Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/21/20231136

Research on soccer prediction model based on machine learning combined with domain knowledge

Junxian Song

Department of Mathematics, Harbin Institute of Technology, Harbin City, Heilongjiang Province, 150001, China

120L041004@stu.hit.edu.cn

Abstract. With the global popularity of soccer and the increasing collection of data, more and more research focuses on using machine learning (ML) algorithms to predict match outcomes. However, accurately predicting soccer match results is a complex task that requires considering multiple factors such as team strength and player status. This paper aims to predict soccer match results using ML techniques in Python. To achieve this goal, a series of methods are employed to collect, analyze, and utilize historical match data, combined with knowledge and experience from the soccer domain, to identify factors relevant to predicting soccer match outcomes. By establishing appropriate feature extraction and selection methods, this paper is able to capture information closely related to match results. Based on this foundation, the paper proceeds with model development and enhances its performance through training and parameter tuning. Specifically, ML algorithms such as Support Vector Machines (SVM) and various techniques are applied to optimize the models and improve their predictive accuracy. Emphasis is placed on model stability and generalization, taking into account issues of overfitting and underfitting during the training process, with appropriate regularization and cross-validation techniques. Finally, a comprehensive performance evaluation is conducted on the established models, and a comparative analysis of different algorithms is performed. The experimental results demonstrate the excellent performance of the models proposed in this paper in predicting soccer match results, showcasing the significant potential of ML in this domain.

Keywords: machine learning, domain knowledge, extreme gradient boosting, support vector machines, performance evaluation.

1. Introduction

As one of the most popular sports worldwide, soccer attracts billions of viewers and possesses a vast market value. One of the captivating aspects of soccer lies in the complex interplay of various factors and chance events that significantly influence the final outcome of soccer matches, thereby rendering the prediction of soccer match results challenging. Moreover, the ability to predict soccer match results holds immense significance for coaches, club managers and fans alike. It aids coaches in formulating effective strategies, enables club managers to enhance team management, and allows fans to speculate on match outcomes. Over the past few decades, researchers have been attempting to predict soccer match results using various methods. However, despite some promising outcomes,

DOI: 10.54254/2755-2721/21/20231136

predicting soccer match results remains a challenging problem. Soccer matches are influenced by multiple factors, including team strength, player injuries, tactical strategies, and match venues. Furthermore, the dynamic nature and uncertainty of soccer matches further increase the difficulty of prediction. Therefore, predicting soccer match results remains a challenging problem that requires continuous research and improvement. The overall structure of this paper is outlined as follows.

In Section 2, decisive factors influencing soccer match outcomes will be investigated. By analyzing historical match data and employing feature engineering techniques, relevant factors related to predicting soccer match results will be identified and extracted, thereby enhancing our understanding of the interrelationships and impacts among different factors.

In Section 3, various machine learning (ML) algorithms, such as extreme gradient boosting and support vector machines (SVM), will be evaluated to compare their performance in predicting soccer match results. Through a comprehensive assessment, the most effective algorithm for predicting soccer match outcomes will be determined.

Section 4 focuses on improving the accuracy of soccer match result predictions by exploring different feature selection methods and ML model algorithms. The aim is to identify the optimal combination that enhances prediction performance.

2. Related work

In the past few decades, numerous researchers have devoted their efforts to utilizing ML techniques for predicting soccer match results. These studies can be primarily categorized into two directions: prediction models based on statistical methods and those based on ML methods.

Prediction models based on statistical methods predominantly employ traditional statistical approaches such as regression analysis and Poisson regression models. These methods analyze historical match data, consider various factors (e.g. team strength, offensive capabilities, defensive capabilities) as predictor variables, and establish mathematical models to predict match outcomes. For instance, Reep and Benjamin proposed a Poisson regression-based model to predict the number of goals in soccer matches [1]. Maher employed regression analysis to predict win or loss outcomes [2]. Although these statistical methods have achieved certain predictive performance, their predictive capabilities are somewhat limited due to assumptions of linear relationships and specific model forms.

In comparison, prediction models based on ML methods are more flexible and adaptable. ML approaches learn from large volumes of historical match data, automatically discover features and patterns, and construct prediction models. Commonly used ML algorithms for predicting soccer match results include random forests [3], SVM [4], neural networks [5, 6], among others. It is worth mentioning that O'Donoghue et al. attempted to predict soccer match outcomes using decision tree algorithms and achieved a certain level of prediction accuracy [7]. However, due to the complexity and dynamics of soccer matches, individual ML algorithms often fail to capture all factors and correlations comprehensively.

To further enhance prediction accuracy, some researchers have employed ensemble learning methods by combining multiple ML models. For example, Dixon and Coles proposed a Bayesian-based ensemble model that combines multiple Poisson regression models to predict match outcomes [8]. Their research indicates that ensemble models outperform individual models in certain aspects of predictive performance. Furthermore, some studies have explored the integration of domain knowledge with ML by introducing additional features and constraints to improve prediction accuracy and interpretability [9-11].

3. Feature engineering

The dataset used in this paper is composed of multiple publicly available datasets from Kaggle and the official website of the English Premier League (EPL). The dataset includes all matches from 2010 to 2023, with 113 statistical features per match and the team's overall season performance up until the respective match. The statistical features included in the dataset cover various aspects of football

DOI: 10.54254/2755-2721/21/20231136

matches, including team passing, shooting, clearances, and more. This dataset provides comprehensive information for analysis and prediction purposes.

3.1. Domain knowledge integration

The main idea of this part is to utilize features to describe the strengths and weaknesses of each team in a match. Relevant knowledge in the field of soccer suggests to consider the following aspects:

Table 1. Relevant knowledge in football field.

Features	Caption
Offensive ability	The team's scoring ability.
Defensive ability	The team's ability to prevent opponents from scoring.
Recent performance	A team's current state based on their overall performance in recent matches.
Opponent strength	Evaluates the strength of the opponent team based on their current league ranking.
Home advantage	The additional benefits a team possesses when playing at their home stadium.

Both offensive ability and defensive ability are evident and intuitive. Offensive ability involves a team's capacity to create scoring opportunities during attacks, such as ball control, passing, shooting, and creating offensive combinations. Its evaluation is based on multiple indicators, including average goals scored, average shots taken, average shot conversion rate, and average number of passes. Defensive ability involves a team's capacity to organize, block, intercept, and restrict opponent attacks during defense. Its evaluation can consider indicators such as the team's average goals conceded, average interceptions made, average clearances, etc. Obviously, the stronger the offensive and defensive ability of a team, the more likely they are to beat their opponents.

Recent performance can be considered based on a team's match results, scoring, win rate, draw rate, goals conceded, and number of victories within a certain time frame. Specifically, the evaluation of recent performance can be determined by comparing a team's results and statistical data in the past few matches. A team's recent good performance may indicate their competitive state, tactical coordination, and individual player's technical abilities.

Opponent strength refers to the level of strength possessed by the opposing team in a match. Evaluating opponent strength can take into account factors such as the opposing team's current league ranking, recent match results, win rate, and goal difference.

Home advantage refers to the additional advantages a team has in home games. In football matches, the home team benefits from the home atmosphere, familiarity with the field conditions, and support from home fans, leading to improved performance [12]. Home advantage can influence the team's psychological state, morale, and overall performance, making them more confident and proactive during the game. According to the research conducted by Dubitzky et al., home teams have been found to win 45.42% of the matches, draw 27.47% of the matches, and lose 27.11% of the matches[13], quantifying the advantage of the home team in football matches.

3.2. Feature engineering method

The approach used in this paper for feature modelling is to represent each match through three feature groups for each team. The three feature groups for each team are the offensive ability feature group, defensive ability feature group, and opponent strength feature group. The paper employs the recent feature extraction method for partial feature extraction to describe the recent performance of teams. Each of the three feature groups of each team consists of n features, where n represents the depth of obtaining feature values from the match time series.

Therefore, the most recent feature extraction method $2\times3\times n$ is used to describe the total number of prediction features used for matching, which reflects 2 teams, each with 3 feature groups.

First, the specific value of n is determined. Four possible depths, $n=12$, $n=9$, $n=6$, and $n=3$, were selected to generate features based on their predictive performance. $n=9$ is the best setting. Next, this

DOI: 10.54254/2755-2721/21/20231136

paper proceeds with feature selection to determine the specific form of each feature in every group. To accomplish this task, the paper employs the min-max normalization method to linearly scale feature values and map them to a specified range. Information gain is utilized to identify features that contribute significantly and are essential for the target variable in the classification task. Subsequently, the chi-square test is applied to further evaluate the association between features and determine if there are significant correlations. Lastly, the correlation coefficient is used to measure the linear relationship between features, resulting in a heatmap of correlation coefficients. By calculating these indicators, a score is assigned to each feature, which is then used to construct the final selection.

Taking the example of the offensive ability feature set, the correlation matrix of each feature is shown in Figure 1.

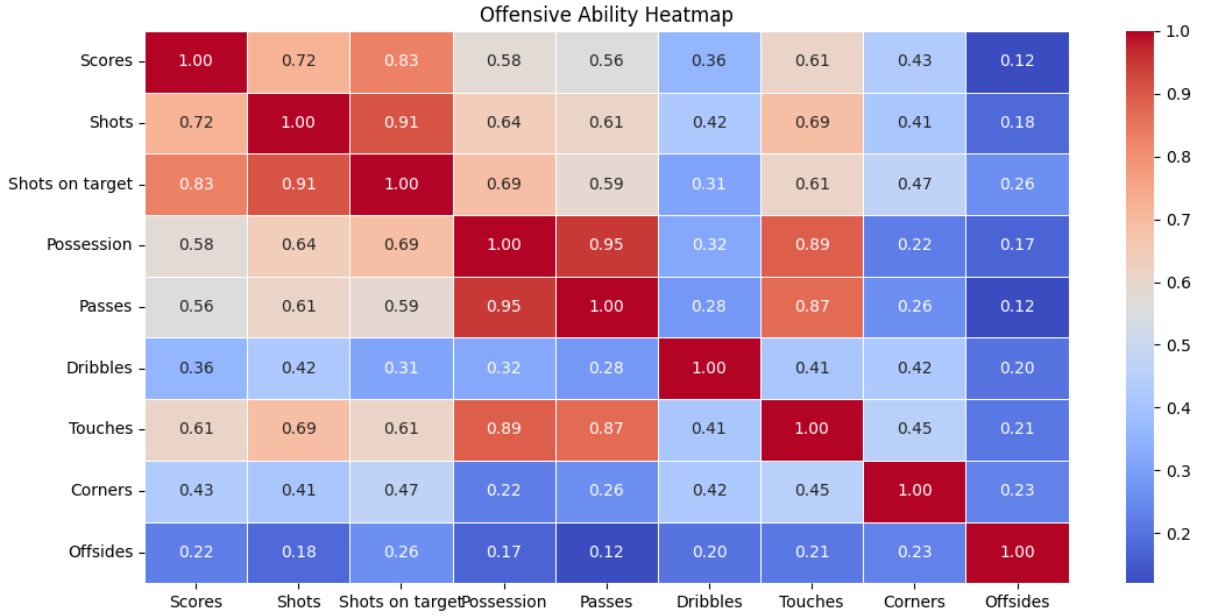

Figure 1. Correlation matrix of the offensive ability. (Photo/Picture credit: Original)

According to the analysis results in Figure 1, it can be observed that highly correlated features may contain similar information. To mitigate issues of redundancy and multicollinearity, this study chose to discard the features "Shots on target" and "Possession". Furthermore, through the adoption of a filtering method, the feature "Offsides" with a low correlation coefficient was also eliminated. By integrating domain knowledge and the results of feature selection, this study ultimately established the following formula to describe the offensive ability:

$$A_{oa} = 0.72 \times A_{Shots} + 0.36 \times A_{Dribbles} + 0.43 \times A_{Corners} + 0.28 \times A_{Passes} + 0.31 \times A_{Touches} \quad (1)$$

Finally, home advantage will be considered as a coefficient to exert an influence on the home team's feature group. Based on the research by Dubitzky, a coefficient of h=0.18 will be introduced to multiply with the home team's feature group. This approach allows for explicit consideration of the potential impact of home advantage on the home team's performance. By incorporating this method, the model can more accurately predict and explain the home team's outcomes, and enhance the comprehensive analysis of match results. The specific equations are as follows.

$$H_{oa} = (1 + h)A_{oa} \quad (2)$$

where H_{oa} are the offensive ability of the home team and A_{oa} are the offensive ability of the away team.

DOI: 10.54254/2755-2721/21/20231136

4. Predictive models

4.1. Machine learning algorithms

In this paper, two learning algorithms, SVM and extreme gradient lifting (xgboost), are used to construct the prediction model from the data set.

SVM is a widely applied in pattern recognition, classification, and regression problems. Its basic concept involves finding the optimal hyperplane for classification in the feature space. SVM maps the data to the high-dimensional feature space, and transforms the low dimensional linear inseparable problem into the high-dimensional linear separable problem. Then, it finds the optimal hyperplane that maximizes the margin between positive samples and negative samples in the feature space. This approach yields a classifier with good generalization ability. SVM performs well when dealing with high-dimensional data and small sample sizes. By conducting computations in the feature space, SVM can handle data with a large number of features and effectively mitigate overfitting issues in scenarios with a small sample size.

Extreme gradient lifting (xgboost) is a ML algorithm based on gradient lifting tree, which is famous for its powerful prediction performance and wide application. XGBoost iteratively trains multiple weak learners (decision trees) and combines them into a powerful ensemble model. The algorithm optimizes the objective function by approximating the loss function using gradient information, aiming to minimize prediction errors. XGBoost has demonstrated outstanding performance in various data mining and ML competitions and is the preferred algorithm in many award-winning solutions. Its successful application in Kaggle competitions signifies its remarkable advantage in terms of predictive performance.

Based on the above analysis, SVM and XGBoost were employed in this study as learning algorithms to construct predictive models from the dataset. SVM offers advantages in handling high-dimensional data and small sample sizes, while XGBoost exhibits powerful predictive performance and extensive applicability. By utilizing these algorithms, this study aims to develop accurate and robust predictive models for various applications in the field of soccer match prediction.

4.2. Model construction and parameter tuning

Based on the three performance rating characteristics mentioned in Section 3 (one for each team), A goal prediction model is defined to predict the goals of the home team and the away team respectively.

$$H_{\tilde{g}} = \frac{\alpha}{1+exp(-\beta(H_{oa}-A_{da})+\chi A_{os})} \tag{3}$$

$$A_{\tilde{g}} = \frac{\alpha}{1+exp(-\beta(A_{oa}-H_{da})+\chi H_{os})} \tag{4}$$

where $H_{\tilde{g}}$are the predicted home-team goal and $A_{\tilde{g}}$are the predicted goals away-team's score. H_{oa} are the offensive ability of the home-team and $A_{oa}$are the away-team's offensive ability. H_{da} are the defensive ability of the home-team and A_{oa} are the away-team's defensive ability. H_{os} are the opponent strength of the home-team and $A_{os}$are the away-team's opponent strength.α, β, χ are three constants.

The prediction model can predict the number of goals scored by the home team based on their offensive ability at home and their defensive ability away from home. The higher the predicted result of the model, the more goals the home team will score.

People need to consider the defensive strength of the visiting team and the offensive strength of the home team, and then predict the number of goals scored by the home team in a specific game, while incorporating the away team's league ranking as a correction factor for their strength. The higher the resulting value, the more goals the home team is expected to score. A similar approach is adopted for predicting the number of goals scored by the away team. Both Equation 3 and Equation 4 employ the sigmoid function. This function form is commonly used to describe a process that starts from a small value, accelerates, and eventually approaches a plateau. Since it can be known that the number of

DOI: 10.54254/2755-2721/21/20231136

goals in football matches is typically low (usually not exceeding 7), this paper set α to 7 in our experiments, while the values of β and χ will be determined during the learning phase. In addition, three goal difference can almost guarantee a certain victory based on domain knowledge. Therefore, teams often no longer strive to score more goals but tend to protect their lead. As a result, the sigmoid function naturally becomes the choice for the target prediction model.

Next, the Section will estimate the model parametersβ and χ. A subset of data, denoted as the learning set L, will be selected, which includes the results of past football matches sorted in chronological order, from the earliest to the most recent matches. Then, an ratings table will be derived from L to track the performance ratings of each team. The algorithm is applied iteratively on the learning set L, optimizing the selected parameters. For each match, the predicted number of goals for the home and away teams is generated based on Equations 3 and 4, using the selected model parameters.

Finally, individual objective prediction errors between the observed and predicted values are calculated, and the parameters corresponding to the minimum error are selected as the final parameters. The prediction error results for different parameter groups are shown in Figure 2.

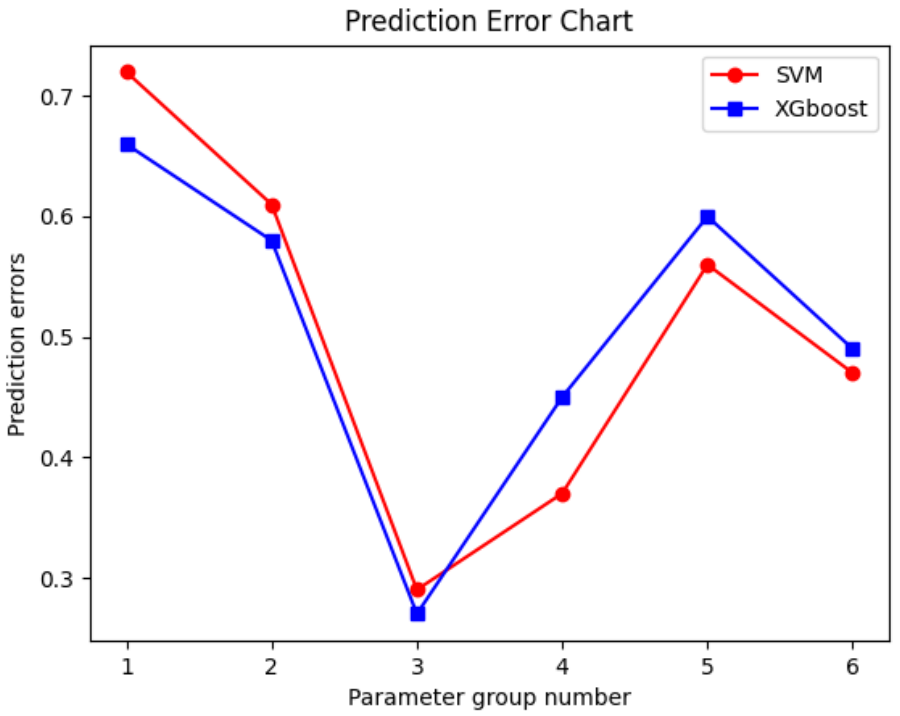

Figure 2. Prediction errors of different parameter groups. (Photo/Picture credit: Original)

Based on Figure 2, the parameter group 3, corresponding to the minimum prediction error, was ultimately selected as the optimal parameter setting, with $\beta = 2$and $\chi = 1$.

5. Experiments and results

This Section illustrates the goal prediction model using the English Premier League Round 36 match between Manchester United (home team) and Wolverhampton Wanderers (away team) that took place on May 13, 2023. The actual result of the match was a 2-0 victory for Manchester United. Based on the computed feature values for both teams in Section 3 and Section 4, the model predicted a goal count of 2.31 for Manchester United and 0.21 for Wolverhampton Wanderers. Figure 3 presents the visualization of the corresponding goal prediction model function.

DOI: 10.54254/2755-2721/21/20231136

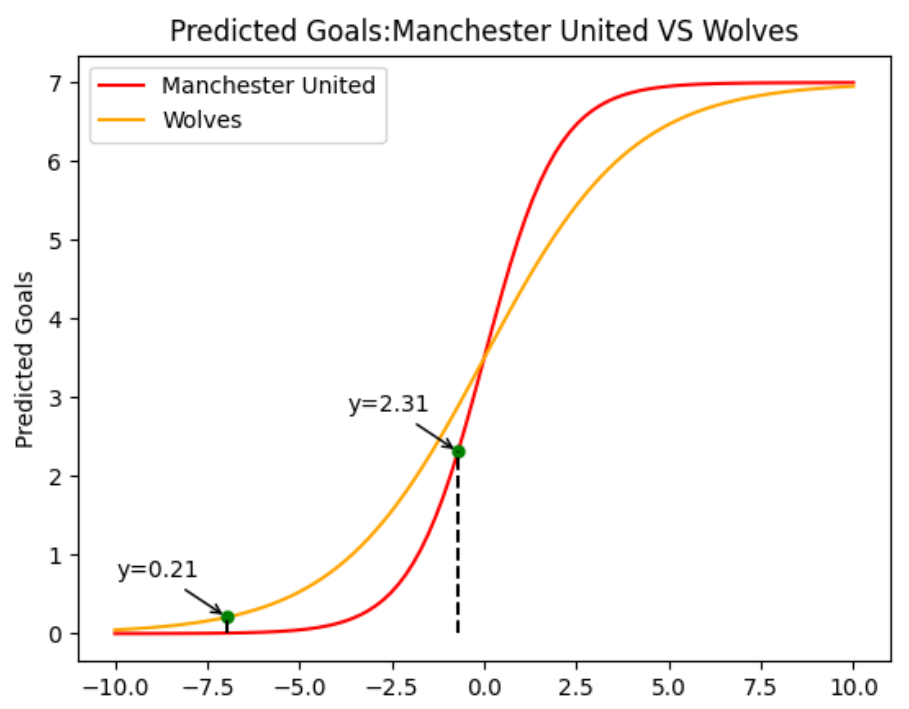

Figure 3. Predicted goals in Manchester United VS Wolves. (Photo/Picture credit: Original)

Furthermore, the experimental results demonstrate that both SVM and XGBoost perform excellently under the new feature evaluation framework, with corresponding errors less than 0.3. These findings strongly validate the effectiveness and feasibility of the proposed approach in this study. Therefore, it can be concluded that by combining domain knowledge with ML algorithms, this paper can construct models with higher accuracy and predictive capability, providing valuable support for decision-making and applications in relevant fields. Future research can further explore and optimize these methods to enhance prediction performance and expand their applicability in practical settings.

6. Conclusion

In conclusion, this paper successfully predicts soccer match results using ML techniques in Python. By collecting, analyzing, and utilizing historical match data, and incorporating domain expertise in soccer, this paper identified relevant factors for predicting soccer match outcomes and developed corresponding prediction models. The models are constructed using ML algorithms with optimization and parameter tuning to enhance prediction accuracy.

Experimental results demonstrate that the established models exhibit excellent performance in predicting soccer match results, highlighting the significant potential of ML in the field of soccer. By integrating domain knowledge with ML algorithms, this paper, successfully constructed models with high accuracy and predictive capabilities for soccer match outcome prediction, providing valuable support for decision-making and applications in relevant domains. Future research can further explore and optimize these methods to enhance prediction performance and expand their applicability in practice.

References

[1] C.Reep and B.Benjamin, “Skill and Chance in Association Football”, Journal of the Royal Statistical Society, vol 131 , pp 581-585, 1968.

[2] M.J.Maher, “Modelling association football scores”, Statistica Neerlandica, vol 36, pp 109-118,1982.

[3] H.Rue S.Martino and N.Chopin, “Approximate Bayesian inference for latent Gaussian models

DOI: 10.54254/2755-2721/21/20231136

by using integrated nested Laplace approximations", Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol 71(2), pp 319-392, 2009.

[4] P.Promvijittrakarn and T.Charoenpong, "A method of soccer-team identification by histogram feature vector and support vector machine", International Workshop on Advanced Imaging Technology, vol 12592, p 12, 2023.

[5] M.Şahin and R.Erol, "A Comparative Study of Neural Networks and ANFIS for Forecasting Attendance Rate of Soccer Games", Mathematical and Computational Applications, vol 22, p 43, 2017.

[6] A.Decuyper A.Troncoso and D.Martens, "Forecasting Association Football Match Outcomes in a Simulated Environment: A Neural Network Approach", Journal of Sports Analytics, vol 5(2), pp 85-96, 2019.

[7] Bloomfield Polman and O'Donoghue, "The 'Bloomfield Movement Classification': Motion Analysis of Individual Players in Dynamic Movement Sports", International Journal of Performance Analysis in Sport, vol 4, pp 20-31, 2004.

[8] M.J.Dixon and S.G.Coles, "Modelling Association Football Scores and Inefficiencies in the Football Betting Market", Journal of the Royal Statistical Society , vol 46(2), pp 265-280, 1997.

[9] O.Hubáček G.Šourek and F.Železný, "Learning to predict soccer results from relational data with gradient boosted trees", Springer, vol 108, pp 29-47, 2019.

[10] D.Berrar P.Lopes and W.Dubitzky, "Incorporating domain knowledge in machine learning for soccer outcome prediction", Springer, vol 108, pp 97-126, 2019.

[11] Y.Cho J.Yoon and S.Lee, "Using social network analysis and gradient boosting to develop a soccer win–lose prediction model", Engineering Applications of Artificial Intelligence, vol 72 pp 228-240,2018.

[12] H. Rue and Øyvind Salvesen, "Focus on Sport: Prediction and Retrospective Analysis of Soccer Matches in a League", Journal of the Royal Statistical Society, vol 49, pp 399-418, 2000.

[13] D.Berrar P. Lopes J.Davis and W.Dubitzky, "Guest editorial: special issue on machine learning for soccer", Springer, vol 108, pp 1-7, 2019.