# Research on Match Trends in Tennis Competitions Based on Bayesian Online Change Point Detection and Random Forest Prediction Model

Xuanming Dong[1], Ce Liang[1], Weijie Cai[2], Yintao Wang[1, *]

[1]School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an, China, 710129

[2]School of Mathematics and Statistics, Northwestern Polytechnical University, Xi'an, China, 710129

* Corresponding Author Email: wangyintao@nwpu.edu.cn

**Abstract.** The prediction of the result of the competition is helpful for the coaches and athletes to adjust their strategy and state during the competition and to summarize and study the competition mode of the opponent after the game. For a long time, the prediction of tennis has focused on the prediction of winning and losing results, and the prediction results are often low, so our work is devoted to providing more reliable data analysis and match trend identification methods for this sport. Based on the data of The Champions Wimbledon 2023, Random forest algorithm was used to predict the mutation points obtained by Bayesian online change point detection algorithm. In this process, "momentum" was taken into account, three categories of features were constructed, and the resulting mutation points were screened. Finally, a high accuracy rate (higher than 0.8) was obtained, indicating that such mutation points contained useful information and had practical significance. Finally, we use an approximate expression of Taylor expansion and the method of undetermined coefficient to correct the weights for the features according to the extreme inversion in the schedule. This mutation point provides a new observation angle and idea for the coaches and athletes during a game of daily training, which is conducive to enlightening the athletes to achieve breakthroughs.

**Keywords:** Bayesian online change point detection, Random forest, Bi-LSTM.

## 1. Introduction

The sports prediction model is an important resource for sports analysts and coaches. The prediction of the outcome in the prediction model is the common idea of most existing models, but this prediction of the final result is not conducive to the coaching staff to guide the players in the game, and in most ball games, the guidance in the game is related to the adjustment of strategic thinking, which is an extremely important part. There are three types of prediction models for tennis results, namely regression models, point-based models, and pair-to-pair comparison models.

Here we highlight the research results based on the point model. Point-based models refer to a general class of tennis forecasting models that begin by specifying probabilities of winning a point on serve and return [1]. Under an IID assumption for point outcomes, the probability of winning a match is a function of the probabilities of winning a point on serve and return and can be written down in closed form. Newton and Keller [2] were two of the earliest authors to give a full treatment of the point-based model along with formulae for predicting a match win given a player's probabilities of winning a point on serve and return, with the possibility of a tiebreaker to determine set wins. The authors did not propose a model for the serve and return probabilities but only suggested considering some adjustment for opponent ability. Barnett and Clarke [3] proposed a tournament-specific average adjusted for player advantage on serve and opponent advantage on return. On top of them, Spanias and Knottenbelt [4] built more complex state models. Knottenbelt, Spanias, and Madurska set up the common opponent model using mean values.

In the competition of high-level athletes, in-competition guidance is very important, while the previous articles mostly focused on the prediction of the final result, i.e. the outcome. Moreover, according to Deg's article [1], the accuracy of the three methods ranged from 59% to 72%, and the

direct prediction of the outcome was not universal, that is, the high accuracy could not be guaranteed by any method for different games. This article is not a direct prediction for each point or game or set. In the case that each point conforms to the IID distribution and is independent of each other, referring to Newton's article, one-dimensional data is also selected as the research object, and the Bayesian online change point detection is used to find the points with mutations in the probability distribution before and after. Compared with calculating the probability of each point recursively by the analytic method, searching for a certain anomaly reduces the amount of data and increases the representativeness of the data. In a sense, this provides a meaningful means of intervention (which will be illustrated by the use of machine learning algorithms, that is, these mutation points have high-quality features that are significantly different from other points, which cannot be found using traditional probabilistic models). Through data visualization and the LSTM algorithm, we find the so-called "hidden mutation point" (its definition is accurately defined below), so that the mutation point is more consistent with the actual competition scene, and the change of accuracy rate before and after removal is calculated.

## 2. The mutation point determination model and the mutation point screening model

### 2.1. Bayesian online change point detection

Let $x_t \in \mathrm{Rd}$ denote the $t$-th observation in a data sequence, and let $x_{s:t}$ denote the sequence $x_s, x_{s+1}, \ldots, x_{t-1}, x_t$ for $s \leq t$. We assume that our T data points $x_{1:t}$ can be partitioned such that the data within each partition are i.i.d. samples from some distribution.

Bayesian online change point detection [5-6] works by modeling the time since the last change point called the *run length*. The run length at time $t$ is denoted $r_t$. Logically, it can take one of two values.

$$r_t = \begin{cases} 0 & \text{if change point at time t} \\ r_{t-1}+1 & \text{else} \end{cases} \tag{1}$$

After deduction, we get that:

$$p(x_{t+1} \mid x_{1:t}) = \sum_{l=0}^{t} p(x_{t+1} \mid r_t=l, x_{(t-l):t}) p(r_t=l \mid x_{1:t}) \tag{2}$$

$$p(r_t \mid x_{1:t}) = \frac{p(r_t, x_{1:t})}{\sum_{r_{t'}} p(r_{t'}, x_{1:t})} \tag{3}$$

Define two symbols $\mathbf{v}$ and $\chi$ as hyperparameters. To allow us to compute the posterior predictive without the need for integration and without calculating the posterior over η, we employ the following efficient method, shown in equations (4), (5), and (6). And This is what Adams and MacKay mean in their paper. [7]

$$v_t^{(0)} = v_{\text{prior}} \tag{4}$$

$$\chi_t^{(0)} = \chi_{\text{prior}} \tag{5}$$

$$v_t^{(l)} = v_{t-1}^{(l-1)} + 1 \tag{6}$$

$$\chi_t^{(l)} = \chi_{t-1}^{(l-1)} + u(x_t) \tag{7}$$

where both $\mathbf{v}$ and $\chi$ are hyperparameters.

The specific implementation process of the algorithm is as follows:

1. Set priors and initial conditions.

$$p(r_0)=\begin{cases}1 & \text{if change point at time t=0}\\ p(r_0=\tau) & \text{else}\end{cases} \tag{8}$$

$$v_1^{(0)}=v_{prior} \tag{9}$$

$$\chi_1^{(0)}=\chi_{prior} \tag{10}$$

$$t=1 \tag{11}$$

2.Observe new datum $x_t$.

3. Compute UPM predictive probabilities. This calculation is for each possible run length value *l*,

$$\pi_{t-1}^{(l)}=p\left(x_t \mid v_{t-1}^{(l)},\chi_{t-1}^{(l)}\right) \tag{12}$$

where the value of *l* corresponds to each possible run length.

4. Compute growth probabilities.

$$p(r_t=l,x_{1:t})=p(r_{t-1},x_{1:t-1})\pi_{t-1}^{(l)}\left(1-H(r_{t-1})\right) \tag{13}$$

where $H(\tau)$ is $f(\tau))/(S(\tau)$, and $f(\tau)$ denotes the probability that the current run length is $\tau$. $S(\tau)$ is the *survival function* at $\tau$.

5. Compute change point probability.

$$p(r_t=0,x_{1:t})=\sum_{r_{t-1}} p(r_{t-1},x_{1:t-1})\pi_{t-1}^{(l)}H(r_{t-1}) \tag{14}$$

6. Compute the evidence. This is just the normalizer in Equation 3.

7. Compute the RL posterior. Equation 3.

8. Update sufficient statistics. Equation 4-7.

9. Perform Prediction. Equation 2.

10. Set t=t+1. Return to Step 2.

### 2.2. Random forest

Random forest is a classic Bagging model whose weak learner is the decision tree model. As shown in the Figure 1.

The Random forest often follows two basic principles when establishing each tree: data randomness and feature randomness. [8]

1. Data randomization: Randomly extract data from all data as the data of one of the decision trees for training. Because there is a callback extraction, some data may be selected several times, some data may never be selected.

2. Feature random: If the feature dimension of each sample is M, a constant k<M is specified, and k features are randomly selected from M features. An optimal feature is selected from this subset for partitioning, and in general, $k=\log_2 M$. [9]

Suppose the integration contains T base learners $\{h_1,h_2,\ldots h_T\}$, where $h_i$ represents a single output of $h_i(x)$ on example x.

The most commonly used combination strategy for classification tasks is voting. Assuming that the set of classes is $\{c_1,c_2,\ldots,c_N\}$, for the sake of discussion, here the predicted output of $h_i$ on sample x is represented as an n-dimensional vector $(h_i^1(x), h_i^2(x),\ldots h_i^N(x))^T$, where $h_i^j(x)$, represents the output of h on $c_j$. [10]

$$H(x)=\begin{cases}c_j, & \sum_{i=1}^{T} h_i^j(x)>0.5\sum_{k=1}^{N}\sum_{i=1}^{T} h_i^j(x)\\ \text{reject}, & \text{others}\end{cases} \tag{15}$$

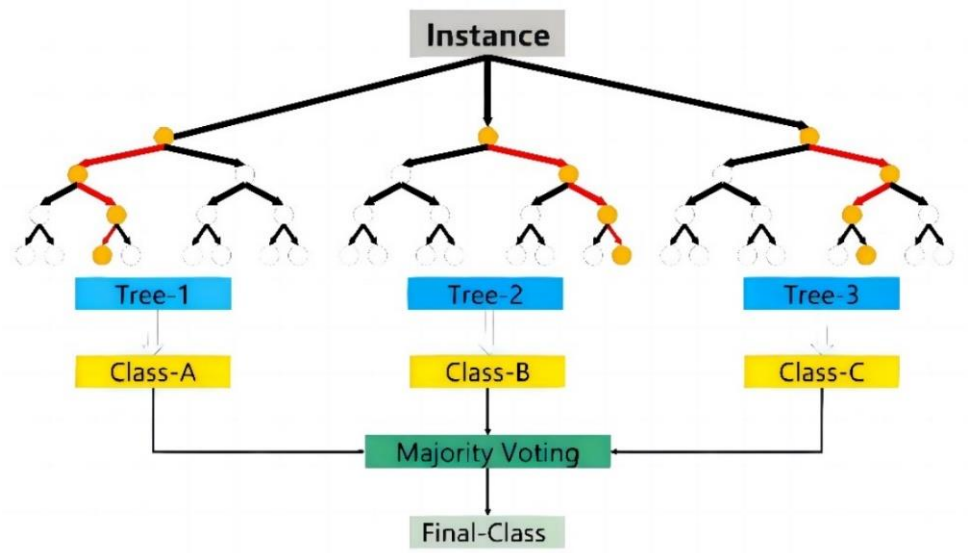


**Figure 1.** Random forest

### 2.3. Bi-LSTM

Bi-LSTM is a long short-term memory network (LSTM) with forward and backward connections.

Bi-LSTM uses two separate LSTM layers, one to process inputs in chronological order and the other in reverse chronological order, capturing the features of the input sequence in both forward and reverse directions, respectively. [11] Specifically, the forward LSTM processes the input sequence from left to right in time steps, as shown in Figure 2, with the hidden state $h_t$ and cell state $C_t$ for each time step calculated by the following formula respectively. [12-15]

$$i_t=\sigma(W_i[x,x_t]+W_i[h,h_{t-1}]+bi) \tag{16}$$

$$f_t=\sigma(W_f[x,x_t]+W_f[h,h_{t-1}]+b_f) \tag{17}$$

$$o_t=\sigma\left(W_o[xx_t]+W_o[h,h_{t-1}]+b_o\right) \tag{18}$$

$$\widetilde{C}_t=\tanh(W_c[x,x_t]+W_c[h,h_{t-1}]+b_c) \tag{19}$$

$$C_t=ft\odot C_t\text{-}1+i_t\odot\widetilde{C}_t \tag{20}$$

$$h_t=o_t\odot\tanh(C_t) \tag{21}$$

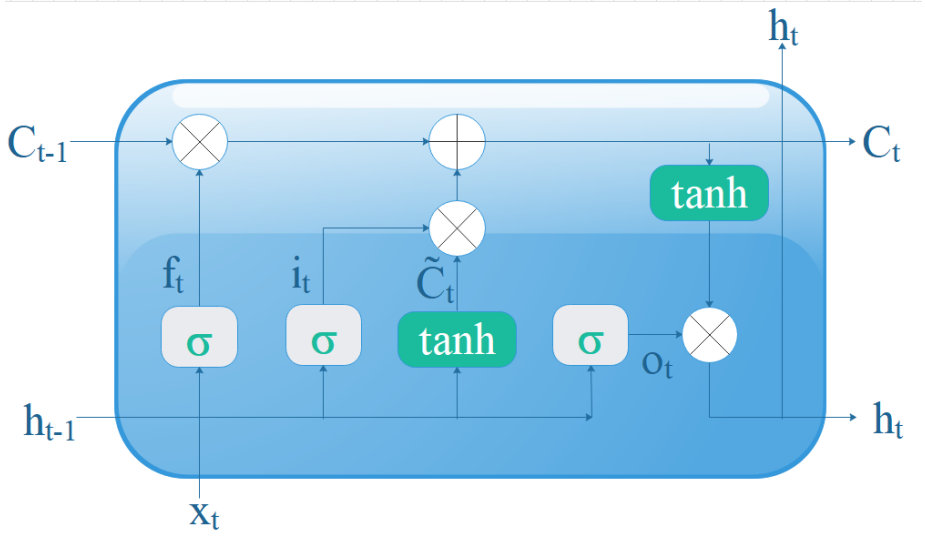


**Figure 2.** the LSTM unit controlled by the gates mechanism

where $i_t$, $f_t$, $o_t$, and $\widetilde{C}_t$ are respectively the input gate, the forgotten gate, the output gate, and the candidate state representing the current information, $\sigma$ and tanh are sigmoid and hyperbolic tangent functions respectively, $\odot$ stands for the element-by-element product. $W_f$, $W_i$, $W_o$, and $W_c$ represent the weight coefficient of h in the process of forgetting gate, input gate, output gate, and feature extraction, respectively.

### 2.4. Taylor expansion model

For a three-variable system F:

$$F=F(e, m, a) \tag{22}$$

Next, we perform the Taylor expansion of this function [16-17] to get the equation:

$$F=F(e)+\frac{F'(e)}{1!}(m+a)+\frac{F''(e)}{2!}(m+a)^2+\ldots+\frac{F^{(n)}(e)}{n!}(m+a)^n \tag{23}$$

F(e) is a constant, the original is equal to $F(e)+\sum_{n=1}^{+\infty}\sum_{k=0}^{+\infty}C_n^k a_n m^k a^{n-k}$. Define that $a_n=\frac{F^{(n)}(e)}{n!}$. Suppose that m is much smaller than a or that both m and a are minimal values, there should be equal to $\sum_{n=1}^{+\infty} n a_n a^{n-1} m+\sum_{n=2}^{+\infty} C_n^2 a_n a^{n-2} m^2$. Suppose that a1>0, a2<0, an→0(n≥3), then the original formula is approximately equal to $F(e)+a_1+2a_2am+a_2m^2+3a_3am^2$.

## 3. Results

### 3.1. Construct the expressiveness score, and carry out the Bayesian online change point detection with a point as the time unit

Our research object is the 2023 Wimbledon men's singles tournament, and all data comes from https://www.merriam-webster.com/dictionary/”momentum”.

After referring to Clinical Chemistry’s paper[18] and Jurnal Penelitian Pembelajaran’s paper[19], select the features from the dataset and construct the expression score formula as follows:

$$P_{er20}=P1_{ace}+P1_{winner}+P_{unf\text{-}err}+P_{net\text{-}pt\ won}+P_{scoreAD}+P_{score}+P_{serve}+P_{doubltfault}+P_{netpt} \tag{24}$$

$$P_{er2}\begin{cases}P_{er20}, & \text{if player 2 serve}\\ P_{er20}\times 0.9, & \text{if player 1 serve}\end{cases} \tag{25}$$

Suppose that the whole process is memoryless, and $H(r_t)$ is a constant. The selected mutation points were visualized, and the results were shown in Figure 3.

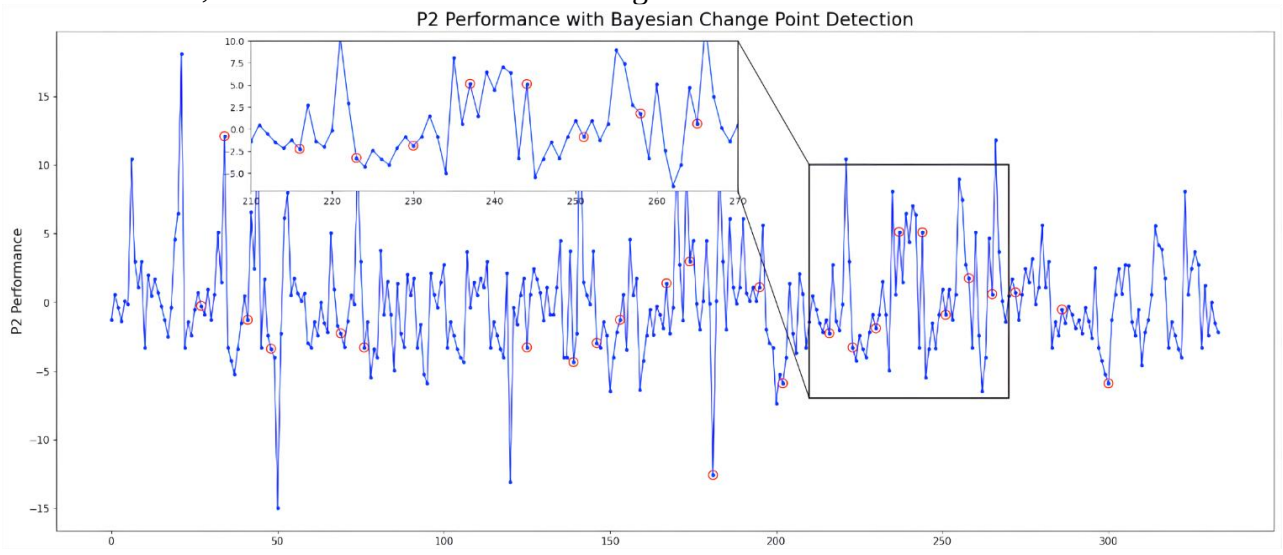


**Figure 3.** 1701’s selected mutation points

Figure 3 shows the selected mutation points during the course of the game. And these are mutation sites that have not been further screened.

### 3.2. Use the Random forest model to predict change points

According to the visualization results, the mutation point data is a small number of data, accounting for a small proportion, but the Random forest algorithm can overcome the problem of data imbalance and solve the binary classification problem well.

Select the features and select and construct the features according to three categories: expressiveness score, technical and tactical level, and “momentum”. The accuracy before and after parameter adjustment is obtained, and the results are shown in Table 1.

**Table 1.** Random forest accuracy means adjusting parameters

| before | after | Max_depth | Min_sample_leaf | Max_features | cv |
|---|---|---|---|---|---|
| 84% | 87% | 5 | 2 | 1 | 10 |

From Table 1, we can see that the Random forest accuracy is 87% which means our features selected are valuable.

The importance of derived features is shown in Figure 4.

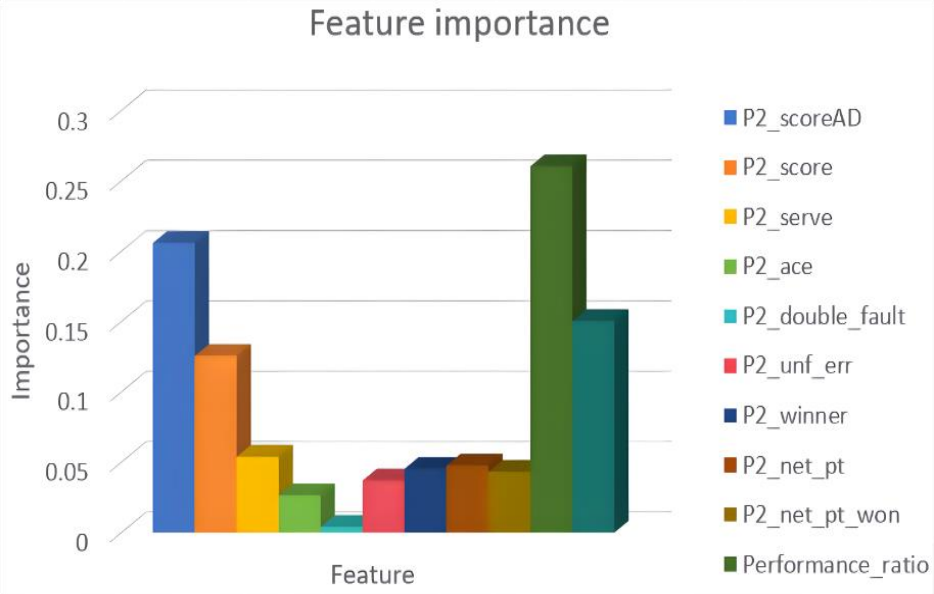


**Figure 4.** The importance of derived features

ScoreAD refers to score, performance_ratio refers to “momentum”, and the rest refers to Technical and tactical level. It conforms to the above feature construction strategy.

From the aspect of feature importance, "momentum" has great influence on prediction and is an important basis for distinguishing abrupt points from non-abrupt points.

### 3.3. Testing the generalization ability of the model

We selected several other competitions[20], repeated the process of 3.1 and 3.2, and obtained the accuracy before and after adjustment, as shown in Table 2:

**Table 2.** The comparison of results before and after was eliminated

| Game code | before | after | Game code | before | after |
|---|---|---|---|---|---|
| 1310 | 75% | 84.92% | 1501 | 68.08% | 83.07% |
| 1316 | 73.68% | 85.07% | 1504 | 73.47% | 85.71% |
| 1405 | 71.13% | 84.22% | 1601 | 76.35% | 86.17% |
| 1408 | 72.13% | 83.77% | 1602 | 72.53% | 83.69% |

It can be seen from the above table that our model has a good generalization degree in the same type of competition and has a certain universal significance.

Because of its high accuracy, it not only shows that the feature construction strategy of 3.2 is reasonable but also shows that there are high-quality features at the mutation point that are significantly different from other points, which is difficult to find in the traditional probability model.

### 3.4. Search and removal of recessive mutation points

Because not all mutations detected by the Bayesian online change point detection method are of practical significance, we made a partial correction that the mutation sites were divided into recessive mutation sites and dominant mutation sites.

Taking the actual state of the athlete at the mutation point as the classification standard, if we divide the athlete's state into a stable state and an unstable state, then the mutation point in the athlete's stable state should be recessive, and in fact, should be removed. We take the difference in the athlete's expressive force at each point as the representation of stability. The difference is shown in Figure 5.

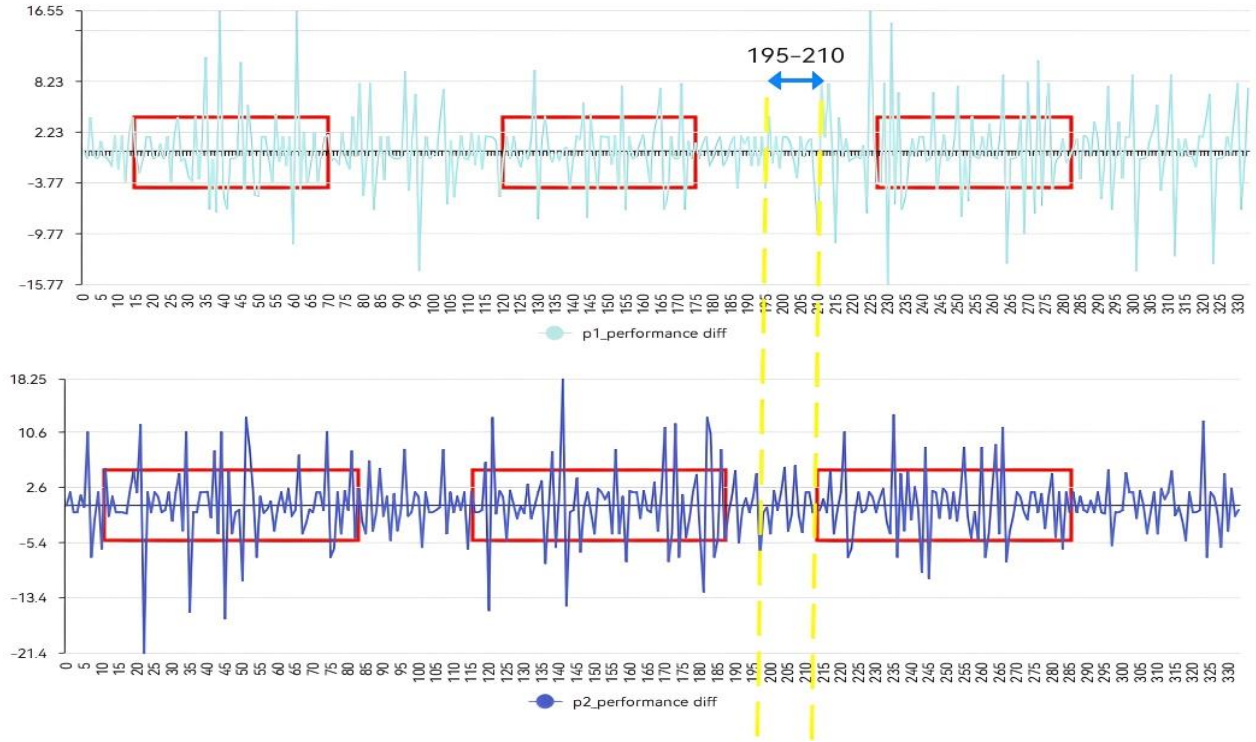


**Figure 5.** Difference of performance

When the expressive force difference fluctuates slightly, its state can be considered to rise or fall steadily, if there is a mutation point at this time, it is considered to be a recessive mutation point. On the contrary, when it fluctuates greatly, it can be considered that its state change trend is sometimes good and sometimes bad. Given the actual game scenario, the mutation point in this case is just the so-called dominant mutation point. After visualizing the difference of expressive force with the change of point, it can be seen in Figure 5 that there are some phased changes in the degree of fluctuation, which are marked by red wireframe. So based on the above analysis, for example, we can remove the point between the two yellow dashed lines.

The gating mechanism of the Bi-LSTM algorithm can realize the flow, forgetting and updating of data well, so the algorithm can better deal with the complex patterns and dependencies in time series. After carrying out the time series prediction based on the Bi-LSTM algorithm for expressive difference, it is obvious that the predicted value is consistent with the actual value size and trend at and near the peak value of the actual data on the test set, and RMSE of the test set =2.9561. As shown in Figure 6.

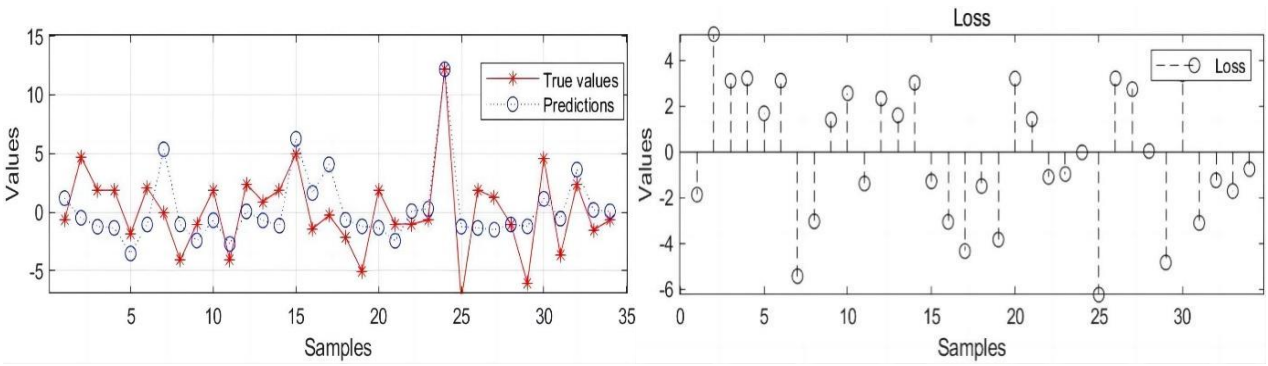


**Figure 6.** Test set

The results indicate that the generation of violent fluctuations has its regularity or even periodicity. Through the above analysis, it is proved that the definition of hidden mutation points is in line with reality and the rationality of removing hidden mutation points.

Taking Contest 1701 as an example, combined with the points that have been detected as mutation points, we removed some recessive mutation points and used the Random forest algorithm again to predict mutation points, and the data showed that its accuracy was improved. Then, we tested the model with the data of other men's singles matches in 2023 Wimbledon and the result is shown in Table 3.

**Table 3.** Statistics before and after removal of recessive mutation points

| Game code | before | after |
|---|---|---|
| 1310 | 84.92% | 86.66% |
| 1316 | 85.07% | 84.00% |
| 1405 | 84.22% | 86.23% |
| 1408 | 83.77% | 84.89% |

As shown in the Table 3 ,in other competitions, we found that the accuracy rate either improves or does not change much, indicating that the corresponding data at the hidden mutation point is useless data or even interference data.

### 3.5. Construct the weight factor to optimize feature about the "momentum"

According to the Taylor expansion model, set the weight factor equal to F(even performance, performance score ratio, cumulative performance score). The weight factor here corresponds to performance_ratio. Supposing the first derivative of F(even performance) is greater than 0, the second derivative is less than 0, and the actual performance score ratio is far less than the cumulative performance score. So it can apply an approximation of F. The alternating points of 4-0kill and 0-4kill are taken as the zeros of F and brought into the solution for a1,a2,a3. From this, we get the function expression of F, and then calculate the weight factor of alternating points and its neighboring points, and finally reconstruct the feature of performance_ratio.

We took the match with the most the case of 4-0 and 0-4 in all matches repeated the process from 3.1 to 3.4, and finally got the prediction accuracy of the mutation point as shown in Table 4.

**Table 4.** Comparison of results before and after weight adjustment

| | |
|---|---|
| before | 85% |
| after | 88% |

From Table 4, we know that after the weight adjustment, the accuracy has improved slighted. This also proves that it is reasonable for us to adjust the weight of the selected features.

## 4. Conclusions and outlooks

In this paper, we mainly solve the problem of mutation detection of player performance scores in tennis. We build a Bayesian online change point detection model to find the mutation points in the course of the match and then use the Random forest model for training prediction. Then we used the Bi-LSTM model to search and remove recessive mutation sites to improve the accuracy of the Random forest model. Finally, we use the Taylor expansion model to apply weights to the influencing factors of the expressiveness score, so as to further improve the training accuracy of the Random forest model.

In this paper, we construct feature engineering, which consists of three parts: ScoreAD refers to score, performance ratio refers to “momentum”, and the rest refers to the Technical and tactical level. This construction method can be applied not only to tennis but also to other sports.

For any sport, we need to determine which features of the current data will be affected by the historical data of this feature based on the sports category and competition rules, and this feature corresponds to "momentum", which plays an important role in the competitions. What’s more, we should use machine learning algorithms to study some of the quantities related to extreme reversals and periodic changes in the game.

Since we have only conducted a theoretical exploration of this research perspective based on mutation points detected by specific algorithms and have made necessary simplification of the problem, there are still the following outlooks: The prior probability of Bayesian online change point detection is difficult to determine. Additionally, the mechanism of the mutation point is not clear and needs more practice tests. The last is that it does not take into account the historical performance of the athlete in other competitions and the impact of his relationship with a particular opponent on the current competition.

## References

[1] Kovalchik S A. Searching for the GOAT of tennis win prediction[J]. Journal of Quantitative Analysis in Sports, 2016, 12(3): 127-138.

[2] Newton P K, Keller J B. Probability of winning at tennis I. Theory and data[J]. Studies in applied Mathematics, 2005, 114(3): 241-269.

[3] Barnett T, Clarke S R. Combining player statistics to predict outcomes of tennis matches[J]. IMA Journal of Management Mathematics, 2005, 16(2): 113-120.

[4] Knottenbelt W J, Spanias D, Madurska A M. A common-opponent stochastic model for predicting the outcome of professional tennis matches[J]. Computers & Mathematics with Applications, 2012, 64(12): 3820-3827.

[5] Lu J, Wang C, Zhang J, et al. A sequential Bayesian change point detection procedure for aberrant behaviours in computerized testing[J]. British Journal of Mathematical and Statistical Psychology, 2024, 77(1): 31-54.

[6] Prabpon N, Homsud K, Vatiwutipong P. Nonparametric Bayesian online change point detection using kernel density estimation with nonparametric hazard function[J]. Statistics and Computing, 2024, 34(2): 1-14.

[7] Adams R P, MacKay D J C. Bayesian online changepoint detection[J]. arxiv preprint arxiv:0710.3742, 2007.Sun Z, Wang G, Li P, et al. An improved Random forest based on the classification accuracy and correlation measurement of decision trees[J]. Expert Systems with Applications, 2024, 237: 121549.

[8] Siqueira R G, Moquedace C M, Fernandes-Filho E I, et al. Modelling and prediction of major soil chemical properties with Random Forest: Machine learning as tool to understand soil-environment relationships in Antarctica[J]. Catena, 2024, 235: 107677.

[9] Arshad A, Mirchi A, Vilcaez J, et al. Reconstructing high-resolution groundwater level data using a hybrid random forest model to quantify distributed groundwater changes in the Indus Basin[J]. Journal of Hydrology, 2024, 628: 130535.

[10] Ozsahin D U, Ameen Z S, Hassan A S, et al. Enhancing explainable SARS-CoV-2 vaccine development leveraging bee colony optimised Bi-LSTM, Bi-GRU models and bioinformatic analysis[J]. Scientific Reports, 2024, 14(1): 6737.

[11] Xue Y, Yao J. Evaluation of Stock Market Risk Model Based on Random Forest+ Two-Way LSTM[C]//Proceedings of the 2nd International Academic Conference on Blockchain, Information Technology and Smart Finance (ICBIS 2023). Atlantis Press, 2023: 912-922.Gozuoglu A, Ozgonenel O, Gezegin C. CNN-LSTM Based Deep Learning Application on Jetson Nano: Estimating Electrical Energy Consumption for Future Smart Homes[J]. Internet of Things, 2024: 11148.

[12] F. Qu, Y. Shi, C. Xie, J. Zhang, W. Li and C. Lin. Improvement of power equipment defect text quality based on improved BI-LSTM[J]. CSEE Journal of Power and Energy Systems, 2024.

[13] Redhu P, Kumar K. Short-term traffic flow prediction based on optimized deep learning neural network: PSO-Bi-LSTM[J]. Physica A: Statistical Mechanics and its Applications, 2023, 625: 129001.

[14] Abdul-Hassan N Y, Kadum Z J, Ali A H. An Efficient Third-Order Scheme Based on Runge–Kutta and Taylor Series Expansion for Solving Initial Value Problems[J]. Algorithms, 2024, 17(3): 123.

[15] Chen L, Zhao J, Lian H, et al. A BEM broadband topology optimization strategy based on Taylor expansion and SOAR method—Application to 2D acoustic scattering problems[J]. International Journal for Numerical Methods in Engineering, 2023, 124(23): 5151-5182. Diamandis E P. Unveiling the Right Side: How to Win Wimbledon Championships: Creating Beklof and Vamos[J]. Clinical Chemistry, 2009, 55(6): 1253-1254.

[16] Jatra R, Fernando D D. The understanding of court tennis rules for participants of licensing tennis umpire training[J]. Jurnal SPORTIF: Jurnal Penelitian Pembelajaran, 2019, 5(1): 70-79.

[17] Gupta K, Krishnamurthy V, Deb S. What elements of the opening set influence the outcome of a tennis match? An in-depth analysis of Wimbledon data[J]. IIMB Management Review, 2024.