

The Building of Players' Momentum Models in Tennis

Bingji Wang^{*,†}, Xingzhi Xu[†], Tiansong Li[†]

Haide College, Ocean University of China, Qingdao, China, 266100

* Corresponding Author Email: wangbingji@stu.ouc.edu.cn

[†] These authors contributed equally.

Abstract. This passage develops a model that can quantify the momentum of tennis players and predict match fluctuations. First, data preprocessing is performed: missing values are filled with means, non-numeric variables are assigned values, and variables are merged using the difference method to reduce data dimensions. Then, passage uses Principal Component Analysis (PCA) to reduce data dimensions and evaluate player performance through K-means clustering, which is displayed in the matches. To explore the relationship between win rates and momentum, passage uses beta distribution and Markov chain to simulate match results, finding that win rate fluctuations are highly correlated with score differences and match progress, quantifying momentum. Next, based on the LSTM algorithm, passage develops a prediction model that uses momentum scores to predict fluctuation changes, with Root Mean Square Error (RMSE) for the training and test sets being 0.02137 and 0.02396, respectively. Passage uses the CHAID model to analyze influencing factors and provide recommendations controllable by coaches and players. Finally, this passage uses the BO-LSTM algorithm to predict momentum fluctuations in women's tennis matches, introducing world ranking and age factors, and the model demonstrates good generalization capability.

Keywords: Principal Component Analysis, K-Means, Markov Chain, BO-LSTM.

1. Introduction

Since 2013, Novak Djokovic has dominated Wimbledon Grand Slam tournaments, but his impressive streak ended in the 2023 final when 20-year-old Carlos Alcaraz defeated the 36-year-old Djokovic. The match was notable for its dramatic shifts in momentum, where Djokovic seemed to be in control at various stages, yet the game dynamics changed unpredictably. This phenomenon, commonly referred to as "momentum," is challenging to quantify due to its dependence on multiple factors. Accurately measuring and predicting momentum has become a crucial area of interest in competitive sports analytics.

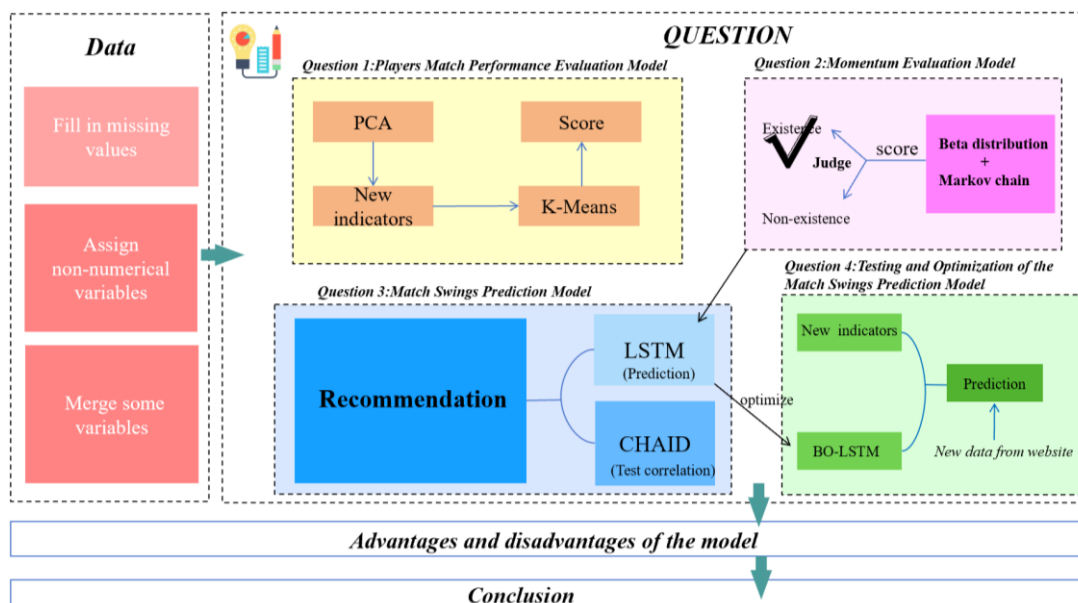


Figure 1. Overall structure of the model

This passage's approach begins with data preprocessing, including addressing missing values, assigning values to non-numeric variables, and merging variables to reduce dimensionality. The Passage employs Principal Component Analysis (PCA) for dimensionality reduction and K-means clustering to evaluate player performance. The passage examine real-time fluctuations in win rates and their correlation with other indicators to quantify momentum and demonstrate that match outcomes are not random. An LSTM (Long Short-Term Memory) model is developed to predict match fluctuations, and the CHAID (Chi-squared Automatic Interaction Detector) model is used to analyze indicator correlations and provide actionable recommendations. Data from women's tennis matches are also analyzed, incorporating two additional indicators and using an optimized BO-LSTM (Bayesian Optimization LSTM) model for predictions. The study concludes with a discussion of the models' strengths and limitations and includes practical recommendations for coaches, with a workflow illustration provided in Figure 1 [1, 2].

2. The building process of players' momentum models

2.1. Data preprocessing

We used all data (from COMAP's 2024C) after the first two rounds of the men's 2023 Wimbledon. For the data, we cannot directly perform the data analysis and need to perform data preprocessing operations. To ameliorate the condition of data set, there are four steps: data classification, data cleaning, variable quantification and variable merge, as shown in Figure 2.

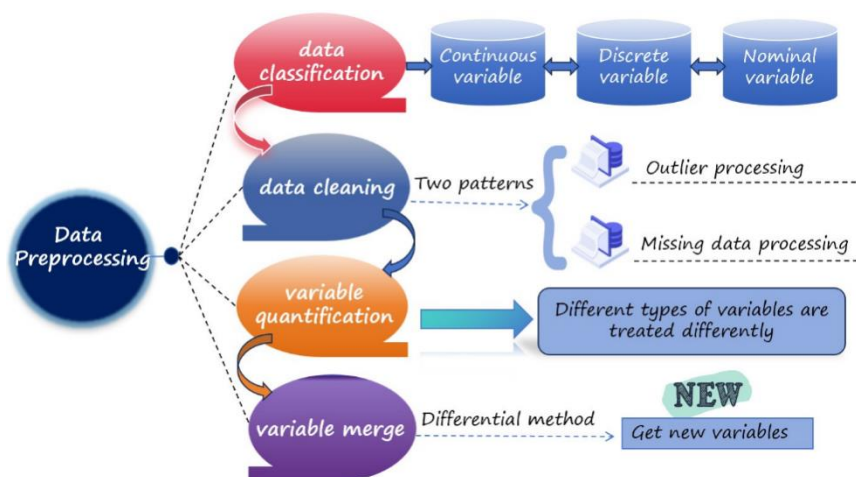


Figure 2. Data Pre-processing Flow Chart

In the subsequent process of model establishment, we consider these new indicators formed by differences, significantly reducing the dimensionality of the original data to enhance the practicality of the model.

2.2. Establishment of the Players Match Performance Evaluation Model

Since the number of variables is too large and there is some correlation between the variables, we can use principal component analysis for appropriate dimensionality reduction and principal component selection by calculating the cumulative variance contribution ratio.

The standardization of data processing, where the formula is as follows:

$$Z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}, i = 1, 2, \dots, n; j = 1, 2, \dots, p \quad (1)$$

Calculate the correlation matrix of standardized data. After finding the p eigenvalues, arrange them in order from largest to smallest and find the corresponding eigenvectors respectively.

Calculate the variance contribution rate and cumulative contribution rate. For example, the variance contribution rate of the k -th principal component is

$$\alpha_k = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \dots + \lambda_p}, k = 1, 2, \dots, p \quad (2)$$

Then the cumulative contribution rate of Y_k is

$$Y_k = \frac{\sum_{i=1}^k \lambda_i}{\sum_{i=1}^p \lambda_i} \quad (3)$$

Calculate the score of each principal component and the formula is as follows.

$$F = e_i^T X_j, i = 1, 2, \dots, k; j = 1, 2, \dots, p \quad (4)$$

Then we get the cumulative contribution rate, as shown in Table 1.

Table 1. The cumulative contribution rate

Index	Value	Index	Value	Index	Value	Index	Value
Y_1	0.150143	Y_6	0.638216	Y_{11}	0.880576	Y_{16}	0.990023
Y_2	0.267701	Y_7	0.696796	Y_{12}	0.914482	Y_{17}	0.995434
Y_3	0.378544	Y_8	0.750601	Y_{13}	0.944764	Y_{18}	1.000000
Y_4	0.473738	Y_9	0.800021	Y_{14}	0.968537	Y_{19}	1.000000
Y_5	0.558934	Y_{10}	0.840970	Y_{15}	0.981011		

We select 9 principal components as the new variables since Y_9 is more than 80%. Then we build the K-Means clustering algorithm to assign weights under each variable. We select the first three principal components after dimension reduction and visualize the data and make a 3D-scatter plot as shown in Figure 3.

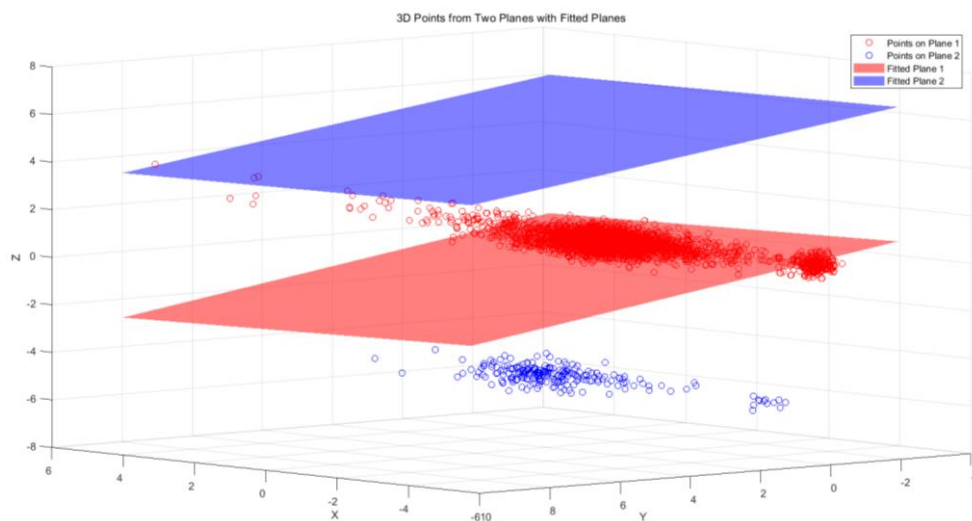


Figure 3. 3D-scatter Plot of the First Three Principal Component

From Figure 3, there is obvious stratification by retaining the three variables and the two levels are clearly different. Therefore, we integrate the discovery into the K-Means clustering model. It includes data preprocessing, centers selection randomly, loss function definition and iterations.

2.3. Establishment of the Momentum Evaluation Model

To analyze match outcomes and quantify momentum, we can use a probabilistic model. The chance of a player scoring on a serve or return correlates with their overall ability to win the game, set, and match. In matches where players have similar strengths, the outcome is independent of serve order. As shown in figure 4, let p be the probability of scoring a point while serving in a single game, and $1-p$ the probability of scoring while receiving. This leads to a Markov chain model where the server winning a point moves up, and losing a point moves down [3, 4].

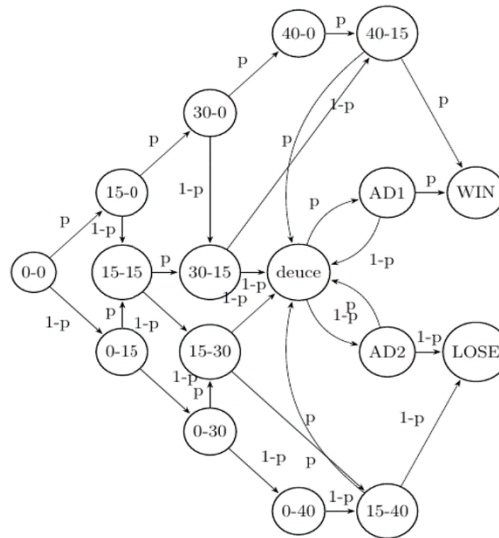


Figure 4. A Markov chain diagram of tennis match scoring

There are multiple ways that a game can end. Given that each point is a Bernoulli variable with only two mutually exclusive outcomes, we can treat a collection of points as a Binomial variable that follows a Binomial distribution. Thus, the probability of winning can be calculated as follows [5]:

$$P(x; n, p) = P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x} \quad (5)$$

2.4. Establishment of the Match Swings Prediction Model

Recently, the widely-used BP neural network, renowned for its excellent non-linear properties, flexible learning methods, and strong anti-interference capabilities, is chosen for predicting "momentum" fluctuations. We leverage the LSTM model for more efficient forecasting of future data. The LSTM network structure is shown in Figure 5 [6].

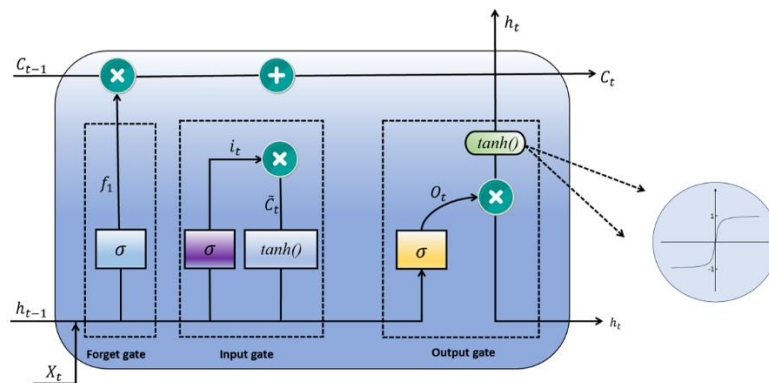


Figure 5. Network Structure Graph of LSTM

STEP 1: We choose the data from only the final match as the sample data for this model. In order to ensure that the LSTM model can converge quickly and maintain the consistency of the data, the data is normalized at the input by using Min-Max method and denormalized in the output.

STEP 2: We randomly split the sample data into training and testing sets, with a ratio of 7:3. Through extensive experiments, we find that having 10 neurons in the hidden layer works well. In this model, considering data stability, we opt for the Adam optimizer, known for its greater stability compared to the SGD optimizer. Additionally, other parameters are set as follows: 1 layer for the LSTM, a learning rate of 0.005, training for 1000 epochs, and utilizing the regression Layer as our loss function [7].

STEP 3: To better showcase the predictive performance of the LSTM model on the swings of match momentum, this paper adopts the Root Mean Square Error (RMSE) as the evaluation metric

to measure the gap between the predicted results and the actual values. The calculation method is as follows:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \quad (6)$$

Then we employ the CHAID (Chi squared Automatic Interaction Detector) decision tree to identify key variables that influence the momentum of a match. Each variable is tested against the quantified data of momentum using the Chi-Square Statistic.

Based on the principles of this algorithm, we test the variables to identify the most critical influencing factors. Utilizing the CHAID decision tree, we examine all match momentum against performance indicators of players on the court to select the most crucial influencing metrics.

To improve the predictive performance of the model, we optimize the model mentioned above. By reviewing the data, we introduced new indicators - world ranking and age. Then, we also optimized the LSTM neural networks. We introduce the Bayesian Optimization algorithm (BO-LSTM), an LSTM model with Bayesian Optimization can enhance the overall performance and robustness of the model while reducing manual intervention and improving the efficiency of hyperparameter tuning.

The core idea behind Bayesian Optimization is treating the target function to be optimized as a black-box function, which can be represented as:

$$X^* = \arg \max_{X \in A} f(X) \quad (7)$$

It continuously explores the parameter space to minimize the target function. After each attempt with a new parameter configuration, Gaussian Process Regression is used to update the inference of the target function. An optimization algorithm is then employed to select the next set of parameters. Through iterative steps, it gradually converges to the global optimal solution.

Based on the flow of this algorithm, we can understand that the entire process is repeatedly executed until the difference between the predicted values and the optimal values is less than a threshold, thereby obtaining the optimal hyperparameters for the neural network.

The BO-LSTM composite prediction model is built on four core components: data preprocessing, model training, Bayesian optimization, and model prediction. These stages collaborate to improve predictive accuracy. Data preprocessing cleans and prepares the input data, which is vital for effective model training. The LSTM network is then trained on sequential data, and Bayesian optimization refines the model parameters for optimal performance. The overall process is depicted in Figure 6.

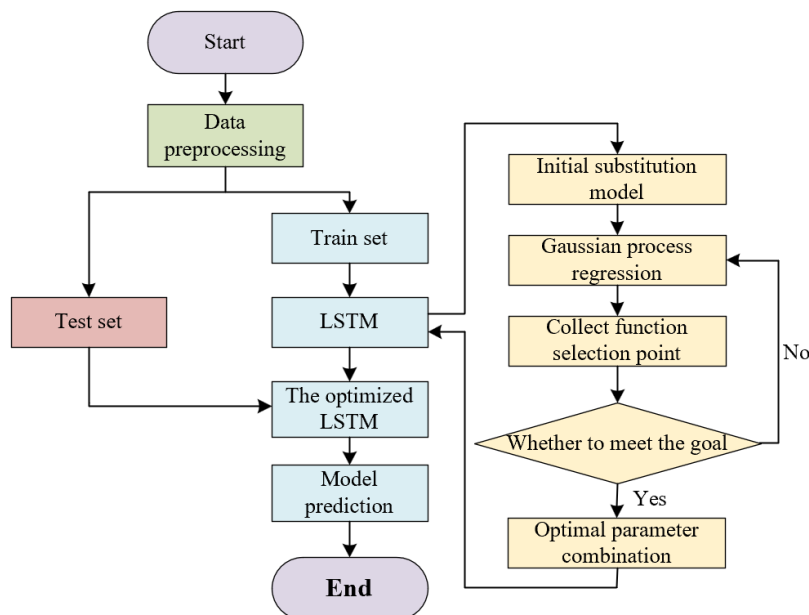


Figure 6. Flow Chart of the BO-LSTM Algorithm

We introduced new indicators, such as world ranking and age, and optimized the LSTM neural networks. The process begins with preprocessing the raw data and splitting it into training and testing sets. We then use the LSTM neural network to learn from the training data, employing Bayesian optimization to fine-tune the hyperparameters for optimal settings. The fine-tuned LSTM network is subsequently used to predict data in the testing set, enabling effective forecasting.

3. Results

3.1. The results of Players Match Performance Evaluation Model

Based on the principle of K-Means clustering model, we use mathematical software to program and finally get the following K-Means clustering scatter plot as shown in Figure 7 and Figure 8.

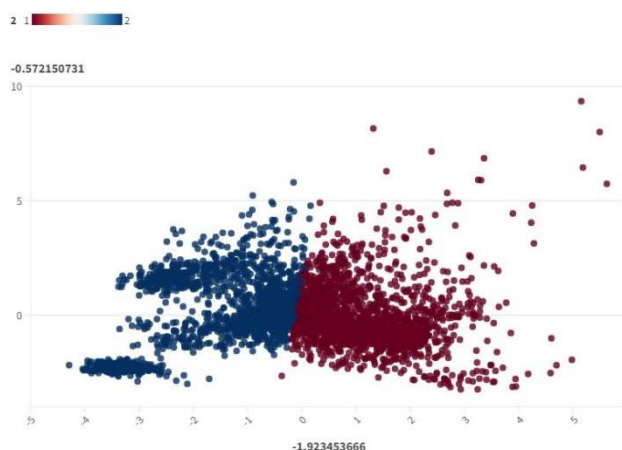


Figure 7. Player 1 Serve

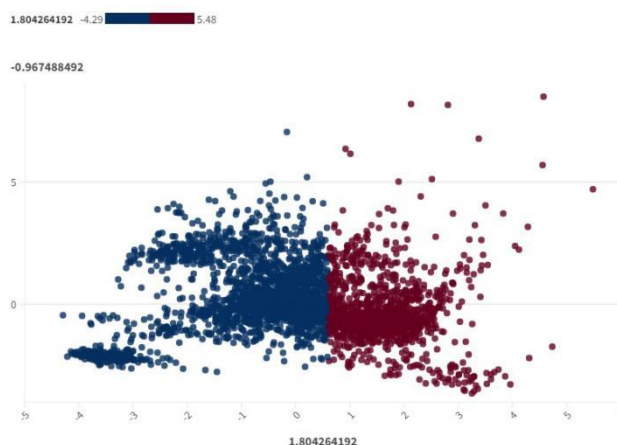


Figure 8. Player 2 Serve

The closer the value is to 1, the better athlete 1 performs; the closer to 2, the better athlete 2 performs. Figures 7 and 8 show clustering diagrams for when players 1 and 2 serve, respectively. This separation highlights each player's performance while serving. We calculate the average distance from each point to its cluster center, then determine the difference between each distance and the average. This difference is standardized and converted into a percentage, representing each player's performance score at different times. Using these scores, we select two matches for simulation, resulting in the line charts shown in Figures 9 and 10.

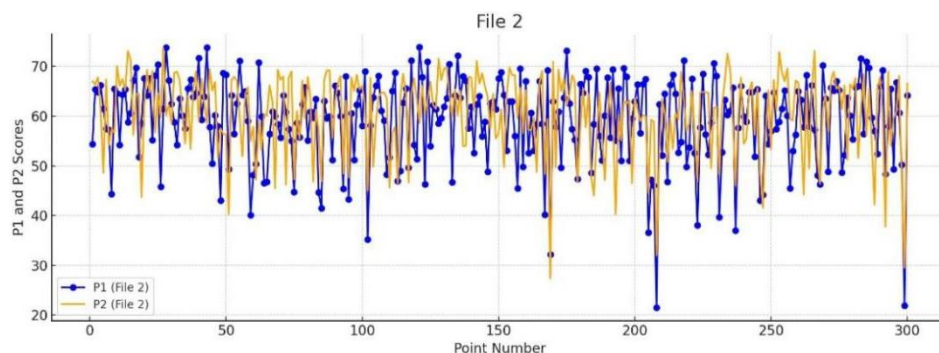


Figure 9. The Performance Score of Players in 2023-Wimbledon-1301

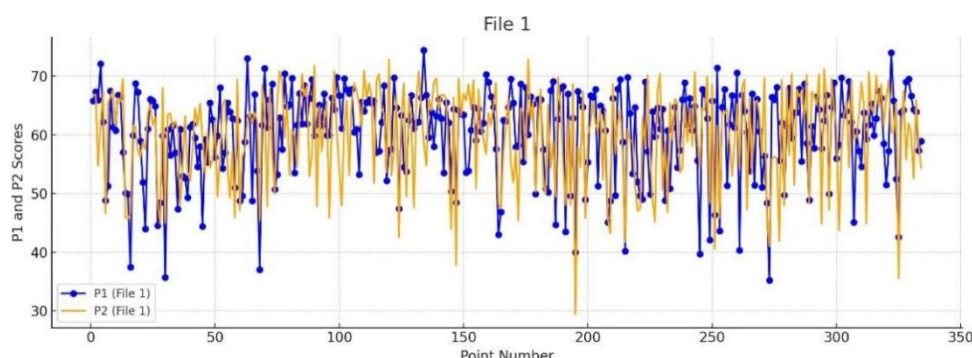


Figure 10. The Performance Score of Players in 2023-Wimbledon-1701

As can be seen from Figure 8 and Figure 9, the players' performance scores fluctuate somewhat as the game proceeds. In 2023-Wimbledon-1301, player 1 is better at the tennis game scoring points of 45,120,174 and more. In 2023-Wimbledon-1701, player 1 is better at the tennis game scoring points of 63,137,322 and more.

3.2. The results of the Momentum Evaluation Model

We utilize a data set structure that includes the server and winner for each point in the match and set up a sliding window: for each point in the match, we consider that point and up to the previous 14 points (a total of 15 points or fewer, depending on the current point's position in the match). The points within this range from a "sliding window," use to calculate the serving win probability for the current point.

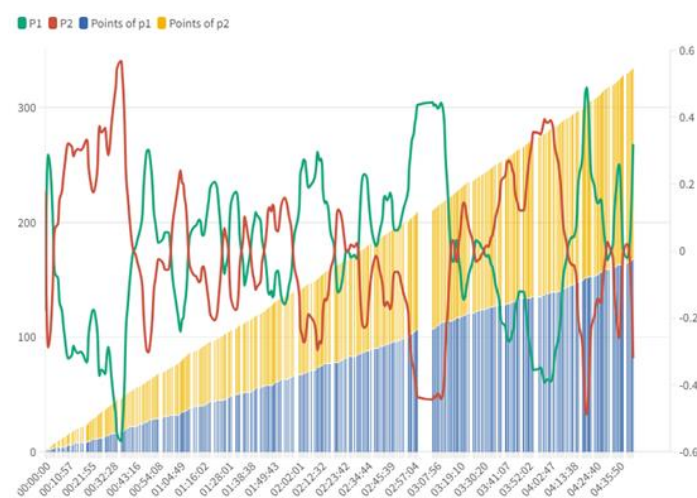


Figure 11. Momentum Matching Time Score Situation

Next, we calculate the real-time win probabilities for both players and then subtract this from the data processed through the beta distribution (Assuming that one player's advantage on the court implies the other's disadvantage). We then smooth these differences and match them with scoring

and time milestones in the match for visualization. The results are shown in Figure 11 and Figure 12 [8, 9].

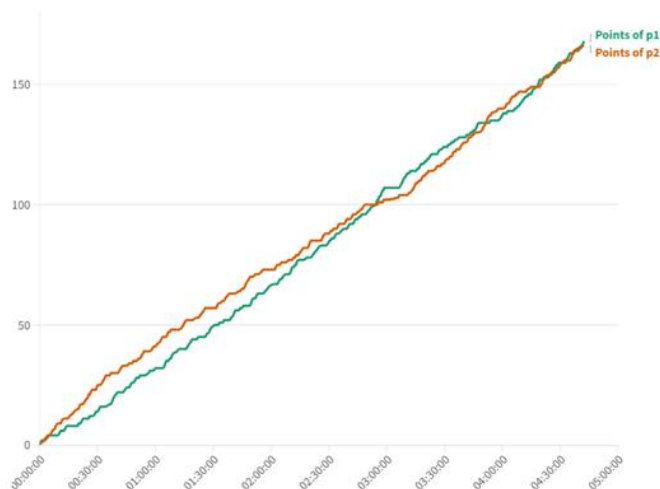


Figure 12. The Time and Score Line Chart

3.3. The results of Match Swings Prediction Model

We choose the Root Mean Square Error (RMSE) as the evaluation metric to measure the difference between predicted and actual values. The numerical values for the training and testing sets are 0.0273 and 0.0305, respectively. We plot the line chart as shown in Figure 13 and Figure 14 [10].

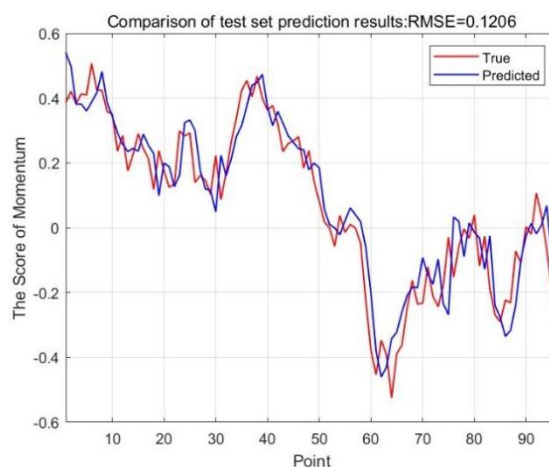


Figure 13. The Prediction Results of Train Set

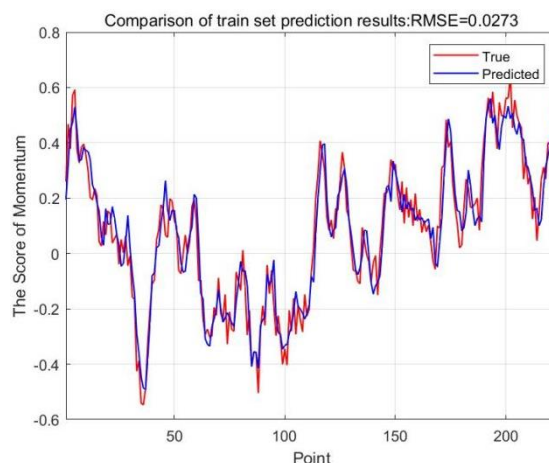


Figure 14. The Prediction Results of Test Set

It can be observed that the predicted fluctuation curve almost completely overlaps with the actual fluctuation curve, demonstrating that the model optimized using the BO-LSTM algorithm exhibits excellent predictive performance.

4. Conclusion

This passage innovatively utilizes PCA dimensionality reduction combined with the k-means algorithm to quantify multidimensional complex data on athlete performance. From a probabilistic point of view, this article also constructs a Markov chain for momentum quantification and uses BO-LSTM time series model for prediction. The quantification and prediction approaches proposed in the article not only make it possible to quantify tennis performance, but also provide valuable insights for athletes and coaches to develop more scientific competition and training strategies. With the above analysis, passage can give the following recommendations for coaches and players: To increase momentum, players should aim for an early lead with a larger score difference. They should control the match pace by managing scoring time within a comfortable range. Players should focus on reducing unforced errors, aiming for depth during returns, and avoiding body/center serves.

References

- [1] Klaassen F J G M, Magnus J R. Modeling and predicting the outcomes of tennis matches [J]. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 2023, 182 (4): 1393-1410.
- [2] Wang L, Li Y. Integrating k-means Clustering and LSTM for Enhanced Ship Heading Prediction in Oblique Stern Wave [J]. *Journal of Marine Science and Engineering*, 2023, 11 (11): 2185.
- [3] Haeri A, Moslehi F, Mousavian Anaraki S A. A hybrid reciprocal model of PCA and k-means with an innovative approach of considering sub-datasets for the improvement of k-means initialization and step-by-step labeling to create clusters with high interpretability [J]. *Pattern Anal*, 2021.
- [4] Sim M K, Choi D G. The Winning Probability of a Game and the Importance of Points in Tennis Matches [J]. *Research Quarterly for Exercise and Sport*, 2019.
- [5] Deribit B. Dynamics of Bitcoin Implied Volatility Surfaces in Deribit Options Exchange [R], 2023.
- [6] GUM Framework for HMM and LSTM Integration. Hybrid hidden Markov LSTM for short-term traffic flow prediction [R], 2023.
- [7] Wang J. Study on predicting moisture content in drying process based on LSTM neural network [J]. *Liaoning Chemical Industry*, 2023, 52 (12).
- [8] Truman J P, Fonseca V P, Soares A. A Comprehensive Data Pipeline for Comparing the Effects of Momentum on Sports Leagues [C], 2024.
- [9] Goyal A, Simonoff J S. Hot Racquet or Not? An Exploration of Momentum in Grand Slam Tennis Matches [J]. 2020.
- [10] Chai T, Draxler R R. Root mean square error (RMSE) or mean absolute error (MAE): Arguments against avoiding RMSE in the literature [J]. *Geoscientific Model Development*, 2014.