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

Research on momentum quantification and influencing factors based on machine learning

Zimu Jia [1,*], Zhuoye Li [2]

[1] School of Mathematics, Hohai University, Nanjing, China, 211100

[2] College of Water Conservancy and Hydropower Engineering, Hohai University, Nanjing, China, 211100

* Corresponding author: jzm200404@gmail.com

Abstract. From the score data of the Wimbledon tennis matches, we observed unexpected score fluctuations and even set-level variations for the leading side. This article aims to evaluate which player performs better at any given moment during the match and quantify the extent of their performance advantage, which this paper refers to as momentum. Firstly, this paper defines momentum as the player's winning probability, convert match data into a Markov chain, and establish a time-predictive model based on the score timeline using Hidden Markov Models. Upon importing match data into our model, this paper successfully computes the magnitude of momentum for each player. Subsequently, this paper utilizes exploratory factor analysis, linear regression analysis, and Pearson correlation coefficient analysis to calculate the degree of association between momentum impact factors and player performance, observing that variables such as serve_no, point_victor, p1_winner, winner_shot_type, p1_net_pt_won, p1_distance_run, rally_count, and speed_mph exhibits strong correlations with momentum. For example, higher serve speeds are likely to increase a player's chance of winning, but break points faced, unforced errors, and distance run by the player may reduce the overall probability of a player winning. When applying the model output in actual matches, coaches and players can formulate strategies for serving selection and tactical arrangements based on the information provided by the model to maximize the utilization of momentum changes and develop targeted strategies against opponents' weaknesses and habits.

Keywords: Hidden Markov Models; Exploratory Factor Analysis; Linear Regression Model.

1. Introduction

In the men's singles final of the 2023 Wimbledon Championships, 20-year-old Spanish newcomer Carlos Alcaraz defeated 36-year-old Novak Djokovic. Data from this match reveals that the dominant player sometimes experiences remarkable fluctuations in points or even game scores, attributed to momentum. Momentum in tennis typically refers to the psychological and physiological enhancements experienced by players, including self-efficacy, motivation, attention and activity levels, pace, posture, or frequency.

Morgulev et al. [1] conducted a regression analysis on the hit rates of basketball players in three consecutive free throws and found no evidence to support the existence of momentum. Jamil et al. [2] suggested that mid-season breaks lasting less than 13 days do not affect technical performance, but longer breaks may impact momentum and lead to deteriorated performance. Munoz et al. [3] concluded that the influence of momentum in sports leagues depends on background and motivational factors. Vergin et al. [4] argued that momentum is overly emphasized by both sports participants and observers as a causal factor in sports outcomes. Noel et al. [5] proposed a shift towards examining momentum from a dimensional perspective and separating momentum from binary notions of winning and losing. Permutt et al. [6] explored the short-term performance enhancement following negative events and discussed the effectiveness of breaks and the existence of momentum. Den Hartigh et al. [7] provided empirical insights into the dynamic nature of team momentum. Steeger et al. [8] used entropy as an unbiased measure to refute the concept of momentum in sports, suggesting that the theory of momentum across a season is a fallacy that should not influence behavior.

The aforementioned studies were unable to quantify the magnitude of momentum and the specific factors through which momentum influences match outcomes. By establishing a model capable of describing momentum, quantifying the momentum of players during matches, and identifying the factors influencing match outcomes, this paper has demonstrated the existence of momentum and its impact on match results.

2. The fundamental of model

2.1. The structure of HHM model

Hidden Markov Models (HMMs) [9] are statistical models used to describe the underlying structure of sequential data, where the states are not directly observable but can emit observable symbols with certain probabilities. HMMs are widely applied in various fields such as speech recognition, natural language processing, bioinformatics, and finance. The model structure is illustrated in the Figure 1.

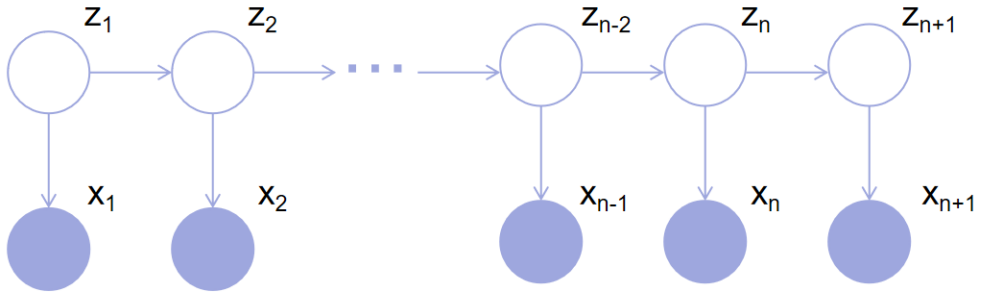

Figure 1. Structure of HHM model

The specific formula of the model HHM is as follows.

$$P_{st} = P(x_i = t \mid x_{i-1} = s) \tag{1}$$

Where represents a conditional probability, indicating the probability of the current state being t given that the previous state was s.

Since there are multiple states at each different time step, there are several possibilities for transitioning from the previous state to one of the current states. Therefore, all the conditional probabilities form a matrix known as the "transition probability matrix." For example, if there are n states at each time step, each state from the previous time step may transition to any of the n states at the current time step. Hence, there are a total of n*n possibilities, organized into a matrix as follows.

$$\begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1j} & \cdots \\ p_{21} & p_{22} & \cdots & p_{1j} & \cdots \\ \vdots & \vdots & & \vdots & \\ p_{i1} & p_{i2} & \cdots & p_{ij} & \cdots \\ \vdots & \vdots & & \vdots & \end{bmatrix} \tag{2}$$

2.2. The constitution of EFA Model

Exploratory Factor Analysis (EFA) [10] is a statistical technique used to explore the underlying structure of a set of variables. It is commonly employed in fields such as psychology, sociology, and education to identify latent factors that may explain the correlations among observed variables. The goal of EFA is to uncover the latent constructs or factors that best account for the patterns of covariation observed in the data. Its structure diagram is shown in Figure 2.

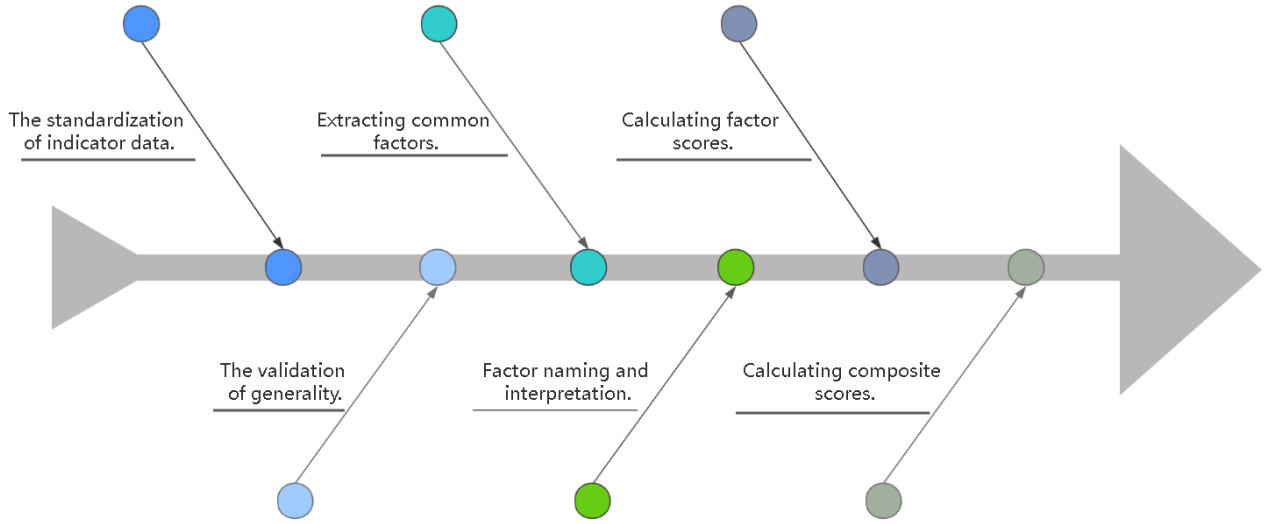

Figure 2. Structure of EFA model

First, this paper selects a series of feature variables for factor analysis. Then, factors are typically rotated to achieve a loadings matrix where each row represents an observed variable and each column represents a factor. The values in the matrix indicate the degree of correlation between each observed variable and the latent factor. Negative values denote negative correlations, positive values denote positive correlations, and the magnitude of the values reflects the strength of the correlation.

Finally, this paper obtains the communalities of the factors, which represent the portion of the variance of each variable that can be explained by the common factors. These values range from 0 to 1. A value close to 1 indicates a strong correlation between the variable and other variables, while a value close to 0 suggests a weak correlation between the variable and other variables.

2.3. The principle of linear regression model

Linear regression [11] is a fundamental statistical method used for modeling the relationship between a dependent variable Y and one or more independent variables $X_1, X_2, \ldots, X_P$. It assumes that this relationship can be well approximated by a linear function. The formula of the Linear Regression Model is as follows.

$$w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_D \\ w_{D+1} \end{bmatrix} X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,D} & 1 \\ x_{2,1} & x_{2,2} & \cdots & x_{2,D} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{N,1} & x_{N,2} & \cdots & x_{N,D} & 1 \end{bmatrix} y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \tag{3}$$

Where vector w represents the weights of various features in the model; each row of matrix X corresponds to a sample, with D+1 features in each sample denoting the values of the sample across different dimensions; vector y represents the actual values.

$$L(w) = (Xw - y)^T (Xw - y) = \| Xw - y \|_2^2 \tag{4}$$

The formula represents the vector representation of the loss function for multivariate linear regression which is used to assess the model's goodness of fit.

Linear regression provides a simple yet powerful framework for modeling the relationship between variables. Despite its simplicity, it remains widely used in various fields for its interpretability and ease of implementation.

2.4. Pearson Correlation Coefficient Analysis.

Pearson correlation coefficient analysis [12] is a widely used method in statistics to assess the strength and direction of the linear relationship between two continuous variables. The Pearson correlation coefficient is a measure of the linear relationship between two variables, ranging in value from -1 to +1. A negative sign indicates negative correlation, a positive sign indicates positive correlation, and values close to 0 indicate almost no correlation.

$$r_{xy} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{(\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2})(\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \tag{5}$$

Where X_i and Y_i are the individual data points for variables X and Y, $\bar{X}$ and $\bar{Y}$ are the means of variables X and Y, respectively. The above formula represents the calculation of the Pearson correlation coefficient, which is the covariance of two variables divided by the product of their standard deviations.

3. Results

3.1. The probability of winning at each time point

This paper utilized a Hidden Markov Model (HMM) to model the scoring data along the timeline. This paper uses the data from the first match in the men's singles final of the 2023 Wimbledon Championships and assume that the player serving has a much higher probability of winning the point. This paper hypothesizes that the probability of the serving player winning the match is 60%, while the probability of the receiving player winning the match is 40%. Then, this paper constructs a state transition matrix to record the transitions between the current state and the next state, and convert it into transition probabilities.

This paper calculates the probability of a player winning at each state based on the known transition probabilities and the advantage probability of serving. Initially, for each state in the transition matrix, this paper initializes its win probability to 0. Then, considering all the next states corresponding to that state along with their probabilities, this paper computes the probability of a player winning at that state. Finally, this paper visualizes the probability of both players winning at each time point in the course of the first match, as shown in the figure 3.

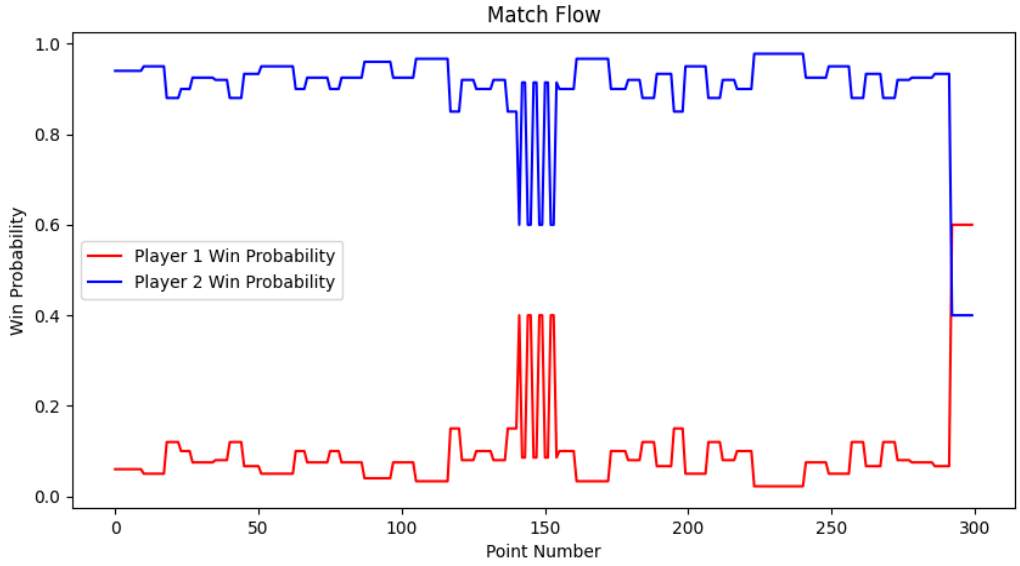

Figure 3. Win Probability Over Time Graph.

The graph depicts the win probability of both players at each time point during the match. The horizontal axis represents the time point, denoted as "point_no," while the vertical axis represents the win probability of the players. A higher win probability at a given time indicates better performance by the player, whereas a lower probability suggests poorer performance.

3.2. The results of factor analysis

This paper obtained the loadings matrix [13] through feature exploratory factor analysis (EFA) method, which reveals the relationship between observed variables and latent factors. The results are shown in the figure 4.

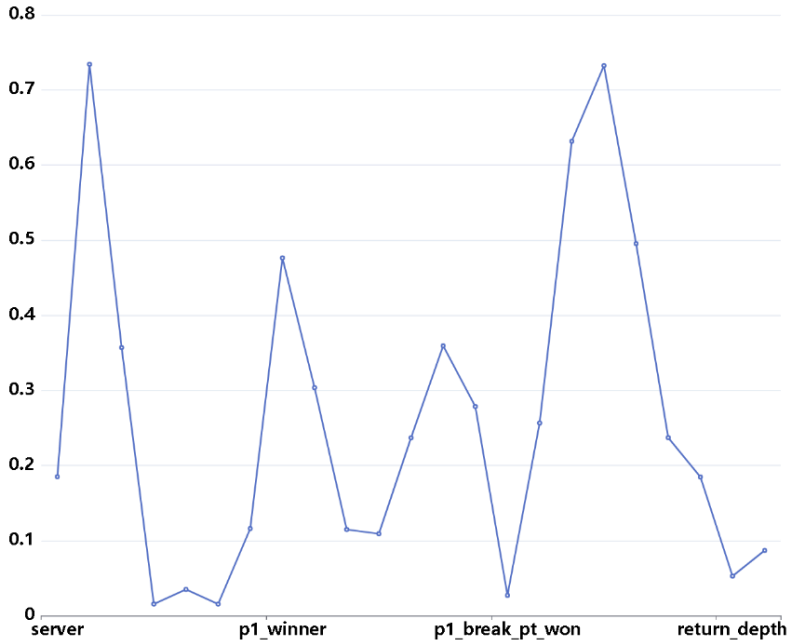

Figure 4. Line Plot of Communalities Factor.

From the above figure, it can be observed that variables such as serve_no, point_victor, p1_winner, winner_shot_type, p1_net_pt_won, p1_distance_run, rally_count, and speed_mph exhibit relatively high communalities of factor variance, approaching or exceeding 0.5. This indicates that there are some strong correlations among these variables.

On the other hand, variables like server, p1_points_won, game_victor, set_victor, p1_ace, p1_double_fault, p1_unf_err, p1_net_pt, p1_break_pt, p1_break_pt_won, p1_break_pt_missed, speed_mph, serve_width, serve_depth, return_depth, and p1_win_percentage demonstrate lower communalities of factor variance, approaching or falling below 0.3. This implies weaker correlations among these variables.

3.3. The conclusion of linear regression

This paper designated the target variable ‘p1_win_percentage’ as the dependent variable for the model. Subsequently, this paper created a linear regression model and fitted it using the training set to enable prediction of the target variable ‘p1_win_percentage’ in the test set. This paper then outputted the impact of each feature on the target variable, i.e., the feature coefficients. A positive coefficient indicates a positive correlation between the feature and the target variable, while a negative coefficient indicates a negative correlation. Finally, this paper used the trained linear regression model to predict the test set and calculated the mean squared error between the actual target values and the predicted values to evaluate the model's performance.

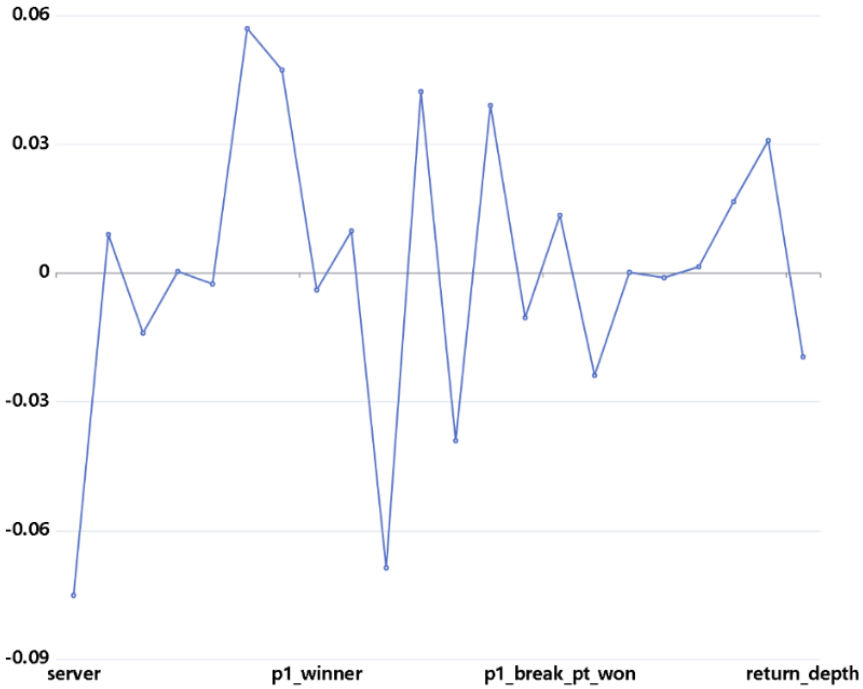

Figure 5. Line Plot of Feature Coefficients

From the data depicted in figure 5, it is evident that features such as server, point_victor, game_victor, p1_winner, p1_double_fault, p1_net_pt, p1_break_pt, return_depth, and p1_break_pt_missed exhibit a negative correlation with the target variable. On the other hand, features like serve_no, p1_points_won, set_victor, p1_ace, winner_shot_type, p1_unf_err, p1_net_pt_won, p1_break_pt_won, p1_distance_run, speed_mph, serve_width, and serve_depth show a positive correlation with the target variable. The mean squared error of the model is 0.009617541688135634, indicating a good level of fit.

3.4. The outcomes of Pearson correlation coefficient analysis

This paper selects specific feature columns from the dataset, then calculate the correlation between the selected features. Then, this paper generates a correlation matrix containing the Pearson correlation coefficients between the features. The visual representation of this matrix is shown in the figure 6.

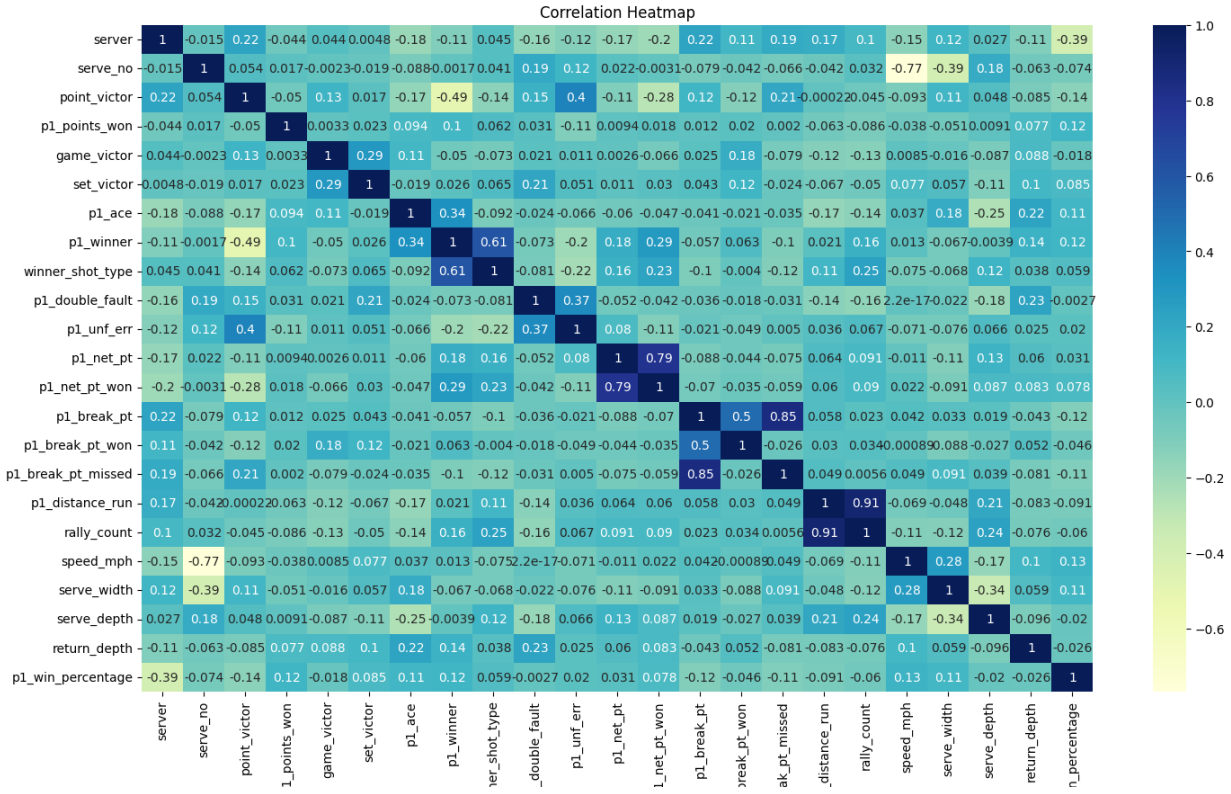

Figure 6. Correlation coefficient heatmap

The darker areas in the graph indicate stronger correlations, while the lighter areas represent weaker correlations. It can be observed that features such as serving points won, serve speed, and serve width show positive correlations with momentum, with "speed_mph" exhibiting a particularly significant positive correlation (0.13), suggesting that higher serve speeds are likely to increase a player's chance of winning. Conversely, variables such as break points faced, unforced errors, and distance run by the player show significant negative correlations with "p1_win_percentage," especially for "p1_break_pt" (-0.12), indicating that failure to convert break points significantly reduces the overall probability of a player winning. Other statistics, such as serve attempts, demonstrate relatively weaker correlations with momentum, suggesting a less strong association between these factors and a player's probability of winning a match.

4. Conclusions and Outlooks

This paper establishes a Hidden Markov Model (HMM) to describe momentum and quantifies the momentum of both players at any given moment during the match. Additionally, through exploratory factor analysis, linear regression analysis, and Pearson correlation coefficient analysis, we can observe that at the beginning of the match, variables such as server, double fault, ace, net point, and momentum exhibit the highest correlation, indicating that momentum primarily depends on these factors early in the match. As the match progresses, the correlation between break point, server speed, and momentum increases, suggesting that they become key factors for sustaining momentum during the middle phase of the match. In the late stages of the match, the correlation between points won and momentum increases significantly. At this point, every point becomes crucial, making it a key determinant for successfully winning the match. The practical application of this model in matches allows coaches to be aware of changes in momentum for both sides, enabling timely tactical adjustments for favorable strategies.

The HMM model in this study only considers two states (game score and set score) and does not include more detailed indicators such as serving side, serve angle, and other features. Future research will consider additional influencing factors in actual matches and incorporate them into the model to improve prediction accuracy.

References

[1] Morgulev E, Azar O H, Bar-Eli M. Searching for momentum in NBA triplets of free throws [J]. Journal of Sports Sciences, 2020, 38 (4): 390 - 398.

[2] Jamil M, McErlain-Naylor S A, Beato M. Investigating the impact of the mid-season winter break on technical performance levels across European football–Does a break in play affect team momentum? [J]. International Journal of Performance Analysis in Sport, 2020, 20 (3): 406 - 419.

[3] Munoz E, Chen J, Thomas M. Momentum in repeated competition: Exploiting the fine line between winning and losing [J]. Available at SSRN 3391748, 2019.

[4] Vergin R C. Winning Streaks in Sports and the Misperception of Momentum [J]. Journal of Sport Behavior, 2000, 23 (2).

[5] Noel J T P, Prado da Fonseca V, Soares A. A Comprehensive Data Pipeline for Comparing the Effects of Momentum on Sports Leagues [J]. Data, 2024, 9 (2): 29.

[6] Permutt S. The efficacy of momentum-stopping timeouts on short-term performance in the National Basketball Association [D]. 2011.

[7] Den Hartigh R J R, Gernigon C, Van Yperen N W, et al. How psychological and behavioral team states change during positive and negative momentum [J]. PloS one, 2014, 9 (5): e97887.

[8] Steeger G M, Dulin J L, Gonzalez G O. Winning and losing streaks in the National Hockey League: are teams experiencing momentum or are games a sequence of random events? [J]. Journal of Quantitative Analysis in Sports, 2021, 17 (3): 155 - 170.

[9] Mor B, Garhwal S, Kumar A. A systematic review of hidden Markov models and their applications [J]. Archives of computational methods in engineering, 2021, 28: 1429 - 1448.

[10] Goretzko D, Pham T T H, Bühner M. Exploratory factor analysis: Current use, methodological developments and recommendations for good practice[J]. Current psychology, 2021, 40: 3510 - 3521.

[11] Maulud D, Abdulazeez A M. A review on linear regression comprehensive in machine learning [J]. Journal of Applied Science and Technology Trends, 2020, 1 (2): 140 - 147.

[12] Baak M, Koopman R, Snoek H, et al. A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics [J]. Computational Statistics & Data Analysis, 2020, 152: 107043.

[13] de Winter J C F. EFA and PCA: Remarkable Phenomena and Practical Advice [J]. 2024.