Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

Momentum Evaluation and Situation Prediction Model Based on Integrated Machine Learning Model - a case of tennis match

Yadi Peng [1, *, #], Xinzhou Du [2, #]

[1] School of Materials Science and Engineering, Chang 'an University, Xi'an, China, 710018

[2] School of Electronic and Control Engineering, Chang 'an University, Xi'an, China, 710018

* Corresponding author: peng754528@163.com

[#]These authors contributed equally.

Abstract. The quantitative analysis of momentum is of great guiding significance to the adjustment of competition strategy and state of coaches and athletes in the field of sports. However, the existing momentum research is mainly explained from the perspective of economics, psychology and other theories, without quantitative analysis. Or after quantitative analysis, the model can only predict the results of the whole game, and cannot accurately predict the changes during the game. Therefore, using data from the 2023 Wimbledon men's singles final as a data set, we propose a momentum evaluation model, a state prediction model and an integrated machine learning model composed of XGBoost, LightGBM, GBDT, to resume the match flow, identify which player perform better at a specific time, and predict state fluctuation. Finally, the 2023 Wimbledon men's singles final match is perfectly visualized and the state fluctuation at each time in this match is accurately predicted. The results suggest that models we established have high prediction accuracy and high stability.

Keywords: Machine learning model; Momentum quantification; XGBoost; GBDT.

1. Introduction

In the sports domain, one word, momentum, is put forward to describe the potential of an athlete to win a championship in a match under the influence of internal and external factors such as tactics, physical strength, mental state. The quantification of momentum provides accurate match feedback for tennis coaches and players, helps the personalized and innovative development of tennis training and preparation, and is of great significance to the theoretical development and practical exploration of tennis competition. [1] [2]

According to the causes of momentum, momentum generally is divided into Strategic Momentum (SM) and Psychological Momentum (PM).[3]-[5]

SM is generated by strategic factors, such as the hitting skills, the use of tactics. PM is a relatively vague concept. In the pertinent literature on the psychological interpretation of sporting events, it is summarized as "success breeds success", [6] that is, winning a round or a game increases an athlete's confidence and enhances the likelihood of winning the next.

At present, the research on momentum is mainly based on psychology and theoretical analysis, and there are few researches on actual match data analysis due to many influencing factors, such as gender, site condition. Most studies only predict the outcome of the game, which cannot accurately predict the situation.

Jiang et al. [7] constructed the optimal regression prediction equation model and Zou [8] constructed a multiple linear regression prediction model [8] to predict the outcome of the game. Elia Morgulev [6] explains momentum theoretically from the economic, psychological, physical and other aspects, but does not have quantitative analysis. More often than not, momentum in quantitative research is operationalized through streakiness (i.e., serial dependency) in performance. Arkes et al [9] uses the way addressed and operationalized the question of an outstanding performance by a basketball player streak shooting – hot hand. Namely, the methodology strives to establish whether initial success breeds further success in sports competitions. Depken, Craig A. et al. [10] separate PM from SM and establish a fully rational model of best-of-three contests. And results show that both SM and PM contribute to the outcomes of best-of-three tennis contests. But their models were not able to predict how the situation would change with limited indicators. Zhao et al. [11] extracted 28 technical indicators and physical performance indicators, and used generalized linear model and data technology inference method to model and analyze the four Grand Slams of men's tennis. However, there were too many data indicators, which was not conducive to predicting the situation in a certain period of normal competition.

Therefore, most of the researches are based on theoretical analysis, and the research of data analysis has some problems such as inaccurate prediction, large error and single method, so it is urgent to develop an efficient combination model.

In conclusion, we firstly establish a momentum evaluation model based on Principal Component Analysis (PCA) to extract vital indicators to resume matches, and then propose a state fluctuation detection model based on K-means to detect fluctuation in momentum during a match, and finally put forward with a mixed model composed of three different machine learning model to perfectly predict the process and result of a match even at some point.

2. Model and methods

To avoid complicated description and intuitively reflect our work process, the flow chart is shown as Figure 1.

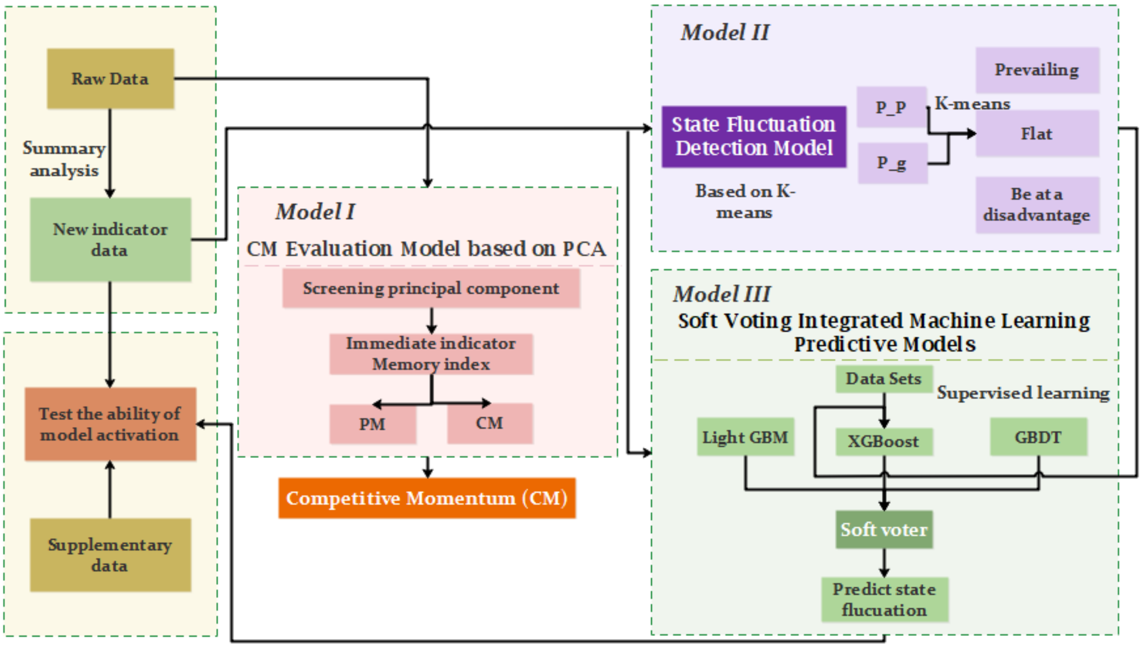


Figure 1. Our work

2.1. Data Collection

We collected and used the Wimbledon men's singles match data to build and solve our models. The data sources are shown in Table 1.

Table 1. Data source summary

Website	Data format
https://www.wimbledon.com/en_GB/scores/results/index.html	.xlsx
https://github.com/JeffSackmann/tennis_slam_pointbypoint/blob/master/2023-usopen-matches-doubles.csv	.csv

2.2. Data Cleaning

(1) Processing of Missing Data

By observing the original data, it is found that there are indeed some missing data, and after determining the data type and effectiveness, it is determined that these missing data will not be used by us, so we delete the missing rows.

(2) Unification of Data Standards

We define p1-score and p2-score as the scores of both players in a single match. It can be found from the specific data of the match that the values of p1-score and p2-score are based on the relationship between the points scored in winning a game and the corresponding integrals. The specific score table is shown in Table 2.

Table 2. Scores are converted to corresponding integrals

Integral	Score
Love	0
15	1
30	2
40	3
AD	4

According to Table 2, we keep the integral 0 in the data set unchanged, replace 15 with 1, 30 with 2, 40 with 3, and AD with 4, so as to unify the data and facilitate subsequent processing.

2.3. Indicators of Short-term Memory Effects

42 indications are visible from the data set that are connected to players' movements, hits, and equal points throughout the game. Metrics pertaining to PM and SM need to be filtered from the data set. Based on the literature [12], We summarize these indicators into three indicators that have short-term memory effects: *runs-success*, *runs-fail*, *HB (Hold & Break).* The specific definition and processing are shown in Figure 2.

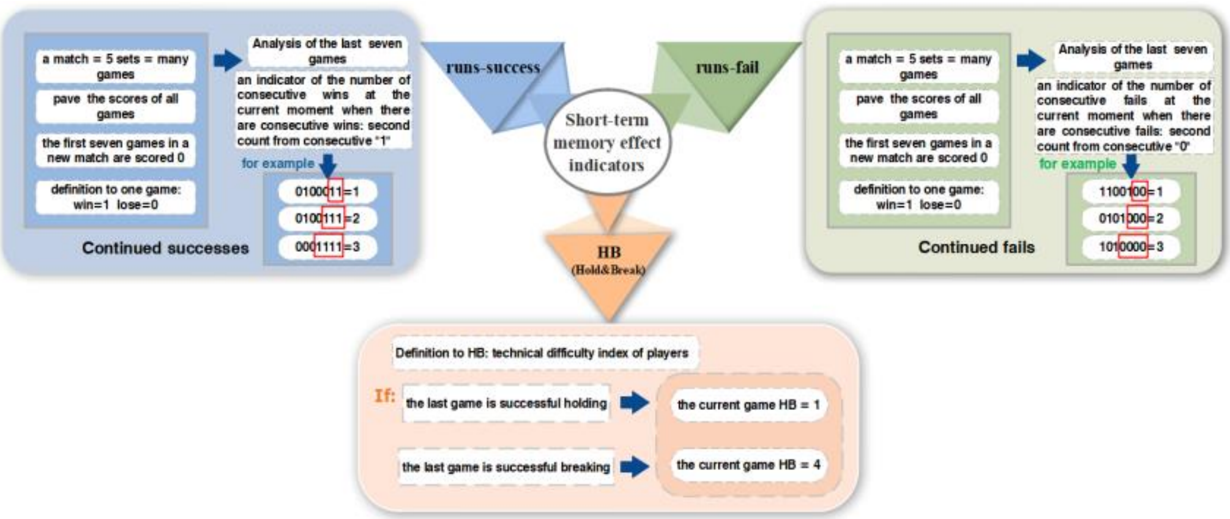


Figure 2. Three short-term memory effect indicators

Runs-success and runs-fail: This is a measure of the number of consecutive successes (and fails) of a player at the current moment in a match. Define success(win) as 1 and fail(lose) as 0. Based on the historical data of the last seven games, we get: in this set of numbers consisting of 0 and 1 with a chronological order, from left to right is the historical data to the current data; our indicator counts from the second in a row of 1s or 0s, as shown in Figure 2.

HB (Hold & Break): This is a measure of the technical difficulty of the player in a match. Using hold rate and break rate to express. According to the statistical analysis of the technical data of top tennis players from websites, Association of Tennis Professional. (https://www.atptour.com/en/scores/results-archive.) and literatures[11]-[14], the technical difficulty of break is higher than that of guarantee, that is, guarantee rate is higher than break rate. Therefore, HB is used as the technical index, and the score of guarantee and break is 1, 4 respectively.

2.4. Competitive Momentum Evaluation Model Based on PCA

Principal component analysis is to utilize the principle of statistical analysis to downscale the relevant multi-dimensional indicators and extract as fully as possible a few key indicators that reflect the information of the original indicators [15]. The dimensions of 26 indicators related to competitive data were reduced by PCA to extract the key indicators.

2.4.1. Principal Component Renaming.

We renamed 10 principal components based on the aforementioned study of the positive and negative correlations as well as the importance of the indicators of the most relevant principal components as shown in Table 3.

Table 3. Principal component renaming

Principal Component	Renaming	Principal Component	Renaming
1	Victory rush	6	Court activity
2	Confidence	7	Service strategy
3	Crucial moment	8	Crucial break
4	Momentarily ahead	9	Ace
5	Volley	10	Stress

2.4.2. Competitive Momentum (CM).

Momentum is characterized as Competitive Momentum since momentum is employed throughout the research to characterize athletes' competitive states in sports. Competitive momentum is defined to be made up of psychological and strategic momentum. In order to distinguish the existence time of competitive momentum, psychological momentum and strategic momentum are further divided into immediate and short-term.

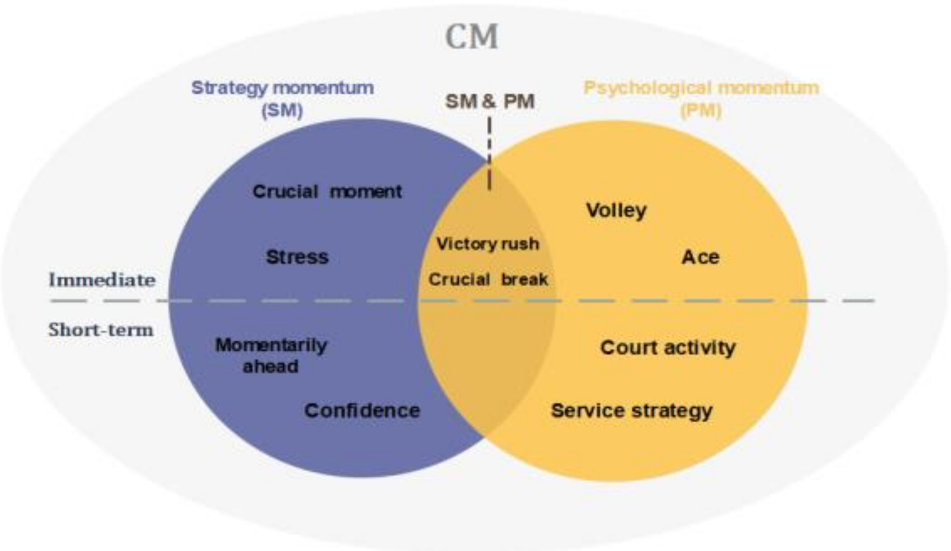


Figure 3. Venn diagram for principal component classification

Based on the classification process based on their meanings and characteristics, the ten main components identified using principal component analysis were used as measures of psychological and strategic momentum, and ultimately competitive momentum was calculated. (Details in Figure 3).

3. Classification and Prediction Models

3.1. State Classification Model Based on K-means Sliding Window

To accurately predict when the flow of a game changes in favor of one player to another, it is necessary to define the advantages and disadvantages of the game. A complete assessment of the state cannot be made solely on the basis of a player's winning streak or momentum change. For example, in a match where player A's victory represents 1 and his loss represents 0, the result of player A in this match can be expressed as "1111011". On the whole, player A always leads the score in this game. The state is always in player A's favor. Player A's winning streak is interrupted in the fifth game, and the momentum will also decline.

Based on K-means Sliding Window algorithm, two prominent metrics in time level are proposed to comprehensively analyze the situation changes. Short-term winning percentage of point ($p-p$) with short-term memory response and long-term winning percentage of game ($p-g$) with long-term effect, which are calculated schematically as shown in Figure 4.

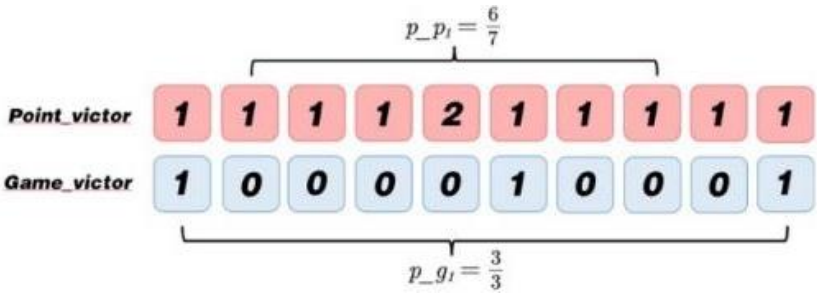


Figure 4. Indicator calculation

3.2. State Prediction Model Based on Soft Voting Fusion

XGBoost [16], LightGBM [17] and GBDT [18] models are selected for state prediction. In order to improve the robustness and generalization ability of the prediction models, soft voting is used to integrate the prediction results of the three models for final prediction.

4. Experiments and Conclusions

4.1. Data Set Partitioning

Every point from all Wimbledon 2023 men’s matches after the first 2 rounds were used in the dataset. 80% of the data in the data set excluding the final is taken as the training set, and the remaining 20% of the data is taken as the verification set.

Table 4. model parameter of GBDT, LightGBM, XGBoost

parameter name	parameter values of GBDT	parameter values of LightGBM	parameter values of XGBoost
Cross validation	5	5	5
Number of base learners	100	100	100
learning rate	0.1	0.1	0.1
Minimum number of samples for internal node split	2	/	/
Sample feature sampling rate	/	1	1
Minimum number of samples at a leaf node	1	/	/
Tree feature sampling rate	/	1	1
Node feature sampling rate	/	/	1
The maximum depth of the tree	10	10	10

The performance metrics used to evaluate the classification model mainly include accuracy, precision, recall and F_1 score. [19]

4.2. Model and method validation

(1) Results Analysis of PCA

As is shown in Table 5, the contribution rate of the first ten principal component reaches 79.621%. Considering that the eigenvalue is greater than or equal to 1, the number of principal components is determined. Finally, 10 principal components are extracted. These 10 principal components can greatly summarize the original 26 indicators.

Table 5. Principal component contribution rate

Component	Eigenvalue	Variance explanation rate (%)	Cumulative variance explanation rate (%)	Weight
1	357.089	13.734	13.734	17.249%
2	294.886	11.342	25.076	14.245%
3	215.454	8.287	33.363	10.408%
4	204.177	7.853	41.216	9.863%
5	196.214	7.547	48.762	9.478&
6	192.064	7.387	56.149	9.278%
7	183.43	7.055	63.204	8.861%
8	147.97	5.691	68.895	7.148%
9	146.165	5.622	74.517	7.061%
10	132.699	5.104	79.621	6.41%

- Expression formula of principal components

According to the weight of each principal component, the formula of the final total index can be obtained, as shown in (1):

$$\begin{aligned} Z = {} & 0.172Z_1 + 0.142Z_2 + 0.104Z_3 + 0.099Z_4 + 0.095Z_5 \\ & + 0.093Z_6 + 0.089Z_7 + 0.071Z_8 + 0.071Z_9 + 0.064Z_{10} \end{aligned} \tag{1}$$

Where Z as a general index of evaluation.

Therefore, we use the score of Z as the evaluation index of momentum.

(2) Game Flow Capture

After renaming momentum as CM and using principal component analysis in 2.4.1, we utilized *Z* score to assess competitive momentum and ultimately create a PCA-CM model. Through this evaluation model, the CM of the two players of the 2023 Wimbledon men's singles final were preliminarily analyzed, as shown in Figure 5.

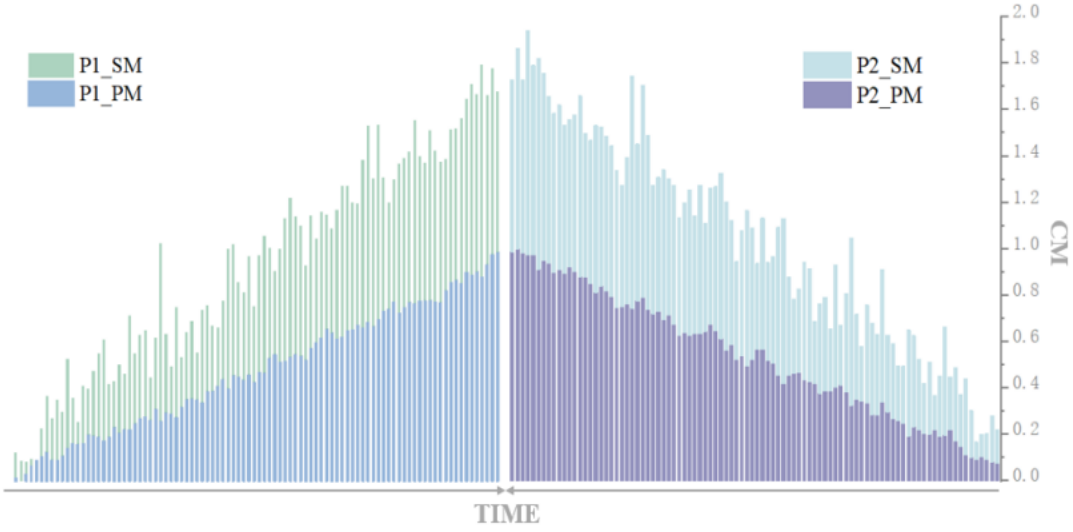


Figure 5. The CM of match_id_1301 players

As observed all the games in the third set, and get the PM and SM of the two players respectively, so as to get the index to measure the players' CM. From Figure 5, the horizontal coordinate is the time of the game progress, the ordinate is the superposition of the normalized PM and SM respectively.

According to the analysis of Figure 5, the overall growth trend of PM of player 2 is faster than that of player 1, and the fluctuation of PM is greater than that of player 1, indicating that player 2's psychological state is greatly affected with the progress of the game, and player 1's psychological state maintains a relatively stable increase during the game .From the perspective of SM, the overall SM of player 2 is higher than that of player 1, and several peaks appear significantly higher than that of player 1, indicating that player 2 is more likely to be a veteran and has richer tactical experience on the field than player 1.

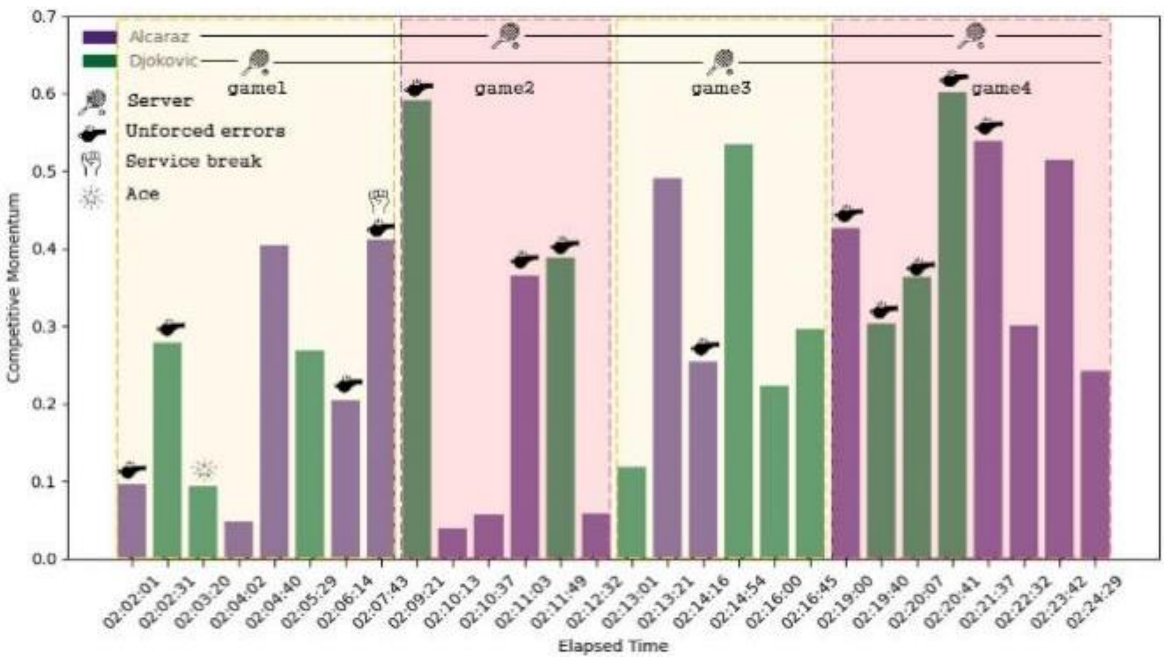


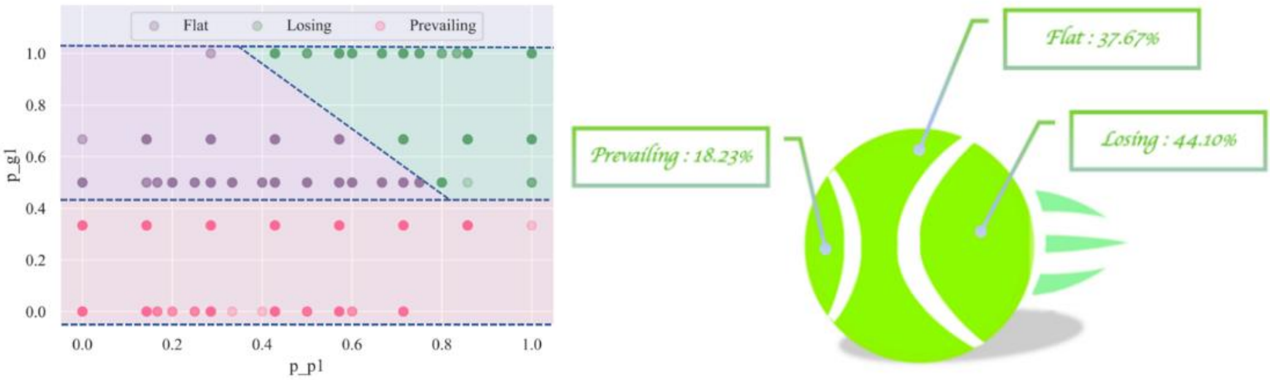
Figure 6. 2023 Wimbledon men’s singles final third set partial flow chart

To explain the state of the game and the performance of the players, we visualized the game in the third round based on the established model, as shown in Figure 9.

From Figure 6, We can see exactly how the players are doing at certain times. For example, at 02:02:01, Alcaraz won the point, performing better, and the CM values corresponding to the bar chart and whether he had runs-success reflet how well he did. Besides, we can also find out how he batted in the inning.

(3) K-means State Classification Results

Using the existing data, the p-p and p-g indicators of each player at each point were calculated, and the p-p and p-g data were cluster analyzed. The results shows that the state is divided into three categories, namely, prevailing, the flat hand and the losing. The classification results are shown in Figure 7.



(a) Classification Results (b) Proportion of various types

Figure 7. Difficulty rate classification

(4) Model Training Results

The models are trained respectively, and feature screening is carried out according to the importance of the features obtained by each model. Finally, 15 key features were selected comprehensively. The final 15 key features are used to train each model again, so as to obtain the importance of each feature and the comprehensive importance are shown in Table 6, The prediction accuracy of each model is 80%, 81.9% and 82% respectively, indicating that the prediction effect of each model is very accurate.

Table 6. Model evaluation results

	Mode Name	accuracy	recall	precision	F_1
training set	GBDT	0.923	0.923	0.924	0.923
validation set		0.803	0.803	0.805	0.803
test set		0.82	0.82	0.823	0.82
training set	LightGBM	0.868	0.868	0.869	0.868
validation set		0.796	0.796	0.799	0.797
test set		0.8	0.8	0.803	0.801
training set	XGBoost	0.99	0.99	0.99	0.99
validation set		0.804	0.804	0.806	0.804
test set		0.819	0.819	0.821	0.82

4.3. Sensitivity Analysis

The feature data was changed by adding Gaussian noise with different standard deviations. Then the momentum score is obtained through the PCA-CM model, and the momentum change percentage is further calculated.

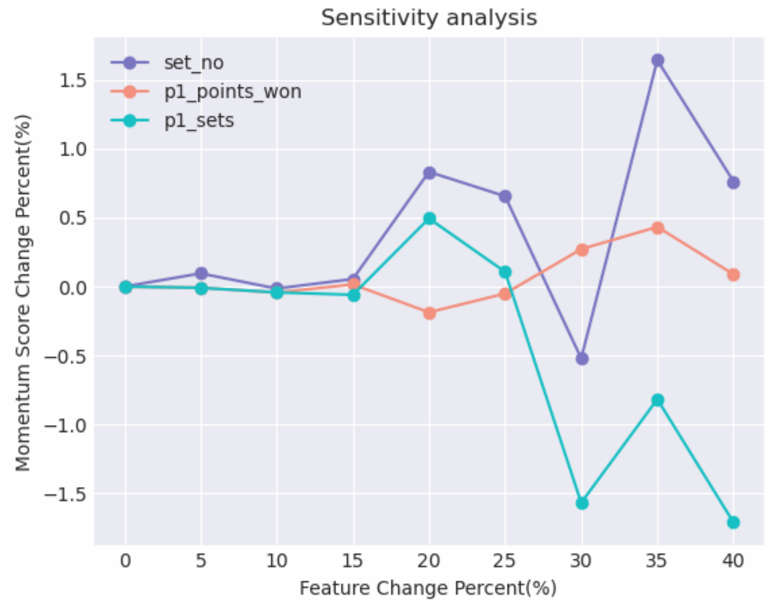


Figure 8. Sensitivity analysis

As shown in the Figure 8, feature set-no, p1-points-won, p1-sets changes in the range of 40% alone, and its maximum momentum change rate only fluctuates in the range of 1.5%, which well indicates the high stability level of CM evaluation model.

In addition, the CM evaluation model is the pre-model of the state fluctuation prediction model, so it can be concluded that the state fluctuation prediction model also has a high level of stability.

5. Conclusion

First, ten principal components were filtered using the PCA model in order to determine the momentum level. It is also categorized according to the degree of immediate and long-term impacts, psychological and strategic momentum, and literature integration to provide a compelling explanation. We offer and visually assess the competitive momentum, which is a combination of psychological and strategic momentum.

Second, the probit regression model well proves that there is a strong correlation between momentum and runs-success. In addition, it is proposed that K-means-SFD model can effectively distinguish the overall state, which is of great significance to the study of the state change.

Last but not least, a state prediction model is built through soft voting integrated machine learning. Through the analysis of momentum, we suggest that players can enter the competition state in advance through pre-competition simulation and warm-up.

References

[1] Cui, Y., Gómez, M. Á., Gonçalves, B., & Sampaio, J. (2019). Clustering tennis players' anthropometric and individual features helps to reveal performance fingerprints. European Journal of Sport Science, 19 (8), 1032 – 1044.

[2] Cui, Y., Gómez, M. Á., Gonçalves, B., Liu, H., & Sampaio, J. (2017). Effects of experience and relative quality in tennis match performance during four Grand Slams., International Journal of Performance Analysis in Sport 17 (5), 783 – 801.

[3] Meier P, Flepp R, Ruedisser M, et al. Separating psychological momentum from strategic momentum: Evidence from men's professional tennis. Journal of Economic Psychology, 2020, 78 (prepublish).

[4] A. C D, M. J G, A. D S. Set-level Strategic and Psychological Momentum in Best-of-three-set Professional Tennis Matches. Journal of Sports Economics, 2022, 23 (5): 598 - 623.

[5] A. C D, M. J G, A. D S. Set-level strategic and psychological momentum in best-of-five matches in professional tennis. Applied Economics Letters, 2023, 30 (6): 735 - 739.

[6] Elia M. Success breeds success: Physiological, psychological, and economic perspectives of momentum (hot hand). Asia Journal of Sport and Exercise Psychology, 2023, 3 (1): 3 - 7.

[7] JIANG Qi - fei. Analysis of Model Construction on Regression Prediction in Men's Tennis Singles Technology and Comprehensive Strength [J]. Journal of Jilin Institute of Physical Education, 2015 (2): 39 - 43.

[8] Zou Xin. Comprehensive Strength Evaluation and Regression Model Forecast Analysis of World Women's Hard Court Singles [J]. Sports Science and Technology Literature Bulletin, 2022, 30 (1): 67 - 70.

[9] Arkes, J., & Martinez, J. (2011). Finally, evidence for a momentum effect in the NBA.

[10] Depken, Craig A. et al. “Strategic and Psychological Momentum in Professional Tennis.” *Other Economics Research eJournal* (2020): n. pag.

[11] Zhao Yue, Cui Yi-xiong, LIU Hong-You, et al. Analysis of winning factors of Professional male Tennis players based on Grand Slam Tournament performance [J]. Journal of Beijing Sport University, 2022, 45 (4): 78 - 90.

[12] Marion G, Stephane L, Muriel T, et al. Success and decline: top 10 tennis players follow a biphasic course. Medicine and science in sports and exercise, 2011, 43 (11): 2148 - 54.

[13] O’Donoghue. Break points in Grand Slam men’s singles tennis. International Journal of Performance Analysis in Sport, 2012, 12 (1): 156 - 165. Education, Sport, Health and Physical Activity (ESHPA): International Journal. 2020, 313 - 323.

[14] Córdoba, Infantes et al. “Game profiles based on the statistics of top 25 men´s tennis.” (2020).

[15] He Gong, Jie Guo, Ye Mu, Ying Guo, Tianli Hu, Shijun, Tianye Luo and Yu Sun. Atmospheric PM2.5 Prediction Model Based on Principal Component Analysis and SSA–SVM. Sustainability 2024, 16 (2), 832.

[16] DING Changwei, WANG Xin, CHEN Tongjun, YU Ting. Application of Bayesian optimized XGBoost in seismic interpretation of small-scale faults. JOURNAL OF CHINA COAL SOCIETY. 2023, 48 (6), 2530 – 2539.

[17] Wang, Rong hua et al. Power System Transient Stability Assessment Based on Bayesian Optimized LightGBM. *2019 IEEE 3rd Conference on Energy Internet and Energy System Integration (EI2)* (2019): 263 - 268.

[18] Tian yang L, Jun hua W, Ting F, et al. Risk prediction for cut-ins using multi-driver simulation data and machine learning algorithms: A comparison among decision tree, GBDT and LSTM [J]. International Journal of Transportation Science and Technology, 2023, 12 (3): 862 - 877.

[19] DING Jianli, SUN Yue. Multiclassification Prediction of Flight Delay Based on LightGBM [J] College of Computer Science and Technology, Civil Aviation University of China, Tianjin 300300, China. 2021, 53 (06): 847 - 854.