Time Series Approaches to Predict Soccer Match Outcome

MSc Research Project
Data Analytics

Lithin Dominic Joseph
Student ID: 20187963

School of Computing
National College of Ireland

Supervisor: Dr. Christian Horn

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Lithin Dominic Joseph
Student ID:	20187963
Programme:	Data Analytics
Year:	2021
Module:	MSc Research Project
Supervisor:	Dr. Christian Horn
Submission Due Date:	31 / 01 / 2022
Project Title:	Time Series Approaches to Predict Soccer Match Outcome
Word Count:	5969
Page Count:	18

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Lithin Dominic Joseph
Date:	30th January 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Time Series Approaches to Predict Soccer Match Outcome

Lithin Dominic Joseph
20187963

Abstract

EPL is one of the most viewed and fan followed series in soccer history. Since the sports gambling business and artificial intelligence have evolved to such an extent, forecasting the outcome of a sporting event using a computerized system has become extremely relevant. Numerous data analysts worked on predicting EPL games' results in recent years. Nevertheless, all the approaches are almost similar, and not many changes have happened in this area. Even though the game held for many years, the researchers do not address time series analysis adequately. This study will focuses on the time series factor of the individual teams in the EPL and predicts the result of matches between the possible pairs of clubs considering home and away matches. Statistical time series methods are compared with LSTM and Bayesian Models.

1 Introduction

1.1 Motivation and Project Background

Soccer is the most well-known and profitable sport on the planet. Predicting the outcome of a game is fascinating to both managers and soccer fans worldwide. Since, the evolution of machine learning new technology, soccer game prediction has grown in popularity among scholars and fans. Bookmakers also keep watch of the game and attempt to anticipate the outcome. A soccer game prediction is a demanding and challenging task due to the unpredictability of the game's outcome. Team cooperation, each player's skill and performance, weather, and the chance of injury are all elements to consider when forecasting the result of a match. The unpredictable nature of the events during a soccer game presents a significant challenge for researchers. Rather than talents or other performance indicators, luck can determine a team's success. Furthermore, even with sufficient web data, everything necessitates a significant amount of pre-planning. Researchers analyse data to forecast a team's match outcome, team performance, player performance, or seasonal outcome. Moreover, to achieve a high level of accuracy, each forecast necessitates a sophisticated model. Different soccer leagues, such as of Spanish Football League named La Liga, Qatar Football League (QFL), and the English Premier League, have previously adopted neural networks, decision trees, and KNN. According to a study by Razali et al. (2017), a Bayesian network can forecast the result of a soccer match. Weather, disasters, diseases, events, and sports have all been predicted using Bayesian networks. According to a study conducted by Rosli et al. (2018), the decision tree model may accurately forecast the winning team.

Together with Spanish and Italian leagues, the English Premier League (EPL) is considered as world's largest well-known and followed soccer leagues. Furthermore, substantial commercial advertising corporations and sponsors are particularly interested in each EPL team's performance. These factors increased the interest of researchers in predicting soccer match prediction, especially the English Premier League. Even though there are game result data of around 30 years, the time-series aspect of the match results is not exploited as expected. Most of the works in this area are predictions based on regression and classification algorithms. This research will concentrate on the time series aspects of the EPL match data to propose models to predict soccer match results. Previous works done to predict the result of soccer data show a high accuracy of 90, which is cross-checked in this work. Moreover, this study provides a comparison of the output accuracy.

1.2 Research Question

This research is based on the following aspects

Qn How accurately can each team's EPL match outcome be predicted based on the time series?

Sub Qn Tries to recreates, and refining existing Bayesian algorithms, compares the results, and critically evaluates them.

In this research, datasets are available from the free website. Different time series methods are applied to get a promising model to predict the game result. Moving Average, the Lag method is used while defining the model. The sliding window-based LSTM model is defined and evaluated its performance. Moreover, each model is compared with other models and critically evaluated. These time-based models are then compared to the previous model's Bayesian models. This study aims to produce a model that can forecast the result based on different clubs' hidden time and seasonal aspects and compare it with other models.

2 Related Work

This section will explore the major experiments done in the area of soccer game. First, discuss about the various soccer game prediction attempts done by different researchers across the glob. Then, this is followed by the critical analysis of the comparison work done on machine learning soccer outcome prediction. Finally, this will deal with the previous researches that proposed methods to predict the player statistics and team line up of a soccer match.

2.1 Prediction of Football Match outcome

This study gives more importance to potential soccer result predicting algorithms published over the years. For example, Razali et al. (2017) mentions about a model to forecast the result of a soccer game in EPL using Bayesian Network. They gathered the necessary dataset for the study from online website. Using the Bayesian algorithms, the authors were able to predict EPL soccer match result from 2010 to 2013. They forecast three possible outputs; Home win, Away win and Draw. From the selected range of dataset, they produced an average accuracy of 75% s. 10-fold cross validation is used for the validation of the algorithm. To proceed with experiments, the researchers used a machine learning tool named WEKA. However, this algorithm is used for classifying the past soccer match and conceptualised as a prediction algorithm. Rahman et al. (2018) used Bayesian Approaches to classify the soccer match result based on win, loss, or draw. This work is an extension of the previous work mentioned. They followed the exact methods to predict a soccer match result. To foresee the result of a soccer game, they employed three different models; Naive Bayes, Tree Augmented Naive Bayes, and General Bayesian Network. The data was gathered from an English soccer match result website, and it covered the years 2014 to 2017. Using Tree Augmented Nave Bayes, they reached a 90% accuracy rate in their experiments (TAN).

Chazan-Pantzalis and Tjortjis (2020) produced a realistic final league table forecast for the world's major soccer leagues. They also attempted to identify the traits of a great soccer defender. They used web scraping to acquire the essential soccer match data from multiple online portals. After the cleaning and pre-processing of the soccer data, they divided the dataset to two sections fortraining and testing. The project's predicting field is bi-variate, and it indicated the performance of a soccer team in current year compared to the previous one. The authors employed different classifiers to categorise the teams to the goal class, and they performed 10-fold cross-validation on the data to ensure efficacy. Finally, they concluded that Random Forest with an accuracy of roughly 70% is the better classifier than others. Using multiple linear regression, the authors also attempted to identify the main variables of a great defensive player. They verified all of the linear regression assumptions. They concluded that team performance and interception, tackles, and clearance skills significantly affect individual defenders' performance. They also believe that the use of wearables and a camera will aid in obtaining more precise data on personal performance.

Yoonjae Cho (2018) defines a new framework named SWLPS, that forecast the results of a soccer match. For the study, the authors use Social Network Analysis (SNA) and Gradient Boosting algorithms are used. Passing of the ball between players is used in the study for developing a SNA based algorithm. They states that the new framework is able to predict the result using GB with good accuracy.

Arabzad et al. (2014) designed a model for forecasting last week outcome of the Iran Pro League's 2012-2013 season. They gathered this information from previous seven years' game summary.To forecast the outcome, they used an artificial neural network. They noted the inescapable circumstances that affect soccer prediction, despite forecasting the future 5 out of 6 teams in the prior week's table. The study's limits included the duration among events, the effect of other professional leagues running concurrently, the amount of money spent by the club, the average age of the players, and the precipitation and conditions at the time.

Huang and Chang (2010), calculated the winning chances of teams of the 2006 soccer world cup using the match summary of each teams group stage fixtures. The Multi-Layer Perceptron (MLP) technique was used to produce the model. They necessary football match summary of world cup data gathered from the official website of FIFA and evaluated different statistics of the 64 matches. There are many variables like Goals For, Goals Against, Penalty Kicks, Fouls Suffered, Yellow Card, Corner Kicks, Off-sides, and so on for each match. For testing and training, the data set was split into two halves. They summarised that scoring a goal in soccer game is hard, and there are many chances to get a game result in draw. Without considering ties, they had a 77 % accuracy rate.

da Costa et al. (2021) tries to figure out a algorithm to predict the score each team can score in a match. They put efforts in finding the hurdles of predicting this value. Also, they compare the result with betting teams and argue that they outperformed the betting teams algorithm. Similarly, The purpose of work done by Odachowski and Grekow (2012) is to check the changes in bookmaker odds and its result in the forecasting of a soccer match. According to the authors, bookmakers' wagers are based on the proper idea of the opposing teams. They gathered and processed data from different betting online sites available in internet. They developed distinct classification algorithms for every game outcome that is possible: home win, away win, and a draw. Ultimately, home and away win is attained with an accuracy of around 70%. They further claim that the outcome of estimating draw output was not particularly impressive. They finish the study by claiming that by establishing a 70% accuracy rate, odd values of bookmakers' is very potent and can be used to forecast the result of a soccer match.

Rotshtein et al. (2005) propose a method that can forecast the outcome of a pair of clubs using the details of their past game results. The authors used fussy rules to train data and produced the required emulation results. First, they created a model based on fuzzy logic rules using the change in value between goals scored and goals conceded. Then, the researchers found out that applying this approach on world-class competitions is not worthy because weather, injury, and booked and reserve players characteristics were not considered. In their study, Hervert-Escobar et al. (2018) present a technique that uses a Bayesian Model-based on rank positioning to predict soccer games in America. More than 200 thousand match results from all over the world are being used to test the model. The objective variable was categorised into three parts: won, lost, and draw. Using the target dataset, they discovered that various leagues have different motives that impact the outcome. Prasetio and Harlili (2016) forecast the English Premier League season of 2015/2016 using general Logistic regression model. Authors attempted in the forecasting of win, loss for the home team and the significant characteristics in a match win. They used logistic regression coefficients to determine which factors were important in the outcome of a soccer match. Also, they used data from numerous seasons to train the algorithm, which turned out to be 69 per cent accurate. They gathered the data for the study from EA Sports' FIFA 2015 soccer simulation game. They argue that using soccer game data for the machine learning analysis helped them to increase efficiency and reduced the time consumption. They say that analysing soccer game data saves time and energy compared to real-world data. Moreover,They claim that players' attributes like heading, passing, shooting skills, and stamina may be employed.

Samba (2019) tried to evaluate the neural network's ability of forecasting the result of a soccer match in this study. To build the model, he gathered data from over 24000 games involving 41 factors.They separated the attributes into two groups: group and group independent. The data was again divided into three sections for training, validation, and testing: 60%, 20%, and 20%, respectively. Then he created a variety of neural networks with varying depths and numbers of neurons per layer. They comes to the conclusion that a neural network with three output neurons is much more precise than one with only one.

2.2 Comparison of Soccer Prediction Algorithms

Rosli et al. (2018) analyses various soccer outcome forecast algorithms in this research. They evaluated different major machine learning techniques such as Decision Trees, Neural Networks (NN), Bayesian Networks, and KNN, using deep learning tool WEKA and GMDH Shell DS tools. Then the authors used data from the EPL from 2013 to 2016, including critical criteria like Full Time Home and Away Goal, Home and Away team shot, Home and Away team shots on Target. They summarise that the Decision Tree gave the most precise output and it is the best technique for predicting the outcome of a soccer match, with a 99 percent accuracy rate. Other approaches, such as NN, Bayesian Network, and KNN, received accuracy below 90%.

Gabriel Fialho (2019) provides a detailed review of the different data analytic and machine learning algoirthm applied in the are of sports like football, soccer, horse riding and athletics. Also they come up with a new AI based algorithm to predict the result of soccer games. The model is based on the neural networks and they mapped all the outcomes of a game in to vector of 0 and 1. They analysed the machine learning algorithms like KNN, Support Vector Machine (SVM) and Fuzzy system in their work.

Azeman et al. (2021) used two different data analytic algorithms to forcast the result of the EPL. They recommended using a Multi-class NN and a Multi-class Decision Forest to forecast the soccer result. The authors gathered the EPL data from 2005-2006, including 380 matches from 20 teams. They compared the two algorithms after their investigation. They came to the conclusion that Multi-class Decision Multi-class Neural Networks were outscored by Forest. In thos study, authors used data from the website [1] , an English soccer results page. Home Team Win, Draw, and Away Team Win were the research's goal classes for prediction. According to Razali et al. (2017), multi-class decision forests and multi-class neural networks have the highest accuracy (88 percent) (71 percent). They finish the paper by stating that because the soccer data set contains categorical variables and tabular values, Random Forest works effectively with it.

2.3 Player and Team performance Analysis

The key performance qualities of winning and losing clubs in the UEFA Champions League are distinguished by Peñas et al. (2011). For the analysis, researchers just used group phase records from the 2007-2010 seasons. The teams' aggregate shots, shoots, effective passes, efficiency, pass, crossing, and off-sides committed and obtained were recorded. According to the writers, the team that wins has a better average in the important match variables: aggregate attempts on target, efficiency, and assists. They further argue that in the variable, the losing team has more yellow and red cards.Shots on goal, crosses, ball control, venue, and opponent quality were found to be the elements that differentiated winning and losing teams.

Yuesen Li et al. (2020) introduced a Linear Support Vector Machine (LSVM) to rank the teams in Chinese Football League (CFL). Moreover, the tried to select the key parameters that will affect the result of the game. They collected data of 1200 games over 4 years and used for the model. From the 165 match features they selected most meaningfull 22 features and applied the LSVM model.

Al-Mulla and Alam (2020) analysed the performance of soccer players in 864 Qatar Stars League (QSL) matches between 2012 and 2019. Various machine learning models were explored for this classification job, with the logistic regression-based model proving to be the top performer, with over 80% accuracy. Surprisingly, they discovered that the importance of defenders in value of soccer game outcome cannot be overstated and that playing fair games enhances the likelihood of victory games in the Qatar League. In their study, they used machine learning techniques to anticipate the results of soccer matches in the QSL regarding player evaluation results.

[1] URL: `https://www.football-data.co.uk`

To mine data on soccer players' abilities, Wang et al. (2009) used an updated Apriori methodology and three connection principles. Researchers devised a method for determining the primary areas in which the players may improve. The inter dependencies and linkages between a player's varied skills are depicted by association rules. The improved Apriori Algorithm AIS is used to extract the dataset's association rules. To reveal the most prevalent technical actions by players, the authors identified five different movements: I1 (steal), I2 (straight attempts), I3 (dribble burst), I4 (midfield push), and I5 (moving the football). Investigators observed that the most successful combination of steals, areal shots, and dribble break, as well as their variations, were the best methods.

The work done by António M. Lopes (2021) introduce a mathematical modelling and visualisations to compare the competitiveness of different international leagues across the world. They collected the data for the analysis from an open website in internet [2]. The competitiveness of soccer leagues in 4 nations is investigated using principles from systems theory and multidimensional scaling. The competitiveness is depicted using 2-dimentional maps that includes different teams perfomance over the a specific period.

Puchun (2016) proposed an enhanced association rule mining method that may be employed during team training sessions and applied to soccer tactics. They first explained association rules, support, and the confidence parameter to establish the main factors. To improve the association rules, they apply particle swarm optimisation. They were able to derive K direct association rules from the information in the end. Researchers end the investigation by putting the system to the test with a soccer match between Spain and Germany in the 2008 European Cup final.

Thinh et al. (2019) built a method that could monitor and quantify player development from soccer match film records. The small number of participants in the video and the cpu utilization necessary to analyze the play in real-time are also disadvantages of this strategy. Efficient Convolution Operators are employed in a real-time game to track the players. They used two cameras to capture the video, then processed it in two different machines in tandem before being blended.

Wang et al. (2020) attempts to create a ball circulating system to evaluate teamwork in a soccer team were attempted initially. They built an unique model for analyzing a team's performance based on game highlights using this system. They offered a detailed report based on spatiotemporal monitoring data of emphasized Everton player and coaching personnel productivity.

Dolores is the name of a novel soccer outcome forecasting model introduced by Constantinou (2019).It could accurately determine the likelihood of any soccer match played anywhere on the planet. Authors has been trained this model with over 50 international soccer competitions, which comprise various levels from over 35 nations. According to the researchers, previous research has shown that critical elements such as player transfer, availability of players, injury, and a new coach can improve the model's accuracy.

2.4 Discussion

Over the years, many researchers tried to predict soccer match result based on the previous data. Some predict the individual match, while others predict the result of one week or one season. Moreover, some brave attempts happened to compare the algorithms to predict the soccer match result. Even though EPL games have been conducted continuously for the past three decades, time series aspects are not much explored.

[2] url:http://www.worldfootball.net/

3 Research Methodology

This section discusses the dataset's details, machine learning techniques, and performance measures to compare the models. The starting point was an experiment to reproduce the work done by Razali et al. (2017) and Rahman et al. (2018) implemented in Weka 3.8.5 [3], using a Python re-implementation. In Weka and Python implementation, 10-fold cross-validation is used to validate. For time series prediction of the future soccer match moving average, a sliding window and LSTM model is used. KDD (Knowledge Discovery in Database) methodology is applied in this study. The steps included in the research is depicted in Figure 1

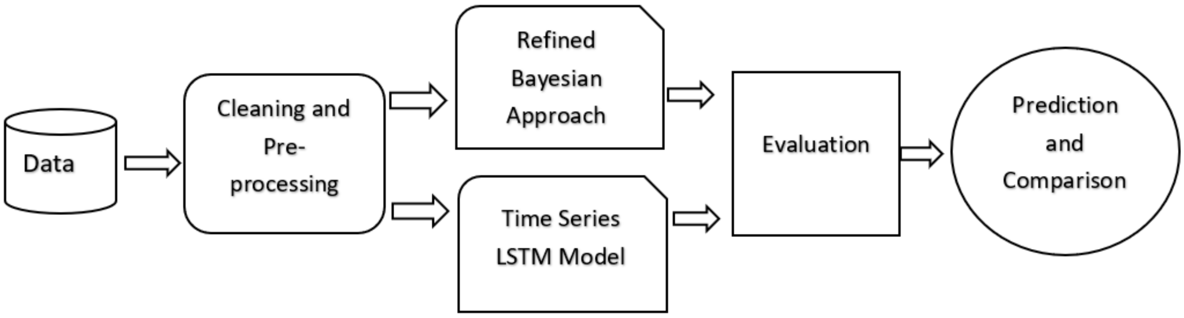

Data
Cleaning and Pre-processing
Refined Bayesian Approach
Time Series LSTM Model
Evaluation
Prediction and Comparison

Figure 1: Research Framework

3.1 Dataset

For predicting a soccer match availability of a proper dataset is critical. Since the game has certain unpredictability, each game statistic is required for the study. The dataset chosen for this study is English Premier League match results from 1993/94 to 2020/21. This dataset is available and can download from the open website [4]. EPL will have 20 participating teams, and each team will play two games (Away and Home) with every other team in each season. Thus, there is a total of 380 games per season. Each downloaded CSV file contains 380 rows containing match details like playing teams, date, individual game statistics and a few betting values, which are irrelevant for the study. For the comparison with Rahman et al. (2018), we used the same dataset as the author's: three seasons 2014/15,2015/16, and 2016/17. They used the result of 3 seasons which are 2014-2015, 2015-2016, 2016- 2017. All the other models use the results from 1993/94 till 2020/21.

3.2 Data Pre-processing

The downloaded dataset is a collection of season-wise CSV files which contain details of all the matches in that season. There are 380 rows in each file, with 68 columns out of 20 is relevant for the study. For the cleaning and transformation of the dataset, Jupyter lab is used. Python libraries like 'pandas', 'NumPy', 'matplotlib', 'glob' is loaded initially.

[3] URL: `https://www.cs.waikato.ac.nz/ml/weka`

[4] URL: `https://www.football-data.co.uk`

The 'pandas' read_csv function is used along with the glob library to retrieve the rows from each file. As the loaded data contains many unwanted columns, those columns are removed using pandas. The dataset is checked for null values and unexpected values. For the Bayesian algorithms, all the game statistics and team details columns are selected, while for the time series, only the date, home team, away team and FTR (Full Time Result). The values of the column FTR are A (away team) or D(draw) or H (home team), which are transformed to 1, 0 and -1, respectively, for the time series analysis. Teams that played only a few numbers seasons in the EPL are omitted from the time series prediction.

4 Design Specification

The techniques used in this study are Bayesian algorithms like Naïve Bayes and Tree-Based Naïve Bayes for comparing the results and LSTM (Long Short-Term Memory) Sequential model for time series prediction. All the models are used for predicting the outcome of a game concerning the home team.

4.1 Refined Bayesian Model

First, tried to replicate the exact Bayesian models introduced in the paper Rahman et al. (2018). Then critically analysed the flaws in the model and remodelled the Bayesian models. In the original model authors applied General Naive Bayes (K2 algorithm) and Naïve Bayes, and Tree-based Naïve Bayes. It uses 10-fold cross-validation for comparing the algorithms and validation, and the dataset has divided into testing and training sets. The division of the dataset was done with a ratio of 9:1. Here the 90% of the dataset is used for training, and the remaining ten percentage is for testing. These models are redefined and compared after adequately cleaning and removing highly correlated variables.

4.2 Novel LSTM Model

The EPL matches have been happening throughout the year and conducted for the past 29 years. The year-wise trend is taken here for predicting the match result for each pair of teams. The match results of every pair over the past years filtered out, and then two separate time series were created for home and away games of each pair in the series. Then, I applied time series techniques on the cleaned processed dataset. Each step can be identified from the below Figure 2.

The cleaned data from the pre-processing stage transformed to an appropriate structure to proceed with the time series. The game result value of each game is converted to values between 1 and -1. Then moving average is found out for each pair of the team, and the last average value is used for predicting the result. Then, a simple sequential model with sliding window values with different window sizes was applied. The LSTM model is applied to the dataset along with a sliding window to improve the model further.

The Keras TimeseriesGenerator produces the input sample for the sequential model. TimeseriesGenerator will accept input and output values, and both are the game result values viewed from the home team. The value of the input size for the generator function is the window size. The output from the algorithms is a value between 1 and -1. The predicted value is rounded based on the range it falls in to compare the output with the actual result, a discrete value of 1, 0 or -1. Different rounding strategies are applied to each algorithm to determine the best accuracy. The rounding range value Γ is introduced to specify each rounding ranges. The predictions in the interval [Γ, 1] are rounded to 1, predictions in interval (Γ, -Γ) are rounded to 0 and [-Γ, -1] is rounded to -1. The range of Γis (0, 1). Accuracy in this study is the percentage of correct predictions relative to the total number of predictions.

4.3 Evaluation

Accuracy and Mean Absolute Error are the evaluation metrics used in this study. Accuracy is the ratio of total correct prediction to the total number prediction in percentage. This metric used to compare all the models, including Bayesian models. Mean absolute error was used to compare the relative performance of the sequential models, including the accuracy.

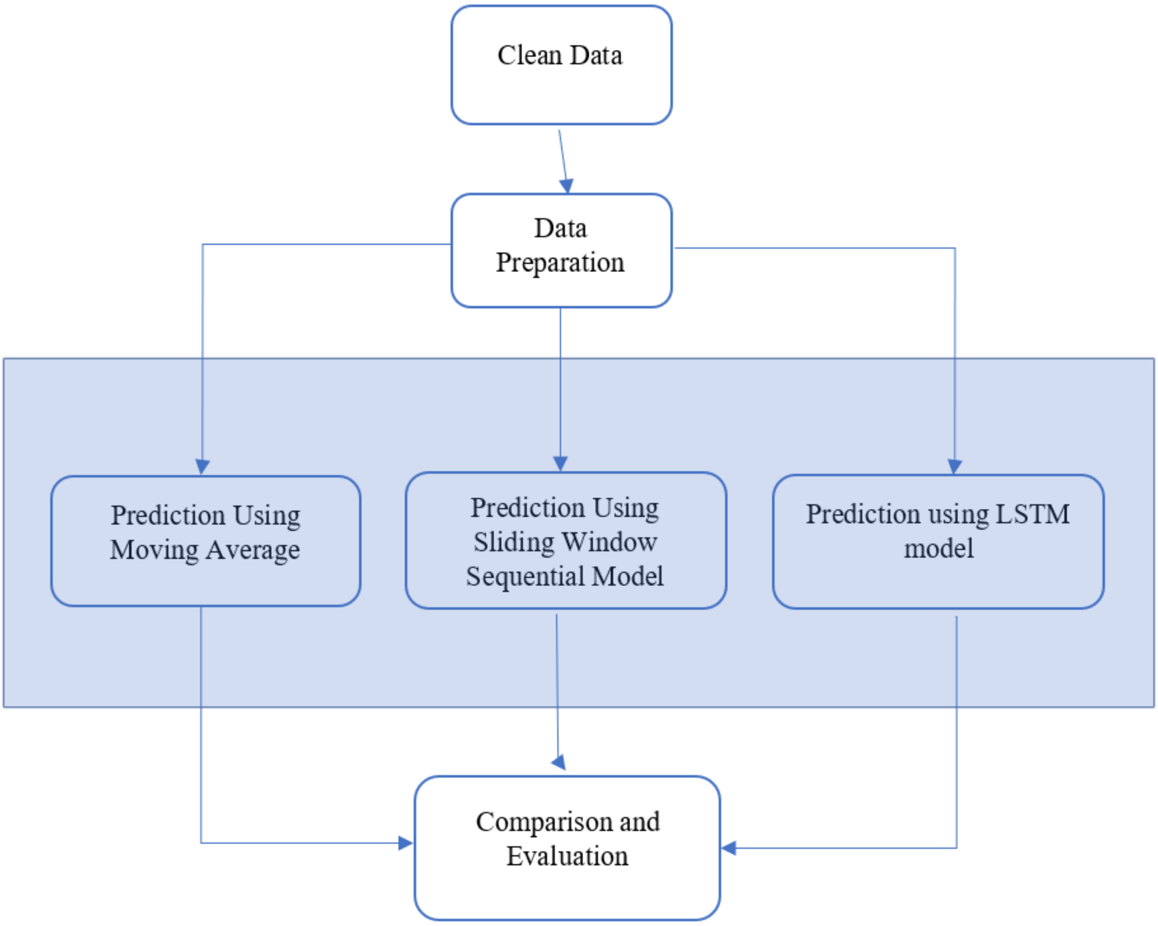
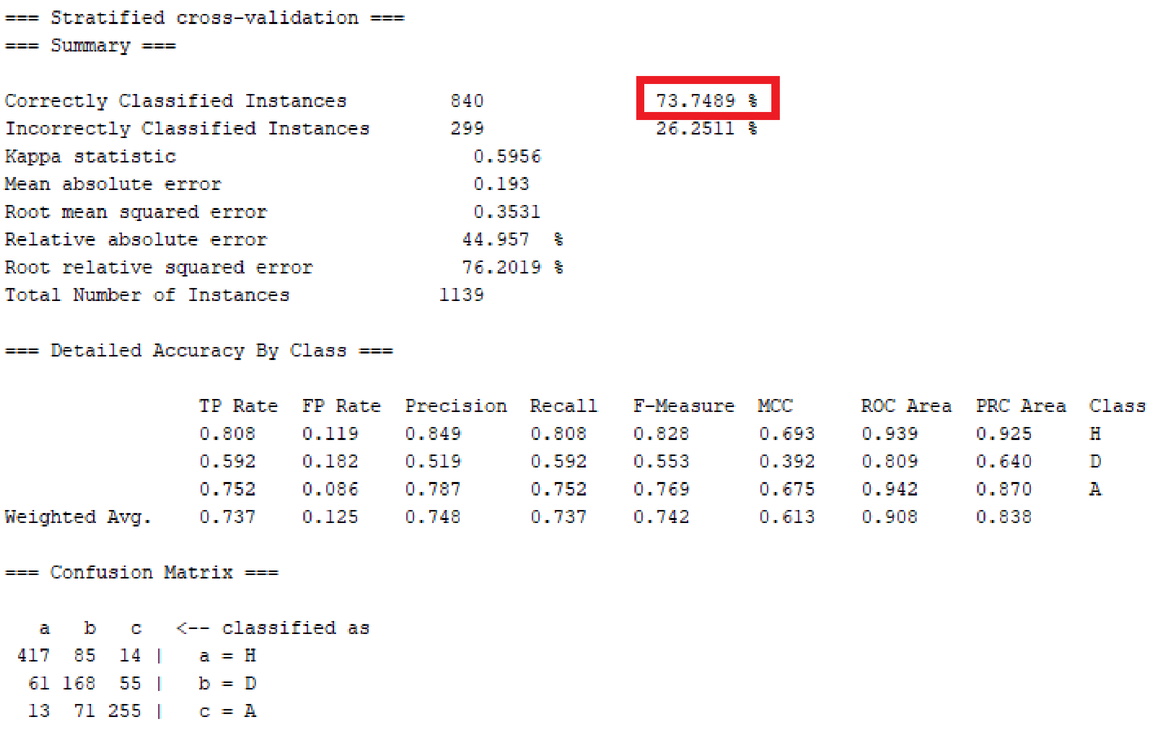

Clean Data
Data Preparation
Prediction Using Moving Average
Prediction Using Sliding Window Sequential Model
Prediction using LSTM model
Comparison and Evaluation

Figure 2: Design of Time series based Models.

5 Implementation

5.1 Bayesian Approach

This study faithfully implements the Bayesian algorithms as in the works of Razali et al. (2017) and Rahman et al. (2018) twice, once in Python and once in Weka. This approach uses the same dataset is used by the authors. The input to the software is a CSV file containing the relevant columns. The data files of three years are merged into one CSV file and loaded to the Weka application. There are options to choose the predicting variable and learning algorithms. The Algorithms used are, General Naïve Bayes (k2), Naïve Bayes and Tree Based Naïve Bayes. Each algorithm will run and be saved for later reference. The below figure is the result of General Naïve Bayes. Here, 10-fold cross-validation is used to validate and compare the algorithms. The data set was split into testing and training with 10 and 90 percentage of data, respectively.



```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         840               73.7489 %
Incorrectly Classified Instances       299               26.2511 %
Kappa statistic                          0.5956
Mean absolute error                      0.193
Root mean squared error                  0.3531
Relative absolute error                 44.957  %
Root relative squared error             76.2019 %
Total Number of Instances             1139

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.808    0.119    0.849      0.808    0.828      0.693    0.939     0.925     H
                 0.592    0.182    0.519      0.592    0.553      0.392    0.809     0.640     D
                 0.752    0.086    0.787      0.752    0.769      0.675    0.942     0.870     A
Weighted Avg.    0.737    0.125    0.748      0.737    0.742      0.613    0.908     0.838

=== Confusion Matrix ===

   a   b   c   <-- classified as
 417  85  14 |   a = H
  61 168  55 |   b = D
  13  71 255 |   c = A
```

Figure 3: Output from Weka for General Naïve Bayes.

Table 1: Comparison of the results from Weka reconstuction

Algorithm	**Rahman et al. (2018) (%)**	**Weka(%)**
General Bayesian Alg	75.2	77.50
Naive Bayes	74.3	67.45
Tree-based NB	90.00	97.30

After executing all three algorithms, the works done by the Razali et al. (2017) and Rahman et al. (2018) reproduced successfully. The K2 algorithm has a 73.7% Prediction accuracy while Naïve Bayes has 67.5%, and Tree-based Naïve Bayes has 97.3%. The comparison of the performance can be visible from the Table 1. The results of Bayesian Network by Razali et al. (2017) had an 75.09% average percentage of accuracy of where as on the Python re-implementation has an average accuracy of 64.6% However, this approach has a few drawbacks.

By cross-checking the training dataset, it is evident that there are a few columns that are correlated to the final result. The fields like Half time result (HTR), Full Time Home Goals, Full Time Away Goals, Half Time Home Goals and Half Time Away goals will affect machine learning. Predicting the final match result with these variables is pointless and does not need machine learning algorithms.

5.2 Refined Bayesian Approach

The near perfect reconstruction of the research of Razali et al. (2017) and Rahman et al. (2018) indicate that there was a flow in that research. In a counter experiment the correlated columns are removed from the training and testing dataset. The dataset only has the game statistics without any goal count and half-time result. The same procedure is re-iterated with the new cleaned dataset and validated with 10-fold cross-validation. A drastic drop in the prediction accuracy is noticed in the result. In this experiment, the General Naïve Bayes has an accuracy of 51.1% while Naïve Bayes has 52.2 and Tree Augmented algorithm has 52.7 percentage of accuracy.

5.3 Prediction using Moving Average

This model is implemented to predict the result of 2020 EPL games based on the data from 1993 to 2019. As the time series need many games, the teams that played fewer games are removed. This analysis was done with Jupyter Notebook 6.4.6 with python version 3.9.5. Each possible pair from the selected teams was taken and created the match result list over the past completed games. Then, this list is converted to time series using pandas functions. Moving average is calculated on this time series using EWM (Exponential Weighted Means) of python. The last mean calculated for the function is taken as the prediction for 2020. This predicted value is then rounded on a different basis and then calculated the mean absolute error and accuracy in percentage.

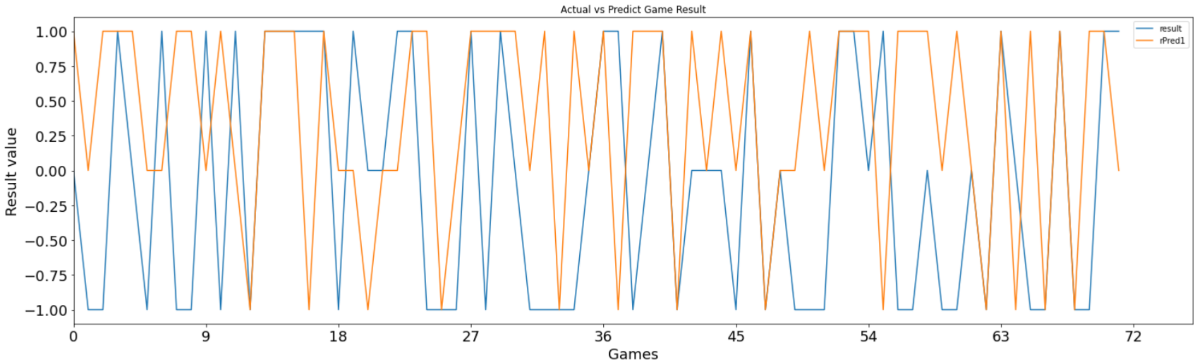


Figure 4: Prediction vs Actual Result of Moving average based Prediction.

First, the game prediction was calculated using the Γ as 0.5 and got the accuracy of

36.1 in percentage. Then changed the rounding range value Γ to 0.33. The change in Γ improved the accuracy to 40.2%, with a mean absolute error of 0.845. Different span value is applied for the ewm function and found out that 4 is the optimum value that gives the best result with a Γ value of 0.33. Figure 4. shows the graph of predicted and actual values of the games in the 2020 season. The x-axis shows each game that happened, and Y-axis shows the result.

5.4 Prediction Based Simple Sequential Model

Here, the time series data require modifications before it can be used for training a sequential model. This deep learning algorithm requires input data to collect samples, which contain an input and output component. For creating this transformation, TimeseriesGenerator from Keras deep learning library was used. This method will automatically convert the input time-series into samples fed into the simple sequential model for training. An instance of the TimeseriesGenerator was created with specifying input and output parameters. Here, input and output parameter is the list game result over the past years—additional parameters like batch and size, input_size specified along with the parameters. The input_size is a parameter specifying the number of lags is used in the input sample. The total number of samples in the generator will be short from the input values by the input lag size. This TimeseriesGenerator can train the deep learning model using the model.fit_generator method. The simple sequential model has two Dense layers of output shape of 100 and 1. The activation function applied is relu, and a window size of 5 is given as the input dimension. The model has trained 200 epochs with steps in each cycle is defined as 1. This model is used to predict the 2020 game result. The last input value is passed to the predicted function and got the predicted result to predict the result. These steps are done for each pair of teams in the game and find the predicted output.

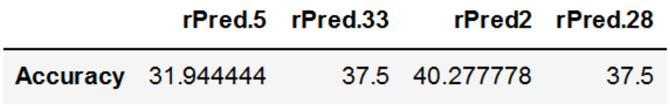
Further, the predicted result is rounded based on different ranges, and the best performing rounding is found. The splitting range of value Γ of 0.2 got an accuracy of 40.27% while rounding range value of 0.33 gave an accuracy of 37.5%. The absolute mean error of the accurate model is 0.875. Result and graph is given in Figure 5.

5.5 5.5 Prediction Based on LSTM Model

To improve further, the LSTM model is applied instead of the Dense layer at the beginning with an output value of 100. The input shape of the model is (window_size, 1). This model was applied with different window sizes, and it was found out that 5 is the optimum value for the window size, which gives maximum accuracy. Also, different rounding strategies were applied to the predicted values to see the accuracy, and it is found that splitting the predicted values with Γ of 0.5 gives the better accuracy of 43.5 in percentage. The absolute mean error of this model is 0.62. Its accuracy for each Γ value is represented in Figure 6 and the graphical representation of the actual and predicted values plotted in Figure 7.

6 Evaluation

In this part, section 6.1 critically evaluate the Bayesian models based on accuracy and section 6.2 evaluate noval approach based on accuracy and Mean Absolute Error.



	rPred.5	rPred.33	rPred2	rPred.28
Accuracy	31.944444	37.5	40.277778	37.5

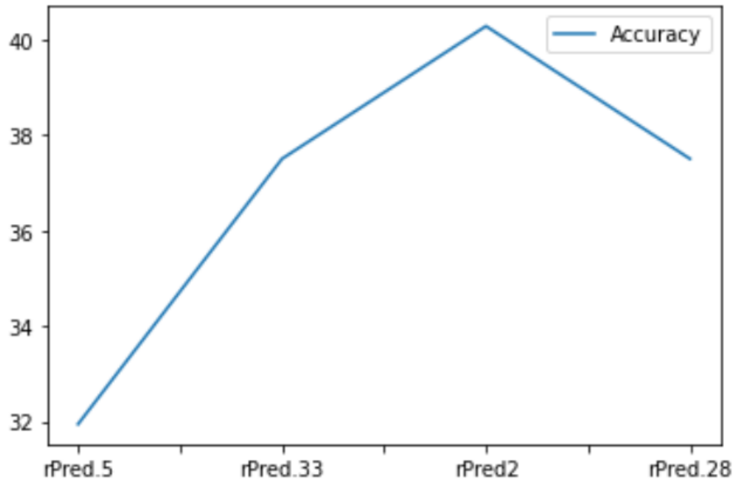
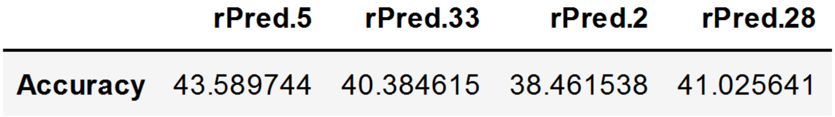


Figure 5: Accuracy vs Rounding method (Γ).



	rPred.5	rPred.33	rPred.2	rPred.28
Accuracy	43.589744	40.384615	38.461538	41.025641

Figure 6: Accuracy for each Γ of LSTM.

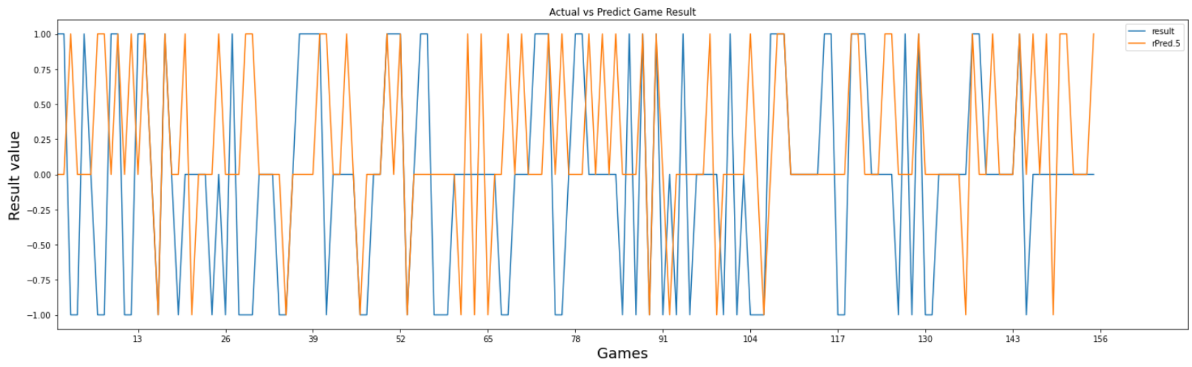


Figure 7: Accuracy vs Rounding method for LSTM.

6.1 Analysis of Bayesian Models

The three Bayesian models described by Rahman et al. (2018) has good accuracy because of the presence of input variables that can directly describe the output of a game. After the removal of these variables, the prediction accuracy dropped drastically. The result are given in below Table 2. It should be noted that the Bayesian model predicts the result based on secondary information about the same game. The data fields used in this model given in Table 3.

Table 2: Accuracy of Bayesian Algorithms Before and After Refining

Algorithm	**Before Cleaning(%)**	**After Cleaning(%)**
General Bayesian Algorithim	77.5	51.18
Naive Bayes	67.45	52.02
Tree-based Naive Bayes	97.37	52.72

Table 3: Data Fields used in Bayesian Model

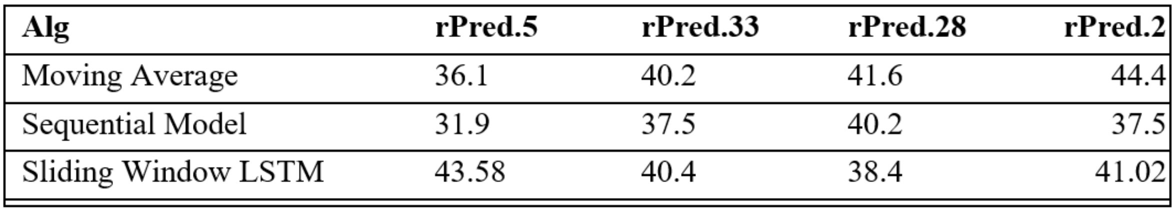
Field Name	**Description**	**Values**
HomeTeam	Name of Home Team	Liverpool
AwayTeam	Name of Away Team	Everton
FTHG	Full Time Home Goal	3
FTAG	Full Time Away Goal	2
HTAG	Half Time Away Goal	2
HTAG	Half Time Away Goal	1
HTR	Half Time Result	H
HS	Home Shot	14
AS	Away Shot	6
HST	Home Shot Target	5
AST	Away Shot Target	4
HF	Home Foul	2
AF	Away Foul	6
HC	Home Card	3
AC	Away Card	2
HY	Home Yellow	1
AY	Away Yellow	2
HR	Away Red	0
AR	Away Red	0

There for the prediction of the game result is only possible after finding all the game statistics. There is no need to predict or classify the match result when the match statics are available in a real scenario. Addressing this issue is irrelevant in machine learning studies.

6.2 Evaluation of Time Series Models

The time series models try to predict the game result based on the results of the previous games. All the three models predicted the game showed almost the same accuracy.

Accuracy of the models given in Figure 8. The moving average prediction algorithm showed a maximum of 40.2% accurate prediction while using Γ as 0.33 when considering a span of 4 values. The lag method tried supervised learning on the time series and did the same rounding but could not improve the accuracy. The Γ 0.2 gave better performance for the sequential model-based prediction algorithm. The higher level LSTM sequential model also performed similar to the other models, and it gave higher accuracy of 43.58% when the rounding range used is Γ value 0.5. It means that when LSTM used, the predicted values of a home win and home loss is almost near to the 1. While for other models, the predicted values of a home win and loss is more distributed. Also, when checking the mean absolute error of each model, LSTM has the least value. The MAE values of the models given in Table 4.



Alg	rPred.5	rPred.33	rPred.28	rPred.2
Moving Average	36.1	40.2	41.6	44.4
Sequential Model	31.9	37.5	40.2	37.5
Sliding Window LSTM	43.58	40.4	38.4	41.02

Figure 8: Accuracy of Each Algorithm for Each Γ

Table 4: Absolute Mean Error of the Better Models

Algorithm	Γ value	AME
Moving Average	0.2	0.845
Sequential Model	0.28	0.875
Sliding window based LSTM	0.5	0.62

6.3 Discussion

From the above experiments, TimeseriesGenerator based LSTM model performs well compared to other models even though the performance is almost similar. The rounding method used for prediction value may be the reason for all three time series-based methods' comparatively low performance. Also, eliminating poorer performing teams that where only shortly in the EPL automatically eliminated the easily predictable instance. It is very complicated to predict the output of equally strong teams in a league than one strong and one weak team. On the other hand, a traditional Bayesian algorithm can predict only 50% accurately with game field values actually after the game. Also, it has all the strong and weak teams in the algorithm, which is an indication that the game results are by their nature random.

7 Conclusion and Future Work

First, this research reproduced and refined existing Bayesian-based methods on predicting the soccer output and critically evaluated the method. Then, the study thoroughly studied the time aspect of the soccer match result of the EPL soccer league and introduced three separate models to predict the future game result. All three models could predict the result with a 40% prediction accuracy. However, soccer is a highly random game and depend on numerous factors like players emotions; 40 per cent of accuracy is acceptable to predict a game. Improving the LSTM model by modifying the model parameters and adding more layers to the model might improve the prediction result.

There are many aspects for future work. These models can be made deeper to improve the result. Moreover, researchers can try different model parameters and layers in the LSTM. Also, studying more on the rounding method will help improve this methods' efficiency. Moreover, a time series analysis can be done, including all the EPL clubs over the years. Also, there is scope to implement these methods in other leagues across the world. Furthermore, for future work, it might be interesting to consider the odds by betting organisations and comparing with the proposed prediction method.

8 Acknowledgement

I would like to express my sincere gratitude and to my supervisor Dr. Christian Horn who made this work possible. His patience, motivation and the immense knowledge helped me all the time of the research and writing of this thesis.

Also I am thankful to my family, friends and colleagues for their support and motivation during the coursework.

References

Al-Mulla, J. and Alam, T. (2020). Machine learning models reveal key performance metrics of football players to win matches in qatar stars league, *IEEE Access* **8**.

António M. Lopes, J. T. M. (2021). Modeling and visualizing competitiveness in soccer leagues, *Applied Mathematical Modelling* **92**: 136–148.

Arabzad, S. M., Tayebi Araghi, M., Soheil, S.-N. and Ghofrani, N. (2014). Football match results prediction using artificial neural networks; the case of iran pro league, *International Journal of Applied Research on Industrial Engineering* **1**: 159–179.

Azeman, A. A., Mustapha, A., Razali, N., Nanthaamomphong, A. and Abd Wahab, M. H. (2021). Prediction of football matches results: Decision forest against neural networks, *2021 18th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, pp. 1032–1035.

Chazan-Pantzalis, V. and Tjortjis, C. (2020). Sports analytics for football league table and player performance prediction, *2020 11th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pp. 1–8.

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

da Costa, I. B., Marinho, L. B. and Pires, C. E. S. (2021). Forecasting football results and exploiting betting markets: The case of "both teams to score", *International Journal of Forecasting* .

Gabriel Fialho, Aline Manhães, J. P. T. (2019). Predicting sports results with artificial intelligence – a proposal framework for soccer games, *Procedia Computer Science*, Vol. 164, pp. 131–136.

Hervert-Escobar, L., Matis, T. I. and Hernandez-Gress, N. (2018). Prediction learning model for soccer matches outcomes, *2018 Seventeenth Mexican International Conference on Artificial Intelligence (MICAI)*, pp. 63–69.

Huang, K.-Y. and Chang, W.-L. (2010). A neural network method for prediction of 2006 world cup football game, *The 2010 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8.

Odachowski, K. and Grekow, J. (2012). Using bookmaker odds to predict the final result of football matches, Vol. 7828, pp. 196–205.

Peñas, C., Lago Ballesteros, J. and Rey, E. (2011). Section iii – sport, physical education recreation differences in performance indicators between winning and losing teams in the uefa champions league, *Journal of Human Kinetics* **27**: 135–146.

Prasetio, D. and Harlili, D. (2016). Predicting football match results with logistic regression, *2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)*, pp. 1–5.

Puchun, W. (2016). The application of data mining algorithm based on association rules in the analysis of football tactics, *2016 International Conference on Robots Intelligent System (ICRIS)*, pp. 418–421.

Rahman, M., Mustapha, A., Fauzi, R. and Razali, N. (2018). Bayesian approach to classification of football match outcome, *International Journal of Integrated Engineering* **10**.

Razali, N., Mustapha, A., Yatim, F. and Aziz, R. (2017). Predicting football matches results using bayesian networks for english premier league (epl), *IOP Conference Series: Materials Science and Engineering* **226**: 012099.

Rosli, C., Saringat, M., Razali, N. and Mustapha, A. (2018). A comparative study of data mining techniques on football match prediction, *Journal of Physics: Conference Series* **1020**: 012003.

Rotshtein, A., Posner, M. and Rakityanskaya, A. (2005). Football predictions based on a fuzzy model with genetic and neural tuning, *Cybernetics and Systems Analysis* **41**: 619–630.

Samba, S. (2019). *Football Result Prediction by Deep Learning Algorithms*, PhD thesis, Tilburg University.

Thinh, N. H., Son, H. H., Phuong Dzung, C. T., Dzung, V. Q. and Ha, L. M. (2019). A video-based tracking system for football player analysis using efficient convolution operators, *2019 International Conference on Advanced Technologies for Communications (ATC)*, pp. 149–154.

Wang, B., Yin, Z. and Wang, L. (2009). Research of association rules in analyzing technique of football match, *2009 2nd International Conference on Power Electronics and Intelligent Transportation System (PEITS)*, Vol. 3, pp. 178–180.

Wang, Y., Wang, H. and Qiu, M. (2020). Performance analysis of everton football club based on tracking data, *2020 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS)*, pp. 49–53.

Yoonjae Cho, Jaewoong Yoon, S. L. (2018). Using social network analysis and gradient boosting to develop a soccer win–lose prediction model, *Engineering Applications of Artificial Intelligence* **72**: 228–240.

Yuesen Li, Runqing Ma, B. G., Gong, B., Cui, Y. and Shen, Y. (2020). Data-driven team ranking and match performance analysis in chinese football super league, *Chaos, Solitons Fractals* **141**: 110330.