Journal of Artificial Intelligence Research 73 (2022) 1285-1322 Submitted 11/2021; published 04/2022

The Application of Machine Learning Techniques for Predicting Match Results in Team Sport: A Review

Rory Bunker RORY.BUNKER@G.SP.M.IS.NAGOYA-U.AC.JP
Graduate School of Informatics
Nagoya University
Furocho, Chikusa Ward
Nagoya, Aichi 464-8601, Japan

Teo Susnjak T.SUSNJAK@MASSEY.AC.NZ
School of Mathematical and Computational Sciences
Massey University
Massey University East Precinct Albany Expressway, SH17, Albany
Auckland 0632, New Zealand

Abstract

Predicting the results of matches in sport is a challenging and interesting task. In this paper, we review a selection of studies from 1996 to 2019 that used machine learning for predicting match results in team sport. Considering both invasion sports and striking/fielding sports, we discuss commonly applied machine learning algorithms, as well as common approaches related to data and evaluation. Our study considers accuracies that have been achieved across different sports, and explores whether evidence exists to support the notion that outcomes of some sports may be inherently more difficult to predict. We also uncover common themes of future research directions and propose recommendations for future researchers. Although there remains a lack of benchmark datasets (apart from in soccer), and the differences between sports, datasets and features makes between-study comparisons difficult, as we discuss, it is possible to evaluate accuracy performance in other ways. Artificial Neural Networks were commonly applied in early studies, however, our findings suggest that a range of models should instead be compared. Selecting and engineering an appropriate feature set appears to be more important than having a large number of instances. For feature selection, we see potential for greater inter-disciplinary collaboration between sport performance analysis, a sub-discipline of sport science, and machine learning.

1. Introduction

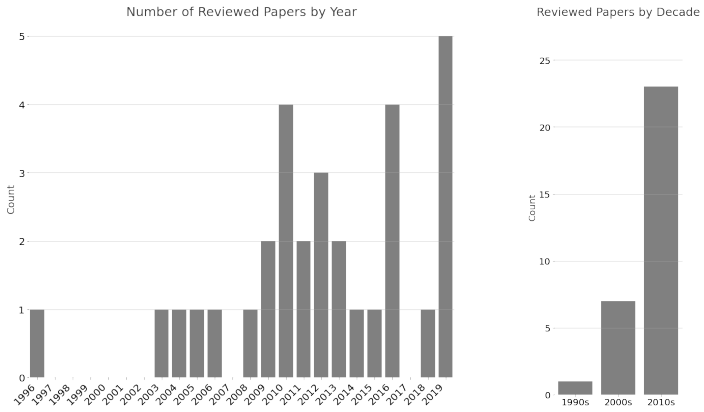
Sport result prediction is an interesting and challenging problem due to the inherently unpredictable nature of sport, and the seemingly endless number of potential factors that can affect results. Indeed, it is this unpredictable nature that is one of the main reasons that people enjoy sport. Despite its difficulty, predicting the results of sports matches is of significant interest to many different stakeholders, including bookmakers, bettors, and fans. Interest in match result prediction has grown with the increased availability of sport-related data online and with the emergence of online sports betting. Sport experts and former players often make predictions on upcoming matches, which are commonly published in the media.

The application of machine learning (ML) in sport has recently branched widely, and research surveys exist on predicting sports injuries (Van Eetvelde et al., 2021), applying predictive analytics to sport training (Rajšp & Fister, 2020), and using predictions for determining optimal team formations (Ishi & Patil, 2021). Numerous studies also exist that focus on in-play predictions, which predict the occurrence of specific events during a contest, e.g., a goal being scored by a specific player. However, the focus of the present survey is specifically on the application of ML in predicting the outcomes of matches in team sports. The prediction of final match results is of interest not only to the likes of bookmakers and bettors but also to players, team management and performance analysts in order to identify the most important factors in achieving winning outcomes. One important task within outcome prediction in sport is to select the best set of features for the predictive model. If the model used is interpretable and incorporates some feature selection mechanisms or, alternatively, a feature selection method is applied prior to applying the model, the most important predictive features can be extracted. Although some features, e.g., the match venue, officials, weather, etc., are external to the sport match, in-play features may identify areas in which teams can adjust their tactics/strategy to improve performance. Sport performance analysis, a sub-discipline of sport science, considers performance indicators (PIs), which are a selection or combination of action variables that aim to define some or all aspects of a performance, which, in order to to be useful, should relate to a successful performance or outcome (Hughes & Bartlett, 2002). In this context, the match outcome is, of course, one measure of performance. Therefore, important features (PIs) are useful for players, team management and sport performance analysts to identify ways in which they can improve their performance and achieve winning outcomes. As we will discuss further below, in this domain we see potential for greater collaboration between sport performance analysis researchers and machine learning researchers going forward.

In the academic literature, match result prediction in sport has been considered by the statistics and operations research communities for some time; however, the application of ML techniques for this purpose is more recent. The first study in this domain appears to have been published in 1996 (Purucker, 1996); however, it was not until the 2010s when research activity in this area intensified, as shown in Figure 1(b). This review considers research papers in this field over the last three decades, but places an inclusion constraint requiring that papers must have made use of at least one ML technique. To that end, knowledge-based systems (e.g., fuzzy logic- and rule-based systems), and ratings methodologies (e.g., Elo ratings) (Rotshtein et al., 2005; Tsakonas et al., 2002; Min et al., 2008) were not considered to be in-scope. Due to the large number of papers that have been published recently, we needed to constrain the scope of this review further to only focus on the prediction of *match results* in *team sports*. Thus, we did not include studies related to sports played by individuals or in pairs, e.g., horse racing (Davoodi & Khanteymoori, 2010), swimming (Edelmann-Nusser et al., 2002), golf (Wiseman, 2016), tennis (Somboonphokkaphan & Phimoltares, 2009) and javelin (Maszczyk et al., 2014). Although this review has a narrower scope compared to some prior reviews (Table 1), its contribution lies in providing a more in-depth analysis of the application of ML for sport result prediction in team sports than the previous survey articles. The present review introduces the reader to a broad overview of approaches in which ML techniques have been applied for match result prediction in team sports. A distinct contribution of this work, in comparison to prior

reviews, lies in our discussion of how the characteristics of particular team sports potentially play a role in the ability of ML to be able to accurately predict match results. For instance, we explore whether the predictability of matches may depend on whether it is an invasion (time-dependent) sport or a striking/fielding (innings-dependent) sport. We also discuss how the occurrence and increments of points or goals may also affect the predictability of match results. Furthermore, we comment on what some of the key drivers of successful studies have been with respect to how ML has been applied, and how those elements have contributed to higher predictive accuracies. In addition, this study makes a meaningful contribution in identifying future research trends and opportunities in this field, while combining the findings into a useful set of recommendations for other professionals in this domain. Lastly, as mentioned, we also highlight the possibility for greater collaboration between researchers from sport performance analysis and those from machine learning. The lack of collaboration between these fields has often resulted in unclear and inconsistent terminology, and there are significant opportunities to advance research with a more unified approach.

The remainder of this paper is structured as follows. In Section 2, we outline our methodological approach and inclusion criteria for this study. Then, in Section 3, we review the literature in ML for match result prediction in team sport, categorizing sports by type (invasion sports and striking/fielding sports) and further sub-categorizing studies based on the individual sport. We also present tabular summaries of the results of the surveyed studies by sport. Following this, we provide critical analysis and discussion in Section 4, before concluding in Section 5.



(a) Papers by year

(b) Papers by decade

Figure 1: Figures depicting the number of papers surveyed by (a) year and (b) decade.

2. Methodology

In this section, we firstly outline our inclusion criteria, which specifies the scope of this review. Secondly, we categorize the algorithms into groups in order to identify trends in usage over time. Thirdly, we describe the measure of accuracy, which has been (by far) the most widely-used metric of performance in this application domain, and we also discuss

why it is indeed an appropriate metric for match result prediction in team sport. Finally, we describe recurring future research themes extracted from the surveyed studies.

2.1 Inclusion Criteria

We considered studies that used ML to predict match results in team sports, which were published between 1996 and 2019 (inclusive). To be included in this review, a study must have applied at least one ML technique. Thus, studies that only considered fuzzy logic-, ratings- (e.g., Elo) or statistical-based methods were not included.

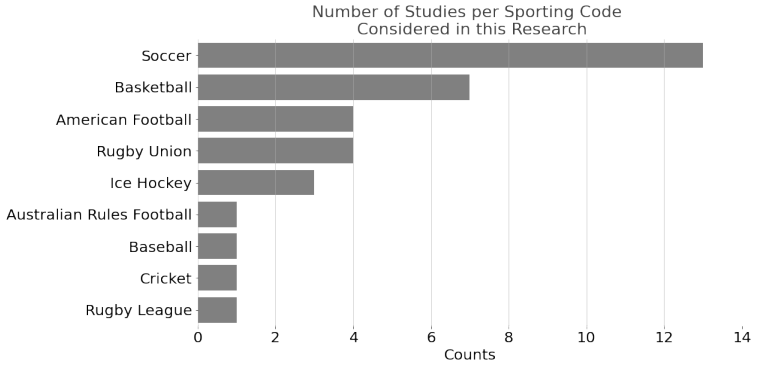


Figure 2: Number of surveyed studies by sporting code

A total of 31 papers were considered in this review. The distribution of papers by year and decade of publication can be seen in Figure 1, which shows an explosion in the level of interest in this field from 2009 onward. The 31 papers covered nine distinct sports and, given that some papers studied multiple sports, there were a total of 35 sport studies. The distribution of the number of studies by sport is shown in Figure 2.

We identified four related review papers that were published in the same time period that we considered. Table 1 shows the methods applied, the types of sports considered, and what the applied methods covered were aiming to achieve. Taken together, we consider these aspects to define the scopes of the prior survey articles. Both Haghighat et al. (2013) and Keshtkar Langaroudi and Yamaghani (2019) included studies that used ML as well as knowledge-based systems (e.g., systems based on fuzzy and rule-based logic). Haghighat et al. (2013) covered both team and non-team sports, focusing on match result prediction specifically, while Keshtkar Langaroudi and Yamaghani (2019) also included studies that considered tactical decision making. The review of Razali et al. (2018) had a narrower scope compared to other reviews (and compared to the present review), and focused on studies applying one specific ML model, a Bayesian Network (BN), for predicting the match results of one sport, soccer. The survey of Beal et al. (2019) had a wide scope, considering artificial intelligence methods (of which machine learning, data mining, and knowledge-based methods are sub-categories) for match result prediction, as well as tactical decision making, player investments, and injury prediction. Beal et al. (2019) also considered fantasy sports in addition to team sports.

Unlike prior work, our review provides more up-to-date coverage and draws out more critical insights compared to previous studies, e.g., Haghighat et al. (2013), who did not, for

Survey	Methods	Sports	Purpose
Haghighat et al. (2013)	Machine Learning & Knowledge-based Systems	Team Sports & Non-Team Sports	Match result prediction
Razali et al. (2018)	Bayesian Networks	Soccer	Match result prediction
Beal et al. (2019)	Artificial Intelligence	Team Sports & Fantasy Sports	Match result prediction, tactical decision making, player investments, injury prediction
Keshtkar Langaroudi and Yamaghani (2019)	Machine Learning & Knowledge-based Systems	Team & Non-Team Sports	Match result prediction, tactical behavior
This review	Machine Learning	Team Sports	Match result prediction

Table 1: Comparison of the scope of this review with that of similar surveys covering a similar time period

instance, discuss the characteristics of the sports that they considered, and how these may have had an impact on obtained accuracies. They also did not consider whether accuracies obtained have improved over time, nor did they discuss the appropriateness of different models for the purpose of sport result prediction, e.g., in terms of their interpretability.

For this paper, we sought to define the scope of the survey such that it is sufficiently narrow so that the analysis is adequately in-depth but, at the same time, its scope is not too wide such that the number of surveyed papers is unmanageable. In this review, we focus on match result prediction, which is a team-level measure of performance. The scope of Keshtkar Langaroudi and Yamaghani (2019) was somewhat ill-defined in that they also included, e.g., coverage of Tilp and Schrapf (2015), who considered the use of ANNs for analyzing tactical defensive behavior in handball, which is not related to the prediction of player- or team-level results. Although narrower in scope, our review provides a wider coverage of surveyed papers than Keshtkar Langaroudi and Yamaghani (2019) and Haghighat et al. (2013), surveying more than double and triple the number of studies, respectively. In addition, unlike the above-mentioned review papers, we cover the studies that arose from the 2017 Open International Soccer Database Competition, which was important in terms of creating a benchmark dataset for soccer match result prediction (the review of Haghighat et al., 2013 noted that there were no such benchmark datasets at the time of their review).

2.2 Algorithm Grouping

ML algorithms were grouped into families of algorithms to identify trends in their usage patterns. All variants of ANN, such as Back Propagation (BP), Feed-Forward, as well as Self-Organising Maps and Long Short-Term Memory (LSTM) ANNs, were grouped under the same umbrella of algorithms. CART, C4.5 and WEKA's J48 algorithm were grouped under the Decision Tree family of methods. RIPPER, FURIA and ZeroR were merged into the Rule Sets category, while the Local Weighted Learning (LWL) algorithm was merged with k-Nearest-Neighbors (kNN). Additionally, AdaBoost, XGBoost, LogitBoost, RobustBoost, and RDN-Boost were grouped in the Boosting category, while Sequential Minimal Optimization (SMO) (Kohonen, 1990) was combined with Support Vector Machines (SVMs). Methods that combined several different families of ML algorithms into a single decision-making architecture were termed Ensemble. Although Naïve Bayes and BNs share some common theoretical foundations, a decision was made to keep them separate since the latter includes the ability to incorporate domain knowledge to a larger degree, which was a

motivating factor for its usage in some studies. A histogram depicting the usage patterns of these algorithm groups, sorted according to their frequency of use, can be seen in Figure 3.

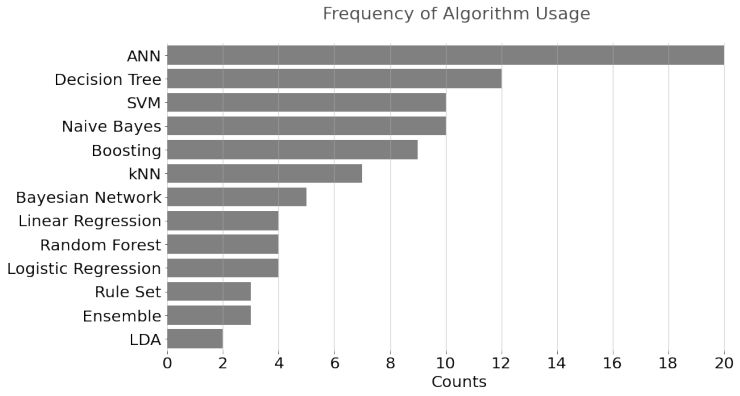


Figure 3: Histogram of usage patterns of the algorithm groups covered in this survey

2.3 Performance Evaluation Metrics

To allow for meaningful between-study comparisons and interpretations, we consider the accuracy measure as the primary evaluation metric, which the vast majority of surveyed studies used. Accuracy is defined as the number of correct predictions divided by the total number of predictions or, equivalently, the proportion of the sum of total true positive and true negative classifications divided by all samples in the dataset. Although the charts presented in the remainder of this paper only include studies that reported accuracy as the measure of performance, all results, including one paper that reported balanced accuracy and three that reported the average ranked probability score (RPS), are presented in the results summary tables in Section 3. It should be noted that the results from binary-class and multi-class problems are not directly comparable since a given level of accuracy becomes harder to achieve as the number of classes increases. Thus, the accuracies of studies that used a three-class formulation instead of a two-class formulation are excluded from the charts for comparative purposes but, again, are reported in the summary tables.

2.4 Future Research Themes

In order to draw out insights across all papers in terms of what the general future research direction trends might be, a set of recurring general themes were first extracted from all of the papers. Subsequently, all of the text referring to future research directions across all of the papers was encoded based on the extracted themes, and a histogram was rendered that depicts the frequency of each theme.

3. Literature Review

In this section, we review the studies identified by our inclusion criteria, distinguishing between invasion sports and striking/fielding sports (Figure 4). Subsection 3.1 covers studies

related to invasion sports, including American Football, Rugby Union, Soccer, Basketball and Ice Hockey, and Subsection 3.2 considers the striking/fielding sports of Baseball and Cricket. The results of the surveyed studies are also summarized in tables containing the competition, models applied, the number of matches in the original dataset, as well as the best performing model and the number of features used in that model.

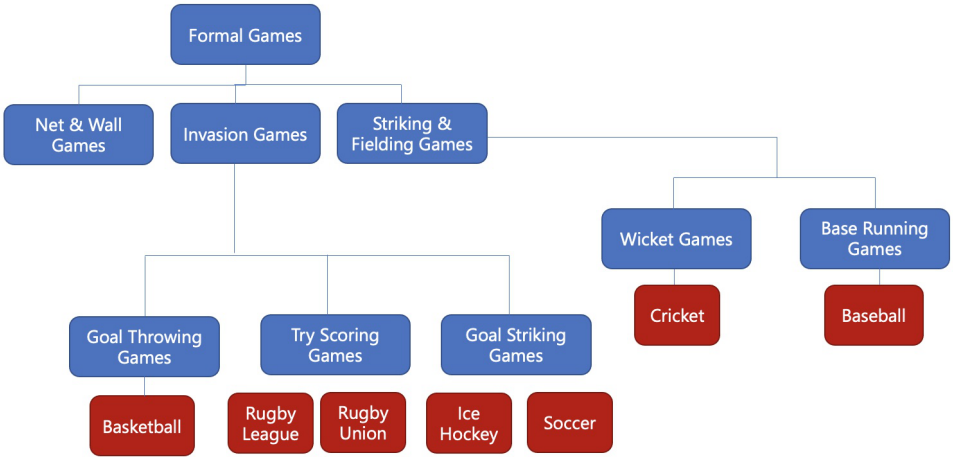


Figure 4: Classification of formal games in this survey based on the categorizations provided by Read and Edwards (1992) and Hughes and Bartlett (2002).

3.1 Invasion Sports

Invasion sports are time dependent in that they have matches with fixed periods of time, commonly divided into halves or quarters. The aim for teams in invasion sports is to move into the opposition team's territory by maintaining possession, creating space, and attacking a goal or target in order to convert scoring opportunities into points while, at the same time, defending their own space and goal to prevent the opposition from scoring (Mitchell, 1996).

3.1.1 American Football

Purucker (1996) used an ANN and unsupervised learning techniques to predict the results of National Football League (NFL) football matches, using data from 90 matches from weeks 11 to 16 of the 1994 NFL competition. Six features were included: victories, yardage differential, rushing yardage differential, turnover margin, time in possession, and betting odds (the inclusion of betting line odds improved upon initial results). An ANN trained with BP was used, with BP providing the best performance among the different network training methods. The unsupervised learning methods that were applied were: the Hamming error (Hamming, 1950), Adaptive Resonance Theory (ART) (Carpenter & Grossberg, 2003), and Self-Organizing Maps (SOM). The SOM provided the best performance among the unsupervised methods, however, it could not match the performance of the ANN. Matches from weeks 12 to 15 were used to predict week 16 matches, with the ANN correctly predicting 11 out of the 14 matches (78.6%) in week 16. Weeks 12 to 14 were also used to predict week

15, with the ANN correctly predicting 10 of the 14 matches (71.4%) in week 15. The author recognized that the dataset consisted of a small number of matches and features, and mentioned that improvements could be gained by fine-tuning the encoding, ANN architecture, and training methods.

Kahn (2003) predicted NFL matches using data from 208 matches in the 2003 season. A BP-trained ANN was used, and the features included were: total yardage differential, rushing yardage differential, time in possession differential, turnover differential, a home or away indicator, home team outcome, and away team outcome. The numeric features were calculated based on the 3-week historical average (the feature's average value over the past 3 weeks), as well as its average value over the entire season. Using the average over the entire season achieved higher accuracy. Weeks 1 to 13 of the 2003 competition were used as training data, and weeks 14 and 15 as the test set. Accuracy of 75% was achieved, which was slightly better than expert predictions on the same matches. It was suggested that, in future work, betting odds and team rankings could be included as features, and matches from previous seasons could be used for model training.

David et al. (2011) used a committees-of-committees approach with ANNs to predict NFL matches, where many networks were trained on different random partitions of the data. For training, 500 ANNs were used, and the best 100 were used in each committee, of which 50 were ultimately used. The mean was used to determine the vote of each committee and to combine their predictions. The features used were differentials between the home and away teams based on: passing yards, yards per rush play, points, interceptions, and fumbles. The differentials between the home and away teams were to incorporate the well-known home advantage phenomenon. The season average of these features were used, apart from the first five rounds of the season, where the weighted average between the current season's and previous season's features was used. In particular, 100% of the feature value from the previous season was used for week 1, then 80% of the previous season and 20% of the current season in week 2, and so on until week 6, at which point only the current season value was used. A total of 11 inputs were used in the ANN for each game, and the Levenber-Marquadt (LM) routine was used to train the network. Principal Components (PCA) and derivative based analyses were applied to determine which features were most influential. It was found that 99.7% of the variance in their data was due to the passing and rushing differentials. The results were compared to bookmakers and predictions from thepredictiontracker.com, and were found to be comparable to the predictions of bookmakers, and better than most of the online predictions. The avenues for future work were to use different types of ANN, e.g., RBF, to include additional statistics (e.g., possession, strength-of-schedule, kicking game and injuries), to investigate how to best predict games that are early in the season, and to apply the method to other levels of football (e.g., NCAA) as well as to other sports.

Delen et al. (2012) used an SVM, CART Decision Tree and ANN to predict the results of NCAA Bowl American Football matches. The Cross-Industry Standard Process for Data Mining (CRISP-DM) framework (Wirth & Hipp, 2000) was used as the experimental approach, and the dataset contained 28 features and 244 matches. A classification approach that predicted home win/away win was compared with a numeric prediction approach that predicted points margin (home team points minus away team points). CART provided the best performance, with 86% accuracy (using 10-fold cross validation), which was statistically significantly better than the other models. The classification approach produced

Paper	Models used	No. of features	No. of Matches	Accuracy of best model
Purucker (1996)	ANN trained with BP*, Unsupervised: Hamming, ART, SOM	6	90	75% (week 14 and 15 combined)
Kahn (2003)	ANN trained with BP	10	208	75%
David et al. (2011)	Committees of ANNs trained with LM	11	unknown	unknown
Delen et al. (2012)	SVM, CART*, ANN	28	244	86%

Table 2: American Football Studies (* denotes the best performing model).

better results than numeric prediction. When the models were trained on the 2002/2003 to 2009/2010 seasons and then tested on the 2010-2011 season, CART again had the best performance, achieving 82.9% accuracy. The suggested directions for future research were to include more variables or represent them in different forms, to make use of other classification and regression methods (e.g., rough sets, genetic algorithm based classifiers, ensemble models), to experiment with seasonal game predictions by combining static and time series variables, and to apply the approach to other sports.

3.1.2 Rugby Union

O'Donoghue and Williams (2004) compared the predictive ability of human experts with computer-based methods for 2003 Rugby World Cup matches. Multiple Linear Regression, Logistic Regression, an ANN, and a Simulation Model were the computer-based models used. Data from the previous four world cups were used to develop predictive models of match results based on three variables: team strength, determined by synthesising world rankings (actual world rankings had not yet been introduced at the time of the 2003 Rugby World Cup), distance travelled to the tournament as a measure of home advantage, and the number of recovery days between matches. The computer-based models correctly predicted between 39 and 44.5 of the 48 matches, while the 42 experts correctly predicted an average of 40.7 matches. However, there was far greater variation in the accuracies of the experts, with the most successful person correctly predicting 46 of the 48 matches. The most accurate computer-based model was the Simulation Model.

O'Donoghue et al. (2016) compared the accuracy of 12 Linear Regression models in predicting match results at the 2015 Rugby World Cup. The models differed in terms of: (i) whether or not the assumptions of the predictive modeling technique were satisfied or violated, (ii) whether all (1987-2011) or only recent Rugby World Cup tournaments' (2003-2011) data were used, (iii) whether the models combined pool and knockout stage match data, and (iv) whether the models included a variable that tried to capture a relative home advantage. The common independent variable in all models was the relative quality, which was the difference from the higher ranked team's world ranking points. The dependent variable was the points margin. All models were executed 10,000 times within a simulation package that introduced random variability. The best model achieved accuracy of 74% and, notably, match outcomes in international Rugby appeared to be more difficult to predict

Paper	Competition	Models used	No. of features	No. of Matches	Accuracy of best model
O'Donoghue and Williams (2004)	2003 Rugby World Cup	Multiple linear regression, Logistic Regression, ANN, Simulation model*	3	48	93%
Reed and O'Donoghue (2005)	English Premiership Rugby	Multiple linear regression, ANN*, discriminant function analysis	7	498	46.1%
McCabe and Trevathan (2008)	Super Rugby	ANN trained with BP* and CGD	19	unknown	74.3%
O'Donoghue et al. (2016)	2015 Rugby World Cup	Linear regression*, Simulation	1	280	74.4%

Table 3: Rugby Union Studies (*denotes the best performing model).

than in previous years. The best model used data from all previous Rugby World Cups in a way that violated the assumptions of Linear Regression, using only one independent variable and ignoring the relative home advantage, while generating separate models for the pool and knockout stage matches.

3.1.3 Soccer

Joseph et al. (2006) found that incorporating expert knowledge into a BN model can result in strong performance, especially when the sample size is small. A decision tree (MC4) and kNN model were also used to predict the results of Soccer matches played by the English Premier League (EPL) team, Tottenham. Their dataset consisted of 76 matches. Four variables were included in the expert model, while 30 variables were used in their original, general model. The expert BN was found to provide the best performance, achieving 59.2% accuracy when predicting a home win, away win or draw (a 3-class problem). In the future, the authors proposed to develop a more symmetrical BN model using similar data but for all the teams in the EPL, to incorporate player-quality features (e.g., players that have performed at international level), and to add additional nodes such as attack quality and defence quality.

Buursma (2010) used data from 15 years of Dutch Soccer to predict match results, and was interested in: which variables were important for predicting match results, how the probabilities for the match results can be predicted from these variables, and what bets should be placed to maximize profit. The author applied the following models in WEKA: Classification via Regression, Multi-class Classifier, Rotation Forest, Logit Boost, BN, Naïve Bayes, and ZeroR. There were three match outcomes to be predicted: home win, draw, and away win. The feature set consisted of 11 features, and all features were either aggregated or averaged across a team's previous 20 matches (by experimentation, 20 was found to be the best number of matches to average across). Classification via Regression and the Multi-class Classifier had the best prediction accuracy, both achieving 55%. For future work, the author considered including more features, e.g., yellow/red cards, the number of players each team has, the management, player budgets and home ground capacities, and was also interested in applying the model to other similar sports such as basketball, baseball and ice hockey.

Huang and Chang (2010) used a BP-trained ANN to predict the results of the 64 matches in the 2006 Soccer World Cup tournament. The output of the ANN was the relative ratio between two teams for each match, which could then be converted into an odds ratio. Eight features were selected based on the domain knowledge of the authors: goals for, shots,

shots on goal, corner kicks, direct free kicks on goal, indirect free kicks on goal, possession, and fouls conceded. Accuracy of 76.9% was achieved based on 2-class prediction, i.e., not including draws. The ANN was found to have difficulty predicting a draw, which was a frequent outcome in the group stages.

In contrast to Joseph et al. (2006), Hucaljuk and Rakipović (2011) found that incorporating expert opinion did not produce any improvement for soccer match result prediction. Their dataset consisted of 96 matches (6 rounds) in the regular part of the European Champions League competition, and was split into three different training and test datasets: a 3 round-3 round training-test split, a 4 round-2 round split, and a 5 round-1 round split. Feature selection resulted in 20 features in their basic feature set. An additional feature set consisted of the 20 basic features plus variables selected by experts. Naïve Bayes, BNs, Logit Boost, kNN, a Random Forest, and a BP-trained ANN were compared, with the ANN performing the best, achieving accuracy of 68%. Perhaps surprisingly, the expert-selected features were not found to yield any improvement. It was mentioned that further improvements could be gained by refining the feature selection, modeling player form, and obtaining a larger dataset.

Odachowski and Grekow (2012) analysed fluctuations in betting odds and used ML techniques to investigate the value in using such data for predicting soccer matches. The odds for home win, away win and draw for the preceding 10 hours, measured at 10-minute intervals, were tracked over time (the data was obtained from the Betfair and Pinnacle Sports websites). A total of 32 features were computed from this time-series, e.g., maximum and minimum changes in betting odds, overall changes in odds, and standard deviations. The 10-hour period was divided into thirds and the features were also calculated based on these sampling periods. The authors balanced their dataset such that there were an equal number of home wins, away wins, and draws (372 matches in each class). Six classification algorithms from WEKA were compared: BN, SMO, LWL, Ensemble Selection, Simple CART, and a Decision Table, and feature selection methods in WEKA were applied. It was found that draws were especially difficult to correctly predict, with only around 46% accuracy obtained when attempting 3-class classification. However, accuracy of around 70% was obtained when ignoring draws. Discretization and feature selection methods were found to improve results. The authors suggested that additional features describing changes in betting odds could be included in future work.

Tax and Joustra (2015) aimed to identify easily retrievable features that have utility for predicting soccer match results. In particular, they sought to explore whether there were different levels of utility between features broadly classified as public match data sources and those derived from bookmaker odds. Their experiments covered data from the 2000 to 2013 seasons of the Dutch Eredivisie competition. The researchers conducted numerous experiments comparing the Naïve Bayes, LogitBoost, ANN, Random Forest, CHIRP, FURIA, DTNB, J48 and Hyper Pipes algorithms from the WEKA toolkit. The experiments were performed in conjunction with feature selection methods such as PCA, Sequential Forward Selection and ReliefF. The best results on the public match data sources were achieved by the Naïve Bayes and ANN classifiers in combination with PCA, achieving accuracy of 54.7%. FURIA achieved the highest accuracy using bookmaker odds features with 55.3%; however, this was ultimately not found to be statistically significant. A marginal improvement in accuracy was realized with LogitBoost and ReliefF when both bookmaker odds and public

match-data features were used, producing an accuracy of 56.1%. While this was also not at a statistically significant level, it nonetheless pointed to the potential utility in combining a broader variety of features for further investigation.

Prasetio (2016) used Logistic Regression to predict EPL soccer results in the 2015/2016 season. Their data set consisted of six seasons, from the 2010/2011 to the 2015/2016 seasons (2,280 matches). Home offense, away offense, home defence, and away defence were used as the input features, and it was found that home defence and away defence were significant (the authors did not mention how these offense and defence ratings were constructed). Despite this, the model that included all four variables was found to yield higher accuracy. Four different training-test splits were trialed, producing four different sets of model coefficients. The best performing model achieved 69.5% accuracy based on a 2-class problem (excluding draws). In the future, they remarked that the results could be used to assist management with game strategy, or the trained models could be turned into a recommendation system for evaluating player purchase decisions.

Danisik et al. (2018) applied a Long Short-Term Memory (LSTM) neural network (Hochreiter & Schmidhuber, 1997) to predict match results in a number of soccer leagues. Classification, numeric prediction and dense approaches were compared, and were contrasted with an average random guess, bookmakers' predictions, and the most common class outcome (home win). Player-level data were included, which was obtained from FIFA video games. Incorporating player-level and match history data, a total of 139 features were included (134 in the dense model). Four EPL seasons, 2011/2012 to 2015/2016, were considered, comprising a total of 1,520 matches. The average accuracy obtained for a 3-class classification problem was 52.5%, achieved with the LSTM Regression model, using the 2011/2012, 2012/2013 and 2015/2016 seasons as the training dataset and the 2013/2014 season as the validation dataset. The accuracy obtained for a two-class problem (excluding draws) was 70.2%. It was stated that betting odds and additional match-specific player and team features could be included, and the use of convolution to transform input attributes during training and a deeper exploration of the ability of LSTMs to leverage features like specific tactics could be investigated.

The Open International Soccer Database: Prediction Challenge. In 2017, a significant development took place for ML researchers interested in Soccer. A comprehensive, open-source database called the Open International Soccer Database (Dubitzky et al., 2019) was compiled and made public. The database contains over 216,000 matches from 52 leagues and 35 countries. The motivation behind this project was to encourage ML research in Soccer by building an up-to-date knowledge base that can be used on an ongoing basis for the prediction of real-world soccer match outcomes, as well as to act as a benchmark dataset that makes comparisons between experiments more robust. In order to maximize the utility of this database, a deliberate design choice was made to collect and integrate only data that are readily available for most soccer leagues worldwide, including lower leagues. The consequence of this is that the database lacks fields that are highly-specialized and sophisticated. Subsequent to its creation, the 2017 Soccer Prediction Challenge was conducted, and a competition was held based on this dataset, the results of which were published in a special issue of the *Machine Learning* (Springer) Journal. The challenge involved building a single model to predict 206 future match results from 26 different Soccer leagues, which were to be played between March 31 and April 9, 2017. Unlike most prior studies, which used accuracy

as a performance metric, this competition used the average ranked probability score (RPS), which measures how good forecasts are compared to observed outcomes when the forecasts are expressed as probability distributions. In the remainder of this section, we summarize three of the ML-related papers from this competition.

Hubáček et al. (2019b) experimented with both relational- and feature-based methods to learn predictive models from the database. Pi-ratings (Constantinou et al., 2012), which capture both the current form and historical strengths of teams, and a rating based on PageRank (Page et al., 1999) were computed for each team. XGBoost (regression and classification) algorithms were employed as the feature-based method, and RDN-Boost was used as the relational method. The feature-based classification method with XGBoost performed best on both the validation set and the unseen challenge test set, achieving 52.4% accuracy on the test set. Avenues for future work were to augment the feature set, to weight aggregated data by recency, to include expert guidance in forming relational concepts (e.g., using active learning), and to identify features that are conditionally important given occurrences of certain features at higher levels of the tree.

Constantinou (2019) created a model combining dynamic ratings with a Hybrid BN. The rating system was partly based on the pi-rating system of Constantinou and Fenton (2013), computing a rating that captures the strength of a team relative to other teams in a competition. Pi-ratings update based on how the actual goal differences in a match compare to the expected result (based on existing ratings). The rating calculation involves placing more emphasis on the result rather than the margin, so the effect of large goal differences are dampened. Unlike the original pi-ratings, this version also incorporated a team form factor, searching for continued over- or under-performance. Four ratings features (two for the home and away teams) were used as the BN inputs. The model provided empirical evidence that a model can make good predictions for a match between two specific teams, even when the prediction was based on historical match data that involved neither of those two teams. The model achieved accuracy of 51.5% on the challenge test data set. The author recognized the limited nature of this data set, and mentioned that incorporating other key factors or expert knowledge, e.g., player transfers, key player availability, international competition participation, management, injuries, attack/defence ratings, and team motivation/psychology may be beneficial.

Berrar et al. (2019) developed two methods, with two feature sets for result prediction: recency and rating features. Recency feature extraction involved calculating the averages of features over the previous nine matches, based on four feature groups: attacking strength, defensive strength, home advantage and opposition strength. Rating features were based on the performance ratings of each team, and were updated after each match based on the expected and observed match results and the pre-match ratings of each team. The XGBoost and kNN algorithms were applied to each of these two feature sets, with both performing better on the rating feature set. The best performance overall was obtained with XGBoost on the rating features, although this result was achieved post-competition. Their best model that was submitted for the competition was a kNN that was applied to the rating features, which achieved accuracy of 51.9% on the unseen competition test set. It was mentioned that the generally small number of goals per match and narrow margins of victory in soccer meant that it is difficult to make predictions based on goals only. The authors concluded that innovative feature engineering approaches and effective

incorporation of domain knowledge are critical for sport result prediction modeling, and are likely to be more important than the choice of ML algorithm. It was again recognized that the competition dataset was limited, and that data about various game events (e.g., yellow and red cards, fouls, possession, passing and running rates, etc.), players (e.g., income, age, physical condition) and teams or team components (e.g., average height, attack running rate) would likely help to improve results.

Overall, the Open International Soccer Database competition produced a number of innovative methods and approaches. Notably, researchers commonly combined some form of ratings-based method with ML techniques. Despite having access to a very large number of matches available in the competition dataset, all of the studies found that accuracy levelled off after a certain point, which perhaps indicates that having a broad range of predictive features is critical for predicting match results in sport. As mentioned, a significant weakness of the competition dataset is that it does not contain in-play features that occur during matches.

3.1.4 Basketball

Loeffelholz et al. (2009) predicted National Basketball Association (NBA) match results in the 2007/2008 season. There were 620 matches that were used for training and testing, and 30 that were used as the validation set, treated as "un-played" games. The features included were: field goal percentage, three-point percentage, free-throw percentage, offensive rebounds, defensive rebounds, assists, steals, blocks, turnovers, personal fouls, and points. To predict the un-played games, averages of features in the current season were found to result in better performance than averaging the features across the past five matches. The authors also investigated ANN fusion using Bayesian Belief Networks and Neural Networks. Four different types of ANN were applied: a Feed-Forward NN (FFNN), a Radial Basis Function NN (RBF-NN), a Probabilistic NN (PNN) and a Generalized NN (GRNN). The best model, a FFNN with four shooting variables, correctly predicted the winning team 74.3% of the time on average, which was better than USA Today, who achieved 68.7%. An iterative Signal-to-Noise Ratio (SNR) method was used for feature selection, selecting four out of the 22 original variables. Although fusion did not result in higher accuracy on this dataset, the authors mentioned that it still warranted further investigation. They also mentioned that different features could be used in the baseline model, and the models could be adjusted to determine whether they can beat the betting odds rather than only predicting the winning team.

Zdravevski and Kulakov (2009) obtained two seasons of NBA data (1,230 matches), the first of which was used as the training dataset, and the second of which was used as the test dataset. All of the algorithms in the WEKA machine learning toolkit were applied with their default parameter settings. A set of 10 features was selected by the authors. Classification accuracy of 72.8% was achieved with Logistic Regression. It was stated that, in future work, it would be preferable to compare their predictions to those of experts, and that it might be possible to cluster training and test data to use different models on each cluster in order to account for winning and losing streaks. It was also mentioned that aggregations or ensembles of classifiers (e.g., voting schemes) could be investigated, and that

Paper	Competition	Models used	No. of features	No. of Matches	Accuracy of best model
Reed and O'Donoghue (2005)	English Premier League	Multiple Linear Regression, ANN*, Discriminant Function Analysis	7	498	57.9% (3-class)
Joseph et al. (2006)	English Premier League	Bayesian Network, Expert Bayesian Network*, Decision Tree, kNN	4	76	59.2% (3-class)
McCabe and Trevathan (2008)	English Premier League	ANN trained with BP* and CGD	19	unknown	54.6% (3-class)
Buursma (2010)	Dutch Eredivisie League	WEKA: MultiClassClassifier with ClassificationViaRegression*, RotationForest, LogitBoost, Bayesian Network, Naïve Bayes, ZeroR	11	4590	55% (3-class)
Huang and Chang (2010)	2006 Soccer World Cup	ANN trained with BP	8	64	62.5% (3-class), 76.9% (2-class)
Hucaljuk and Rakipović (2011)	European Champions League	Naïve Bayes, Bayesian network, LogitBoost, kNN, random forest, ANN*	20	96	68% (3-class)
Odachowski and Grekow (2012)	Various leagues	BayesNet*, SVM, LWL, Ensemble Selection, CART, Decision Table	320	1,116	70.3% (2-class), 46% (3-class)
Tax and Joustra (2015)	Dutch Eredivisie League	WEKA: NaïveBayes, LogitBoost*, ANN, RandomForest, CHIRP, FURIA, DTNB, J48, HyperPipes	5	4284	56.1% (3-class)
Prasetio (2016)	English Premier League	Logistic Regression	4	2280	69.5% (2-class)
Danisik et al. (2018)	Various leagues	LSTM NN classification, LSTM NN regression*, Dense Model	139	1520	52.5% (3-class), 70.2% (2-class)
Hubáček et al. (2019b)	52 leagues	XGBoost classification*, XGBoost regression, RDN-Boost	66	216,743	52.4% (3-class)
Constantinou (2019)	52 leagues	Hybrid Bayesian Network	4	216,743	51.5% (3-class)
Berrar et al. (2019)	52 leagues	XGBoost*, kNN	8	216,743	51.9%** (3-class)

Table 4: Soccer Studies (*denotes the best performing model). Accuracies for 2-class (win,loss) and 3-class (win,loss,draw) problems are denoted. **Berrar et al. (2019)'s best performing model was kNN, but post-competition they mention that they improved on this with XGBoost. The accuracy of 51.9% is for their in-competition result - the XGBoost accuracy was not reported in their paper but it would have been slightly higher than this.

automatic feature selection methods should be used rather than features being manually selected by the authors.

Ivanković et al. (2010) used a BP-trained ANN to predict the results of the Serbian First B Basketball League, using five seasons from 2005/2006 to 2009/2010 (890 matches). The authors used the CRISP-DM framework for their experimental approach, and investigated how the successful shot percentage in six different regions of the court affected match results. The input dataset was divided into training and testing data (75%:25%), and 66.4% accuracy was obtained on the test set. The authors then reverted back to the data preparation phase of the CRISP-DM framework to see whether adding additional variables (offensive rebounds, defensive rebounds, assists, steals, turnovers, and blocks) could improve results. This improved accuracy to just under 81%. It was concluded that actions in the zone directly under the hoop, in particular, rebounds in defence and scoring in this zone, were crucial to determining the outcome of the game. It was mentioned that, in future work, a richer data set and new software solutions may help to ensure that all relevant events are included.

Miljković et al. (2010) predicted basketball match results using data from 778 games in the regular 2009/2010 NBA season. The features were divided into game (in-play) features, which directly relate to events within the match (e.g., fouls per game and turnovers per game), and those that relate to standings (e.g., total wins and winning streaks). Naïve Bayes achieved 67% accuracy (10-fold cross validation), and was found to be the best performing model when compared to kNN, a Decision Tree and SVM. Their future research plans included applying their system to other sports, and to experiment with other models such as ANNs.

Cao (2012) created an automated data collection system, obtaining six years of NBA matches from the 2005/2006 season to the 2010/2011 season. The dataset, comprising around 4,000 matches, was divided into training, test, and validation sets. Four models were compared: Simple Logistic Regression, Naïve Bayes, SVM, and a BP-trained ANN. The feature set of 46 variables was selected based on the domain knowledge of the author. All models were found to produce similar accuracies, with Simple Logistic Regression, which performs automatic feature selection, achieving the highest accuracy (67.8%). The best expert predictions on teamrankings.com were slightly better, achieving 69.6% accuracy. The author suggested that, in the future, clustering could be used to group players by positional group, or to identify outstanding players, and outlier detection methods could be used to identify outstanding players or team status. Investigating the impact of player performance on match results, and comparing different feature sets derived from box-scores and team statistics were also mentioned as avenues for future work.

Shi et al. (2013) investigated the viability of ML in the context of predicting the results of individual NCAAB matches which, up to this point, had been dominated by statistical methods. They used the WEKA toolkit to compare the accuracies of the ANN, C4.5, RIPPER and Random Forest algorithms. Experiments were conducted using data from six seasons, together with an expanding window approach so that initially training was performed on the 2008 season and testing on the 2009 season. Thereafter, the combination of all previous seasons comprised the training set until the 2013 season. The authors concluded that, on average, the ANN with default parameter settings provided the best accuracies, though statistical tests to confirm this were not provided. The top-ranked features in

Paper	Competition	Models used	No. of features	No. of Matches	Accuracy of best model
Loeffelholz et al. (2009)	NBA	ANN (types: FFNN*, RBG, PNN, GRNN, fusions of these)	4	650	74.3%
Zdravevski and Kulakov (2009)	NBA	All models in WEKA (Logistic Regression*)	10	1,230	72.8%
Ivanković et al. (2010)	Serbian First B	ANN trained with BP	51	890	81%
Miljković et al. (2010)	NBA	kNN, Decision Tree, SVM, Naïve Bayes*	32	778	67%
Cao (2012)	NBA	Simple Logistic Regression*, Naïve Bayes, SVM, ANN	46	4,000	67.8%
Shi et al. (2013)	NCAAB	ANN*, C4.5 Decision Tree, RIPPER, Random Forest	7	32,236	74%
Thabtah et al. (2019)	NBA	ANN, Naïve Bayes, LMT Decision Tree*	8	430	83%

Table 5: Basketball Studies (*denotes the best performing model).

terms of importance were location, the "four factors" (Oliver, 2002) and adjusted offensive and defensive efficiencies (kenpom.com). The authors remarked that they experienced an upper limit of 74% accuracy that they could not improve beyond, and noted that feature engineering and selection hold promise for an improvement in results.

Thabtah et al. (2019) used Naïve Bayes, an ANN, and an LMT Decision Tree model to predict the results of NBA matches, focusing on trialing various different subsets of features in order to find the optimal subset. Their dataset, obtained from Kaggle.com, consisted of 21 features and 430 NBA finals matches from 1980 to 2017 and a binary win/loss class variable. Defensive rebounds were found to be the most important factor influencing match results. The feature selection methods used were: Multiple Regression, Correlation Feature Subset (CFS) selection (Hall, 1998), and RIPPER (Cohen, 1995). Defensive rebounds were selected as being important by all three feature selection methods. The best performing model (83% accuracy) was trained on a feature set consisting of eight features, which were selected with RIPPER and trained using the LMT model. The authors suggested that the use of a larger dataset, more features (e.g., players, coaches), and other models (e.g., function-based techniques and deep learning methods) are potential avenues for further research.

3.1.5 Ice Hockey

Weissbock et al. (2013) noted that Ice Hockey had not received much attention in ML research with respect to predicting match results. The continuous nature of this sport makes it difficult to analyze, due to a paucity of quantifiable events such as goals. This characteristic was cited as a possible reason for the lack of attention Hockey had received historically. The authors focused on exploring the role of different types of features in predictive accuracies across several types of ML algorithms. They considered both traditional statistics as features, as well as performance metrics used by bloggers and statisticians employed by

teams. They used WEKA's implementations of ANN, Naïve Bayes, SVM and C4.5 for training classifiers on datasets describing National Hockey League (NHL) match results in the 2012/2013 season. The entire dataset amounted to 517 games. The authors concluded that traditional statistics outperformed the newer performance metrics in predicting the results of single games using 10-fold cross validation, while the ANN displayed the best accuracy (59%). Research into extracting more informative features and predicting the winners of NHL playoffs was cited as future work, as well as incorporating knowledge from similar sports such as soccer.

Weissbock and Inkpen (2014) combined statistical features with features derived from pre-game reports to determine whether the sentiment of these reports was useful for predicting NHL Ice Hockey matches. Data from 708 NHL matches in the 2012/2013 season were collected, and the pre-game reports were obtained from NHL.com. Both natural language processing and sentiment analysis based features were used in the experiments. The three statistical features, identified from their previous research (Weissbock et al., 2013), were: cumulative goals against and their differential, and the match location (home/away). The following algorithms from the WEKA toolkit were applied with their default parameters: ANN, Naïve Bayes, Complement Naïve Bayes, Multinomial Naïve Bayes, LibSVM, SMO, J48 Decision Tree, JRip, Logistic Regression, and Simple Logistic Regression. Three models were compared: models that used only the statistical features, models that used only the pre-game report text, and models trained with the features from sentiment analysis. It was found that models using only the features from pre-game reports did not perform as well as models trained with only the statistical features. A meta-classifier with majority voting was implemented, where the confidence and predicted output from the initial classifiers was fed into a second layer. This architecture provided the best accuracy (60.25%). This cascading ensemble approach on the statistical feature set provided superior performance to using both feature sets, suggesting that the pre-game reports and statistical features provided somewhat different perspectives. The authors commented that it was difficult to predict matches with a model trained using data from one or two seasons prior, probably due to player and coaching changes, etc.

Gu et al. (2019) reported that ensemble methods provided encouraging results in the prediction of outcomes of NHL Hockey matches over multiple seasons. The data extraction was automated and scraped from several websites containing NHL matches from the 2007/2008 to 2016/2017 seasons (1,230 matches). Data was merged from several sources including historical results, opposition information, player-level performance indicators, and player ranks using PCA. A total of 26 team performance variables were also included in their model. The kNN, SVM, Naïve Bayes, Discriminant Analysis and Decision Tree algorithms were applied, along with several ensemble-based methods. The ensemble-based methods (Boosting, Bagging, AdaBoost, RobustBoost) achieved the highest accuracy on the test set (91.8%). In terms of future research, the authors mentioned that additional data could further improve predictions, and different or additional features could be used, e.g., player ranking metrics, moving-averaged or exponentially-smoothed team and player performance metrics, psychological factors, the strategic or tactical judgements of coaches/experts, and players' physical or mental assessments. They also mentioned that the ML problem could be re-formulated so that a different outcome is predicted, e.g., whether a team will make

the playoffs or which team will win the championship, by training a model on the regular season data.

Paper	Competition	Models used	No. of features	No. of Matches	Accuracy of best model
Weissbock et al. (2013)	NHL	Naïve Bayes, SVM, ANN*, C4.5 Decision Tree	11	517	59%
Weissbock and Inkpen (2014)	NHL	Weka: ANN, Naïve Bayes, Complement Naïve Bayes, Multinomial Naïve Bayes, LibSVM, SMO, J48, JRip, Logistic, Simple Logistic, Simple Naïve Bayes, Cascading Ensemble*	6	708	60.3%
Gu et al. (2019)	NHL	KNN, SVM, Naïve Bayes, Discriminant Analysis, Decision Trees + ensembles of these: Boosting*, Bagging, AdaBoost, RobustBoost	19	1,230	91.8%

Table 6: Ice Hockey Studies (*denotes the best performing model).

3.1.6 Multiple Invasion Sports

McCabe and Trevathan (2008) used data from 2002 to 2007 and considered four different sports: Australian National Football League (NFL) Rugby League, Australian Football League (AFL) Australian Rules Football, Super Rugby (Rugby Union), and EPL soccer. ANNs trained with BP and Conjugative-Gradient Descent (CGD) were applied, with the former found to be slightly more accurate but had longer training time. The features used were the same across all of the four sports, i.e., sport-specific features derived from in-play events within matches were not included. The average accuracy achieved with the BP-trained ANN was 67.5%, higher than the expert predictions, which ranged from 60% to 65%. For future work, the authors mentioned that other sports could be considered, more features could be included, and the points margin could instead be predicted.

Reed and O'Donoghue (2005) predicted the results of EPL Soccer and English Premiership Rugby, building seven models including Multiple Linear Regression, ANN, and Discriminant Analysis. Their dataset consisted of the matches of three Soccer teams and two Rugby teams across three seasons, and contained seven features: match venue, rest, the positions of the team and opposition team in the league table, distances travelled to the match, and form. These features were used to predict both Rugby and Soccer matches, i.e., variables specific to in-play match events in Rugby or Soccer were not included. Given that a draw is a much more common result in Soccer than in Rugby, it is surprising that the accuracy obtained for Soccer (57.9%) was higher than for Rugby (46.1%). The models outperformed the predictions of human experts, and the ANN was found to achieve the best accuracy. The authors stated that motivational, injury and other variables could be included in future studies, and complex pattern recognition technology could be trialed rather than inflexible, simple linear models.

3.2 Striking & Fielding Sports

As opposed to invasion sports, which are time dependent, striking and fielding games are innings dependent. Perhaps the most widely-played striking and fielding sports are baseball, which is widely played in countries such as the United States, Japan, Korea and the Dominican Republic, and cricket, which is widely played in England and the countries of

Paper	Sport-Competition	Models used	No. of features	No. of Matches	Accuracy of best model
McCabe and Trevathan (2008)	Rugby League (NRL)	ANN trained with BP* and CGD	19	unknown	63.2%
McCabe and Trevathan (2008)	Australian Rules Football (AFL)	ANN trained with BP* and CGD	19	unknown	65.1%

Table 7: Other invasion sport studies (*denotes the best performing model)

the current/former British Commonwealth (e.g., Australia, New Zealand, South Africa, and nations in the Caribbean and Indian subcontinent).

3.2.1 Baseball

Valero (2016) performed a comparative study for prediction of Baseball matches, using 10 years of Major League Baseball (MLB) data. Lazy Learners, ANNs, SVMs and Decision Trees were the candidate models. The CRISP-DM framework was used as the experimental approach, and a classification approach (win/loss for the home team) was compared with a numeric prediction approach that predicted the run difference between the home and away teams. Feature selection and ranking methods in WEKA were applied to rank the original set of 60 features. The model used only the top three ranked variables: home field advantage, Log5 ratings, and the Pythagorean Expectation, and it was found that adding additional features did not improve results. The Pythagorean Expectation, developed by the Baseball statistician Bill James (James, 1984), represents the expected number of wins for a team given their runs scored and runs allowed, and Log5 ratings are essentially the same as Elo Ratings (Elo, 1978). SVM produced the best accuracy for both the classification and numeric prediction approaches. Consistent with the results of Delen et al. (2012), the classification approach performed significantly better than the numeric prediction approach. The SVM classification model achieved accuracy of around 59%. However, when using the 2005-2013 seasons as training data and the 2014 season as test data, the model's predictions were not significantly more accurate than predictions derived from match betting odds. The authors highlighted the difficulty in predicting outcomes in Baseball using statistical features alone, but suggested that experiments using the Japanese or Korean Baseball leagues could be useful. In future work, the authors also considered adjusting their model parameters, refining features, extending their datasets, and applying their model to other sports such as Basketball, Football, and Water Polo.

3.2.2 Cricket

Pathak and Wadhwa (2016) predicted the results of One-Day International (ODI) Cricket, and included four features based on the prior work of Bandulasiri (2008): toss outcome, match venue (home or away), time (day or night), and whether the team batted first or second. Three classification models were applied: Naïve Bayes, Random Forest and SVM. Data for matches from 2001 to 2015 were collected from cricinfo.com. A separate model was constructed for each team, analyzing that particular team with respect to all other teams. A training-test split of 80%:20% was used. To mitigate the effect of imbalanced datasets

Paper	Sport-Competition	Models used	No. of features	No. of Matches	Accuracy of best model
Pathak and Wadhwa (2016)	Cricket (ODIs)	Naïve Bayes, Random Forest, SVM*	4	unknown	61.7% (balanced accuracy)
Valero (2016)	Baseball (MLB)	Lazy Learners, ANN, SVM*, Decision Tree	3	24,300	59%

Table 8: Striking/Fielding sport studies (*denotes the best performing model)

(some teams had a high ratio of wins to losses), the three models were evaluated based on balanced accuracy and the Kappa statistic. SVM was found to perform the best across all teams, with an average balanced accuracy of 61.7%. It was suggested that, in future work, newer classification methods could be used, additional features could be included, and the approach could be applied to other forms of cricket, e.g., test matches and T20, and to other sports, e.g., baseball and football.

4. Discussion

In this section, we provide some discussion, from a machine learning perspective, on what can be observed across the surveyed articles. In Section 4.1, we discuss the most commonly applied ML techniques in this domain. In Section 4.2, we discuss data-related aspects of ML for team sport match result prediction, including feature selection and engineering, e.g., methods, dataset size, and feature subset comparison. We also highlight the fact that in-play features are specific to each sport, and that, in this respect, there are opportunities for inter-disciplinary collaboration between machine learning and sport performance analysis researchers. In addition, we discuss dataset size in terms of the number of instances, model training and validation approaches, cross-validation with chronologically-ordered instances, and class variable definition and comparison. In Section 4.3, we then describe evaluation of performance in terms of the most common evaluation metric (accuracy), as well as benchmark datasets and other ways to evaluate performance. Then, in Section 4.4, we discuss between-sport differences in predictability, based on the inherent characteristics and points-scoring systems of different sports. Finally, in Section 4.5, the most common future research directions that were extracted from the surveyed studies are analyzed and discussed.

4.1 Machine Learning Algorithms

ANNs are one of the most predominant ML techniques used in the analysis of data in sports (Schumaker et al., 2010). Indeed, in the course of our review, we found that a number of studies, especially early ones, used an ANN as the only predictive model, and did not compare its performance with any other models. This may well have been a consequence of the availability of specific software, tools and algorithms at the time of a study was conducted, rather than being motivated by a belief that ANNs are inherently better for

predicting match results in team sports. In a 2019 article in MIT Technology Review[1], it was found that, in an analysis of 16,625 AI-related papers, there has been a notable shift away from knowledge-based methods (those that derive rules or logic) over the past two decades. In the first half of the 2000s, there was an increase in use of ANNs, but from 2004 to 2014 their use declined. In the second half of the 2010s, ANNs again gained in popularity, probably due to the emergence of deep learning. Nonetheless, the common application of ANNs for team sport results prediction, often as the sole predictive model, prompts us to investigate whether the evidence in the literature suggests that ANNs have performed better than other models in practice.

Our research found that the majority of studies (65%) considered ANNs in their experiments, as shown in Figure 3. In earlier studies in particular, 23% of the papers considered ANNs as their only predictive model. The greater propensity for researchers to use ANNs in this domain has also resulted in the majority of studies attributing their highest accuracy to ANNs (Figure 5). It should also be noted that studies on two sports in this graph (Rugby League and Australian Rules Football) considered ANNs only. Three sports (American Football, Ice Hockey and Basketball), which found their best accuracy performances with alternative algorithms, also used ANNs, however, the alternative algorithms outperformed them. Therefore, the evidence does not suggest that ANNs have consistently performed better than other ML algorithms in predicting the match results of team sports. Indeed, the broader ML literature, real-world applications, and ML competitions (e.g., those held on Kaggle.com) do not support blanket statements that would give primacy to ANNs over all other algorithms. It is unclear why, historically, researchers displayed a preference for ANNs given that they are not straightforward to parameterize optimally and often over-fit, especially in the absence of sufficiently large datasets, which tends to be the norm rather than the exception in this domain. Another disadvantage of ANNs is that they are not easy to interpret and are therefore less useful to performance analysts and coaches seeking to draw out insights than other, more interpretable, ML models.

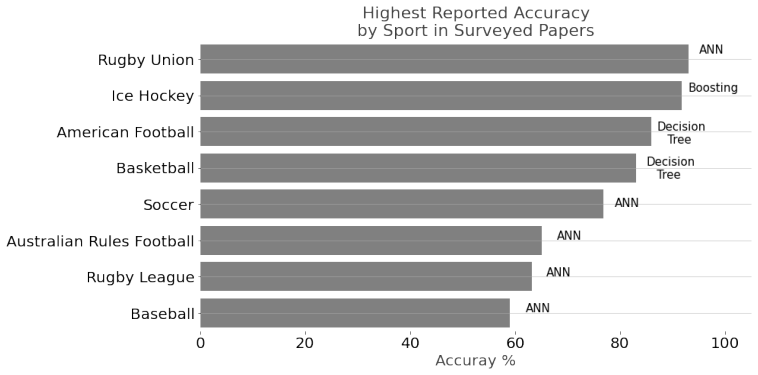


Figure 5: Highest recorded accuracies in studies by the different sporting codes covering the review period. The highest accuracy is reported along with the associated algorithm

1. https://www.technologyreview.com/2019/01/25/1436/we-analyzed-16625-papers-to-figure-out-where-ai-is-headed-next/

Decision Trees were the second most commonly applied technique (Figure 3). The appeal of these algorithms is obvious in that they are fast to train and usually do not possess a cumbersome number of tunable parameters. Importantly, they do not generate black-box models, but instead embody varying degrees of interpretability, depending on their implementation. This property can offer utility to professionals beyond just the ability to make predictions, but also in providing insight to coaches, management, and athletes. For instance, interpretable models like Decision Trees can aid in the identification of the most important performance indicator variables that influence match results, which is valuable in terms of developing appropriate strategies and focus areas. Their widespread use has resulted in both American Football and Basketball reporting their highest accuracy results using CART and Logistic Model Trees (LMT), respectively. Other variants of Decision Tree that have been popular in the literature include C4.5 (Quinlan, 1993) and its corresponding implementation in WEKA (Witten et al., 1999), J48.

Ensemble methods, Boosting algorithms, and Random Forests together form a top-three category of techniques used in the surveyed studies (Figure 3). Given the differing degrees of resilience of this family of algorithms to over-fitting, it is unsurprising that they have been used liberally, and have registered top accuracy results in Ice Hockey, in a recent study that highlighted the potential of ensemble-based solutions (Gu et al., 2019). A useful aspect of some of these models is their ability to achieve high accuracy while retaining interpretability. In future research in this domain, we see the potential for the application of Alternating Decision Trees (Freund & Mason, 1999), which possess the accuracy-improving benefits of boosting while retaining the interpretability of a decision tree structure.

Bayesian algorithms, e.g., Naïve Bayes and BNs, were one of the most popular sets of techniques used in the surveyed studies. Though these algorithms are not listed as performing the best in individual sports (Figure 5), they have been found in some comparative studies to offer better accuracies than alternative algorithms (Joseph et al., 2006; Miljković et al., 2010; Odachowski & Grekow, 2012). The popularity of Naïve Bayes in particular can be attributed to its common usage as a benchmarking algorithm when assessing the learnability of a new problem, which is tied to its ability to generate classifiers that do not over-fit. BNs were also shown to be useful when incorporating expert knowledge and when using small datasets (Joseph et al., 2006).

4.2 Data

In this subsection, the data-related aspects of the surveyed studies are discussed. In particular, we discuss feature selection and engineering, dataset size in terms of the number of matches/instances, model training and validation, and the use of cross-validation where instances are temporally ordered, which is the case for sports matches. Appropriate ways of defining the class variable, and comparing different class variables, are also discussed.

4.2.1 Feature Selection & Engineering

In early studies in particular, model features were often selected manually by researchers based on their knowledge of the specific sport. More recently, data-driven feature selection methods, including various filter-based techniques, have become more commonplace. These have ranged from CFS selection (Hall, 1998), to algorithms such as ReliefF (Kira & Rendell,

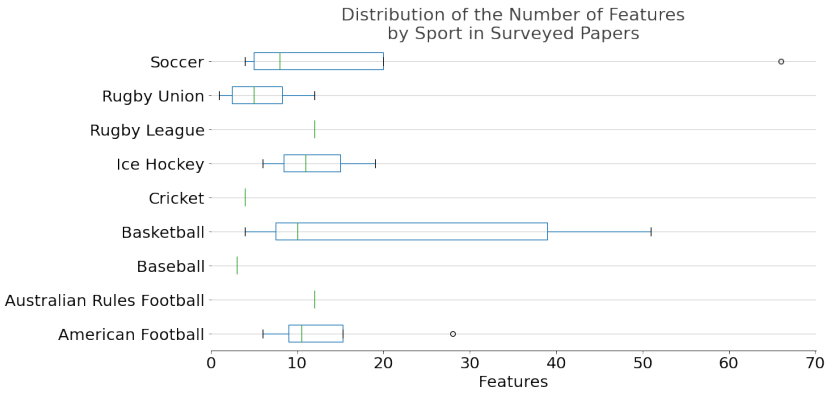


Figure 6: Distribution of the number of features used for machine learning by sport across all surveyed papers

1992), which have a greater contextual awareness and consider the existence of dependencies between features. Others have used feature-importance outputs from ML algorithms such as RIPPER (Cohen, 1995) to inform which features should be retained in the training of final models, while some (e.g., Loeffelholz et al., 2009) have used ANN-specific methods including signal-to-noise ratios (Bauer Jr et al., 2000). Sport experts have also been consulted to select what they consider to be the most important predictive features.

Compared to some other domains, the number of features used across the various sports have generally been on the modest side, and their distributions can be seen in Figure 6. If we introduce time as a dimension to discern if trends exist, we can see in Figure 7 that for Soccer, American Football, and Ice Hockey, a general tendency towards using larger feature sets can be observed. These are also sports that have, on average, seen some of the largest improvements in accuracies.

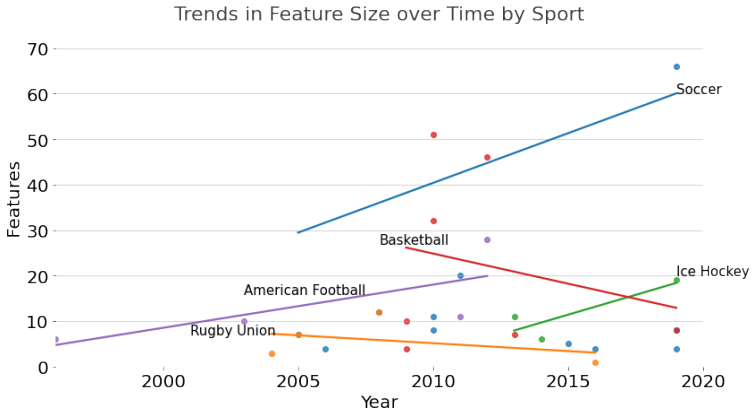


Figure 7: Differences in the distribution of predictive accuracies reported by sport.

In addition to the application of feature selection techniques, studies with a robust experimental process have generally compared several different feature subsets, e.g., betting odds, in-play features, features external to the match, player-level features, expert-selected

features, features extracted from pre-game reports, and features constructed from ratings. The performance of each feature subset can be compared, and can also be compared with the entire original set of features.

In-play features are sport-specific. Researchers have often proposed applying their predictive models to other sports. However, given that each sport has unique features that are associated with outcomes, it is generally not possible to directly apply a model to a dataset from a different sport. Rather, it is necessary to go through an entirely new experimental process on the dataset for that sport. Therefore, we recommend that, to approach prediction problems in a structured way, researchers should follow an experimental framework such as the CRISP-DM framework, the Knowledge Discovery in Databases (KDD) framework (Fayyad et al., 1996) or the Sport Result Prediction CRISP-DM (SRP-CRISM-DM) framework (Bunker & Thabtah, 2017) (the latter is an extension of CRISP-DM specifically designed for the application of ML to sport results prediction).

Feature selection & opportunities for inter-disciplinary collaboration. Nearly 90% of the studies we surveyed listed engineering additional richer features as one of their endeavours for future work (Figure 11). This should come as no great surprise, since generating more descriptive and therefore discriminatory features, with the help of domain knowledge, is generally accepted as the best strategy for improving predictive accuracy using ML (Domingos, 2012). The ability to generate more effective decision boundaries between instances of different classes is, to a larger degree, determined by richer features rather than by algorithms of increasing sophistication. To that end, domain expertise plays an important role in crafting more descriptive features. Such domain expertise could be obtained in consultation with coaches or athletes, or potentially from academic literature, e.g., from the field of sport performance analysis. Much of the research sport performance analysis considers so-called performance indicators (PIs). PIs represent the in-play features in sport result prediction models, and are usually augmented with features external to sport matches that may have an influence on outcome, e.g., weather, venue, travel, and player availability. Sport performance analysts, however, are not usually concerned with external variables because they are usually outside of the control of coaches and players. We suggest that there is an opportunity for knowledge transfer from the field of sport performance analysis, likely to be found in the identification and development of new model features. Given that feature engineering has been identified in this review as an area that is of highest priority for future research, greater collaboration between sport performance analysis and machine learning may result in meaningful advances. The two disciplines do not appear to have reached the point where a significant level of interchange is currently taking place, so researchers are encouraged to expand their collaborations in this respect.

4.2.2 Dataset Size

The distribution of the size of the datasets, in terms of number of instances, and their evolution over time, can be seen in Figure 8. It can be observed that, generally, the datasets have tended to be of a relatively small size, due to the limited amount of historical data in particular sports. This, together with fact that the data tends to be highly structured, potentially limits the ability of the datasets to capture the signal, which would likely result

in more accurate models. Sports that have seen an increase in dataset sizes are Basketball and, more acutely, Soccer.

It appears, however, that having access to a large dataset in terms of number of matches has not necessarily led to higher accuracy. This was particularly evident from the results of the 2017 Open International Soccer Database Competition, where model accuracies were modest despite the dataset containing over 216,000 matches. This particular pattern is depicted in Figure 9, where accuracy trends over time are rendered for all sports. Arguably, the growth in both the quantity and quality of features used in each sport (Figure 6) has grown more than the size of the datasets, which suggests that feature engineering and robust feature selection are likely to be key drivers of improvements in predictive accuracy.

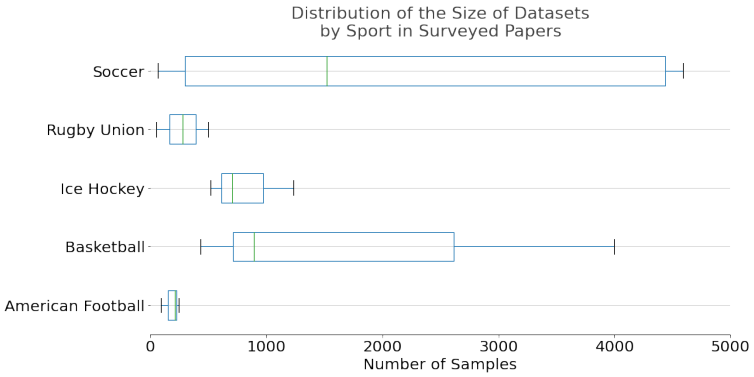


Figure 8: Distribution of the number of samples in datasets used for machine learning, by sport, across all surveyed papers.

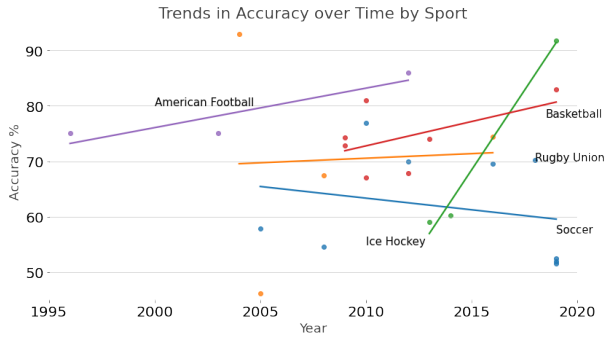


Figure 9: Trends in accuracy over time by sport.

4.2.3 Model Training & Validation

Successful studies have formulated experimental designs that have tested a number of different training-validation splits in their data. For instance, a certain number of historical seasons have commonly been used to train the model(s), and validation is performed on a current or future season. Or, if only one season of data was available, models are generally trained on a certain number of competition rounds and then validated on a future round in

that season. Researchers have often trialed various training-test splits and compared their accuracies.

4.2.4 Cross-Validation With Chronologically-Ordered Match Instances

Cross-validation was used in a number of studies. However, in match result prediction in sport, this can be problematic if the technique is not appropriately modified. Standard cross-validation randomly shuffles instances, so its application may mean that future matches are used to predict past matches. This pitfall was identified in a number of surveyed studies, and the predictive accuracies of these studies may have been compromised as a consequence.

4.2.5 Class Variable Definition & Comparison

A comparison of the performance of models, predicting both the points margin (home team points minus away team points) and discrete match outcomes (win/loss/draw or win/loss), has been investigated by a number of researchers (Delen et al., 2012; Valero, 2016; Danisik et al., 2018). Both approaches form a distinctive approach to formulating the ML task. However, results have been mixed, with no approach consistently shown to be superior. For instance, Delen et al. (2012) and Valero (2016) found that the classification approach performed better, while Danisik et al. (2018) found that a numeric prediction approach was superior on their dataset. Given these mixed results, we would recommend that, where possible, future researchers perform this type of comparison as part of their experimental process.

4.3 Performance Evaluation

In this subsection, common performance evaluation approaches, including evaluation metrics, are discussed. We discuss why predictive accuracy, the most common evaluation metric in this domain, is indeed an appropriate evaluation metric for team sport match result prediction. We also discuss benchmark datasets, and other ways in which model performance can be evaluated if no such benchmark dataset is available.

4.3.1 Evaluation Metric: Predictive Accuracy

Predictive accuracy is the performance evaluation metric that has been used by the vast majority of researchers in this domain. This is understandable since it is intuitive and interpretable, and datasets in this domain tend not to be overly imbalanced. In particular, when each instance in a dataset represents a match between two teams (home and away), there is generally a slight class imbalance in favor of the home team due to the well-known home advantage phenomenon. Over a whole season (or multiple seasons) in a competitive league, class imbalance generally will not exist to the degree that predictive accuracy becomes inappropriate to use as a performance evaluation metric.

4.3.2 Benchmark Datasets

A number of studies highlighted the inherent difficulties in comparing the results of studies in this domain. This difficulty arises since studies usually differ in at least one of the following dimensions: the sport(s) considered, the input dataset, the model predictors, the

class variable, and/or the matches, seasons or competitions considered. Part of the challenge lies in the scarcity of available benchmark datasets. Although this has been resolved for soccer, with the creation of the Open International Soccer Database (Dubitzky et al., 2019), as mentioned, this dataset is limited in that it does not include features derived from in-play events. Sport datasets are becoming increasingly available on websites such as Kaggle.com, and these could come to act as benchmark datasets for other sports in the future.

4.3.3 Other Ways to Evaluate Performance

Despite an absence of benchmark datasets in sports other than soccer, there are a number of other ways for researchers to evaluate their experimental results, e.g., by comparing their results to some baseline measure. Common approaches that have been used in the literature include comparing the predicted outcomes with:

- *Predictions derived from betting odds:* The outcome with the lowest betting odds is used as the class variable. Betting odds have also proven useful for match result prediction, even when included as the sole model predictor (Tax & Joustra, 2015). However, Hubáček et al. (2019a) pointed out that, if the purpose of the model is to generate profit through betting strategies, betting odds should not be included as a model predictor if one wants to "beat the house."

- *Predictions of experts:* ML model predictions can be compared to the predictions made by experts on the same set of matches, which are often published online or in the media.

- *ZeroR (Majority-class selection):* This is a simple classification rule that always selects the majority class. In most cases, this will predict a home-team victory due to the existence of the well-known home-advantage phenomenon.

- *Random prediction:* A randomly selected match outcome.

Given the specific nature of most datasets in sporting domains in terms of their time periods and features, comparisons with the above are useful for researchers when reporting their experimental results.

4.4 Between-Sport Differences in Predictability

Given the existing data and possible confounding factors, it is difficult to determine whether predicting match outcomes in certain sports is inherently more challenging than in others. Luck will always be a factor. Some research into disentangling the size of the luck (randomness) component from the skill component when predicting outcomes has already been conducted Aoki et al. (2017). The authors note that, in Soccer, the teams will win only 50% of the time when favoured by bettors, while the favored teams win 60% of the time in Baseball, and 70% of the time in both American Football and Basketball. Their own modelling research concludes that, out of four teams sports that they considered, Basketball appears to be the sport in which skill plays the largest role in the final results, and therefore has the most predictability. Basketball was followed in order by Volleyball, Soccer and Handball.

Possible hypotheses. Based on the data we have gathered, we offer some hypotheses with respect to the causes of the differences in predictability between the different sports. One possible explanatory factor may simply be that sufficiently large datasets and rich feature sets, which support high predictive accuracies, have not been equally available to researchers across different sports. Another possible reason is that sports have received highly imbalanced amounts of attention in the ML literature, which could in itself be due to the fact that non-ML techniques already perform adequately in predicting match results. For these reasons, it is inappropriate to attribute lower accuracies in sports that have received less research attention, e.g., Cricket, Rugby League, and Australian Rules Football (Figure 2), to something that is intrinsically more non-deterministic in these sports compared to other sports. However, a little more can arguably be inferred from sports such as Soccer, Basketball, American Football, Rugby Union and Ice Hockey, which have garnered more attention.

Low-scoring sports can have higher unpredictability. Soccer has received the largest share of research, yet its highest recorded predictive accuracy was 78% (Figure 5), and it comes fifth with respect to accuracy across all sports. Rugby Union, on the other hand, has received limited research attention, yet it ranks the highest of all surveyed sports, with an accuracy of 93% (O'Donoghue & Williams, 2004) recorded as its best outcome. Interpreting this perhaps requires some caution. Low-scoring sports tend to embody a higher degree of random chance as a determinant of results and this, in part, would explain some of the variation. This is also supported by Aoki et al. (2017), who showed that pure chance can be the single factor of outcomes in as many as 18% of matches in a season.

Lower competitiveness suggests higher predictability. The differential in accuracies highlighted above may, however, to some degree, be attributable to the characteristics of the sports and, even more, to the competitive contexts to which the predictive experiments were applied. For instance, Rugby Union is a much smaller global sport than Soccer, being played by fewer nations, while historically being dominated by a handful of them. Understandably, then, the best results reported for Rugby Union originated from a study of the 2003 Rugby World Cup, which suggests a context with a higher degree of predictability.[2] The context for Soccer and the competitions from which the results were collected were markedly different. In particular, the results were drawn from both national and international events in which the depth of competition was arguably greater, and that ultimately may have created conditions in which accurate prediction of results was less deterministic. Some evidence supporting the notion that a great deal of the predictability of sport results is naturally determined by the inherent depth of the competitions under observation, rather than on the sports themselves, is supported by Figure 10, which shows a much higher variability in the predictive accuracies for Rugby Union than for Soccer. We can see that in Rugby Union studies that considered national premiership (Reed & O'Donoghue, 2005) and international franchise competitions (McCabe & Trevathan, 2008), which exhibit a greater degree of equality between teams, the accuracy was correspondingly much lower, at 46% and 74%, respectively. In addition, Soccer is a low-scoring sport compared to Rugby Union

2. Rugby Union is, however, known to be becoming less predictable as the gap between low-ranked and high-ranked nations is narrowing over time and, consequently, more upsets are occurring. This was evident in the reduced predictability of the 2015 Rugby World Cup compared to previous tournaments (O'Donoghue et al., 2016).

and, returning to this point, we postulate that this may also be a contributing factor to generally lower accuracies in its studies. In low-scoring sports, there is a larger element of randomness in outcomes, which decreases the performance of predictive models.

Sports with more possible outcomes are less predictable. Associated with this is the higher likelihood of matches ending in draws. Given that draws are not improbable, many researchers predicting soccer match results have formulated the ML task as a three-class problem (win/loss/draw), rather than a binary-class (win/loss) problem. Of course, in general, as the number of classes increase, the learning problem becomes more difficult and thus accuracies tend to decrease.

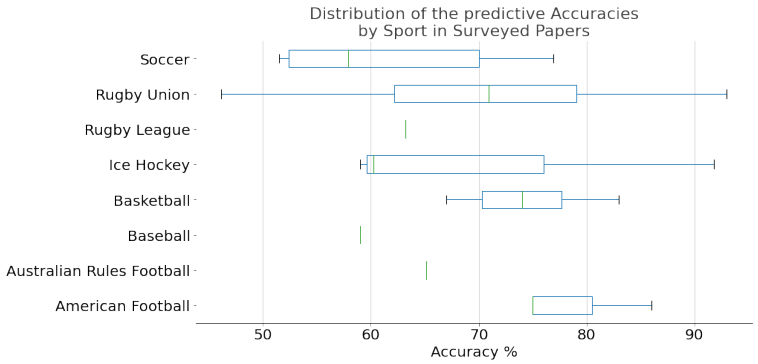


Figure 10: Differences in the distribution of predictive accuracies by sport.

Better features & richer datasets tend to increase predictability. Ice Hockey, on the other hand, is a counter-example, even though a binary-class approach was used in all studies. The sport is relatively low-scoring (although higher scoring than Soccer), yet its most recent and best result (92% accuracy) demonstrated considerably higher accuracy in comparison, as well as in comparison to American Football (86% accuracy) and Basketball (83% accuracy), which are both high-scoring sports. Studies on Ice Hockey, American Football and Basketball generally considered American national league competitions, e.g., the NHL, the NFL, and the NBA, which generally exhibit high degrees of competitiveness between teams. To some degree, this had the effect of controlling for the lopsidedness in expectations of outcomes that can exist in some sports and global competitions. When examining the trends over time in these three sports, some potential patterns emerged, hinting at an explanation for Ice Hockey's unexpectedly higher accuracies compared to the other two sports. We can see that the best performing result in Basketball had double the number of features of the previous study (74% accuracy Loeffelholz et al., 2009). However, the dataset was smaller and comparable ML algorithms were used, with standard Decision Trees performing the best out of the suite of methods explored. The best-performing result in American Football saw an 11 percentage-point improvement over the next best result in the NFL (Kahn, 2003), which may be attributed to a threefold increase in the total number of features used compared to the prior study, since both the total size of the dataset remained essentially the same, and the ML algorithms and training procedures did not represent a considerable increase in innovation. On the other hand, the leap from 60% accuracy (Weissbock & Inkpen, 2014) to 92% accuracy for Ice Hockey can be attributed to multiple factors. In particular, the authors used triple the number of features, double the

number of instances, and used more sophisticated algorithms and training approaches in the form of ensembling.

Effects of differing point-scoring systems on predictability. Across invasion sports, points-scoring systems and the manner in which points are attributed to scoring events differ, and this can affect the predictability of each sport. For instance, goals in Soccer increment the score by one, and it is not possible for a team to increment the score by more than one from any one scoring play. Ice Hockey is similar in this respect. On the other hand, it is possible in Basketball to increment the score by different amounts in each scoring play since there are three-pointers, in-circle shots (two points), and free-throws (one point). Different possibilities for increments in the score also exist in Rugby Union, Rugby League, American Football, and Australian Rules Football. Sports that have different possibilities for increments in score have more possible permutations in the final match scores and thus the result, an assertion that is also supported by Aoki et al. (2017).

Predictability is ultimately multi-factorial. In summary, although some invasion sports do embody characteristics such as being low-scoring (which makes outcomes harder to predict accurately, especially if a multi-class formulation is used), the competitive depth of the types of competitions in which the matches take place is likely to also be an important factor. A reasonable assumption is that, in general, the match outcomes of sports that are highly competitive, low-scoring, and have less possible increments in score will be more difficult to predict. However, some evidence does suggest that these difficulties can be mitigated to an extent when large and rich feature sets are used in conjunction with datasets containing a large number of instances, and cutting-edge training procedures are employed that combine multiple algorithms into robust, ensemble-based solutions.

4.5 Common Future Research Directions

Figure 11 depicts the percentage of surveyed studies that cited a specific future research direction (only research themes that were cited more than once are shown on the chart).

As mentioned, Figure 11 shows that early 90% of surveyed studies cited engineering additional features, which was by far the most common intended future research direction. Next, experimenting with alternative ML algorithms was cited as a future undertaking by nearly 40% of the surveyed studies. Each ML algorithm embodies within it assumptions about characteristics of the problem dataset, which may or may not hold, and the extent of the disconnect between the two will also affect the generalizability of different algorithms to varying degrees. A reasonable course of action, when an improvement in accuracy is sought, is to investigate different families of algorithms, which is the intention that Figure 11 appears to indicate that researchers are heading towards. What is more, non-parametric algorithms have a tendency to over-fit, and this is exacerbated on smaller datasets. A quarter of the studies signaled their intention to increase the size of their datasets, which, in this instance, would be the correct course of action for studies that have experienced this difficulty.

Meanwhile, improving training methods by reformulating the ML problem can also have a significant effect on accuracy, and nearly 40% of the studies intend to explore this. The types of training modifications that were cited for future work included using different combinations of previous season results (where appropriate) and giving greater weight to

datasets that are more recent. Some proposed applying clustering to the datasets before applying ML, and using alternative algorithms on different datasets.

Given that each sport has unique characteristics and potential inflection points that can act as markers for winning outcomes, it is somewhat surprising that a quarter of the studies plan to apply their ML methodology, including features, to other sporting codes. This seems rather counter-intuitive, given that 90% of these papers intend to generate more custom-designed features that are more tailored to their respective sporting codes.

Furthermore, it was also unanticipated to find that only 10% of the studies cited improvements in feature selection as an area of pursuit for future research. The problem of over-fitting is amplified when the pool of features is too large with respect to the size of the datasets. Given that most of the available datasets in this domain are not large, coupled with a strong consensus across all studies to pursue engineering and generation of more features, the problem of over-fitting is likely to plague many studies unless efforts to employ feature selection or dimensionality reduction techniques are afforded equal attention.

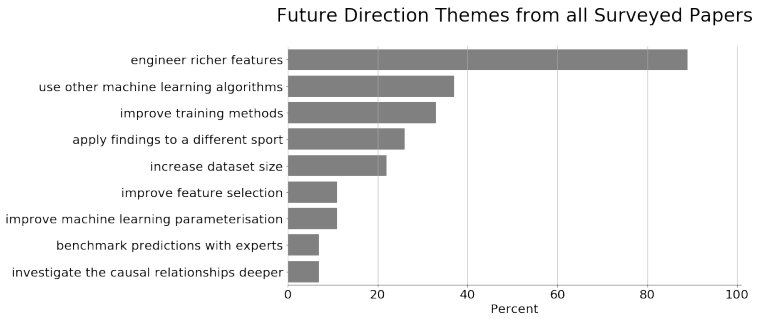


Figure 11: Commonly cited future research directions.

5. Conclusions

This survey reviewed studies published between 1996 and 2019 that applied machine learning (ML) methods to predict match results in team sport. In contrast to previous review articles in the same period, the scope of this review was defined to be narrow enough to allow for sufficiently in-depth analysis, but not overly wide so as to result in an unmanageable number of surveyed papers. The surveyed papers were categorized and sub-categorized by sport type (invasion sport or striking/fielding sport) and sport, respectively, and were analyzed and summarized in tabular formats. Various characteristics of the surveyed studies were analyzed including the types of models and experimental approaches used, the best performing models, the number of features included, and the total number of instances available. From this, we then discussed commonly applied ML models, data-related ML considerations including feature selection and engineering, dataset size, training/validation approaches, the use of cross-validation when considering the temporal order of match instances, and the definition of the class variable. Aspects related to performance evaluation from the surveyed papers were also discussed, including appropriate evaluation metrics, benchmark datasets, and other common evaluation approaches, e.g., predictions derived from betting odds, predictions of experts, majority-class selection, and random prediction.

We then discussed differences in the inherent predictability of sports due to their characteristics and points-scoring systems, before discussing the most common future research directions across the surveyed studies.

Overall, in this survey we found that, although ANNs were commonly applied especially in early studies, often as the sole predictive model, this may have been due to available software and tools at the time, since the evidence suggested (like other domains) that ANNs do not necessarily have primacy over other ML models. Thus, future researchers are recommended to compare a a set of candidate models.

Some recent studies (outside of the time-period considered in this review) have applied deep learning techniques to predict sport results (e.g., Chen et al., 2020 and Rudrapal et al., 2020); however, a limitation in this domain may be in their lack of interpretability. We suggest that Alternating Decision Trees (Freund & Mason, 1999), which combine the accuracy-boosting benefits of ensembles while retaining the interpretability of a decision tree structure, could be a suitable model to apply to predict match results in team sports.

Selecting an appropriate feature set and engineering additional features is crucial for prediction, and appears to be more important for accuracy than access to a large number of matches/instances. We see opportunity for greater collaboration between researchers from sport performance analysis and machine learning to define sport-specific in-play features, also known as performance indicators.

Researchers should consider their model training and validation approach, e.g., by using a hold-one-out approach to train their model on all prior matches to predict one future match, or by using historical seasons to predict the current season, or by using historical competition rounds to predict the current round (or a future round). Care needs to be taken if cross-validation is used since the standard method shuffles instances, which may result in future matches (inappropriately) being used to predict past matches.

Researchers also need to consider whether to use a discrete class variable (e.g., win/loss or win/loss/draw) or a numeric class variable (e.g., points margin). These two types of class variables were compared in a number of the surveyed studies, and this approach might warrant investigation by future researchers in this domain.

We found that accuracy is, by far, the most common evaluation metric used in this domain, which seems appropriate given that it is both interpretable and intuitive, and match result datasets generally are not particularly imbalanced. Comparing the accuracies of studies remains difficult even within the same sport due to different datasets, seasons, and predictive features being used. Such comparisons are even more difficult for studies that consider different sports with distinct characteristics and points-scoring systems.

Although a lack of benchmark datasets in this domain has been addressed to a certain extent for soccer through the Open International Soccer Database, this dataset is limited in that it does not contain in-play features. There remains no generally-accepted benchmark datasets for other sports. There are, however, other performance evaluation approaches with which researchers can gauge the performance of their models, e.g., by comparing their model predictions to those obtained from betting odds, from experts, or based on majority-class selection (usually a home team victory) or random prediction.

Due to the explosion in the number of papers published in the last decade in this domain, we recommend that future surveys focus on the application of machine learning for match result prediction in one specific sport and conduct a systematic review using,

e.g., the Preferred Reporting Items of Systematic reviews and Meta-Analyses (PRISMA) framework.

References

Aoki, R. Y., Assuncao, R. M., & Vaz de Melo, P. O. (2017). Luck is hard to beat: The difficulty of sports prediction. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1367–1376.

Bandulasiri, A. (2008). Predicting the winner in one day international cricket. *Journal of Mathematical Sciences & Mathematics Education*, *3*(1), 6–17.

Bauer Jr, K. W., Alsing, S. G., & Greene, K. A. (2000). Feature screening using signal-to-noise ratios. *Neurocomputing*, *31*(1-4), 29–44.

Beal, R., Norman, T. J., & Ramchurn, S. D. (2019). Artificial intelligence for team sports: a survey. *The Knowledge Engineering Review*, *34*.

Berrar, D., Lopes, P., & Dubitzky, W. (2019). Incorporating domain knowledge in machine learning for soccer outcome prediction. *Machine Learning*, *108*(1), 97–126.

Bunker, R. P., & Thabtah, F. (2017). A machine learning framework for sport result prediction. *Applied computing and informatics*.

Buursma, D. (2010). Predicting sports events from past results. In *14th Twente Student Conference on IT*.

Cao, C. (2012). Sports data mining technology used in basketball outcome prediction..

Carpenter, G., & Grossberg, S. (2003). Adaptive resonance theory, the handbook of brain theory and neural networks. *MA Arbib (ed)*, 87–90.

Chen, M.-Y., Chen, T.-H., & Lin, S.-H. (2020). Using convolutional neural networks to forecast sporting event results. In *Deep Learning: Concepts and Architectures*, pp. 269–285. Springer.

Cohen, W. W. (1995). Fast effective rule induction. In *Machine learning proceedings 1995*, pp. 115–123. Elsevier.

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

Constantinou, A. C., Fenton, N. E., & Neil, M. (2012). pi-football: A bayesian network model for forecasting association football match outcomes. *Knowledge-Based Systems*, *36*, 322–339.

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

Danisik, N., Lacko, P., & Farkas, M. (2018). Football match prediction using players attributes. In *2018 World Symposium on Digital Intelligence for Systems and Machines (DISA)*, pp. 201–206. IEEE.

David, J. A., Pasteur, R. D., Ahmad, M. S., & Janning, M. C. (2011). Nfl prediction using committees of artificial neural networks. *Journal of Quantitative Analysis in Sports*, *7*(2).

Davoodi, E., & Khanteymoori, A. R. (2010). Horse racing prediction using artificial neural networks. *Recent Advances in Neural Networks, Fuzzy Systems & Evolutionary Computing*, *2010*, 155–160.

Delen, D., Cogdell, D., & Kasap, N. (2012). A comparative analysis of data mining methods in predicting ncaa bowl outcomes. *International Journal of Forecasting*, *28*(2), 543–552.

Domingos, P. M. (2012). A few useful things to know about machine learning.. *Commun. acm*, *55*(10), 78–87.

Dubitzky, W., Lopes, P., Davis, J., & Berrar, D. (2019). The open international soccer database for machine learning. *Machine Learning*, *108*(1), 9–28.

Edelmann-Nusser, J., Hohmann, A., & Henneberg, B. (2002). Modeling and prediction of competitive performance in swimming upon neural networks. *European Journal of Sport Science*, *2*(2), 1–10.

Elo, A. E. (1978). *The rating of chessplayers, past and present*. Arco Pub.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R., et al. (1996). *Advances in knowledge discovery and data mining*, Vol. 21. AAAI press Menlo Park.

Freund, Y., & Mason, L. (1999). The alternating decision tree learning algorithm. In *icml*, Vol. 99, pp. 124–133. Citeseer.

Gu, W., Foster, K., Shang, J., & Wei, L. (2019). A game-predicting expert system using big data and machine learning. *Expert Systems with Applications*, *130*, 293–305.

Haghighat, M., Rastegari, H., & Nourafza, N. (2013). A review of data mining techniques for result prediction in sports. *Advances in Computer Science: an International Journal*, *2*(5), 7–12.

Hall, M. A. (1998). Correlation-based feature subset selection for machine learning. *Thesis submitted in partial fulfillment of the requirements of the degree of Doctor of Philosophy at the University of Waikato*.

Hamming, R. W. (1950). Error detecting and error correcting codes. *The Bell system technical journal*, *29*(2), 147–160.

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

Huang, K.-Y., & Chang, W.-L. (2010). A neural network method for prediction of 2006 world cup football game. In *The 2010 international joint conference on neural networks (IJCNN)*, pp. 1–8. IEEE.

Hubáček, O., Šourek, G., & Železnỳ, F. (2019a). Exploiting sports-betting market using machine learning. *International Journal of Forecasting*, *35*(2), 783–796.

Hubáček, O., Šourek, G., & Železnỳ, F. (2019b). Learning to predict soccer results from relational data with gradient boosted trees. *Machine Learning*, *108*(1), 29–47.

Hucaljuk, J., & Rakipović, A. (2011). Predicting football scores using machine learning techniques. In *2011 Proceedings of the 34th International Convention MIPRO*, pp. 1623–1627. IEEE.

Hughes, M. D., & Bartlett, R. M. (2002). The use of performance indicators in performance analysis. *Journal of sports sciences*, *20*(10), 739–754.

Ishi, M. S., & Patil, J. B. (2021). A study on machine learning methods used for team formation and winner prediction in cricket. In Smys, S., Balas, V. E., Kamel, K. A., & Lafata, P. (Eds.), *Inventive Computation and Information Technologies*, pp. 143–156, Singapore. Springer Singapore.

Ivanković, Z., Racković, M., Markoski, B., Radosav, D., & Ivković, M. (2010). Analysis of basketball games using neural networks. In *2010 11th International Symposium on Computational Intelligence and Informatics (CINTI)*, pp. 251–256. IEEE.

James, B. (1984). *The Bill James Baseball Abstract, 1984*. Ballantine Books New York.

Joseph, A., Fenton, N. E., & Neil, M. (2006). Predicting football results using bayesian nets and other machine learning techniques. *Knowledge-Based Systems*, *19*(7), 544–553.

Kahn, J. (2003). Neural network prediction of nfl football games. *World Wide Web electronic publication*, 9–15.

Keshtkar Langaroudi, M., & Yamaghani, M. (2019). Sports result prediction based on machine learning and computational intelligence approaches: A survey. *Journal of Advances in Computer Engineering and Technology*, *5*(1), 27–36.

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

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*(9), 1464–1480.

Loeffelholz, B., Bednar, E., & Bauer, K. W. (2009). Predicting nba games using neural networks. *Journal of Quantitative Analysis in Sports*, *5*(1).

Maszczyk, A., Gołaś, A., Pietraszewski, P., Roczniok, R., Zajac, A., & Stanula, A. (2014). Application of neural and regression models in sports results prediction. *Procedia-Social and Behavioral Sciences*, *117*, 482–487.

McCabe, A., & Trevathan, J. (2008). Artificial intelligence in sports prediction. In *Fifth International Conference on Information Technology: New Generations (itng 2008)*, pp. 1194–1197. IEEE.

Miljković, D., Gajić, L., Kovačević, A., & Konjović, Z. (2010). The use of data mining for basketball matches outcomes prediction. In *IEEE 8th International Symposium on Intelligent Systems and Informatics*, pp. 309–312. IEEE.

Min, B., Kim, J., Choe, C., Eom, H., & McKay, R. B. (2008). A compound framework for sports results prediction: A football case study. *Knowledge-Based Systems*, *21*(7), 551–562.

Mitchell, S. A. (1996). Improving invasion game performance. *Journal of Physical Education, Recreation & Dance*, *67*(2), 30–33.

Odachowski, K., & Grekow, J. (2012). Using bookmaker odds to predict the final result of football matches. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pp. 196–205. Springer.

O'Donoghue, P., & Williams, J. (2004). An evaluation of human and computer-based predictions of the 2003 rugby union world cup..

Oliver, D. (2002). Basketball on paper. brassey's..

O'Donoghue, P., Ball, D., Eustace, J., McFarlan, B., & Nisotaki, M. (2016). Predictive models of the 2015 rugby world cup: accuracy and application. *International Journal of Computer Science in Sport*, *15*(1), 37–58.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web.. Tech. rep., Stanford InfoLab.

Pathak, N., & Wadhwa, H. (2016). Applications of modern classification techniques to predict the outcome of odi cricket. *Procedia Computer Science*, *87*, 55–60.

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

Purucker, M. C. (1996). Neural network quarterbacking. *IEEE Potentials*, *15*(3), 9–15.

Quinlan, J. R. (1993). C4. 5: programs for machine learning..

Rajšp, A., & Fister, I. (2020). A systematic literature review of intelligent data analysis methods for smart sport training. *Applied Sciences*, *10*(9).

Razali, N., Mustapha, A., Utama, S., & Din, R. (2018). A review on football match outcome prediction using bayesian networks. In *Journal of Physics: Conference Series*, Vol. 1020, p. 012004. IOP Publishing.

Read, B., & Edwards, P. (1992). Teaching children to play games. *Leeds: White Line Publishing*.

Reed, D., & O'Donoghue, P. (2005). Development and application of computer-based prediction methods. *International Journal of Performance Analysis in Sport*, *5*(3), 12–28.

Rotshtein, A. P., Posner, M., & Rakityanskaya, A. (2005). Football predictions based on a fuzzy model with genetic and neural tuning. *Cybernetics and Systems Analysis*, *41*(4), 619–630.

Rudrapal, D., Boro, S., Srivastava, J., & Singh, S. (2020). A deep learning approach to predict football match result. In *Computational Intelligence in Data Mining*, pp. 93–99. Springer.

Schumaker, R. P., Solieman, O. K., & Chen, H. (2010). *Sports data mining*, Vol. 26. Springer Science & Business Media.

Shi, Z., Moorthy, S., & Zimmermann, A. (2013). Predicting ncaab match outcomes using ml techniques-some results and lessons learned. In *ECML/PKDD 2013 Workshop on Machine Learning and Data Mining for Sports Analytics*.

Somboonphokkaphan, A., & Phimoltares, S. (2009). Tennis winner prediction based on time-series history with neural modeling. In *Proceedings of the International Multi-Conference of Engineers and Computer Scientists*, Vol. 1.

Tax, N., & Joustra, Y. (2015). Predicting the dutch football competition using public data: A machine learning approach. *Transactions on Knowledge and Data Engineering*, *10*(10), 1–13.

Thabtah, F., Zhang, L., & Abdelhamid, N. (2019). Nba game result prediction using feature analysis and machine learning. *Annals of Data Science*, *6*(1), 103–116.

Tilp, M., & Schrapf, N. (2015). Analysis of tactical defensive behavior in team handball by means of artificial neural networks. *IFAC-PapersOnLine*, *28*(1), 784–5.

Tsakonas, A., Dounias, G., Shtovba, S., & Vivdyuk, V. (2002). Soft computing-based result prediction of football games. In *The First International Conference on Inductive Modelling (ICIM'2002). Lviv, Ukraine*. Citeseer.

Valero, C. S. (2016). Predicting win-loss outcomes in mlb regular season games–a comparative study using data mining methods. *International Journal of Computer Science in Sport*, *15*(2), 91–112.

Van Eetvelde, H., Mendonca, L. D., Ley, C., Seil, R., & Tischer, T. (2021). Machine learning methods in sport injury prediction and prevention: a systematic review. *Journal of experimental orthopaedics*, *8*(1), 1–15.

Weissbock, J., & Inkpen, D. (2014). Combining textual pre-game reports and statistical data for predicting success in the national hockey league. In *Canadian Conference on Artificial Intelligence*, pp. 251–262. Springer.

Weissbock, J., Viktor, H., & Inkpen, D. (2013). Use of performance metrics to forecast success in the national hockey league.. In *MLSA@ PKDD/ECML*, pp. 39–48.

Wirth, R., & Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, pp. 29–39. Citeseer.

Wiseman, O. (2016). *Using Machine Learning to Predict the Winning Score of Professional Golf Events on the PGA Tour*. Ph.D. thesis, Dublin, National College of Ireland.

Witten, I. H., Frank, E., Trigg, L. E., Hall, M. A., Holmes, G., & Cunningham, S. J. (1999). Weka: Practical machine learning tools and techniques with java implementations..

Zdravevski, E., & Kulakov, A. (2009). System for prediction of the winner in a sports game. In *International Conference on ICT Innovations*, pp. 55–63. Springer.