Machine Learning
https://doi.org/10.1007/s10994-024-06608-w

Evaluating soccer match prediction models: a deep learning approach and feature optimization for gradient-boosted trees

Calvin Yeung[1] · Rory Bunker[1] · Rikuhei Umemoto[1] · Keisuke Fujii[1,2,3]

Received: 27 September 2023 / Revised: 3 April 2024 / Accepted: 4 August 2024

Abstract

Machine learning models have become increasingly popular for predicting the results of soccer matches, however, the lack of publicly-available benchmark datasets has made model evaluation challenging. The 2023 Soccer Prediction Challenge required the prediction of match results first in terms of the exact goals scored by each team, and second, in terms of the probabilities for a win, draw, and loss. The original training set of matches and features, which was provided for the competition, was augmented with additional matches that were played between 4 April and 13 April 2023, representing the period after which the training set ended, but prior to the first matches that were to be predicted (upon which the performance was evaluated). A CatBoost model was employed using pi-ratings as the features, which were initially identified as the optimal choice for calculating the win/draw/loss probabilities. Notably, deep learning models have frequently been disregarded in this particular task. Therefore, in this study, we aimed to assess the performance of a deep learning model and determine the optimal feature set for a gradient-boosted tree model. The model was trained using the most recent 5 years of data, and three training and validation sets were used in a hyperparameter grid search. The results from the validation sets show that our model had strong performance and stability compared to previously published models from the 2017 Soccer Prediction Challenge for win/draw/loss prediction. Our model ranked 16th in the 2023 Soccer Prediction Challenge with RPS 0.2195.

Editors: Philippe Lopes, Werner Dubitzky, Daniel Berrar, and Jesse Davis.

✉ Keisuke Fujii
fujii@i.nagoya-u.ac.jp

Calvin Yeung
yeung.chikwong@g.sp.m.is.nagoya-u.ac.jp

Rory Bunker
rory.bunker@g.sp.m.is.nagoya-u.ac.jp

Rikuhei Umemoto
umemoto.rikuhei@g.sp.m.is.nagoya-u.ac.jp

[1] Graduate School of Informatics, Nagoya University, Nagoya, Japan

[2] Center for Advanced Intelligence Project, RIKEN, Osaka, Japan

[3] PRESTO, Japan Science and Technology Agency, Saitama, Japan

Published online: 23 August 2024

Keywords Sports · Match result prediction · Match outcome forecasting · Football

1 Introduction

Soccer, also known as “Association Football” or ”Football”, is widely recognized as the most popular sport worldwide in terms of both spectatorship and player numbers. As a generally low-scoring sport, especially at the professional level, small goal margins, competitive leagues, and draws being a common outcome make predicting soccer match results a challenging task (Bunker & Susnjak, 2022; Yeung et al., 2023), especially when only goals are available in the dataset (Berraret al., 2019). The inherent unpredictability of outcomes is, of course, one of the primary reasons soccer attracts such a large number of fans. Despite its challenging nature, given the popularity of the sport, there is a wide range of stakeholders who are interested in the prediction of soccer match results including fans, bookmaking companies, bettors, media, as well as coaches and performance analysts.

Models for predicting match results have traditionally emerged from mathematical sub-disciplines, including statistics (Maher, 1982; Dixon & Coles, 1997) and operations research (Hvattum & Arntzen, 2010; Wright, 2009). With the rise of machine learning (ML) in the last two decades, ML models have gained traction as a prevalent method for predicting soccer match outcomes. The lack of publicly-available benchmark datasets has, however, meant that it has been challenging for researchers to evaluate their results against other studies. Match features used in models, which are derived from events that occur within matches, are often contained in vendor-specific event data streams that are generally only available to professional teams (Decroos et al., 2019). The Open International Soccer Database (Dubitzky et al., 2019), despite not containing such match features, has enabled researchers to compare their models in a like-for-like manner on a large number of soccer matches (over 216,000 matches across 52 leagues). The 2017 Soccer Prediction Challenge (Berrar et al., 2019) was held, with participants using the Open International Soccer database to predict 206 unplayed matches. Some of the top-ranked participants in the 2017 Soccer Prediction Challenge used gradient-boosted tree models and/or rating features (Berraret al., 2019; Constantinou, 2019; Hubáček et al., 2019), which suggested that condensing a wide range of historical match information into ratings was of benefit, as was using the accuracy-enhancing benefits of boosting. Subsequently, other studies (Razali et al., 2022, 2022; Robberechts & Davis, 2019) have used the Open International Soccer Database, in some cases improving upon the 2017 challenge results (Razali et al., 2022).

Deep learning has, over the past few years, gained in popularity for the prediction of match results in soccer, given its success in many domains including computer vision, trajectory analysis, and natural language processing. In soccer, deep learning has been helpful in predicting the locations and types of subsequent events (Simpson et al., 2022; Yeung et al., 2023), the outcomes of shots (Yeung & Fujii, 2023), and in using match video to detect and track players and/or the ball, to detect events, and to analyze matches Akan & Varlı (2023). Two types of models have thus emerged as potential state-of-the-art models for predicting sports match results. This study investigates both of the two approaches: deep learning, discussed in Sect. 3.1, and boosted decision tree models, discussed in Sect. 3.2.

A subsequent soccer prediction challenge competition was held in 2023,[1] using a similar dataset. However, unlike the 2017 Soccer Prediction Challenge, the 2023 competition required two tasks, the first of which, “exact score prediction,” involved predicting match

[1] https://sites.google.com/view/2023soccerpredictionchallenge.

results in terms of exact goals scored (for each team), and the second, "probability prediction," involved prediction in terms of the probabilities for a win, draw, and loss.

In the current study, initially, consistent with Razali et al. (2022), a CatBoost model (Prokhorenkova et al., 2018) with pi-ratings (Constantinou & Fenton, 2013) as the model features were found to be the best-performing model for win/draw/loss probability prediction. The Pi-Ratings is a rating system dedicated to dynamically evaluating soccer team performance. It factors in recent matches, home advantage, and the significance of winning over score margins. However, to explore the potential of deep learning models for match result prediction in soccer, we then developed a deep learning-based model for win/draw/loss probability prediction that utilizes a combination of cutting-edge techniques. Specifically, the proposed method incorporates modules from the TimesNet time series model (Wu et al., 2022), Transformer, a neural language processing model (Vaswani et al., 2017), and a neural network. The specifics regarding the dataset utilized and model training for the 2023 SoccerNet Prediction Challenge were outlined in Sects. 4.1 and 3, respectively. The results from the validation sets show that our model outperformed all previously published models from the 2017 Soccer Prediction Challenge for win/draw/loss probability prediction. All methods and experiments were conducted prior to the submission deadline of the 2023 Soccer Prediction Challenge, with the exception of Sect. 4.4. The ranking and performance metrics for the prediction challenge were provided by the organizers subsequent to the submission deadline.

The main contributions of this study are as follows. First, we reviewed features that had been utilized in previous studies and proposed a method to select the feature set with the highest information gain and lowest inter-correlation. Second, we proposed the Inception+TE+MLP model, a deep-learning model that was preferable in soccer match result prediction to the existing models. Third, the proposed deep learning model Inception+TE+MLP was compared with existing models, and real-world data were used to evaluate the approach's effectiveness. Finally, this paper served as the technical report of the 2023 Soccer Prediction Challenge.

The remainder of this paper is organized as follows. In Sect. 1.1, we describe the two tasks of the 2023 Soccer Prediction Challenge. Then, in Sect. 2, we discuss research related to the current study. Including the existing literature related to the 2017 Soccer Prediction Challenge, as well as studies subsequent to the competition that also used the Open International Soccer Database. Afterward, we detail the two approaches used in this study for soccer match result prediction in Sect. 3. Following this, the experimental results are presented and discussed in Sect. 4. Finally, the paper is concluded in Sect. 5.

1.1 2023 Soccer prediction challenge

As mentioned in the introduction, the 2023 Soccer Prediction Challenge made available to participants a dataset similar to that of the 2017 competition to train their models. This training dataset comprised match results from 51 soccer leagues from 2001 to April 4, 2023, encompassing a total of over 300,000 matches. The dataset of matches to be predicted included 736 matches from 44 leagues, spanning the period from April 14 to April 26, 2023. Within these datasets, nine distinct features were provided: the season, league, date of the match, names of the home and away teams, the goals scored by the home and away teams, the difference in goals scored between the home and away teams, and the outcome of the match(win/draw/loss). Participating contestants were granted the option to leverage supplementary publicly available data. As mentioned in the introduction, the 2023 Soccer Prediction Challenge necessitated the completion of two distinct tasks: exact score prediction and probability prediction.

Task 1: Exact scores prediction Predicting match results based on the exact goals scored by the home and away teams was a task introduced in the 2023 competition, which was not required in the 2017 competition. The evaluation metric used for Task 1 was the Root Mean Squared Error (RMSE), which is defined as follows:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{1}$$

where N denotes the total number of data points, y_i denotes the actual observed values, and $\hat{y}_i$ represents the values predicted by the model. Models with a lower RMSE are preferred.

Task 2: Probabilities prediction The prediction of match results based on win, draw, and loss probabilities, which was the sole task in the 2017 Soccer Prediction Challenge, was the second task in the 2023 competition. The evaluation metric used for Task 2 models was the Ranked Probability Score (RPS) (Epstein, 1969; Constantinou & Fenton, 2013), which is given by:

$$\text{RPS} = \frac{1}{r-1}\sum_{i=1}^{r-1}\left(\sum_{j=1}^{i}(p_j - a_j)\right)^2, \tag{2}$$

where r denotes the number of potential match outcomes (e.g., $r = 3$ if there are three possible outcomes: home win, draw, and away win). The RPS values always lie within the interval [0, 1], with a lower RPS indicating a better prediction. In particular, an RPS value of 0 indicates a perfect prediction by a model, whereas a value of 1 represents a prediction that was completely incorrect.

2 Related work

In this section, we review related research on deep learning for match results prediction in soccer and also studies that have used the Open International Soccer Database (Dubitzky et al., 2019) to build their models.

Despite its successful application in a number of application domains, there are still relatively few studies that have applied deep learning models for soccer match result prediction. Danisik et al. (2018) used an LSTM model to predict soccer match outcomes, comparing classification, numeric prediction, and dense approaches, and also with baselines based on the average random guess, bookmaker odds-derived predictions, and home win (the majority class). Data from the English Premier League was used, and player-level data was obtained from the FIFA video game for the classification and numeric prediction approaches. The average accuracy obtained with the LSTM regression model was 52.5%. Jain et al. (2021) used Recurrent Neural Networks and LSTM networks for soccer match result prediction. Using English Premier League data, the authors manually engineered several relevant features, e.g., winning and losing streaks, points, and goal differences. Their reported accuracy was 80.75%, however, it should be noted that this was for 2-class, not a 3-class prediction. Rahman (2020) used deep neural networks and artificial neural networks for soccer match result prediction using primarily rankings and results, achieving 63.3% accuracy when predicting 2018 FIFA World Cup matches. Recently, Malamatinos et al. (2022) used k-Nearest-Neighbors, LogitBoost, Support Vector Machine, Random Forest, and CatBoost, along with Convolutional Neural Networks and Transfer Learning with tabular data that was encoded and converted to image models, to predict the results of Greek Super League matches. The best-performing model, CatBoost, which was found to outperform the Convolutional Neural Network, was also applied to predict English Premier and Dutch Eredivisie league matches. Also recently, Joseph (2022) used statistical time series approaches to predict English Premier League outcomes and compared their performance with LSTM and Bayesian methods.

In more recent times, driven by the popularity of the Transformer model (Vaswani et al., 2017), attention mechanisms have gained prominence in modeling soccer-related data. Notably, attention mechanisms have been harnessed in various ways for soccer data analysis. For instance, Zhang et al. (2022) introduced an attention-based LSTM network to predict soccer match result probabilities. Another notable contribution comes from Simpson et al. (2022), who proposed the Seq2event model, based on the transformer architecture, for modeling sequential soccer data. This work was subsequently extended and refined by Yeung et al. (2023). Given this backdrop, this study aimed to investigate the effectiveness of attention mechanisms in the context of soccer match results prediction. Furthermore, we intended to delve into investigating the potential of state-of-the-art time series models, leveraging the inherent sequential nature of soccer match results data.

The standard evaluation metric used in the 2017 Soccer Prediction Challenge, and subsequent studies that have used the Open International Soccer Database, has been the Ranked Probability Score (RPS) (Epstein, 1969; Constantinou & Fenton, 2012). However, other evaluation metrics such as cross-entropy (Hubáček et al., 2022) and accuracy have also been used. Since the RPS is sensitive to distance Constantinou and Fenton (2012), where the ordered nature of win/draw/loss has been considered, the RPS is a preferable metric in evaluating match outcome prediction model performance. However, subsequently, Wheatcroft (2021) suggested that the ignorance score may be a more appropriate metric for match outcome prediction model evaluation.

Among the studies from 2017 Soccer Prediction Challenge participants, Tsokos et al. (2019) used a Bradley-Terry model, Poisson log-linear hierarchical model, and an integrated nested Laplace approximation, achieving an RPS of 0.2087 and accuracy of 0.5388. Hubáček et al. (2019) used relational- and feature-based methods, with pi-ratings (Constantinou & Fenton, 2013) and PageRank ratings (Page et al., 1998) computed for each of the teams in each match. XGBoost (Chen & Guestrin, 2016) was employed as the feature-based method and was used for both classification and regression, and boosted relational dependency networks (RDN-Boost) (Natarajan et al., 2012) were used as the relational method. Classification with XGBoost, achieving RPS and accuracy of 0.2063 and 0.5243, respectively, performed best on the validation set and the challenge test set. Constantinou (2019) proposed a Hybrid Bayesian Network, using dynamic ratings based on the pi-rating system developed in previous work (Constantinou & Fenton, 2013) but that also incorporated a team form factor to identify continued over- or under-performance. In the modified pi-rating calculation, the (win, draw, loss) match outcome was emphasized to a greater extent than the goal margin in order to dampen the effect of large goal margins. The Hybrid Bayesian Network was applied to four rating features—two each for the home and away teams. The model was able to make accurate predictions for a match between two teams even when the prediction was based on historical match data that involved neither of the two teams, with the model achieving accuracy of 0.5146 and an RPS of 0.2083 on the challenge test data set. Berraret al. (2019) created two types of feature sets: recency and rating features. Recency features consisted of the averages of features over the previous nine matches, based on four feature groups: attacking strength, defensive strength, home advantage, and opposition strength. XGBoost and k-Nearest-Neighbors were applied to each of the two feature sets, with both models performing better on the rating features than the recency features. XGBoost, when applied to the rating features, provided the best performance, although this result of 0.5194 accuracy and 0.2054 RPS was obtained after the competition had concluded. The k-Nearest-Neighbors model applied to the rating features, which achieved accuracy of 0.5049 and an RPS of 0.2149 on the competition test set, was the best result achieved during the competition.

The top-performing participants in the 2017 Soccer Prediction Challenge commonly applied machine learning to rating features. At least in the absence of match-related features on in-game events, ratings, therefore, seem to be an effective means of condensing a large amount of historical match information into a concise set of model features. It was also evident that gradient-boosted tree models such as XGBoost exhibited strong performance in the competition.

Subsequent to the 2017 Soccer Prediction Challenge, other researchers have made use of the Open International Soccer Database. Robberechts and Davis (2019) compared the performance of result-based Elo ratings and goal-based offensive-defensive models in predicting match results in FIFA World Cup and Open International Soccer database matches. The ELO ordered logit achieved an RPS of 0.2035 and an accuracy of 0.5146, while the ELO plus offensive-defensive model ordered logit obtained RPS and accuracy of 0.2045, and 0.5146, respectively. However, bookmaker odds-obtained predictions achieved slightly better performance, with a lower RPS of 0.2020 and slightly higher accuracy of 0.5194. Razali et al. (2022) also used the Open International Soccer Database, comparing the performance of gradient-boosted tree models such as XGBoost, LightGBM, and CatBoost on goal- and result-based Elo ratings and pi-ratings. The authors found that CatBoost applied to the pi-ratings features yielded the best performance (RPS = 0.1925), which was better than the results achieved by the 2017 Soccer Prediction Challenge participants. Razali et al. (2022) used a deep learning-based approach by applying TabNet, an interpretable canonical deep tabular data learning architecture, to pi-ratings and achieved a slightly higher RPS of 0.1956, which was still better than the 2017 Soccer Prediction Challenge participants. Hubáček et al. (2022) compared several statistical Bivariate and Double Poisson and Weibull distributions-based statistical models and ranking systems, specifically, Elo ratings, Steph ratings, Gaussian-OD ratings, as well as the soccer-specific rating systems of Berrar ratings (rating feature learning in Berrar et al. (2019)) and pi-ratings. Through their experiments using the Open International Soccer database—with matches before July 2010 forming a validation set that was used for hyperparameter tuning, and matches after this date forming a test set of 91,155 matches—the authors found that Berrar ratings provided the lowest RPS (0.2101) among the different statistical models and rating systems. However, the other models also provided performance very close to that of Berrar ratings, suggesting the existence of some limits to match prediction performance. Since Hubáček et al. (2022) did not use any boosting or deep learning methods, we can tentatively conclude that—at least on datasets such as the Open International Soccer Dataset that do not contain features derived from match events other than the goals scored—gradient-boosted tree models and/or deep learning models can outperform the rating systems themselves and statistical model.

In this study, we devised two distinct approaches. Firstly, in the context of the gradient-boosted tree model, it is noteworthy that several methods have demonstrated superior performance by leveraging unique features. Consequently, our primary emphasis within this approach lies in pinpointing the optimal feature set through the utilization of feature selection algorithms. Secondly, taking into account the remarkable achievements of deep learning models in recent years, our objective was to probe the potential of these models in the realm of predicting soccer match outcomes. This endeavor involved an exploration of the predictive prowess of deep learning methodologies.

3 Methods

In this section, we first introduce the deep learning and boosted decision tree methods in Sects. 3.1 and 3.2, respectively.

3.1 Approach 1: Deep learning (DL)

The deep learning approach leverages time series-based features elaborated in Sect. 3.1.1 and employs a transformer-based model to predict probabilities as described in Sect. 3.1.2.

3.1.1 Features engineering

In this section, we elucidate the process of engineering features for each match. The recency features extraction method proposed by Berraret al. (2019) was utilized. In which, given a particular match of interest at time t, the n previous matches at time $t - i$ of both the home and away teams are considered, where $i \in \{1, 2, ..., n\}$ and $n = 5$. For each of the $t - i$ matches, the following features were derived:

- Team ID: a randomly assigned ID for the team
- Attacking strength: goals scored in match $t - i$.
- Defensive strength: goals conceded in match $t - i$.
- Strength of opposition: average goal difference of the opponent as of match $t - i$, calculated across its prior n matches.
- Home advantage: a binary variable that takes a value of 1 if match $t - i$ was a home game and -1 if it was an away game.

The derived features can be represented as a matrix (Table 1 shows the transpose of this matrix).

3.1.2 Inception+TE+MLP model

In this section, we elucidate the components of the Inception+TE+MLP model and their respective objectives. The Inception+TE+MLP model was designed to proficiently extract pertinent information from input features, encode them into a vector representation, and subsequently decode this vector to produce the desired output. The specifics of each component are elaborated upon below and Fig. 1 illustrates the refined concept of the deep learning approach.

Table 1 Example of results from the recency feature extraction method

		Recency				
	Feature group	t − 1	t − 2	t − 3	t − 4	t − n
Home team	Attacking strength	5	2	4	1	3
	Defensive strength	0	0	0	0	1
	Strength of opposition	−1.8	1	−1.6	−0.2	−0.2
	Home advantage	−1	1	1	1	−1
	Team ID	436	436	436	436	436
Away team	Attacking strength	2	1	1	0	0
	Defensive strength	1	3	0	2	1
	Strength of opposition	0.8	0	−1.2	−2	0.2
	Home advantage	−1	1	1	−1	−1
	Team ID	609	609	609	609	609

The feature values for the 2020–2021 English Premier League match between Manchester City (1st) and Sheffield United (20th) are presented herein

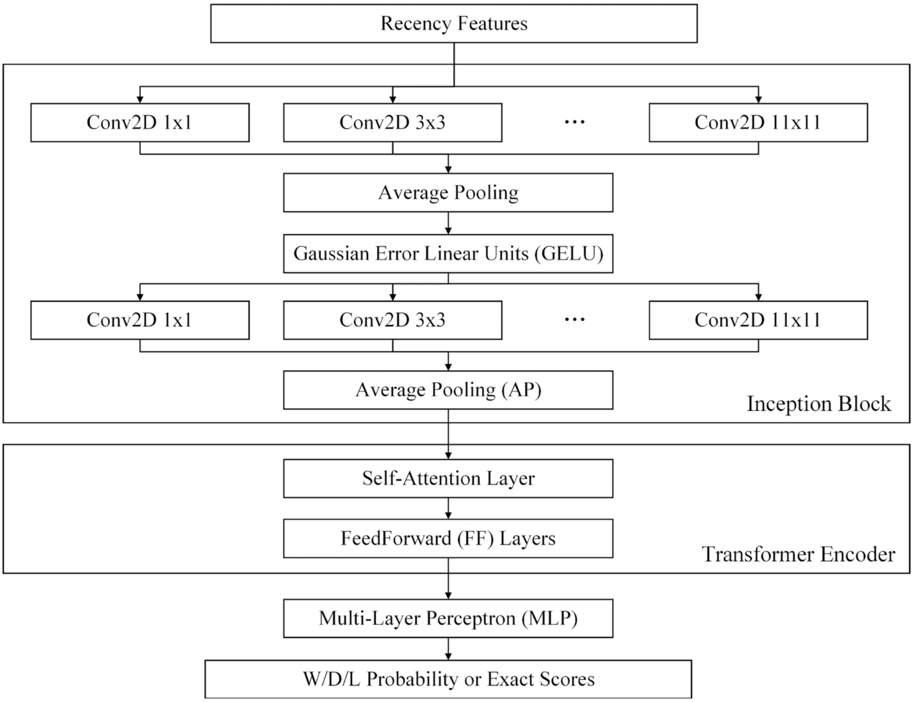

Fig. 1 Overview of the Inception+TE+MLP model. The Conv2D represents 2D convolutional layers with a kernel size of n×n, while GELU serves as an activation function, and FF layers denote multiple linear layers

Information extraction within the model Inspired by Timesblock in the TimesNet model (Wu et al., 2022), we employ the inception block (Szegedy et al., 2015) to extract additional features from the recency features matrix from Sect. 3.1.1, while retaining its shape. The inception block has traditionally been employed to allow subsequent layers to better capture information, and TimesNet has shown the effectiveness of applying the inception block to multi-dimensional time series. The recency features matrix from Sect. 3.1.1 can be viewed as an 8-dimensional time series.

Encoding and decoding In recent times, the Transformer Encoder (TE) (Vaswani et al., 2017) has become a popular method to embed soccer time series and sequential data into an informative vector (Yeung et al., 2023; Simpson et al., 2022; Yeung & Bunker, 2023). This vector can then be decoded by a Multi-Layer Perception (MLP), to infer the target variable(s). In our case, we want to infer the probabilities of each of the possible match outcomes. Moreover, conventional Recurrent Neural Network models such as Long-Term Short-Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997) and Gated Recurrent Unit (GRU) (Chung et al., 2014) could be used as the encoder. The performance of these encoders was compared with the TE in Sect. 4.3.

3.2 Approach 2: Feature selection and CatBoost

The feature selection and CatBoost approach leverage features engineered in prior studies (Baboota & Kaur, 2019; Berraret al., 2019; Tsokos et al., 2019) and more, as discussed in Sect. 3.2.1 and detailed in Table 2. This methodology utilizes feature selection techniques to ascertain the most suitable feature set, subsequently integrating it into the boosted decision tree model, CatBoost (Prokhorenkova et al., 2018), as elaborated in Sect. 3.2.2.

3.2.1 Features engineering and selection

In this section, we describe and explain the features that could be utilized and how they were selected.

Potential feature set In the 2017 Soccer Prediction Challenge, the first- and second-placed participants: Berraret al. (2019) and (Hubáček,Šourek, &Železný2019), respectively, both employed gradient tree boosting models. Nevertheless, they utilized distinct sets of engineered features. Given this disparity, this study aims to investigate and identify the optimal feature set. We constructed a potential feature set by concatenating features from noteworthy methodologies employed in the 2017 challenge, as outlined by Berraret al. (2019) (1st and 5th place), (Hubáček,Šourek, &Železný2019) (2nd place), and Tsokos et al. (2019) (4th place), as well as considering findings from other research focused on match result prediction (Baboota & Kaur, 2019). The comprehensive compilation of potential features is tabulated in Table 2, totalling 205 distinct features.

Feature selection The process of feature selection is grounded in assessing, e.g., the correlation or information gain between input features and the target variable, the instance distances, and sometimes, the correlation among the different input features. Initially, four prevalent feature filtering techniques from WEKA were adopted, which take into account both the information content and correlation with the target variable. These include the Chi-square, Symmetrical Uncertainty, Correlation, and Information Gain attribute evaluation methods. Following consideration of the median ranking of the features across these filter methods, the 20 features most relevant were chosen. Subsequently, the ReliefF feature selection method (Kira & Rendell, 1992; Kononenko, 1994) in the scikit-rebate library in Python (Urbanowicz et al., 2017) was employed to factor in the distances between instances within the feature space. This resulted in an additional set of top 20 features. In the next phase, the elimination of duplicated features derived from the two aforementioned

Table 2 Features considered for feature selection

Features	Description	Reference
GD	Cumulative goal difference during the season	
GS	Cumulative goals scored during the season	
GC	Cumulative goals conceded during the season	N/A
elo	Elo ratings	https://footballdatabase.com/methodology.php
Streak	Performance in most recent (n = 6) matches	
Weighted_Streak	Performances in most recent (n = 6) matches, weighted according to match recency	
Form	Performance relative to opponents	Baboota and Kaur (2019)
attacking_strength_i	Goals scored in match $t - i$, where $i \in \{1,2,...9\}$	
defensive_strength_i	Goals conceded in match $t - i$, where $i \in \{1,2,...9\}$	
strength_opposition_i	Home (away) team's opponent's average goal difference in their 9 most recent matches at time $t - i$, where i ∈ {1,2,...9}	
home_advantage_i	Whether the home team and away team played at home (1) or not (-1) in match at time $t - i$, where i ∈ {1,2,...9}	
H_Off_Rating	Berrar rating home offensive rating	
H_Def_Rating	Berrar rating home defensive rating	
A_Off_Rating	Berrar rating away offensive rating	
A_Def_Rating	Berrar rating away defensive rating	
EG	Berrar rating predicted goals	Berraret al. (2019)
newly_promoted	1 if newly promoted, 0 otherwise	
newly_demoted	1 if newly demoted, 0 otherwise	
days_since_previous	Number of days since the previous match	
Form2	Numbers of points gained in the last 3 matches, divided by 9	
point_tally	Number of points in the current season (up to but not including the current match)	
point_per_match	Average points scored per match in the current season (up to but not including the current match)	
previous_point_tally	Number of points in the previous season	
previous_GS	Number of goals in the previous season	

Table 2 (continued)

Features	Description	Reference
previous_GC	Number of goals conceded in the previous season	
previous_GD	Goal difference in the previous season	
days_since_first_match[1]	Number of days since the first match in the league	
quarter[1]	Calendar year quarter	Tsokos et al. (2019)
L_up_i	Points difference from the first i teams in the league table, where $i \in \{1,2,\ldots 5\}$	
L_down_i	Points difference from the bottom i teams in the league table, where $i \in \{1,2,\ldots 5\}$	
Home_venue_win_pct	Home winning percentage in the current and last 2 seasons	
Away_venue_win_pct	Away winning percentage in the current and last 2 seasons	
win_pct	Winning percentage in the current and last 2 seasons	
Home_venue_draw_pct	Home draw percentage in the current and last 2 seasons	
Away_venue_draw_pct	Away draw percentage in the current and last 2 seasons	
draw_pct	Draw percentage in the current and last 2 seasons	
Home_venue_GS_avg	Average number of goals scored at home in the current and last 2 seasons	
Away_venue_GS_avg	Average number of goals scored away in the current and last 2 seasons	
GS_avg	Average number of goals scored in the current and last 2 seasons	
Home_venue_GC_avg	Average number of goals conceded at home in the current and last 2 seasons	
Away_venue_GC_avg	Average number of goals conceded away in the current and last 2 seasons	
GC_avg	Average number of goals conceded in the current and last 2 seasons	
home_venue_goal_difference_avg	Average goal difference at home in the current and last 2 seasons	
away_venue_goal_difference_avg	Average away goal difference in the current and last 2 seasons	
goal_difference_std	Average goal difference in the current and last 2 seasons	
Home_venue_GS_std	Standard deviation in the number of goals scored at home in the current and last 2 seasons	
Away_venue_GS_std	Standard deviation in the number of goals scored away in the current and last 2 seasons	
GS_std	Standard deviation in the number of goals scored in the current and last 2 seasons	

Table 2 (continued)

Features	Description	Reference
Home_venue_GC_std	Standard deviation in the number of goals conceded at home in the current and last 2 seasons	
Away_venue_GC_std	Standard deviation in the number of goals conceded away in the current and last 2 seasons	
GC_std	Standard deviation in the number of goals conceded in the current and last 2 seasons	(Hubáček,Šourek &Železný2019)
home_venue_goal_difference_std	Standard deviation in the goal difference at home in current and last 2 seasons	
away_venue_goal_difference_std	Standard deviation in the away goal difference in the current and last 2 seasons	
goal_difference_std	Standard deviation in the goal difference in the current and last 2 seasons	
win_pct	Winning percentage in the last 5 matches	
draw_pct	Draw percentage in the last 5 matches	
GS_AVG	Average number of goals scored in the last 5 matches	
GC_AVG	Average number of goals conceded in the last 5 matches	
GS_STD	Standard deviation in the number of goals scored in the last 5 matches	
GC_STD	Standard deviation in the number of home team goals conceded in the last 5 matches	
home_venue_goal_scores_avg[1]	Average number of goals scored at home venues across the league in the last 2 seasons	
away_venue_goal_scores_avg[1]	Average number of goals scored at away venues across the league in the last 2 seasons	
home_venue_goal_scores_std[1]	Standard deviation in the number of goals per match scored at home venues across the league in the last 2 seasons	
away_venue_goal_scores_std[1]	Standard deviation in the number of goals scored per match at away venues across the leaguein the last 2 seasons	
home_venue_win_pct[1]	Winning percentage across the league in the last 2 seasons	
home_venue_draw_pct[1]	Draw percentage across the league in the last 2 seasons	
team_cnt[1]	Number of teams in the league in the previous season	
gd_std[1]	Standard deviation in the goal difference across the league in the last 2 seasons	
rnd_cnt[1]	Number of rounds in the league in the previous season	
Round[1]	The round in the current season	
PageRank	PageRank computed based on the current and previous 2 seasons	

Table 2 (continued)

Features	Description	Reference
pi_rating	Pi-rating computed based on the current and previous 4 seasons	(Hubáček,Šourek &Železnỳ2019)

All features are calculated for both the home team and the away team unless otherwise specified c

[1] Not based on the home or away team (computed once only)

feature selection methods (filter and ReliefF) led to a compilation of up to 40 features. Ultimately, the Correlation Subset Feature Selection (CFS) method (Hall, 1999) was utilized, which seeks to identify a feature set with the highest average correlation to the target feature, while simultaneously minimizing the average inter-feature correlation.

3.2.2 CatBoost

In this section, we provide an overview of the conventional boosted decision tree model and highlight the significant enhancements introduced by CatBoost (Prokhorenkova et al., 2018) compared to conventional models.

The boosted decision tree model is a form of decision tree (Loh, 2011; Friedman et al., 2000; Rokach & Maimon, 2005) integrated with gradient boosting (Friedman, 2001). In this model, a decision tree h divides the feature space $X^n \in \mathbb{R}^n$ into multiple (J) disjointed regions (tree nodes) N based on feature values. Mathematically, a decision tree h can be represented as follows, where $\hat{y}$ approximates the target variable $y \in \mathbb{R}$:

$$h(x) = \sum_{j=1}^{J} \hat{y}_j \mathbb{1}_{\{x \in N_j\}} \tag{3}$$

Gradient boosting (Friedman, 2001) operates iteratively on a sequence of approximations $F^t : \mathbb{R}^n \to \mathbb{R}$, $t = 0, 1, ...$ Each approximation F^t is obtained by updating the previous approximation F^{t-1} with a scaled version of a new decision tree h^t trained on the residual. This process is optimized through gradient descent (Friedman et al., 2000):

$$h^t = \underset{h \in H}{\operatorname{argmin}} \mathbb{E} L(y, F^{t-1} + h) \tag{4}$$

Recent studies (Berraret al., 2019);(Hubáčcek,Šourek, &Železný2019) have demonstrated strong performance using XGBoost (Chen & Guestrin, 2016), a type of boosted decision tree, for predicting soccer match results. However, CatBoost (Prokhorenkova et al., 2018) is a more recent boosted decision tree model that addresses the limitations of traditional models, particularly information leakage (Zhang et al., 2013). CatBoost introduces two key innovations. Firstly, it proposes ordered target encoding to handle categorical features more effectively. For the ith feature in the kth training sample, the encoded categorical feature $\hat{x}^i_k$ is calculated as:

$$\hat{x}^i_k = \frac{\sum_{x_j \in D_k} \mathbb{1}_{\{x^i_j = x^i_k\}} y_j + ap}{\sum_{x_j \in D_k} \mathbb{1}_{\{x^i_j = x^i_k\}} + a} \tag{5}$$

where $D_k = \{x_j : \sigma(j) < \sigma(k)\}$ for training samples and $D_k = D$ for testing samples, and σ, D, a, and p represent a random permutation function, dataset, parameter, and average target value, respectively.

Secondly, CatBoost introduces ordered boosting to obtain an unbiased estimation of y. This involves training n different models $M_1, ...M_n$, where each model M_i is trained with the first i data points ordered by permutation function σ. At each gradient step t, the residual of jth sample is computed using model M_{j-1}. Given these advancements, we opted for CatBoost as our model of choice. The CatBoost model was trained using the feature set described in Sect. 3.2.1.

4 Experiments and results

In this section, we commence by elucidating the methodology for model training and validation, which is detailed in subsection 4.1. Subsequently, we delve into the analysis and selection of models for Task 1 (exact score prediction) and Task 2 (probability prediction) in subsections 4.2 and 4.3, respectively. Finally, in subsection 4.4, we deliberate upon the chosen model, and evaluate and discuss its performance relative to the top-placed performers in the 2023 Soccer Prediction Challenge.

4.1 Model training and validation

Since the challenge set results were not available at the time of conducting the experiment (during the 2023 Soccer Prediction Challenge submission period), we utilized training and validation sets retrieved from the provided data. The objective of splitting the training sets was to simulate the data used for training models to predict the challenge prediction set's outcome, while the validation sets aimed to replicate the challenge prediction set itself. We employed multiple training and validation sets because certain models could exhibit high variability in performance when different years of data are used.

The dataset provided by the 2023 Soccer Prediction Challenge was partitioned into three distinct training sets and three corresponding validation sets. For the three respective training datasets, five years' worth of data was utilized, encompassing seasons up to round $x - 1$ of the 2018–19, 2019–20, and 2020–21 seasons. Conversely, the three respective validation sets were constructed using rounds x and $x + 1$ of the 2018–19, 2019–20, and 2020–21 seasons. The x and $x + 1$ correspond to the league rounds within the prediction set that spans the period from April 14 to April 26, 2023.

To ascertain the effectiveness of the proposed methodologies, i.e., (1) Deep Learning and (2) Feature Selection combined with CatBoost, the performance of each methodology was contrasted against baseline models, as well as ablated models, which investigate the effect of the exclusion or replacement of specific model components. For Task 1 (exact score prediction), our baseline models consisted of two straightforward statistical approaches, leveraging historical data to make predictions. Specifically, we calculated the mean scores for both the home and away team within each team and league in the training set. In the validation set, these mean scores were then utilized as predictions for both home team scores (HS) and away team scores (AS). We denoted the approach where the mean score of each team was used as the "team average" method, while the approach utilizing the league-wide mean score was termed the "league average" method. Furthermore, models from previous research efforts were included, such as the Berrar ratings (rating feature learning) (Berraret al., 2019), XGBoost applied to Berrar ratings, and CatBoost applied to pi-ratings (Razali et al., 2022).

In Task 2 (probabilities prediction), we employed two basic baselines: team win/draw/loss (W/D/L) percentages and a rule-based benchmark that consistently predicted a home team victory. The W/D/L percentages provided a straightforward method for prediction by analyzing past outcomes. By examining the frequency of wins, draws, and losses for each team, we could estimate the likelihood of different outcomes in future matches. Additionally, the rule-based benchmark that always predicted a home team victory reflected the well-known phenomenon of home advantage in sports, where the probability of the home team winning was generally higher compared to the away team.

Table 3 Exact scores prediction (Task 1) model results on the validation sets detailed in Sect. 4.1

Model	Avg RMSE	Sigma
Berrar ratings	1.0047	0.0434
Team average	1.0206	0.0540
XGBoost+Berrar ratings	1.0212	0.0381
League average	1.0346	0.0347
CatBoost+selected feature set (Approach 2)	1.2162	0.0053
CatBoost+pi-ratings	1.2356	0.0355
TE+MLP (Approach 1)	1.5063	0.0317

The model that was preferred and used in the challenge is shown in **bold**

Additionally, models from earlier literature were integrated, including the best-performing model from the 2017 Soccer Prediction Challenge—Berrar ratings coupled with XGBoost (Berraret al., 2019)—alongside the best-performing model published in a study after the conclusion of the 2017 challenge, which applied a CatBoost model to pi-ratings features (Razali et al., 2022).

Furthermore, within the domain of Inception+TE+MLP model (Approach 1), ablated models were explored, which, as mentioned, are models in which the effect of the exclusion or replacement of specific components is investigated. For instance, a model that excludes Information Extraction (TE+MLP), as well as models that substitute the encoder with LSTM or GRU.

All Models for Task 1 (exact scores prediction) and Task 2 (probabilities prediction) were trained to minimize Eqs. 1 and 2, respectively. The hyperparameters for Inception+TE+MLP model (Approach 1) for Task 2 are listed in Table 7.

4.2 Exact scores prediction results

In this subsection, the performance of exact score prediction using approaches 1 and 2 is compared with that of the baselines mentioned in Sect. 4.1. The optimal feature set, chosen through approach 2, is presented in Table 8.

Through analyzing Table 3, it was evident that the Berrar ratings (rating feature learning) exhibited superior performance compared to the competing models. Notably, the team average statistical baseline followed, while the remaining models comprised primarily gradient tree boosting and deep learning models. This suggests that, despite the capability of gradient tree boosting and deep learning models in capturing complex relationships in data, these models proved unsuitable for accurately modeling exact match scores in soccer. In contrast, the Berrar ratings-based model, which incorporates team performance and domain knowledge related to matches, emerged as the more suitable choice for this task.

Given that Berrar ratings consistently outperformed the proposed approaches and other baselines, it was designated as the final model for Task 1 (exact score prediction). For further details on how the Berrar ratings are calculated, please refer to Berraret al. (2019).

Table 4 Probabilities prediction (Task 2) model results on the validation sets detailed in Sect. 4.1

Model	Avg RPS	Sigma
CatBoost+pi-ratings	0.2085	0.0083
Inception+TE+MLP (Approach 1)	0.2098	0.0051
LSTM+MLP	0.2105	0.0050
TE+MLP	0.2111	0.0062
GRU+MLP	0.2116	0.0052
XGBoost+Berrar ratings	0.2141	0.0046
W/D/L percentage	0.2303	0.0015
CatBoost+selected feature set (Approach 2)	0.2416	0.0028
Home win	0.4450	0.0031

The model that was preferred and used in the challenge is shown in **bold**

4.3 Probabilities prediction results

The performance of approaches 1 and 2 in probabilities prediction was then compared to the baseline models and ablated models previously described in subsection 4.1. The optimal feature set selected in Approach 2 is listed in Table 8.

Even though Table 4 clearly indicated that among the various models, the CatBoost+pi-ratings model exhibited the highest performance, with the lowest average RPS loss, further examination revealed nuances. Upon scrutinizing the results of the three distinct validations detailed in Sect. 4.1 independently, we arrived at the conclusion that the second-best performing model in Table 4, Inception+TE+MLP (Approach 1), should be employed for the challenge set.

Specifically, the RPS values for the CatBoost+pi-ratings model in the 2018–19, 2019–20, and 2020–21 validation sets were 0.1991, 0.2120, and 0.2145, respectively. Meanwhile, Inception+TE+MLP achieved RPS values of 0.2072, 0.2102, and 0.2141 in the respective validation sets. A comparison of the CatBoost+pi-ratings model's performance across the validation sets reveals that its performance in 2018–19 was approximately 7% lower than in the other two validation sets (this discrepancy is noteworthy considering the comparison in Table 6 where the difference between the 1st and 16th is also around 7%). This suggested a potential risk of over-fitting, where the model might not have generalized well or might have excessively captured the variability present in the matches of the 2018–19 validation set.

However, drawing definitive conclusions required additional context. Therefore, we opted to disregard the results of the CatBoost+pi-ratings model on the 2018–19 validation set when considering the final submission model. Consequently, Inception+TE+MLP outperformed the CatBoost+pi-ratings model and other baseline models in both the 2019–20 and 2020–21 validation sets, making the Inception+TE+MLP model (Approach 1) the preferred final model for the challenge.

Upon further scrutiny of the ablated version of Inception+TE+MLP model (Approach 1), the model devoid of the inception block (TE+MLP) displayed weaker performance. Moreover, in comparing the encoder LSTM, TE, and GRU, the LSTM exhibited the most favorable performance. However, due to the considerable training time required for hyperparameter grid searching in LSTM, TE was selected as a more practical alternative given the limited time available to meet the competition deadline.

Table 5 Task 1 result compared to the top-performing approach

Team	RMSE
TeamNateWeller (1st)	1.6235
Berrar ratings (13th)	1.8169

Table 6 Task 2 result compared to the top-performing approach

Team	RPS
Bookmakers (1st)	0.2063
Inception+TE+MLP model (Approach 1) (16th)	0.2195

Lastly, the performance of the CatBoost+Selected feature set model fell short of the performance of the CatBoost+pi-ratings model. Consequently, the prospect of augmenting the model with additional engineered features seemed less promising. Instead, focusing on enhancing pi-ratings appeared a more viable strategy for further work. Nevertheless, the incorporation of additional engineered features in Inception+TE+MLP model (Approach 1) is also an avenue for potential further research.

4.4 2023 Soccer prediction challenge comparative evaluation

The final model was trained using 5 years of data until April 14, whereas the provided training set only provided data until April 4, 2023. To address this gap, as previously mentioned, data from April 4, 2023, to April 14, 2023, was manually appended to the training set.

Our performance in the 2023 Soccer Prediction Challenge in the two required challenge tasks, compared to the top-performing team, is summarized in Tables 5 and 6. In Task 1, the Berrar ratings (rating feature learning) (Berraret al., 2019) were surpassed by the first-place team by a margin of 11.91% ranked 13th. As for Task 2, our Inception+TE+MLP model (Approach 1) involving deep learning was outperformed by a bookmaker consensus type model by 6.42% ranked 16th. Given that the 2023 Soccer Prediction Challenge permitted the utilization of alternative features and training instances beyond the provided dataset, the integration of additional features stands as a potential avenue for enhancing model performance. This is particularly relevant to Inception+TE+MLP model (Approach 1) for Task 2, where alternative features could be seamlessly incorporated.

Upon scrutinizing the validation and results from the 2023 soccer prediction challenge, it is apparent that, in terms of exact score prediction, machine learning models presently lag behind rule-based counterparts such as team ratings. Furthermore, the consideration of alternative features, encompassing elements like expert opinions, team formations, and player details—elements potentially encapsulated in betting odds—has the potential to bolster the efficacy of deep learning models in match outcome probability prediction. Nevertheless, there is a practical challenge associated with the use of additional information as input. The inclusion of additional information would necessitate conducting a grid search on both feature sets and hyperparameters. This process could be computationally intensive and time-consuming, potentially leading to a significant increase in resource requirements. Moreover, the increased complexity introduced by additional features may also lead to issues such as overfitting, making the models less generalizable to unseen data.

5 Conclusion

In this study, our objective was to assess the performance of a deep learning model and determine the optimal feature set for a gradient-boosted tree model in predicting soccer match results in terms of win/draw/loss (W/D/L) probabilities as well as exact scores. To achieve this aim, we introduced a deep learning-based model and leveraged features from prior studies, coupled with feature selection algorithms, to identify the most effective feature set. Our models were trained, validated, and tested using data from the 2023 Soccer Prediction Challenge. The results revealed that, in terms of W/D/L probabilities, the deep learning model was outperformed by betting odds consensus model predictions by a margin of 6% and ranked 16th in the 2023 Soccer Prediction Challenge. Moreover, it was noteworthy that pi-ratings still retained their status as the most suitable features for using with gradient-boosted tree models. When it comes to predicting exact scores, the Berrar ratings (rating feature learning) and simple statistical baseline models exhibited superior performance compared to both the deep learning and gradient-boosted tree models.

Future endeavors could focus on enhancing model interpretability. Given that both deep learning and boosted decision tree models fall under the category of black-box models, efforts to enhance interpretability would greatly benefit coaches and analysts in identifying the pivotal features for achieving victory. Furthermore, consideration could be given to alternative features, such as betting odds that encapsulate expert opinions, team compositions, player characteristics, and more. Despite this, we anticipate that our study will underscore the efficacy of employing deep learning methodologies in predicting soccer match outcomes, thereby inspiring forthcoming research in this domain. The advancements made in predicting soccer match outcomes can potentially be extrapolated to other team sports. For comprehensive data on publicly accessible team sports match results, please refer to Table 9.

Model Hyperparameters

See Table 7.

Table 7 Optimal hyperparmeters for Task 2 Inception+TE+MLP model (Approach 1)

Hyperparameter	Grid-searched value	Optimal value
team_id_embedding_dim	1,2,4,8,16	1
TE_dim_feedforward	1,8,64,512,2048,4096	1
TE_dropout	0,0.1,0.2,0.3	0
MLP_num_layer	1,2,3,4,5,6,7,8,9,10,11,12	10
MLP_dropout	0,0.1,0.2,0.3	0.2

Approach 2 selected feature set

See Table 8.

Table 8 Optimal feature set from Approach 2

Target feature	Optimal feature set
Home team goals	EG_HT, GS_avg_HT
Away team goals	Home_venue_GS_avg_AT, GC_avg_HT, pi_rating_AT, GC_AVG_HT, previous_GD_AT
W/D/L probability	EG_HT, EG_AT, point_per_match_HT, win_pct_AT, pi_rating_HT, pi_rating_AT

An explanation of the features is available in Table 2. HT or AT denotes the features that are calculated based on Home Team or Away Team, respectively

Datasets of team sports match results

See Table 9.

Table 9 Datasets of team sports match results

Sports	Tournaments	Years	Reference
American Football	National Football League (NFL)	2009 ~	https://github.com/maksimhorowitz/nflscrapR
Australian Football	Australian Football League (AFL) Mens	2012 ~	https://github.com/jimmyday12/fitzRoy?tab=readme-ov-file
	Australian Football League (AFL) Womens	2019 ~	
Baseball	Major League Baseball (MLB)	2010–2021	https://sports-statistics.com/sports-data/ mlb-historical-odds-scores-datasets/
Basketball	National Basketball Association (NBA)	-	https://github.com/seemethere/nba_py
	National Collegiate Athletic Association (NCAA)	2013–2018	https://www.kaggle.com/datasets/ncaa/ncaa-basketball
Cricket	T20 internationals and more	2002–2024	https://cricsheet.org/matches/
Ice Hockey	National Hockey League (NHL)	2000–2020	https://www.kaggle.com/datasets/martinellis/nhl-game-data
Rugby	Premiership Rugby	-	https://github.com/seanyboi/rugbypy
	Rugby World Cup		
	Six Nations Championship		
Tennis	Association of Tennis Professionals (ATP) Men's Tour	2000–2024	http://www.tennis-data.co.uk/alldata.php
	Women's Tennis Association (WTA) Women's Tour	2007–2024	

The references encompass datasets, web scraping tools, and APIs designed for retrieving match results for specific tournaments. However, it's worth noting that certain years remain unspecified due to the lack of information provided within the API or scraping tools, which couldn't be verified

Acknowledgements This work was financially supported by JST SPRING (Grant Number JPMJSP2125). Calvin Yeung would like to take this opportunity to thank the "Interdisciplinary Frontier Next-Generation Researcher Program of the Tokai Higher Education and Research System."

Author Contributions All authors contributed to the study conception and design. Data preparation, modeling, and analysis were performed by all authors. The first draft of the manuscript was written by all authors and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding This work was financially supported by JST SPRING, Grant Number JPMJSP2125.

Availability of data and materials The data of the research was collected and provided by the 2023 Soccer Prediction Challenge

Code availability The code for the model is available at https://github.com/calvinyeungck/Soccer-Prediction-Challenge-2023.

Declarations

Conflict of interest The authors have no conflict of interest to declare that are relevant to the content of this article.

Consent to participate Not applicable.

Consent for publication Not applicable.

Ethics approval Not applicable.

References

Akan, S., & Varlı, S. (2023). Use of deep learning in soccer videos analysis: survey. *Multimedia Systems, 29*(3), 897–915.

Baboota, R., & 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.

Berrar, D., Lopes, P., Davis, J., & Dubitzky, W. (2019). Guest editorial: special issue on machine learning for soccer. *Machine Learning, 108*, 1–7.

Berrar, D., Lopes, P., & Dubitzky, W. (2019). Incorporating domain knowledge in machine learning for soccer outcome prediction. *Machine Learning, 108*, 97–126.

Bunker, R., & Susnjak, T. (2022). The application of machine learning techniques for predicting match results in team sport: A review. *Journal of Artificial Intelligence Research, 73*, 1285–1322.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM sigkdd international conference on knowledge discovery and data mining* (pp. 785–794).

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv:1412.3555

Constantinou, A. C. (2019). Dolores: A model that predicts football match outcomes from all over the world. *Machine Learning, 108*(1), 49–75.

Constantinou, A. C., & Fenton, N. E. (2012). Solving the problem of inadequate scoring rules for assessing probabilistic football forecast models. *Journal of Quantitative Analysis in Sports*, *8*(1).

Constantinou, A. C., & Fenton, N. E. (2013). Determining the level of ability of football teams by dynamic ratings based on the relative discrepancies in scores between adversaries. *Journal of Quantitative Analysis in Sports, 9*(1), 37–50.

Danisik, N., Lacko, P., & Farkas, M. (2018). Football match prediction using players attributes. *2018 World symposium on digital intelligence for systems and machines (DISA)* (pp. 201–206).

Decroos, T., Bransen, L., Van Haaren, J., & 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).

Dixon, M. J., & Coles, S. G. (1997). Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 46*(2), 265–280.

Dubitzky, W., Lopes, P., Davis, J., & Berrar, D. (2019). The open international soccer database for machine learning. *Machine Learning, 108*, 9–28.

Epstein, E. S. (1969). A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology (1962–1982)*, *8*(6), 985–987.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors). *The Annals of Statistics, 28*(2), 337–407.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of statistics*, 1189–1232.

Hall, M. A. (1999). Correlation-based feature subset selection for machine learning. *Thesis submitted in partial fulfilment of the requirements of the degree of Doctor of Philosophy at the University of Waikato*.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.

Hubáček, O., Šourek, G., & Železný, F. (2019). Learning to predict soccer results from relational data with gradient boosted trees. *Machine Learning, 108*, 29–47.

Hubáček, O., Šourek, G., & Železný, F. (2022). Forty years of score-based soccer match outcome prediction: an experimental review. *IMA Journal of Management Mathematics, 33*(1), 1–18.

Hvattum, L. M., & Arntzen, H. (2010). Using elo ratings for match result prediction in association football. *International Journal of Forecasting, 26*(3), 460–470.

Jain, S., Tiwari, E., & Sardar, P. (2021). Soccer result prediction using deep learning and neural networks. In *Intelligent data communication technologies and internet of things: Proceedings of ICICI 2020* (pp. 697–707).

Joseph, L. D. (2022). *Time series approaches to predict soccer match outcome (Unpublished doctoral dissertation)*. National College of Ireland.

Kira, K., & Rendell, L.A. (1992). A practical approach to feature selection. In *Machine learning proceedings* (pp. 249–256). Elsevier.

Kononenko, I. (1994). Estimating attributes: Analysis and extensions of relief. *European conference on machine learning* (pp. 171–182).

Loh, W.-Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1*(1), 14–23.

Maher, M. J. (1982). Modelling association football scores. *Statistica Neerlandica, 36*(3), 109–118.

Malamatinos, M.-C., Vrochidou, E., & Papakostas, G. A. (2022). On predicting soccer outcomes in the Greek league using machine learning. *Computers, 11*(9), 133.

Natarajan, S., Khot, T., Kersting, K., Gutmann, B., & Shavlik, J. (2012). Gradient-based boosting for statistical relational learning: The relational dependency network case. *Machine Learning, 86*, 25–56.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). *The pagerank citation ranking: Bring order to the web* (Tech. Rep.). Technical report, Stanford University.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). Catboost: unbiased boosting with categorical features. *Advances in Neural Information Processing systems*, *31*.

Rahman, M. A. (2020). A deep learning framework for football match prediction. *SN Applied Sciences, 2*(2), 165.

Razali, M. N., Mustapha, A., Mostafa, S. A., & Gunasekaran, S. S. (2022). Football matches outcomes prediction based on gradient boosting algorithms and football rating system. *Human Factors in Software and Systems Engineering, 61*, 57.

Razali, N., Mustapha, A., Arbaiy, N., & Lin, P.-C. (2022). Deep learning for football outcomes prediction based on football rating system. In *Aip conference proceedings* (Vol. 2644).

Robberechts, P., & Davis, J. (2019). Forecasting the fifa world cup-combining result-and goal-based team ability parameters. *Machine learning and data mining for sports analytics: 5th international workshop, MLSA 2018, co-located with ECML/PKDD 2018, Dublin, Ireland, September 10, 2018, proceedings 5* (pp. 16–30).

Rokach, L., & Maimon, O. (2005). Top-down induction of decision trees classifiers-a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *35*(4), 476–487.

Simpson, I., Beal, R.J., Locke, D., & Norman, T.J. (2022). Seq2event: Learning the language of soccer using transformer-based match event prediction. In *Proceedings of the 28th ACM sigkdd conference on knowledge discovery and data mining* (pp. 3898–3908).

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1–9).

Tsokos, A., Narayanan, S., Kosmidis, I., Baio, G., Cucuringu, M., Whitaker, G., & Király, F. (2019). Modeling outcomes of soccer matches. *Machine Learning, 108*, 77–95.

Urbanowicz, R. J., Olson, R. S., Schmitt, P., Meeker, M., & Moore, J. H. (2017). *Benchmarking relief-based feature selection methods*. arXiv:https://arxiv.org/abs/1711.08477

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Wheatcroft, E. (2021). Evaluating probabilistic forecasts of football matches: The case against the ranked probability score. *Journal of Quantitative Analysis in Sports, 17*(4), 273–287.

Wright, M. (2009). 50 years of or in sport. *Journal of the Operational Research Society, 60*, S161–S168.

Wu, H., Hu, T., Liu, Y., Zhou, H., Wang, J., & Long, M. (2022). Timesnet: Temporal 2d-variation modeling for general time series analysis. arXiv:2210.02186

Yeung, C., & Bunker, R. (2023). An events and 360 data-driven approach for extracting team tactics and evaluating performance in football. *StatsBomb Conference 2023*.

Yeung, C., Bunker, R., & Fujii, K. (2023). A framework of interpretable match results prediction in football with fifa ratings and team formation. *PLoS ONE, 18*(4), e0284318.

Yeung, C., & Fujii, K. (2023). A strategic framework for optimal decisions in football 1-vs-1 shot-taking situations: An integrated approach of machine learning, theory-based modeling, and game theory. arXiv:2307.14732

Yeung, C., Sit, T., & Fujii, K. (2023). Transformer-based neural marked spatio temporal point process model for football match events analysis. arXiv:2302.09276

Zhang, K., Schölkopf, B., Muandet, K., & Wang, Z. (2013). Domain adaptation under target and conditional shift. In *International conference on machine learning* (pp. 819–827).

Zhang, Q., Zhang, X., Hu, H., Li, C., Lin, Y., & Ma, R. (2022). Sports match prediction model for training and exercise using attention-based lstm network. *Digital Communications and Networks, 8*(4), 508–515.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.