PDF Download
3686397.3686398.pdf
04 January 2026
Total Citations: 0
Total Downloads: 980

Published: 24 June 2024

Citation in BibTeX format

ICISDM 2024: 2024 the 8th International Conference on Information System and Data Mining
June 24 - 26, 2024
CA, Los Angeles, USA

RESEARCH-ARTICLE

Survey of Soft Computing Methods to Predict Soccer Win/Loss Probability

KAYANNA A MORGAN, University of North Dakota, Grand Forks, ND, United States

EMANUEL SYLVESTER GRANT, University of North Dakota, Grand Forks, ND, United States

EUNJIN KIM, University of North Dakota, Grand Forks, ND, United States

Open Access Support provided by:

University of North Dakota

ICISDM '24: Proceedings of the 2024 8th International Conference on Information System and Data Mining (June 2024)
https://doi.org/10.1145/3686397.3686398
ISBN: 9798400717345

Survey of Soft Computing Methods to Predict Soccer Win/Loss Probability

Kayanna A. Morgan*
School of Electrical Engineering and Computer Science, University of North Dakota
kayanna.morgan@und.edu

Emanuel S. Grant
School of Electrical Engineering and Computer Science, University of North Dakota
emanuel.grant@und.edu

Eunjin Kim
School of Electrical Engineering and Computer Science, University of North Dakota
eunjin.kim@und.edu

ABSTRACT

Worldwide sports activities have one of the largest supporter/fan bases of all the many areas of entertainment. The sport of soccer is one of the richest sports in the world and commands the position of being the most popular sport in the world. That single data point forms the basis for an interesting set of data analytical research questions. First on that list would be: How is soccer determined to be the most popular sport in the world? How do the factors of advertising/marketing, televising/streaming, teams' win/loss, players, location, etc. contribute to the popularity of the sport? What impact does the popularity of soccer have on the players' performance? Can data analytics predict teams' performance? The report presented herein is a documentation and review of prior research efforts that have applied soft computing methods to predict teams' win/loss probability in the sport of soccer. This review comments on the success of the data analytics applied and makes an assessment of how this can affect the popularity of the sport.

CCS CONCEPTS

• **Applied computing**; • **Operations research**; • **Decision analysis**; • **Multi-criterion optimization and decision-making**;

KEYWORDS

soft computing, data analytics, sport popularity, sports' win/loss prediction, literature survey

ACM Reference Format:
Kayanna A. Morgan, Emanuel S. Grant, and Eunjin Kim. 2024. Survey of Soft Computing Methods to Predict Soccer Win/Loss Probability. In *2024 the 8th International Conference on Information System and Data Mining (ICISDM 2024), June 24–26, 2024, Los Angeles, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3686397.3686398

1 INTRODUCTION

The definition of the term sport or sports is not as simple as it looks. In his research article on that topic [1], Professor Sutula reviewed multiple special literatures in which the term sport is used and concluded with the definition "as a special sociocultural phenomenon, is a historically determined activity of people connected with the use of physical exercises, which is aimed at preparing and participating in a specially organized system of competitions, as well as individual and socially significant results of such activity." [1] Sport is a billion-dollar field that has attracted interest from many viewers because of its impact on the quality of human life in terms of providing satisfaction to players and audiences [2]. Its rise in popularity has led to an increase in recent interest from the data science and analytical community [3, 4]. The development of data analytics in sports has played a role in how athletes, coaches, and organizations view the team and plan their game plays. Sports analytics have been favored by the statistical community and less discussed by the machine learning community [4]. While this has aided in real-time analysis, there is growing interest in machine learning algorithms to derive models based on probabilistic reasoning, which would contribute to sports development [3, 4].

From the list of sports, one that has seen great success in the application of data analytics is Soccer. Soccer is a popular sport worldwide that continues to grow. The WorldAtlas website lists the sport of soccer as the most popular sport with over 3.5 billion fans, with its origins going back 3000 years. It is estimated that there are over 240 million registered players and it generates the largest revenue, having a net worth of US$350 billion. As such, soccer has garnered a lot of attention, bringing to light the difficulty of making win predictions because of the many factors the match depends on such as skills, current score, team morale, etc. [5]. Despite this, there have been many proposed algorithms for win forecasts in soccer, however, these algorithms encounter challenges such as emotion or lack of adequate data sets [5].

Research papers on this topic can be classified into three categories with respect to win probability predictions. Statistical models [6], machine learning models [7], and probabilistic models [8]. Additionally, a few researchers have proposed using fuzzy logic [9] to improve machine learning methods. This paper will comprise the proposed soft computing methods [10] for determining a team's win probability in soccer.

The decision to compile this study was driven by a deep fascination with the power of data-driven insights and their potential to revolutionize the sports industry. As data science and analytics become more prevalent in society it is fascinating to see how it would affect the sports environment to improve the gameplay and entertainment value. By exploring the various methodologies and statistical models, the research offered valuable insights and actionable recommendations for professionals in the field. This may be

*Place the footnote text for the author (if applicable) here.

ICISDM 2024, June 24–26, 2024, Los Angeles, CA, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1734-5/24/06
https://doi.org/10.1145/3686397.3686398

used by betting corporations when developing their odds or the coaching staff to develop strategies for their gameplay against opponents. Furthermore, the research will highlight possible gaps in the current methodologies thus opening additional research pathways. Ultimately, the goal is to contribute to the broader understanding of the use of soft-computing methods in sports and the impact these methods have on the industry.

The remainder of this report is as follows: The next section presents a compilation of all methods used in the reviewed articles. Next are documents of the statistical articles selected for this review process. The next section presents the machine learning algorithms articles, followed by the articles that used the fuzzy logic approach. A review of the work is then presented in the discussion section, which precedes the conclusion.

2 BACKGROUND ON SOFT COMPUTING METHODS REVIEWED

2.1 Machine Learning

Machine learning (ML) predicts future data by discovering patterns from past data. As such, this differentiates ML from other analytical approaches [11]. Whereas the traditional approaches require predefined rules and processes to examine the datasets [11], ML may be used to make decisions under uncertainty [12]. Many of the methods used in machine learning to forecast sporting results are complicated in the sense that they use multiple assumptions, may be confusing, and arduous because they require large statistical sampling [5]. In other words, the data the machine should be learning from has not been agreed upon. This leaves the decision up to the researchers to determine what factors they deem important for the machine to consider. The machine learning methods implemented in the surveyed research are defined below.

Artificial neural networks (ANN) [13] have been used increasingly to identify, classify, and predict performance in the realm of soccer [13]. ANNs are computational intelligent models used for prediction in complex systems. Some difficulties associated with the use of ANN include the vast amount of match data available and it is challenging to understand and interpret to simplify objectively [13]. Also, the attributes to be used for the training process have been neglected [13]. That is, it has not been determined what attributes make significant contributions when determining the predictions.

Radial basis function (RBF) [13] is a three-layered neural network with an input layer that is nonlinear, one hidden layer that performs the functions of radial stimulation, and a linear output layer [13].

Support Vector Machine (SVM) is an algorithm that is categorized under supervised learning. It analyzes data for classification and regression analysis. It sorts data into two categories and determines which category a new data point belongs in [14].

Naïve Bayes Classifier is an algorithm based on the Bayes Theorem commonly used to classify objects. It has the assumption that all attributes of a data point are independent of each other. The more data it has the more accurate the prediction [14].

Random Forest is an ensemble method, that is, it is made up of many individual decision trees. This leads to better predictive performance than any algorithm would do alone. Each tree prediction is combined with others in the ensemble to produce a more accurate prediction [14].

2.2 Fuzzy Computing

Although humans are presented with an immense amount of information at any given moment, the mind is capable of organizing the data and concentrates only on relevant information [15]. Note that if a computer had to process the same amount of data as humans, it would ‘choke’. The mind’s ability to only process relevant data is related to the Fuzzy process [15]. Zadeh describes Fuzzy logic as being dissimilar to classical logical systems as its purpose is to model reasoning that does not contain distinct values, but rather a degree of values [16]. This condition is an aspect that mimics the human way of making rational decisions within uncertain environments [16]. In a similar way that the human mind reduces the amount of information it must process, fuzzy logic does the same by assigning membership degrees between 0 and 1, with values closer to 1 considered relevant information to be processed [15]. This allows the system to process decisions quickly and effectively [15]. It removes the distinct linguistic evaluation such as true/false, and yes/no and describes intermediate linguistic values known as fuzzy rules instead [15]. Below are the fuzzy methods used in the reviewed articles.

A Conditional fuzzy inference model contains a set of fuzzy rules which are acquired from the in-sample with a power assigned to each rule based on the frequency and accuracy of said rule [15]. All rules are then ranked accordingly with the strongest rule applied to the out-of-sample [15].

The Adaptive Network Fuzzy Inference System (ANFIS) is one of the most popular fuzzy inference systems used to develop fuzzy rules; however, the defuzzification methods may vary [15]. One method of defuzzification involves applying the maximum operator to the qualified fuzzy outputs that meet some of the firing qualities. Next, the aggregated output is calculated by a function such as the mean of maxima [15]. Another defuzzification approach, known as the Sugeno approach, utilizes a general linear model, based on inputs. It is estimated and the aggregate output is the weighted average of all the rules involved [15].

In the ANFIS method, the conditional approach that utilizes the first defuzzification method from the prior paragraph is often overlooked but it displays promising qualities for match predictions [15]. However, an issue that may be encountered is overfitting, where the ANFIS uses a double-pass algorithm to optimize the rule specification. Overfitting occurs when the out-of-sample performance is considerably worse than the in-sample performance, which leads to an overall undesirable performance [15].

3 STATISTICAL MODELS

3.1 Optimal Sports Math, Statistics, and Fantasy

Kissel applied six sports model approaches to Major League Soccer (MLS) results for the 2015 season. These methods are based on linear regression techniques and probability estimation methods [17]. The model evaluation was based on the winning percentage, R2 goodness of the estimated victory margin, and the regression error. All draws were counted as half of a win and half of a loss because ties are common in MLS [17].

The first model introduced is the game scores regression model which predicts the game outcomes based on the average number of goals scored and allowed by each team. The variables used include the home team victory margin, the average number of goals scored, and the average number of goals allowed for each team per game. This method correctly predicted the game-winner 64% of the time [17].

The second method is the team statistics model which includes the variables shots on goal, corner kicks, and offside penalties for each team per game. The success rate for this method is 66.1% and the estimated goal spreads were accurate to within half of a goal in 90 matches and one goal in 172 matches [17]. The projections missed by more than three goals 23 times with the largest discrepancy being a 5-0 win where the projection predicted the losing team to win by 0.7 goals [17].

Thirdly, the logistic probability model considers the home and away team ratings. These ratings are determined by the maximum likelihood estimates. The home team victory analysis is derived from a second analysis where the authors regress the actual home team spread on the estimated probability of the input variable. The success rate of the model was 65% [17]. In the playoffs, the model's favorites won 12 of 17 matches, losing three and drawing two [17].

Another use of a linear regression model was the team rating model. This model used team ratings derived from the prior logistic probability model [17]. The variables used include strength rating and team rating for each team as well as home-field advantage. This model resulted in a 65% success rate where visiting teams were only favored in 13 matches [17].

Next, the logit spread model is a probability model that predicts the home team victory margin based on the inferred team rating metric. The model alters the home team victory margin to a probability value between 0 and 1 using the cumulative distribution function. Then it estimates model parameters from the logistic regression analysis using the variables home field advantage, home team parameter, and away team parameter. The home team winning margin is determined by performing a regression analysis on the actual spread as a function of the estimated spread to determine the second set of model parameters. This model was one of the more accurate predictive models with a success rate of 65.5% [17]. The model's favorites won 13 matches and had 2 losses and 2 draws in the playoffs [17].

Lastly, the logit points model predicts the home team victory margin by taking the difference between home and away teams' predicted goals. Similar to the logit spread model, the predicted goals are determined based on the inferred team ratings using the parameters home field advantage, home team rating, away team rating corresponding to the home team goals, and away team rating corresponding to away team goals. This model matched the Logit Spread Model with a 65.5% success rate and was 12-3-2 when predicting playoff matches [17].

3.2 Predicting Match Outcomes in Association Football using Team Ratings and Player Ratings

Like Kissel, Arntzen, and Hvattum [18] focus on the rating of the teams. However, unlike the previous study, these authors compare the importance of using team ratings or the players' ratings. The authors state that it is common to evaluate the team strength of each team involved in the match [18]. They continue to add that existing articles have barely touched on whether the player ratings are more informative than team ratings regarding the match outcome predictions [18]. Specifically, while player ratings provide more information in the predictions, they may also be noisier than team ratings, thus impacting predictions [18]. The authors aim to compare the ordered logit regression (OLR) model, used for generating a pre-game forecast, and the competing risk model which is commonly used for in-game predictions [18].

The experiment comprised data collected from the four divisions of the English League, Championship League, Premier League, League One, and League Two as well as matches from English League One. The information ranged over 10 years, from 2009 to 2019. In total, 129 matches were considered after data cleaning that involved removing games played on neutral ground and the early matches of League One [18].

The seasons that were played between 2009 to 2014 are used for initial calculations. Next, the season which started in 2014 to the season that ended in 2017 is used for the initial observation for the statistical model, where each new game day the player ratings are updated [18]. The final set of seasons, ranging over the years 2017 to 2019 are used to evaluate the predictions created by the statistical models [18].

Ratings are based on the ELO rating system [8]. The authors observed that for both models, player ratings provide better results than team ratings but, the differences are not statistically significant for any combination of model and metric [18]. Overall, the team rating and players' rating are complementary. Note that they still perform well as separate entities when determining the win probabilities [18]. However, the OLR [18] is only applicable to pre-game forecasts while the competing risk model can provide updated predictions during the matches by analyzing the current environments as well as pre-game forecasts [18].

4 4MACHINE LEARNING

4.1 Predicting Wins, Losses, and Attributes' Sensitivities in the Soccer World Cup 2018 Using Neural Network Analysis

Hassan and colleagues state that predicting the results of soccer matches has garnered much attention from the machine-learning community in recent years [13]. Because of the increased accessibility to match data due to player tracking, it has been shown that match performance is affected by several different contextual, situational, and positional attributes [13].

The authors aim to use a neural network model to process the big data collected from the devices and sensors used to track and analyze the players. Additionally, they estimated the sensitivity of the match attributes affecting the chances of winning or losing using the neural network model [13]. 57 matches were analyzed from the FIFA World Cup of which only 55 were considered. These were matches that ended with a clear winner, thus draws were not included, or were randomly used in the validation phase [13].

The attributes used include but are not limited to, speed, sprints, number of passes, passes received, and fouls. The goals scored and conceded were not included because of their direct correlation to the match outcome, which negatively affects the validity of the results [13].

The network successfully predicted losses with 72.7% accuracy and wins with 83.3% accuracy [13]. While other predictions focus on location, home or away, and goals gained or conceded, this study used many attributes to determine their sensitivity which can be used in real-time matches [13].

4.2 Prediction of Winning Team using Machine Learning

Ajgaonkar and colleagues propose a hybrid model to predict the outcome of soccer games using the English Premier League. They use multiple machine learning classifiers, Support Vector Machine (SVM) [14], Naïve Bayes [14], and Random Forest [14], which are compared amongst one another, and the most accurate prediction is chosen [14]. For further enhancement of the model prediction accuracy, optimization can be done on the classifier [14].

No losses are considered in the training data, only home wins, away wins, and draws. Data is collected from past games from recent seasons. This amounts to 3000 records with approximately 8-10 attributes [14]. Some attributes are deemed unnecessary for the prediction of the results, thus a box plot is used to understand how the values in the data are scattered [14].

In the experiments, 20% of the collected data is used as test data while the other 80% is used as training data. The results show 63% accuracy for SVM, 57% accuracy for Naïve Bayes, and 55% accuracy for Random Forest [14]. The authors discovered that the Attacking Strength and Defensive Strength of both teams are significant variables but should not be the only variables used [14]. Thus, they suggest additional attributes such as Corners for both teams, Home Shots on Target, and Away Shots Total.

4.3 Prediction of Match Outcomes with Multivariate Statistical Methods for the Group Stage in the UEFA Champions League

Parim and colleagues categorized their variables into 2 groups. The first category, notational, includes attributes such as dribbles, loss of control, and tackles [19]. The second category, situational, includes attributes such as match status, game location, and opponent quality [19]. The goals of the authors were to predict the wins, losses, and draws of the teams using 20 in-game variables. These variables include accurate passes, aerial wins, clearances, corners, defensive aerials, lost balls, dribbles attempted, shots on target, total shots, and touches in tandem with the quality of the oppositions, whether they are stronger, balanced, or weaker [19]. Also, the 'scoring first' variable is considered.

The authors used matches over 10 years from 2010 to 2020. This consisted of 1920 total observations with the dependent variable divided into three classes (win, draw, and loss). One-way analysis of variance (ANOVA) [19] was used to identify the statistically significant performance indicators. To determine the strength of the teams, the k-means clustering analysis was applied using the performance indicators. The probability of the teams winning was predicted using decision tree analysis on the performance indicators.

From the 1920 games played, the k-means clustering determined that 338 faced off against weaker teams (cluster 1), 787 faced off against balanced teams (cluster 2) and 795 faced off against stronger opponents (cluster 3) [19]. The decision tree analysis was applied to each cluster where it predicted that teams from cluster 1 had a 68% chance of winning, however, if they scored the first goal the prediction rose to 92% [19]. As a result, an important factor highlighted in this study is the "first goal". Another highlighted indicator was the shots on target, where it was estimated that the winning chance increased by at least 14% [19].

5 FUZZY LOGIC

5.1 Football Predictions Based on a Fuzzy Model with Genetic and Neural Tuning

One of the earlier studies that included fuzzy logic was conducted by Rotshtein and colleagues. They deduced that soccer fans and experts were already making predictions using common-sense assumptions that follow the IF-THEN criteria of a fuzzy model [5]. The example introduced by the authors states that

IF Team A has won all their previous matches,

AND Team B has lost all their previous matches,

AND Team A has won all the matches they have played against Team B,

THEN it is expected that Team A should win [5].

According to the authors, a prediction model is developed in two stages. The first stage deals with the construction of the fuzzy model that connects the result of a soccer match with the previous results of both teams [5]. As such, the authors used a generalized fuzzy approximator. The second stage introduces the tuning aspect for the fuzzy models where the optimal parameters will be determined from the sample used [5]. The model proposed by the authors takes the previous games played by both teams to forecast the match results. The model relies on the method of acknowledging the characteristics of a past-future nonlinear dependence by a fuzzy knowledge base [5].

The simulation is needed for the prediction of the match results using the goal differential. They define the goal difference using five levels which cover high-score win (big win), high-score loss (big loss), low-score loss (small loss), low-score win (small win), and draw [5]. The factors used to control the results of the match include the outcome of the five previous matches played by Team A, the outcome of the five previous matches played by Team B, and lastly, the outcome of the last five games played between Team A and Team B [5].

For tuning, parameters are chosen based on their membership values of fuzzy terms and rule weights by combining genetic (offline) and neural (online) optimization algorithms [5]. The genetic algorithm provides a rough offline hit into the global minimum while the neural network is used for the online improvement of parameter values [5]. For the neural tuning, they injected the IF-THEN fuzzy rules into their neural network.

The authors sampled a Finnish Championship tournament characterized by a minimum of sensations. The training data contained 1056 match results collected over eight years from 1994-2001. The testing data consisted of 350 matches from 1991-1993 [5].

The results obtained were more precise for the extreme categories (big win, and big loss). Thus, the least accurate results occurred for the small win and loss categories [5]. When compared to Genetic tuning, Neural tuning produced more accurate predictions in a shorter time frame. For Genetic tuning, after 52 minutes it was able to make a prediction with approximately 86% accuracy for big loss and big win [5]. However, after 7 minutes of Neural tuning, it was producing predictions of approximately 91% accuracy for big loss and approximately 95% accuracy for big wins [5].

Note, that this model did not take any additional factors into account such as injuries, benched players, suspensions, and more [5].

5.2 A Conditional Fuzzy Inference Approach in Forecasting

While Hassanniakalager and colleagues use this approach to forecast soccer matches and company stocks, this paper will only focus on their approach and results for the match predictions. The authors attempt to forecast the result and goal difference at some of the biggest soccer championships (the English Premier League, Italian Seria A, and Spanish La Liga) from 2005 to 2016. They are only interested in making strong decisions and thus, their model avoids decision-making under uncertainty. Furthermore, the authors benchmark their proposed method against machine learning methods and fuzzy extraction methods [15].

The authors propose a combination of the ANFIS and Sugeno defuzzification methods. Note that when it comes to large-scale issues a crisp output based on a simple linear model is highly favored; however, if a dataset comes with remarkable noise, the uncertainty may be controlled using a selective feature for the aggregation of the fuzzy rules [15].

The rules generated for fuzzy logic provide certain benefits such as being applied by non-experts as the number of rules necessary may be small [15]. Additionally, the chosen rules will not 'fall victim' to overfitting [18]. Machine learning and complex models, unlike fuzzy logic, are prone to overfitting. This is due to its high need for extensive experimentation [15].

For the calculation of the membership function, the Gaussian membership function is used because it provides a smooth-shaped curve around the center of clusters [15]. The conditional fuzzy inference offers eligibility criteria for the defuzzification method which removes any rule which has a weak membership value from the fuzzy rules [15]. The fuzzy rules are then ordered from strongest to weakest, with the stronger rules being closer to the center of the cluster [15]. In other words, the closer the data point to the center of the cluster, the higher its membership value is between 0 and 1. To combat overfitting, the authors utilize a combination of endogenous and exogenous thresholds to ensure the applied rules for forecasting are correctly fitted [15].

For the experiment, the authors start with 83 potential inputs, however, if fed into a predictive model the training time increases while the number of fuzzy rules created may increase up to 283, which is impractical [15]. Thus, to obtain the most practical subsets of potential predictors, the inputs are processed through a Relevance Vector Machine (RVM) which reduces the dimensions of the input vector [16]. The in-sample comprises three seasons whilst the out-of-sample uses two seasons [15].

Using the Premier League as an example, the in-sample would consist of 2006-2007, 2007-2008, and 2008-2009 as its three seasons, and the out-of-sample would comprise 2009-2010 and 2010-2011 seasons [15]. Note that the estimations made in the first out-of-sample season do not roll over to the second out-of-sample season [15]. Thus, the second out-of-sample season acts as robustness to the models [15]. Additionally, the first three home and away games for each team are disregarded to allow for equal starting points for all teams [15].

The combinations of conditional fuzzy inference and the RVM approach improve the accuracy of the underlying RVM system [15]. It generates predictions based on the strongest fuzzy rules which lead to improved predictability by approximately 10% in most of the exercises conducted in the experiment [15].

5.3 A Neuro-Fuzzy Logic Model Application for Predicting the Result of a Football Match

Onwuachu and Enyindah explore the use of advanced non-linear modeling techniques such as neural networks and fuzzy logic models to solve the problems associated with making predictions. They propose a Neuro-fuzzy logic model to predict the results of a soccer match. The proposed model consists of two phases; the first phase uses a neural network (NN) to analyze the major factors that affect the results of a match into five categories [20]. These categories include team strategy (such as tactics and playing system), tactical skills (such as passing, control, and shooting), physical abilities (such as aerobatic capacity, speed, and agility), psychological effects (such as motivation and confidence), and the current status (such as nutrition, injuries, and training level) [20]. The feed-forward algorithm was used to calculate the optimal weights of the individual factors that make up the categories. The weights were adjusted to produce new weights as feedback for the input layer [20].

The NN model was trained using the Levenberg-Marquardt back-propagation with each training being improved upon by ensuring the inputs and outputs match [20]. Once the generalization stops rising, training ceases. With a NN fitting tool, training and performance evaluation are aided using mean square error and regression analysis [20]. Multiple training executions result in different outputs because of the different initial conditions and sampling [20].

The result of the neural network is then passed to the second phase which uses a fuzzy logic model to predict the results of the soccer match. The fuzzy logic module comprises four main modules, the fuzzification module, the fuzzy inference engine, the fuzzy rule base, and the defuzzification module [20]. It normally accepts a crisp input value, assesses it, and then maps it into fuzzy membership function values [20]. Note that the fuzzy engine is responsible for calculating the fuzzy membership function values and relating them to the fuzzy rules to determine the best fuzzy output [20]. Meanwhile, the defuzzification module transforms the fuzzy output into a crisp output value which is more suitable for decision-making and control [20].

The authors use a Membership Function Editor to view and edit the membership functions associated with the input and output of the fuzzy inference system [20]. Based on the description from the previous action, the authors use a Rule Editor to graphically construct the rule statements automatically following the if-then criteria [20]. Then the Rule Viewer graphically presents the whole fuzzy inference process which may be used to immediately interpret the complete process [20]. It demonstrated the impact the factors have on the final output as it displays one computation at a time in detail [20]. The authors were able to present that the Neuro-fuzzy logic technique is a viable option for match predictions.

5.4 Fuzzy-Based Model for Predicting Football Match Results

Soccer, like other sports, is a game where luck can affect the outcome of a match [21]. However, unlike other sports, its prediction may vary because weaker teams may defeat their stronger opponents, barely or by a great goal differential [21]. As a result, Omomule and colleagues propose a model that includes fuzzy logic to predict the results of soccer matches with higher prediction accuracy. The authors use the IF-THEN criteria which contains a set of soccer clubs, a set of attributes such as manager profile, players' quality, and weather, assigned to each club, a set of events such as a change in manager, injury, and suspension and the specified time each event occurred [21]. The fuzzy rules range from very low at less than 0.1 to very high measuring greater than or equal to 0.8 but less than or equal to 1 [21].

The fuzzy rules are based on the Sugeno inference system. The authors used a case study using data collected from the English Premier League 2017-2018 season which comprised 20 sample teams [21]. The fuzzy framework included 13 input variables, 60 fuzzy rules, and three linguistic variables (low, average, high) except for the location variable which only accounts for Home and Away [21]. Additionally, the minimum method was used for implication, the maximum method for aggregation, and the wtaver method for defuzzification [21]. The 'wtaver' method refers to the weighted average of all rules output and is the default for Sugeno systems in MATLAB.

The first team in the English Premier League, Manchester United, was used as the simulation for the implementation stage [21]. 13 inputs and 1 output were mapped from their crisp values to linguistic variables using the Gaussian membership function [21]. A loss function is used to calculate the quality of the predictive model by finding the difference between the expected results and the predicted outcome [21]. Ultimately, the authors utilize the Average Testing Error to determine their model performance with the correlation analysis displaying the accuracy of the predictions which amounted to approximately 0.075 and 89.27 % respectively on the predictors that affect the soccer match [21].

6 DISCUSSION

Soccer is a universally popular sport, thus there is more data available for consideration when conducting the win probability prediction [13, 18]. This may negatively impact the experiments as it is time-consuming and tedious to determine the relevant attributes [13]. This was one issue encountered by Hassan [13] and colleagues in the World Cup 2018 study. However, it highlighted the knowledge that attributes should not be excluded from the process but determining the most sensitive ones may even improve the predictions in tandem with locations and goals scored and conceded [13]. Although the more common parameters regard team strengths and goals scored and conceded, the authors believed that these parameters would negatively impact their prediction and thus left them out. They were able to have a success rate of 72% for losses and 82% for wins compared to the statistical models which developed their probability based solely on goals scored versus goals conceded having a success rate of 64% [13, 17].

However, that is not to say that goals scored are an unimportant attribute, in fact, because soccer is a low-scoring game it should be considered. In the study conducted by Parim and colleagues [19] that focused on the UEFA championship league, they mentioned that the number of goals scored in the top four leagues in Europe was 2.66 per game. Goals are not a common action for the sport; thus, it may be more accurate to account for the first goal scored and scoring chances for each team. While the scoring chances may not lead to a goal, it raises the chances of one especially since it indicates the dominant team and the pressured team. By continuously making these chances, the pressured team may succumb to mistakes allowing for a greater winning chance for the dominant team. The authors went on to confirm these findings stating if a strong team faced off against a weak team, they had a 68% chance of winning however if that team went on to score the first goal their chances rose to 92% [19]. Additionally, the winning chances increased by at least 14% for the team that had more shots on goal [19].

As mentioned earlier, strength is one of the more common parameters used to determine a team's win probability. One of the studies that focused on the strength of the home and away teams is the prediction of winning teams using machine learning. However, compared to the other studies, the authors did not consider loss when conducting their experiment [14]. Considering the results of the other studies which include loss, it is assumed that this negatively impacted the predictions of the three machine learning networks used by only having it learn the win and draw patterns of the teams. The highest accuracy received in this study was 63% for SVM [14]. They do indicate that while strength is important it should not be the only parameter used for the predictions [14]. The strengths considered focus on the overall team aspects while the multivariate study conducted on the championship league determines a team's strength compared to the opponent they will face. This led to a more accurate prediction with the assumption being that the team's strong and weak points differ from each other, thus it is not accurate to label a team as strong with no relation to the strengths of the other teams within the league.

It is important to note, that a significant amount of the machine learning studies mentioned had to sieve through thousands of data for training. Additionally, there has been no agreement on the attributes necessary to effectively determine the win probabilities. As such, the use of fuzzy logic in tandem with machine learning should be considered. The fuzzy logic studies had a significant decrease in the number of attributes needed by selecting parameters based on their membership values, and the accuracy was not negatively impacted. In fact, with the use of fuzzy logic, the negative side effects such as overfitting are removed from training and accuracy may increase. This may be witnessed in Hassanniakalager and colleagues' conditional fuzzy-inference study where their hybrid fuzzy-inference and RVM approach improved the accuracy of the underlying RVM system by approximately 10% [15].

Furthermore, some of the highest accuracy seen was with the use of a fuzzy-based system by Omomule and colleagues which amounted to 89.27% [21]. As such, it can be assumed that a hybrid system that incorporates machine learning and fuzzy-based rules may provide the best win probability prediction amongst standard machine learning systems, statistical models, or hybrids.

7 CONCLUSION

Many types of sports activities require physical preparation and exertion to increase the quality of life for both the spectators and players because of the satisfaction it provides. One important sport is Soccer with over 100 national teams worldwide and its viewership reaching billions. As a result of its popularity, it has attracted the attention of researchers in an attempt to find accurate models to predict its win probability.

Because of the availability of data, there is still an issue surrounding the necessary attributes and those that would act as noise in the prediction models. Thus far, it can be determined that the sensitivity of the attributes should be what determines which parameters are used, however, that is a time-consuming process. Additionally, the factor of whether the attributes used should be a comparison, such as a team strength relative to another, or be standalone such as team strength based on their attack and defensive strength.

As such, fuzzy logic can be viewed as an advantageous technique to use among the many artificial intelligence methods introduced as it contains practices that result in higher accuracy and smoother control [15]. Attributes would be chosen based on their membership values, lessening the amount of data needed for the prediction. The use of a combination of fuzzy-based systems and machine learning methods may provide the most accurate predictions while simultaneously removing issues such as overfitting

REFERENCES

[1] V. Sutula, "General Definition of the Concept Sports," Journal of Physical Fitness, Medicine & Treatment in Sports. 4, July 2018

[2] E. Terrell, "Sports Industry: A Research Guide, " US Library of Congress, June 2020.

[3] F. M. Clemente, Special Issue "Data Analytics in Sports Sciences: Changing the Game" (Collection), Entropy, ISSN 1099-4300 24, July 2022.

[4] G. B. Mgaya, H. Liu, and B. Zhang, "A survey on applications of modern deep learning techniques in team sports analytics," Advances in Intelligent Systems and Computing, pp. 434–443, 2021.

[5] A. P. Rotshtein, M. Posner, and A. B. Rakityanskaya, "Football predictions based on a fuzzy model with genetic and neural tuning," Cybernetics and Systems Analysis, vol. 41, no. 4, pp. 619–630, 2005.

[6] A. Jose, M. Philip, P.Tumkur Prasanna, L. Munivenkatappa, M. Munivenkatappa, "Comparison of Probit and Logistic Regression Models in the Analysis of Dichotomous Outcomes". Current Research in Biostatistics. 10. 1-19. 10.3844/amjbsp.2020.1.19., 2020.

[7] A Karanasiou, D. Pinotsis., "Towards a legal definition of machine intelligence: the argument for artificial personhood in the age of deep learning." In Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law, ACM, New York, NY, USA, 119–128. https://doi.org/10.1145/3086512.3086524 June 2017.

[8] R. Levy, Probabilistic models of word order and syntactic discontinuity. Ph.D. Dissertation. Stanford University, Stanford, CA, USA., 2005.

[9] D. Dubois, H. Prade,. Fuzzy logic. Encyclopedia of Computer Science. John Wiley and Sons Ltd., GBR, 739–742., Jan. 2003.

[10] H. F. J. Ho, Y. K. Wong, A. B. Rad, "Novel adaptive control algorithms via soft computing methods.", Ph.D. Dissertation. Hong Kong Polytechnic University (People's Republic of China), 2005.

[11] N. Chmait and H. Westerbeek, "Artificial Intelligence and machine learning in sport research: An introduction for non-data scientists," Frontiers in Sports and Active Living, vol. 3, 2021. doi:10.3389/fspor.2021.682287

[12] M. K. Langaroudi and M. R. Yamghani, "Sports Result Prediction Based on Machine Learning and Computational Intelligence Approaches: A Survey," Journal of Advances in Computer Engineering and Technology, vol. 5, no. 1, pp. 27–36, Feb. 2019.

[13] A. Hassan, A.-R. Akl, I. Hassan, and C. Sunderland, "Predicting wins, losses and attributes' sensitivities in the Soccer World Cup 2018 using neural network analysis," Sensors, vol. 20, no. 11, p. 3213, 2020.

[14] Y. Ajgaonkar, A. Patil, K. Bhoyar, and J. Shah, "Prediction of Winning Team using Machine Learning," International Journal of Engineering Research & Technology (IJERT), vol. 09, no. 3, pp. 461–466, Feb. 2021.

[15] A. Hassanniakalager, G. Sermpinis, C. Stasinakis, and T. Verousis, "A conditional fuzzy inference approach in forecasting," European Journal of Operational Research, vol. 283, no. 1, pp. 196–216, 2020.

[16] L. A. Zadeh, "Fuzzy logic," Computer, vol. 21, no. 4, pp. 83–93, 1988.

[17] R. Kissell, "Chapter 9 Soccer - MLS," in Optimal Sports Math, statistics, and Fantasy, London: Academic Press, 2017, pp. 229–252.

[18] H. Arntzen and L. M. Hvattum, "Predicting Match Outcomes in association football using team ratings and player ratings," Statistical Modelling, vol. 21, no. 5, pp. 449–470, 2020.

[19] C. Parim, M. Ş. Güneş, A. H. Büyüklü, and D. Yıldız, "Prediction of match outcomes with Multivariate Statistical Methods for the group stage in the UEFA Champions League," Journal of Human Kinetics, vol. 79, no. 1, pp. 197–209, 2021.

[20] U. C. Onwuachu and P. Enyindah, "A neuro-fuzzy logic model application for predicting the result of a football match," European Journal of Electrical Engineering and Computer Science, vol. 6, no. 1, pp. 60–65, 2022.

[21] Omomule T.G., Ibinuolapo A. J., Ajayi O.O., "Fuzzy-Based Model for Predicting Football Match Results," International Journal of Scientific Research in Computer Science and Engineering, Vol.8, Issue.1, pp.70-80, 2020.