Problem Chosen	2024 MCM/ICM Summary Sheet	Team Control Number
C		2418588

The Power of Momentum: How to Win the Match with Data

Summary

Djokovic and Alcaraz at Wimbledon 2023 gave us an exciting match with ups and downs, which prompted us to think: is it possible to quantify the performance of the players in the match and thus infer the process of the development of the match? In response to this question, we launched an in-depth study.

For **Task 1**, we constructed a model based on the **IG-LSTM network** architecture to measure players' **momentum** level and match development: **Player Momentum Model (PMM)**. The features were filtered by an information gain-based decision tree and combined into three secondary metrics: **Scoring Performance (SP)**, **Fatigue Level (FL)** and **Mental State (MS)**, which were inputted into the neural network for training. The final neural network output matches the real competition situation, indicating that PMM can better measure the players' momentum and visualize the competition situation by analyzing the key data.

For **Task 2**, in order to verify the close relationship between players' momentum and the match situation, we use **Correlation Analysis** and **Stochastic Simulation** to verify the relationship. In Correlation Analysis, we calculate player' s momentum and the **Fitting error rate**, **Pearson Correlation Coefficient** and **Spearman's rank correlation coefficient** of the match situation through the data in the dataset. In the final match between Djokovic and Alcaraz, the three metrics were: **0.0451, 0.9267, 0.9571**. In Stochastic Simulation, based on the mean and variance of the data in real matches, we randomly generated 5000 items of data and match results, and the difference between the calculated momentum and the randomized match results is vastly different, with the above 3 metrics being: **0.7446, 0.1283, 0.1631** respectively. therefore, the swings in play are not randomized.

For **Task 3**, we extracted the key data points of momentum changes for analysis, applied the **decision tree based on information entropy** to filter the important features, and constructed an ANN neural network model **Momentum Swings Prediction Model (MSPM)** to analyze it.The accuracy of MSPM in predicting swings of match on the training set reached **94.8%**. We formulated the strategy of the match for the players who were about to compete.

For **Task 4**, we used the MSPM in Task 3 to test the 5 matches in the test set, and the average prediction accuracy obtained was: **90.70%**. In order to further improve the performance of the model, we added more parameters such as audience inference, weather condition, court surface, etc. We re-trained and re-tested the optimized **MSPM-Pro**, and obtained an average prediction accuracy of **96.64%**. Further more, we adjusted different parameters to enable the MSPM-Pro model to analyze more types of matches, and the model performed well in predicting tennis matches, but was slightly less accurate in predicting other types of matches, such as table tennis and badminton.

Memo in **Task 5** covers analysis results and model usage, as well as specific recommendations customized for coaches and players.

Keywords: PMM; IG-LSTM network; Decision tree; MMT; MSPM.

Contents

1 Introduction

1.1 Problem Background

In the final of the men's singles tournament at the 2023 Wimbledon Open, Novak Djokovic and Alcaraz gave us an incredible up-and-down match. In order to quantify the changes in the match situation, we would like to use a model that simulates the match trend and predicts the swing of play by using the data observed during the match, providing targeted data to support players to improve their performance in the match.

1.2 Problem Restatement

Through in-depth analysis and research on the background of the problem, combined with the specific constraints given, the restate of the problem can be expressed as follows:

- Judge the performance level of a player at this point in the game by using his or her data during the game. Compare and contrast the performance of different players to generate a model that can simulate the trend of the game.
- Analyze the relationship between momentum and the match situation, and determine whether the momentum of the player changes randomly or is affected by other factors during the match.
- Identify the key factors that influence the change of a player's momentum during a game and construct a model that can predict the swing of play. provide pre-game advice to players based on these key factors.
- Use the model in different matches, including different tennis matches, different types of oppositional sports, test the generalizability of the model for different matches, and think about what metrics should be added to the building blocks of the model to improve the accuracy of the model.
- Generate a one- to two-page memo summarizing the outcomes and offering guidance to coaches regarding the significance of "momentum" and strategies to prepare players for responding to events influencing the flow of play in a tennis match.

1.3 Our Work

In this problem, we need to process and analyze the data of the players in the game to derive the features to measure the performance of the players and the flow of play. Our work mainly includes the following:

1. We constructed a model based on IG-LSTM(Information Gain-LSTM) named Player Momentum Model(PMM), which is used to measure the swings of player's momentum and match trend, and integrates the three indicators of Scoring Performance, Fatigue Level and Mental State.

2. We verified whether the change of match situation is random or not, and analyzed it by both Correlation Analysis and Stochastic Simulation methods, and assessed whether swings in play is related to momentum.

3. We used a decision tree based on information entropy to extract the key features that lead to the occurrence of swing of play, and designed an ANN neural network model: Momentum Swing Prediction Model(MSPM) to predict the changes of the match situation and provided important indicators that coaches are concerned about and practical recommendations for players.

4. We use the MSPM to analyze several other Wimbledon men's matches and test the accuracy of the model. After that, we further optimized the model by adding new features into our model. We constructed a model MSPM-Pro with higher generality, and tested it through more different tournaments of matches and sports.

5. We summarized the data generated by PMM and MSPM(-Pro), and generalized the results of the analysis to give coaches and players suggestions on how to improve momentum and performance in match.

In order to avoid complicated description, intuitively reflect our work process, the flow chart is shown in figure 1 below:

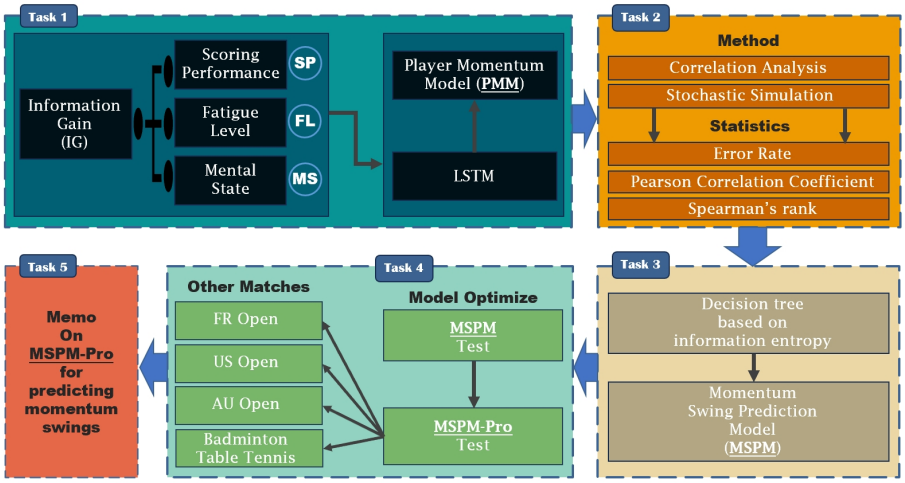


Figure 1: Flow Chart of Our Work

2 Assumptions and Justifications

- **Assumption 1:** The greater the player's momentum, the better their performance in the competition.

  Justification: A player may feel they have the momentum, or "strength" during a match. We aim to use momentum to measure the player's status, assuming that momentum correlates with the player's performance in the game.

- **Assumption 2:** Assuming a greater difference in momentum between two athletes, the one with higher momentum is more likely to win.

  Justification: The greater the difference in momentum between two players, the one with higher momentum will perform better and, therefore, is more likely to achieve victory.

- **Assumption 3:** Disregarding any influence between two matches.

  Justification: A player's momentum in the current match will not be affected by their following matches, and the momentum in the next match will also not be influenced by the result of the previous match.

- **Assumption 4:** Fatigue Level does not naturally recover during non-rest periods.

  Justification: During the competition, we do not consider the player's Fatigue Level recovering unless it occurs within the specified rest periods.

- **Assumption 5:** Disregarding the quality of the venue, facilities, and any unforeseen circumstances affecting the players.

  Justification: Ensuring fairness in the competition beyond the format (court direction etc), we do not consider issues like the venue's quality or unexpected events affecting the players, such as sudden illnesses.

3 Notations

The key mathematical notations used in this paper are listed in table 1.

Table 1: Notations used in this paper

Symbol	Description	Unit(Example)
interruption_of_play	whether play is interrupted	0 or 1
Time_per_score	time spent on each score	second(s)
MMT	momentum	\
SP	Scoring Performance	\
FL	Fatigue Level	\
MS	Mental State	\
NS	Number of Sets won	\
NG	Number of Games won	\
SA	Score Accumulation	\
TC	Time Consumed	second(s)
TD	Total Distance	meter(m)
TS	Total Shot	\
Ace	hit an untouchable winning serve	0 or 1
BPW	Win the game of opponent's serve	0 or 1
UE	Unforced Error	0 or 1
$MMT_i, i = 1, 2$	player i 's momentum, $i = 1, 2$	\
ΔMMT	difference of momentum	\
P	scoring metric	± 1
S	label of momentum Swing	0 or 1

4 Data Pre-processing

4.1 Data format conversion

Due to the fact that the "elapsed_time" field represents the time elapsed since the start of timing for a match, subtracting the values of this field in consecutive lines allows us to obtain the time data for each team to score 1 point, named "Time_per_score". Considering the possibility of interruptions in matches where the data may have an impact on momentum, a new additional field ("interruption_of_play") is introduced to indicate whether there was an interruption before this score. Here, an interruption in the match is defined as a time difference of more than 10 minutes, or 600 seconds, between the start times of the competitions for two consecutive points. If there is an interruption, the value of this field is 1; otherwise, it is 0.

4.2 Missing value replacement

We observed that the "speed_mph" field has missing values. Here we first make a supplementary assumption:

- Wimbledon Tennis Open is a top-notch global tennis event, and the participating players are highly skilled. They can maintain a stable swing and serve speed during a match.

Under this assumption, we calculate the average serve speed of player1 and player2 for each match. If the "speed_mph" field data is missing for a player's serve during a particular match, it should be replaced with the average value of "speed_mph" for that player in the same match. In addition, the "serve_width", "serve_depth" and "return_depth" fields also have missing values. We assign numerical labels to them separately, and replace the missing values with 0. The specific process is shown in the next part.

4.3 Numerical conversion of professional terms

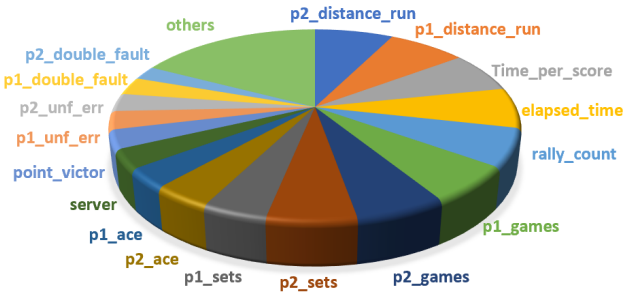
We will sequentially replace the valid values of the "serve_width," "serve_depth," and "return_depth" fields with positive integers, converting the mentioned four fields from float to int. Specifically, we replace the values of "serve_width" B, BC, BW, C, W with 1, 2, 3, 4, 5 respectively; CTL, NCTL with 1, 2; and D, ND with 1, 2.

5 Task 1: Player Momentum Model(PPM)

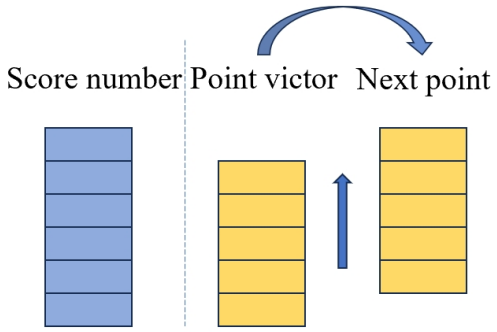
5.1 Data processing and feature screening

As Merriam-Webster explains, momentum means "strength or force gained by motion or by a series of events". In this question, the momentum of a tennis player on the court at a certain moment is constituted by factors such as their skill level, physical condition, and other relevant data, which are reflected in the provided CSV file. However, due to the diverse and detailed nature of these features, they are not directly suitable for representing the momentum of a player at a specific moment. Therefore, we have conducted an initial screening of the provided features.

Events occurring in a match have a temporal aspect, involving the progression of time. Neural networks excel in handling time-related data and capturing complex temporal relationships. Therefore, using a neural network can better simulate the changes in momentum during a match. The model we employ is the "Long Short-Term Memory Neural Network (LSTM)" model selected through information gain filtering.



(a) Feature Importance



(b) Data column shifting up

Firstly, using the information gain of different features, calculate the importance of each feature in influencing momentum. For a specific feature, such as x_1, its information gain $IG(x_1)$ can be expressed using the following formula:

$$IG(x_1) = E(B) - \sum_j \frac{n_j}{n} E(A) \tag{1}$$

Here, n represents the total number of samples in the parent node B, and n_j is the number of samples in the child node A. E represents information entropy, which can be further expressed as:

$$E(t) = -\sum_{i=1}^{c} p(i,t) \cdot log_2(p(i,t)) \tag{2}$$

Where$p(i,t)$ represents the proportion of samples belonging to class i in node t.

The model we have established is the **Player Momentum Model (PMM)**, with a particularly important feature:

△ We consider the victor of the next point within the same match as the current point's feature. Specifically, we shift the "point_victor" column in the dataset one row up according to the match instances. This shifted column serves as the training target for the momentum (MMT) function in machine learning.

This distinctive approach will be continuously applied in the specific construction outlined in the following sections, so it is specifically highlighted here, as shown in the above figure.

5.2 Feature construction and momentum function

Utilizing a decision tree based on information gain for feature selection, we obtained the ranking of the impact of 46 original features. The hierarchical clustering plot below provides a more intuitive

representation. Based on this result, we selected features with higher importance and practical significance. On this basis, we constructed 3 new effective features: **Scoring Performance**, **Fatigue Level**, and **Mental State**.

Considering different impacts of the indexes and correlation between indexes, We further utilize weighted parameter α to measure the impact of these three secondary indicators. Based on this, the general expression for momentum is as follows:

Figure 3: PMM overview

5.2.1 Feature 1: Scoring Performance(SP)

The first feature is the player's scoring performance(SP). The player's scoring situation is the most direct representation of their advantage in the current game, which is a crucial constituent feature of momentum. Simultaneously, it is also the most intuitive reflection in the data. In the provided dataset, the fields 'p1_sets,' 'p2_sets,' 'p1_games,' and 'p2_games' respectively indicate the number of sets and games won by player 1 and player 2 in the current match. Additionally, cumulative scoring data for the players is available. The equation for scoring performance (SP) is as follows:

$$SP = \beta_{11} \cdot NS + \beta_{12} \cdot NG + \beta_{13} \cdot SA \tag{3}$$

In this context, $beta_{1j}(j = 1, 2, 3)$ represents the weights assigned to the player's number of sets won (NS), number of games won (NG), and score accumulation (SA), respectively.

Generally, a significant lead in points is more indicative of superior scoring performance than a marginal advantage. Therefore, the corresponding weights are also larger. Additionally, according to the notes from question one, the server is more likely to win the point. We will reflect this serving advantage in the changes to score accumulation. The specific rules are as follows:

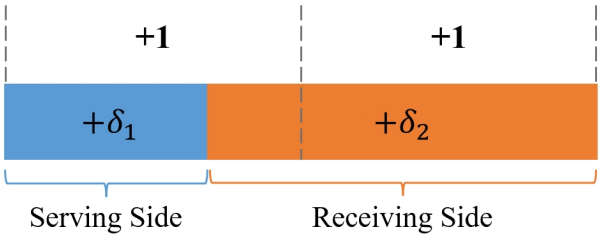


Figure 4: Rule of score accumulation

If the player is the server and wins the point, the cumulative score increases by δ_1 (where $0.5 < \delta_1 < 1$); if the player is the server but loses the point, the opponent's cumulative score increases by δ_2 (where $1 < \delta_2 < 1.5$).

5.2.2 Feature 2: Fatigue Level(FL)

The second feature is the player's fatigue level (FL). After a period of game, the player's physical capabilities tend to diminish, significantly impacting their momentum. Specifically, fatigue level can be reflected through the total match time (Time Consumed), the player's total distance covered (Total Distance), and the total number of shots taken (Total Shots). It can be observed that as the match time increases, the player's total distance and the total shots also increase, contributing to a higher fatigue level(TF). Here, we primarily consider the impact of these three features on the player's fatigue level.

Firstly, further processing is conducted on the cleaned data. The elapsed_time includes the total time elapsed from the start of the match and the rest periods. As the glossary of key terms/concepts, there is a 90-second break after the 3rd, 5th, 7th... games, and at least a 120-second break after each set. Therefore, we subtract 90 seconds and 120 seconds at these corresponding points to obtain the total consumed(TC). Moreover, during mid-game breaks, the player's fatigue level also undergoes some recovery. After a 90-second break, fatigue is alleviated by γ_1, and after a 120-second break, it is alleviated by γ_2. For example, after the first 90-second break, the time consumed(TS) is alleviated from the original t to $(1-\gamma_1)t$.

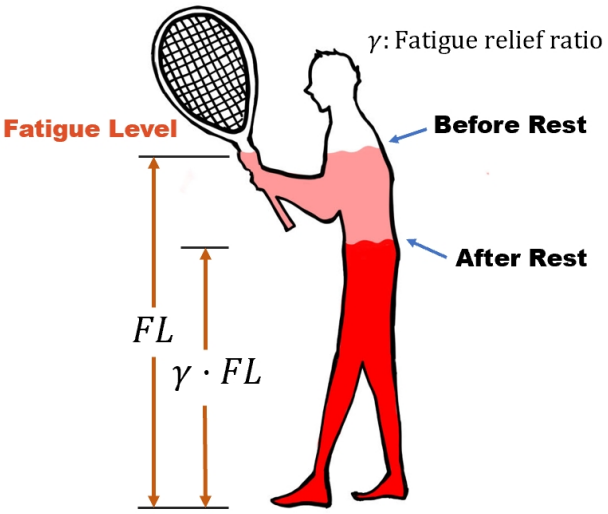


Figure 5: Changes in fatigue level during rests between games

Considering that it is difficult for fatigue to naturally recover during the course of the game, we assume that the player's fatigue level does not naturally recover within a match. Therefore, when considering the distance covered and the number of shots, we need to take into account the distance covered and the number of shots in each set that has already been played in the match. Consequently, we perform a cumulative sum on the original 'distance_run' and 'rally_count' field. We also introduce the total distance covered (Total Distance) and the total number of shots (Total Shots), which can be alleviated through rest. After data processing, we can obtain the time consumed (TC), the total distance (TD), and the total shots (TS) for each point played.

Next, we proceed with normalization. Given that data such as time consumed (TC) tends to be relatively large, we utilize the sigmoid function as a normalization tool. Consequently, the expression for fatigue level (FL) is as follows:

$$FL = \beta_{21} \cdot TC + \beta_{22} \cdot TD + \beta_{23} \cdot TS \tag{4}$$

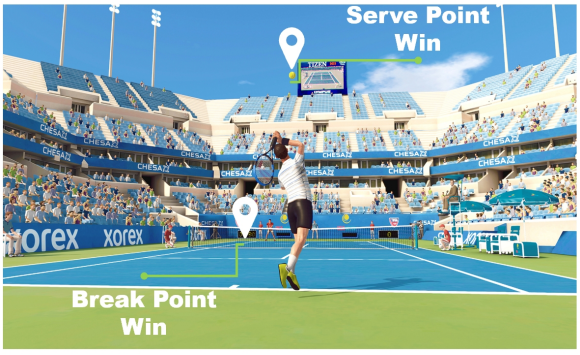
where $\beta_{2j}, j = 1, 2, 3$ represent the weights assigned to time consumed (TC), total distance (TD), and total shots (TS) respectively.

5.2.3 Feature 3: Mental State(MS)

The third feature is the player's current mental state(MS). During a match, there are numerous factors that can influence the player's mindset. Factors such as the server's success in the first serve, scoring on the serve, the receiver's success in breaking serve, and whether both players commit double faults or unforced errors are all reflected in the provided dataset. Additionally, other factors such as the weather conditions on the day, court comfort, audience interference, etc., can also to some extent impact the player's mental state.

For better understanding, let's briefly analyze the 2023 Wimbledon Championships final mentioned in this question. Grand Slam winner Novak Djokovic is renowned for his defensive counterattacks, known for quick defensive reactions. Among the four major tennis opens worldwide, the US Open and the Australian Open are played on hard courts, while the Wimbledon Championships are held on grass courts. In comparison, grass courts have a lower friction coefficient and are less suitable for rapid player movements.

(a) Features affecting Mental State 1



(b) Features affecting Mental State 2

Data indicates that in this particular match, Djokovic covered a relatively short distance, suggesting less running on the court. However, there were as many as 40 unforced errors. In this match, he failed to capitalize on his defensive strengths, committing numerous errors, and criticism from the audience also affected his mental state during the match.

In the model for Problem 1, we will first analyze only the factors that are in the provided dataset; other factors, such as environmental factors, will be discussed later. A serve point reflected in the "p1_ace"will boost the morale of the player; on the contrary, if the player is broken by the opponent, the morale of the player will fall. Comparatively speaking, it is normal for players to make errors, and therefore the mental state of the player is less affected. Thus, the equation for the player's current mental state (MS) is as follows:

$$MS = \eta \cdot (\beta_{31} \cdot Ace + \beta_{32} \cdot BPW - \beta_{33} \cdot UE) \tag{5}$$

Overall, the above 3 features cover more than a dozen important fields of metrics in the dataset, reflecting player's momentum from different aspects, so we eventually generated expression for momentum(MMT):

$$MMT = f(SP, FL, MS) \tag{6}$$

5.3 Calculation and result analysis

We input the filtered features as well as the three newly constructed features into the LSTM model. Since the match is an ongoing process and earlier data can still have an impact on later results, the LSTM is able to capture and handle this long-term dependency and thus accurately simulate what happens in different matches.

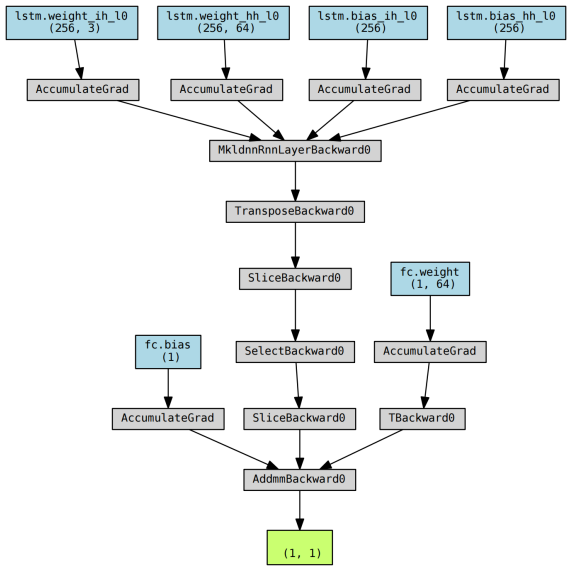


Figure 7: LSTM model feedback layer

At the same time, the neural network can help to determine the weighting coefficients of each item in the momentum equation we defined, and obtain the momentum function that is compatible with the real scoring situation and the match results. The LSTM model involves forgetting and updating the information, which is handled as follows.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f), \quad i_t = \sigma \cdot [h_{t-1}, x_t] + b_i) \tag{7}$$

$$\tilde{C}_t = tanh(W_C \cdot [h_{t-1}, x_t] + b_C), \quad o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{8}$$

$$C_t = f_t \circ C_{t-1} + i_l \circ \tilde{C}_t, \quad h_t = o_t \circ tanh(C_t) \tag{9}$$

Through the gating mechanism, the LSTM model updates our momentum function continuously in chronological order, and the model is trained to obtain the weight coefficients of each feature. At this point the model construction is complete.

Next, we substituted the real game situation data into our Player Momentum Model (PMM), and obtained the momentum changes of player 1 and player 2 in the scoring process of each game. Due to the large amount of data and space limitation, we have chosen 4 representative matches to plot the momentum curves. The dark green curve represents the momentum of player 1, while the light green curve represents player 2.

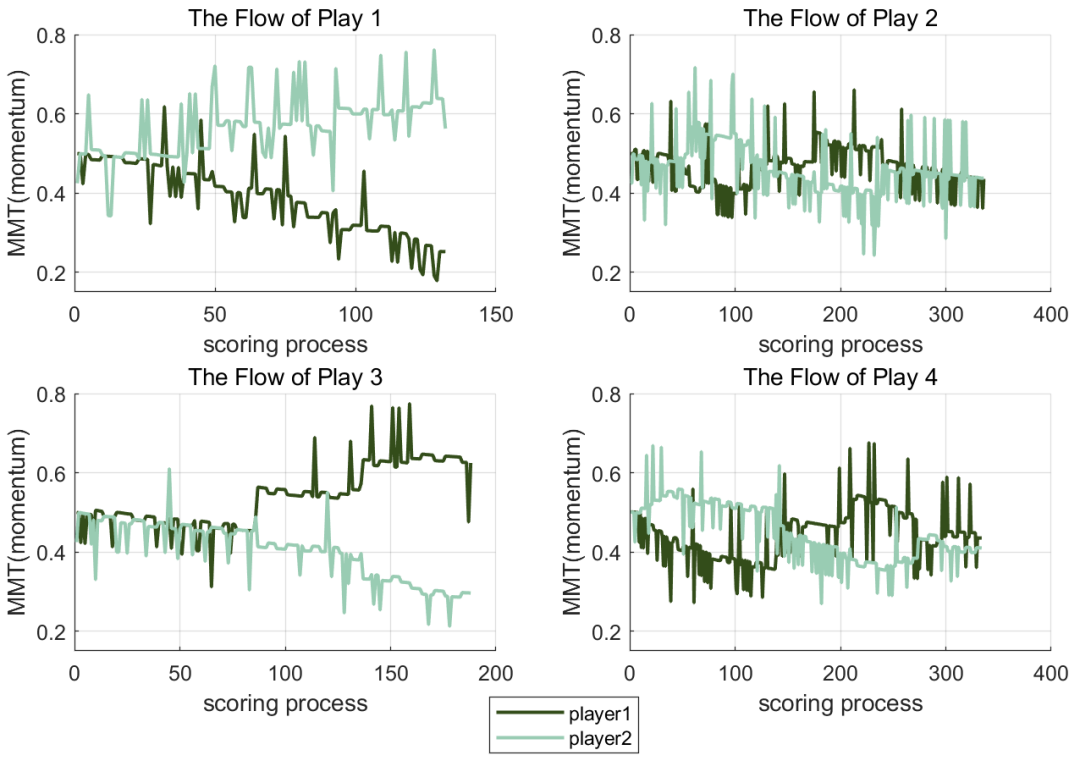


Figure 8: Momentum curve for player 1 and player 2 in 4 different games

In order to verify the accuracy of the PMM, we intercepted several segments of the momentum graph and added real scores, based on the "point_victor" field data in the original features, to the graph for comparative analysis.

As shown in figure 9, most of the momentum data points are in the same half of the axis as the true score data points. Therefore, we can conclude that the player with the higher momentum basically won the ball. This is consistent with our hypothesis and shows the exact scoring process of the match.

6 Task 2: Assess the claim of random momentum

The PMM model in Question 1 depicts the flow of the match with the help of momentum curves. We do not believe that the coach's claim that changes in the momentum of the players during a match are random is reasonable. Instead, changes in momentum are related to events and characteristics of the game and reflect the overall winning trend of the game. In the following, we use both correlation analysis and stochastic simulation to verify the accuracy of the idea.

In both of these models, we will use the calculation of **Pearson Correlation Coefficient** and **Spearman's rank correlation coefficient**, and the calculation method for Pearson Correlation Coefficient and Spearman's rank correlation coefficient is provided below in advance.

Pearson Correlation Coefficient:

$$\rho = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{10}$$

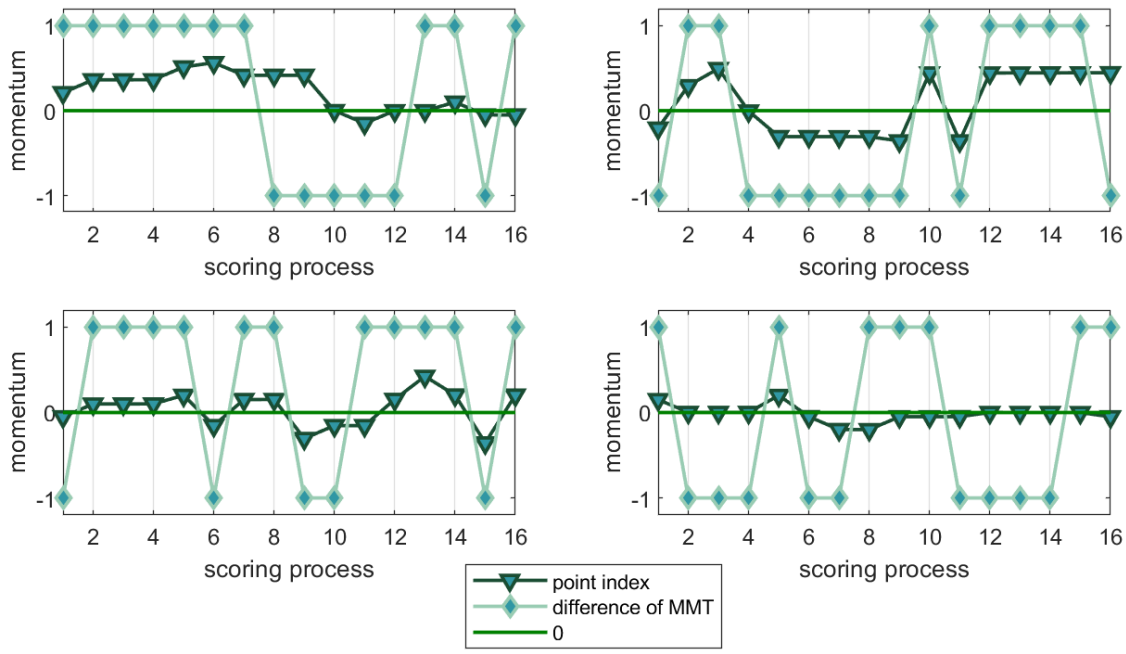


Figure 9: Comparison of momentum(MMT) and point index

In the formula, cov represents the covariance of X and Y, σ_X and σ_Y represents the standard deviation, which can be further calculated by the following formula:

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}, \quad \sigma_X = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}, \quad \sigma_Y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}} \tag{11}$$

Spearman's rank correlation coefficient:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{12}$$

6.1 Method 1: Correlation Analysis

In order to study whether there is a correlation between momentum and winning or losing a match, i.e., whether a player has a higher winning rate when his momentum is higher, we analyze whether the two players have higher or lower momentum and whether they can win the match. For this purpose, we introduce a new indicator - momentum difference ΔMMT to describe the level of momentum of these two players, and the momentum difference satisfies the following equation:

$$\Delta MMT = MMT_2 - MMT_1 \tag{13}$$

The scoring metric P is also introduced to describe the player's win or loss, with a value of -1 when player 1 wins and 1 when player 2 wins.There are a total of four combinations of ΔMMT and P, as shown in the following figure 10.

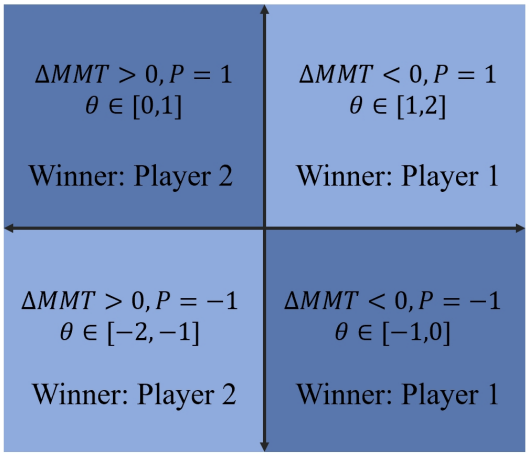


Figure 10: Model 1 mechanism overview

When $P = 1, \Delta MMT > 0$, player 2 has higher momentum and wins the game, coinciding; when $P = -1, \Delta\ MMT < 0$, player 1 has lower momentum but wins the game, not coinciding. For the purpose of the following discussion, we introduce one more indicator θ, which satisfies:

$$P - \Delta MMT = \theta \tag{14}$$

From the data in the table: it can be noted that when the range of θ is controlled in $[-1, 1]$, the momentum high and low coincides with the result of the game; when θ is not in $[-1, 1]$, the momentum high and low does not coincide with the result of the match. Therefore, we consider to use a segmented function $h(\theta)$ to further simplify the data, $h(\theta)$ is satisfied:

$$h(\theta) = \begin{cases} 0, & -1 \le \theta \le 1 \\ 1, & otherwise \end{cases} \tag{15}$$

It is called the Validation function. $h(\theta) = 0$ when the momentum high and low coincide with the result of the match, while $h(\theta) = 1$ does not coincide, so $h(\theta)$ of each point is summed up, that is, to obtain the $\sum_{i=1}^{N} h(\theta_i)$, representing the number of points that do not coincide, and finally, considering its percentage in the total number of points , which is denoted as the error rate ∇err, and finally obtains:

$$\nabla err = \frac{\sum_{i=1}^{N} h(\theta_i)}{N} \tag{16}$$

When the value of ∇err is lower, it means that there are more cases in which the momentum high and low coincide with the results of the match, indicating that the correlation between momentum and the win or loss of the match is stronger; on the contrary, when the value of ∇err is higher, it means that there are more cases in which the momentum high and low do not coincide with the results of the match, indicating that the weak correlation with the match winners and losers. After calculation, the correlation and fitting error rate are shown in the following table 2.

It can be found that our PMM error rate is low and the momentum function can match the real race situation well. This indicates that the change of player's momentum in the game is closely related

Table 2: Error rate and correlation coefficient of Method 1

match ID	Fitting error rate	correlation coefficient (Pearson)	correlation coefficient (Spearman)
1314	0.0532	0.9126	0.9452
1408	0.0428	0.9351	0.9661
1503	0.0725	0.9077	0.9336
1602	0.0561	0.9115	0.9412
1701	0.0451	0.9267	0.9571

to the indicators in the match, and has a strong correlation with the win or loss of the match. Therefore, the change of momentum is not random but regular.

6.2 Method 2: Stochastic Simulation

The idea of this method is that based on the mean and variance of the three effective features, SP,FL and MS, using the computer to simulate the match randomly and generate the corresponding feature values and score situation. The simulation results are then substituted into our PMM model to obtain the momentum based on the randomized match scenario and fitted to the randomized simulated win/loss scenario. Finally, this fitting result is compared with the fitting result in Problem 1 to determine the difference in fitting effect.

To make the analysis more concise and clear, we show the final match between Djokovic and Alcaraz ($match\_id = 1701$) as a case study. First, the mean and variance of the 3 features of this match, SP, FL and MS are calculated and shown in the table 3 below.

Table 3: The mean and variance of the 3 features SP, FL and MS

Statistic	Scoring Performance (p1)	Scoring Performance (p2)	Fatigue Level (p1)	Fatigue Level (p2)	Mental State (p1)	Mental State (p2)
Mean	0.51203	0.48797	0.35684	0.35703	0.49342	0.48767
Variance	0.02334	0.02335	0.00531	0.00528	0.02548	0.02476

The purpose of calculating the mean and variance is to allow the computer simulation to be based on the true level of play of both sides of the game and to prevent extremes. After entering the data into the simulation program, we simulated $5,000$ goals, with a random win/loss scenario for each goal. Next, these simulations were added to PMM and momentum was generated. Similar to Task 1, we also fitted the momentum data to the simulated score data to obtain the fit error rate and correlation coefficient. The fit of the randomized simulation to the real game is shown below.

As can be seen from the table, the momentum generated by the random simulation fitted the win/loss scenarios poorly and much less well than the real matches. Therefore, the coach's claim that the flow of the play and win/loss bias is random is unreasonable.

Table 4: Error rate and correlation coefficient of Method 2

Fitting mode	Fitting error rate	correlation coefficient (Pearson)	correlation coefficient (Spearman)
Real match	0.0821	0.9109	0.9033
Stochastic simulation	0.7446	0.1283	0.1631

7 Task 3: Momentum Swings Prediction Model(MSPM)

The PMM model successfully translates the in-game metrics into a specific function that quantifies the favorable situation of the match, and Task 3 will identify the extremes of the momentum function and the influencing factors behind the change in monotonicity.

It is important to emphasize that since the analysis started from Task 2, we have unified the momentum of players into the momentum of the match itself. Therefore, to address Task 3 concerning the identification and prediction of momentum swings, what we have established is the **Momentum Swings Prediction Model (MSPM)**.

7.1 Indicator selecting and model construction

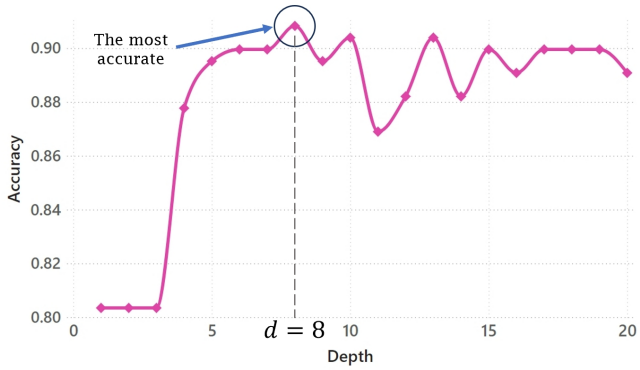


Figure 11: Calculation of the depth of decision tree

In order to determine which factors may cause a swing in momentum, we focus on those matches in which a swing in momentum occurs. In order to quantify the swing in momentum of a match, we define the variable S. Here we still use the momentum difference ΔMMT as defined in Task 2, where $S = 1$ when ΔMMT changes from a negative to a positive value, and vice versa $S = -1$.

From the definition, we know that a change in the value of S indicates a swing in the momentum of play: when the value of S is 1, the advantage of play is shifted from player 1 to player 2; when $S = -1$, the flow of play change from favoring player 2 to player 1. Next, we extract these swings of play for the next screening step.

Here we use a decision tree model based on information entropy to filter out the features that have a greater impact on momentum swings. We adjust the depth of the decision tree. We set the tree depth from 1 to 20, and calculate the training accuracy of different tree depth. The results are shown above in figure 11:

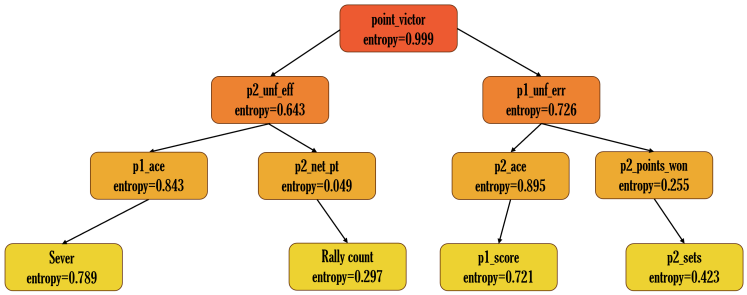


Figure 12: The top 4 layers of the decision tree

Obviously, the best depth of decision tree is 8. So we set the tree depth to 8, and compile our decision tree script again to select the essential features contributing to the swings of play. The screened features are in figure 12.

These factors in the decision tree are the essential indicators that contribute to judging the swings of momentum. In order to be able to predict the swings of the match trend by these features, we decided to design a neural network to make the prediction.

There are no more than three changes in momentum, and all can be be quantified by the variable S:

(1). shifting from 1 to 2; (2). shifting from 2 to 1; (3). staying the same.

Thus, the training task becomes a multi-classification problem.

7.2 Model application and calculation

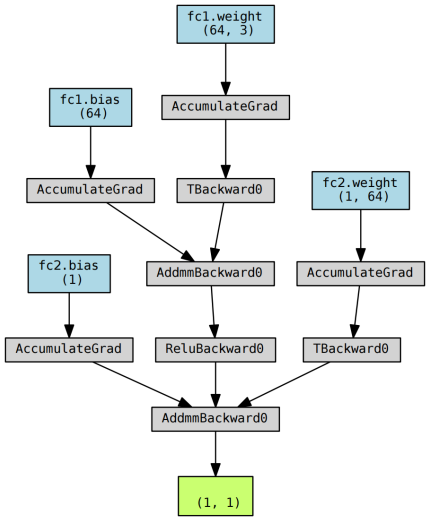


Figure 13: Neural network feedback layer

Before applying the match data to the model for computation, we first use a non-repetitive sampling

method to clearly divide the training and testing sets for the MSPM. The reason for doing this is that, Task 4 requires testing the model on other matches, so the training and testing sets must not overlap. For the testing set, we still use the 5 matches analyzed in the correlation case study from Task 2, with "match_id" being 1314, 1408, 1503, 1602, and 1701.

The brief structure of the neural network we designed and the change in accuracy of the process of choosing Adam as the optimizer for training are shown in the following two figures.

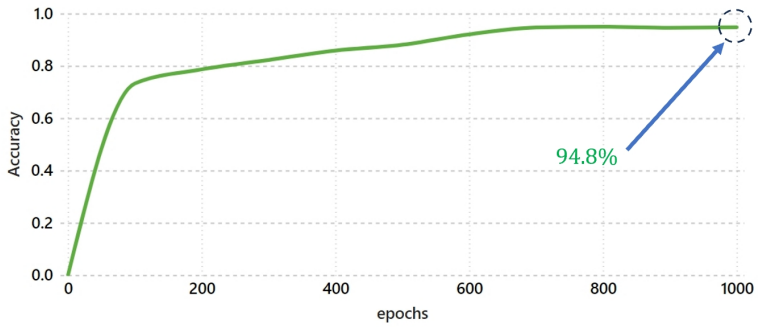


Figure 14: Training accuracy of Momentum Swing Prediction Model

In the end, the neural network was able to achieve a prediction accuracy of 94.8% on the training set. It shows that the filtered features as well as the constructed model can predict the swings in the given match very well.

7.3 Advice based on momentum swings

Our model is able to well connect the indicators and match trends, and found that there is a pattern in the change of momentum of the players. So now we will give the important metrics that coaches are concerned about that affect the momentum of the match based on the previous analysis and provide substantial suggestions for coaches and players.

★Off the field:

Model Fit: Coaches can input data from the player's past matches into our model to get the metrics that play a key role in their momentum swing, and target the weak areas for improvement. For example, if a player's momentum drops after an unforced error in a match, the coach should target to the player's quick reaction, quality of return and defensive counterattack.

Targeted training: In preparation for training, focus on several key indicators such as SP, FL and MS that we give in PMM. For example, Ace scoring has a great effect on the morale of player, and it also keeps players in a good state of mind. Therefore, the coach should strengthen the serve training to improve the success rate and deterrence of the serve.

Strategize: Coaches should analyze players' strengths and next opponent's weaknesses with PMM based on past game data to develop appropriate offensive, defensive and psychological tactics with the aim of reducing the opponent's momentum and moving the match situation in the direction of our side's favor.

★On the Field:

Analyzing Momentum: coaches can generate momentum and favorable trends in the game for their own and opposing players in real time from the sidelines using the MSPM and various in-game

metrics updated in real time. Use the changeover breaks between games to work with players to analyze when the momentum of the match favors their player in matchups that have already taken place, and in which situations the momentum drops back or picks up.

Recognize patterns: Determine if the player's momentum is cyclical, or if he/she performs better on key shots (e.g. break points), etc., to clarify your side's advantage.

Compare opponents: Use MSPM to obtain the fatigue level and current mindset of opposing players, etc., analyze their patterns of momentum changes and weaknesses in certain in-match indicators, and target offense and defense.

Adjustment of strategy: According to the match situation of both sides, improve the stability of key shots, seize the opponent's physical decline and psychological pressure to strengthen the attack and quickly kill the game.

8 Task 4: Testing, Optimization and Generalizing

8.1 Model testing

We used the model in Task 3 to test other matches in the dataset (match_id=1303 1304 1501 1701) to verify the accuracy of the model in other matches. After we input the key features of these matches into the model, we compared the output with the actual matches to get the accuracy of the predicted momentum swing as follows:

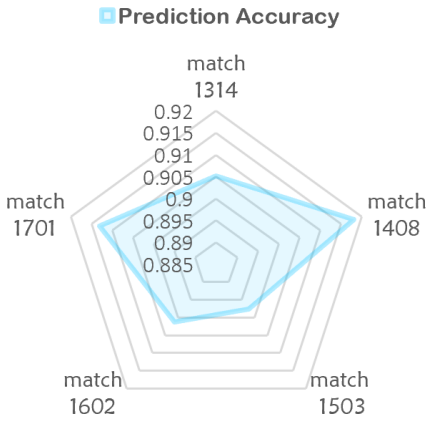


Figure 15: MSPM prediction accuracy radar chart

8.2 Model optimization: MSPM-Pro

In order to further improve the prediction accuracy of the model in predicting different matches, we will consider more factors that can affect momentum swing. Including: audience interference, weather condition, court surface, etc. We collected information about these influencing factors from different sources and quantified them to analyze how they would affect momentum. Due to the similarity of the methodology, we will not elaborate on the process of data cleaning and processing here, but only

present how these metrics are incorporated into the model and the process of MSPM optimization. We refer to the model with the new metrics added and optimized by training and testing as **MSPM-Pro**.

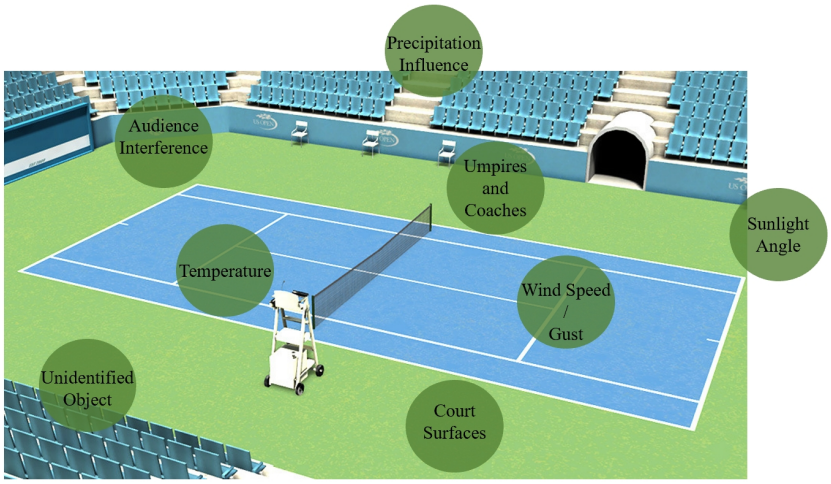


Figure 16: Field simulation and environmental factors

In conjunction with the PMM in Task 1 to improve the indicators affecting momentum, we analyze the direct effect of environmental indicators on SP, FL, and MS.

Audience interference mainly affects MS. Grand Slam tournaments such as Wimbledon, where athletes compete on an individual basis and are required to travel to the respective venue. The audience is usually dominated by local fans, with a few being fans of the athlete. Therefore, when an athlete loses a point or makes an unforced error, the audience may boo the athlete or even throw unknown objects. This can affect the athlete's state of mind.

The weather condition includes precipitation influence, sunlight angle,temperature, wind speed, etc, which may have influence on SP, FL and MS.

If there is rainfall before the match, it may lead to a slippery field, affecting normal play; if there is rainfall in the middle of the match, it may lead to an interruption of the match, suspending the athletes' momentum growth, but at the same time, giving the athletes more sufficient time to rest and adjust their status, reducing their fatigue level FL.

The angle of sunlight also affects the scoring performance of players. Players have a better field of vision in backlit conditions; while when players are facing the sunlight, their eyes may be exposed to direct sunlight during the serving process, leading to serving errors.

Under high temperature conditions, players experience greater physical exertion and a faster rise in fatigue FL.

It can be seen that environmental metrics can influence these characteristics, such as SP, FL and MS, which in turn can counteract the raw metrics.

For example, when a player's MS metrics drop, i.e. the mindset is negatively impacted, there may be a lack of concentration, which can lead to double fault or unforced error, which can reduce the athlete's momentum.

After adding these metrics to the model, we retrained the model and the final output accuracy on the test set is in figure 17.

The comparison shows that after taking these factors into account, the model's prediction accuracy of swing of play in Wimbledon Open has been further improved.

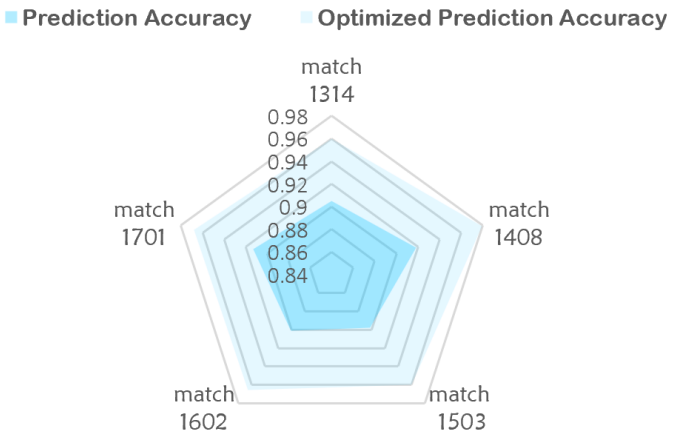


Figure 17: MSPM-Pro prediction accuracy radar chart

8.3 Generality discussion

Next, we test the generality of the model. We apply the model to other types of tournaments including: the French Open men's tournament, the French Open women's tournament, the US Open men's tournament, the US Open women's tournament, the Australian Open men's tournament, as well as badminton and table tennis tournaments.

⋆ Women's matches

In general, men and women differ in physical and emotional functions such as strength and endurance. In tennis, the system of play and the balls used in matches are also different. For example, in Grand Slam tournaments women use a three-set, two-win system, while men use the same five-set, three-win system as in this title. The women's match ball is lighter in weight to suit the women's strength and serve style.

In order to enhance the generalizability and persuasiveness of the model, we include additional player information and other data from several websites, and use the data available therein as model training data. We will show the data sources in Reference to facilitate a coherent narrative in the following sections.

⋆ Other tournaments

Novak Djokovic is a player who has achieved greatness in the Grand Slams, which in addition to Wimbledon include the Australian Open, the French Open, and the US Open.Therefore, next we target the model expansion to the other three major open tournaments for both men's and women's tennis. In the previous section of the analysis, we have included the venue types in the discussion and added them to the optimized momentum swings prediction model (MSPM-Pro). the Australian Open, and the US Open venue types are hardcourt, while the French Open venue type is red clay. The Australian Open, and the US Open site types are hardcourt, while the French Open site type is red clay.

⋆ Other sports

Now we know that the MSPM-Pro performs moderately well in other Grand Slam events and women's tournaments. So let's analyze the effect of MSPM-Pro in other sports events. Unlike basketball, soccer and other team confrontation type sports, tennis belongs to individual confrontation type sports, and both of them are very different in the system and the number of people. Therefore, we analyze badminton and table tennis which are two sports with similar match formats to tennis.

Therefore, in different types of tennis matches, the model needs to modify parameters such as: ground material, environmental factors, player characteristics, etc.; while for badminton and table tennis, due to the difference in the system of the game, it needs to make targeted modifications to the features, including: scoring forms, length of the game, etc.

Based on the above analysis, we updated the features of the model when predicting different problems, and finally, the accuracy of the model for swing prediction in these different matches is as follows:

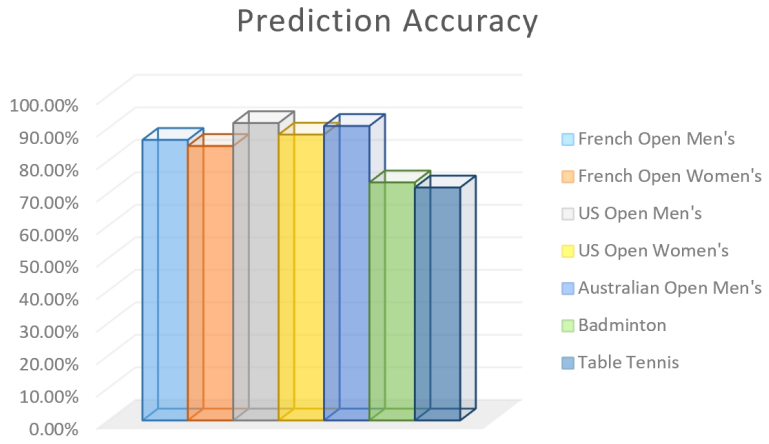


Figure 18: MSPM-Pro prediction accuracy in other competitions

It can be seen that for tennis matches, the model still has good accuracy after modifying the corresponding parameters; while for other types of matches such as badminton and table tennis, the model performs slightly worse. It shows that when we want to use the model to predict other matches, we still need to optimize the model structure and adjust the relevant parameters according to the characteristics of the matches.

In summary, MSPM-Pro demonstrates good generality, but there is still room for further optimization and improvement, which will be elaborated upon in the model evaluation section.

9 Task 5: Memo

Due to space constraints, we put the memo at the end of the paper for easy browsing.

10 Sensitivity Analysis

To test the sensitivity of the PMM model we used in Task 1, we simulated how changes in multiple metrics would affect the model.

We varied the following metrics while keeping all other metrics constant: unforced errors(UE), double fault(DF), number of ace(NA). The number of these metrics in the dataset was slowly increased and the mean of the output momentum was calculated. The result is shown in figure 19:

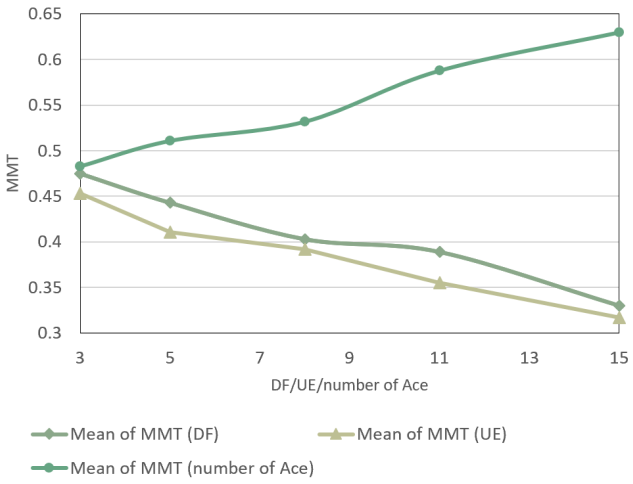


Figure 19: Sensitivity Analysis Result

From real world experience and relevant research, we know that when the ratio of unforced errors and double faults increases, it will affect the players' mentality and lead to lower momentum values; while when the number of aces hit increases, it will increase the players' self-confidence and thus have a positive effect on momentum. Thereby having a positive effect on momentum. As can be seen from Figurex, the experimental results are as expected.

According to the discussion in Task 3, we know that: unforced errors, double fault, and number of ace play a key role in the change of players' momentum, and the sensitivity of the model to these three indicators is as expected. Therefore, the PMM model in this paper is reliable.

11 Model Evaluation

11.1 Strengths

1. When we constructed the model, we used multi-dimensional features for analysis, so the variables used in the model are more representative and the model has a higher degree of accuracy.

2. Our model in Task 1 is based on Information Gain-LSTM, which can efficiently filter out the key features and better simulate the race process through the unique memory mechanism of LSTM.

3. We used official data to obtain indicators, such as unforce error and double fault, which means that our data sources are more comprehensive and scientific.

4. From our sensitivity analysis results, it can be seen that the model we have developed are stable and can be widely applicable to different matches.

11.2 Weaknesses

1. The generality of our model is not so good for other matches except tennis, and we need to adjust the parameter characteristics according to different matches.

2. As we are not sports professionals, our advice and assistance to coaches and players may not be comprehensive and will need to be considered in the light of our professional knowledge.

Reference

Table 5: Data source

Data Base	Website
ATP French Open' Tennis Data	https://www.flashscore.com/tennis/atp-singles/french-open/results/
WTA Australian Open' Tennis Data	https://www.sofascore.com/gauff-sabalenka/efnbsIZfc#id:11568964
WTA Frence Open's Tennis Data	https://www.flashscore.com/tennis/wta-singles/french-open/#/vXVNtTpp/draw
ATP Austrialian Open' Tennis Data	https://www.flashscore.com/tennis/atp-singles/australian-open/#/4ALCG38U/draw
ATP US Open' Tennis Data	https://www.flashscore.com/tennis/atp-singles/us-open/#/tSQdSudj/draw
Table Tennis Competition Data	https://github.com/ptalon91/TTStats
Tennis Competition Data	https://github.com/serve-and-volley/atp-world-tour-tennis-data
Badminton Competition Data	https://github.com/juanliong14/badminton_data_analysis
Weather Data	www.wunderground.com

Memo on MSPM-Pro for Predicting Momentum Swings

Data: February 5, 2024 MCM Team 2418588

Presentation of analysis results

In **Task 1**, we established momentum, which is related to three characteristics: scoring performance(SP), fatigue level(FL), mental state(MS), which is related to number of sets won(NS), number of games won(NG), score accumulation(SA), time consumed(TC), total distance(TD), total shot(TS), Ace, unforced error(UE).And then, the athlete's momentum can be quantified

In **Task 2**, from the correlation analysis and stochastic simulation, it is known that there is a strong correlation between the level of momentum and the ability to win the game.

In **Task 3**, we train with a neural network that can do a good job of predicting swings in the match, so that we can rationalize our suggestions for players.

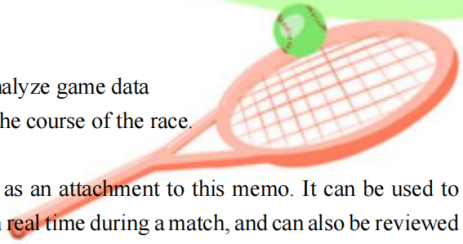
In **Task 4**, we used the model for other Grand Slam tournaments and other sports and found that it can still have good results; it is also possible to make the model fit better by adding new features and changing parameters.

Momentum function role

1. Able to visualize the competition

2. An effective tool for players and coaches to analyze game data

3. Advise athletes on adjustments by predicting the course of the race.



Advice to Coaches

We have created **a zip package of MSPM-Pro** as an attachment to this memo. It can be used to generate momentum functions for both athletes in real time during a match, and can also be reviewed after a match as an effective means of helping players train and peers learn to communicate. The following are suggestions for coaches.

- **Model Fit:** Data from the athlete's past matches can be input into our model to get the indicators that play a key role in the athlete's momentum change, and to target the weak points for improvement.
- **Formulate tactics:** Coaches should use **PMM** to analyze the strengths of the athlete and the weaknesses of the next opponent based on data from past matches to formulate appropriate offensive, defensive and psychological tactics.
- **Comparing opponents:** Use **MSPM** to obtain the fatigue level and current mentality of opposing athletes, etc., analyze their momentum change patterns and weaknesses in certain in-game indicators, and provide targeted guidance to athletes on offense and defense.

Advice to players

- **Targeted training:** In preparation training, focus on several key indicators such as SP, FL and MS, and conduct targeted training for several aspects such as serve, reaction and endurance to improve the quality of serve reception, reaction speed, and to strengthen the mindset.
- **Identify patterns:** Judge whether your momentum is cyclical or you perform better on key shots (e.g. break points), etc., to clear your strengths; at the same time, correctly recognize your weaknesses and try to overcome them.
- **Adjustment strategy:** According to the play of both sides of the game, improve the stability of the key ball, seize the opponent's physical decline and psychological pressure and other opportunities to strengthen the attack, quickly kill the game.