Molecular & Cellular Biomechanics 2025, 22(2), 723.
https://doi.org/10.62617/mcb723

Article

Biomechanical spatio-temporal data analysis of football based on machine learning

Peng Zhou[1,*], Wenchao Hou[2], Yiqi Zhu[2], Weijie Zhang[2], Yitian Zhang[2]

[1] Physical Education Department, Nanjing University of Information Science and Technology, Nanjing 210044, China
[2] Reading Academy, Nanjing University of Information Science and Technology, Nanjing 210044, China
*** Corresponding author:** Peng Zhou, xiao20150400@163.com

CITATION

Zhou P, Hou W, Zhu Y, et al. Biomechanical spatio-temporal data analysis of football based on machine learning. Molecular & Cellular Biomechanics. 2025; 22(2): 723. https://doi.org/10.62617/mcb723

ARTICLE INFO

Received: 5 November 2024
Accepted: 21 November 2024
Available online: 10 February 2025

Abstract: With the advent of the era of big data, how to analyze the massive data of players' passing, shooting and position in football matches more effectively has become a new topic for the development of football. Machine learning algorithm has been widely used in various fields in recent years, relying on its strong data processing ability. Football match data analysis based on machine learning can effectively mine the effective characteristics of football data and better assist coaches' tactical arrangement, personnel arrangement and player evaluation. In this paper, machine learning algorithms such as clustering algorithm, classification algorithm, Markov chain model and kernel density estimation algorithm are used to analyze the spatio-temporal data of players' passing, shooting and position in football match. Compared with the traditional data analysis methods based on simple statistics, the method in this paper has more intuitive visualizations and deeper data insights. This approach is instrumental in guiding tactical planning and personnel strategies in football. Additionally, by integrating biomechanics into the analysis, we enhance our understanding of player performance. Biomechanical factors such as movement patterns, force application, and body mechanics play a critical role in a player's effectiveness on the field. Incorporating physiological data, including players' heart rate and movement intensity, allows for a more holistic perspective on performance, addressing potential fatigue and injury risks. By analyzing how biomechanics influence spatio-temporal data—such as optimal angles for passing, shooting mechanics, and body positioning during play—we can provide actionable insights into player training and development. This comprehensive approach not only improves tactical decision-making but also fosters player longevity and performance optimization.

Keywords: big data; machine learning; football data analysis; physiological data; biomechanics; player performance optimization

1. Introduction

With the development of sensor technology, video positioning technology and other football match data acquisition technology, the cost of time and space related to football match and event data collection has been gradually reduced, and football has gradually entered the "big data era". Using big data to analyze football matches has become a new topic in the development of competitive football in the new era [1,2]. Relying on its strong data processing and analysis ability, machine learning can analyze the process of football matches, predict match results, evaluate football players and other functions. According to the analysis results, football coaches can optimize tactical arrangements, football players can evaluate sports performance, and fans can have a clearer understanding of football matches through analysis. In this paper, the open-source programming language Python and machine learning

algorithms such as clustering, kernel density estimation and Markov chain are applied to analyze the temporal and spatial data such as passing, shooting and position of football matches on the open football data set.

2. Related work

2.1. Big data in football match

Football big data has a wide range, including the time and space data, the result data and the physiological data of the players in the football match and so on [3]. This paper focuses on the analysis of time and space data in football matches. The existing well-known football match big data providers include Wyscout, StatsBomb, Opta and so on. The data analyzed in this paper comes from the open-source datasets of Wyscout [4] and StatsBomb [5]. The time and space data of a football match describe the information of each ball processing in a match, including passing, shooting, standing, dribbling, confrontation and other events, as well as the corresponding player number, event stamp, coordinates on the field, the use of the left and right feet, success and other information.

Taking the StatsBomb football match data as an example, the content and format of football big data will be explained in detail below. StatsBomb football data include 70 seasons from 2003 to 2023, including leagues, World Cups, European Cups and so on. StatsBomb season data information has competition_id, season_id and other characteristics. Their specific meanings are shown in the **Table 1**.

Table 1. StatsBomb football competition data.

StatsBomb competition data	
competition_id	Competition number
season_id	Season number
country_name	The country where the event is held
competition_name	The name of the competition
competition_gender	Gender of the competition (male/female)
competition_youth	Whether the competition is a youth competition
competition_international	Whether the competition is an international competition
season_name	The name of season

StatsBomb match data information contains 64 features such as match number (match_id) and match date (match_date), which describe the details of all competitions in a season, as is shown in **Table 2**. Using competition_id = 72 and season_id = 107 as screening criteria to retrieve the 2023 Women's Football World Cup, we get a table with 64 rows and 52 columns, representing all the match information of the tournament. The following table lists the first six features of the match information.

Table 2. StatsBomb football match data.

StatsBomb match data	
match_id	Match number
match_date	Match date
kick_off	Kick-off time
home_score	The score of home team
away_score	The score of away team
competition_stage_name	Stage of competition

StatsBomb event data, as is described in **Table 3**, unfolds the data in a match in time, including the following data. (1) player position: provides real-time position information of players in the match, which can be used to analyze the trajectory and position strategy of players. (2) passing data: record the passing between players, including the starting point, the end point, the passing distance, the type of passing and so on. This is very useful for analyzing the team's passing fluency and the ability to create opportunities. (3) shooting data: including shooting position, shooting mode, scoring and other information. This can be used to analyze the offensive effect of the team and the shooting skills of the players. (4) foul data: record the fouls that occurred in the match, including the location of fouls, fouls players and other information, help to understand the team's physical fitness, discipline and defensive performance. Other event data, such as interception, clearance, steals, etc., provide a more comprehensive analysis of the match. The following table lists the partial characteristics of the event data.

Table 3. StatsBomb football event data.

StatsBomb event data	
timestamp	Timestamp of the event
duration	The first/second half
type_id	Type of the event
possession_team_name	Name of possession team
player_id	The player number involved in the event
player_name	The player name involved in the event

2.2. Machine learning and football

Machine learning algorithm is a kind of algorithm that automatically analyzes and obtains rules from data and uses laws to predict unknown data [6,7]. With the arrival of the era of big data, machine learning has been widely used in data mining, computer vision, natural language processing and other fields. Machine learning has significantly enhanced football match data analysis, offering deeper insights into player performance, team tactics, and game dynamics. Leveraging large-scale spatio-temporal data such as player positions, ball movement, and interactions, machine learning models uncover patterns often missed by traditional statistics [4]. Due to its strong ability, the combination of machine learning and football have gain increasing attention. In the following, we summarize some related works.

(1) Performance analysis of players [8–10]. The application of machine learning in player performance analysis can deeply mine match data, so as to evaluate the technical level and tactical execution ability of players. For example, by tracking a player's running trajectory, passing accuracy and shooting efficiency, coaches and analysts can identify a player's strengths and improvement. This helps to personalize the training program so that players can get a giant development in key areas.

(2) Analysis of the characteristics of the opponent's competition [11–13]. Through the analysis of the historical match data of the belligerent opponent, such as passing data, shooting data, position data and so on, we can reveal the opponent's tactical orientation, common offensive path and defensive strategy. This in-depth opponent analysis provides a basis for the team to formulate competition strategies, so that it can better deal with the strengths and weaknesses of opponents and improve the chances of winning the match.

(3) Injury risk prediction [14–16]. Using machine learning to analyze the health data of players, we can predict the injury risk that players may face. This helps the team to take personalized preventive measures, adjust the training plan, minimize the possibility of players' injuries, and ensure that the team has a stronger squad.

(4) Competition result prediction [17]. Through the in-depth analysis of the historical data of teams and players, the machine learning model can provide the prediction of the results of the match. This has practical reference value for fans, gambling companies and clubs in decision-making, resource allocation and strategic planning, and increases the scientific nature of decision-making.

In summary, machine learning techniques offer a robust framework for football data analysis, covering clustering, event recognition, and predictive modeling. As football data collection methods continue to improve, such as through advanced GPS tracking, video analytics, and wearable sensors, machine learning models will reveal new dimensions of tactical insights, optimize training, and support data-driven decision-making. These techniques are reshaping how teams analyze, plan, and engage with the game, broadening the scope and impact of sports analytics in modern football.

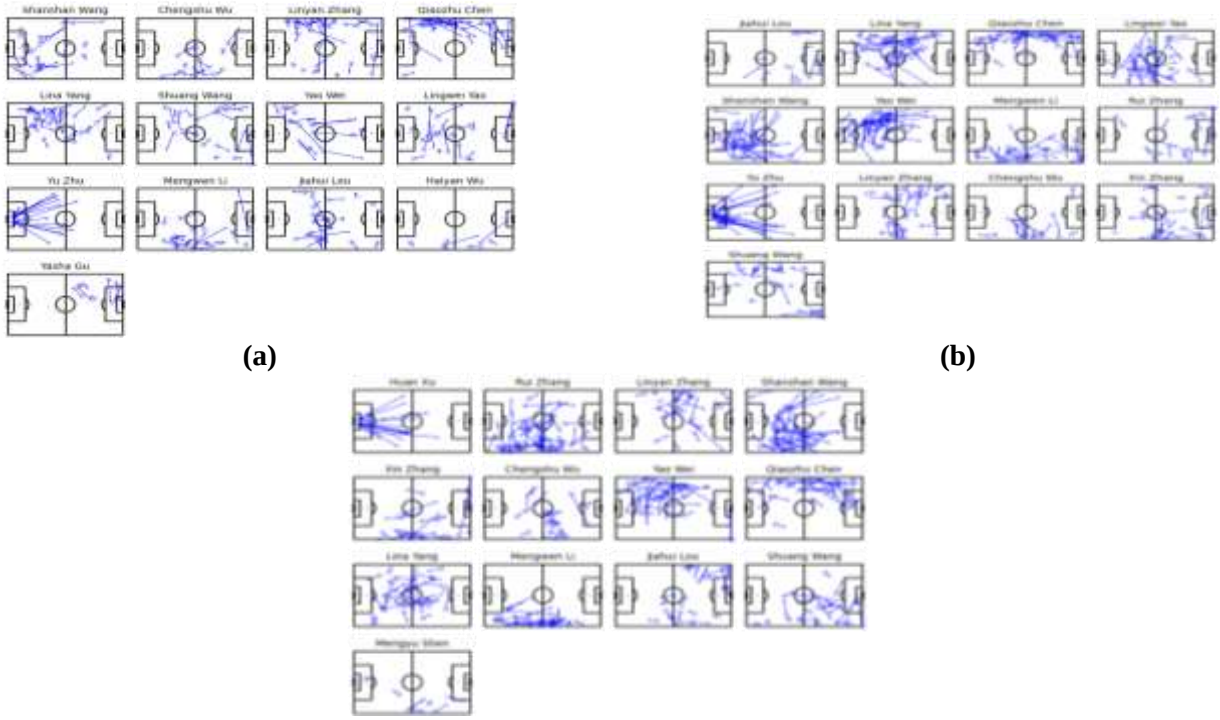
3. Big data analysis method of football based on machine learning

Football big data is huge and complex, including the time and space data, the result data, the physiological data of the players in the football match and so on [18,19]. This paper focuses on the analysis of the time and space data in the football match, including passing, shooting, position and other data.

This section uses machine learning algorithms to analyze the published data of the 2023 Women's World Cup by StatsBomb and Wycout. We will use the visualization function of Python language to visualize the passing route, shooting, position and other data in the 2023 Women's World Cup. The clustering algorithm is also used to analyze the passing choices of each team in the 2023 Women's World Cup, so as to analyze the similarities and differences of women's football tactics in various countries. The kernel density estimation algorithm (KDE) is used to draw the heat map of players' activity in the stadium, so as to analyze the situation of the match. Finally, the Markov chain model is used to analyze the threat degree of players, so as to quantify the actual effect of attack.

3.1. Visual analysis of players' passing route

By visualizing the passing route choice of the players in each position in **Figure 1**, we can directly observe the passing direction, passing times, passing starting position and other information, and can also analyze the tactical dynamics on the court and the cooperative relationship between the players.



(a) **(b)** **(c)**

Figure 1. Selection of passing routes for Chinese women's football players in group stage. **(a)** China vs. England; **(b)** China vs. Haiti; **(c)** Denmark vs. China.

Visual analysis of football passing routes helps football teams, coaches and analysts understand the match in depth. By visualizing the passing route, we can clearly analyze the main direction of the players passing the ball in each position, which is helpful to identify the overall tactical orientation of the team, such as whether they are more inclined to attack through midfield or focus more on the flank pass. The visualization of the passing route can also show the passing frequency of the players in each position. The passing route frequently chosen by a player may reflect the team's core passing strategy. Observing the starting position of the passing route helps to understand how the team attacks. The starting position of the passing of players in different positions can reveal the tactical characteristics of the team, such as organizing the attack in midfield or passing a long pass from the backcourt to the front court.

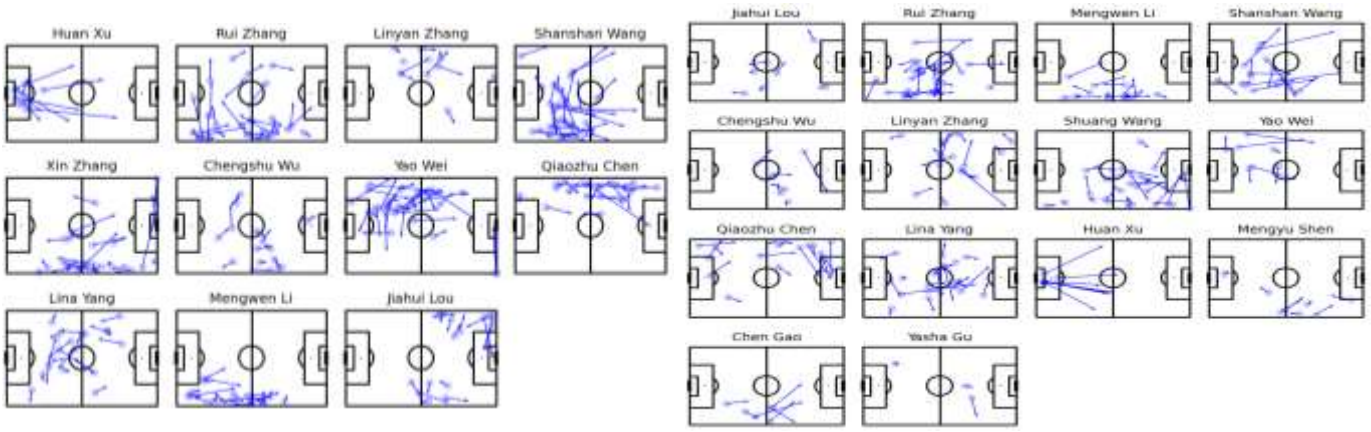
Visualization of the passing route is also helpful to analyze the cooperative relationship between players. By observing which players pass the ball frequently, the tacit combination in the team can be identified, which is very important to improve the tacit understanding and overall coordination of the team. Visual analysis of the passing route can also help the team discover the pressure points and gaps of the opponent. By identifying areas in which opposing players are actively defending, the team can adjust its tactics and look for loopholes of opponent. With the progress of the match, the

visualization of the passing route can also help teams and coaches track tactical changes, help to make timely adjustments and decisions in the match, in order to better adapt to the opponent's strategic changes.

Moreover, we can visualize the passing route in different stage of a match to analyze the tactical shift. The passing route maps for the Chinese team reveal a clear tactical shift between the first and second halves of the match, as is visualized in **Figure 2**. In the first half, the team's passing patterns are concentrated in the midfield and defensive areas, with players like Huan Xu and Xin Zhang actively distributing the ball from the back. This suggests a cautious approach focused on maintaining possession and building up play from defense, likely aimed at controlling the game's pace and assessing the opponent's tactics. However, in the second half, the passing routes show increased penetration into the opponent's half, particularly through players such as Rui Zhang and Yao Wei. This shift indicates a more offensive strategy, with the team attempting to exploit spaces in the opponent's defense and create scoring opportunities. The change from a conservative, possession-focused approach in the first half to an aggressive, forward-oriented style in the second half reflects a tactical adjustment aimed at shifting the game's rhythm. This adaptation demonstrates the team's flexibility in responding to the dynamics of the match, moving from defensive stability to offensive pursuit in response to the evolving situation on the field.

Through the visual analysis of the passing route, the football team can have a more comprehensive understanding of the tactical situation in the match and make more appropriate decisions. And optimize the tactical layout of the team, improve the chances of winning the match, and provide strong support for the team's tactical research and training.

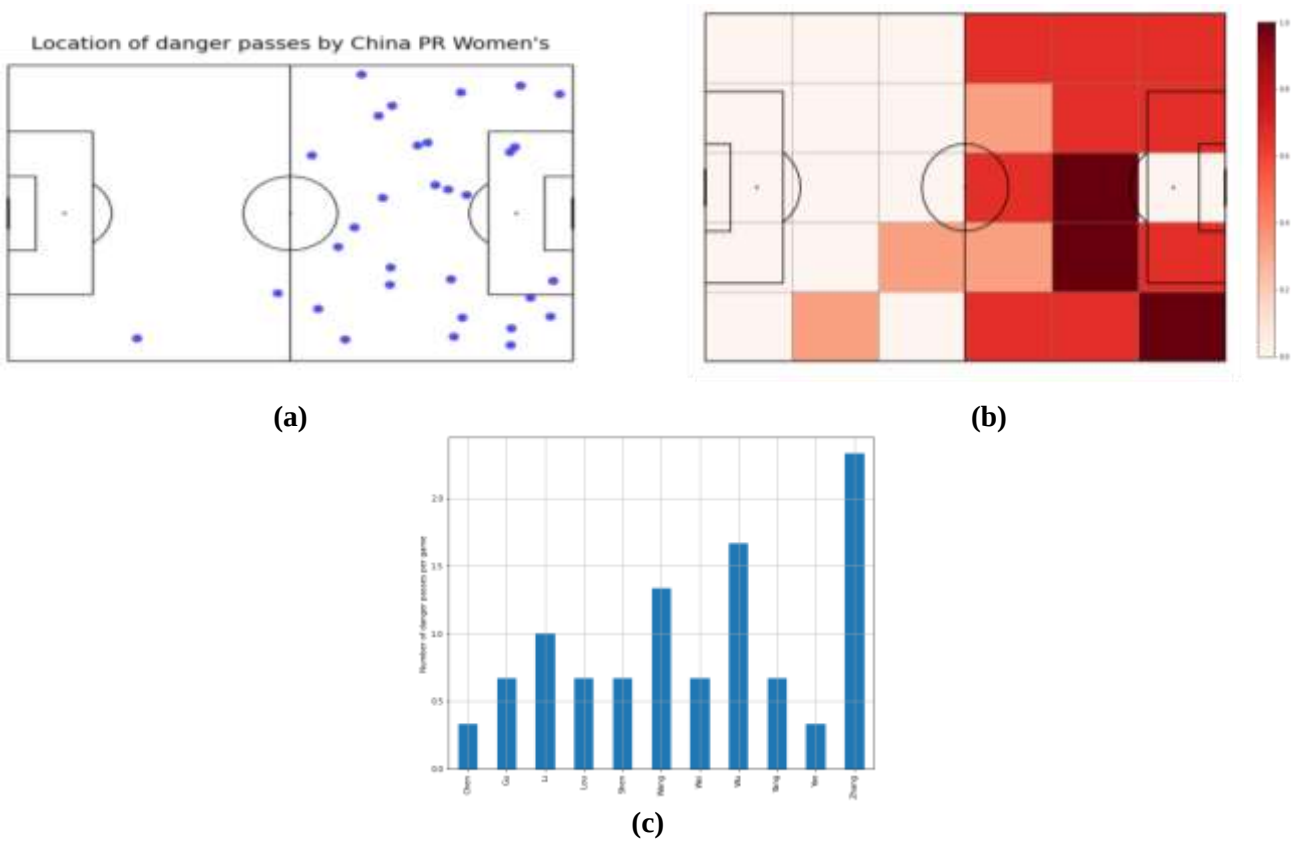
A killer pass is a pass that causes the last action to be a shot. This paper analyzes the killer passing data of Chinese women's football team in the 2023 Women's World Cup. Select the data of the three matches in which the Chinese women's football team participated, and then screen the passing data that caused the last action to shoot, and the visual results are shown in **Figure 2**.



(a) **(b)**

Figure 2. Analysis chart of killer pass of Chinese women's football team. **(a)** China's passing routes in the first half of Denmark vs. China match; **(b)** China's passing routes in the second half of Denmark vs. China match.

As shown in **Figure 3**, the location information, regional information and player information of the Chinese women's football team causing the threat pass can be obtained intuitively. The area of "threatening passing" of Chinese women's football team in the 2023 Women's World Cup is concentrated on the right side of the forbidden area, indicating that the right attack of Chinese women's football team is a great threat. According to the statistics of the players who caused the "threat pass", we can find that the player Zhang Linyan caused the most "threat pass". This information can help the Chinese women's football team to summarize after the match and make targeted arrangements for the subsequent matches.



(a) **(b)**

(c)

Figure 3. Analysis chart of killer pass of Chinese women's football team. **(a)** Position of killer pass; **(b)** Statistical chart of killer passing area; **(c)** Killer passer contribution ranking.

3.2. Visualize the player's shooting position

Visualization of players' shooting position provides an intuitive data display for players, coaches and fans, which is conducive to an in-depth understanding of players' shooting habits, offensive effects and opponents' defensive loopholes. By visualizing the shooting position of players in the form of charts or thermal maps, coaches can visually evaluate the shooting effects of players in different positions and understand the areas where players are more likely to score goals. It helps to optimize the tactics and guide the players to attack in the dominant area. In addition, the analysis of the shooting position near the opponent's goal can reveal the loopholes in the opponent's defense. Understanding the weakness of the opponent's goalkeeper or the weak defensive area of the opposing team is helpful to develop a more targeted attack strategy. Visualizing the shooting position also helps coaches to discover the shooting

strengths of each player. Some players may perform better at a distance from the goal, and understanding these characteristics will help to reasonably arrange the players' positions and roles. For fans, the intuitive data presented by heat maps and charts make it easier for them to track the team's offensive performance and improve the viewing of the match.

Figure 4 shows the activity range of the players of both sides in the Women's World Cup and the shooting location of the players in the strong teams of the world. As can be seen from the picture, compared with the strong European teams, the shooting position of the Chinese women's football team is concentrated outside the forbidden area and at the penalty spot, indicating that there is a certain gap between the Chinese women's football team and the European strong teams in terms of penetration. This data-driven analysis method not only helps to optimize the offensive side of the team, but also injects new vitality into the development of football analysis field.

(a) **(b)** **(c)**

Figure 4. Visualization of shooting position data. **(a)** China vs. England; **(b)** China vs. Haiti; **(c)** Denmark vs. China.

3.3. Application of clustering and classification algorithm in football data analysis

Clustering and classification algorithms are both essential in football data analysis, each offering unique benefits. Clustering is used to analyze real-time match data, helping teams identify similar game situations to adjust tactics effectively. By clustering player interactions, teams can detect well-coordinated pairs or groups, enhancing team synergy. Additionally, clustering helps clubs analyze players' skills and performance, aiding in targeted recruitment strategies to strengthen the team. Classification algorithms, meanwhile, focus on predictive insights based on historical data. They can predict outcomes like pass success rates or shot accuracy, allowing teams to optimize passing routes and make informed decisions. Classification also evaluates player performance, enabling personalized training and tactical adjustments. Together, clustering and classification provide a comprehensive, data-driven approach to improve decision-making, strategy, and team development in football.

The application of clustering and classification algorithm in football data analysis not only improves the understanding of teams and matches, but also provides a more

precise and scientific tool for decision-making, tactical adjustment and team construction.

3.3.1. Clustering and classification algorithm

(1) Clustering Algorithm. Clustering is to divide a data set into different classes or clusters according to a specific standard (such as distance), so that the similarity of data objects in the same cluster is as large as possible. at the same time, the difference of data objects not in the same cluster is as large as possible [19]. Different from classification, sequence tagging and other tasks, clustering divides samples into several categories through the internal relationship between data without knowing any sample tags in advance. *K*-means is the most common clustering algorithm, and its algorithm flow is as follows:

Algorithm 1 K-means Clustering Algorithm

1: **Data: Dataset**: $\mathcal{X} = \{x_1, x_2, \dots, x_M\}$, number of clusters$K$
2: **Results**: Cluster assignments $\{c_i\}_{i=1}^{M}$ and cluster centers$\{\mu_j\}_{j=1}^{k}$
3: **Initialize**: K cluster centers $\mu_1^{(0)}, \mu_2^{(0)}, \dots, \mu_k^{(0)}$ randomly;
4: **Set** iteration counter $t = 0$;
5: **Repeat**
Assignment Step:
for each data point $x_i \in \mathcal{X}$ do:
Assign $x_i$to the nearest cluster;

$$c_i^{(t)} = argmin_j \|x_i - \mu_j^{(t)}\|^2$$

end
Update Step
for each cluster $j = 1, \dots, k$ do
Update the cluster center μ_j

$$\boldsymbol{\mu_j^{(t+1)}} = \frac{1}{|\{x_i : c_i^{(t)} = j\}|} \sum_{i:c_i^{(t)}=j} \boldsymbol{x_i} \$$$

end
Increment iteration counter $t = t + 1$
6: **Until** *convergence* of $\{\mu_j\}_{j=1}^{k}$ or *maximum* iterations reached
7: **Return** Cluster assignments $\{c_i\}_{i=1}^{M}$ and cluster centers $\{\mu_j\}_{j=1}^{k}$

The *K*-means algorithm aims to partition a dataset into k clusters by minimizing the within-cluster variance, defined by the objective function:

$$J(c, \mu) = \sum_{i=1}^{M} \left\| x_i - \mu_{c_i} \right\|^2 \quad (1)$$

where$x_i$represents the i -th data point, μ_{c_i} denotes the center of the cluster to which $x_i$is assigned, and M is the total number of data points. This objective function encourages data points within each cluster to be as close as possible to their respective cluster centers. The *K*-means algorithm converges to a local minimum of the objective function, and each iteration either reduces the objective function or leaves it unchanged. Given that the objective function is lower-bounded by 0, and is discrete (since the data points are assigned to a fixed number of clusters), the algorithm is guaranteed to converge in a finite number of iterations, although it may not necessarily converge to the global minimum.

(2) Classification Algorithm. Classification is a process of assigning data objects into predefined categories or classes based on specific criteria (such as feature similarity), ensuring that each data object is accurately labeled according to its characteristics. Unlike clustering, which groups data without prior knowledge of labels, classification requires labeled data to train a model that can distinguish between different classes. This process leverages known sample tags to learn patterns in the data and to predict the class of new, unseen data. Classification is widely used in tasks such as image recognition, sentiment analysis, and medical diagnosis. Among various classification algorithms, Random Forest [20] is one of the most popular and robust methods due to its high accuracy and versatility. Random Forest is an ensemble learning technique that builds multiple decision trees during training and combines their outputs to improve prediction accuracy and reduce overfitting. By averaging the predictions of many trees, Random Forest provides a stable and reliable classification model, even when dealing with noisy or complex data. It is particularly effective for handling large datasets with numerous features and can automatically manage feature importance, offering interpretability in identifying which features contribute most to the classification decisions. This makes Random Forest suitable for a wide range of applications, including medical diagnosis, fraud detection, and image classification, where accuracy and resilience to overfitting are crucial.

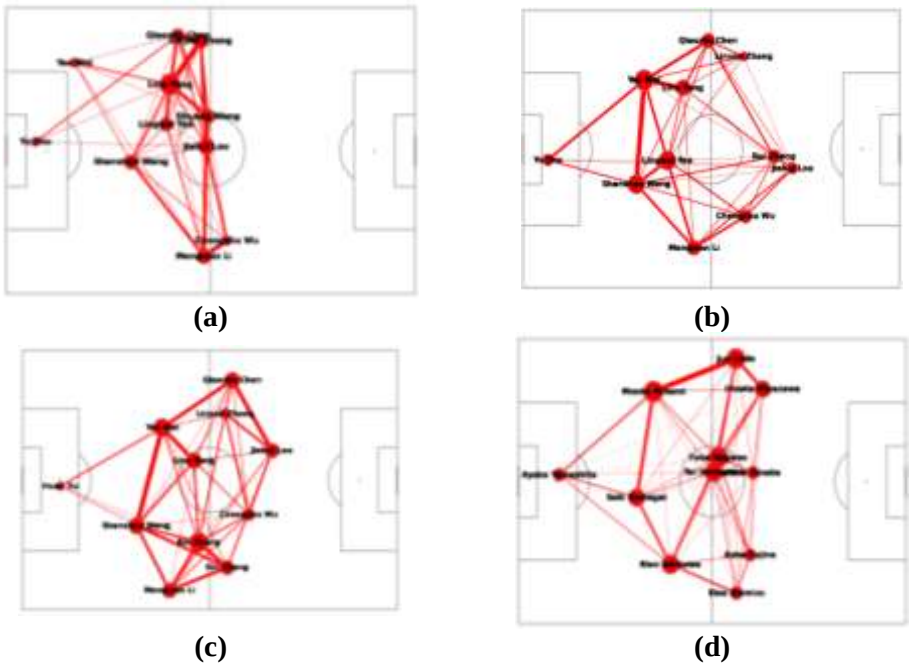
3.3.2. Position-relation analysis of players on the field based on clustering algorithm

This section combines position data and passing data to carry out position-contact analysis to analyze the position of team players in the match and the relationship between them. The *K*-means clustering analysis of the player position data in the match is carried out to get the clustering center, and the number of passes between players is counted, the position-contact data is obtained, and the player position-contact map can be visualized. Based on the player's position-contact data, it can help the coach evaluate the team's offensive and defensive strategies and identify the players' roles and responsibilities in the match. This is essential to ensure that each player is able to perform his or her specific duties in order to achieve the overall tactical objectives. By looking at the wiring chart, the coach can quickly identify weaknesses in the formation and make adjustments, such as changing the position of players, adjusting tactics or making substitutes.

Based on the above method, the data of the Chinese women's football group stage of the 2023 Women's Football World Cup are analyzed, and the player position-contact map is shown in the following figure.

Figure 5 shows the standing map of the formation of the Chinese women's football match in the 2023 Women's World Cup. From the standing map of the formation, we can see that when facing England, the formation of the Chinese women's football team is concentrated on its own half-court side, and is transferred to its own left by the opponent, and the distance between the two lines is relatively narrow. This picture shows that the Chinese women's football team is passive in the face of the English women's football team, and the opponent's attack is concentrated on the right. When facing the Haitian women's football team, the formation of the Chinese women's football team shows that there is a large distance between the midfield line

and the frontline. When facing the Danish women's football team, Chinese women's football team is attacking in favor of the left back.



(a) **(b)** **(c)** **(d)**

Figure 5. Standing map based on clustering algorithm. **(a)** China vs. England standing map of Chinese; **(b)** China vs. Haiti standing map of Chinese; **(c)** Denmark vs. China standing map of Chinese; **(d)** Japan vs. Zambia standing map of Japanese.

Another thing we can observed from the standing map is centralization. To quantitatively evaluate how central a football team is, we define the centralization index, denoted as C, as follows:

$$C = \frac{\sum_{i=1}^{N}(P_{\max} - P_i)}{10 \times T} \tag{2}$$

where

C: Centralization Index.

$P_{\max}$: Maximum number of successful passes made by a single player.

P_i: Number of successful passes made by player \(i\).

N: Total number of players.

$T = \sum_{i=1}^{N} P_i$: Total sum of successful passes by all players.

The centralization index measures the extent to which a team's passing depends on one or a few players, highlighting the concentration of passing activity. By normalizing the differences in pass counts across players, it provides a clear measure of team dynamics and balance. Calculating the centralization index for the China vs. England match, we find that China's index is 0.12, relatively high compared to England's 0.08, indicating that a particular player played a key role for China in that match.

In conclusion, it can be seen that the formation position in the match can be analyzed through the formation standing map, so as to assist the coach team to observe the degree of tactical execution in the match.

3.3.3. Analysis of passing characteristics based on clustering algorithm

This section processes all the match data of Chinese women's football team, Japanese women's football team and Spanish women's football team based on *K*-means clustering algorithm, and analyzes their competition characteristics. The data of the whole match are analyzed by *K*-means cluster analysis. The number of the total cluster is a hyper-parameter which requires fixing by human knowledge. We introduce silhouette coefficient, which for a single sample is calculated as follows:

$$s(i) = \frac{b(i) - a(i)}{max(\, a(i), b(i))} \tag{3}$$

where $a(i)$ is the average distance between sample i and all other points in the same cluster (intra-cluster distance).$b(i)$ is the minimum average distance between sample i and points in a different cluster, taken over all other clusters (nearest-cluster distance). The silhouette coefficient for the entire dataset is then calculated by averaging $s(i)$ over all samples:

$$S = \frac{1}{N}\sum_{i=1}^{N} s\,(i) \tag{4}$$

where *N* is the total number of samples. The silhouette coefficient vs. number of clusters (K) is visualized in **Figure 6**.

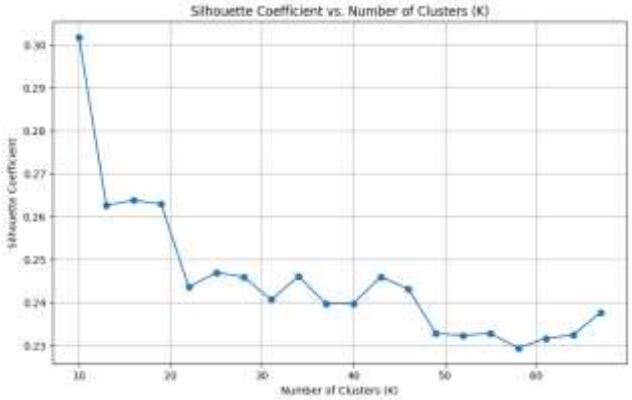


Figure 6. Cluster chart of 2023 women's world cup teams passing.

Based on the silhouette coefficient plot, selecting $K = 50$ as the number of clusters strikes a balance between cluster quality and granularity. As *K* increases, we observe a general decrease in the silhouette coefficient, indicating that clustering quality slightly diminishes. However, around $K = 50$, the silhouette coefficient stabilizes at a moderate value, which suggests that this choice maintains reasonable intra-cluster cohesion and inter-cluster separation. Although a lower *K* would yield higher silhouette scores, $K = 50$ provides finer granularity in the clustering, capturing more detailed patterns within the data. This choice thus allows for a more nuanced analysis, enhancing interpretability while preserving an acceptable clustering quality, making $K = 50$ a suitable selection for our analysis. The clustering results are visualized, and the arrow color is set according to the passing success rate, as shown in **Figure 7**.

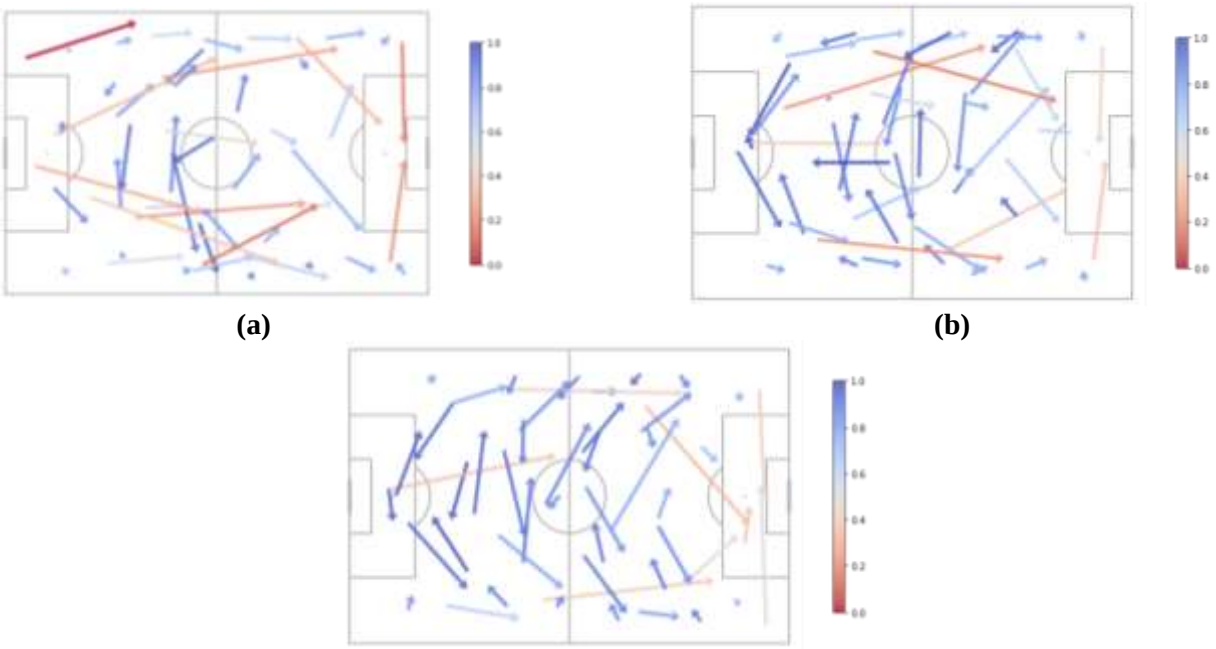


Figure 7. Cluster chart of 2023 women's world cup teams passing. **(a)** Chinese women's football team; **(b)** Japanese women's football team; **(c)** Spanish women's football team.

It can be seen from the results that the "intuitive killer pass" of the Chinese women's football team is less, that is, the pass in the opponent's attack zone 3 is less, and the success rate is lower than that of the strong teams in the world. In addition, it can be seen that the offensive way of the bottom pass is common in the tactical play of the women's football team in the world.

3.3.4. Analysis of passing characteristics based on classification algorithm

In this section, we predict the passing success rate and evaluate pass difficulty for the Chinese women's football team using a Random Forest classification algorithm. We collected passing data from the World Cup, using the coordinates of the pass starting and ending points as inputs. The model outputs a binary result, with 1 indicating a successful pass and 0 indicating a failed attempt. We used 70% of the data as the training set and the remaining 30% as the test set, with resulted ROC curve shown in **Figure 8**.

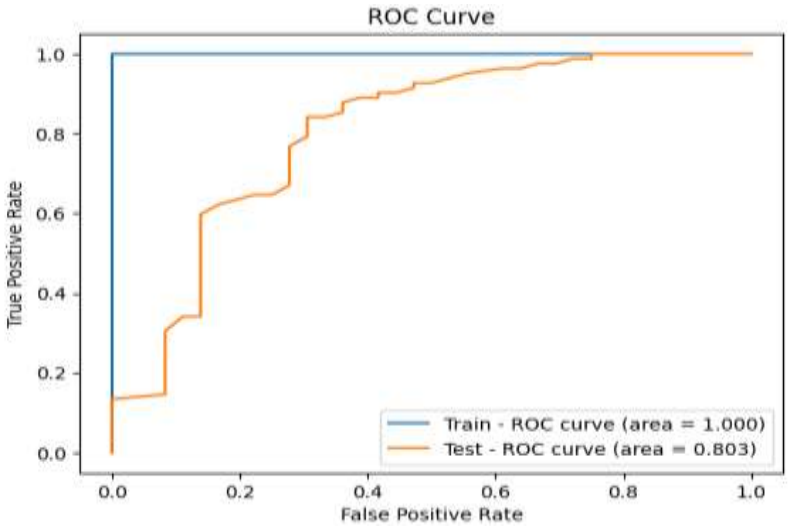


Figure 8. ROC curve.

The ROC curve is a graphical representation that illustrates the performance of a binary classification model at various threshold settings. It plots the True Positive Rate (TPR) against the False Positive Rate (FPR), showing the trade-off between sensitivity and specificity as the decision threshold is varied. The AUC represents the area under the ROC curve and provides a single scalar value that summarizes the model's overall performance. The definition of AUC is shown as:

$$\text{AUC} = \int_0^1 \text{TPR}\left(\text{FPR}^{-1}(x)\right)dx \tag{5}$$

where TPR and FPR denotes True Positive Rate (TPR) and False Positive Rate (FPR). Readers can refer to [21] to get more details with respect to ROC curve and AUC.

The ROC curve indicates that the model does have a certain level of predictive capability, as shown by the testing AUC of 0.803. This AUC value suggests that the model can distinguish between successful and unsuccessful passes with reasonable accuracy on unseen data, making it a useful tool for predicting pass success rates. Although there is a notable difference between the training and testing AUC values, which suggests some overfitting, the testing AUC still reflects a good level of generalization. This indicates that the model can capture meaningful patterns related to passing success, providing valuable insights for tactical analysis and planning. Further refinement of the model could improve its predictive performance, but even in its current state, it demonstrates useful predictive power in analyzing pass outcomes. Additionally, we used the trained model to predict the difficulty of each passing attempt, represented as the probability of success (probability of being 1), and visualized the results in **Figure 9**.

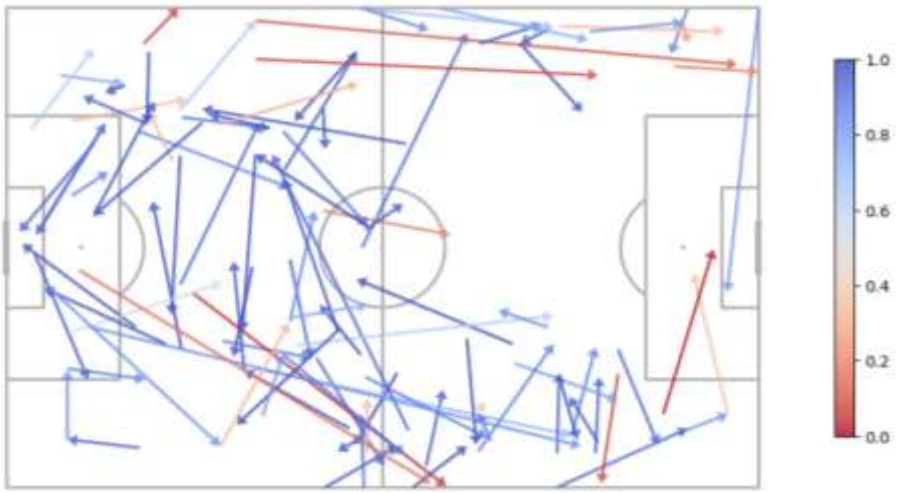


Figure 9. Difficulty predicted by our model.

The visualization illustrates the predicted pass difficulty across different areas of the field, with each pass color-coded according to its success probability. Blue arrows represent passes with a higher likelihood of success (lower difficulty), while red arrows indicate passes with a lower probability of success (higher difficulty). We observe that passes directed toward central areas and shorter distances generally have higher success probabilities, as indicated by darker blue shades. In contrast, longer passes or those directed toward congested areas near the opponent's goal tend to have lower success probabilities, shown in red. This visualization helps coaches and analysts identify high-risk passes, allowing them to adjust strategies to minimize turnovers and optimize passing routes based on difficulty levels, thus enhancing team efficiency and control in key areas of the field.

3.4. Application of kernel density estimation (KDE) in football data analysis

3.4.1. Kernel density estimation (KDE)

Kernel density estimation (KDE) is used to estimate unknown density functions in probability theory, which is one of the nonparametric test methods. It was proposed by Emanuel Parzen [22], also known as Parzen window. Let $(x_1, x_2, \ldots, x_n)$ be an independent sample of the same distribution taken from a univariate distribution. The given point x has an unknown probability density f, and its kernel density estimator can be expressed as follows.

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) \tag{6}$$

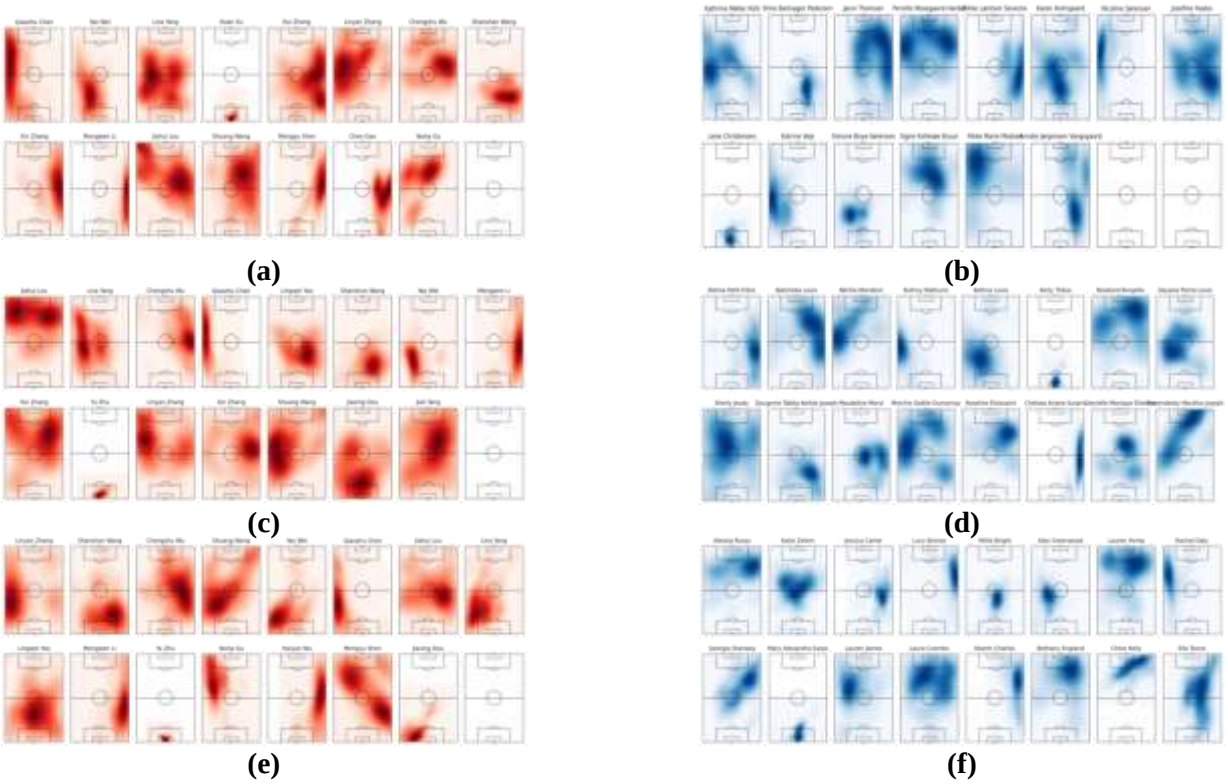
where K is a nonnegative kernel function.

3.4.2. Thermal map analysis of player activity based on kernel density estimation algorithm

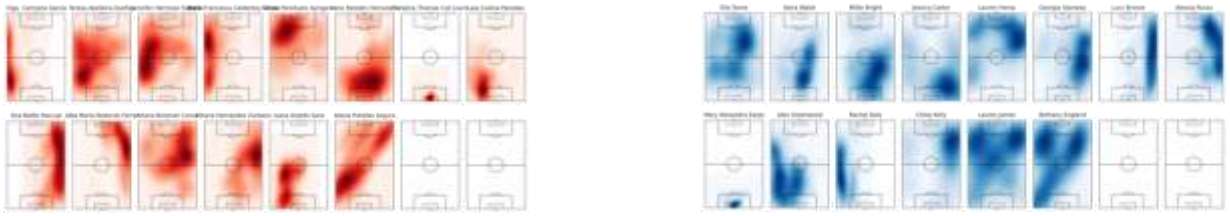
The problem of visualizing the thermal map of player activity can be regarded as a kernel density estimation (KDE) problem. A heat map of a player can be generated by placing a core around each event on the court, such as passing, shooting, and steals.

This heat map can visually show the activity density of players in the match, thus helping coaches to understand the scope of activities and behavior preferences of players. The distribution information of player position data can be obtained through kernel density estimation, and then the thermal map of player activity can be obtained. We utilized Seaborn's kdeplot, a KDE analysis tool implemented in Python, for kernel density estimation, the choice of kernel function and bandwidth significantly impacts the accuracy of results. Seaborn's kdeplot uses the Gaussian kernel by default, which is commonly chosen for its smoothness and effectiveness in continuous data visualization. For the bandwidth parameter, which controls the level of smoothing, we utilized the bw_adjust parameter in kdeplot to fine-tune the balance between bias and variance. By adjusting this parameter, we achieved a KDE that accurately represents the data distribution without excessive smoothing, ensuring reliable visualization results.

Figure 10 visualizes the data of the players' activities of both sides in the Chinese women's football World Cup and the players' activities of the world's top teams. It can be analyzed that the flank attack of the Chinese women's football team mostly depends on two full-backs, and most of the forward players of the Chinese women's football team are interspersed to the flank. The application of kernel density estimation in football data analysis provides an intuitive and effective way to help teams understand the match and the performance of players more comprehensively. This method provides a powerful tool for tactical adjustment, player evaluation and opponent analysis.



(a) **(b)** **(c)** **(d)** **(e)** **(f)**



(g) **(h)**

Figure 10. Heat map of player activity area. **(a)** Denmark vs. China; **(b)** Denmark vs. China (Denmark); **(c)** Haiti vs. China; **(d)** Haiti vs. China (Haiti); **(e)** England vs. China; **(f)** England vs. China (England); **(g)** Spain vs. England; **(h)** Spain vs. England (England).

3.5. Analysis of Football attack treat degree based on Markov chain model

In this section, we proposed a novel approach based on Markov chain model to evaluating threat at different regions in the pitch. Compared to conventional methods which evaluate threat based on goal and assist, our approach separates the calculation of XT into several sub-process, which models the match more effectively. Based on Wyscout's Russia 2018 FIFA World Cup open-source datasets, we will show our approach in detail.

3.5.1. Markov chain model

Markov chain describes the transition of stochastic processes from one state to another in the state space. The process has the property of "no memory": the probability distribution of the next state can only be determined by the current state and has nothing to do with its previous state. The Markov property can be expressed as follows:

$$P\{X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \ldots, X_1 = i_1, X_0 = i_0\} = P\{X_{n+1} = j | X_n = i\} \quad (7)$$

3.5.2. Analysis of football attack treat degree

Events in a football match can be regarded as a random process.

With the assumption of "no memory", the football match can be modeled as a Markov chain, as shown in **Figure 11**.

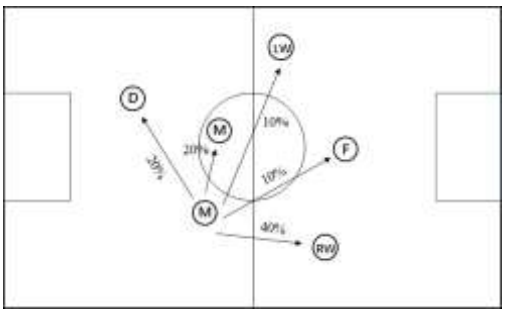


Figure 11. Football event model based on Markov chain.

In order to analyze the threat degree of football attack by using Markov chain model, we divide the football field into 16 × 12 small areas and define the following parameters:

(1) $xT_{x,y}$: threat at (x, y)

(2) $s_{x,y}$: Probability of shooting at (x, y)

(3) $g_{x,y}$: Probability of scoring a goal at (x, y)

(4) $m_{x,y}$: Probability of passing or dribbling at (x, y)

(5) $T_{(x,y)->(z,w)}$: The probability of transferring the ball to (z, w) at (x, y)

Thus, a recursive formula for $xT_{x,y}$ can be obtained:

$$xT_{x,y} = (s_{x,y} \times g_{x,y}) + (m_{x,y} \times \sum_{z=1}^{16}\sum_{w=1}^{12} T_{(x,y)->(z,w)} \times xT_{z,w}) \tag{8}$$

From Equation (8), we notice that the calculation of xT at (x, y) depends on the xT at other regions. Therefore, we solve the problem through an iterative approach as follows:

$$xT_{x,y}{}^{(t+1)} = (s_{x,y} \times g_{x,y}) + (m_{x,y} \times \sum_{z=1}^{16}\sum_{w=1}^{12} T_{(x,y)->(z,w)} \times xT_{z,w}{}^{(t)}) \tag{9}$$

When the iteration above converges to its fixed pointed, Equation (7) holds. Also, we can reformulate Equation (9) in matrix form as follows:

$$\begin{pmatrix} xT_{1,1} \\ xT_{1,2} \\ \vdots \\ xT_{16,12} \end{pmatrix}^{(t+1)} = \begin{pmatrix} s_{(1,1)} \\ s_{(1,2)} \\ \vdots \\ s_{(16,12)} \end{pmatrix} \odot \begin{pmatrix} g_{(1,1)} \\ g_{(1,2)} \\ \vdots \\ g_{(16,12)} \end{pmatrix} + \begin{pmatrix} 0 & T_{(1,1)->(1,2)} & \cdots & T_{(1,1)->(16,12)} \\ T_{(1,2)->(1,1)} & 0 & \cdots & T_{(1,2)->(16,12)} \\ \vdots & \vdots & \ddots & \vdots \\ T_{(16,12)->(1,1)} & T_{(16,12)->(1,2)} & \cdots & 0 \end{pmatrix} \begin{pmatrix} xT_{1,1} \\ xT_{1,2} \\ \vdots \\ xT_{16,12} \end{pmatrix}^{(t)} \tag{10}$$

From Equation (10), we can show that the iterative process is a stationary iterative process. Thus, the fixed point of Equation (10) is asymptotically stable [23].

Based on the Wyscout open-source datasets, iterate according to Equation (7) and calculate the threat degree of each small area $xT_{x,y}$ after the football field is divided in steady state, and visualize the threat degree xT, as shown in the **Figure 12**.

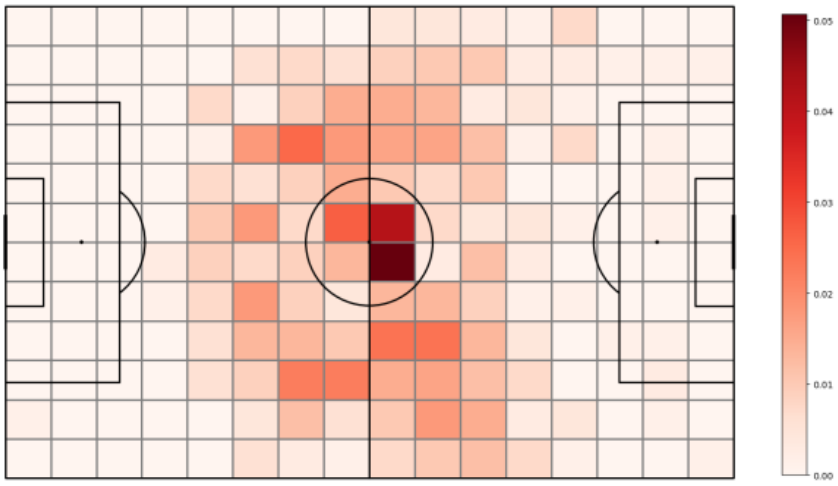


Figure 12. Transition possibility matrix at 90th area.

Figure 13. Visualization of xT matrix.

Based on the threat degree matrix illustrated in the **Figure 13**, we can quantitatively analyze the threat level in each area of the football field under steady-state conditions. This analysis enables a more strategic approach to tactical arrangements by identifying high-threat zones where the opposing team is more likely to create scoring opportunities. Focusing on these high-threat areas, the coaching team can optimize passing strategies to either avoid these zones or exploit them depending on the game plan. For instance, areas with high threat degrees may suggest the need for tighter defensive coverage or more aggressive interception strategies. Offensively, players can be encouraged to maneuver the ball through less risky areas or prepare to capitalize on open spaces near these zones. Furthermore, this analysis supports the design of targeted training sessions for goalkeepers and defenders, particularly in high-threat zones near the goal area. Goalkeepers can undergo specialized drills to improve reflexes, positioning, and decision-making when shots are likely to come from these high-risk regions. Defenders, similarly, can be trained to anticipate passes or positioning that might lead to threats from these specific areas, improving overall defensive coordination.

4. Conclusion

In this paper, machine learning algorithms such as clustering, classification and kernel density estimation are used to analyze the temporal and spatial data such as passing, shooting and position of football matches on the open football data set, and the “killer pass” is taken as a quantitative index. This paper makes a visual analysis of the shooting areas and main passing types of the Chinese women’s football team, and uses the Markov chain model to calculate the offensive threat degree of each area on the field. The method in this paper has more intuitive visual effect and more in-depth data insight. It is of guiding significance for the tactical arrangement and personnel arrangement of the football match.

Author contributions: Conceptualization, PZ and WH; methodology, PZ; software, YZ (Yiqi Zhu); validation, PZ, WH, and WZ; formal analysis, YZ (Yitian Zhang); investigation, WZ; resources, YZ (Yiqi Zhu); data curation, WZ; writing—original draft preparation, PZ; writing—review and editing, PZ and WZ; visualization, WH;

supervision, PZ; project administration, WH; funding acquisition, WZ. All authors have read and agreed to the published version of the manuscript.

Ethical approval: Not applicable.

Conflict of interest: The authors declare no conflict of interest.

References

1. Rico-gonzález M, Pino-ortega J, Méndez A, et al. Machine learning application in soccer: a systematic review [J]. Biology of sport, 2023, 40(1): 249-63.
2. Baboota R, Kaur H. Predictive analysis and modelling football results using machine learning approach for English Premier League [J]. International Journal of Forecasting, 2019, 35(2): 741-55.
3. Liu S. Application status and prospect of big data in China's football field [J]. Contemporary sports technology, 2022, 12(10): 169-72.
4. Pappalardo L, Cintia P, Rossi A, et al. A public data set of spatio-temporal match events in soccer competitions [J]. Scientific data, 2019, 6(1): 236.
5. Tureen T, Olthof S. "Estimated Player Impact"(EPI): Quantifying the effects of individual players on football (soccer) actions using hierarchical statistical models; proceedings of the StatsBomb Conference Proceedings, F, 2022 [C]. StatsBomb.
6. Alpaydin E. Machine learning [M]. MIT press, 2021.
7. Jordan M I, Mitchell T M. Machine learning: Trends, perspectives, and prospects [J]. Science, 2015, 349(6245): 255-60.
8. Yiğit A T, Samak B, Kaya T. Football player value assessment using machine learning techniques; proceedings of the Intelligent and Fuzzy Techniques in Big Data Analytics and Decision Making: Proceedings of the INFUS 2019 Conference, Istanbul, Turkey, July 23-25, 2019, F, 2020 [C]. Springer.
9. Behravan I, Razavi S M. A novel machine learning method for estimating football players' value in the transfer market [J]. Soft Computing, 2021, 25(3): 2499-511.
10. Al-asadi M A, Tasdemır S. Predict the value of football players using FIFA video game data and machine learning techniques [J]. IEEE access, 2022, 10: 22631-45.
11. Wang Z, You S. Analysis method and development trend of football tactics based on location data under the background of big data [J]. journal of shanghai university of sport, 2021, 45(09): 60-9+98.
12. García-aliaga A, Marquina M, Coteron J, et al. In-game behaviour analysis of football players using machine learning techniques based on player statistics [J]. International Journal of Sports Science & Coaching, 2021, 16(1): 148-57.
13. Al-asadi M A M. Decision support system for a football team management by using machine learning techniques [J]. Xinyang Teachers College, 2018, 10(2): 1-15.
14. Rossi A, Pappalardo L, Cintia P, et al. Effective injury forecasting in soccer with GPS training data and machine learning [J]. PloS one, 2018, 13(7): e0201264.
15. Majumdar A, Bakirov R, Hodges D, et al. Machine learning for understanding and predicting injuries in football [J]. Sports Medicine-Open, 2022, 8(1): 73.
16. Oliver J L, Ayala F, Croix M B D S, et al. Using machine learning to improve our understanding of injury risk and prediction in elite male youth football players [J]. Journal of science and medicine in sport, 2020, 23(11): 1044-8.
17. Joseph A, Fenton N E, Neil M. Predicting football results using Bayesian nets and other machine learning techniques [J]. Knowledge-Based Systems, 2006, 19(7): 544-53.
18. Kumar G. Machine Learning for Soccer Analytics [D], 2013.
19. Sun J, Liu J, Zhao L. Research on clustering algorithm [J]. Journal of Software., 2008, 19(1): 48-61.
20. Breiman L. Random Forests [J]. Machine Learning, 2001, 45(1): 5-32.
21. Hoo Z H, Candlish J, Teare D. What is an ROC curve? [Z]. BMJ Publishing Group Ltd and the British Association for Accident …. 2017: 357-9
22. Parzen E. On estimation of a probability density function and mode[J]. The annals of mathematical statistics, 1962, 33(3): 1065-1076.
23. Brin M, Stuck G. Introduction to dynamical systems [M]. Cambridge university press, 2002.