Proceedings of Machine Learning Research vol 245:1–8, 2024 2024 International Conference on Machine Learning and Intelligent Computing

Prediction of Momentum in Tennis Using Random Forest Based on Bayesian Optimization and LSTM-ARIMA model

Yan Dong
Hangzhou Dianzi University, Hangzhou, 310018, China

Ruokai Zhang[*] 21041530@HDU.EDU.CN
Hangzhou Dianzi University, Hangzhou, 310018, China

Yihang Jin
Hangzhou Dianzi University, Hangzhou, 310018, China [*]

Editors: Nianyin Zeng and Ram Bilas Pachori

Abstract

Athletes' performances are often influenced by an intangible factor, momentum, which reflects the ability to perform exceptionally or consistently at a specific moment. Our model quantifies momentum and predicts match win rates, aiding athletes and coaches in optimizing their game strategies. We analyzed factors such as break points and winning streaks, employing a Random Forest Model to evaluate momentum's influence. Through the SHAP model, we established a quantifiable relationship with momentum and considered previous momentum using exponential weighted moving averages (EWMA). We developed a Gaussian Distribution Maximum Distance (GDMD) Threshold and utilized an LSTM-ARIMA model to predict momentum differences and identify turning points. The most critical factors were winning break points, running distance, and runs of success. Players are advised to be aware of their opponents' turning points and conserve energy to break them. Potential improvements include considering external factors like audience impact and expected goals, as well as incorporating more data to enhance model generalization capability.

Keywords: Random Forest, SHAP, GDMD Threshold, LSTM-ARIMA Model

1. Introduction

The 2023 Wimbledon Men's Singles final became a focal point as the 20-year-old Spanish sensation, Carlos Alcaraz, defeated 36-year-old Novak Djokovic, marking Djokovic's first defeat at Wimbledon since 2013. In the match, Djokovic initially dominated with a 6-1 advantage, but Alcaraz staged a comeback in the second set tiebreak, causing multiple shifts in the direction of the game and making momentum (Dietl and Nesseler, 2017) a key point of focus.

In previous studies, researchers have acknowledged the significant role of momentum in tennis matches (Meier et al., 2019). However, their investigations have primarily been qualitative, lacking a quantitative formula for momentum (Morgulev, 2023). To address this gap, we utilize a random forest model to derive a quantitative formula for momentum. This formula can guide coaches and players in devising strategic tactics and allocating physical resources effectively. Additionally, we employ an LSTM-ARIMA model to forecast the development of momentum in matches.

[*] Yan Dong & Ruokai Zhang & Yihang Jin contributed equally to this work.

© 2024 Y. Dong, R. Zhang[*] & Y. Jin.

2. Method

2.1. Quantitative Formula for Momentum Based on Random Forest

In a match, a player's winning depends on various factors (Moss and O'Donoghue, 2015). We believe their skill level remains stable throughout. Momentum reflects their performance at key moments, influenced by multiple factors. This paper only considers their prior performance in the match.

We use an exponentially weighted averaging algorithm to update momentum systematically. This technique assigns varying weights to data, prioritizing recent data over historical data, aligning well with competitive environments where recent performance matters more. The Exponential Weighted Moving Average (EWMA) algorithm is expressed as follows (Crowder and Hamilton, 1992):

$$D_t = \alpha D_{t-1} + (1-\alpha)\, d_t \tag{1}$$

In the formula provided, α represents the rate of weight decline, often called the attenuation factor. A smaller α leads to a faster decline, meaning less influence from past time periods on the present one. d_t is the observed value at time t, while D_t and D_{t-1} are the exponentially weighted averages at times t and $t-1$ respectively. Adjusting α allows for controlling how much the current state ignores historical data. With its trend sensitivity and data smoothing capabilities, this method is particularly useful for detecting trends and subtle changes, fitting well with the objectives outlined in this article. To quantify the magnitude of momentum, we have defined Formula 2:

$$M_t = \alpha M_{t-1} + (1-\alpha) \sum_{i=0}^{n} \beta_i x_i \tag{2}$$

We use the EWMA method to track momentum changes smoothly. M_t denotes momentum magnitude at time t. Parameters α and $1-\alpha$ determine the influence of prior momentum and past performance on current momentum. The term $\sum_{i=0}^{n} \beta_i x_i$ indicates the combined impact of performance indicators on momentum, with α set to 0.3.

We use Random Forest (RF) Karabadji et al. (2023) to determine β and interpret the results with SHAP. RF constructs multiple decision trees for classification or regression tasks, forming a "Forest" of models. Prediction results from the trees are aggregated, and a new sample's final prediction is calculated using averaging. The regression decision formula is represented as follows:

$$\hat{f}_K(x) = \frac{1}{K} \sum_{k=1}^{K} t_k(x) \tag{3}$$

In this context, $\hat{f}_K(x)$ represents the combined regression model, with t_k denoting an individual decision tree regression model, and K representing the total number of regression trees (N estimators). For models like Random Forest, several hyperparameters significantly affect predictive accuracy. Thus, tuning these hyperparameters, known as hyperparameter optimization, is crucial. However, this optimization poses a combinatorial problem, not suited for gradient descent optimization methods used for general parameters. Adjusting individual hyperparameters requires retraining, making computation resource-intensive.

To overcome these challenges, automated hyperparameter tuning processes aim to find optimal settings in less time through informed strategy-based searches. Bayesian optimization is a favored method for optimizing objective functions, constructing a probability model based on past evaluations to minimize the objective function. It's widely used for hyperparameter tuning in machine learning, offering better generalization and requiring fewer iterations compared to Grid Search (GS) and Random Search (RS) methods. (Zhang et al., 2021)

SHAP (Kim and Kim, 2022), or Shapley Additive Explanations, interprets machine learning models using Shapley values. In Random Forest models, each Decision Tree node provides conditions for dataset splits. These criteria help find optimal conditions for splitting data during classification, revealing each feature's contribution to reducing classification errors. However, this method struggles to explain individual predictions. To address this, SHAP calculates the importance of individual variables in predicting outcomes, introduced by Lundberg and Lee through an additive feature attribution method with a linear explanatory model for binary variables (Lundberg and Lee, 2017).

$$g\left(z'\right) = \phi_0 + \sum_{i=1}^{M} \phi_i z' \tag{4}$$

Here, g is a linear function of binary features.z' represents the set of observed ($z_i^{'} = 1$) or unknown ($z_i^{'} = 0$) features. The variable M denotes the number of simplified input features, and ϕ_i indicates the contribution of each feature.

2.2. Gaussian Distribution Maximum Distance Threshold

Following initial modeling, we quantitatively assess player momentum using performance data, where differences in momentum signal advantage. However, during stalemates, these differences fluctuate, obscuring who has the upper hand. Hence, a threshold, known as the "turning point," is essential to identify significant momentum shifts.

We propose the Gaussian Distribution Maximum Distance Method to determine this threshold, tailored for time series data with Gaussian distribution. Similar to Otsu's Method (Liu and Yu, 2009), it segments data into positive and negative subsets, then applies Otsu's Method independently to find thresholds for each.

This modification acknowledges the dual nature of time series data, enhancing accuracy, especially for datasets with positive and negative aspects. Histograms are computed and normalized separately for positive and negative data, and thresholds are iterated from minimum to maximum values, calculating between-class variance for transitional and non-transitional periods.

$$\sigma_b^2 = \frac{\omega_1 \omega_2 (\mu_1 - \mu_2)^2}{(\omega_1 + \omega_2)^2} \tag{5}$$

Here, σ_b^2 represents the between-class variance, while w_1 and w_2 are the weights of the two classes separated by the threshold, and μ_1 and μ_2 are the means of the two classes.

In this formula, the between-class variance is a measure of the spread between two classes separated by a threshold. The weights w_1 and w_2 represent the probabilities of the two classes, and μ_1 and μ_2 are the means of these classes. The formula aims to find the optimal threshold that

maximizes the between-class variance, leading to effective separation of classes in the context of image segmentation or thresholding applications.

We use an iterative approach to obtain the threshold T. The iterative method proceeds as follows: a threshold, denoted as T, is set. The momentum values are then partitioned using T, resulting in two groups: inflection points and non-inflection points. The between-group variance is calculated using the distributions obtained from this partition. The process is repeated by traversing through various values of T. The optimal threshold is determined as the one that maximizes the between-group variance.

2.3. Swings prediction Based on LSTM-ARIMA Model

The ARIMA model (Piccolo, 1990) is a classic time series forecasting model used for analyzing and predicting time series data. It combines autoregressive (AR) and moving average (MA) methods while considering the differencing of the time series. After ARIMA model predictions revealed unexpected stability in the forecast data, contrary to our expectations, we decided to augment our approach by incorporating the Long Short-Term Memory (LSTM) model. The LSTM model (Yu et al., 2019) is a specialized type of recurrent neural network (RNN) designed for handling and predicting time series data. LSTM has memory cells that can capture long-term dependencies, enabling it to better handle long-term memory and information in time series data.

To assess the weights of the two models, we established a loss function:

$$L = \sum (y_{true} - y_{pred})^2 \tag{6}$$

We use the player's previous momentum data to predict the subsequent changes in their momentum and calculated the loss function. After obtaining the loss function, we could determine the weights of the two models:

$$\epsilon_1 = \frac{L_2}{L_1 + L_2} \tag{7}$$

$$\epsilon_2 = \frac{L_1}{L_1 + L_2} \tag{8}$$

Here, L_1 is the loss function of the ARIMA model, L_2 is the loss function of the LSTM model, ϵ_1 is the weight of the ARIMA model, and ϵ_2 is the weight of the LSTM model. After computation, we found that $\epsilon_1 = 0.21$ and $\epsilon_2 = 0.79$. Thus, we obtained the final weighted combination prediction:

$$P = \epsilon_1 \times P_{ARIMA} + \epsilon_2 \times P_{LSTM} \tag{9}$$

3. Result

3.1. Model Evaluation and Interpretation

The model performance obtained through the above method yields an MSE of 0.0036, $RMSE$ of 0.0601, and R^2 of 0.8336. It is evident that the model exhibits strong fitting between predicted and actual data, indicating its effectiveness in capturing underlying patterns within the dataset.

The global interpretation results of SHAP indicate that, in the model, features with higher importance include the player's ability to successfully break serve, set differentials, game differentials, point differentials, and consecutive points. All of these are positively correlated with the outcome,

meaning that successfully breaking serve, leading in sets, games, and points, as well as achieving consecutive points, significantly enhance the player's momentum. On the other hand, total moving distance and non-forced errors are negatively correlated with the outcome, indicating that a decrease in stamina within a set and personal errors tend to weaken the player's momentum to some extent. Through the above methods, the final importance of various indicators is obtained, and we utilize our model to visualize the data from the last game in the match. Figure 1 demonstrates the normalized impact on tennis matches' momentum, highlighting the influence of various factors on the game's flow. Figure 2 presents the scoring rate curve during the match, which aids in understanding the scoring situations at critical moments. Figure 3 illustrates the changes in momentum and the occurrences of key points throughout the match, thereby revealing the dynamic shifts in the game's progress.

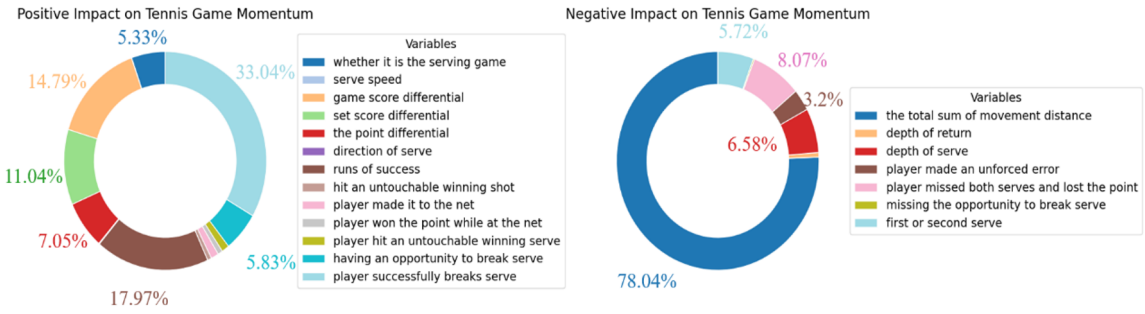

Figure 1: Normalized Impact on Tennis Matches' Momentum.

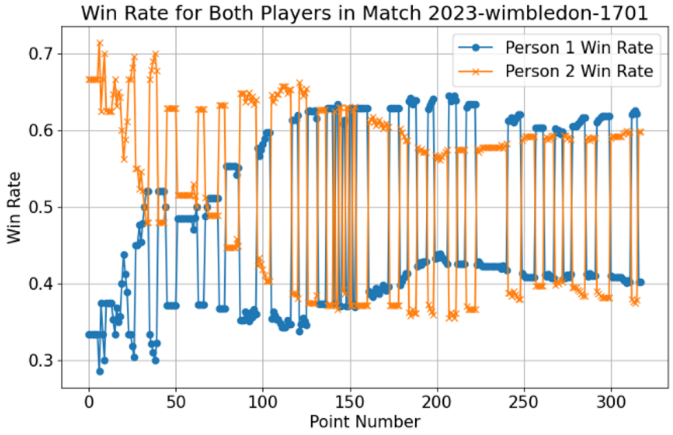

Figure 2: Scoring Rate Curve Graph During the Match.

3.2. Swings prediction

In Figure 4, we can see that upon completion of the iteration, identify the T value that maximizes the between-class variance as the threshold, corresponding to the turning point. Following that, perform the same procedure for the negative data, resulting in the identification of two turning points.

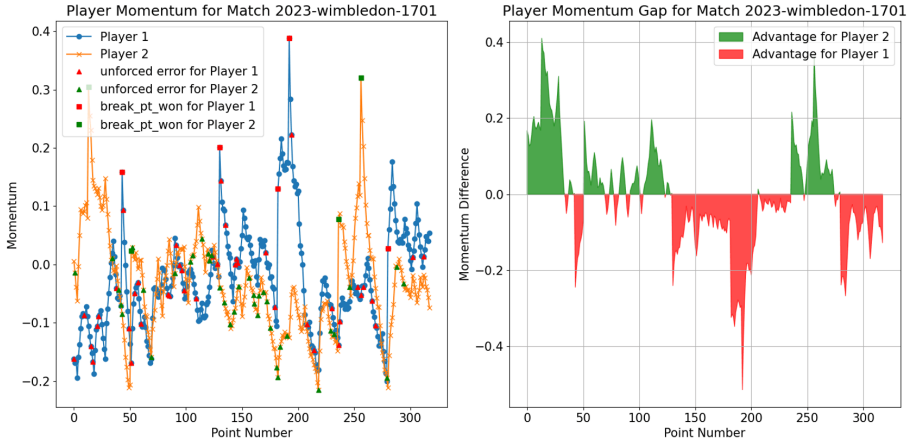

Figure 3: Changes in Momentum and Key Points Occurrences During the Match.

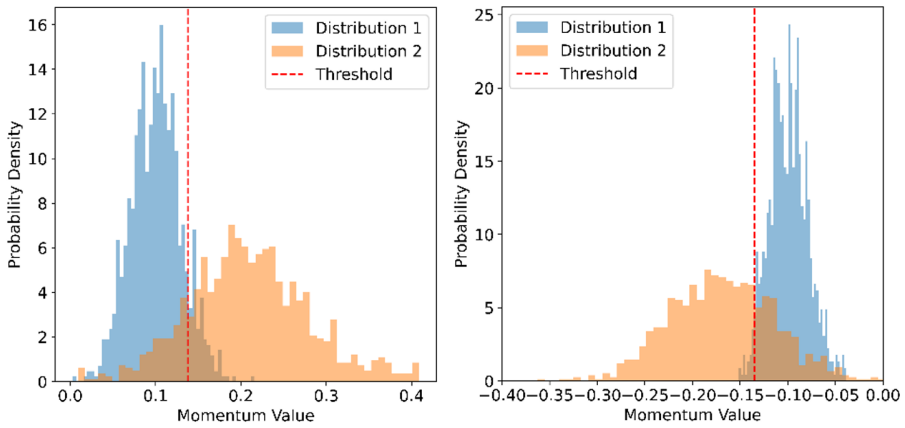

Figure 4: Using Threshold to Divide Two Distributions.

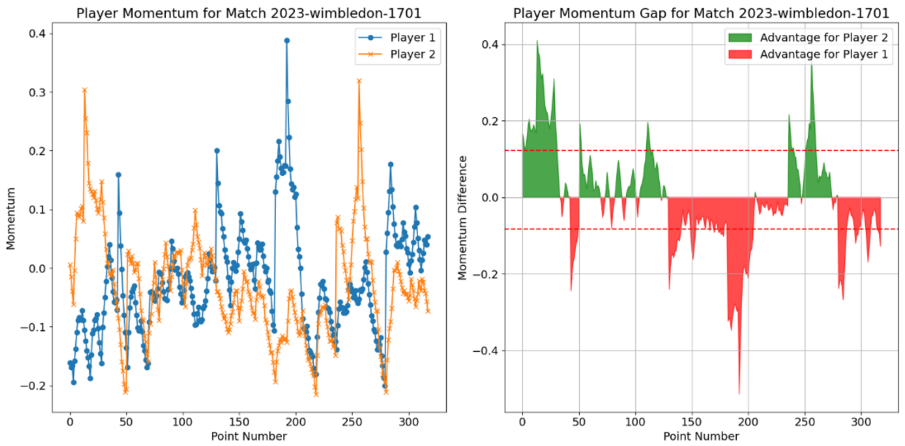

Figure 5: Momentum Gap with Threshold.

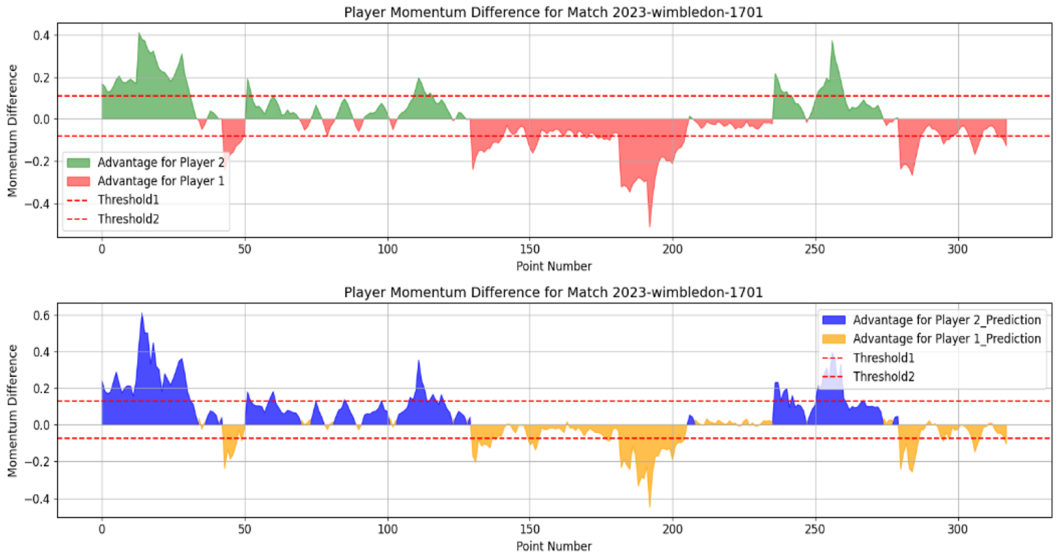

Figure 6: Momentum Gap with Threshold (Upper: Actual Results, Lower: Predicted Results).

In Figure 5, two distinct thresholds are evident, each indicating a turning point for the respective players. When the performance difference surpasses threshold 1, signaling a turning point for player 1, we observe a pronounced surge in player one's winning rate. Similarly, when the difference falls below threshold 2, marking a turning point for player 2, we again witness a sharp rise in player one's winning rate. Notably, within the range defined by these two thresholds, their winning rates tend to stabilize.

The upper part of Figure 6 displays the actual player momentum difference in a match, while the lower part illustrates the predicted values using a mixed model. The accuracy of the predictions is evident from the calculated $RMSE$ and MSE. The calculated $RMSE$ and MSE provide clear insights into the predictive accuracy. The final prediction yields an MSE of 0.0005, an $RMSE$ of 0.0214 and R^2 of 0.93, indicating a high level of accuracy in the predictions. This suggests that the model is effective in forecasting changes during matches.

4. Conclusion

In conclusion, our research thoroughly analyzes momentum's influencing factors, providing practical insights for players and coaches. We believe these findings will significantly improve training and game performance. Here are key takeaways for coaches:

- Momentum is closely tied to turning points, which can be identified using the GDMD Threshold.
- Winning break points, running distance, and serving greatly impact momentum.
- Managing running distance strategically can affect the duration and outcome of turning points.
- Serving games are crucial; winning them significantly boosts momentum.

- Note that momentum trends and winning probability may diverge, particularly when players face unforced errors despite strong momentum.

References

Stephen V. Crowder and Marc D. Hamilton. An ewma for monitoring a process standard deviation. *Journal of Quality Technology*, 24(1):12–21, 1992. doi: 10.1080/00224065.1992.11979369.

Helmut Dietl and Cornel Nesseler. Momentum in tennis: controlling the match. Technical Report 365, January 2017.

Nour El Islem Karabadji, Abdelaziz Amara Korba, Ali Assi, Hassina Seridi, Sabeur Aridhi, and Wajdi Dhifli. Accuracy and diversity-aware multi-objective approach for random forest construction. *Expert Systems with Applications*, 225:120138, 2023. doi: 10.1016/j.eswa.2023.120138.

Yesuel Kim and Youngchul Kim. Explainable heat-related mortality with random forest and shapley additive explanations (shap) models. *Sustainable Cities and Society*, 79:103677, 2022. doi: 10.1016/j.scs.2022.103677.

Dongju Liu and Jian Yu. Otsu method and k-means. In *2009 Ninth International Conference on Hybrid Intelligent Systems*, volume 1, pages 344–349, 2009. doi: 10.1109/HIS.2009.74.

Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017.

Philippe Meier, Raphael Flepp, Maximilian Rüdisser, and Egon Franck. Investigating the conditions for psychological momentum in the field: Evidence from men's professional tennis. *SSRN Electronic Journal*, 08 2019. doi: 10.2139/ssrn.3438123.

Elia Morgulev. Success breeds success: Physiological, psychological, and economic perspectives of momentum (hot hand). *Asian Journal of Sport and Exercise Psychology*, 3(1):3–7, 2023. doi: 10.1016/j.ajsep.2023.04.002. Judgment And Decision Making In Sports: Advances And Future Perspectives.

Ben Moss and Peter O'Donoghue. Momentum in us open men's singles tennis. *International Journal of Performance Analysis in Sport*, 15(3):884–896, 2015. doi: 10.1080/24748668.2015.11868838.

Domenico Piccolo. A distance measure for classifying arima models. *Journal of Time Series Analysis*, 11(2):153–164, 1990. doi: 10.1111/j.1467-9892.1990.tb00048.x.

Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. *Neural Computation*, 31(7):1235–1270, 07 2019. doi: 10.1162/neco_a_01199.

Wengang Zhang, Chongzhi Wu, Haiyi Zhong, Yongqin Li, and Lin Wang. Prediction of undrained shear strength using extreme gradient boosting and random forest based on bayesian optimization. *Geoscience Frontiers*, 12(1):469–477, 2021. doi: 10.1016/j.gsf.2020.03.007.