

RESEARCH PAPER

Football Analytics: Understanding and Applying Match Event Data for Performance Analysis

Richard Arambam and Bishwa Ranjan Roy*

Department of Computer Science, Assam University, Silchar, Assam, India

*Corresponding author: bishwa.ranjan.roy@aus.ac.in

Received: 10 Oct., 2025

Revised: 29 Nov., 2025

Accepted: 02 Dec., 2025

ABSTRACT

This paper presents football analytics, focusing on match event data and its use in performance analysis. Match event data consists of precise, time-stamped recordings of on-ball events such as passes, shoots, and tackles, which capture essential aspects of games. The study reviews data sources, analysis methodologies and key performance metrics. Practical applications, including player evaluation and tactical assessment, demonstrate how event data inform strategic decisions. The results show how important it is to include event data in performance analysis. The findings also highlight the problems, limits, and future directions of football analytics.

Keywords: Football analytics, Event data, Machine Learning

Data analytics is the process of collecting, organizing, and analyzing raw data to extract meaningful insights that inform decision-making and improve performance. It identifies patterns, trends, and correlations within data, optimize processes, predict future outcomes and transform data into actionable information that supports strategic and operational decisions^[1].

Sport analytics is the systematic collection of sports data, analyzing and interpreting data to obtain a deeper understanding of the game, improve performance, and inform decision-making in sports^[2]. It is the application of data analysis techniques and statistical methods to understand and improve various aspects of sports, including match outcome, player performance, team strategy, scouting, injury prediction, and fan engagement^{[3]-[6]}.

Moneyball, the story of the Oakland Athletics (A's), a Major League Baseball team is a classic example of sports analytics^[7]. They adopted sabermetrics, a statistical analysis method used in baseball, to assess players according to their slugging percentage (SLG), on-base percentage (OBP), and other performance

How to cite this article: Arambam, R. and Roy, B.R. (2025). Football Analytics: Understanding and Applying Match Event Data for Performance Analysis. *IJISC.*, 12(02): 161-170.

Source of Support: None; **Conflict of Interest:** None



indicators that scouts had not previously taken into account. Despite having a smaller budget than other teams, it emphasizes how sports analytics may be used to obtain a competitive edge. It challenged the traditional scouting methods, which relied heavily on subjective opinions and gut feelings. Instead, they focused on objective statistical analysis to identify undervalued players. Moneyball completely changed how teams in other sports evaluate players and develop their teams. It showed how effective data driven decision making can be and how analytics can help teams with varying resources compete on an even playing field.

Football analytics focuses specifically on the analysis of data within the context of football (soccer). It offers an insight into the use of positional data, exploring how we can be collected, modelled, analyzed and interpreted^[8]. Over the past two decades, the field has evolved from relying on basic match statistics to incorporating advanced metrics and machine learning models that provide deeper insights into player and team behavior. This transformation has been driven by the increasing availability of high-quality data and advancements in data science.

Football analysis has traditionally relied on simple descriptive data like goals scored, assists, possession %, and pass completion rates. While these measures provided an overview of performance, they frequently failed to convey the game's contextual and dynamic nature. With the rise of modern football analytics, more sophisticated measurements like expected goals (xG), expected assists (xA), and expected threat (xT) have emerged to assess the quality of acts rather than just their outcomes. These measurements allow analysts to determine how likely a shot is to result in a goal or how useful a pass is in generating scoring opportunities.

One of the most important applications of football analytics is player performance evaluation. Analysts can analyze a player's contribution in ways that standard measurements cannot. For example, defensive activities such as interceptions, pressing intensity, and off-ball movements can be quantified to provide a more complete evaluation of players. Similarly, analytics is vital in team strategy and tactical analysis. Coaches and analysts use data to discover strengths and weaknesses, optimize formations, and examine opposing strategies.

Despite its advantages, football analytics confronts a number of problems. Modeling is tough due to the game's intricacy, which is defined by constant movement and interactions between players. Furthermore, domain knowledge is required for interpreting analytical data in order to obtain useful insights. Nonetheless, as data collecting technologies and analytical tools progress, football analytics will play an increasingly important part in molding the sport's future.

LITERATURE REVIEW

The most popular sport in the world, football (or soccer) has billions of supporters worldwide. The dynamic nature of the game and the wide range of circumstances that can affect its results make it a great subject for analysis. The expertise of coaches, players, and scouts as well as their subjective evaluations were major components of the old approach to football administration and strategy. On the other hand, the introduction of a data driven paradigm that improves decision-making at all levels of the game with the arrival of football analytics has completely changed the sport.

Early research in football analytics primarily focused on descriptive statistics and basic performance indicators. Traditional metrics such as goals scored, assists, possession percentage, and pass completion

rates were widely used to evaluate team and player performance. These studies provided a foundational understanding of the game but were limited in their ability to capture the complexity and context of football actions.

Football analytics has evolved tremendously in the last two decades. Initially, the focus was on fundamental statistical analysis, including possession percentages, pass completion rates, and shot accuracy. As technology and data collection methods advanced, more complex analysis were possible. Football analytics today uses a variety of approaches, including traditional machine learning and advanced deep learning models. These concepts are applied in various game scenarios, including evaluating player performance, analyzing tactics, preventing injuries, and predicting match outcomes.

The application of machine learning techniques has significantly advanced football analytics. Researchers have employed supervised learning methods such as logistic regression, decision trees, random forests, and support vector machines to predict match outcomes and player performance. For example, classification models have been used to predict whether a team will win, lose, or draw a match based on historical data and match features. Similarly, regression models have been applied to estimate player ratings and expected contributions. Clustering techniques, such as k-means and hierarchical clustering, have been used to group players based on playing style and positional roles. More recently, deep learning approaches, including neural networks and recurrent neural networks (RNNs), have been explored for sequence modeling in football. These models are particularly useful for analyzing event sequences and capturing temporal dependencies in the game. Despite their potential, deep learning models often face challenges related to interpretability and data requirements.

Another important area of research is the development of player performance and rating systems. Traditional rating systems often relied on subjective assessments or limited statistics, which could be biased or incomplete. Modern approaches leverage event and tracking data to create objective and comprehensive player ratings. Researchers have proposed various models that combine multiple performance indicators, such as passing accuracy, defensive actions, and goal contributions, into a single rating score. Some studies have also incorporated contextual factors, such as match difficulty and team strength, to provide a more balanced evaluation. These models are particularly useful for scouting and recruitment, as they allow clubs to compare players across different leagues and competitions. However, the design of rating systems remains challenging due to the need to balance simplicity, interpretability, and accuracy.

Various researchers have made contributions to this field's development. Researchers, stakeholders, and practitioners are always interested in the prediction of football match outcomes based on Machine Learning algorithms. It is thought that Charles Reep was the first notational analyst in Britain. In 1968, he and B. Benjamin produced a statistical analysis of football play patterns based on 578 games played between 1953 and 1967. Lately, there has been a surge in articles and research material in this field. In this section, the work produced by various authors and academics are discuss.

E. Eryarsoy and D. Delen, models such as Naïve Bayes, Decision Trees, Gradient Boosting, and Random Forest were used and applied to a dataset covering a 34-week regular season. The results indicated that ensemble methods significantly outperform individual models; however, predicting draw outcomes remains a major challenge due to league competitiveness and dynamic team behavior^[9].

R. Beal *et al.* used Random Forest on a dataset of 1770 matches across six seasons emphasized the importance of integrating traditional statistical data with contextual and human insights to improve predictive accuracy^[10].

L. Carloni et.al evaluated various machine learning models, including Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Naïve Bayes, Random Forest, and Artificial Neural Networks (ANN). The findings revealed that ANN achieved the highest predictive accuracy, particularly in the over/under betting market^[11].

F. Rodrigues and Pinto utilized various algorithms like Naive Bayes, KNN, Random Forest, Support Vector Machines, Decision Tress, XgBoost, Multinomial Logistic Regression, and Artificial Neural Networks on five seasons of English Premier League data. The study achieved an overall prediction accuracy of 65.26% by Random Forest and demonstrated improved profitability compared to traditional approaches^[12].

Regression-based and ensemble techniques were applied on the Turkish Super League, where feature selection improved prediction accuracy from 75.04% to 85%, with ANN outperforming other models^[13]. Similarly, multiple algorithms, Bayesian Networks, LogitBoost, KNN, Random Forest, and ANN applied on UEFA Champions League data, achieving an accuracy of approximately 65% while highlighting challenges related to data collection and team variability^[14].

Artificial Neural Networks have been widely applied in football analytics. S. M. Arabzad et.al demonstrated the effectiveness of ANN in predicting league outcomes despite the inherent uncertainties of football matches^[15]. Also, hybrid approaches combining Naïve Bayes, LogitBoost, Neural Networks, Random Forest, and Genetic Programming on Dutch Eredivisie data showed improved prediction performance, particularly for draw outcomes^[16]. Using exploratory data analysis and feature engineering, Baboota and Kaur have developed a prediction model for the English Premier League. They have employed machine learning techniques including Gaussian Naïve Bayes, SVM, Random Forest, and Gradient Boosting on English Premier League data found that Gradient Boosting outperformed other models, especially in predicting challenging outcomes such as draws^[17].

Machine learning models have been used in international competitions in addition to league-based research. Random Forest outperformed regression-based models utilizing FIFA World Cup data (2002–2014), and Lasso regression also produced competitive results^[18]. Similarly, neural networks for FIFA World Cup 2018 predictions reported high predictive accuracy, supported by strong ROC performance and low error rates^[19]. Football analytics has also made use of large-scale datasets; over 216,000 matches showed that ANN and SVM models attain competitive accuracy, while fuzzy logic systems can be used successfully in sports betting scenarios^[20].

Recent advances in football analytics have emphasized hybrid and data-driven approaches. Machine learning models, clustering techniques such as K-means and fuzzy C-means, and fan sentiment analysis to improve prediction accuracy and provide novel insights into match outcomes^[21]. Furthermore, Chazan-Pantzalis and Tjortjis used numerous algorithms, including XGBoost, SVM, Random Forest, and regression models on multi-source datasets, successfully forecasted league standings and player performance, emphasizing the necessity of integrating varied data sources^[22].

More recent research has emphasized the growing importance of advanced analytics in football. B *et al.* used Logistic Regression, ensemble voting methods, and Random Forest on data from FIFA and Opta emphasized the need of adding advanced analytics and adjusting to changing player roles and team strategy^[23]. Furthermore, S. Manish *et al.* used regression models, neural networks, and support vector regression for player performance prediction to show that feature selection and dimensionality reduction considerably improved model accuracy^[24].

Overall, the literature shows that ensemble models and neural networks frequently outperform traditional statistical methods in football prediction challenges. Despite the progress made in football analytics, several challenges remain. One of the primary issues is the limited availability of high-quality data, as most detailed datasets are proprietary and expensive. This restricts access for researchers and limits the reproducibility of studies. Another challenge is the lack of standardization in data formats and analytical methods. Different data providers use varying definitions and structures, making it difficult to compare results across studies. Additionally, the complexity of football as a dynamic and low-scoring sport poses challenges for modeling and prediction. Interpretability is also a significant concern, particularly with the use of complex machine learning models. While these models may achieve high predictive accuracy, their outputs are often difficult to interpret, limiting their practical applicability in decision-making.

FOOTBALL EVENT DATA

Football event data refers to detailed match data that captures every action occurring during a game. Table 1 presents common examples of football event data and their corresponding descriptions. Passing events record whether passes are successful or unsuccessful along with attributes such as distance and direction, which help evaluate team build-up play and passing efficiency. Shooting events include attempts on target, off target, and goals, providing insights into attacking effectiveness and scoring opportunities.

Defensive actions such as tackles and interceptions describe attempts to regain possession and assess defensive performance. Fouls capture both committed and received infractions, reflecting player discipline and match intensity. Ball possession events measure the duration and frequency of control, which are commonly used indicators of team dominance.

Goalkeeping performance is represented through save events, while crossing data evaluates wide play strategies based on the number and success rate of crosses. Set piece events, including free kicks and corner kicks, highlight structured attacking opportunities. Player status events document substitutions and disciplinary actions such as yellow and red cards, which significantly influence match outcomes.

Additionally, player movement heatmaps provide spatial visualization of player positioning and movement patterns throughout the match. Collectively, these event types offer comprehensive temporal and contextual information that supports performance analysis, tactical evaluation, and machine learning-based prediction models in football analytics.

Table 1: Examples of Event Data

Event Data Type	Description
Passes	Successful/unsuccessful, distance, direction
Shots	On target, off target, goals
Tackles	Successful or unsuccessful tackles
Fouls	Fouls committed and fouls received
Ball Possessions	Time and frequency of possession
Saves	Saves made by the goalkeeper
Crosses	Number and success rate of crosses
Interceptions	Regaining possession
Set Pieces	Free kicks, corner kicks
Player Status Events	Substitutions, yellow/red cards
Player Movement Heatmaps	Visualizing positions and movement patterns

DATA COLLECTION AND SOURCES

Data is typically collected by tracking providers like StatsBomb, Opta, Wyscout and is used for performance analysis, scouting, and predictive modeling.

The study utilized the Wyscout 2017/2018 dataset, which includes detailed match and player data from five major European football leagues: Serie A, Premier League, La Liga, Ligue 1, and Bundesliga. Sourced from Kaggle^{[25]-[27]}.

CHARACTERISTICS OF FOOTBALL MATCH EVENT DATA

The dataset encompasses 98 clubs, 1,824 matches, and 3,011 players, making it a reliable resource for analyzing player performance across various leagues and positions. The event-level data comprises 17 variables detailing technical and tactical aspects, including match identifiers, event types, spatial coordinates, and temporal attributes. These features facilitate comprehensive analyses of player actions, providing insights into passing sequences and overall match involvement. Table 2 shows the descriptive statistics of the dataset of events and players.

Table 2: Data Descriptive Statistics

Metric	Count
Total unique events	3,071,034
Total unique players	3,011
Total unique matches	1,826
Total unique teams	98

A descriptive study showed that the dataset had more than 3.07 million events, which were captured over 1,826 matches and included 3,011 participants. The positional split showed that midfielders made up the biggest group, with 16,779 player records. Defenders came next with 15,397, forwards with 10,013, and goalkeepers with 3,448. Shown in Fig. 1. There were 98 teams participating, which made sure that all playing roles and league participation were well represented. This variety makes the dataset especially useful for investigations that need a wide range of tactical and positional differences in how players operate.

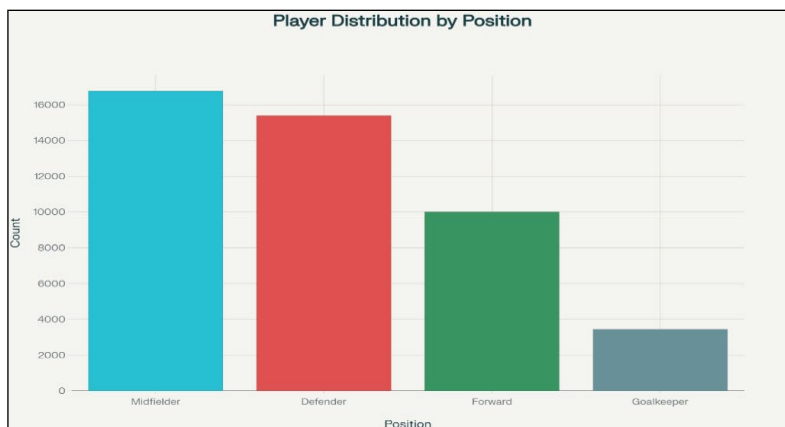


Fig. 1: Player distribution by Position

The Role Cluster contains 26 different types as shown in the graph below, which gives a full picture of how players are spread out across distinct, precise positional groups in the dataset. Fig. 2 shows that most players are in important centre offensive and midfield positions like “central FW,” “central MF,” and “left FW.” It also shows a range of hybrid and specialised positional groupings that include defenders and midfielders. This level of detail shows how complicated player responsibilities are and gives useful information for more detailed position-based analysis. Understanding how important and different these clusters are makes it easier to rate players more accurately in tactical situations.

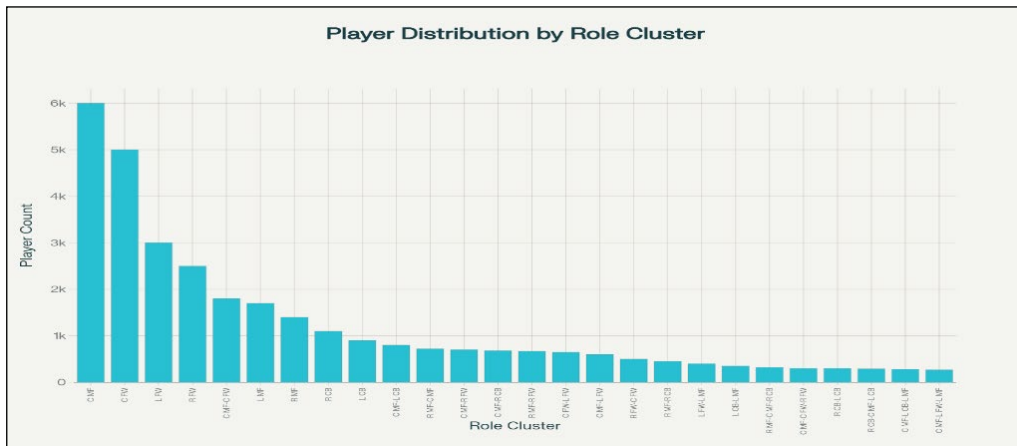


Fig. 2: Player distribution by Role Cluster

6. Challenges and Limitations in Football Analytics

Although sports analytics has completely changed how teams and organizations make choices, there are a number of drawbacks. Accurate data collecting is one of the major challenges. Particularly in intricate settings with plenty of variables, sports data might be inaccurate, inconsistent, or incomplete. For example, it can be expensive or challenging to acquire high quality data needed for player movement tracking, injury analysis, or player effect on the field^[28].

Another difficulty is the complexity of evaluating large datasets. Sports create massive volumes of data, ranging from individual statistics to real-time game results. Obtaining useful insights from this data necessitates complex algorithms and models. To maintain accuracy, these models must be modified on a constant basis, which can take time and resources^[29].

Integrating data from a variety of sources, including wearables, video analysis, and traditional statistics, can be challenging. Each source has its own set of variables and methods for interpretation, making it difficult to develop unified models^[30].

FUTURE DIRECTION IN FOOTBALL ANALYTICS

Recent advancements in machine learning have significantly enhanced football analytics, particularly through the adoption of deep learning and reinforcement learning techniques. These approaches enable the modeling of complex player team interactions and dynamic match conditions, leading to improved predictive performance.

Furthermore, the integration of multi-modal data sources has emerged as an important research direction. Information derived from social media sentiment analysis, video-based match analytics, and betting market trends is increasingly combined with traditional event data to produce more comprehensive and accurate outcome predictions.

In addition, collaborative efforts among data scientists, football analysts, and industry stakeholders have accelerated progress in this domain. Such collaborations emphasize effective feature engineering, model interpretability, and the development of ensemble learning techniques to enhance robustness and practical applicability of predictive models.

CONCLUSION

The Wyscout dataset is good for advanced football analytics and building standardized player-rating systems that take into account the player's position. This study connects theoretical models with real-world data applications, establishing the framework for future research on hybrid, machine-learning-based rating systems. As football analysis keeps evolving, it is important to combine cleaner datasets with position-sensitive algorithms in order to get accurate and useful information.

REFERENCES

1. Tsai, C.-W., Lai, C.-F., Chao, H.-C. and Vasilakos, A.V. 2015. Big data analytics: A survey. *Journal of Big Data*, **2**(1): 21.
2. Bhosale, S., Ray, S., Bhosale, S. and Ray, S. 2023. A review paper on the emerging trends in sports analytics in India. *World Journal of Advanced Research and Reviews*, **19**(2): 461–470.
3. Singh, N. 2020. Sport analytics: A review. *The International Technology Management Review*, **9**(1): 64–69.
4. Van Eetvelde, H., Mendonça, L.D., Ley, C., Seil, R. and Tischer, T. 2021. Machine learning methods in sport injury prediction and prevention: A systematic review. *Journal of Experimental Orthopaedics*, **8**(1): 27.
5. Ghar, S., Patil, S. and Arunachalam, V. 2021. Data-driven football scouting assistance with simulated player performance extrapolation. In *Proceedings of the 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 1160–1167). IEEE.
6. Mahmoud, A.B. 2024. Digital marketing analytics in sports. In *Sports analytics: Data-driven sports and decision intelligence* (pp. 55–74). Springer.
7. Lewis, M. 2004. *Moneyball: The art of winning an unfair game*. W.W. Norton & Company.
8. Memmert, D. and Raabe, D. 2023. *Data analytics in football: Positional data collection, modelling and analysis*. Routledge.
9. Eryarsoy, E. and Delen, D. 2019. Predicting the outcome of a football game: A comparative analysis of single and ensemble analytics methods.
10. Beal, R., Middleton, S.E., Norman, T.J. and Ramchurn, S.D. 2021. Combining machine learning and human experts to predict match outcomes in football: A baseline model. *Proceedings of the AAAI Conference on Artificial Intelligence*, **35**(17): 15447–15451.

11. Carloni, L., De Angelis, A., Sansonetti, G. and Micarelli, A. (2021). A machine learning approach to football match result prediction. In C. Stephanidis, M. Antona and S. Ntoa (Eds.), *HCI International 2021 – Posters* (Communications in Computer and Information Science, Vol. 1420, pp. 473–480). Springer.
12. Rodrigues, F. and Pinto, Â. 2022. Prediction of football match results with machine learning. *Procedia Computer Science*, **204**: 463–470.
13. Tüfekci, P. 2016. Prediction of football match results in Turkish Super League games. In C. Stephanidis (Ed.), *Communications in Computer and Information Science* (Vol. 427, pp. 515–526). Springer.
14. Hucaljuk, J. and Rakipović, A. 2011. Predicting football scores using machine learning techniques. *Proceedings of the 34th International Convention MIPRO*, 1623–1627.
15. Arabzad, S.M., Tayebi Araghi, M.E., Sadi-Nezhad, S. and Ghofrani, N. 2014. Football match results prediction using artificial neural networks: The case of Iran Pro League. *Journal of Applied Research on Industrial Engineering*, **1**(3): 159–179.
16. Tax, N. and Joustra, Y. 2015. Predicting the Dutch football competition using public data: A machine learning approach.
17. Baboota, R. and Kaur, H. 2019. Predictive analysis and modelling football results using machine learning approach for English Premier League. *International Journal of Forecasting*, **35**(2): 741–755.
18. Schauburger, G. and Groll, A. 2018. Predicting matches in international football tournaments with random forests. *Statistical Modelling*, **18**(5–6): 460–482.
19. Hassan, A., Akl, A.-R., Hassan, I. and Sunderland, C. 2020. Predicting wins, losses and attributes' sensitivities in the Soccer World Cup 2018 using neural network analysis. *Sensors*, **20**(11): 3213.
20. Fialho, G., Manhães, A. and Teixeira, J.P. 2019. Predicting sports results with artificial intelligence: A proposal framework for soccer games. *Procedia Computer Science*, **164**: 131–136.
21. Kınalhoğlu, İ. and Kuş, C. 2023. Prediction of football match results by using artificial intelligence-based methods and proposal of hybrid methods. *International Journal of Nonlinear Analysis and Applications*, **14**(1).
22. Pantzalis, V.C. and Tjortjis, C. 2020. Sports analytics for football league table and player performance prediction. In 2020 11th International Conference on Information, Intelligence, Systems and Applications (IISA) (pp. 1–8). IEEE.
23. Chandra, B., Jennet, D. and Keshav, M. 2024. Prediction of football player performance using machine learning algorithm.
24. Manish, S., Bhagat, V. and Pramila, R. 2021. Prediction of football players performance using machine learning and deep learning algorithms. In 2021 2nd International Conference for Emerging Technology (INCET) (pp. 1–5).
25. Pappalardo, L., Cintia, P., Ferragina, P., Massucco, E., Pedreschi, D. and Giannotti, F. 2019. PlayerRank: Data-driven performance evaluation and player ranking in soccer via a machine learning approach. *ACM Transactions on Intelligent Systems and Technology*, **10**(5): 1–27.

26. Decroos, T., Bransen, L., Van Haaren, J. and Davis, J. 2019. Actions speak louder than goals: Valuing player actions in soccer. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 1851–1861). <https://doi.org/10.1145/3292500.3330758>
27. Pulis, M. and Bajada, J. 2022. Reinforcement learning for football player decision making analysis.
28. Davis, J., Bransen, L., Devos, L., Jaspers, A., Meert, W., Robberechts, P., Van Haaren, J. and Van Roy, M. 2024. Methodology and evaluation in sports analytics: Challenges, approaches, and lessons learned. *Machine Learning*, **113**(9): 6977–7010.
29. Yang, L., Amin, O. and Shihada, B. 2024. Intelligent wearable systems: Opportunities and challenges in health and sports. *ACM Computing Surveys*, **56**(7): 1–42.
30. Ju, X., Chen, X., Zhang, Y., Zhou, Q. and Zhao, J. 2024. Tactical optimisation and opponent analysis for football teams with big data mining and machine learning. *International Journal of Medicine & Science of Physical Activity & Sport*, **24**(94).