International Journal of Mathematical, Engineering and Management Sciences
Vol. 10, No. 1, 249-264, 2025
https://doi.org/10.33889/IJMEMS.2025.10.1.014

Ram Arti Publishers

Intelligent Strategic Planning Method based Algorithm (ISPMA) for Estimation of Soccer Sports Match Outcome

Karun Madan
Department of Computer Science and Applications,
Panjab University, Sector 14, Chandigarh, India.
Corresponding author: karanmadan9@gmail.com

Kavita Taneja
Department of Computer Science and Applications,
Panjab University, Sector 14, Chandigarh, India.
E-mail: kavitatane@gmail.com

Harmunish Taneja
Department of Computer Science and Applications,
DAV College, Sector 10, Chandigarh, India.
E-mail: harmunish.taneja@gmail.com

(Received on March 10, 2024; Revised on May 12, 2024 & June 14, 2024 & August 4, 2024 & September 6, 2024;
Accepted on September 29, 2024)

Abstract
Estimating the soccer match outcome with adequate accuracy is still one of the biggest challenges in the sports domain. In this work, the proposed novel Intelligent Strategic Planning Method based Algorithm (ISPMA) carries out dynamic estimation of soccer team performance in terms of the match outcome and noticeably outperforms the existing state of the art methods due to its unique features. In this work, the output of the four feature selection machine learning techniques i.e. Pearson correlation, forward selection, Extra tree classifier, and CHI-square is firstly unified before feeding these selected features as an input to the seven classifiers i.e. SVM (Support Vector Machine), Naïve Bayes, KNN (K-Nearest Neighbor.), Decision Tree, Random Forest, Logistic Regression, and AdaBoost. The dataset comprises eleven seasons of the English premier league and 3762 matches have been used to train the model and 418 matches to test the same. Such a reasonable size of soccer dataset is not common in previous studies. Another unique feature of this work is the time of estimation as estimation can be done during the progression of the game based on match statistics associated with the first half of the match. The proposed method uses a novel approach by computing the average values of the selected set of features for the victory of the team to estimate match results. By using these computed average values, ISPMA generates strategic planning based suggestions for the second half of the match. The strategic planning generated by the proposed method facilitates estimating the team performance and shifting the momentum from one team to another and can assist the coach, managers, and the team in carrying out effective decision-making for better match outcome.

Keywords- Soccer match outcome, Feature selection, Strategic planning, Estimation, Team performance.

1. Introduction

For the last few years, machine learning technologies have emerged as one of the most important branches of computer science. It has been used in diverse fields. The sports domain is not an exception. Being the most watched as well as most played sport worldwide, enormous amounts of soccer data and statistics have been recorded from some major tournaments. Humans cannot process such a large set of soccer data. Therefore, the estimation of match outcome becomes one of the great examples of the AI problem. A huge amount of money is involved in sports like soccer, so the low performance of the team

leads to huge financial losses (Omomule et al., 2020). Estimation of soccer match outcome has long been the interest of soccer clubs, coaches, players, and sports organizations.

Artificial intelligence has provided many solutions in diverse fields including predictive systems (Fialho et al., 2019). Many intelligent computing techniques like artificial neural networks, Regression, KNN, Random Forest, Naïve Bayes, Decision Tree, and SVM are giving significant results in a variety of domains (Berrar et al., 2019; Joseph et al., 2022). These intelligent computing techniques are promising to address many issues and problems related to the domain of sports management i.e. game tactics, strategic planning, and decision-making (Madan et al., 2022). These predictive systems can be used in many domains like disease prediction, weather forecast prediction, insurance, marketing, stock exchange prediction, etc. (Anand and Bansal, 2016; Razali et al., 2017).

The objective behind designing predictive systems is to make better decision-making and forecasting based on patterns found in historical data (Pu et al., 2024). But in the process, when accuracy has been compromised due to any reason, it can bring about extremely costly and dangerous consequences. Improved accuracy leads to added reliability (Prasetio and Harlili, 2016). This added reliability in making clear-cut estimations brings many advantages and better decisions (Mattera, 2023). A lot of efforts are going on to improve the accuracy to make the estimations reliable (Bouaguel, 2022). However, without adequate accuracy, one cannot replace AI with humans. Several data mining and machine learning methods can be applied for the estimation of soccer match outcome (Divekar et al., 2023). As almost every sport has vast uncertainty, this accuracy is even lower in comparison to other domains.

In this work, a novel unified intelligent strategic planning method based algorithm (ISPMA) for estimating soccer match outcome has been proposed. The unified output of four feature selection techniques i.e. Extra tree, forward selection, CHI-square, and Pearson correlation is fed to seven classification models i.e. SVM, Naïve Bayes, KNN, Decision Tree, Random Forest, Logistic Regression, and AdaBoost classifier. The relationship between the final common set of selected features and the target variable is also presented. The proposed method computes the average values of the selected set of features for the victory. These average values are used to estimate the team's performance in terms of match outcome with optimum predictive accuracy. Another unique feature of this work is the time of estimation as estimation is done based on match statistics associated with the first half of the match.

The rest of the paper is ordered as follows. Section 2 presents the literature review, which is further compared in terms of various machine learning models deployed and their achieved accuracy in soccer sports. Section 3 comprises the methodology used in this work, where the combinations of feature selection and classification models are evaluated on the given data set. Section 4 presents the proposed algorithm (ISPMA). Section 5 represents the results and analysis. In Section 6, finally, the conclusion is given.

2. Review of Literature

For a long time, researchers have been trying to estimate the match results in soccer. Along with that, a huge amount of sports data is available for such analysis. Consequently, there is already some related literature available in the domain. In this section, an overview of these related studies has been presented in the context of soccer sports.

In the earlier studies, Cho et al. (2018) presented a conceptual soccer match prediction framework using gradient boosting and social networking analysis based on passing distribution data using social network analysis and gradient boosting. A Logistic Regression model was opted for predicting the soccer match

results. The dataset used for the study has been taken from the English First Division competition. The prediction accuracy of 70.1% has been reported. On the similar lines, Baboota and Kaur (2018) investigated various classification models for performing predictive analysis for the results of soccer matches. Exploratory data analysis and feature engineering have been applied for accurate prediction. Among the classification models, Gradient boosting performed best and an accuracy score of 0.59 was achieved from this model. Zaveri et al. (2018) performed their work using the last 5 seasons of the Spanish La Liga competition. An analysis of the tactics, strengths, and weaknesses was also provided. After analyzing the home and away teams, the tactics were suggested to maximize the winning chances of the team. They have used Logistic Regression and reached an accuracy of 71.63%. In their work, Esme and Kiran (2018) developed a model to forecast the soccer match outcome (home team win, a draw or away team win) of the competition matches using historical statistical data and the betting odds. The k parameter value of the k-nearest neighbor algorithm has been used between 1 and 20 throughout the work. The dataset has been taken from six seasons of the Turkey Super League from 2010-11 to 2015-16. Hervert-Escobar et al. (2018) proposed a predictive model to forecast the final results by using data from 52 soccer leagues around the world. The proposed model was two dimensional; the first was a Bayesian model based on the team rankings and the second utilizes history collected from the team's data. The ranking has been computed using a fusion of FIFA procedure and the total number of conceded goals scored and goals scored.

In the same way, Yang (2019) worked on predicting the soccer match results based on the starting lineup soccer player data. The dataset has been taken from the first 3 months of the English Premier League season 2018-19. ICT index and BPS index have been used with supervised machine learning techniques like Naïve Bayes, K-Nearest Neighbor, and Support Vector Machine to achieve 80.8% predictive accuracy. Bilek and Ulas (2019) investigated the performance indicators and factors that influence the soccer match results on the bases of current form and the quality of opponent team. The data comprises the English Premier League for the single 2017-18 season. The highest accuracy of 64.8% has been achieved using the decision tree classifier. Stübinger et al. (2020) exploited soccer data between seasons 2006-07 to 2017-18 from several major leagues including Premier League, France's Ligue1 and Ligue2, Germany's Bundesliga and Bundesliga2, Serie A and Serie B (Italy), Segunda Division, Primera Divison and Football League Championship (England). The proposed work exploited machine learning to estimate final result of the match based on player and match attributes. The maximum predictive accuracy achieved using the integrated model is 81.77%. Less than five years ago, Eryarsoy and Delen (2019) used ensemble analytics methods for forecasting the soccer match results. The last 10 years of game-level data of the Turkish Super League between 2007-2017 seasons has been applied for this work. A variety of classifiers have been employed to recognize the most prospective method. In this investigation, Random Forest outperformed the other classifiers by achieving accuracy scores of 0.7460. Knoll and Stübinger (2020) performed a simulation study on five seasons of European football leagues from 2013-14 to 2017-18. The work focused on the use of advanced betting strategies in match prediction. After applying several machine learning models, the maximum accuracy of 75.62% has been reported using random forest. Similarly, Hassan et al. (2020) provided a soccer match outcome prediction approach by using a supervised type neural network on the basis of the radial base function using seventy five attributes. The analysis of the match attribute sensitivities has been done to optimize the attribute analysis benefits.

By employing two dissimilar statistical models, Arntzen and Hvattum (2021) generated predictions in soccer matches. The first represented the before-match variation in the ELO ratings of the both competing teams, and the second was the variation in the average of the player ratings for the initial line ups of both the teams. For both the statistical models, ordered-logit regression and the competing risk modeling system have been used respectively. Wheatcroft (2021) provided the relationship between match outcome

and match statistics (predicted and observed). Strong predictors like corners, shots-on-target, and shots-off-target, etc. have also been identified. For predicting the match, BA and GAP ratings have also been used. On the other hand, Geurkink et al. (2021) intended to recognize influential predictive type variables in the Belgian Soccer Division. The work has shown a predictive accuracy of 89.6% using the Extreme Gradient Boosting algorithm. This accuracy has been obtained as a result of the additional inclusion of a broad spectrum of the variables for forecasting soccer match outcome. In the same way, Malamatinos et al. (2022) worked on data taken from the Greek Super League for predicting soccer match outcome. For better results, data cleaning, data augmentation, and feature engineering have been performed. Their experiment suggested CatBoost (CB) as the most accurate among other models by offering the highest accuracy of 67%. Similarly, Rodrigues and Pinto (2022) used 4 seasons of English Premier League data for the training purpose and one season (2016/17) for the testing to predict the soccer match outcome. The proposed model predicted the final match outcome with 65.26% accuracy. The profit margin of football bets obtained was also reasonably higher.

Among the recent research, Aarons et al. (2023) employed Generalized Linear Models, AdaBoost, and Random Forest to predict the match outcome with a maximum of 92.7% accuracy. The dataset of more than 800 matches of Australian Football League from 5 seasons between 2017 and 2021 has been utilized. Data informed type feature sets showed better results than the data driven feature sets in majority of the cases. Parker (2023) employed four classifiers i.e. Linear Regression, Support Vector Machine, Naïve Bayes, and Decision Tree on data comprised of 2612 matches of the Mexican league from 2012 to 2020 seasons. The proposed model predicted the match outcome on the bases of total goals scored by the soccer team with the highest accuracy of 75% (draws included) and 84% (draws excluded). Yeung et al. (2023) worked on an interpretable and generalized machine learning framework that required decisions of the coach and quality features of the players for the prediction of match outcome. Data of over 1700 matches from 5 seasons of English Premier League have been used. Divekar et al. (2023) used multinomial probit regression to predict the impact of covariates for forecasting the match outcome for English Premier League using Bayesian method. The model offered the effect of different events like crosses, cards, and corners on the final match outcome. On the similar lines, Kinalioglu and Kus (2023) proposed hybrid classification methods to predict soccer matches. The unified method used clustering and classification algorithms together on data taken from European leagues. The prediction results showed the highest accuracy of 81.76%. The dataset also included fans' opinions collected from various social media platforms along with general statistical data. Holmes and McHale (2024) proposed a model for estimating the final result of football matches. The proposed system is a player based rating framework, which adjusts the player match-day ratings. The work introduced a multinomial model to reveal how a particular soccer player will be playing against a specific opponent. Similarly, Chandra et al. (2024) proposed highly influential attributes and outlined the strategic planning for calculating the new parameters derived from these attributes. The main objective of proposed study is to build a model for forecasting the performance in terms of assists made and goals scored, etc. to enhance the team dynamics and final match results.

Researchers have been attempting different methodologies for estimation of team performance in terms of match outcome. So, there is a need of a strategic planning method, which can estimate the team performance in terms of match outcome with adequate accuracy and reliability.

Related work has been further compared in terms of the achieved accuracy, the approach used and the technique used as shown in **Table 1**.

Table 1. Comparative study of related work based on machine learning for the prediction of soccer match outcome.

S. No.	Reviewed study	Approach used	Machine learning models	Highest accuracy
1.	Esme and Kiran (2018)	The proposed prediction model used a k-nearest neighbor to measure the similarity between various competitions and events based on the betting odds. A risk analysis option has also been added to the model to minimize the margin of error in game prediction.	k-nearest neighbor	57%
2.	Baboota and Kaur (2019)	The work employed exploratory data analysis along with feature engineering to estimate the match results of English premier league. The major constraint is the limited amount of data.	Gradient Boosting	59.0%
3.	Hervert-Escobar et al. (2018)	A procedure has been proposed for forecasting the soccer match results of 2018 FIFA World Cup using a hybrid Bayesian network based on the rank position.	Hybrid Bayesian Networks	69%
4.	Cho et al. (2018)	Presented the conceptual framework win-lose prediction system using UEFA Champions league data.	Gradient Boosting	70.1%
5.	Zaveri et al. (2018)	The proposed work predicted the match results of Spanish La Liga using the FIFA 18 game dataset by providing an analysis of the strengths and weaknesses of teams.	Logistic Regression	71.63%
6.	Yang (2019)	The BPS and ICT index have been covered by employing supervised machine learning techniques. Finally, Random Forest has been found as the best classifier in the given project.	Random Forest	80.8%
7.	Bilek and Ulas (2019)	Predicted match outcome by making use of team performance indicators and situational variables. Prediction is done in line with the superiority of opponent in famous English premier league.	Decision Tree	64.8%
8.	Stübinger et al. (2020)	The baseline of the simulated study is various features and skills on the player-level taken from five major leagues across Europe. The work suggested that the strategies that depend on tree based machine learning methods do better than the linear approach based strategies.	Integrated Model	81.77%
9.	Eryarsoy and Delen (2019)	Predicted match results based on win/loss/draw and points/no points by using ensemble methods on the dataset of the Turkish Super League.	Random Forest	74.60%
10.	Knoll and Stübinger (2020)	Predicted outcome of matches of five top European football leagues by exploiting match characteristics and also used statistical arbitrage stock market.	Random Forest	75.62%
11.	Geurkink et al. (2021)	The work intended to identify and use the most influential 100predictive variables for victory and defeat in eminent Belgian soccer division matches. Variance Inflation Factor and BoraShap have been applied to avoid the multicollinearity problem.	Extreme Gradient Boosting	89.6%
12.	Malamatinos et al. (2022)	Investigated matches from the Greek Super League, English Premier League, and Dutch Eredivisie. Transfer learning is tested by encoding images from tabular data.	CatBoost	67.73%
13.	Kinalioglu and Kus (2023)	The proposed method for predicting soccer match outcome uses statistical information along with the opinions of fans collected from social media platforms.	Classification and Clustering Hybrid methods	81.76%
14.	Parker (2023)	The Model predicted the failure or success of the football team based on the goal-scoring performance. The proposed work has been split into two sections: one including draws and the other excluding the draws.	Linear Regression Support Vector Machine Naïve Bayes Decision Tree	84%
15.	Aarons et al. (2023)	Score-agnostic performance indicator based machine learning models have been used to forecast the match outcome for the decision support system for the coaches.	Generalized Linear Models AdaBoost Random Forest	92.7%

Several data mining and machine learning methods can be applied for estimation of soccer match outcome. However, creating a system with adequate accuracy is still one of the biggest challenges in the context of soccer sport. Most of the previous work could not achieve desired accuracy. The thorough analysis of the previous work indicates the inadequate dataset size, improper selection of features and time of the estimation could be the reason in achieving inadequate accuracy.

3. Methodology

For this work, the dataset contains 4180 matches taken from eleven seasons of the prestigious English Premier League from 2012 to 2022 (Football-Data.co.uk, 2022). The dataset is publicly available at http://www.football-data.co.uk/englandm.php. Firstly, data cleaning and feature encoding have been done as the data preparation part. The dataset has been split into 90% train data and 10% test data. The model has been trained on 3762 matches and 418 matches are used to test the model. The dataset includes a total of 32 features as shown in **Table 2**.

Table 2. Full set of features.

S. No.	Feature name	Description of features	Type
1.	*Div*	League Division Season	Object
2.	*Date*	Match Date	Object
3.	*HomeTeam*	Home Team	Object
4.	*AwayTeam*	Away Team	Object
5.	*Referee*	Match Referee	Object
6.	*FTHG*	No. of Home Team Goals after Full-Time	Int64
7.	*FTAG*	No. of Away Team Goals after Full-Time	Int64
8.	*HTHG*	No. of Home Team Goals after Half-Time	Int64
9.	*HTAG*	No. of Away Team Goals after Half-Time	Int64
10.	*HS*	Total no. of Shots by Home Team	Int64
11.	*AS*	Total no. of Shots by Away Team	Int64
12.	*HST*	Total no. of Shots on Target by Home Team	Int64
13.	*AST*	Total no. of Shots on Target by Away Team	Int64
14.	*HF*	Total no. of Fouls Committed by Home Team	Int64
15	*AF*	Total no. of Fouls Committed by Away Team	Int64
16.	*HC*	Total no. of Corners by Home Team	Int64
17.	*AC*	Total no. of Corners by Away Team	Int64
18.	*HY*	No. of Yellow Cards by Home Team	Int64
19.	*AY*	No. of Yellow Cards by Away Team	Int64
20.	*HR*	No. of Red Cards by Home Team	Int64
21.	*AR*	No. of Red Cards by Away Team	Int64
22.	*B365H*	Bet365 win odds of the Home Team	Float64
23.	*B365D*	Bet365 Draw odds	Float64
24.	*B365A*	Bet365 win odds of Away Team	Float64
25.	*WHH*	William Hill win odds of Home Team	Float64
26.	*WHD*	William Hill draw odds	Float64
27.	*WHA*	William Hill win odds of Away Team	Float64
28.	*VCH*	VC Bet home win odds	Float64
29.	*VCD*	VC Bet draw odds	Float64
30.	*VCA*	VC Bet away win odds	Float64
31.	*HTR*	Half Time Results (A=Away Win, D=Draw, H=Home Win)	Object
32.	*FTR*	Full Time Results (A=Away Win, D=Draw, H=Home Win)	Object
33.	*HTR_num*	Hal Time Results (Categorical to Numeric data conversion)	Int64
34.	*FTR_num*	Full-Time Results (Categorical to Numeric data conversion)	Int64

The output of four machine learning based feature selection techniques employed in this study is unified before feeding these selected features as input to seven classifiers i.e. Naïve Bayes, KNN, Logistic regression, SVM, Random Forest, Decision Tree, and AdaBoost classifier for estimating team performance in terms of match outcome. The six selected features are *HST*, *AST*, *HTHG*, *HTAG*, *HS* and

AS. *HST* and *AST* denote shots-on-target in total by the host team and visiting team respectively. *HTHG* and *HTAG* denote the goal scores of the host team till the half-time interval and the goal scores of the away team till the half-time interval respectively. *HS* and *AS* are the shots made in total by host and visiting teams respectively.

The four machine learning based feature selection techniques used in this study i.e. Pearson correlation, forward selection, Extra tree, and CHI-square produced varied results as shown in **Table 3**. The output of all the feature selection techniques is unified before feeding these selected features as input to seven classifiers i.e. Naïve Bayes, KNN, Logistic Regression, SVM, Random Forest, Decision Tree, and AdaBoost classifier for estimating team performance in terms of match outcome.

Accuracy scores of all seven models have been recorded without using any feature selection and after using the unified feature selection.

Table 3. Feature selection techniques used in the study and their respective selected features.

Feature selection techniques used	**Selected set of features (top 10)**
Extra tree classifier	*HTHG, HTAG, HS, AS, HST, AST, HF, AF, HC, AC*
Forward selection	*HTHG, HTAG, HS, AS, HST, AST, HF, AF, HC, AC*
CHI-square	*HTHG, HTAG, HS, AS, HST, AST, B365H, B365A, WHH, WHA*
Pearson correlation	*HTHG, HTAG, HS, AS, HST, AST, B365H, B365A, B365D, VCH*

4. Intelligent Strategic Planning Method based Algorithm (ISPMA) for Estimation of Match Outcome

The proposed ISPMA discovers the relationship between the final common set of selected features and the target variable. Based on team performance in the first half of the match, the proposed algorithm computes the average values of the selected set of features for the victory of the team and generates strategic planning suggestions based on these average values.

***Algorithm* 1:** Intelligent Strategic Planning Method based Algorithm (ISPMA)

Input_Match_Data (Representing the team performance during the first half and second half of the match in the form of shots, fowls, cards, goal scores, and related information).
Data_Preprocessing
Data Cleaning
Feature Encoding
Deploy_Feature_Selection_techniques
Multiplexing_Feature_Selection_Output
Outputs from four Feature selection techniques multiplexed and routed as a single output
Selected_Set_of_Features
Set *HST* = Shots-on-target in total by the host team
Set *AST* = Shots-on-target in total by the away team
Set *HTHG* = Goal scores of the host team till the half-time interval
Set *HTAG* = Goal scores of the away team till the half-time interval
Set *HS* = Shots made in total by the host team
Set *AS* = Shots made in total by the away team
Model_Deployment_using_Train_test_split
Train_data =90%
Test_data = 10%
Deploy_Classification_models
Investigate_Combination_of_Intelligent_techniques
Investigate various combinations of Classifiers and feature selection techniques
Record Accuracies
Select the optimum combination
Determine_Average_Values_for_Victory

Find the relationship among the selected set of features and target variables
Get average values of the selected set of features for the victory of the team after the first half of the match
Calculation_of_Accuracy_Scores
Generate_Estimation_of_Match_Results
Estimation of match results are generated using average values of the selected features after the first half of the match
Possible_Match_Outcome
Category1 - Home team Victory
Category2 - Away team Victory
Category3 – Draw
Strategic_Planning_Decision_Making
If (*Average_Value_HST* **≥6): *Estimated_Result*=Home_team_Victory**
Strategic_Plan∈{Defensive_Play}
Else if (*Average_Value_HST*≤**4): *Estimated_Result*=Away_team_Victory**
Strategic_Plan∈{Attacking_Play}
Else**: *Estimated_Result*= Draw**
Strategic_Plan∈{Moderate_Play}

if (*Average_Value_AST*≥**5): *Estimated_Result*=Away_team_Victory**
Strategic_Plan∈{Attacking_Play}
Else if (*Average_Value_AST*≤**3): *Estimated_Result*=Home_team_Victory**
Strategic_Plan∈{Defensive_Play}
Else**: *Estimated_Result*= Draw**
Strategic_Plan∈{Moderate_Play}

If (*Average_Value_HTHG* ≥1):***Estimated_Result*= Home_team_Victory**
Strategic_Plan∈{Defensive_Play}
Else
***Estimated_Result* = {Draw, Away_team_Victory}**
Strategic_Plan∈{Moderate_Play, Attacking_Play }

if (*Average_Value_HTAG* **≥1): *Estimated_Result*=Away_team_Victory**
Strategic_Plan∈{Attacking_Play }
Else
***Estimated_Result* = {Draw, Home_team_Victory}**
Strategic_Plan∈{Moderate_Play, Defensive_Play}

If (*Average_Value_HS*≥**15): *Estimated_Result*=Home_team_Victory**
Strategic_Plan∈{Defensive_Play }
Else if (*Average_Value_HS*≤**12): *Estimated_Result*=Away_team_Victory**
Strategic_Plan∈{Attacking_Play }
Else**: *Estimated_Result*= Draw**
Strategic_Plan∈{Moderate_Play }

If (*Average_Value_AS* **≥13): *Estimated_Result*=Away_team_Victory**
Strategic_Plan∈{Defensive_Play }
Else if (*Average_Value_AS*≤**10): *Estimated_Result*=Home_team_Victory**
Strategic_Plan∈{Attacking_Play }
Else**: *Estimated_Result*= Draw**
Strategic_Plan∈{Moderate_Play}

The proposed Intelligent Strategic Planning Method based Algorithm (ISPMA) for estimation of team performance in terms of match outcome discovers the relationship between the target variable and the final set of selected features. The box plots have been drawn using the matplotlib library to exhibit the relation between *FTR* (Full-Time match Result) and all the six selected features (**Figures 1** to **6**). The proposed unified model first computes the average values of six selected features for the victory of the team and generates strategic planning suggestions based on these average values.

Box plots are graphic representations specifically used for data analysis. The dots represent the outliers. The horizontal lines below and above the boxes are the maximum as well as minimum values of the features in the given dataset. Box boundary lines represent 25% and 50% of the dataset values. The

median is the mid value of the dataset and is shown by the line that divides the box into two parts. The proposed algorithm (ISPMA) focuses on these median values termed as average values.

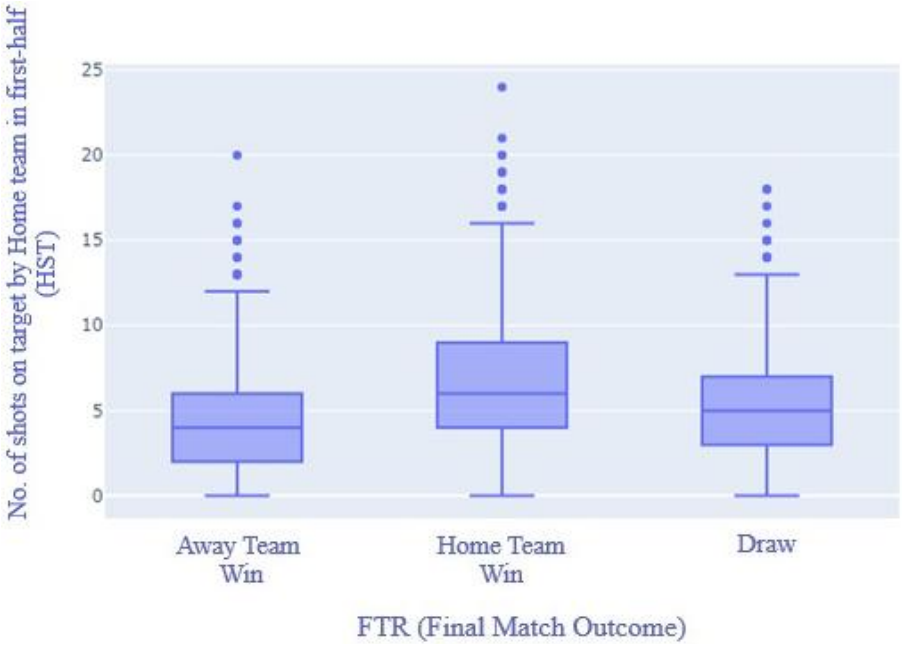

Figure 1. Computed average values of *HST* in relation to *FTR*.

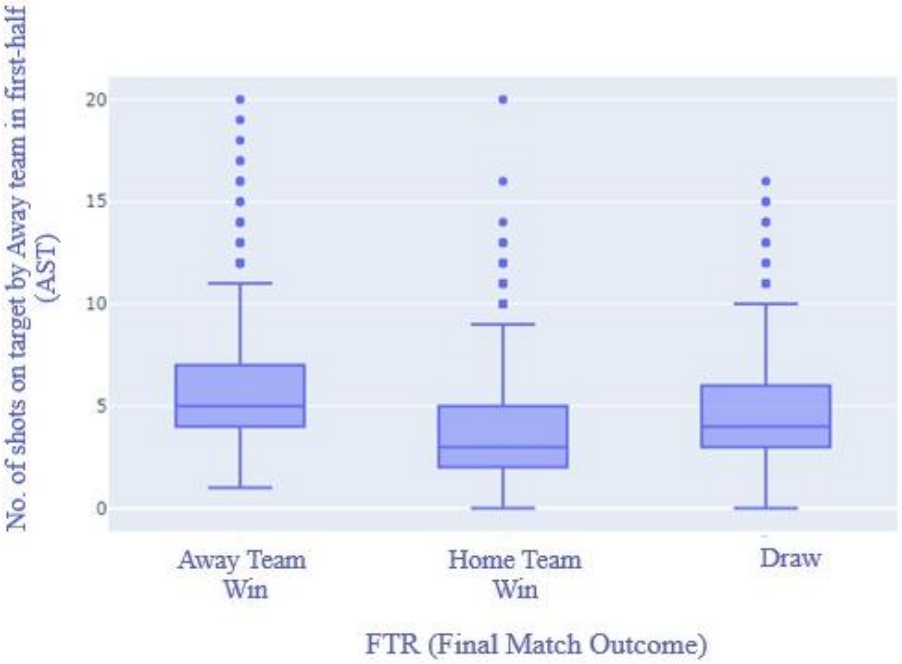

Figure 2. Computed average values of *AST* in relation to *FTR*.

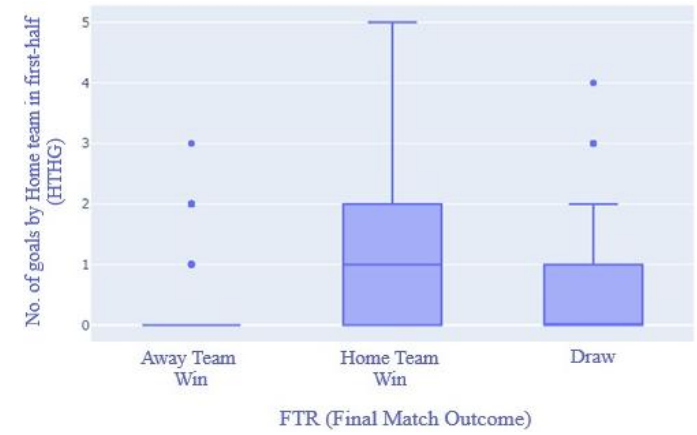

Figure 3. Computed average values of *HTHG* in relation to *FTR*.

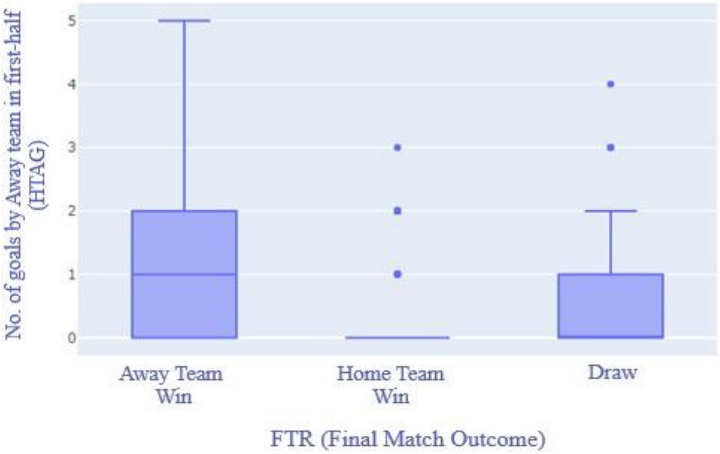

Figure 4. Computed average values of *HTAG* in relation to *FTR*.

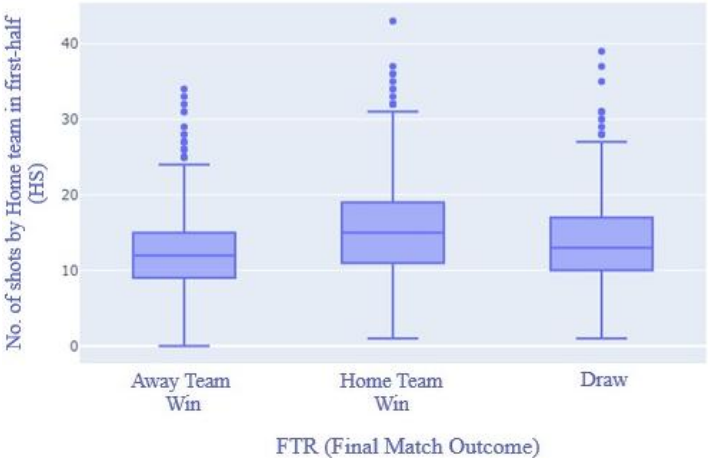

Figure 5. Computed average values of *HS* in relation to *FTR*.

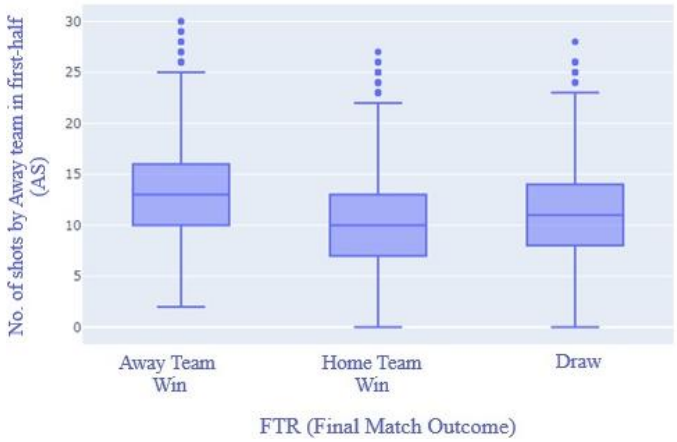

Figure 6. Computed average values of *AS* in relation to *FTR*.

5. Results and Discussion

The proposed Intelligent Strategic Planning Method based Algorithm (ISPMA) estimates the team performance in terms of match outcome with adequate accuracy. After applying four machine learning based feature selection techniques i.e. Extra tree, forward selection, chi-square, and Pearson correlation, the output is unified and six selected features have been obtained.

The achieved accuracy scores by seven classifiers have been recorded before applying any feature selection technique. After applying the above four feature selection techniques in unification, accuracy scores achieved by all seven classifiers have again been recorded. Various machine learning based combinations of classifiers and feature selection techniques have been investigated and compared on the given dataset. By applying the selected set of six features on all the seven classifiers, all the classifiers have shown considerable enhancement in accuracy. While comparing all the seven classification techniques in terms of achieved accuracy before and after applying feature selection, it has been observed that Random Forest gave much promising predictive accuracy than other six classifiers i.e. SVM, Decision Tree, Logistic Regression, Naïve Bayes, KNN and Adaboost (**Table 4**). In the case of SVM, the difference between the accuracy scores recorded before and after applying feature selection is approximately 5%. This difference is nearly 1% in the case of Decision Tree and Random Forest. In logistic regression, there is an enhanced accuracy by nearly 3% after the unified feature selection. Whereas, in Naïve Bayes, the accuracy increased by more than 5%. KNN showed the most improvement in accuracy among all the cases, as after applying feature selection the accuracy increased by approximately 8%. Similarly, AdaBoost has shown an increase in accuracy by approximately 5%. The proposed method provides an estimation of match results with three possibilities i.e. victory for the away team, home team, or draw.

Table 4. Comparison of accuracy scores achieved before and after applying feature selection techniques.

Classifier	Accuracy scores without feature selection	Accuracy scores with unified feature selection	Percentage increase in accuracy
SVM	0.6112	0.6603	4.91%
DECISION TREE	0.5490	0.5550	0.60%
RANDOM FOREST	0.9841	0.9880	0.39%
LOGISTIC REGRESSION	0.6435	0.6770	3.35%
NAÏVE BAYES	0.6196	0.6710	5.14%
KNN	0.5191	0.5980	7.89%
ADABOOST	0.6244	0.6710	4.66%

The relationship among these common sets of chosen features and the target variable has been described in the form of the box plots generated using the matplotlib library. The proposed method computed the average values of the selected set of features for the victory of the team.

The proposed algorithm (ISPMA) computed the average values of *HST* (Shots-on-target in total by the host team) in relation to *FTR* (Full-Time final Result), which indicates if shots-on-target by the home team are 6 or more, then there is a strong prediction of the victory of the home team. If shots-on-target by the home team are 5, then there will be chances of a draw. If this value is 4 or less, it means there are strong chances of away team victory. In the same way, the average values of *AST* (Shots-on-target in total by the away team) indicate if shots-on-target by the away team are 3 or less, then there is a strong prediction of the victory of the home team. If shots-on-target by the away team are 4, then there will be chances of a draw. If this value is 5 or more, it means there are strong chances of away team victory.

The average value of *HTHG* (Home Team goals after Half Time) in relation to *FTR* indicates if the home team goal score on half time is 1 or more, there is a strong prediction of the victory of the home team. If the home team goal score after half-time is 0, it means there are equal chances of either an away team victory or a draw. Similarly, the average values of *HTAG* (half-Time Away team Goals) indicate if the away team goal score on half-time is 1 or more, then there is a strong prediction of the victory of the away team. If the away team goal score after half-time is 0, it means there are equal chances of either a home team victory or a draw.

In the same way, the average value of *HS* (Shots by the home team) in relation to *FTR* indicates if shots by the home team are 15 or more, there is a strong prediction of the victory of the home team. If shots by the home team are 13, then there will be chances of a draw. If this value is 12 or less, it means there are strong chances of away team victory. Similarly, the average value of *AS* (Shots by the home team) indicates if shots by the away team are 10 or less, then there is a strong prediction of the victory of the home team. If shots by the away team are 11, then there will be strong chances of a draw. If this value is 13 or more, it means there are strong chances of the away team victory.

The ISPMA generates strategic planning based suggestions to assist in decision making for the second half of the match. For instance, teams can try to shift the momentum by modifying their approach to little aggressive or defensive based on the suggestions generated by ISPMA.

Everyone is not privileged to have expert or coach due to geographical, economical or commercial factors. This research proposes ISPMA, which can help the team and players to carry out reliable decision-making for better match outcome by assessing the performance of opponents as well as their team, and to make the essential adjustments in their strategy.

In this work, the output of four machine learning based feature selection techniques i.e. Extra tree classifier, forward selection, CHI-square, and Pearson correlation is firstly unified before feeding these selected features as input to the classifier. The reason behind choosing random forest is the unmatched best results obtained by random forest as compared to the other six classifiers deployed. The dataset comprises eleven seasons of the English premier league and 3762 matches have been used to train the model and 418 matches to test the same. In sports like soccer, such a huge dataset for test and train-set is not common in previous studies. Existing research has not adequately explored the timing of the match prediction. Unlike most of the previous research on the estimation of match outcome, one of the unique features of our work is the time of estimation. Dynamically estimating the match outcome based on match statistics associated with the first half of the match appears unconventional but effective approach.

Dynamic estimation can provide better results as estimation can be done during the progression of the game, not at the end or start of the match. Previous research focused only on one or two dimensions while forecasting the match results. The proposed method uses a novel approach by computing the average values of the selected set of features for the victory of the team to estimate match results with 98.80% accuracy. This multi-dimensional approach allowed computed average values of all the features to participate in estimation. As a result of these factors, the proposed method noticeably outperforms the existing state of the art methods in terms of accuracy achieved for the estimation of soccer match outcome (**Figure 7**).

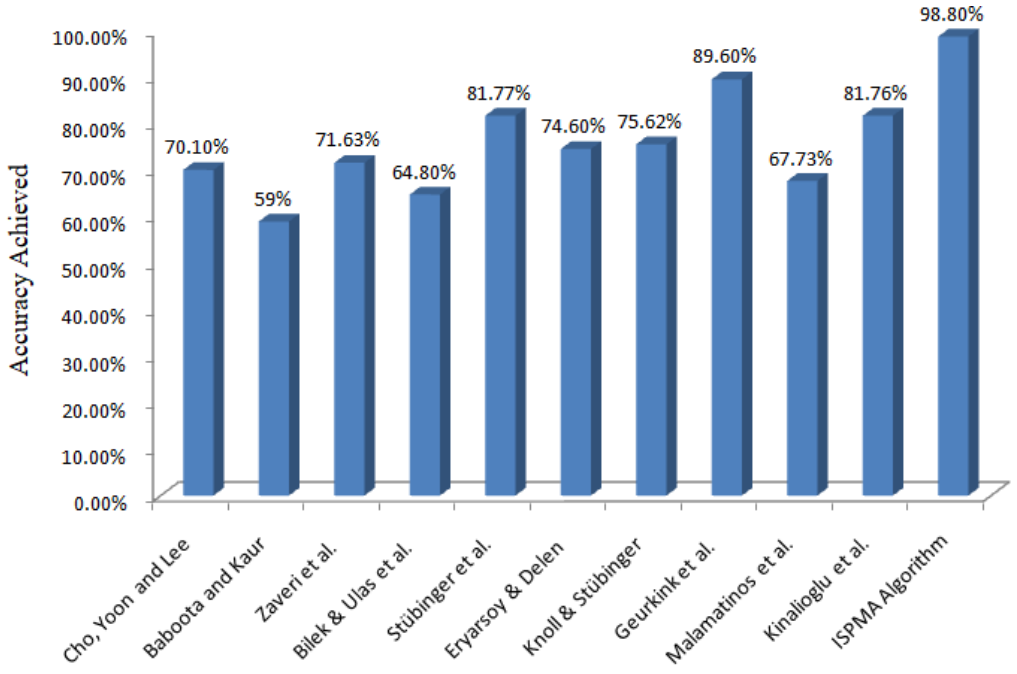

Figure 7. Comparison of proposed intelligent strategic planning method based algorithm (ISPMA) with existing state of the art methods in terms of accuracy.

6. Conclusion

Estimating the outcome of soccer matches has long been a major concern for coaches and sports organizations. The major bottleneck is the low accuracy for such models to be reliable for prediction. In this work, a novel intelligent strategic planning method based algorithm (ISPMA) has been proposed for dynamic estimation of soccer team performance in terms of match outcome. For this work, the dataset comprises eleven seasons of the English Premier League. Initially, data preparation was done on a total of 4180 soccer matches in the form of data cleaning and feature encoding. 3762 matches have been used to train the model and 418 matches to test the same.

In this work, seven classifiers i.e. SVM, Naïve Bayes, KNN, Decision Tree, Random Forest, Logistic Regression, and AdaBoost classifier have been evaluated along with diverse machine learning based feature selection techniques like Extra tree classifier, forward selection, CHI-square, and Pearson correlation to obtain the optimum predictive accuracy for soccer match outcome.

The output of the four machine learning based feature selection techniques is firstly unified before feeding these selected features as an input to the random forest classifier. The reason behind choosing

random forest is the unmatched best results obtained by random forest as compared to the other six classifiers deployed. ISPMA discovers the relationship between the final common set of selected features and the target variable. These relationships are represented in the form of box plots generated using the matplotlib library. Another unique feature of this work is the time of estimation as estimation can be done during the progression of the game, not at the end or start of the match. By using match statistics from the first half of the match, the proposed unified method computes the average values of the selected set of features for the victory of the team to estimate results with three possibilities i.e. win, lose, or draw with 98.80% accuracy. After computing these average values, ISPMA generates strategic planning based suggestions for the second half of the match. The proposed method outperforms the existing state of the art methods in terms of accuracy achieved due to its distinct features.

The ISPMA can assist in decision making as by applying these strategically planned suggestions, momentum can be shifted from one team to another as the trailing team tries to push harder by modifying their approach to a little aggressive and the leading team tries to maintain the lead and momentum by modifying their approach between moderate and defensive. By these average values, teams can assess the performance of opponents as well as their team, make the essential adjustments in player positions, and fine-tune their strategic planning. The proposed intelligent strategic planning method based algorithm (ISPMA) gave promising results, which can help coaches, managers, and the team to carry out reliable decision-making in the match.

Conflict of Interest

The authors confirm that there is no conflict of interest to declare for this publication.

Acknowledgments

The research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The authors also sincerely appreciated the editor and reviewers for their time and valuable comments.

References

Aarons, M.F., Young, C.M., Bruce, L., & Dwyer, D.B. (2023). Real time prediction of match outcomes in Australian football. *Journal of Sports Sciences*, *41*(11), 1115-1125. https://doi.org/10.1080/02640414.2023.2259266.

Anand, A., & Bansal, G. (2016). Predicting customer's satisfaction (dissatisfaction) using logistic regression. *International Journal of Mathematical, Engineering and Management Sciences*, *1*(2), 77-88. https://dx.doi.org/10.33889/ijmems.2016.1.2-009.

Arntzen, H., & Hvattum, L.M. (2021). Predicting match outcomes in association football using team ratings and player ratings. *Statistical Modelling*, *21*(5), 449-470. https://doi.org/10.1177/1471082x20929881.

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.https://doi.org/10.1016/j.ijforecast.2018.01.003.

Berrar, D., Lopes, P., & Dubitzky, W. (2019). Incorporating domain knowledge in machine learning for soccer outcome prediction. *Machine learning*, *108*(1), 97-126. https://doi.org/10.1007/s10994-018-5747-8.

Bilek, G., & Ulas, E. (2019). Predicting match outcome according to the quality of opponent in the English premier league using situational variables and team performance indicators. *International Journal of Performance Analysis in Sport*, *19*(6), 930-941. https://doi.org/10.1080/24748668.2019.1684773.

Bouaguel, W. (2022). Efficient multi-classifier wrapper feature-selection model. Application for dimension reduction in credit scoring. *Computer Science*, *23*(1), 133-155. https://doi.org/10.7494/csci.2022.23.1.4120.

Chandra, B., Shinny, D.J., & Adhitya, M.K. (2024). Prediction of football player performance using machine learning algorithm. *Research Square.* https://doi.org/10.21203/rs.3.rs-3995768/v1.

Cho, Y., Yoon, J., & Lee, S. (2018). Using social network analysis and gradient boosting to develop a soccer win-lose prediction model. *Engineering Applications of Artificial Intelligence*, *72*, 228-240. https://doi.org/10.1016/j.engappai.2018.04.010.

Divekar, C., Deb, S., & Roy, R. (2023). Real-time forecasting within soccer matches through a Bayesian lens. *arXiv preprint arXiv:2303.12401*. https://doi.org/10.48550/arXiv.2303.12401.

Eryarsoy, E., & Delen, D. (2019). Predicting the outcome of a football game: A comparative analysis of single and ensemble analytics methods. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*. (pp.1107-1115). http://hdl.handle.net/10125/59550. Gamification Group, Hawaii.

Esme, E., & Kiran, M.S. (2018). Prediction of football match outcomes based on bookmaker odds by using k-nearest neighbor algorithm. *International Journal of Machine Learning and Computing*, *8*(1), 26-32. https://doi.org/10.18178/ijmlc.2018.8.1.658.

Fialho, G., Manhães, A., & Teixeira, J.P. (2019). Predicting sports results with artificial intelligence–a proposal framework for soccer games. *Procedia Computer Science*, *164*, 131-136. https://doi.org/10.1016/j.procs.2019.12.164.

Football-Data.co.uk. [database file]. (2022) Premium League. UpToDate.RetrievedDecember 26, 2022, fromhttp://www.football-data.co.uk/englandm.php.

Geurkink, Y., Boone, J., Verstockt, S., & Bourgois, J.G. (2021). Machine learning-based identification of the strongest predictive variables of winning and losing in Belgian professional soccer. *Applied Sciences*, *11*(5), 2378. https://doi.org/10.3390/app11052378.

Hassan, A., Akl, A.R., Hassan, I., & Sunderland, C. (2020). Predicting wins, losses and attributes' sensitivities in the soccer world cup 2018 using neural network analysis. *Sensors*, *20*(11), 3213. https://doi.org/10.3390/s20113213.

Hervert-Escobar, L., Matis, T.I., & Hernandez-Gress, N. (2018). Prediction learning model for soccer matches outcomes. In *2018 Seventeenth Mexican International Conference on Artificial Intelligence* (pp. 63-69). IEEE. Guadalajara, Mexico. https://doi.org/10.1109/micai46078.2018.00018.

Holmes, B., & McHale, I.G. (2024). Forecasting football match results using a player rating based model. *International Journal of Forecasting*, *40*(1), 302-312. https://doi.org/10.1016/j.ijforecast.2023.03.002.

Joseph, J., Senith, S., Kirubaraj, A.A., & Ramson, S.R.J. (2022). Machine learning for prediction of clinical appointment no-shows. *International Journal of Mathematical, Engineering and Management Sciences*, *7*(4), 558-574. https://doi.org/10.33889/ijmems.2022.7.4.036.

Kinalioğlu, İ.H., & Kuş, C. (2023). Prediction of football match results by using artificial intelligence-based methods and proposal of hybrid methods. *International Journal of Nonlinear Analysis and Applications*, *14*(1), 2939-2969. https://doi.org/10.22075/ijnaa.2023.26802.3413.

Knoll, J., &Stübinger, J. (2020). Machine-learning-based statistical arbitrage football betting. *KI-KünstlicheIntelligenz*, *34*(1), 69-80. https://doi.org/10.1007/s13218-019-00610-4.

Madan, K., Taneja, K., & Taneja, H. (2022). Intelligent computing based soccer sports management for effective estimation of match outcome. In *2022 International Conference on Decision Aid Sciences and Applications* (pp. 660-664). IEEE. Chiangrai, Thailand. https://doi.org/10.1109/dasa54658.2022.9765291.

Malamatinos, M.C., Vrochidou, E., & Papakostas, G.A. (2022). On predicting soccer outcomes in the greek league using machine learning. *Computers*, *11*(9), 133. https://doi.org/10.3390/computers11090133.

Mattera, R. (2023). Forecasting binary outcomes in soccer. *Annals of Operations Research*, *325*(1), 115-134. https://doi.org/10.1007/s10479-021-04224-8.

Omomule, T.G., Ibinuolapo, A.J., & Ajayi, O.O. (2020). Fuzzy-based model for predicting football match results. *International Journal of Scientific Research in Computer Science and Engineering*, *8*(1), 70-80.

Parker, E. (2023). A predictive framework for forecasting soccer match outcomes by analyzing the goal count achieved by a specific team. *Infotech Journal Scientific and Academic*, *4*(2), 35-47. https://infotechjournal.org/index.php/infotech/article/view/30.

Prasetio, D., & Harlili, D. (2016). Predicting football match results with logistic regression. In *2016 International Conference on Advanced Informatics: Concepts, Theory and Application* (pp. 1-5). IEEE. Penang, Malaysia. https://doi.org/10.1109/icaicta.2016.7803111.

Pu, Z., Pan, Y., Wang, S., Liu, B., Chen, M., Ma, H., & Cui, Y. (2024). Orientation and decision-making for soccer based on sports analytics and AI: A systematic review. *IEEE/CAA Journal of Automatica Sinica*, *11*(1), 37-57. https://doi.org/10.1109/jas.2023.123807.

Razali, N., Mustapha, A., Yatim, F.A., & Ab Aziz, R. (2017). Predicting football matches results using Bayesian networks for English Premier League (EPL). In *IOP Conference Series: Materials Science and Engineering*, *226*(2017), 012099. https://doi.org/10.1088/1757-899x/226/1/012099.

Rodrigues, F., & Pinto, Â. (2022). Prediction of football match results with machine learning. *Procedia Computer Science*, *204*, 463-470. https://doi.org/10.1016/j.procs.2022.08.057.

Stübinger, J., Mangold, B., & Knoll, J. (2020). Machine learning in football betting: prediction of match results based on player characteristics. *Applied Sciences*, *10*(1), 46. https://doi.org/10.3390/app10010046.

Wheatcroft, E. (2021). Forecasting football matches by predicting match statistics. *Journal of Sports Analytics*, *7*(2), 77-97. https://doi.org/10.3233/jsa-200462.

Yang, R. (2019). Using supervised learning to predict English Premier League match results from starting line-up player data. *M.Sc. in Computing (Data Analytics). Technological University Dublin.* https://arrow.tudublin.ie/scschcomdis/177/.

Yeung, C.C.K., 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. https://doi.org/10.1371/journal.pone.0284318.

Zaveri, N., Shah, U., Tiwari, S., Shinde, P., & Teli, L.K. (2018). Prediction of football match score and decision making process. *International Journal on Recent and Innovation Trends in Computing and Communication*, *6*(2), 162-165. https://www.academia.edu/download/56704483/1520497925_08-03-2018.pdf.