Dean&Francis
ISSN 2959-6157

The Advancements and Applications of Machine Learning in Predicting Football Scores

Boqiu Zhang

Portland College, Nanjing University of Posts and Telecommunications, Nanjing, China
p23000119@njupt.edu.cn

Abstract:

Football stands as one of the world's most popular sports, where match prediction significantly contributes to its development and enhances its economic and cultural value. This calls for a comprehensive review of the technology used to predict football scores. The study first examines three traditional machine learning methods with simple single-layer architectures. While these methods are straightforward to understand, they suffer from limitations such as low accuracy and slow computational speed, rendering them less advantageous for practical applications. Subsequently, this paper explores four deep learning approaches, whose multi-layered structures outperform traditional methods in model capacity, learning efficiency, and computational power. Consequently, deep learning has become a primary direction for both current and future research in football scoring prediction. However, these approaches face challenges including generalization, accuracy, and explainability. To address these challenges, this paper proposes adversarial domain adaptation techniques to tackle generalization challenges. The introduction of high-dimensional dynamic feature data enhances accuracy. By integrating models, their respective strengths are leveraged to resolve interpretability challenges, while presenting optimistic prospects for deep learning applications in football scoring prediction. This article provides a comprehensive review of machine learning in football score prediction, which is conducive to readers' in-depth and comprehensive understanding of this field, and provides readers with reference for subsequent independent exploration and innovation.

Keywords: Football score prediction; sport; machine learning.

1. Introduction

Nowadays, Football has become one of the most popular sports worldwide. Global football players are estimated to exceed 200 million, with over a billion fans worldwide. The 2022 FIFA World Cup final in Qatar drew viewership exceeding 1.5 billion globally, setting a new record for sports event viewership. In terms of economic scale, the football industry firmly holds the top position in the sports sector. Currently, the global sports betting market continues to expand, with football betting accounting for approximately 70% of this sector's total revenue. The prediction of football scores can stimulate fans' enthusiasm for betting and guide them to make more reasonable bets, thereby promoting the development of the football betting economy and driving the global sports economy development. The prediction of football scores is also beneficial for team management and analysis, helping teams make correct tactical preparations before the game. Therefore, using modern advanced analysis technology to predict football scores not only benefits team and athlete preparations, but also benefits fans' betting and other activities, which should be considered.

In general, the models in most previous studies have ignored some important information when importing data. This has led to significant limitations in their models for predicting matches. For example, there is an article which uses data about a player's attributes and characteristics to predict the score [1], which argues that the score in football is determined by a player's goal, so analyzing and importing all the player data into the model can predict the football score. Such an approach ignores the importance of overall teamwork in football matches. Additionally, some other articles' models have ignored the impact of players' suspensions and injuries, differences in team styles, home or away status, and other factors on the score. Moreover, most models can only be used for predicting football scores before the game. You know, the situation on the field is constantly changing, and the current betting rules allow fans to place bets during the game and teams to make personnel and tactical adjustments during the game. Therefore, real-time score prediction during the game is also very crucial. And the method mentioned in Rodrigo Alves's article has the function of predicting scores in real time [2]. The method based on convolutional neural networks predicts important events based on past events in the game. It can achieve real-time prediction of the score during the game. However, it also has many limitations. For example, it cannot predict events before the game, and it only imported the data set of the top five European leagues for evaluation, making it difficult to predict other types of football matches.

Thanks to the popularity and development of football, football match score prediction can not only help teams to deploy strategies in advance, but also help audiences to bet and other activities. Therefore, it is necessary to make a comprehensive review of football match score prediction technology. The remainder of this paper is structured as follows: Section 2 discusses several traditional machine learning-based approaches and several deep machine learning-based approaches to football match prediction. Section 3 analyzes the future challenges and expectations for these challenges in football match prediction based on the insights gained from the discussion. Finally, Section 4 summarizes the main findings of this paper.

2. Method

2.1 Traditional Machine Learning-based Approaches

2.1.1 KNN

K-Nearest Neighbors (KNN), is a learning model based on factual data. It assumes that similar data points typically have comparable output values in the feature space, allowing predictions about target matches using historical football score data. An article developed its own web crawler to collect and filter 50 key variables, determined the neighbor count (k) through online searches, then calculated sample similarity using Manhattan distance. Finally, it applied inverse distance weighting (1/distance) to amplify the influence of neighboring samples, thereby improving prediction accuracy [3].

2.1.2 SVM

Support Vector Machine (SVM) is a model that is mostly used for classification and regression tasks. The support vector refers to the key sample points in training data that determine classification boundaries, which are located on the margin of hyperplanes. For example, in football score prediction, a surprise match might become a support vector. Using SVM for football match prediction involves calculating the meaning of historical key team data (such as home/away goals per game, shots, corners, etc.) to generate time-series feature data. The Boruta algorithm is then employed to retain significant features, while Z-score normalization is applied to eliminate dimensionality effects [4].

2.1.3 RF

The core of the random forest (RF) method is Booststrap aggregation. It refers to the random extraction of multiple sub-samples (with repartition) from the original data set, each of which is independently trained and can be used

ISSN 2959-6157

for decision making. To mitigate overfitting risks and enhance model generalizability, multiple sampling processes are required. Additionally, backward wrapper can be employed for feature selection. The process begins with basic feature classification using random forest (RF), followed by 10-fold cross-validation to evaluate feature combination performance, ultimately identifying key factors influencing football scores. This integration of backward wrapper with random forest significantly improves prediction accuracy [5].

2.2 Deep Learning- based Approaches

2.2.1 ANN

Artificial Neural Network (ANN) is a computational model inspired by biological neurons. By simulating the neural connections and information processing mechanisms of human brain neurons, it achieves modeling and prediction of complex data. The overall model architecture of ANN has input layer, hidden layer and output layer. The hidden layer adopts Deep Multilayer structure, which has three small layers with 128,64 and 32 neurons respectively. In football score prediction, ANN can receive large feature vectors, use the LeakyReLU activation function to solve the gradient vanishing problem, and output probability distributions through the Softmax function. When using ANN for football score prediction, Elastic Net can be introduced to address overfitting issues in small sample data. The elastic network integrates L1 and L2 coefficients, which are used for feature selection and overfitting suppression respectively [6]. Although ANN's hierarchical information processing capability mimics the human brain and its regularization techniques help resolve high-dimensional complexity in football data—though weaker than RF in interpretability and small sample adaptability—its probabilistic output and nonlinear modeling capabilities make it highly significant and valuable.

2.2.2 CNN

Convolutional Neural Network (CNN) is a deep learning model specifically designed to process grid-structured data (such as images, videos, matrices), extracting spatial hierarchical features by simulating biological visual cortex mechanisms. The main architecture of the model consists of an input layer, a convolutional layer, and an output layer. Pooling layers and normalization layers can be added directly between the convolutional layer and the output layer to enhance performance. In football prediction research, the system first imports an 11×11 player similarity matrix (19-dimensional feature differences between 22 players from home and away teams). It then applies sliding convolution kernels to scan the matrix, calculates local region-weighted sums, reduces dimensions to retain key features, and maps spatial features into win/lose/bet probability using Softmax output [2]. When predicting football scores with CNN, models can be designed with multiple scoring prediction methods: Method 1: Extracting local interaction patterns (e.g., forward-backline interactions) through convolution layers using player features, and outputting three-category probabilities via fully connected layers. Method 2: Integrating player and team statistics (e.g., average goals per game) into the edge of the player matrix to expand input channels, enabling joint learning of micro-player and macro-team features. Method 3: Employing twin networks where dual-branch CNNs process home and away team data separately, quantifying team gaps through feature comparison layers (e.g., cosine similarity) to enhance interpretability [7]. While converting football matches into player relationship graphs and processing them with CNN provides a foundation for subsequent graph neural network (GNN) applications, the accuracy of score predictions remain relatively low.

2.2.3 RNN

Recurrent Neural Networks (RNN) is a neural architecture specifically designed for processing sequential data. Its core design philosophy involves introducing a "memory" mechanism that enables the network to capture temporal dependencies or sequence patterns in data. Unlike traditional neural networks (such as fully connected networks), RNN neurons not only receive current input but also incorporate the previous moment's output as an additional input. When predicting football match scores, this feature enhances the input of parameters influencing the game, thereby improving accuracy. However, RNN has limitations. For long sequences, the gradient of backpropagation decays exponentially, making it difficult for the network to learn how earlier steps affect final outcomes (for example, the significant contribution of the first match to predictions of the 10th match is largely forgotten). Additionally, gradients may grow exponentially, leading to unstable training. Therefore, more complex RNN architectures like Long Short-Term Memory Network (LSTM) can be employed for improvement. Such a model requires an input layer, an LSTM layer, a fully connected layer, and an output layer. The original data format (sample count, feature count) needs to be converted into 3D sequence format suitable for LSTM. Each LSTM unit receives the current time slice's feature vector and the previous time step's hidden state/memory cell state at each time step, then outputs the final hidden state vector. This vector theoretically encodes information from the required sequence of n previous matches. The final output vector from the LSTM layer is fed through a fully connected layer into a Dense

Layer, where the Softmax activation function is applied to generate match predictions [8].

By incorporating temporal modeling concepts into football prediction, the RNN model constructs a feature sequence reflecting teams' recent performance. This approach demonstrates notably superior predictive capabilities compared to other existing models.

2.2.4 GNN

Graph Neural Network (GNN) is a deep learning model specifically designed to process graph-structured data. Its core concept involves using neural networks to directly aggregate information and perform representation learning on nodes, edges, and overall structures within graphs. In football scenarios, GNNs enable modeling of competitive networks between teams or the impact of player mobility. Its model consists of three modules, Player Interaction Network, Team Interaction Network and Match Comparison Transformer. Each module is set up in three layers, which not only captures enough information, but also compresses the output dimension and reduces the amount of computation. However, traditional neural networks like CNN and RNN struggle with complex node relationships, making them unsuitable for graph data processing. To apply GNN in football prediction, the key lies in modeling multi-level interactions between players and teams. This requires establishing a player interaction network (with players as nodes and player interactions as edges) and a team interaction network (with all teams as nodes and historical win-loss ratios as edges). By combining player embeddings with corresponding team embeddings, GNN enables joint prediction of match outcomes [9]. In football prediction, GNN overcomes the temporal limitations of RNN by directly modeling complex team competition networks, providing deeper relational insights for result prediction.

3. Discussion

3.1 Challenges

In the context of rapid technological advancement, traditional machine learning approaches exhibit multiple performance limitations. Conventional models often struggle to process sufficient data inputs, resulting in compromised accuracy. Moreover, their infrequent use of hierarchical architecture significantly slows computational efficiency, compromising real-time responsiveness. This makes deep learning-based methodologies the prevailing trend in football match prediction systems.

Deep learning-based approaches still face multiple challenges in football prediction. 1. Generality challenges: Football match data exhibits significant contextual dependency. Factors like match type, team style, weather conditions, and refereeing standards can cause data distribution drift, making models unable to adapt to all scenarios. Additionally, dynamic disturbances such as player condition fluctuations, sudden injuries, and transfer events significantly impact final outcomes. 2. Accuracy challenges: While deep learning methods have improved prediction accuracy compared to traditional approaches, severe scoring prediction errors frequently occur. Rare occurrences of high-scoring matches leave insufficient reference samples for model learning, hindering pattern recognition. High-probability random events—such as goalkeeper saves, shots hitting the post, or critical refereeing errors—can dramatically alter match results. 3. Explainability challenges: Deep learning models are often called "black boxes" due to their opaque decision-making logic. This limitation is particularly evident in football prediction. Some models may incorporate irrelevant variables (e.g., jersey colors, player names) that affect predictions, reducing trust in outcomes among teams and spectators.

3.2 Future Prospects

Regarding the challenges mentioned, deep learning-based approaches also hold positive expectations. Cultivating domain-adaptive transfer learning capabilities in model training can effectively address generalization challenges. 1. By leveraging source domains (training base models and fine-tuning or feature alignment to adapt models to target domains), researchers have proposed adversarial domain adaptation for match prediction. Through adversarial training of domain classifiers, this method has significantly reduced distribution differences between the Premier League and La Liga, enhancing cross-league generalization capabilities [10]. 2. Integrating high-dimensional dynamic features serves as an effective approach to combat accuracy challenges. The core concept involves abandoning traditional static data and introducing real-time, granular features: player movement heatmaps, passing networks, pressing intensity, shooting positions (expected goals), key passes, foul locations, real-time scores, remaining time, red-yellow cards, etc. The way of combining the xG sequence with the LSTM model is provided, which can simulate the game process and goal distribution more accurately [11]. 3. Combining deep learning with interpretable mechanism models offers a good solution for explainability. Using deep learning to process complex features (such as video and spatiotemporal data) generates intermediate results (e.g., expected goals xG, probability of key events). These intermediate results are then fed into transparent football knowledge-based models (such

ISSN 2959-6157

as improved Poisson regression and rule-based reasoning systems) to generate final scores and explanations. In this study, the player position features extracted by CNN are input into Bayesian network to generate score prediction [12].

4. Conclusion

This review explored the work of three traditional machine learning approaches (KNN, SVM, RF) and four deep learning approaches (ANN, CNN, RNN, GNN) when applied to football score prediction. This paper also analyzed current challenges in universality, accuracy and interpretability, and in view of these challenges, this paper proposes three expectations: domain adaptive transfer learning, fusion of high-dimensional dynamic features, and combination of deep learning with interpretable mechanism models. In the future, deeper research on various learning-based approaches will be conducted.

References

[1] Danisik N, Lacko P, Farkas M. Football match prediction using players attributes. 2018 World Symposium on Digital Intelligence for Systems and Machines (DISA), IEEE, 2018.

[2] Alves R. SCORE: A convolutional approach for football event forecasting. International Journal of Forecasting, 2025.

[3] Kinalioğlu İH, Kuş C. Prediction of football match results by using artificial intelligence-based methods and proposal of hybrid methods. International Journal of Nonlinear Analysis and Applications, 2023, 14(1): 2939-2969.

[4] Rodrigues F, Pinto Â. Prediction of football match results with machine learning. Procedia Computer Science, 2022, 204: 463-470.

[5] Alfredo YF, Isa SM. Football match prediction with tree based model classification. International Journal of Intelligent Systems and Applications, 2019, 11(7): 20-28.

[6] TeliShinde P, et al. Prediction of football match score and decision making process.

[7] Herath SD, Manivannan S. Predicting the outcome of football matches using convolutional neural network.

[8] Awadallah AA, Khandelwal R. Football match prediction using deep learning (recurrent neural network), 2020: 3-4.

[9] Wang L, et al. Player-team heterogeneous interaction graph transformer for soccer outcome prediction. Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, V. 2, 2025.

[10] Vaswani A, et al. An autoencoder based approach to simulate sports games. International Workshop on Machine Learning and Data Mining for Sports Analytics, Cham: Springer International Publishing, 2020.

[11] Bandara I, et al. Predicting goal probabilities with improved xG models using event sequences in association football. PLoS One, 2024, 19(10): e0312278.

[12] Meng F, Gong X, Zhang Y. RHL-track: Visual object tracking based on recurrent historical localization. Neural Computing and Applications, 2023, 35(17): 12611-12625.