Journal of Computers Vol. 36 No. 1, February 2025, pp. 1-14
https://doi.org/10.63367/19911599202502360100 1

Advanced Real-time Analysis and in-Game Prediction of Tennis Matches

Zhiheng Qian[1], Yixin Shi[1*], Xinyi Tian[2], and Shurui Zheng[3]

[1] School of Foreign Languages, Shanghai Jiao Tong University, Shanghai 200240, China
{n1vnhil, yixinshi}@sjtu.edu.cn

[2] School of Electrical Information and Electronical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
chloeeeett@outlook.com

[3] Qiuzhen College, Tsinghua University, Beijing 100084, China
zhengshurui0627@gmail.com

Received 9 March 2024; Revised 30 June 2024; Accepted 8 August 2024

Abstract. In competitive sports, there exists an intangible phenomenon known as "momentum" that significantly influences the dynamics of a game. Through the application of data analysis and machine learning techniques, this elusive concept of momentum can be quantified and leveraged to forecast game outcomes. Focusing on the domain of tennis, a linear model incorporating four parameters, which are selected based on the regulations and other intrinsic attributes of tennis matches, is devised to encapsulate player performance. By definition, the exponential weighted moving average of player performance serves as a robust metric for measuring momentum, validated through correlation testing. Subsequently, utilizing logistic regression and simulation algorithms, a predictive model is constructed to forecast game progression at both discrete points and the ultimate match result. Experimental findings indicate a high degree of alignment between the model predictions and the actual flow of the game.

Keywords: real-time prediction, data analysis, logistic regression, entropy weight method

1 Introduction

Predicting the outcomes of sports games leverages the growing field of data science, with tennis, a sport globally cherished, attracting significant attention for its predictive analytics. Predominantly, this trend of forecasting sports outcomes hinges on the synergistic use of data science, computing and artificial intelligence techniques, where vast datasets are processed and analyzed to extract patterns and predictions. This combination not only facilitates the assimilation of large volumes of data but also enhances the precision and relevance of the forecasts. The application of machine learning techniques for forecasting sports outcomes has seen prolific development, driven by the demand from sports enthusiasts and professionals alike to glean predictive insights from game data.

Extensive studies have utilized various machine learning models across different sports disciplines. For instance, Hervert-Escobar et al. utilized a Bayesian learning model trained on a dataset comprising 200,000 soccer games to predict match outcomes, demonstrating the effectiveness of probabilistic models in sports predictions [1]. Similarly, Jain et al. developed a Hybrid Fuzzy-SVM model to forecast basketball game results, highlighting the utility of combining fuzzy logic with support vector machines for enhanced prediction accuracy [2]. In the realm of table tennis, Wang et al. assessed the comparative performance of decision trees and neural networks, illustrating the diverse applicability of machine learning algorithms in sports analytics [3].

In tennis, machine learning methods such as random forests and support vector machines have been extensively adopted. Studies by Gao et al. and Bayram et al. demonstrated the application of these techniques in predicting tennis match outcomes by analyzing player performance metrics and historical match data [4-7]. These methods, however, typically focus on pre-game predictions based on extensive historical datasets and often do not provide real-time insights during match play.

Real-time prediction in sports analytics is a burgeoning field that addresses the dynamic aspects of sports

* Corresponding Author

where conditions can change within seconds. Recent advancements have seen the integration of real-time data streams with predictive models to offer in-game insights that are invaluable for sports management and fan engagement.

This paper builds on these foundational studies by proposing a novel methodology that treats tennis matches as continuous time sequences. Our approach leverages minimal real-time data to predict not only the outcomes but also the progression of points within a match. This method allows for an adaptive prediction system that can adjust to the flow of the game, providing more granular and actionable insights during live matches. By analyzing point-by-point data, we aim to develop a robust model that anticipates shifts in momentum and player performance, thereby offering a more detailed analysis of match dynamics than previously available.

2 In Game Performance Indicator

2.1 Linear Model of Performance

To accurately quantify player performance during specific phases of tennis play, and thereby generate a model of momentum, we define performance as an indicator that reflects both a player's behavior and their probability of winning. Given the unique characteristics of tennis, we developed a comprehensive framework of indicators to calculate performance.

Scoring Efficiency: This metric measures a tennis player's ability to convert actions into actual points. Scoring efficiency may be positive or negative, indicating whether a point is won or lost, respectively. The absolute value represents the efficiency of converting actions into points. We quantify this indicator with the rally times of a certain point. If the player loses a point with a low rally time, his efficiency of that point would be low (close to -1, for instance). In another case, if a player wins a point rapidly, his scoring efficiency would be a high value (close to 1, for instance). If a point is very intense, the players rally a lot of time to determine the point victor, then the absolute value of scoring efficiency will be close to 0. -- In other words, as the rally time approaches positive infinity, scoring efficiency will tend to 0.

Winning Streak: This indicator tracks the number of consecutive points a player wins during a match. Higher consecutive points count demonstrates the player's ability to build momentum and pressure, as he consistently wins key points and keep his opponents unsettled.

Serve Efficiency: Serve efficiency gauges how frequently a player successfully serves the ball. As the server is advantaged, faults a player make while playing as the server will produce great mental pressure on a player and cause negative for him in the game. As serving is the most basic skill for a tennis player, we believe that serve efficiency can be used to indicate the players' confidence in the game. Meanwhile, a high serve efficiency means that the player can maintain consistency during the match and create advantageous starts, thereby increasing their chances of winning.

Returner's Win Rate: This indicator is used to count the influence of playing as a returner or sever on players' Performance. Initially, we considered using Break Points as an indicator due to their significance in matches. However, after examining the winning rate of points in the given data, it revealed a substantial advantage for servers. We realized the advantage of the server is much greater than we originally thought. For a player who returns the ball, it is very difficult to even obtain a point, much less to win a game. Consequently, we chose points as a statistical dimension to maximize the weight of this "returner and server" factor in players' Performance defined by us.

This measures the impact of playing as a returner versus a server on a player's performance. Initially, we considered using Break Points as an indicator due to their significance in matches. However, data analysis revealed a substantial advantage for servers. Consequently, we chose a broader statistical dimension that emphasizes the impact of the server and returner roles on performance.

Based on the above framework, we constructed 4 indicators as in Table 1.

Table 1. Indicators depicting performance

Measured feature	Symbol	Meanings
Scoring efficiency	E	Efficiency of winning or losing a point
Winning streak	S	The streak of winning the point
Serve efficiency	P	The efficiency of serving a ball successfully
Returner's win rate	R	Rate of winning while playing as the returner

Upon a detailed analysis of historical data provided by the Wimbledon authority, we have identified a consistent pattern in the winning ratios between service and receiving games in gentlemen's matches. The dataset reveals a distinct advantage for the server, with the winning percentage approximately at 3:7, clearly favoring the server. This statistically significant trend underscores the strategic advantage held by the serving player, indicating a higher likelihood of winning the game when serving compared to receiving. This finding corroborates existing literature that posits a non-negligible benefit for the server in tennis.

Next, we will explain how to incorporate the indicators mentioned above into the calculations with supplementary symbol in Table 2.

Since there are instances of double faults where the rally count is 0, we consider these cases as having a double fault of 1. The positive and negative signs of the rally count indicate points won and lost, respectively. Here, a higher rally count value indicates lower scoring efficiency, while a lower rally count value suggests higher scoring efficiency. In this way, we transform the rally count into a parameter that reflects the absolute value of shot efficiency.

To count *S*, we retrieve the match scores of each game, and count the two players' consecutive wins.

To calculate *P*, we can refer to double fault and serve numbers from the dataset, and add them together when calculate the efficiency. For cases with double faults, we need to include the serve number in calculating the total number of serves.

For players returning the ball, to illustrate *R,* we need to consider the winning rate of the nearly five return games.

Another very important factor, the recent scoring rate, is not counted in performance *p* mentioned here, as we have incorporated the result of a point by the sign of scoring rate.

Table 2. Supplementary symbol table

Symbol	Meanings
r	Current rally count
c	Consecutive scores
p	Point performance

The weights of the four indicators in the linear function are solved by the Entropy Weight Method (EWM).

EWM is based on the concept of Shannon entropy, which is a measure of uncertainty associated with random variables [8]. Entropy was originally developed to weigh the evenness and richness of animal and plant species. The formula Eqn. (4) used to calculate entropy in EWM is based on Shannon's original metric, which was calculated as follows:

$$H = -\sum_{i=1}^{S} (P_i \log P_i), \tag{1}$$

where P_i is the fraction of the population composed of a single species *i* and the logarithm is e-based.

The reasons we chose EWM are as follows: First, EWM is an objective weight calculation technique. Unlike subjective weight calculation which involves the judgment of the decision-maker, EWM involves mathematical algorithms and is not influenced by the subjective ideas of decision-makers. Second, there is no classification degree in this multi-index problem. EWM only considers the numerical discrimination degree of the index, and it meets our needs.

In this method, 4 indicators and 297 samples are set in the evaluation (each representing a point). For ease of description, we use *m* for the number of indicators, *n* for the number of the samples, and let x_{ij} represent the measured value of the i^{th} indicator in the j^{th} sample.

Normalization: Linear normalization is first utilized to make the data dimensionless due to different units of indicators. All the indicators chosen to have optimal data a_i. Hence, Eqn.(2) was used and it is noticeable that x_{ij} vary within the interval [0,1].

$$x_{ij}^{'} = 1 - \frac{\left| x_{ij} - a_i \right|}{\max_i \left| x_{ij} - a_i \right|} \tag{2}$$

Entropy: P_{ij} describes the weight of x_{ij} in the i^{th} indicator, which was calculated as follows:

$$p_{ij} = \frac{x_{ij'}}{\Sigma_{j=1}^{n} x_{ij'}} \tag{3}$$

In the EWM, the entropy value H_i of the i the indicator is defined as:

$$H_i = -\frac{\Sigma_{j=1}^{n} p_{ij} \cdot \log p_{ij}}{\log n} \tag{4}$$

H_i varies within the interval [0,1]. The larger Hi is the greater the differentiation degree of index i is, more information can be derived, and higher weight should be given to.

Divergence and Entropy Weights:

$$g_i = 1 - H_i, \tag{5}$$

$$w_i = \frac{g_i}{\Sigma_{j=1}^{m} g_i}. \tag{6}$$

Eqn. (5) is utilized to compute the degrees of divergence g_i, and Eqn. (6) obtains the entropy weight (w_i) of the i^{th} indicator.

In this way, the performance P_j of the j^{th} point is defined as:

$$P_j = w_1 \cdot E(j) + w_2 \cdot P_c(j) + w_3 \cdot S_s(j) + w_2 \cdot B_p(j). \tag{7}$$

Table 3. Wi for Player 1&2

W_i	Scoring efficiency	Winning streak	Serve efficiency	Returner's win rate
Player 1	0.0995	0.1790	0.1808	0.5407
Player 2	0.0880	0.1689	0.1528	0.5904

Based on our calculation (Table 3 & Fig. 1), we have observed that the performance of the serving player is significantly higher than that of the opponent. This finding aligns with our initial prediction, suggesting that the player serving possesses a notable advantage.

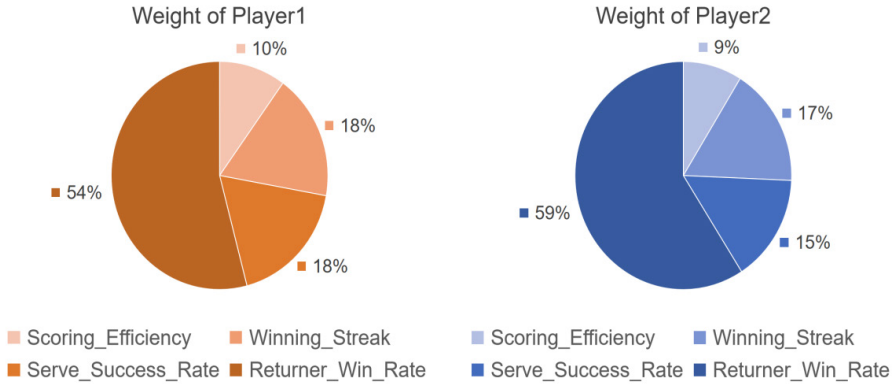

Fig. 1. Weight of Player 1&2

However, to ensure fairness in evaluating the performance of both players, we have made a slight adjustment to the performance function.

By shifting the function (for serving points, substracting average serving performance from P; for returning points, substracting average returning performance), we have ensured that the average performance score is 0 for both the serving and non-serving points. This adjustment allows us to account for any inherent advantages or disadvantages associated with serving and create a balanced assessment of performance for all players involved. Then, to maintain the parameter's comparing ability between the players, we added the original average on each P_j to demonstrate the competence of the players.

The refined performance is:

$$P_j = w_1 \cdot E(j) + w_2 \cdot P_c(j) + w_3 \cdot S_s(j) + w_2 \cdot B_p(j) - K_j, \tag{8}$$

where K_j is a constant such that the average P is the same for serving and returning while conserving the average of P considered as integrity for every player.

Using the formula mentioned above, we have determined the player's performance. We then mapped this result to a scale of 0 to 100 score, ensuring the data falls within this range. The solution is presented as follows:

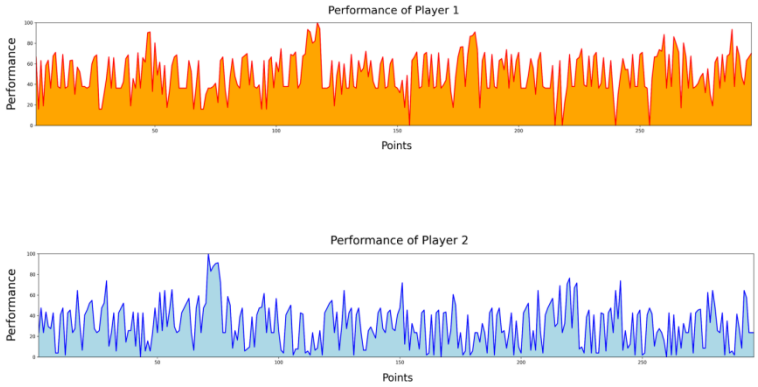

Fig. 2. Performance of Player 1 and Player 2

From Fig. 2, we can see that the performance of the two players varies considerably with the stage of the match throughout the 300-point match.

3.2 Explanatory Ability Test

As shown in Fig. 3, we examined the correlation between leading Performance and winning outcomes in a variety of games. The accuracy of performance p as an indicator of game success has been examined, yielding the following findings:

1. Across all observed cases, it was consistently observed that the participant exhibiting the best performance emerged as the victor. This suggests a strong association between performance levels and outcomes.

2. Notably, in 86.0% of the analyzed cases, the leading player possessed a performance lead exceeding 60% of the total duration of the game. This substantial margin further substantiates the significance of performance differentials in determining success.

3. Moreover, in approximately 41.2% of instances, the victor player maintained a performance lead surpassing 70% of the total game time. This indicates a notable frequency at which a sizable performance advantage was sustained throughout a significant portion of the game.

4. In 14.0% of cases, the leading player consistently maintained a dominant position for over 90% of the game duration, underscoring the potential for an overwhelming performance to secure victory.

Taken together, these findings highlight the strong relationship between leading performance and winning

outcomes in small games. The results demonstrate that maintaining a considerable performance advantage, especially exceeding 60% of the game duration, significantly increases the likelihood of emerging as the victor. These insights contribute to our understanding of the importance of performance differentials in game outcomes.

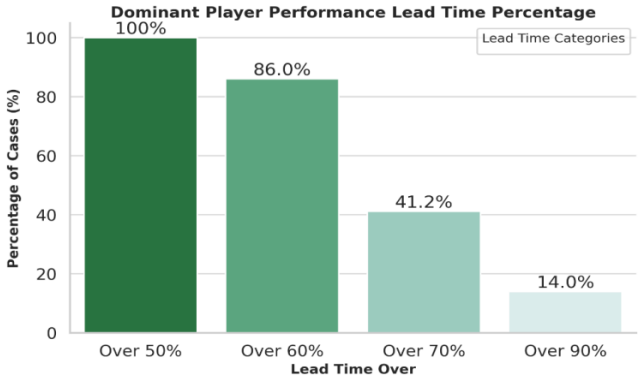

Fig. 3. Assessing game outcome using performance

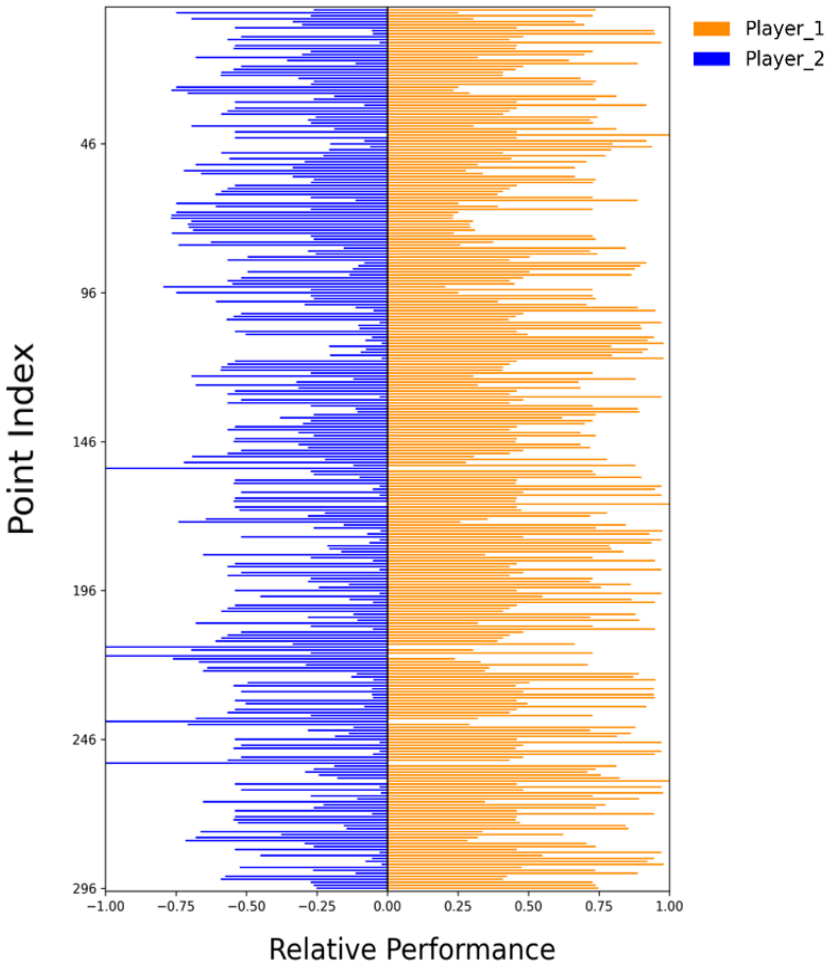

Fig. 4. Relative performance in the match

Fig. 4 visualized and compared the performance of two players of all the points in match 1301.

4 Analyzing Trend of the Game

Momentum at a certain point should be related to the momentum of the last piece of time. That is a weighted function of performance through time, where the weight descends through the distance of time.

Still, we put the data from Match 1301 in Wimbledon 2023 into calculation to examine the dependency of momentum and Success Rate. And we defined our objective function as the Pearson correlation between momentum and the probability of victory.

Exponentially Weighted Moving Average, in short, EWMA, is a method to analyze data in time series. [7] In the EWMA method, earlier data points receive lower weights, while more recent data points receive higher weights, thus giving more recent data a greater impact on the overall average. This approach ensures that newer data has a greater impact on the average, while the influence of older data diminishes over time, which fits the need of our model.

The momentum model can be quantified as follows:

$$M_t = \lambda \cdot M_{t-1} + (1-\lambda)P_t, \tag{9}$$

where M_t is momentum at time t, λ is a decaying factor in EWMA-Exponentially Weighted Moving Average to be determined by simulated annealing [12]. Here, we have the default $\lambda = 0.9$.

With default parameter, the correlation results of player 1 and player 2's momentum was 0.5284 and 0.5637. Based on the Pearson correlation coefficient table (Table 4), we see a moderate correlation between the two results. These moderate correlations suggest that the momentum metric can effectively indicate when the flow of play is likely to shift between players, thereby aiding in the detection of pivotal swings within the game.

Table 4. Pearson correlation coefficient [9]

Scale of Correlation Coefficient	Value
$0 < r \leq 0.19$	Very Low Correlation
$0.2 \leq r \leq 0.39$	Low Correlation
$0.4 \leq r \leq 0.59$	Moderate Correlation
$0.6 \leq r \leq 0.79$	High Correlation
$0.8 \leq r \leq 1.0$	Very High Correlation

We conducted a comprehensive analysis and found that the momentum, processed by EWMA, exhibits a strong correlation with the winning percentage over a specific historical period [10]. This intriguing connection motivated us to delve deeper and explore the potential for further enhancing the momentum by optimizing its parameters.

In pursuit of this objective, we employed the Simulated Annealing method, a powerful optimization technique. Its purpose was to determine the optimal value of λ, which plays a pivotal role in the EWMA calculation. By fine-tuning this crucial parameter, we aimed to maximize the momentum's effectiveness in predicting and reflecting the winning percentage.

Simulated Annealing, known for its ability to efficiently navigate complex and multi-dimensional parameter spaces, was ideally suited for our task [6, 12]. It allowed us to systematically iterate through various values of λ, simulating an annealing process where the system gradually cools and settles into an optimal state.

This iterative search for the optimal value of λ enabled us to find the configuration that yielded the highest correlation between momentum and the winning percentage. By carefully adjusting the parameter settings, we were able to enhance the momentum's accuracy and predictive power.

The first step in the simulated annealing process is to generate a new solution in the solution domain by the function. The next is to determine whether to accept the new solution or not based on an acceptance criterion.

The Metropolis criterion is the most commonly used acceptance criterion in simulated annealing process:

$$P_i(x_{old}, x_{new}) = \begin{cases} 1, & if \quad E(x_{new}) \le x_{old} \\ e^{\frac{E(x_{new})-E(x_{old})}{T}}, & if \quad E(x_{new}) > x_{old} \end{cases}. \tag{10}$$

Metropolis criterion simulates the process by which a solid reaches thermal equilibrium at a constant temperature. At a certain temperature, if the energy of the system in the new state is lower than that in the old one, or in other words, the system is more stable compared with the one in the past, then the new state will be accepted. However, if the energy possessed by the system in the new state, the turn from the current state to the new one will not be rejected instantly, but will be accepted under the probability of $e^{\frac{E(x_{new})-E(x_{old})}{T}}$. At high temperatures, the new state with a larger energy difference in the current state can be accepted; At low temperatures, only new states with less energy difference from the current state are accepted.

The most critical part of the optimization process in simulated annealing is implementing iterations within the solution domain. This process is similar to the Monte Carlo method. Each subsequent round of testing begins based on the results of the previous one. When a new solution is deemed unacceptable, the next round of tests will continue based on the original solution. During the cooling process, the objective function of simulated annealing will asymptotically converge to the optimal solution. Unlike common greedy algorithms, simulated annealing does not only converge to a local optimum because it accepts inferior solutions with a certain probability. It has been proven to be a global optimization algorithm.

By fine-tuning λ using simulated annealing, we obtained the best λ of 0.6467 and with the correlation coefficient of 0.7310, which maximized the explanatory power of momentum in understanding the situation.

Additionally, as showed in Table 5, separate calculations were performed, leading to different results. And for the value in line "Player1 & Player 2", the λ value is calculated based on the data of two players together, and then it is used to optimize the average correlation.

Table 5. Pearson correlation with optimized λ

Group name	λ	Pearson correlation
Player 1	0.6422	0.7480
Player 2	0.7023	0.7118
Player 1&Player 2	0.6467	0.7310

Fig. 5 illustrates a clear inverse relationship between the momentum of two players in a given match. As one player's momentum increases, the other player's momentum tends to decrease. This observation is consistent with the notion that in a competitive setting, momentum typically shifts between opponents, reflecting the dynamic nature of the game.

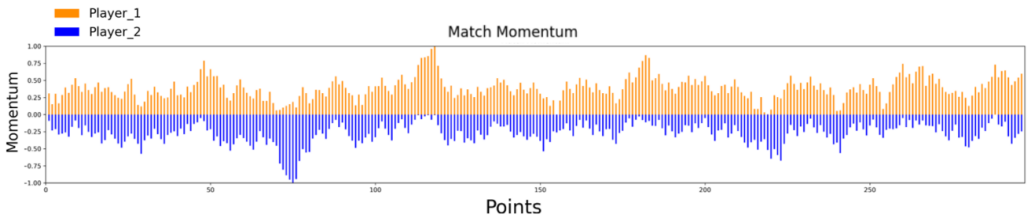

Fig. 5. Momentum of Player 1&2

Mathematically, the relationship between the two players' momentum can be quantified using statistical measures. The variance of the sum of momentum, denoted as $\sigma^2(x+y)$, is less than half the sum of the variances, $^1/_2(\sigma^2(x)+\sigma^2(y))$. In general, when there is no relationship or correlation between the momentum of the two players, we would expect the variance of the sum to be approximately equal to half of the sum of the variances. This is because the individual variances capture the dispersion or spread of each player's momentum, while the sum of the variances represents the combined spread of both players' momentum.

However, if there is a strong relationship or correlation between the momentum of the players, the variance of their sum would tend to be smaller than half the sum of their individual variances. This indicates that the combined momentum of the two players is more clustered or concentrated compared to what would be expected if there was no correlation.

By defining momentum in this manner and quantifying it with statistical measures, such as variance, we can effectively capture and quantify the dominance or control of the game within a tennis match. This approach provides a systematic and objective way to assess and compare the relative strengths of players based on their momentum.

Thus, the momentum defined is a successful figure for quantifying the dominance of the game in a tennis match. Therefore, we have demonstrated the existence of momentum.

5 Point Prediction based on Logistic Regression

During the exploration process in the first two problems, we have already quantified two values - player's performance and momentum. Furthermore, we proved that momentum and the winning rate within the neighborhood of a certain point are strongly correlated, which indicates that it is valid to implement a prediction with momentum. Based on these works, we expected to find a way to predict the next point's winning probability of the two players by a group of momentum of a given point. Once we establish a model to predict the winning probability of a single point, we could obtain a clearer profile of the flow of the game and characterize the trend of global winning possibility.

In above section, we have calculated that the Pearson correlation coefficient between a player's momentum and his winning rate is over 0.7, which means that these two variables exhibit a significantly strong linear correlation. As we expect to get a possibility as the return value, we assume that logistic regression can be applied to establish our model.

Logistic regression is a widely used statistical method for predicting the probability of a binary outcome. It maps the result of linear regression into the range from 0 to 1. In that way, the output is converted into a possibility, which is used by the model to implement classification.

The model can be represented as follows:

$$P_{win} = \frac{1}{1+e^{-(\beta_0+\beta_1 \cdot M_t)}}, \tag{11}$$

To establish a logistic regression model, we utilized the Scikit-learn library in Python (a library for machine learning, which is widely used in Data Science field).

Before fitting a logistic regression model, we must first prepare some data to be used as the label of machine learning. In a supervised learning task, labels are the target of training. The algorithm is designed to fit the model most closely to the given labels. Selecting an appropriate value as the label is very significant, it primarily determines the performance of our model.

At first, we attempted to use the result of each game as the label of logistic regression directly. We judged the model by its accuracy rate on a binary classification task - predicting the win or lose result of the next point, but the result is poor. We applied grid search to find the best parameters. Even under the optimum condition, the accuracy rate only reached 0.56, which means that it has almost no difference compared with randomly guessing the result of a point.

Reconsidering our definition of momentum, we assume that such a result is reasonable. Momentum characterizes the trend of an intangible state of a player, in common words, it is resembled to concepts like "morale" or "confidence". In our model, the better a player's historical performance is, the higher his momentum is. However, regardless of how high a player's momentum is, it is still possible for him to lose the next point. Confusing the concept of momentum with concrete concepts, like winning probability or strength of the player, will lead to mistakes [11, 13]. Analogize with "confidence", life experience has taught us, assumptions like "the more confident, the more strengthful" are invalid. Thus, for the result of a single point, a random factor should be taken into consideration.

To introduce a random factor, we changed the selected label from the actual result to a hypothetical value based on historical winning rate. We firstly recorded the winning rate of the nearest five point, then compare this historical winning rate value. In cases when the differential is no larger than 25%, we assume that the random factor will be the dominant factor for the result of next point. In other cases, we assume that player with higher historical winning rate is more possible to win the next point. With random factor, this label represents the theoretical winner (player who is more likely to win) of the next point. The pseudo code of obtaining this label is shown as follows:

Algorithm 1 Obtaining Labels

Input: rate_p1 , rate_p2
  Function :identifier(x, y)
  if $x > y$ **then**
    if $\frac{x-y}{y} < 0.25$ **then**
      return **random**.randint(0, 1)
    else
      return 1
    end if
  else if $x < y$ **then**
    if $\frac{y-x}{x} < 0.25$ **then**
      return **random**.randint(0, 1)
    else
      return 0
    end if
  else
    return **random**.randint(0, 1)
  end if
  for $i = 0$ **to len**(rate_p1) **do**
    Labels[i] $\leftarrow$ identifier(rate_p1$[i]$, rate_p2$[i]$)
  end for
Output: Labels

For the new labels, the theoretical winning rate of the next point, we applied it to fit logistic regression and obtain the best parameter with grid search again. This time, the accuracy rate reached 0.7, which means that the model is capable to predict the theoretical winner of the next point.

For some points, the theoretical winning probabilities of the two players are very close. Obviously, making a correct classification on these points are very difficult, and meaningless as the eventual data we want to learn from the model is the probability. Therefore, we affirmed that 0.7 precision is ideal enough. The actual predicting ability of the model will be much better than that.

The best parameters of the model obtained by grid search are: C=100, penalty='l1', solver='liblinear'

Here, "C" is the regularization parameter, penalty here stands for regularization penalty term, and solver is the selected optimization algorithm to solve logistic regression. "Liblinear" is an optimization algorithm suitable for binary classification problems.

Our final results are showed in Fig. 6.

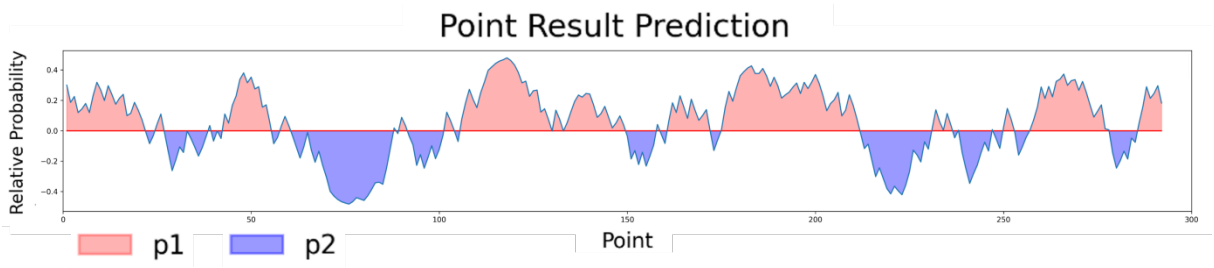

Fig. 6. Point result prediction

6 Real-time Game Result Prediction

In above section, we have accomplished following tasks:

1. Depicting momentum in moment t with player's global performance p from moment 0 to t.
2. Predicting the result of moment t+1 with momentum in moment t.

Based on these achievements, we expected to predict the global winning probability for all moment t. Thereby, we could locate the turning points of the match, and find an optimum strategy for a player to compete against his opponent in the match.

Our idea is to analyze the global winning probability is to search for all the probable result by simulating the whole process of a tennis match.

For all moment t, starting with the predicted point winning probability, we traverse all the probable match flow and calculate the global winning rate as the probability of winning the match. For each iteration, we updated the performance of the two players, and obtained the real time momentum and winning probability of the next point by the mapping pattern we have fit from the original data. In this way, we could predict the players' global winning probability of any given point as in Fig. 7.

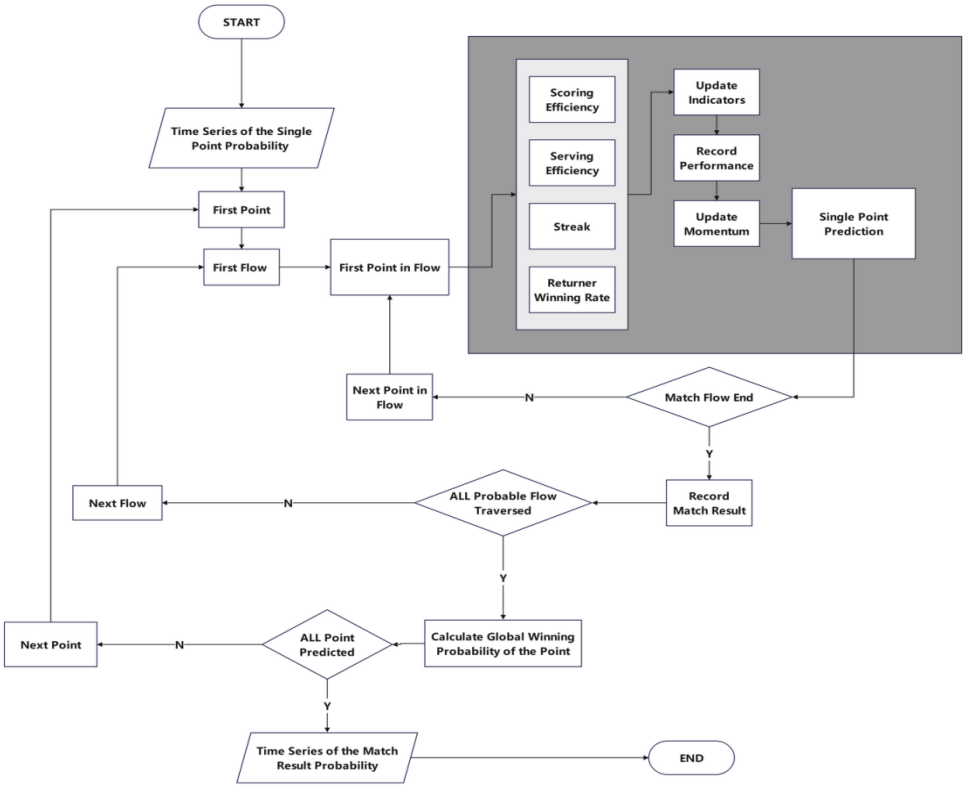

Fig. 7. Flow of simulated tennis

With this model, we found that, in match 1301 of Wimbledon 2023, points No.12, No.28, and No.183 are three turning points of the match. On these points, the flow of the match changed from favoring one player to another player. Among all the turning point, No.183 is the most decisive one. The winning probability of player2

continuously descended from point No.96 to point No.183. In first half of this match, player2 seems to be more likely to win. However, after his advantage reached a peak in point No.100, his disadvantages gradually accumulated and finally lost ground on point No.183. After that, although the probability fluctuated a little, player2 did not get the chance to turn the table.

Apply our simulation algorithm, we can depict and provide a simulated win rate graph for the overall game results. And we then tested the whole model on the 2023 Wimbledon Gentlemen's final, the match Carlos Alcaraz compete against Novak Djokovic.

Firstly, we calculated the weights of the two players as showed in Table 6.

Table 6. Weights of Djokovic and Alcaraz

W_i	Scoring efficiency	Winning streak	Serve efficiency	Returner's win rate
Carlos Alcaraz	0.0868	0.1837	0.1917	0.5378
Novak Djokovic	0.0853	0.1815	0.1927	0.5405

For these two, best λ = 0.659, 0.623, 0.635.

We then derived a prediction of the momentum and the point result of each ball as showed in Fig. 8.

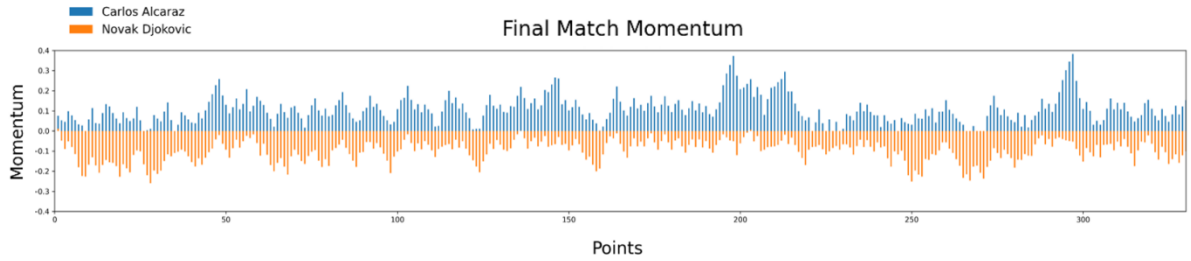

Fig. 8. Momentum of final match

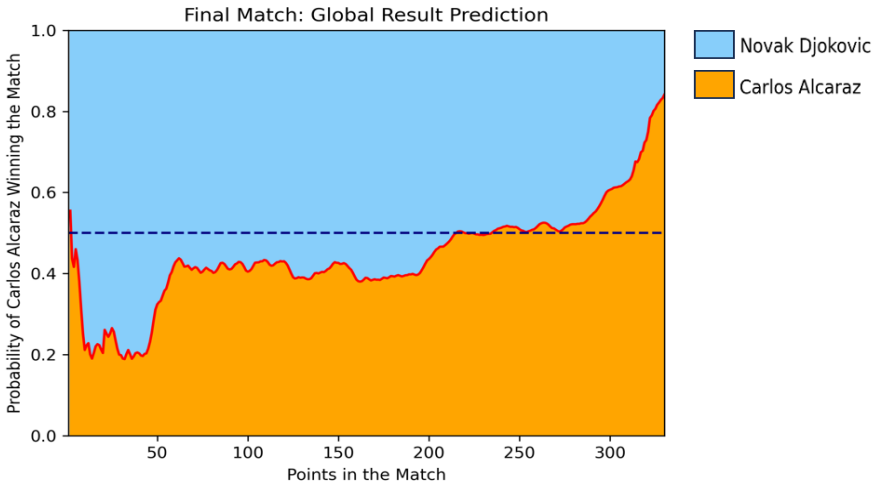

Fig. 9. Global result prediction of final match

Ultimately, based on our model and data, predictions of the outcome at each point are obtained.

As demonstrated in Fig. 8, the match commenced with Djokovic exhibiting high momentum. However, as the match progressed, his momentum showed a gradual decline, while Alcaraz, initially trailing, began to improve his performance and eventually led in the later stages of the game.

The comparative analysis of point-by-point predictions, shown in Fig. 9, initially favored Djokovic, reflecting his strong start in the game. Conversely, Alcaraz was predicted to perform better as the match progressed, which aligns with the pattern of him gaining momentum over time. The predictions indicated that despite Djokovic's early lead, the final stages of the match would tilt in favor of Alcaraz, which was precisely what unfolded in the actual match. Djokovic, who seemed poised for victory initially, saw a shift in the game's dynamics, leading to Alcaraz's consecutive wins and eventual triumph.

This alignment between our model's predictions and the actual game outcomes not only validates the accuracy of our predictive framework but also highlights its potential application in real-time sports analytics. The model's ability to adapt to the flow of the match and update its predictions based on real-time data was crucial in accurately forecasting the shift in momentum between the players.

Further analysis of our data across different matches revealed that the information entropy of the four performance indicators—serve efficiency, winning streaks, return efficiency, and break points—was remarkably consistent. This consistency allowed us to derive a set of common weights for these indicators, which simplifies the adaptation of the model to different matches and player matchups.

The standardized weights in Table 7, derived from analyzing various matches suggest that the model can be generalized to predict outcomes in a wide range of scenarios, not limited to specific player dynamics. This feature is particularly valuable for real-time sports analytics, where conditions can vary significantly from one match to another.

Table 7. Average weights for reference

W_i	Scoring efficiency	Winning streak	Serve efficiency	Returner's win rate
Average	0.1307	0.1184	0.1358	0.6151

8 Conclusion

This article introduced a groundbreaking real-time model designed to accurately capture game trends and predict outcomes using historical data in competitive sports, with a specific application to tennis. Through rigorous experimentation and validation, this model has demonstrated remarkable effectiveness and reliability in analyzing tennis matches. Our methodological approach leveraged advanced statistical techniques and machine learning algorithms to process and interpret large datasets, enabling dynamic predictions that reflect the fluid nature of live sports.

The model was extensively tested across a diverse range of tennis matches, including games with varying levels of complexity and player skill. These tests were not merely theoretical applications but involved real-time data processing, which presented unique challenges such as dealing with data sparsity and ensuring the timeliness of predictions. The positive outcomes of these experiments underscore the model's robustness and its capability to adapt to different game scenarios effectively.

Moreover, the potential of this model extends beyond the realm of tennis. The foundational principles and algorithms used in this study are versatile and can be adapted to other competitive sports with minimal modifications. Sports like basketball, football, and volleyball, which similarly generate rich datasets and require real-time analytical capabilities, could benefit substantially from this model. The transition to other sports would involve adjusting the model to accommodate sport-specific variables and data characteristics, which this research has shown to be a feasible task.

Future research could focus on several enhancements to this model. Firstly, improving the model's predictive accuracy and speed could be explored through the integration of more sophisticated machine learning algorithms and real-time data processing technologies. Secondly, extending the model to include player-specific analytics would provide deeper insights into individual performances and how they impact game outcomes [14]. Lastly, the application of this model in a real-world setting would offer further validation of its practical utility, providing sports analysts and coaches with a powerful tool to enhance their strategies and decision-making processes.

In conclusion, the real-time predictive model developed in this study not only enhances our understanding of game dynamics in tennis but also sets a precedent for the broader application of similar models in various sports disciplines. Its ability to integrate and analyze vast amounts of data in real time represents a significant advancement in sports analytics, promising exciting developments in the field.

References

[1] L. Hervert-Escobar, T.I. Matis, N. Hernandez-Gress, Prediction Learning Model for Soccer Matches Outcomes, in: Proc. 2018 Seventeenth Mexican International Conference on Artificial Intelligence (MICAI), 2018.
[2] S. Jain, H. Kaur, Machine learning approaches to predict basketball game outcome, in: Proc. 2017 3rd International Conference on Advances in Computing, Communication & Automation (ICACCA), 2017.
[3] J. Wang, L. Yu, Use the combination of the decision tree and the artificial neural networks to predict the outcome of table tennis matches, in: Proc. 2010 Sixth International Conference on Natural Computation, 2010.
[4] W. Gu, T.L. Saaty, Predicting the Outcome of a Tennis Tournament: Based on Both Data and Judgments, Journal of Systems Science and Systems Engineering 28(3)(2019) 317–343.
[5] Z. Gao, A. Kowalczyk, Random Forest Model Identifies Serve Strength as a Key Predictor of Tennis Match Outcome, Journal of Sports Analytics 7(4)(2021) 255–262.
[6] F. Bayram, D. Garbarino, A. Barla, Predicting Tennis Match Outcomes with Network Analysis and Machine Learning, in: Proc. SOFSEM 2021: Theory and Practice of Computer Science, 2021.
[7] A.S. Randrianasolo, L.D. Pyeatt, Comparing Different Data Representations and Machine Learning Models to Predict Tennis, in: Proc. FICC 2022 Advances in Information and Communication, 2022.
[8] A. Sekar, Predicting the Winner of a Tennis Match Using Machine Learning Techniques, [dissertation] Dublin: National College of Ireland, 2019.
[9] J.Benesty, J.Chen, Y. Huang, I. Cohen, Noise Reduction in Speech Processing, Springer Science & Business Media, 2009.
[10] H. Dietl, C. Nesseler, Momentum in tennis: Controlling the match, UZH Business Working Paper Series (2017) 365.
[11] P. O'Donoghue, E. Brown, Sequences of service points and the misperception of momentum in elite tennis, International Journal of Performance Analysis in Sport 9(1)(2009) 113–127.
[12] S. Kirkpatrick, C.D. Gelatt, Jr., M.P. Vecchi, Optimization by Simulated Annealing, Readings in Computer Vision: issues, problems, principles, and paradigms, Morgan Kaufmann, San Francisco, 1987 (606-615).
[13] C. Bühren, P.J. Steinberg, The impact of psychological traits on performance in sequential tournaments: Evidence from a tennis field experiment, Journal of Economic Psychology (72)(2019) 12-29.
[14] L. Page, The momentum effect in competitions: field evidence from tennis matches, Econometric Society Australasian Meeting (2009) 1-30.