sphericity to examine the relationships between variables. As shown in Table 2, a KMO value equal to 0.792 and a significance level of P = 0.0000001 indicate significant correlations among variables, validating the effectiveness of factor analysis.

***Step* 3: Explained variance and factor rotation**

According to Table 3, the cumulative variance contribution of the first four factors reaches **89.425%**, indicating a strong explanatory capacity. Finally, we can summarize the four factors as: return game performance factor (F_1), serve game performance factor (F_2), serving stability factor (F_3), and scoring factor (F_4).

Table 2: KMO and Bartlett's test

KMO		0.792
Bartlett	Approx. Chi-Square	965.916
	df	105
	P	0.0000001

Table 3: Variance explained

Name	% of Variance(Rotated)	Cumulative % of Variance(Rotated)	% Weight
F1	0.302	30.213	33.786
F2	0.273	57.521	30.537
F3	0.208	78.322	23.261
F4	0.11	89.425	12.416

***Step* 4: Static metric evaluation**

We define S_j as the average individual level of the j-th player.

$$S_j = 0.338 \times F_1 + 0.305 \times F_2 + 0.233 \times F_3 + 0.124 \times F_4 \tag{1}$$

5.3 Grey Relational Analysis of Dynamic Metrics

Dynamic metrics (see those metrics in Appendices, Figure 19) exhibit certain correlations; thus, through grey relational analysis, the degree of association and weights are determined, further evaluating and elucidating the interactions between metrics over time.

***Step* 1: Data preprocessing.**

The process involves normalizing metrics to ensure a positive orientation and incorporating the optimal reference sequence (X_0), resulting in the decision matrix X_{DM} as follows:

$$(X_0, X_1, \ldots, X_6) = \begin{bmatrix} x_0(1) & x_1(1) & \cdots & x_6(1) \\ x_0(2) & x_1(2) & \cdots & x_6(2) \\ \vdots & \vdots & \ddots & \vdots \\ x_0(n) & x_1(n) & \cdots & x_6(n) \end{bmatrix} \tag{2}$$

***Step* 2: Calculate correlation coefficient.**

The correlation coefficient between each comparison sequence ($X_1, \ldots, X_6$) and the optimal reference sequence (X_0) are calculated. A higher value of $\xi_i(k)$ indicates a stronger correlation:

$$\xi_i(k) = \frac{\min\limits_i \min\limits_k |x_0(k) - x_i(k)| + \rho \cdot \max\limits_i \max\limits_k |x_0(k) - x_i(k)|}{|x_0(k) - x_i(k)| + \rho \cdot \max\limits_i \max\limits_k |x_0(k) - x_i(k)|} \tag{3}$$

***Step* 3: Calculate correlation.**

The correlation, which represents the weight of each metric, is calculated as follows.

$$r_{0i} = \frac{1}{n}\sum_{k=1}^{n}\xi_i(k) \tag{4}$$

***Step* 4: Dynamic metric evaluation.**

Define $D_{i,j}$ as the state of the match for the j-th player at the i-th point:

$$D_{i,j} = 0.211 \times ss_{i,j} - 0.225 \times sls_{i,j} - 0.121 \times ee_{i,j} + 0.142 \times s_{i,j} + 0.103 \times bs_{i,j} - 0.184 \times fa_{i,j} \tag{5}$$

5.4 Adaptive Multifaceted Integration Model

5.4.1 Player's Performance Evaluation

Inspired by the Exponential Moving Average (EMA) [1], we developed the Adaptive Multifaceted Integration Model (AMIM). Rather than simply weighting the results of static metric evaluations (individual level) and dynamic metric evaluations (match state) simultaneously, we believe that in a match, the state of the game does not directly decide the outcome. Instead, it acts upon the individual level, thereby influencing the result. Consequently, we define the performance of the j-th player at the i-th point as $B_{i,j}$.

$$B_{i,j} = \frac{D_{i,j} + (1-\alpha)D_{i-1,j} + (1-\alpha)^2 D_{i-2,j} + ... + (1-\alpha)^{i-1}D_{1,j}}{1 + (1-\alpha) + (1-\alpha)^2 + ... + (1-\alpha)^{i-1}} \cdot S_j = \frac{\sum_{t=1}^{i}(1-\alpha)^{i-t}D_{t,j}}{\sum_{t=1}^{i}(1-\alpha)^{i-t}} S_j \tag{6}$$

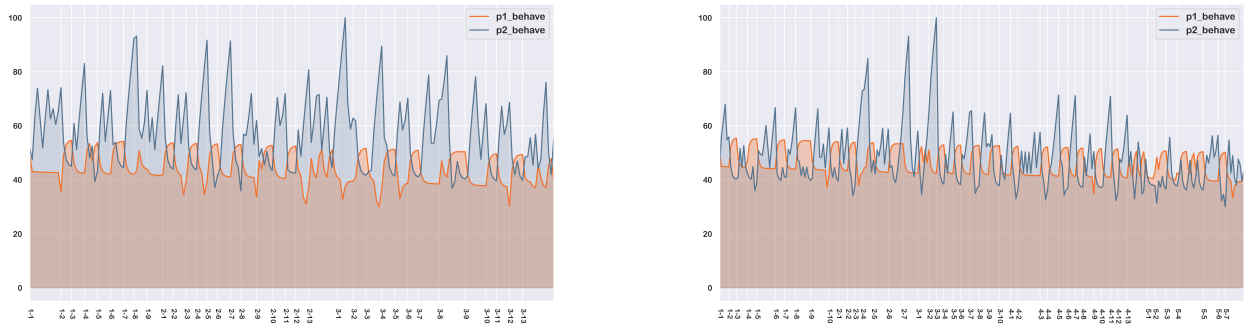
Here, α represents the smoothing coefficient, which controls the extent to which a series of previous match states affects the player's base level. We set α to $\frac{1}{3}$.

In Equation 6, applying EMA to $D_{i,j}$ implies that the match states closer to the i-th point have a greater influence on the current state of the match. Moreover, over a short period, there will not be significant changes in S_j. This multiplication can be considered as the current performance of the j-th player.

5.4.2 Player's Performance Swings in Response to Pointing Events

Following the application of AMIM to 31 matches, match situations can be categorized into four distinct types: Dominating Set, Comeback Set, Rally Set, and Tight Set. Subsequently, we select the most representative match for each of these situations (see Figure 3) to analyze player performance throughout the pointing process within the entire match flow.

For example, the dominating set, as illustrated in Figure 18a (match_id: 2023-wimbledon-1302), involved three sets. Player 2 achieved a dominating victory in each set, largely due to consistently outperforming Player 1 on the majority of points.(The rest two set situations, comeback set and rally set. These figure in Appendices, Figure 18)



(a) Dominating Set (b) Tight Set

Figure 3: Flow of two match situations

5.4.3 Validation of Performance Evaluation

To validate the effectiveness of our model in evaluating player performance, we calculate the difference in performance values between players for each match ($B_{\text{player1}} - B_{\text{player2}}$), using the resulting difference to predict the outcome of each scoring point and comparing it with the actual outcomes. The prediction accuracy for scoring points across all 31 matches was then statistically analyzed (Figure 4). It was observed that our model achieved an average accuracy of **76.5%** across all matches, with the highest accuracy reaching **87%**.

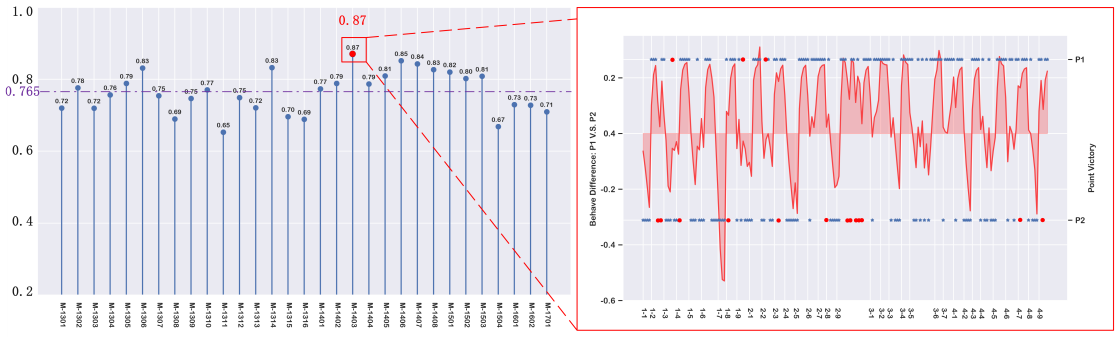


Figure 4: Prediction accuracy for 31 matches

For instance, focusing on predictions with the highest precision, the magnified view reveals that the performance differences under most points are consistent with whether a score is made or not. The discrepancies in certain predictions from actual outcomes may be attributed to the influence of momentum.

Finally, we provide a visual description of the match flow in Figure 5.

6 Task 2: Causality and Correlation Analysis of Momentum

6.1 Basic Definitions

In sports events, momentum is an important yet elusive metric. In the Literature Review section, we have already considered the definition of momentum from two perspectives (PM and BM). Before defining momentum, we first selected relevant metrics from these two dimensions (Table 4).

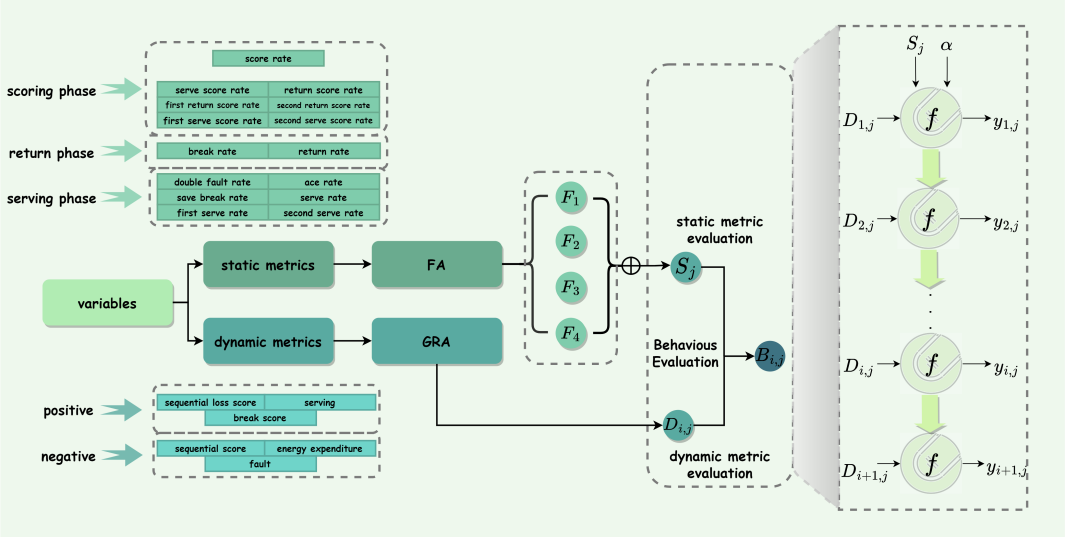


Figure 5: Description of the match flow

Table 4: Psychological metrics and behavioral metrics

Perspective	Metrics	Type	Domain	Explanation
PM	SP	discrete	-1, 1	serving possession
	SS	discrete	0, 2, 3, 4	consecutive scoring
	SL	discrete	0, 2, 3, 4	consecutive losing scores
	FA	discrete	0(F), 1(T)	fault
	FUA	discrete	0(F), 1(T)	unforced fault
BM	E_p	continuous	[-1, 1]	energy expenditure ratio

Here, E_p defines:

$$E_p = \frac{\sum_{i=1}^{N} run_i}{\sum_{i=1}^{M} run_i} \tag{7}$$

Where $\sum_{i=1}^{N} run_i$ represents the cumulative distance covered by a player up to the N$^{\text{th}}$ point, and M denotes the total number of movements made by the player within the game.

6.1.1 Leverage

Leverage, a measure of how the probability of winning changes with scoring or conceding points, holds significant referential value for the analysis of sports events [2]. To better integrate PM and BM, we define leverage L_i as follows:

$$L_i = w_1 \cdot SS_i + w_2 \cdot SL_i + w_3 \cdot FA_i + w_4 \cdot FUA_i + w_5 \cdot SS_i + w_6 \cdot E_{p_i} \tag{8}$$

6.1.2 Momentum Value

Practical considerations indicate that, regardless of the analytical perspective or measurement method employed, there are limitations. Inspired by Briki, Walid, et al. [3], we will measure momentum from an integrative perspective. A detailed momentum framework is shown as Figure 6.

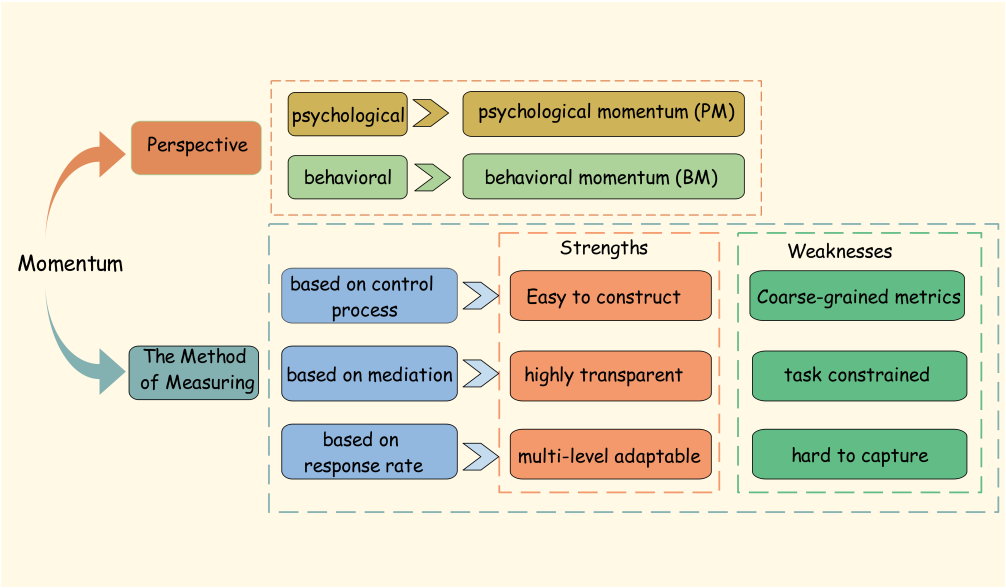


Figure 6: A Comprehensive Momentum Framwork

We define the momentum value m_i:

$$m_i = \frac{L_i + (1-\beta)L_{i-1} + (1-\beta)^2 L_{i-2} + ... + (1-\beta)^{i-1} L_1}{1 + (1-\beta) + (1-\beta)^2 + ... + (1-\beta)^{i-1}} = \frac{\sum_{t=1}^{i} (1-\beta)^{i-t} L_t}{\sum_{t=1}^{i} (1-\beta)^{i-t}} \tag{9}$$

Where β represents the dynamic factor, which we set to $\beta = \frac{1}{3}$.

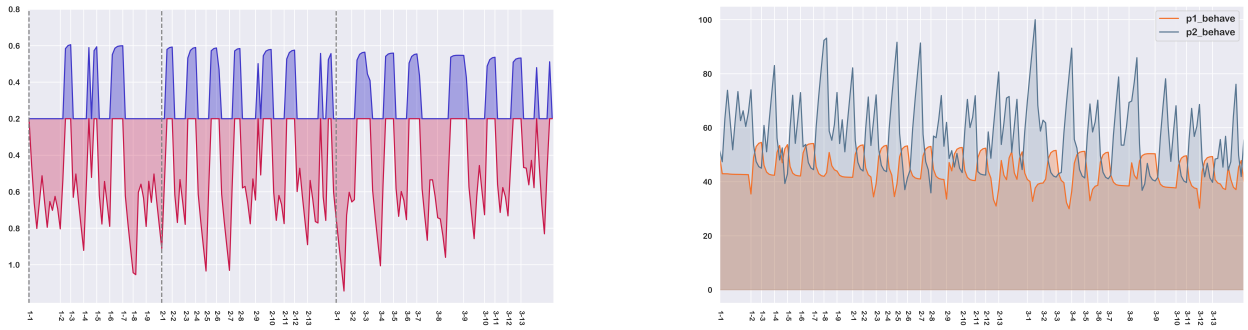
6.1.3 Clutch

The concept of a clutch, as identified by [2], refers to a point that significantly influences the probability of winning the current match. We consider a point to be a clutch when its leverage rate exceeds a predefined threshold.

Taking the dominating set from Task One as an example(Figure 7), we further analyze their momentum values.

It can be observed that momentum value effectively explains the occurrence of a dominating set:

- The momentum value of player2 is greater than that of player1.
- The duration of player2's momentum advantage significantly exceeds that of player1.



(a) Momentum value of dominating set

(b) Performance of dominating set

Figure 7: Analysis of 2023-wimbledon-1302's momentum

6.2 Causality Analysis Between Momentum and Performance

6.2.1 Construct Time Series

To analyze the correlation between the fluctuating values of momentum and the performance differential between two players within the same match, we first serialize and standardize the data to construct time series. After processing, we obtain time series where y can represent m (momentum) or $B_{\text{player1}} - B_{\text{player2}}$.

$$Y_t = \{y_{t1}, \ldots, y_{tM}\} \tag{10}$$

6.2.2 The Granger Causality Test

Firstly, the Augmented Dickey-Fuller (ADF) test method is employed to check the stationarity of each time series, thus avoiding the phenomenon of "spurious regression". For non-stationary time series, differencing is performed until stationarity is achieved. Upon obtaining stationary time series, the Granger causality test model [4] is utilized to examine the causal relationship between the value of momentum and the performance margin across 31 matches. The optimal lag order is determined based on the Akaike Information Criterion (AIC), and the combinations with significant causal relationships are identified as shown in Table 5.

Table 5: Causality of momentum value and performance margin in 31 matches

Match	M-1301	M-1302	M-1303	M-1304	M-1305	M-1306	M-1307	M-1308
Causality	$\Longleftrightarrow$	$\Longleftrightarrow$	$\Leftarrow$	$\Longleftrightarrow$	$\Longleftrightarrow$	$\Longleftrightarrow$	$\Leftarrow$	$\Longleftrightarrow$
Match	M-1309	M-1310	M-1311	M-1312	M-1313	M-1314	M-1315	M-1316
Causality	$\Rightarrow$	$\Longleftrightarrow$	$\Longleftrightarrow$	$\Longleftrightarrow$	$\Longleftrightarrow$	$\Longleftrightarrow$	$\Longleftrightarrow$	$\Longleftrightarrow$
Match	M-1401	M-1402	M-1403	M-1404	M-1405	M-1406	M-1407	M-1408
Causality	$\Leftarrow$	$\Longleftrightarrow$	$\Leftarrow$	$\Rightarrow$	$\Longleftrightarrow$	$\Longleftrightarrow$	$\Leftarrow$	$\Longleftrightarrow$
Match	M-1501	M-1502	M-1503	M-1504	M-1601	M-1602	M-1701	
Causality	$\Rightarrow$	$\Longleftrightarrow$	$\Longleftrightarrow$	$\Leftarrow$	$\Longleftrightarrow$	$\Longleftrightarrow$	$\Longleftrightarrow$	

Where, $\Longleftrightarrow$ denotes that in this match, the value of momentum and the performance margin cause changes in each other, $\Rightarrow$ indicates that the value of momentum leads to changes in the performance margin, and $\Leftarrow$ signifies that the performance margin influences the value of momentum.

It was observed that in 22 matches, changes in momentum and performance differences caused changes in each other. In 3 matches, momentum influenced performance differences, while in 6 matches, performance differences affected momentum. This preliminarily indicates a causal relationship between momentum and performance differences, suggesting that they mutually influence each other in most cases.

6.3 Correlation Analysis Between Momentum and Performance

6.3.1 Feature Magnification

Define function $f_{\alpha,\beta}(x)$:

$$f_{\alpha,\beta}(x) = \begin{cases} e^{\alpha \min(x,\beta)} & \text{if } x \geq 0 \\ -e^{\alpha \min(|x|,\beta)} & \text{if } x < 0 \end{cases} \tag{11}$$

6.3.2 Correlation Analysis

Applying the function $f(\cdot)$ to Y_t, we obtain $\hat{Y_t} = \{f(y_{t1}), \ldots, f(y_{tM})\}$. For ease of analysis, we define $K = \{k_1, \ldots, k_M\}$ (for $y = m$), and $G = \{g_1, \ldots, g_M\}$ (for $y = B_{\text{player1}} - B_{\text{player2}}$). Construct K_s as follows:

$$K_s = \begin{cases} \{0, \ldots, 0, k_1, \ldots, k_{n-s}\} & \text{if } s \geq 0 \\ \{k_{n-s}, \ldots, k_1, 0, \ldots, 0, \} & \text{if } s < 0 \end{cases} \tag{12}$$

where $|s|$ represents the number of zeros, and $-M < s < M$.

We define the inner product of K_s and G as $R(K_s, G) = K_s \cdot G$, and the specific calculation for correlation is as follows:

$$CC(K_s, G) = \frac{R(K_s, G)}{\sqrt{R(K_s, G), R(G, G)}} \tag{13}$$

Given that the direction of swings between the two time series curves might be consistent or opposite, it is necessary to consider cases where the correlation is greater than zero and less than zero, respectively.

$$\begin{aligned} CC_{min} &= \min_{-M<s<M}(K_s, G) \\ CC_{max} &= \max_{-M<s<M}(K_s, G) \end{aligned} \tag{14}$$

The indices of the maximum and minimum values are determined by:

$$\begin{aligned} s_1 &= \arg\min_{-M<s<M}(K_s, G) \\ s_2 &= \arg\max_{-M<s<M}(K_s, G) \end{aligned} \tag{15}$$

The tuple $FCC(K, G)$ is defined as:

$$FCC(K,G) = \begin{cases} (CC_{min}, s_1) & \text{if } |CC_{max}| < |CC_{min}| \\ (CC_{max}, s_2) & \text{if } |CC_{max}| \geq |CC_{min}| \end{cases} \tag{16}$$

This tuple represents three pieces of information: the significant correlation between the two time series, the sequence of swing occurrence, and the direction of swing.

- $FCC(K, G)$ is confined within the range $[-1, 1]$. Values closer to 1 or -1 indicate a stronger correlation between K and G.
- For determining the sequence of swing occurrence, $s > 0$ implies that G fluctuates before K, and vice versa.
- Regarding the direction of swing, if $FCC(K, G) > 0$, it indicates that the swings are in the same direction, i.e., positively correlated; otherwise, they are negatively correlated.

6.4 From Macro to Micro: A Layered Analysis of Momentum

6.4.1 Macro Analysis of FCC

In the analysis of 31 matches, the FCC (swing Characteristic Coefficient) between the momentum of each match and the performance margin, as listed in Table 6, is consistently above 0.76, indicating a strong positive correlation. Moreover, a noticeable pattern is that swings in momentum precede those in performance margin.

Table 6: FCC of momentum and performance margin in 31 matches

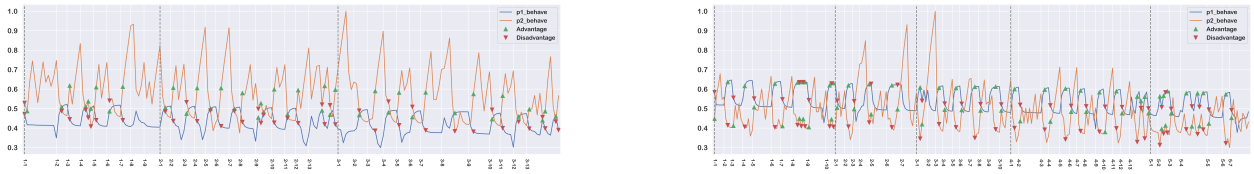
Match	M-1301	M-1302	M-1303	M-1304	M-1305	M-1306	M-1307	M-1308
FCC	0.93	0.905	0.901	0.773	0.915	0.935	0.766	0.791
Match	M-1309	M-1310	M-1311	M-1312	M-1313	M-1314	M-1315	M-1316
FCC	0.891	0.87	0.948	0.797	0.825	0.811	0.804	0.794
Match	M-1401	M-1402	M-1403	M-1404	M-1405	M-1406	M-1407	M-1408
FCC	0.882	0.891	0.81	0.84	0.892	0.826	0.849	0.911
Match	M-1501	M-1502	M-1503	M-1504	M-1601	M-1602	M-1701	
FCC	0.779	0.815	0.857	0.877	0.825	0.832	0.902	

6.4.2 Micro Analysis of Momentum

Before capturing momentum, it is necessary to identify all clutch moments within each match. By observing the performance following clutch moments, we assess the effect of momentum.

Here, taking two match situations from Task One as examples, we elaborate further.

In Figure 8, we observe: when the clutch moment is at an Advantage, under the influence of momentum, the player performs well. Conversely, when the clutch moment is at a Disadvantage, the influence of momentum results in poor performance.



(a) Momentum of dominating set (b) performance of tight set

Figure 8: Capture momentum with clutches

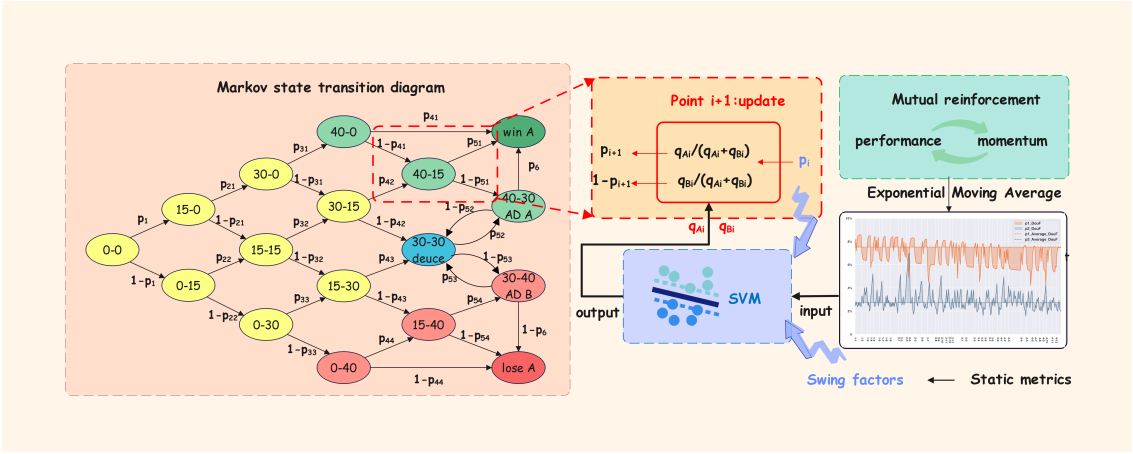


Figure 9: Workflow of Task 3

7 Task 3: Flow of Play with Probability Prediction

In sports events, swings in the match are often time-related. Such phenomenon is considered to be congruent with the Markov chain. However, this approach has a critical limitation: it necessitates the assumption that each set within a match is independent of others, an assumption dictated by the unique scoring rules of tennis competitions.

Starting from this premise, we propose the incorporation of Support Vector Machines (SVM) to refine the Markov chain, thereby facilitating a more accurate analysis of tennis matches.

7.1 Basic Markov Chain Model to Tennis

7.1.1 Game Perspective

Before the analysis, we first define the relevant variables:

- $P(a, b)$**:** The probability that player1 wins a game when the point is (a, b).
- p_i**:** The scoring rate of player1 when serving at the i-th point.
- q_i**:** The scoring rate of player2 when serving at the i-th point.

For a game, there are two situations: regular game and tie-breaker.

***Case* 1: Regular game:** Here we assume this game is player1's serve.

- **Probability transition equation:**

$$P(a, b) = p_i \cdot P(a + 1, b) + (1 - p_i) \cdot P(a, b + 1) \tag{17}$$

- **Boundary conditions:**

$$\begin{cases} P(4,b) = 1 & \text{if } b \leq 2 \\ P(a,4) = 0 & \text{if } a \leq 2 \end{cases} \tag{18}$$

- **30 all:**

$$P(3,3) = p_i^2 P(5,3) + 2p_i(1-p_i)P(4,4) + (1-p_i)^2 P(3,5) = \frac{p_i^2}{p_i^2 + (1-p_i)^2} \tag{19}$$

***Case* 2: Tie-breaker:** Here we assume player1 serves first in this game.

- **Probability transition equation:**

$$\begin{cases} P^*(a,b) = p_i \cdot P^*(a+1,b) + (1-p_i) \cdot P^*(a,b+1) & \text{if } (a+b) mod 2 = 0 \\ P^*(a,b) = q_i \cdot P^*(a,b+1) + (1-q_i) \cdot P^*(a+1,b) & \text{if } (a+b) mod 2 = 1 \end{cases} \tag{20}$$

- **Boundary conditions:**

$$\begin{cases} P^*(7,b) = 1 & \text{if } b \leq 5 \\ P^*(a,7) = 0 & \text{if } a \leq 5 \end{cases} \tag{21}$$

- **60 all:**

$$P^*(6,6) = \frac{p_i(1-q_i)}{p_i(1-q_i) + q_i(1-p_i)} \tag{22}$$

7.1.2 Set Perspective

By treating a game as a point, we can naturally extend the formulas used within a game to a set. Once the pointing rate for each point within a game is determined, the corresponding probability of winning the game is also defined. Here is the definition the relevant variables.

- $P_S(c,d)$**:** The probability of player1 winning a set when the game is (c, d).
- P**:** The probability of player1 winning a game in player1's serve.
- Q**:** The probability of player1 winning a game in player2's serve.

- **Probability transition equation:**

$$\begin{cases} P_S(c,d) = P \cdot P_S(c+1,d) + (1-P) \cdot P_S(c,d+1) & \text{if } (c+d) mod 2 = 0 \\ P_S(c,d) = Q \cdot P_S(c,d+1) + (1-Q) \cdot P_S(c+1,d) & \text{if } (c+d) mod 2 = 1 \end{cases} \tag{23}$$

- **Boundary conditions:**

$$\begin{cases} P_S(6,d) = 1 & \text{if } d \leq 4 \\ P_S(c,6) = 0 & \text{if } c \leq 4 \end{cases} \tag{24}$$

- **At 5-5, the equation becomes:**

$$P_S(5,5) = P \cdot Q + \big(P \cdot Q + (1-P)(1-Q)\big)P^* \tag{25}$$

7.1.3 Match Perspective

By considering a set as a point, we can naturally extend the formulas used in games to matches. Once the probability of winning each game within a set is established, the corresponding probability of winning the set is also determined. Here is the definition the relevant variables.

- $P_M(e, f)$**:** The probability of player1 winning the match when the score is (e, f).
- P_S**:** The probability of player1 winning a set.
- **Probability transition equation:**

$$P_M(e, f) = P_S \cdot P_M(e + 1, f) + (1 - P_S) \cdot P_M(e, f + 1) \tag{26}$$

- **Boundary conditions:**

$$\begin{cases} P_M(3, f) = 1 & \text{if } f \leq 2 \\ P_M(e, 3) = 0 & \text{if } e \leq 2 \end{cases} \tag{27}$$

7.2 Probability Updatable Markov Chain (PUMC) Based On SVM

7.2.1 Theoretical Formula

In Task Two, we have demonstrated the effectiveness of momentum and its interaction with performance difference. Drawing from the definition of momentum value, we further transform static metrics: the application of momentum value on static metrics introduces variability, which we refer to as the swing Factor (SF).

$$SF_s = \frac{L_i + (1-\beta)L_{i-1} + (1-\beta)^2 L_{i-2} + ... + (1-\beta)^{i-1} L_1}{1 + (1-\beta) + (1-\beta)^2 + ... + (1-\beta)^{i-1}} \cdot X_s = \frac{\sum_{t=1}^{i} (1-\beta)^{i-t} L_t}{\sum_{t=1}^{i} (1-\beta)^{i-t}} \cdot X_s \tag{28}$$

where $s = 1, 2, \ldots, 15$. X_s represents a static metric (for example, ar (ace rate), scr (score rate)).

Adaptability Explanation: First, we defined a dynamic factor set, representing the cumulative impact of past matches that have occurred:

$$SF_{set} = \{SF_1, SF_2, \ldots, SF_{15}\}$$

To determine the current round's probability of winning, we need to calculate the probability of winning in the current round given the dynamic factors. This integrates the influence of previous match outcomes, ensuring that the winning probability for each round depends on the overall competition situation:

Define $SF_{subset} = \{x | x \subset SF_{set} \bigcap x \neq \emptyset\}$, then

$$P(correct_i | SF_{subset_{i-1}}, \ldots, SF_{subset_1}) \tag{29}$$

where $correct_i$ denotes a successful prediction for the i-th point

7.2.2 PUMC Solving Algorithm

Based on the above basic data and models, the process of obtaining the probability strategy is as follows:

Algorithm 1: PUMC Solving Algorithm

Initialization: DMetrics = $C_{i\cdots,s}$, Point Number: $P_{\text{num}} = 0$,
FinalState = {'4-0': 1,...,'4-2':1,'0-4':0,...,'2-4':0}

Step 1: Determine the adaptive winning probability in a point

while *No Point Winner* **do**

  for $t \leftarrow 1$ *to* P_{num} **do**

    Calculate the momentum and performance: SF_i;

    Transform(DMetrics);

    $Prob_t$ = SVM(DMetrics);

    State transition: $P(a+1,b) \leftarrow P(a,b)$ or $P(a,b+1) \leftarrow P(a,b)$;

    Update DMetrics C_i;

Step 2: Predict the winning probability of the match at each point

for *game in a set* **do**

  while *No Game Winner* **do**

    for $t \leftarrow 1$ *to* P_{num} **do**

      if (a,b) *not in FinalState* **then**

        $P(a,b) = Prob_t \cdot P(a+1,b) + (1 - Prob_t) \cdot P(a,b+1)$;

      else

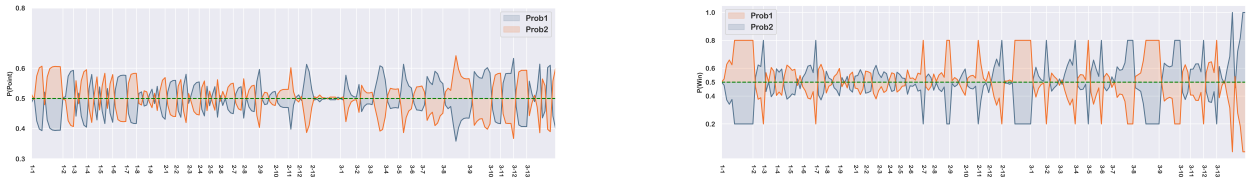
        $P(a,b)$ = FinalState(a,b);

        break;

7.3 Result Analysis

7.3.1 Prediction of Winning Probability

After adjustments through SVM, our model achieved a final accuracy rate of **77.3%**. To further observe when the flow of play is about to change from favoring one player to the other, we consider 0.5 as a pivotal point. A transition is identified when the probability of winning a game shifts from above 0.5 to below 0.5 or vice versa. Here, we take the dominating set as an example (match_id: 2023-wimbledon-1302), with the prediction results shown in Figure 10.



(a) Point level (b) Match level

Figure 10: Prediction of winning probability

- **From point level.** Considering the winning probability from a point perspective, Player 2's probability is mostly above 0.5, indicating a predominant chance of victory across most points. Additionally, there are instances where Player 1 holds the advantage, but Player 2 quickly regains the upper hand in the following point, further elucidating the characteristics of a dominating set.
- **From match level.** Observing the winning probability of the last two points, it becomes evident that Player 1 is almost certain to win the match. For other points, Player 2's winning probability is generally above 0.6, and in some cases, even as high as 0.8, illustrating a good grasp of the dominating set's nature from the match level.

7.3.2 Validation of SVM Effectiveness

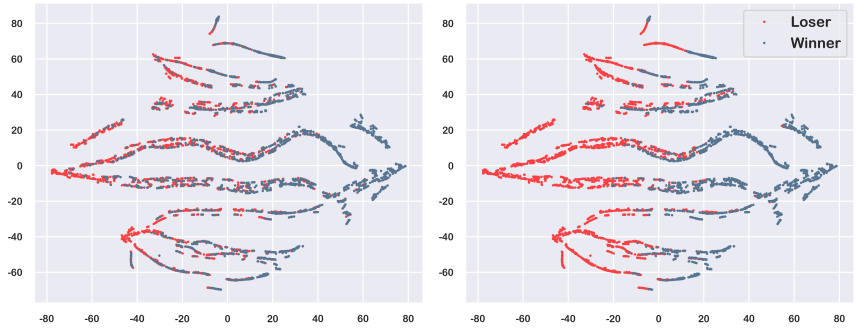


Figure 11: T-SNE visualization of the dataset

Finally, we perform T-SNE dimensionality reduction on the real dataset (left image) versus the results after SVM training (right image) as shown in Figure 11. When reduced to two dimensions, using a vertical line as the divider, we can distinguish two states. It is observed that SVM effectively captures this state. The points where colors are mixed on both sides of the state reflect the swing of the state, and SVM has also captured this volatility to a certain extent.

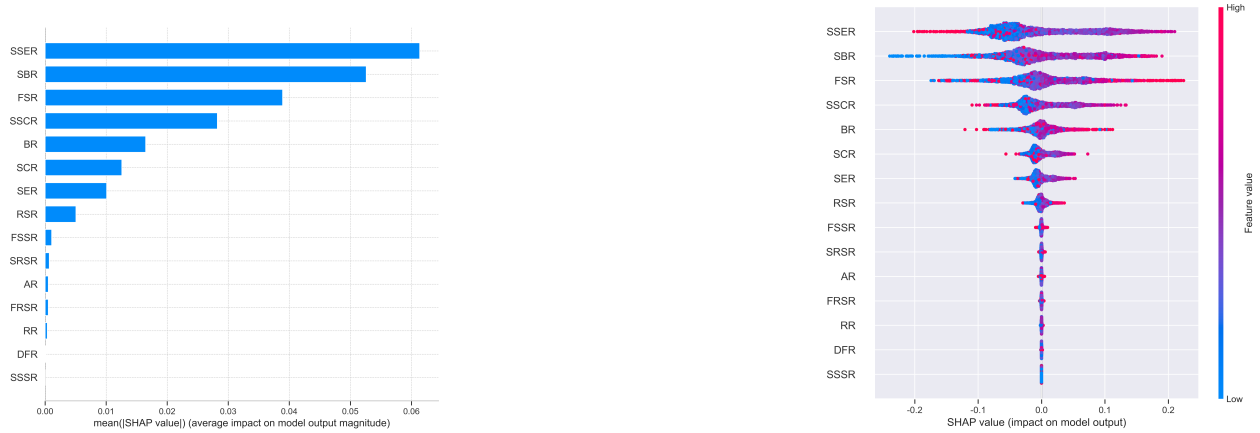
Explanation for the differences in volatility: Since the data is sourced from the Wimbledon Gentlemen's final, we believe that the levels of the players are closely matched, and it is the state of the match that leads to increased volatility in the outcomes.

7.3.3 Capturing the Dominant swing Factors

In the previous section, we updated the serving score using SVM. Besides considering the accuracy of SVM, we also need to interpret SVM's prediction results with other methods to further determine the dominant swing factors of the flow of play.

Here, we employ SHapley Additive exPlanations (SHAP) [5], a method based on the concept of Shapley values from cooperative game theory, to explain the predictions of machine learning models.

- **Feature values.** Features with larger Shapley absolute values are considered important. Here, we measure SHAP feature importance by the average Shapley absolute value. It is found that the most important swing factors are SSER, SBR, FSR, SSCR, while DFR and SSSR have almost no effect on the prediction outcome.
- **SHAP values.** The summary plot combines feature importance and the impact of features. Each point on the summary plot represents the SHAP value for a feature and a specific data



(a) Feature values (b) SHAP values

Figure 12: Capture the dominant factors from 15 swing factors

point. Further analysis reveals that SSER (second serve success rate), SBR (break point save rate), FSR (first serve success rate), SSCR (serving scoring rate) predominantly play a facilitative role in the outcome.

In conclusion, we identify the dominant swing factors influencing the flow of play as: SSER, SBR, FSR, SSCR.

7.3.4 Practical Suggestions for a New Match

Based on our previous research, we found that momentum changes during a match can influence the outcomes of individual games. Through feature selection and model validation, we identified that these momentum factors are primarily affected by first-serve success rate, second-serve success rate, saved break points success rate, and service game winning rate. Therefore, for a player participating in a new match, we offer the following suggestions:

Self-Adjustment: Given the strong correlation between momentum and performance in a match, we observe that momentum often determines a player's specific performance in the current game. It is advisable for the player to forget about negative outcomes from past matches, adjust their mental state, and fully unleash their capabilities on the court.

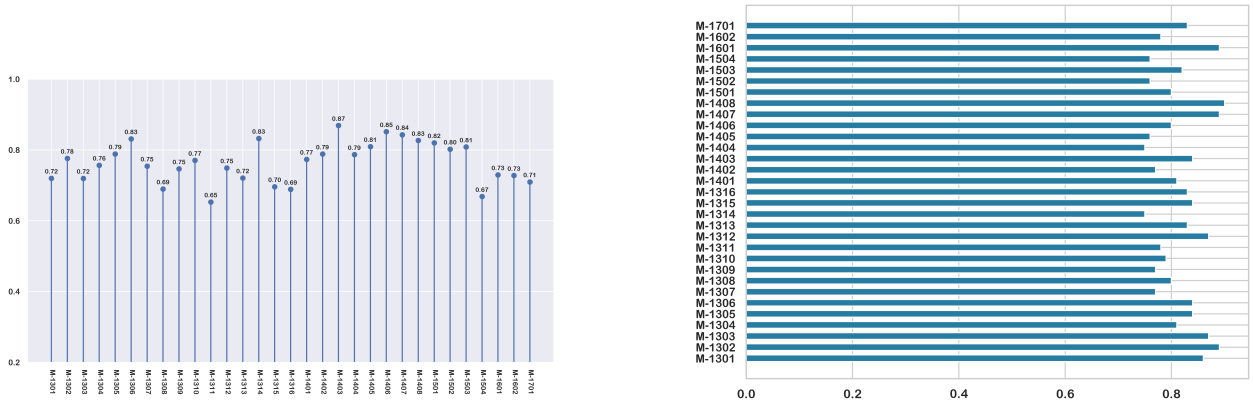
Emphasize Service Games: We found that the service game winning rate has a significant impact on match outcomes. In one's own service games, it is recommended to maximize the first-serve success rate, contributing to a quicker victory over the opponent.

Effort in Saving Break Points: Results indicate that a higher success rate in saving break points further increases the likelihood of winning the match. In situations where the player is at a disadvantage, it is crucial to stabilize break point defenses, as this represents an opportunity for a player to turn the tide.

8 Task 4: Evaluate Model Performance and Eneralizability

8.1 Predictive Capability

To further demonstrate the model's performance, we analyzed 31 matches using two metrics (accuracy and recall). The results are shown in Figure 13.



(a) Accuracy of model prediction

(b) Recall of model prediction

Figure 13: Evaluation of model prediction

- **Accuracy:** Observing Figure 13a, it is noted that the highest accuracy was achieved in the match M-1503 (match_id: 2023-wimbledon-1503) at **92.1%**. The lowest accuracy still reached **72.6%**, indicating that the model is highly accurate in making judgments about swings.
- **Recall:** Observing Figure 13b, it is found that the highest recall was achieved in the match M-1408 (match_id: 2023-wimbledon-1408) at **95.4%**. The lowest recall also reached **75.0%**, showcasing the model's ability to identify as many instances of swings as possible.

8.1.1 Generalization Performance in Women's Tennis Competitions

To further demonstrate the model's generalization capability, we selected data related to the Federation Cup, also known as the Fed Cup [6].

Comparing different rules, we adjusted the model, ultimately obtaining the model's prediction results on this dataset (Figure 14).

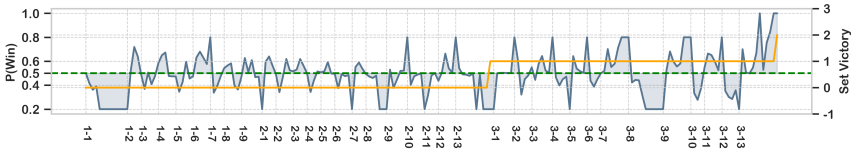


Figure 14: A match level prediction for Wickmayer, Yanina (BEL)

Here, we predict the probability of victory for a player at the match level. The yellow line represents the winning outcome for this player (2-1). Referencing the yellow line, we can analyze each of the three sets individually.

1. **First set:** Initially, the probability of winning was low (0.2), but from the second game onwards, the probability of winning increased, yet fluctuated around 0.5.

2. **Second set:** The early part largely continued the trend of the first set, with a wider swing in the probability of winning in the latter half.
3. **Third set:** Although the final judgment on the outcome was correct, there was the significant swing in the probability of winning throughout the set, especially towards the end of the match.

In summary, analyzing from a match-level perspective reveals that, although our model's final judgment on the match outcome was correct, the prediction process was unstable. We believe the instability can be attributed to the following reasons:

1. Insufficient dataset size
2. Lack of further data cleaning
3. Changes in certain conditions (player gender, court type, etc.)

8.2 Model Improvements

8.2.1 Consideration of Other Factors

- **Intrinsic differences between serving and receiving.**

This discrepancy arises because players have a greater advantage while serving. Therefore, we introduce a variable to adjust for the intrinsic difference between serving and receiving: e.

$$\begin{cases} p^{'} = p + e \\ q^{'} = q + e \end{cases} \tag{30}$$

- **The Impact of Court Surface.**

The type of court surface (hard, grass, clay, and carpet) can significantly influence the characteristics of a tennis match. The primary difference between these surface types is their hardness, which affects the power and intensity of serves. The faster the court surfaces, the higher the probability of a player winning a point on their serve.

Therefore, an adjustment to the probability is made: $p' = p + p_\alpha$

8.2.2 Universality

1. **Difference:** Tennis competitions include singles, doubles, and team events. Doubles and team events require a comprehensive consideration of players' performance and momentum, as well as the physical differences between men and women in mixed doubles.

2. **Specificity:** The Wimbledon Championships is the only Grand Slam event played on grass courts among the four major tournaments. Different court types can influence the probability p through p_α, thereby adjusting the probability of a player winning a game.

9 Sensitivity and Robustness Analysis

In this section, we analyze the sensitivity of the prediction accuracy for each game with respect to the smoothing coefficient and the robustness concerning the dynamic factor.

9.1 Sensitivity Analysis

We choose different smoothing coefficients α as influencing indicators, with the number of matches as the horizontal axis and the accuracy of match predictions as the vertical axis. This allows us to observe the benefits of the model under different smoothing coefficients are shown in Figure 15.

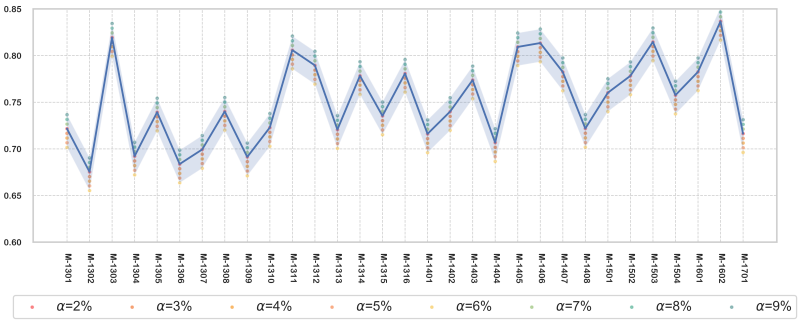


Figure 15: The accuracy of the model under different smoothing coefficients

From the figure, it can be observed that under the influence of different smoothing coefficients, the prediction accuracy for each match exhibits slight swings. Moreover, with a larger smoothing coefficient, the relative prediction accuracy tends to be higher. This is primarily because a larger smoothing coefficient considers a more extensive history of match results, resulting in predictions that better align with actual outcomes. Therefore, it can be inferred that the model demonstrates a higher level of sensitivity.

9.2 Robustness Analysis

We choose different dynamic factors β as influencing indicators, with the percentage swing in prediction accuracy as the horizontal axis and different numbers of match occurrences as the vertical axis. This allows us to observe the benefits of the model under different dynamic factors are shown in Figure 16.

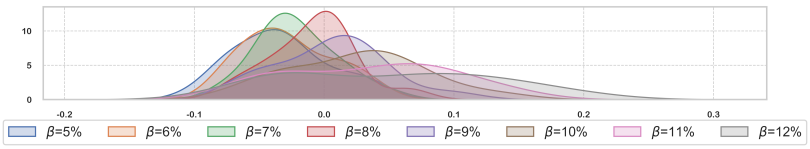


Figure 16: The percentage swing in prediction accuracy of the model under different dynamic factors

From the figure, it can be observed that under the influence of different dynamic factors, the percentage swing in prediction accuracy for each match remains within the range of -10% to 10%. The average swing is centered around 0. Therefore, it can be concluded that the model exhibits excellent stability regarding the fluctuating factors in actual matches. This validates the robustness of the proposed model.

10 Strengths and Weaknesses

10.1 Strengths

1. **Integration of Dynamic and Static Metrics:** By categorizing metrics and applying a moving average weighted dynamic metric to static metrics, the overall evaluation of a player's performance is more accurate.
2. **Analysis of the Correlation Between Momentum swings and Performance swings:** Through the extraction and amplification of swing characteristics, the correlation strength and sequence of swings between momentum and performance are precisely captured.
3. **Mechanistic Analysis of Match Probabilities:** From a statistical perspective, using the Markov model to suggest state transition matrices for match states allows for mechanistic analysis of the match process, yielding more rational and accurate results.
4. **Real-time Update of Winning Probabilities Influenced by Momentum:** Through the swing factor metric, SVM outputs are updated in real time to predict the winning probability of the next point. Considering the impact of momentum makes the model more reasonable and complete.
5. **Universality:** After adjustments and improvements, the model can be applied to any tennis match or other ball sports, with high accuracy.

10.2 Weaknesses and Improvement

10.2.1 Weaknesses

1. **Numerous Factors Influencing Player Performance and Swings:** Such as the number of matches a player has participated in under similar conditions (e.g. weather), the number of matches played by the player in the days leading up to the match, and the importance of the match, which are not considered in the model.

10.2.2 Improvement

1. **Adjustment Factors:** Incorporating factors such as the importance of the match and the frequency of the player's participation in matches into the model, adding adjustment factors to reflect real situations better, can increase the accuracy of the results.

2. **Hyperparameter Selection:** Machine learning models typically require hyperparameters, and the process of obtaining optimal hyperparameters is often empirical. Choosing better hyperparameters can improve the precision of prediction algorithms.

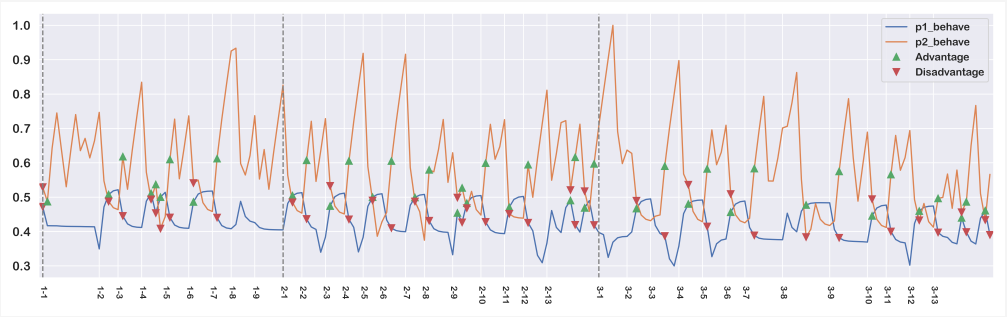
11 Memorandum

To: The Trader
From: Team# 2423183
Date: February 6th, 2024
Subject: Momentum analysis – a integrated and holistic approach

Dear Sir or Madam,

Based on recent research on momentum, to analyze the effect of momentum, we quantified momentum from two perspectives: psychological momentum and behavioral momentum, and conducted a detailed analysis of momentum at three levels: point, game, and set.

First, it is essential to note the causal relationship between momentum and player performance. To uncover this relationship, we selected 21 metrics from both dynamic and static perspectives to evaluate player performance and determined through time series correlation analysis that the swing between the two are positively correlated. In the figure below, after the green key points, momentum increases; after the red key points, momentum decreases. Hence, momentum plays a key role in the swings of the match, and players should seize momentum at critical scoring points to enhance their performance.



Secondly, we developed a probability updatable markov chain (PUMC) based on SVM to predict swing during the match, ultimately achieving a predictive accuracy of 76.5% (with the highest accuracy reaching 92.1%). Through interpretability analysis, we successfully identified four dominant fluctuation factors: the success rate of the second serve, break point save rate, the success rate of the first serve, and the scoring rate on serve.

Lastly, based on the above analysis, we present the following constructive advice:

***Tip* 1: Self-Adjustment.** Given the momentum often determines a player's specific performance in the current game. Players are advised to forget negative outcomes from past matches, adjust their mental state, and fully unleash their capabilities on the court.

***Tip* 2: Focus on Service Games.** We observed that the win rate of service games significantly impacts the match outcome. In one's service games, it is recommended to try to increase the first serve success rate, which helps to defeat the opponent more quickly.

***Tip* 3: Strive to Save Break Points.** Results show that improving the success rate of saving break points further increases the likelihood of winning matches. In disadvantageous situations, stabilizing break point defense is crucial, as it represents an opportunity for players to turn the situation around.

References

[1] Frank Klinker. Exponential moving average versus moving exponential average. *Mathematische Semesterberichte*, 58:97–107, 2011.

[2] Robert Seidl and Patrick Lucey. Live counter-factual analysis in women's tennis using automatic key-moment detection.

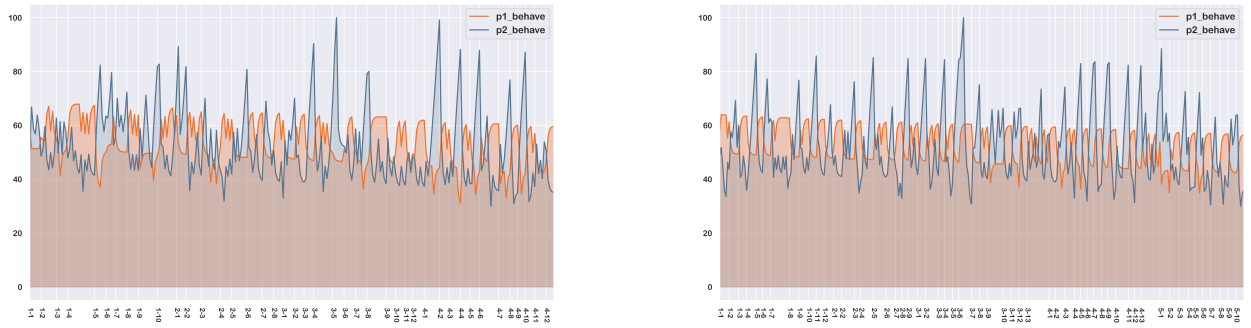
[3] Walid Briki. Rethinking the relationship between momentum and sport performance: Toward an integrative perspective. *Psychology of Sport and Exercise*, 30:38–44, 2017.

[4] Mariusz Maziarz. A review of the granger-causality fallacy. *The journal of philosophical economics: Reflections on economic and social issues*, 8(2):86–105, 2015.

[5] Yanan Zhou, Wei Wu, Huan Wang, Xin Zhang, Chao Yang, and Hongbin Liu. Identification of soil texture classes under vegetation cover based on sentinel-2 data with svm and shap techniques. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:3758–3770, 2022.

[6] Federation cup (fed cup), 2024. Accessed: 2024-02-05.

Appendices



(a) Comeback Set (b) Rally Set

Figure 18: Flow of rest two match situations

Symbol	Explanation	Symbol	Explanation	Symbol	Explanation	Symbol	Explanation
DFR	double fault rate	AR	ace rate	FRSR	first return score rate	SS	sequential score
FSR	first serve rate	BR	break rate	SRSR	second return score rate	S	serving
SSER	second serve rate	RR	return rate	SSCR	serve score rate	BS	break score
SBR	save break rate	FSSR	first serve score rate	RSR	return score rate	SLS	sequential loss score
SER	serve rate	SSSR	second serve score rate	SCR	score rate	EE	energy expenditure
						FA	fault

Figure 19: Explanation of static metrics And dynamic metrics