Received: 22 May 2019 | Revised: 11 May 2020 | Accepted: 25 May 2020
DOI: 10.1002/widm.1380

ADVANCED REVIEW

WIREs DATA MINING AND KNOWLEDGE DISCOVERY WILEY

The use of machine learning in sport outcome prediction: A review

Tomislav Horvat[1] | **Josip Job**[2]

[1]Department of Electrical Engineering, University North, Varaždin, Croatia
[2]Faculty of Electrical Engineering, Computer Science, and Information Technology Osijek, Josip Juraj Strossmayer University of Osijek, Osijek, Croatia

Correspondence
Tomislav Horvat, Department of Electrical Engineering, University North, 104, Brigade 3, Varaždin, Croatia.
Email: tomislav.horvat@unin.hr

Abstract

The increase in the volume of structured and unstructured data related to more than just sport events leads to the development and increased use of techniques that extract information and employ machine-learning algorithms in predicting process outcomes based on input but not necessarily output data. Taking sports into consideration, predicting outcomes, and extracting valuable information has become appealing not only to sports workers but also to the wider audience, particularly in the areas of team management and sports betting. The aim of this article is to review the existing machine learning (ML) algorithms in predicting sport outcomes. Over 100 papers were analyzed and only some of these papers were taken into consideration. Almost all of the analyzed papers use some sort of feature selection and feature extraction, most often prior to using the machine-learning algorithm. As an evaluation method of ML algorithms, researchers, in most cases, use data segmentation with data being chronologically distributed. In addition to data segmentation, researchers also use the *k*-cross-evaluation method. Sport predictions are usually treated as a classification problem with one class being predicted and rare cases being predicted as numerical values. Mostly used ML models are neural networks using data segmentation.

This article is categorized under:
Technologies > Machine Learning
Technologies > Prediction

KEYWORDS
learning algorithms, machine learning, outcome prediction, sport

1 | INTRODUCTION

A huge amount of (un)structured data related to sport events is produced nowadays, which, in turn, leads to an increased interest in sports by general public. In addition to analytical purposes, a large amount of these data is used for predicting future events. Predictions can be made based on past experience or data, that is, based on acquired data leading to the concept of machine learning (ML). In sports, ML algorithms aim to assist coaches and sport managers not only to predict a game outcome but also a player's or team's performance, possible player's injury, scouting, and making sport-betting decisions.

In 1959, Arthur Samuel, an American pioneer in the fields of computer gaming and artificial intelligence, defined ML as a field of study that enables computers to learn without being explicitly programmed (Samuel, 1959). A recent definition defines ML as a process of programming computers to optimize the performance criterion using example data or past experience (Alpydin, 2010). The main task of an ML algorithm is to predict an outcome of an observed process and build a model which will be useful for data approximation. ML is closely related to computational statistics with the background in probability theory, linear algebra, information theory, and cognitive sciences. ML tasks are typically classified into three categories—supervised learning, unsupervised learning, and reinforcement learning (Alpydin, 2010; Kotsiantis, 2007). The goal of supervised learning is to develop a predictive model that, based on both input and output data, predicts future events on previously unseen data. In unsupervised learning, unlike supervised learning, the main goal is to group and interpret data based only on input, so-called unlabeled data. In other words, the goal of unsupervised learning is to find data regularities. Reinforcement learning is a different category of ML (Barto & Sutton, 1997). The training information is provided to the learning system by the environment in the form of a scalar reinforcement signal that constitutes a measure of how well the system operates. The learner is not directly told what to do but rather to figure out which actions yield the best reward (Kotsiantis, 2007). The types of ML algorithms and their short descriptions are illustrated in Figure 1.

Sport predictions are usually treated as a classification problem by which one class is predicted (win/loss/draw) (Prasetio & Harlili, 2016) and rare cases are predicted by numerical values, such as predicting a scoreline or spread. A scoreline is an exact number of points scored by a team while a spread is the difference between scored points. Predicting a scoreline or spread is certainly a much more challenging task than predicting outcomes. Predicting outcomes involving two possible classes is less challenging than predicting outcomes involving three possible classes. Likewise, it is also easier to predict the outcome of individual sports compared to team sports in which the performance of a larger participants' number affects the final outcome. Delen et al. (2012), Soto Valero (2016), and Elfrink (2018) showed that the classification-type models predict the game outcomes better than regression-based classification models. Feature selection and extraction have an important role in the accuracy of ML algorithm. The aim of both approaches is to increase the accuracy of the model, make it easier to understand and visualize the data. In theory, the goal of a feature selection and feature extraction is to find an optimal feature subset (one that maximizes the model accuracy). In this article, we give a broader literature review in the field of sport outcome predictions using classification, analysis, and comparison. For this purpose, several review papers in the related fields have been analyzed (Bunker & Susnjak, 2019; Bunker & Thabtah, 2019; Haghighat et al., 2013; Jović et al., 2015; Koseler & Stephan, 2018). Haghighat et al. (2013) reviewed scientific papers related to sport outcome predictions based on data mining techniques. The authors analyzed the research results and outlined the advantages and disadvantages of individual prediction models. Two major problems were detected. First, low prediction accuracy highlighted the need for further research to achieve reliable prediction results. Second, the difference in used datasets prevents the researchers from comparing achieved results to other researchers' results. Jović et al. have provided a comprehensive overview of the feature selection papers. The paper covers feature selection methods including filters (selecting features based on performance measure regardless of the employed data modeling algorithm), wrappers (feature selection based on the quality of the modeling algorithm performance), embedded (performing feature selection during the modeling algorithm's execution), and hybrid methods. A feature selection is a very important step in ML and its main purpose is to achieve better performance by reducing the

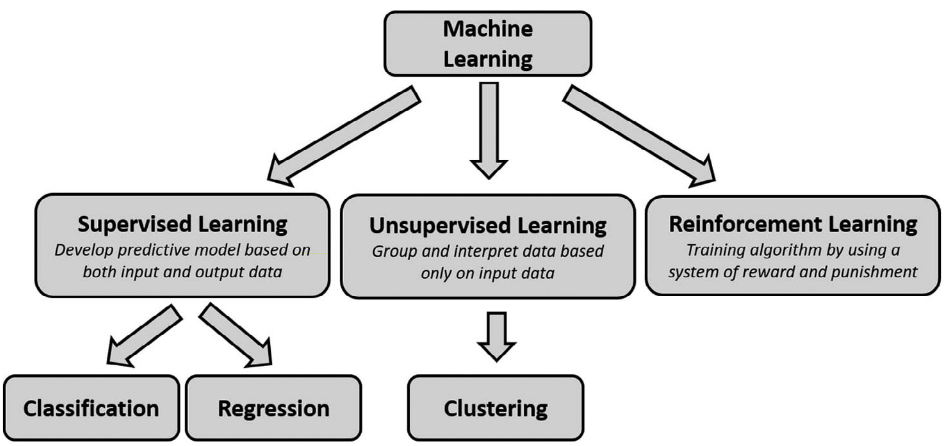

FIGURE 1 The types of ML

redundant features. Koseler and Stephan (2018) perform a systematic literature review of using ML applications in baseball analytics. A total of 32 articles were analyzed. The authors found that two ML algorithms dominate the literature, namely *Support Vector Machine* (SVM) and *k-nearest neighbors* (k-NN) for both classification and regression problems. They also concluded that neural networks, because of their efficiency, will soon become the most used ML model in baseball predictions. Bunker and Susnjak (2019) provided a review of papers using ML to predict results in team sports. The ultimate goal was to answer key research questions regarding sports prediction. The authors concluded that the majority of studies (65%) considered artificial neural networks in their experiments, and that artificial neural networks do not necessarily perform better than other ML algorithms. It was also concluded that earlier researches often selected predictive features manually, usually based on researchers' knowledge and that having a large dataset does not necessarily lead to high accuracy. They concluded that low-scoring sports and sports involving a greater number of possible outcomes provide generally lower accuracies. Authors cite the problem of comparing results, because researches differ in at least one of the following dimensions: Considered sport, the input dataset, the model predictors, or considered competitions. Bunker and Thabtah (2019) noted the problem of using the same prediction models for multiple sports. Besides providing a critical survey of the literature on ML for sport prediction, authors concluded that it is not possible to directly apply a predictive model to a different sport. Therefore, they proposed a novel sport prediction framework, so-called sports result prediction "SRP-CRISM-DM" framework, as an extension of CRISP-DM methodology, through which ML could be used as a learning strategy. In our paper, the emphasis is placed on predicting outcomes in multiple team sports and, unlike in paper by Haghighat et al. (2013), up-to-date data is used. Since a feature selection is an important step in ML, it is vital to explore available feature selection methods.

In our paper, a literature survey and other research results on the ML algorithms in sport outcome prediction are presented and discussed. Over 100 papers concerning sport predictions or extracting useful facts and regularities related to sports have been found. Consideration is given to papers that concern at least one ML algorithm in team sport outcome prediction or papers that extract useful facts and regularities from the input dataset. To better understand the significance of the present review paper, it is necessary to specify the criteria for including/excluding papers. The possible reasons for the exclusion of papers were difficulty in data interpretation, poor work methodology, or lack of clarity. Specific papers with previously mentioned characteristics were included for pointing out possible mistakes in predicting sport outcomes.

Section 2 gives an insight into previous ML research related to sport outcome prediction and brief information about the used datasets and data sources. Section 3 depicts methods used for the feature selection and feature extraction while Section 4 presents the used ML algorithms and obtained results. Finally, Section 5 concludes the paper.

2 | DATA COLLECTION

As stated earlier, huge amounts of (un)structured data related to sport events are being generated. Besides increasing the amount of data, the number of relevant sport statistic databases is also expanding. By analyzing the work of other researchers, it is easy to notice that information sources are usually the official sport organization websites. Table 1 shows data sources, used datasets and used ML algorithms in predicting sport outcomes.

Table 1 chronologically lists papers related to team sports outcome prediction. A total of 39 papers from the period of 1996 to 2020 that cover five team sports were considered and analyzed. Most of the papers are related to basketball, followed by soccer, football, baseball, and cricket. Additionally, the most popular basketball league in predicting outcomes is NBA (75.00%), while the most popular soccer league is EPL (55.56%). The most popular football league is NFL and it includes a total of 80% of the football-related papers. Other leagues appear twice or less than 10%. Most papers involve multiple ML algorithms so the analyzed sample contains more than 100 outcome prediction results. If the authors use different feature sets or datasets, only the best results related to the ML algorithm are analyzed. Result comparison is difficult because researchers use different datasets and leagues of different competitiveness.

The most commonly used ML model in predicting game outcomes are neural networks, for example, feed-forward, multilayer perceptron, convolutional, radial basis function, probabilistic, and so on. Figure 2 shows the number of papers using particular ML algorithm group in predicting game outcomes regardless of sport. ML algorithms are grouped according to their similarities. All variants of neural networks are categorized under *Neural networks*. *Naïve Bayes* and *Bayesian network* are also grouped, as well as all the boosting algorithms (LogitBoost, AdaBoost, XGBoost, etc.). When multiple ML algorithms are used in a particular paper, the best results of each algorithm are included.

TABLE 1 Analyzed papers categorized by the publication year

#	Dataset	Data sources	Type of ML model
Purucker (1996)	National Football League (NFL), season 1994, weeks 11–16 (football)	–	Neural network
Kahn (2003)	NFL, first 15 weeks of season 2013 (football)	NFL official website	Neural network
Hamadani (2006)	NFL, seasons 2003–2005 (football)	Different sources	Logistic regression, SVM
McCabe and Travathan (2008)	English premier league (EPL) and Australian football league (AFL), seasons 2002–2007 (soccer, football)	Different sources	Neural network
Loeffelholz et al. (2009)	The first 650 games of the season 2007 (basketball)	ESPN	Neural network
Miljković et al. (2010)	National Basketball Association (NBA), regular season 2009 (basketball)	NBA official website, yahoo sports	Naïve Bayes, linear regression, k-NN, decision trees, SVM
Zdravevski & Kulakov (2010)	NBA, two consecutive seasons (basketball)	NBA official website	37 ML algorithms
Trawinski (2010)	Asociacion de Clubes de Baloncesto (ACB), season 2008 (basketball)	ACB official website	10 fuzzy rule-based models
Ivanković et al. (2010)	Serbia, first B basketball league, seasons 2005–2009 (basketball)	Basketball Federation of Serbia	Neural network
Buursma (2011)	The last 15 years of soccer in the Netherlands (soccer)	www.football-data.co.uk	Linear/logistic regression, decision trees, LogitBoost, Bayesian network, Naïve Bayes
Blaikie et al. (2011)	NFL and college football, seasons 2003–2010 (football)	ESPN, USA today	Neural network
Hucaljuk and Rakipović (2011)	Champions League (soccer)	Different sources	Naïve Bayes, Bayesian network, LogitBoost, k-NN, random forest, neural network
Cao (2012)	NBA, seasons 2005–2010 (basketball)	NBA official website, basketball reference, Databasebasketball.com	Logistic regression, Naïve Bayes, SVM, neural network
Delen et al. (2012)	College football, seasons 2002–2010 (football)	Different sources	Decision trees, neural network, SVM
Kravanja (2013)	NBA, seasons 2007–2009 (basketball)	Basketball geek	SVM, logistic regression
Torres (2013)	NBA, regular seasons 2006–2012 (basketball)	Basketball reference	Linear regression, maximum likelihood classifier, multilayer perceptron—Back propagation
Zimmermann et al. (2013)	National Collegiate Athletic Association Basketball (NCAAB), seasons 2008–2013 (basketball)	kenpom.com	J48, random forest, Naïve Bayes, multilayer perceptron neural network
Owramipur et al. (2013)	Spain soccer league, season 2008 (soccer)	Football356, soccer-Spain, stat-football	Bayesian network
Lin et al. (2014)	NBA, seasons 1991–1997 (basketball)	NBA official website	Logistic regression, adaptive boost, random forest, SVM, Naïve Bayes
Igiri and Nwachukwu (2014)	EPL, season 2014 (soccer)	–	Neural network, logistic regression
Kampakis and Thomas (2015)	English 20 over country cricket cup, seasons 2009–2014 (cricket)	www.cricinfo.com	Naïve Bayes, logistics regression, random forests, gradient boosted decision trees
Tax and Joustra (2015)	Dutch Eredivisie, seasons 2000–2012 (soccer)	Different sources	Naïve Bayes, LogitBoost, neural network, random forest, genetic programming

TABLE 1 (Continued)

#	Dataset	Data sources	Type of ML model
Avalon et al. (2016)	NBA, season 2013/2014 (basketball)	Basketball reference	Linear regression, Gaussian discriminant analysis, SVM, random forest, adaptive boost
Prasetio and Harlili (2016)	EPL, seasons 2010–2015 (soccer)	www.football-data.co.uk, sofifa.Com/teams	Logistic regression
Soto Valero (2016)	Regular major league basketball (MLB), seasons 2005–2014 (baseball)	Retrosheet, Lahman database	k-NN, neural network, SVM, decision trees
Cheng et al. (2016)	NBA, seasons 2007–2014 (basketball)	NBA official website	Maximum entropy model
Tran (2016)	NBA, seasons 1985–2016 (basketball)	NBA official website, basketball reference	Matrix factorization
Ping-Feng et al. (2017)	NBA, seasons 2008–2010 (basketball)	Basketball reference, NBA official website	SVM
Mustafa et al. (2017)	Indian Premier League 2014/2015, World Cup 2015 (cricket)	Twitter, cricinfo.com, cricbuzz.com and official team pages	SVM, Naïve Bayes, linear regression
Horvat et al. (2018)	Euroleague (EL), seasons 2012–2017 (basketball)	EL official website	k-NN
Elfrink (2018)	MLB, seasons 1930–2016 (baseball)	Retrosheet	Random forest, XGBoost, logistic regression, GLM
Lam (2018)	NBA, seasons 2013–2014 (basketball)	Basketball reference	Bayesian regression
Ganguly and Frank (2018)	NBA, season 2002–2016 (basketball)	–	Neural network
Zaveri et al. (2018)	Spain soccer league, seasons 2012–2016 (soccer)	Different sources	Logistic regression, random forest, neural network, SVM, Naïve Bayes
Horvat and Job (2019)	NBA, seasons 2009–2017 (basketball)	Basketball reference	naïve ML algorithm
Hubáček et al. (2019)	NBA, seasons 2000–2014 (basketball)	NBA official website	Neural network, logistic regression
Knoll and Stübinger (2020)	EPL, Ligue 1, Bundesliga, Seria A, Primera division, seasons from 2013 to 2017 (soccer)	Sportal	Random forest, gradient boosting algorithm, SVM, linear regression
Stübinger et al. (2020)	5 European soccer leagues and the corresponding second divisions, seasons from 2006 to 2017 (soccer)	FIFA index	Random forest, boosting algorithm, SVM, linear regression

In addition to standard ML algorithms, researchers also use other methods such as fuzzy rule-based models, entropy models, matrix factorization models, and so on. Figure 3 emphasizes a growth of interest among researchers in the field of sport outcome predictions. An increased interest in predicting sport outcomes started around 2010. In addition to comparing the final predicting outcomes in sports, this review paper provides an insight into the use of feature selection and feature extraction methods. Any paper of the analyzed that uses some type of feature selection or feature extraction is analyzed.

3 | FEATURE SELECTION AND EXTRACTION

Feature selection and feature extraction represent a very important preliminary step for the ML algorithm because it significantly contributes to the algorithm's success. In brief, feature selection is the process of identifying and removing as many irrelevant and redundant features as possible with the purpose of reducing problem dimensionality and to enable algorithms to operate faster and more effectively (Grossberg, 1988; Han & Kamber, 2006; Yu & Liu, 2004).

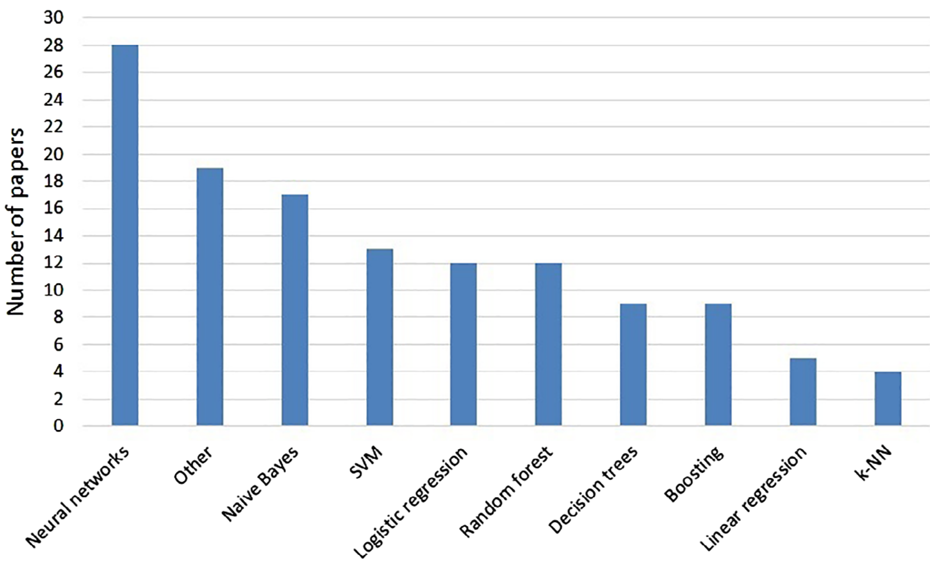

FIGURE 2 Number of papers using a particular ML algorithm group

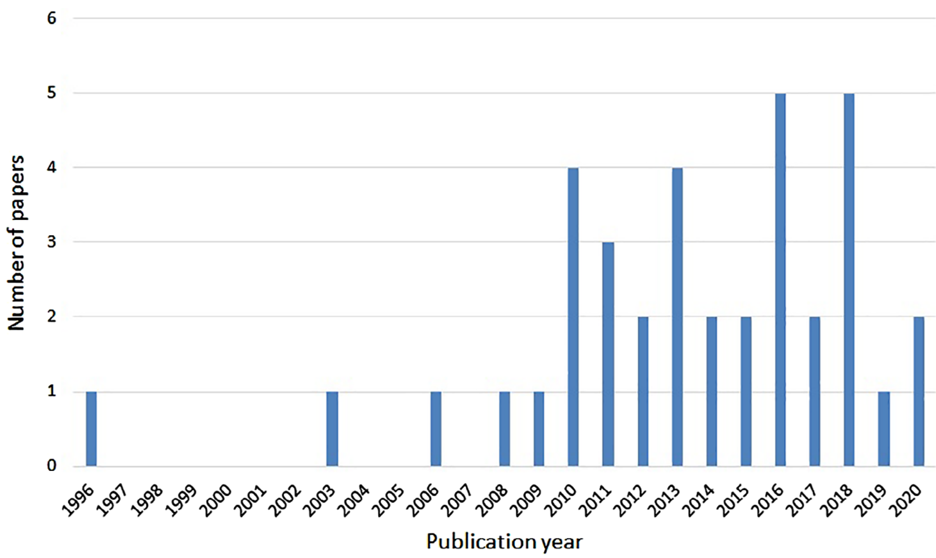

FIGURE 3 Number of analyzed papers by the publication year

Feature extraction is the process of dimensionality reduction by which an initial set of raw data is reduced to more manageable groups for processing (Zhang, 2000). Variables are combined into features effectively reducing the amount of data that must be processed while still accurately and completely describing the original dataset. Feature selection and feature extraction techniques increase the efficiency of the ML algorithm and speed up algorithm performance, which is crucial in analyzing large datasets. The input feature set usually contains some of the features suggested by experts' experience. The features are a combination of other known, publicly available features. These features usually refer to the league standings or represent one of the required, usually expert suggested features.

Most authors use the feature selection and feature extraction methods in the data preparation phase. Authors first define the input feature set based on expert experience and, if necessary, using the feature extraction method, calculate the missing values of defined input features. Research comparing the effect of the ML algorithm before and after the use of the feature selection are very rare. The papers summarized in the next chapters describe the effects of the feature selection and feature extraction methods. Next sections are divided according to the feature selection method, namely, the section in which feature selection is based only on expert's experience and the section which uses more sophisticated feature selection methods.

3.1 | Feature selection based on experts' experience

Each analyzed paper involves the feature selection in the form of selecting an initial feature set(s). This section elaborates only on papers that use the additional feature selection based on the author's experience or a combination of initial feature sets. Each analyzed paper involves the feature selection in the form of selecting an initial feature set(s).

Hucaljuk and Rakipović (2011) used several ML algorithms to predict outcomes in Champions' League group stage games and achieved maximum accuracy of 68.8% using the artificial neural network. The authors used two feature sets (basic and expert) on three different datasets. The expert feature set achieved better accuracy only when using LogitBoost and artificial neural networks. Ganguly and Frank (2018) used four sets of input features—play-by-play feature set, box score feature set, pregame line-ups feature set, and on-courts labels feature set containing a total of 352 features and using a different combination of feature sets for the outcome prediction. Best in-season (internal) accuracy of 88.1% was achieved by using the advanced play-by-play feature set, box score feature set, and neural network the last of which consists of three fully connected encoding layers. The best out of season (external) accuracy of 82% was achieved by using MDN (Mixture Density Network) model where all feature sets were included. MDN model was trained on 4,440 games from 2002 to 2014 seasons and tested in 500 games from 2002 to 2014 (internal) and 500 games from the ones held out during 2016–2017 season (external).

A relatively small number of papers use additional feature selection based on experts experimenting or a combination of initial feature sets, with the purpose of finding optimal feature subsets to maximize the proposed model accuracy. Optimizing the feature sets achieved better prediction.

3.2 | Feature selection based on feature selection methods

More sophisticated feature selection methods are used to maximize relevance and minimize redundancy. Feature selection includes calculating a feature subset containing only relevant features. Feature selection methods are usually classified into filters, wrappers, embedded, and hybrid methods (Jović et al., 2015). This section presents papers that use more sophisticated feature selection methods to improve the prediction results.

Loeffelholz et al. (2009) used the Signal-to-Noise Ratio (SNR) feature selection method and selected 4 out of 22 most representative features from the initial feature set. The reduced feature set yields better results than those of the initial feature set. The authors also claim that due to the use of a small feature set, dimensionality reduction techniques such as Principal Component Analysis (PCA) are not required. Buursma (2011) used an initial set of 10 features eliminating one feature at the time. If the percentage drops, the feature is deemed important and should be left in, and if the percentage rises, the feature is discarded. A total of eight features were selected at this stage. A new set of eight features were added one by one to get the final set of nine features. The extra feature was not accepted unless accuracy was increased. Kravanja (2013) used backward elimination methods for feature selection. For quality measures, Pearson correlation, information gain, and ReliefF method were selected. The author used four sets of features and two ML algorithms, SVM, and logistic regression for predicting NBA basketball game outcomes. SVM achieved the best accuracy of 70.01% using the initial feature set while logistic regression achieved the maximum accuracy of 69.73% using the best relative feature set. In the paper by Lin et al. (2014), the authors initially trained and tested all chosen learning models (logistic regression, SVM, AdaBoost, random forest, and Gaussian Naïve Bayes [GNB]) on 17-dimensional feature vectors and quickly realized that besides logistic regression, all other models suffered from overfitting, and low prediction accuracies. To avoid overfitting, the authors used three feature selection methods. They used forward and backward search for adding or removing features one by one to determine which features result in the highest prediction accuracies and consequently defined a final set of seven features. A heuristic feature selection algorithm was also used to verify that the selected features are the most informative about whether a team will win. The features selected by a backward search were almost the same as those selected by the heuristic method. The best accuracy of 65.15% was achieved by a random forest method. By taking out the win records from the training dataset, model accuracy dropped so the authors suggest that box score statistics may not capture all the elements of a winning team and that further research must be done. Dividing up a season into quartiles resulted in an improvement to 68.75% with logistic regression in the final quartile. The authors have also defined two naïve win prediction methods. The first prediction method was based on the greater difference between average points scored and average points allowed per game (achieved accuracy of 63.5%) while the second method was based on choosing the winning team based on a better win rate (achieved accuracy of 60.8%). In addition, the authors also considered accuracy of 71% which was achieved by the expert's opinion. In the paper by Kampakis and Thomas (2015), the authors used Chi-square, mutual information, Pearson correlation, and recursive feature elimination as a feature selection method from the initial set of 500 features. The selected features were used as input in various ML algorithms. The authors also used PCA to improve the model performance and Naïve Bayes, logistic regression, random forests, gradient boosted decision trees to predict the outcome of English County 20 over Cricket Matches. The performance of each predictive algorithm was assessed during the previous year's data as the training dataset and the year in question as to the test. Tax and Joustra (2015) used different dimensionality reduction methods like PCA, Sequential Forward Selection, ReliefF, and Correlation-based Feature Subset Selection for feature selection. They used two different feature subsets (public data feature set and betting odds feature set). Ping-Feng et al. (2017) used CFS algorithm—a feature subset ranking method for evaluating the importance of a feature subset—for selecting 7 of 17 features. The authors proposed a model based on SVM supported by a decision tree and using a CFS feature selection algorithm which achieved the accuracy of 85.25%. Without feature selection, the authors achieved the accuracy of 67%. Horvat et al. (2018) proposed a model for feature selection based on the feature information gain. The authors presented two feature selection variants and two data preparation algorithms. The first feature selection variant uses features whose information gain is higher than zero while the second feature selection variant uses features whose information gain is higher than the average feature information gain of a particular team. The authors proposed a learning model based on k-NN for predicting Euroleague games. They applied several models using different k and number of seasons. The best accuracy of 83.96% was achieved with $k = 3$ for the dataset of three seasons and the dataset of one season and $k = 5$ or $k = 7$.

It is difficult to conclude which feature selection method provides the best results. The most commonly used are filter feature selection methods which select features based on a performance measure regardless of the modeling algorithm. Filter feature selection methods characterize dimensionality reduction before using the modeling algorithm. The comparison of the results before and after feature selection is not shown in most articles, but the assumption is that results are better after using feature selection. Table 2 shows that the number of features varies from research to research. Most authors use feature selection and feature extraction methods and the average number of used features, including outliers, is approximately 57. After using feature selection and feature extraction, the number of selected features, including outliers, drops to approximately 39. The initial feature set is defined based on the researcher's experience and thus includes feature selection. Researchers generally use basic game statistics related to a particular sport to predict outcomes and often include league standing features obtained by feature extraction. Other feature sets such as psychological state or social media data are rarely used. Many researchers point out that increasing the feature set could lead to better prediction results. Additional features may certainly be presuggested feature sets such as the psychological state or social media information. Both approaches represent a novelty in the field of predicting sport outcomes. The analysis of the papers revealed that early papers in the field of predicting sport outcomes often used only feature selection by the authors themselves (based on their knowledge or past experience). Recently, more and more authors use more sophisticated feature selection methods. The general conclusion evident from the analyzed papers is that feature selection and feature extraction certainly contribute to increasing the efficiency of the ML models. Most papers do not show a comparison table with prediction results using the initial feature set and results obtained by using some type of feature selection. The papers with the presented improvements of the prediction results using additional feature selection are Loeffelholz et al. (2009), Trawinski (2010), Buursma (2011), Ping-Feng et al. (2017), Ganguly and Frank (2018), and Horvat et al. (2018). Other researchers only note that additional feature selection contributes to better prediction results. Ideally, several different feature sets should be tested when proposing an outcome prediction model and, accordingly, the one that returns the best result should be selected. The assumption is that using wrapper or embedded feature selection methods could lead to improved prediction results. Embedded filter selection methods that perform feature selection during the modeling algorithm's execution, hybrid feature selection methods that would include elements of the filter feature selection methods, and the wrapper filter selection methods are of particular interest.

4 | EVALUATION OF THE RESULTS

Neural networks have been the mostly used model since the beginning of using the ML in predicting the sport outcome (Grossberg, 1988; Zhang, 2000). Even though there are many new ML models, the trend of using neural networks continues with neural networks still being the most widely used ML model in predicting sport outcomes. This section will provide an insight into the results of using ML algorithms in predicting sport outcomes. The list of analyzed papers and used ML algorithms is shown in Table 1.

TABLE 2 Used ML algorithms and the best result for the analyzed papers

References	Best accuracy (%)	Type of ML model	Seasons	Number of features	Feature selection	Feature extraction	Model evaluation
Igiri and Nwachukwu (2014)	93.00	Logistic regression	1	9	✓	✓	Segmentation Dataset
Owramipur et al. (2013)	92.00	Bayesian network	1	13	✓	✓	Segmentation Dataset
Mustafa et al. (2017)	87.90	SVM	1	3	✓	✓	10-fold CV
Delen et al. (2012)	86.48	Decision trees	9	28	✓	✓	10-fold CV
Lam (2018)	85.28	Bayesian regression	1	17	✓	x	Segmentation Dataset
Ping-Feng et al. (2017)	85.25	SVM	3	7/17	✓	x	Segmentation Dataset
Horvat et al. (2018)	83.96	k-NN	3	15	✓	x	Segmentation Dataset
Ganguly and Frank (2018)	82.00	Neural network	15	352	✓	x	Segmentation Dataset
Stübinger et al. (2020)	81.77	Integrated model	12	40	✓	✓	Segmentation Dataset
Ivanković et al. (2010)	80.96	Neural network	5	9	✓	x	Segmentation Dataset
Purucker (1996)	78.60	Neural network	1	5	✓	x	Segmentation Dataset
Knoll and Stübinger (2020)	75.62	Random forest	5	39	✓	✓	Segmentation Dataset
Kahn (2003)	75.00	Neural network	1	5	✓	✓	Segmentation Dataset
Zimmermann et al. (2013)	74.46	Multilayer perceptron neural network	5	–	✓	x	Segmentation Dataset
Cheng et al. (2016)	74.40	Maximum entropy model	8	28	✓	x	Segmentation Dataset
Loeffelholz et al. (2009)	74.33	Neural network	1	4/22	✓	x	Segmentation Dataset
Zdravevski and Kulakov (2010)	72.80	Logistic regression	2	10	✓	✓	Segmentation Dataset
Zaveri et al. (2018)	71.63	Logistic regression	5	13	✓	✓	10-fold CV
Trawinski (2010)	71.50	Class-fuzzy-chi-RW algorithm	1	5/15	✓	✓	10-fold CV
Tran (2016)	70.95	Matrix factorization	2	–	✓	x	Segmentation Dataset
Kravanja (2013)	70.01	SVM	3	39	✓	✓	3-fold CV
Cao (2012)	69.67	Logistic regression	6	46	✓	✓	Segmentation Dataset
Prasetio and Harlili (2016)	69.50	Logistic regression	6	4	✓	✓	Segmentation Dataset
Hubáček et al. (2019)	68.83	Neural network	15	–	✓	✓	Segmentation Dataset
Hucaljuk and Rakipović (2011)	68.80	Neural network, LogitBoost	1	20/30+	✓	✓	Segmentation Dataset

(Continues)

TABLE 2 (Continued)

References	Best accuracy (%)	Type of ML model	Seasons	Number of features	Feature selection	Feature extraction	Model evaluation
Torres (2013)	68.44	Neural network	7	8	✓	✓	Segmentation Dataset
Hamadani (2006)	67.08	Logistic regression	3	30+	✓	✓	Leave-out-out CV
Miljković et al. (2010)	67.00	Naïve Bayes	1	32	✓	✓	10-fold CV
Avalon et al. (2016)	65.53	Gaussian discriminant analysis	1	218	✓	✓	20-fold CV
Lin et al. (2014)	65.15	Random forest	7	7/17	✓	✓	Segmentation Dataset
Kampakis and Thomas (2015)	62.40	Naïve Bayes	6	31/500+	✓	✓	Segmentation Dataset
Soto Valero (2016)	58.92	SVM (classification)	10	3/60	✓	✓	10-fold CV
McCabe and Travathan (2008)	58.90 (EPL), 67.20 (AFL)	Neural network	6	19	✓	✓	Segmentation Dataset
Buursma (2011)	57.00	Linear regression	15	9/18	✓	✓	10-fold CV
Tax and Joustra (2015)	56.10	LogitBoost	13	–	✓	*x*	k-fold CV, Segmentation Dataset
Elfrink (2018)	55.52	XGBoost	1	164	✓	✓	k-fold CV

4.1 | Basketball

Loeffelholz et al. (2009) used a variety of neural networks (feed-forward neural network [FFNN], Radial basis function neural network [RBFNN], Probabilistic neural network [PNN], and Generalized regression neural networks [GRNN]) and fusion of neural networks for predicting the success of basketball teams in the NBA league. The authors used four different feature sets obtained by the feature selection method or according to expert's experience. The best result of 74.33% was produced by FFNN supported with the expert feature set based on shooting statistics with four shooting features, followed by PNN and GRNN which produced the average accuracy of 73.34%, RBFNN with the average accuracy of 72%, and fusion of neural networks with the accuracies of 72.24% (PNN Fusion) and 71.57% (Bayes Fusion). The proposed neural networks were more efficient compared to the experts' results (68.67%). Miljković et al. (2010) used a variety of classification techniques such as decision trees, k-NN, Naïve Bayes and SVM for outcome prediction and multivariate linear regression for spread calculation. The database was always up-to-date with previous day data being added to the existing data in the system. The features were divided into game-related features (standard basketball statistics) and league standing features. Using 10-fold cross-validation for classifying the training and testing datasets, the authors recorded the best accuracy of 67% using Naïve Bayes in outcome prediction and 10% in spread prediction. Trawinski (2010) used 10 fuzzy models and 10-fold cross-validation to predict ACB league game outcomes. The author used two datasets. The first dataset refers to the last three games and had six features. The advanced model used eight feature selection method results and picked 5 out of 15 features that have been repeatedly selected by algorithms, included the whole season results and achieved the best accuracy of 71.5%. Ivanković et al. (2010) applied the CRISP-DM methodology and used feed-forward neural network for basketball game outcome prediction getting the accuracies of 80.96 and 66.40%, respectively, depending on the used feature set. The first feature set contained 12 features related to shooting positions while the second one contained nine features related to standard basketball statistics. In addition to outcome prediction, the authors used neural network to conclude that defensive rebounds followed by two-point shots, three-point shots, stolen balls, turnovers, offensive rebounds, free throw shots, block, and assists have the highest effect on the game outcome. Zdravevski and Kulakov (2010) used 37 algorithms in WEKA. The best result of 72.8% was achieved by logistic regression showing that the best classifiers have 5% better precision rate than the referent classifier which favors the team with better rating. The first season was used as the training dataset whereas the second one was used as the testing dataset. The authors also stated that automatization of feature selection could improve prediction results when features are selected by an expert. Cao (2012) used logistic regression, Naïve Bayes, SVM, and multilayer perceptron neural network to predict NBA basketball games. Seasons 2005/2006–2009/2010 were a training dataset and season 2010/2011 the testing dataset with the achieved accuracies of 69.97, 66.25, 67.70, and 68.01%, respectively. The feature set contained a total of 46 features selected or extracted based on the authors' experience. Kravanja (2013) used SVM and logistic regression for predicting NBA basketball games outcome and yielded the best results of 70.01 and 69.73%. Torres (2013) used multilayer perceptron—back propagation neural network, linear regression, and Maximum Likelihood Classifier to predict NBA games outcome with the achieved accuracies of 68.44, 67.98, and 66.81%. The author stated that adding additional features and player individualities could contribute to achieving better results but not significantly. Zimmermann et al. (2013) proposed a model for predicting college basketball game outcomes using decision trees (J48), random forests, Naïve Bayes method, and multilayer perceptron neural network. The authors used previous seasons as a training dataset and a single season as a testing dataset with two types of feature sets. The best accuracy of 74.46% was achieved by using the adjusted efficiency feature set achieved by multilayer perceptron for testing season 2012/2013 and training seasons 2008/2009–2011/2012. Other methods on the same dataset achieved the best results by using the adjusted efficiency feature set (decision trees—70.42%, random forest—71.37%) with the exception of Naive Bayes method where the best result, using four-factor feature sets, was 73.05%. Lin et al. (2014) used logistic regression, adaptive boost, random forest, SVM, and GNB to predict NBA games outcome with the best accuracy of 65.20% achieved by the random forest method. SVM, logistic regression, adaptive boost, and GNB produced the accuracies of 65.10, 64.70, 64.10, and 63.30%, respectively. The main conclusion of the paper is that the win record of past games plays a crucial role in basketball outcome prediction. By taking out the win records from the training dataset model, the accuracy drops so the authors suggest that box score statistics may not capture all the elements of a winning team and that further research needs to be done. Dividing up a season into quartiles resulted in an improvement to 68.75% with logistic regression in the final quartile. The authors have also defined two naïve win prediction methods. The first was based on the greater difference between the average points scored and average points allowed per game (63.5%) while the second method was based on choosing the winning team based on a better win rate (60.8%). In addition, the authors considered the accuracy of 71% which was achieved by the expert's opinion. Tran (2016) proposed a matrix factorization prediction model where a different number of seasons was used as a training dataset and season 2015/2016 for testing. The best result of 70.95% was achieved by using a single season as a training dataset. The author used a rolling learning approach with the previous season results plus played games of the analyzed season used as a training dataset. Cheng et al. (2016) applied the Maximum Entropy principle to a set of features and established the NBA Maximum Entropy model. After that, they used the model to calculate the probability of the home team's win of an upcoming game, made predictions based on this probability, and achieved the accuracy of 74.40%. In the paper by Avalon et al. (2016), the authors used various classification and regression type ML algorithms to predict NBA games outcome. The best result was yielded by Gaussian discriminant analysis (GDA) with the accuracy of 65.53% followed by linear regression (64.26%), SVM coupled with PCA (61.96%), random forest (61.36%), and adaptive boosting (60.23%). The training data included team statistics and game spreads and the output, except for GDA, was game spread. The models were tested by performing 20-fold random sample validation. Predictors output the difference between the scores of the two teams and the winning team is selected. Based on the results, the authors have also calculated top three features for every presented learning algorithm. Ping-Feng et al. (2017) proposed a model based on SVM with the support of a decision tree and using CFS feature selection algorithm achieved the accuracy of 85.25%. Without feature selection, the achieved accuracy was 67%. Lam (2018) proposed a pioneering modeling approach called one-match-ahead forecasting model based on stacked Bayesian regressions which produced the accuracy of 85.28% for predicting winning probabilities using past games and player performance. The problem is defined as a regression, therefore, the model predicts how likely the home/guest team is to win the next game. The authors concluded that the player performance in the past games is sufficient to infer the next game result. The achieved prediction was higher compared to other researcher's prediction result. In paper by Horvat et al. (2018), the authors proposed a model based on k-NN for predicting Euroleague games. They applied several models using different values of k and a number of seasons and two data preparation variants based on different feature sets. The best results, accuracy of 83.96%, were achieved with $k = 3$ for dataset of three seasons and dataset of a single season with $k = 5$ or $k = 7$. The authors also noticed that coefficient k does not have a significant impact on the model accuracy whereas feature selection and grouping of input features have. Ganguly and Frank (2018) proposed the Mixture Density Network (MDN) model and achieved the maximum in-season (internal) accuracy of 86.7% and the maximum out of season (external) accuracy of 82%. MDN model was trained on 4,440 games from seasons 2002–2014 and tested in 500 games from 2002 to 2014 (internal) and 500 games from the held out 2016–2017 season (external). The authors also highlighted three issues when predicting outcomes and those are the lack of context, no measure of prediction uncertainty, and no publicly available datasets for outcome result comparing. It is also noted that almost all currently proposed models do not consider the in-game situation such as best player injuries at an early stage of the game. Hubáček et al. (2019) used logistic regression and two types of neural networks (FFNN and convolutional layer neural network [CLNN]) for predicting NBA outcomes with a purpose of beating a bookmakers betting odds. The FFNN is intended for the team-level feature set, while the CLNN is designed for the player-level feature set. The best accuracy of 68.83% was produced by FFNN, followed by the CLNN (68.80%) and logistic regression (68.70%), all of them additionally using betting odds feature set. Even though the prediction results using betting odds feature set gave better predictions, authors concluded that it is not desirable to use betting odds in the prediction since this introduces high correlation with the bookmaker prediction. By ejecting low-confidence predictions before providing them with the betting strategy, maximum accuracy increased significantly (80.26% for FFNN and 84.35% for CLNN). Interesting research, related to predicting outcomes in the NBA league, has also been given by Horvat and Job (2019). The aim of the paper was not to achieve the maximum possible outcome prediction accuracy, but to define the optimal dataset length. A basic feature set consisting of 13 features related to game statistics and the naive classification ML classification algorithms were used. Best prediction results were achieved by the database segmentation method and the use of one to three training seasons and single testing season. It was also concluded that increasing the training dataset by adding known test phase data also contributes to the increase in accuracy.

Figures 4–8 show the relationship between the outcome prediction results of the analyzed sport and other analyzed sports. Most of the analyzed papers use multiple ML algorithms. The best results of each ML algorithm are included in the analysis. The prediction results of the analyzed sport in relation to other sports are shown in red. The aim of the analysis is to show the relation between the proposed models prediction accuracy of a particular sport and other analyzed sports. Figure 2 shows the relationship between the maximum basketball outcome prediction accuracy in relation to other sports. As is illustrated, the majority of research is related to basketball with prediction results ranging from 60% to 85%. The only results that deviate from other results are 80–85%. A further analysis with the boxplot will indicate whether there are any outliers.

4.2 | Soccer

McCabe and Travathan (2008) used the artificial neural network trained with back propagation and Conjugative-Gradient Descent (CGD) to predict four major leagues outcomes (Australian National Rugby League, AFL, EPL, Super Rugby League) and obtained the best accuracy, for three consecutive seasons, of 68.1, 67.2, 58.9, and 75.4%, respectively. The

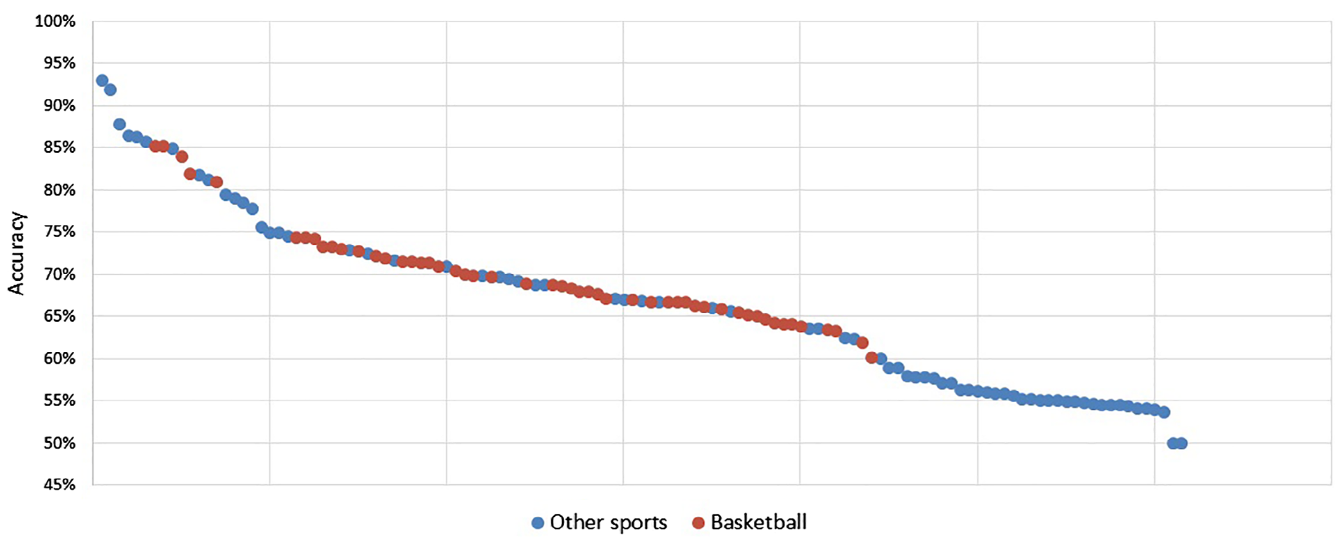

FIGURE 4 Basketball prediction results in relation to other analyzed sports

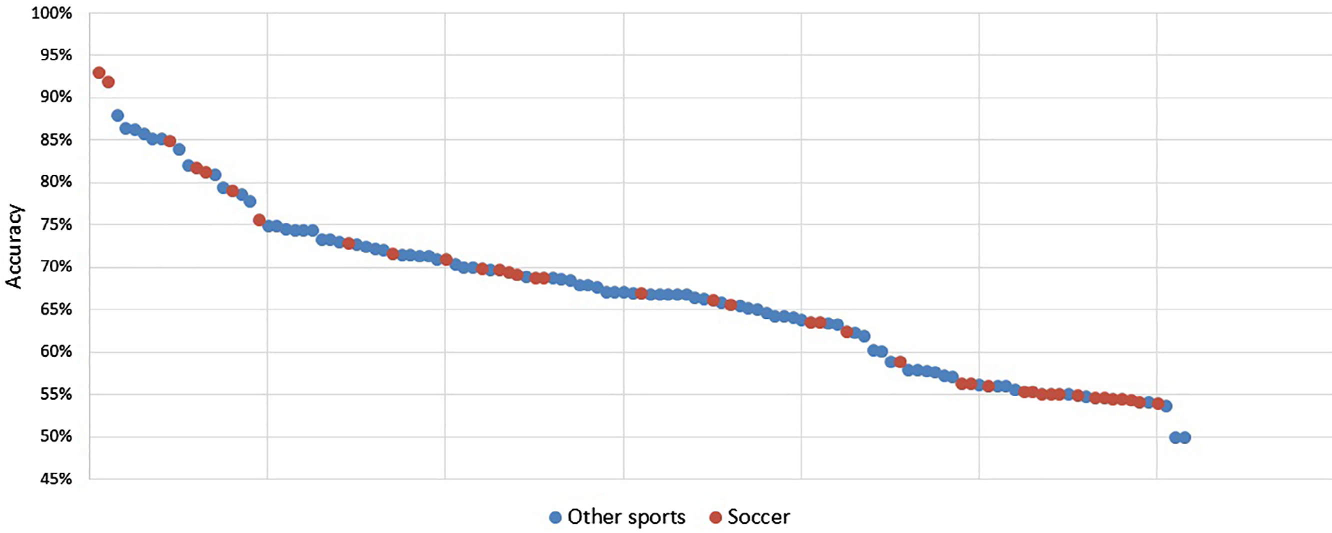

FIGURE 5 Soccer prediction results in relation to other analyzed sports

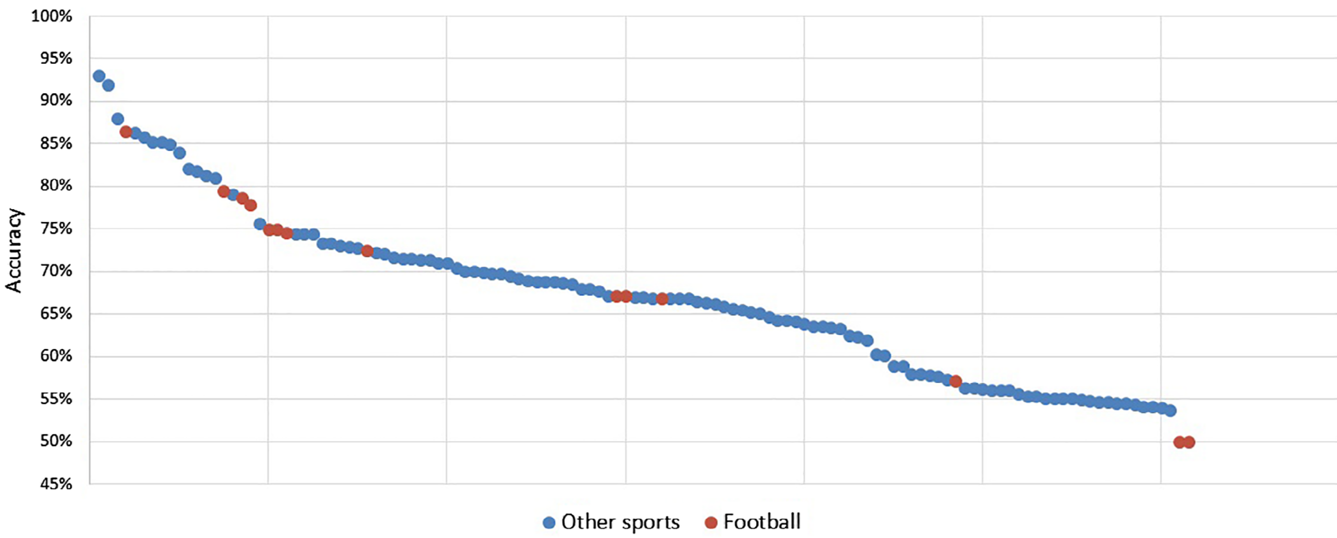

FIGURE 6 Football prediction results in relation to other analyzed sports

training time with back propagation was longer but produced slightly better predictions. The feature sets in the research were the same across all analyzed sports (sport-specific features were not included). Hucaljuk and Rakipović (2011) used several ML models (Naïve Bayes, Bayesian network, LogitBoost, k-NN, random forest, neural network) to predict Champions' League group stage games outcomes and got the maximum accuracies of 56.3, 56.3, 68.8, 62.5, 65.6, and 68.8%, respectively. They used two feature sets (basic and expert) and three different datasets. The expert suggested feature sets produced better accuracy only while using LogitBoost and artificial neural network. Buursma (2011) used linear regression with the accuracy of 55.05%, logistic regression with the accuracy of 54.98%, decision trees with the accuracy of 54.57%, LogitBoost with the accuracy of 54.62%, BayesNet with the accuracy of 54.55%, and Naïve Bayes with the accuracy of 54.43% to predict outcomes in Netherlands football league based on the last 15 years. All features were based on the last 20 team games, which is the number determined as the best for calculating extracted features. Owramipur et al. (2013) proposed a Bayesian network model implemented in NETICA software and used the same data for training and testing purposes, predicting outcomes of team FC Barcelona in Spain soccer league with the accuracy of 92%. This result is not surprising given that the same dataset was used for training and testing and the fact that the team with the most wins, a final league winner, was analyzed. Igiri and Nwachukwu (2014) used the artificial neural network and logistic regression implemented in RapidMiner to predict EPL game outcomes. The artificial neural

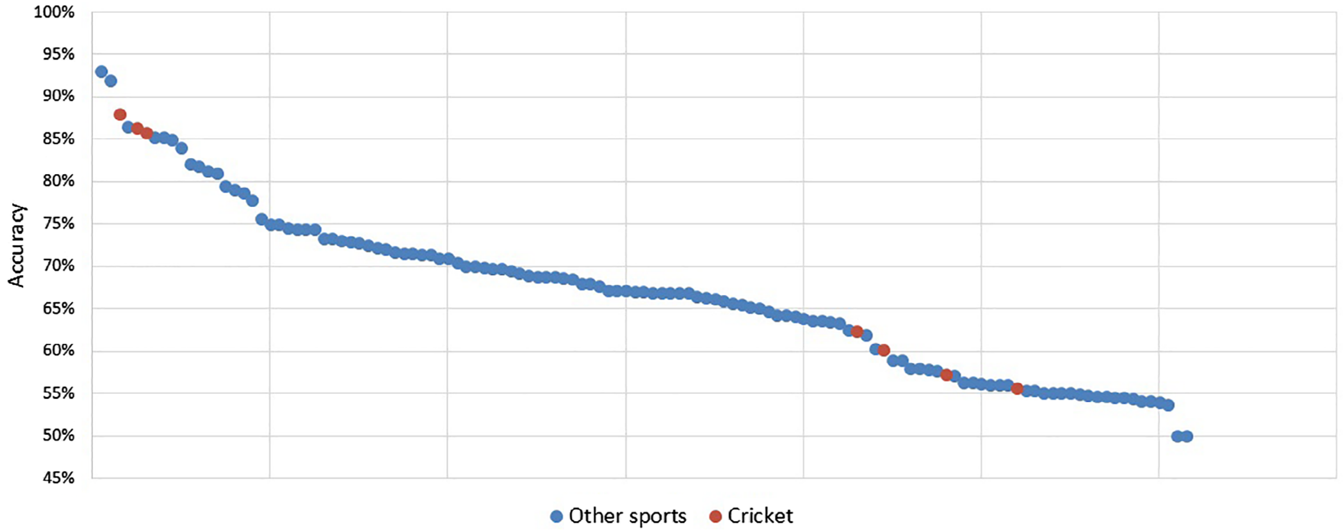

FIGURE 7 Cricket prediction results in relation to other analyzed sports

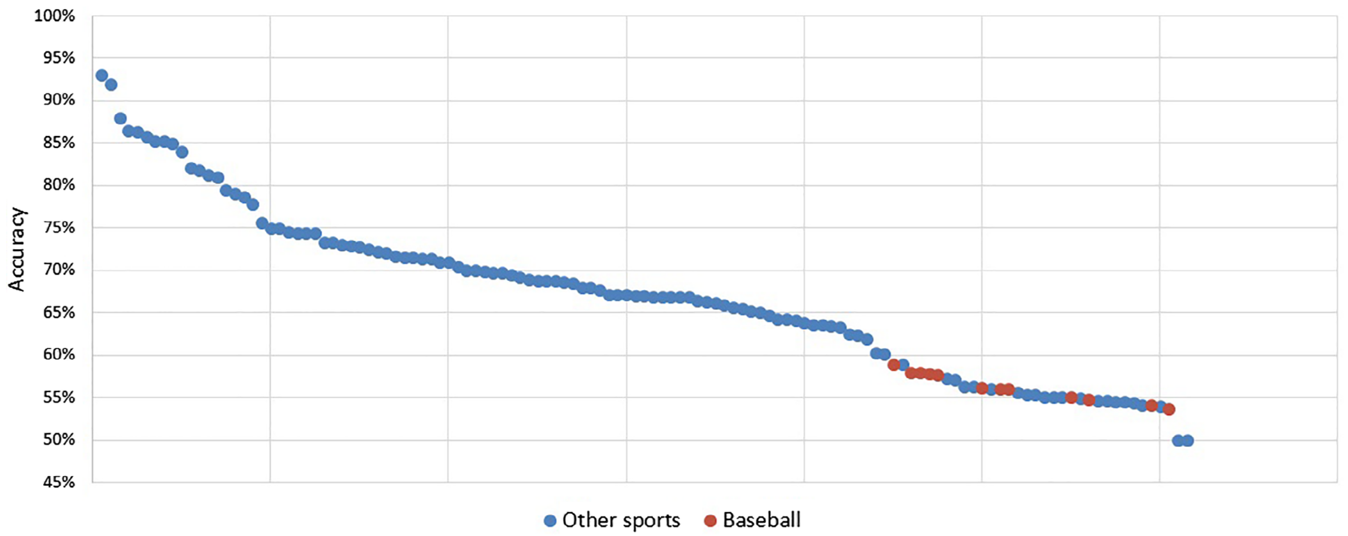

FIGURE 8 Baseball prediction results in relation to other analyzed sports

network can predict a win, loss, or draw compared to logistic regression which can only predict a win or loss. A prediction accuracy of 75.04% was recorded without the appropriate optimization. Optimization and feature weighting increased the accuracy to 85% for the artificial neural network and to 93% for linear regression. The optimization by weighting also showed that the Player Performance Index has a higher weighting compared to other features when implemented with the ANN technique. The highest accuracy of 54.702% for the public data model was achieved using Naïve Bayes or Multilayer Perceptron combined with PCA. Training and testing were done on a self-made dataset from publicly available sources. The model that includes only the betting odds feature set achieved the maximum accuracy of 55.30% using the FURIA classifier. A hybrid model that used both feature sets reached the maximum accuracy of 56.10% when using LogitBoost with ReliefF feature selection method. The authors also concluded that a cross-validation evaluation method is not recommended because of the time-ordered data. Prasetio and Harlili (2016) proposed a model for soccer match prediction with data from EPL and achieved the accuracy of 69.51% using logistic regression. Significant features gathered from other researchers were used as a training dataset for 2010/2011–2015/2016. Features such as “Home Defense” and “Away Defense” were used but it was not described how these features were calculated. Zaveri et al. (2018) used different ML algorithms (logistic regression, random forest, artificial neural network, linear

SVM, and Naïve Bayes) and achieved the maximum accuracies of 71.63, 69.90, 69.20, 66.95, and 63.57%, respectively. In addition to ML algorithms, they used a statistical approach to improve the accuracy. The authors used two sets of features—the Match History Database feature set, which included 12 features related to game performance, and a feature set called Team versus Team, which contained records for all teams against each other during the last five seasons. Use of the match history database and the database containing records for all teams led to a better prediction result. Knoll and Stübinger (2020) proposed a methodology for predicting the outcomes of top five European soccer leagues (EPL, Ligue 1, Bundesliga, Serie A, and Primera Division) and back-tested the prediction results on the betting odds of one of the world's leading online sport betting providers. The feature set contained information about the general game, pass behavior, defense and disciplinary measures, and attack capacities. Average feature values, except defensive behavior, were generally higher for the home team. The best accuracy of 75.62% was achieved using random forest algorithm, followed by gradient boosting algorithm (70.95%), SVM (66.12%), and linear regression (63.60%). The authors also implemented three naive prediction methods. The BET strategy declared a winner team with the lowest betting odd and achieved accuracy of 54.49%. The HOM strategy declared home team as a winner team (46.10%), while RAN strategy randomly picked a winner team and achieved accuracy of 37.82%. The difference between used ML algorithms and naive methods was in the number of classes predicted. In the context of the BET strategy, the lowest betting odd is usually home or away team win, HOM strategy always declares home team as a winner team, while the RAN strategy randomly picks a winner. Average feature values were calculated based on the past three and at least three current season home (away) games. Stübinger et al. (2020) proposed a model for predicting game outcomes in top five European soccer leagues and used different ML algorithms (random forest, boosting algorithm, SVM, and linear regression) based on game and player characteristics with the main goal of beating the betting odds. The best accuracy of 81.77% produced regression model called ALL which integrated of the all aforementioned ML approaches, followed by random forest (81.26%), boosting algorithm (79.12%), linear regression (72.92%), and SVM (69.71%). As in Knoll and Stübinger (2020), the authors implemented three naive approaches which achieved prediction accuracies of 49.91% (BET), 45.44% (HOM), and 36.05% (RAN). Proposed ML framework succeed in beating betting odds by using appropriate betting.

Figure 5 shows a disparity in the results associated with soccer outcome prediction. The outliers are clearly visible in the figure. Excluding outliers, soccer prediction accuracies are generally not high compared to other sports predictions. One of the reasons is certainly the need for predicting three possible outcomes (win/loss/draw). The number of goals scored in a soccer game is much lower than in the other analyzed sports and, consequently, the possibility of a draw runs much higher.

4.3 | Football

Purucker (1996) used several supervised and unsupervised neural network strategies for predicting NFL league outcomes. A total of five features and data from weeks 11 to 16 were used. Supervised back propagation method reached the best result of 78.60% when predicting Week 16 game outcomes. Among unsupervised methods (Hamming, Adaptive Resonance Theory, Self-Organizing Map [SOM]), the SOM method provided the best results. Additionally, the author calculated four features (yards gained, rushing yards gained, turnover margin, and time of possession) in which a team needs to outperform the opposing team to win. Kahn (2003) extended Purucker's (1996) research and used a multilayer perceptron neural network for predicting NFL league outcomes in weeks 14 and 15 on two datasets. The first dataset refers to the mean statistics of the whole season while the second one refers to the mean statistics during the last 3 weeks of the season. In both cases, the season average prediction set was more effective in predicting the outcome. For weeks 14 and 15, the season average prediction set generated 75% correct outcomes while the second dataset generated accuracies of 62.5% for week 14 and 37.5% for week 15. These results were better compared to expert's opinion. Hamadani (2006) used SVM and logistic regression to predict NFL league game outcomes achieving the best accuracy of 67.08% for logistic regression and 66.81% when using SVM. The author used a leave-one-out-classification evaluation method with first 2 weeks of each season used only as a training dataset. Two features sets, win-ratio features, and game statistics feature set were used. The feature selection backward/forward methods were implemented only for logistic regression. Blaikie et al. (2011) used different types of artificial neural networks to get a correlation between game features and final outcome. The authors concluded that more than 66% of derivative information about outcomes comes from scoring. Delen et al. (2012) applied CRISP-DM methodology and used decision trees, neural network, and SVM, both classification and regression types of outcome with the combination of 10-fold cross-validation for predicting college football game outcomes. Using the classification approach, the authors achieved the accuracies of 86.48, 75, and

79.51%, whereas by using the regression approach, the authors achieved the accuracies of 77.87, 72.54, and 74.59%, respectively.

Figure 6 shows the football prediction results compared to other sports. The sample is relatively small, so no special conclusions can be drawn except that there are two outliers and that the mean data show the expected prediction value.

4.4 | Cricket

Kampakis and Thomas (2015) used Naïve Bayes, logistic regression, random forests, and gradient boosted decision trees to predict the outcome of English County 20 over Cricket Matches. The performance of each predictive algorithm was assessed using the previous year's data as the training dataset and the year in question as the test. Each model was tested over six seasons and achieved the accuracies of 62.4% for Naïve Bayes, 60.1% for logistic regression, 55.6% for random forests, and 57.2% for decision trees. In the paper by Mustafa et al. (2017), the authors published an interesting research where they used three different ML algorithms (SVM, Naïve Bayes, logistic regression) that depend on crowd opinion (total number of tweets before the game for each team, fans sentiments, and fans score predictions) on Twitter and implemented in WEKA. The authors collected tweets of 109 games and extracted three features (tweet volume, aggregated fans sentiments, score prediction). The best result of 87.90% was achieved by using SVM method, followed by Naïve Bayes (86.28%) and logistic regression (85.73%). They also indicated that the usage of social networks can be as informative as professional newspaper media.

Figure 7 illustrates the cricket prediction results compared to other analyzed sports with unbalanced results being visible.

Only two papers in different leagues were used in the analysis, so no conclusions can be drawn. In general, cricket does not fall into the category of the most popular sports, so it is not surprising that there are not many available papers related to predicting cricket outcomes. A more specific analysis requires more papers related to at least the same league.

4.5 | Baseball

In the paper by Soto Valero (2016), the author proposed a model for predicting outcomes in MLB regular-season games. The author used a 10-fold cross-validation procedure and both classification and regression type of predicting. *K*-NN method achieved the accuracies of 55.98 and 55.97% for classification and regression, multilayer perceptron achieved the accuracies of 57.98 and 56.17%, and decision trees achieved the accuracies of 57.86 and 57.90%, while SVM achieved the accuracies of 58.92 and 57.66% for classification and regression. Elfrink (2018) used linear regression, random forest, XGBoost, and boosted logistic regression for predicting MLB game outcomes based on both classification and regression methods. Boosted logistic regression was tested only as a classification problem. The author used two different feature sets. The first feature contained basic baseball statistics while the second one contained basic baseball statistics and recent events information. The best classification accuracy of 55.52% was produced by XGBoost method, followed by Boosted logistic regression (54.78%), random forest (54.08%), and linear regression (53.75%) all of which used basic baseball statistics. The best results using a regression problem were also produced by XGBoost by using both feature sets (55.02%), followed by random forest with the basic baseball statistics feature set (53.34%) and linear model with both feature sets included (51.81%). The research confirmed the statement that the classification-type models predict the game outcomes better than the regression-based classification model.

Figure 8 illustrates the baseball outcome prediction result compared to other analyzed sports. The results are expected and there are no outliers. A later analysis will explain the expected baseball results.

4.6 | Summary

A summary of all papers, best accuracy of the proposed algorithms, number of seasons and features, information as to whether the feature selection, or feature extraction method were used and evaluation models are summarized in Table 2. It is evident there that the most authors used two evaluation methods—dataset segmentation and *k*-fold cross-validation (*k*-fold CV). In the dataset segmentation method, the input data is usually portioned into three different

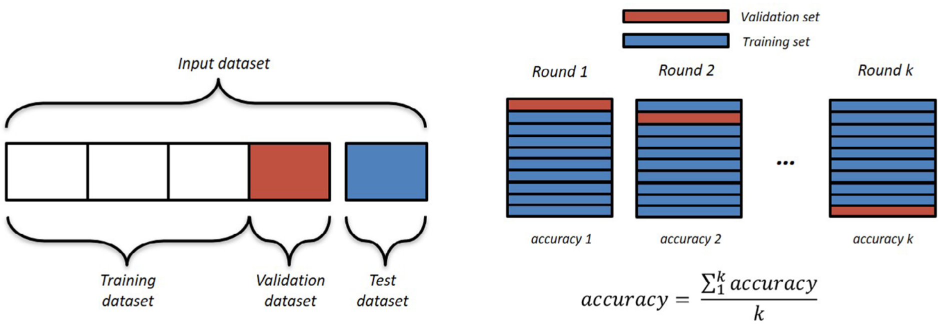

FIGURE 9 Dataset segmentation (left) and k-fold cross-validation (right) evaluation methods

datasets—the training dataset, validation dataset, and testing dataset which should be chronologically ordered (Figure 9). Validation datasets are not always used and are usually employed to tune the final artificial intelligence model parameters. Using a chronologically defined subset of input data is recommended because sport events are not entirely independent events. Historical data can provide useful information and thus help in predicting future events. In k-fold cross-validation (Figure 9), the original dataset is partitioned into k folds. Of k folds, a single fold is retained as the validation dataset for testing the model and the remaining folds are used as training datasets. The cross-validation procedure is repeated k times and finally, the average performance over all folds is calculated. This method has the advantage that all observations are used for training and testing (Gorges et al., 2018) but it is not recommended for predicting sport outcomes.

As noted above, sporting events are not entirely independent, the outcome prediction based on future events or the long history due to dynamics in a team roster is not advisable because it will usually lead to lower prediction results. Among all analyzed sports, basketball is the only sport played indoors. Thus, weather-related conditions do not have an impact on the final game outcome. On the other hand, a basketball team scores a lot more points than teams in other analyzed sports. Predicting sport outcomes is a complex problem due to the range of uncertainties that can, even during the game, influence the final game outcome such as a player's injury, weather conditions, in-game tactical changes, and so on.

Most research involves predicting basketball and soccer outcomes, respectively to sports with 2 (basketball) and 3 (soccer) possible outcomes. It is important to note that a draw in soccer is highly probable outcome but also that predicting outcomes in sports with highly probable three outcomes offer a mathematical loss. The average maximum accuracy per analyzed paper and sport is 73.92% for basketball and 72.43% for soccer. Excluding two outliers (maximum accuracies of 92.00 and 93.00%) in predicting soccer outcomes, the average maximum accuracy dropped to 67.42%. Additionally, for NBA, the most popular basketball league achieved maximum average accuracy per analyzed paper is 72.83%. The biggest problem, even for the NBA where the most research is available, is the use of different datasets, which makes the comparison of results very difficult or almost impossible. Cricket and baseball, even football, contain insufficient analysis of different leagues, making it even more difficult to draw regularities, and conclusions. The worst average prediction results are for baseball, a sport with only two possible outcomes. The result is supported by Table 3, which shows that the smallest difference in the percentage of wins between the best and the worst teams during the regular season is in baseball, particularly MLB league. The general conclusion is that the complexity of outcome prediction depends mostly on the league competitiveness and increases with the number of possible outcomes. Likewise, basketball proved to be more consistent than soccer.

The most suitable result representation for this problem is the box plot (Tukey, 1977, 1990). A box plot is a method for graphically depicting groups of numerical data through their quartiles. A box plot can easily organize large amounts of data and visualize outlier values. The box plot in Figure 10 shows the result dispersion categorized by sport.

As is evident from Figure 10, the baseball result stands out giving significantly lower prediction results but also the baseball box plot is comparatively short. This indicates that research have a high level of uniformity. In comparison, the football and soccer box plot are comparatively high indicating that research have a quite low level of uniformity. The basketball box plot contains outliers, but it is shorter than the football and cricket box plot, which suggests that most of basketball research have a high level of uniformity. Table 3 leads to a conclusion that sport whose outcome is the easiest to predict is football followed by basketball, soccer, and baseball. Baseball, especially MLB, is known to be competitive, the best team of the regular season wins only 63% of all games and the worst team 37%. Analyzing only soccer, it is clear that two results do not fit with others. These two points represent a single season and use the data

TABLE 3 Percentage of wins per league and seasons during a regular season

	NFL (football)		NBA (basketball)		EPL (soccer)		MLB (baseball)	
Season	Winner (%)	Last (%)	Winner (%)	Last (%)	Winner (%)	Last (%)	Winner (%)	Last (%)
2017/2018	81.25	0.00	79.27	25.61	84.21	15.79	66.67	29.01
2016/2017	87.50	6.67	81.70	24.39	78.95	15.79	64.19	39.50
2015/2016	93.75	23.08	89.00	12.12	60.53	7.89	63.58	36.42
2014/2015	75.00	12.50	81.70	19.51	68.42	21.51	61.73	38.89
2013/2014	81.25	12.50	75.61	18.29	71.05	18.42	60.49	39.50
Average	83.75	10.95	81.46	19.98	72.63	15.88	63.33	36.66
Difference	72.80		61.48		56.75		26.67	

segmentation method for evaluation. When using a single season, if most of the season is used for training and a small part for testing, it is possible to get the unrealistically high predictive outcome accuracy. Another problem with the soccer and cricket results is the position of the median, which is, in this case, very low (median line is usually close to the average), which also indicates a low level of uniformity between the papers. Other sports, except cricket, also have a high dispersion, which is expected since not all the papers are related to the same competition or dataset.

An interesting, but due to a small number of samples somewhat inappropriate for display as a boxplot, is the analysis of the prediction results by ML algorithm groups. An appropriate tool to present prediction results by ML algorithm group and by sport is using a matrix bubble chart, that is, a categorical bubble plot in which data is depicted by circles and its color and size. Figure 11 shows a prediction results according to an ML algorithm group and analyzed sport. ML algorithm group and sport form the data pair. Additional dimensions of dataset are represented, thus median of results for each data pair is depicted by the size of the circle and other dimensions, the number of occurrences of specific data pair within the overall dataset is depicted by the intensity of the color. In Figure 11, ML algorithms are grouped according to their similarities, as in Figure 2.

The number in the middle of the circle and the size of the inner circle represent the median of the achieved results by ML algorithm group and by sport, that is, data pair, while the size of an outer circle represents the maximal accuracy achieved by ML algorithms group for a particular sport. The color intensity represents the relevance of the data pair, a dark blue circle means that there are more occurrences of specific data pair within the dataset than if a circle is filled with a lighter blue. Data pairs with no clearly defined difference between the median and maximal accuracy are pairs where median is the same as maximal accuracy (single occurrence of pair) or the median and maximal accuracy are near in value (two occurrences). The similarity in size between an inner and outer circle represents uniformity of the results, for example, more overlapping circles depict high uniformity of the results and it is visible mostly within pairs with a small number of results in the dataset, such as cricket and baseball. Two exceptions, cases of low uniformity in cricket, is logical given that two researches related to different datasets and leagues are used. Generally, the lowest level

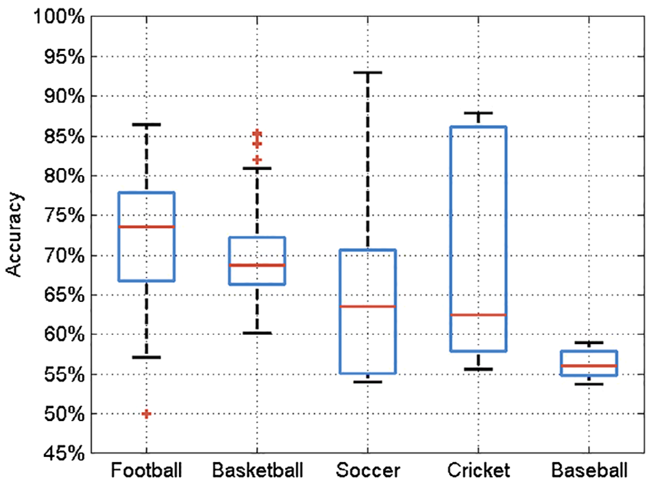

FIGURE 10 Boxplot of the achieved maximum accuracies by sport

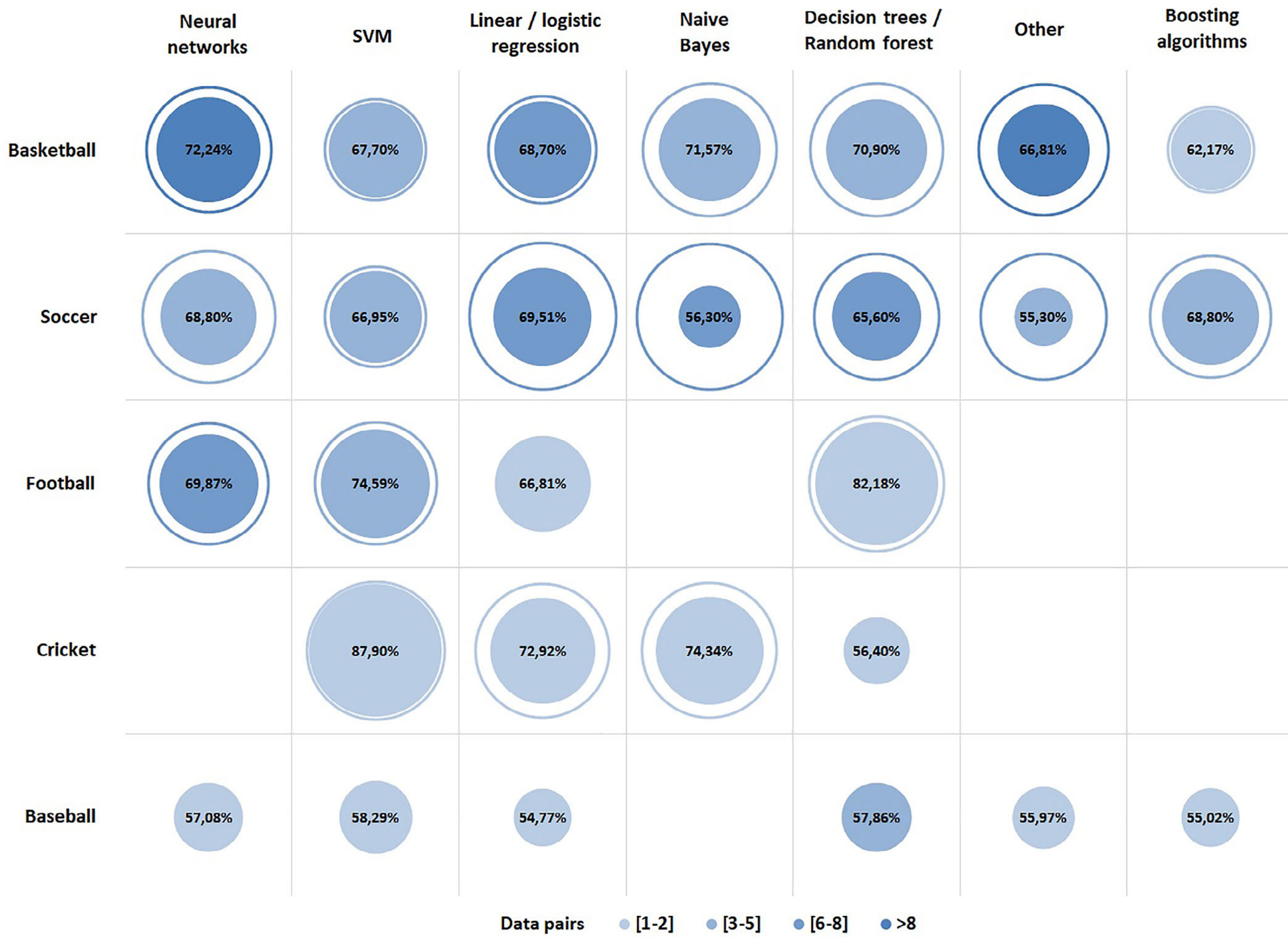

FIGURE 11 Median of accuracy, maximal accuracy, and number of occurrences of the proposed ML algorithm group and sport data pair within overall dataset. Data pair relevance is shown by intensity of the color, accuracies are shown by circle size and median is shown by its value as well

of uniformity is achieved in soccer what is depicted with a difference in the size of the inner and outer circle. Such a result was already suggested by Figure 10 with a comparatively high box plot and relatively low position of the median. The reason for that has already been mentioned and concerns primarily the use of different datasets and leagues of different competitiveness. Likewise, soccer is the only sport with three highly probable final outcomes. The most relevant is basketball with satisfactory uniformity considering that different datasets, features, and most importantly, leagues of different competitiveness were used. Generally, the highest uniformity is achieved by the ML algorithm groups and sports with a smaller number of occurrences within the dataset.

In the following section, the properties related to the input dataset are analyzed (number of seasons, features, references, and prediction algorithms accuracy progress).

Figures 12–14 show only the best results of the individual research. The accuracy of outcome prediction algorithms decreases with increasing the number of seasons and features. A greater dataset (a greater number of seasons or features) does not necessarily mean the decline in the ML algorithm accuracy but only suggests that almost equal or even better research results can be achieved by using smaller datasets, more precisely a dataset that best describes the current state of the analyzed team. Figure 13 has a few outliers. By removing the outliers (Figure 14), a slight increase in accuracy is evident in cases where fewer than 10 features is used. As stated before, a larger subset of features does not necessarily mean lower predictive results, rather it suggests that the analyzed process can be described with a smaller feature subset without compromising the model accuracy

Figure 15 shows the relationship between the number of features and seasons. If the feature selection method was used, the number of the reduced feature set is used. The graph clearly shows that the number of features and seasons

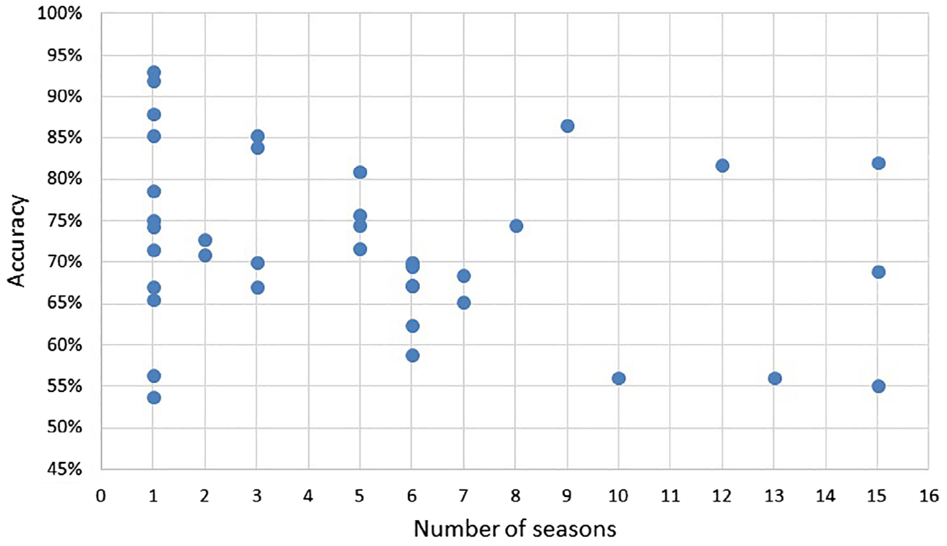

FIGURE 12 Dependence of accuracy to the number of seasons

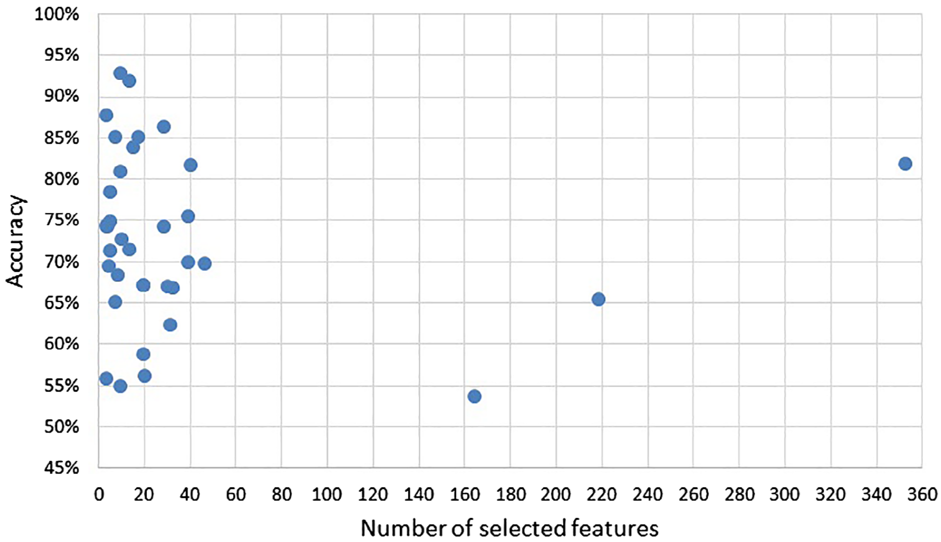

FIGURE 13 Dependence of accuracy to the number of selected features

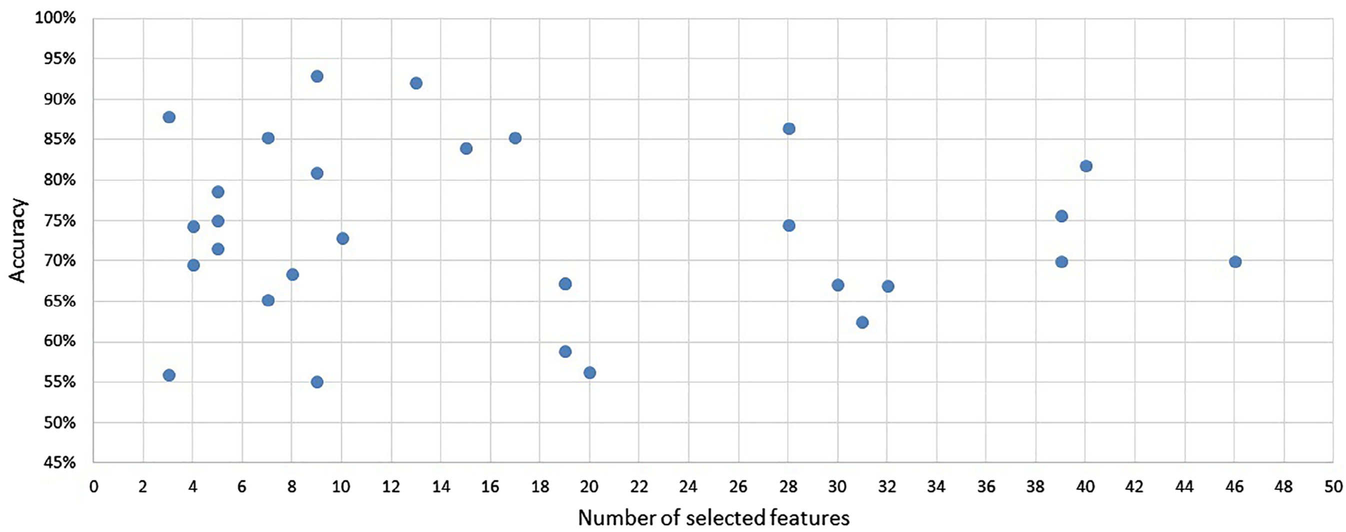

FIGURE 14 Dependence of accuracy to the number of selected features (outliers excluded)

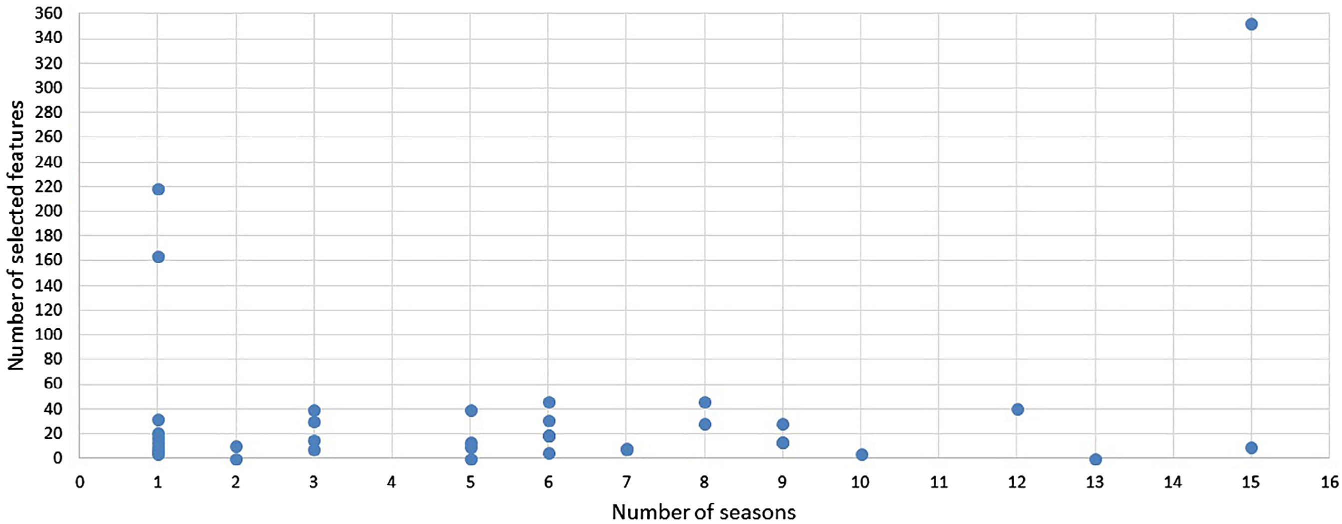

FIGURE 15 Dependence of the number of features and seasons

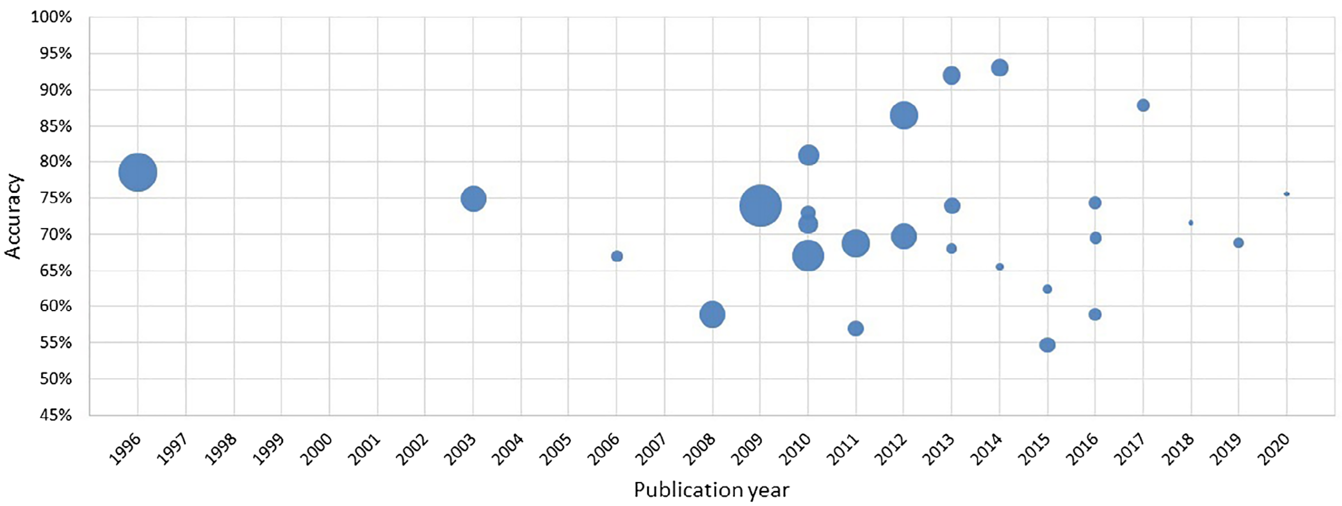

FIGURE 16 Dependence of accuracy vs. time and the number of citations

are usually in inversely proportional relation. There are several papers with fewer number of features and seasons and only one paper with a larger number of features and seasons. The graph also shows that researchers generally prefer the use of fewer training and testing seasons to obtain better prediction results.

Figure 16 shows the dependence of maximum accuracy vs. year of publication, and the citation number according to Google Scholar where a bubble size represents the number of citations. The most cited papers are from 1996 and 2009 and most of the available papers are from 2010 to 2013 period. The graph clearly indicates that the number of citations depends mostly on the publication year. Newer papers, regardless of their quality, have fewer citations than the older ones, which should be taken into account.

Figure 17 shows the progress of ML algorithms throughout the years, unrelated to the type of sport. The trend line suggests the progress of ML algorithms. The progress is understandable given that researchers use their peers' results consequently coming up with new discoveries, so we believe that new methods will be even more accurate. A trend line is calculated by the least square method. Figures 18–22 illustrate the progress of the prediction results by year and the analyzed sports. The best results of all ML algorithms used in the analyzed papers are presented. The progress is evident in the quantity of available papers. Generally speaking, progress in the accuracy of the proposed models is also evident

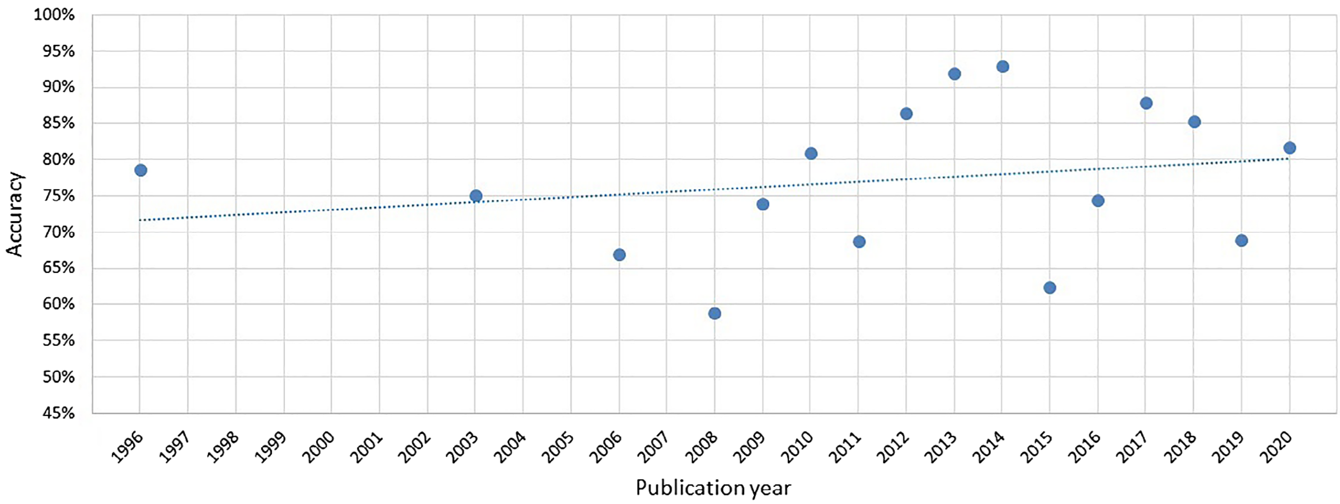

FIGURE 17 Progress of the ML algorithms regardless of sport

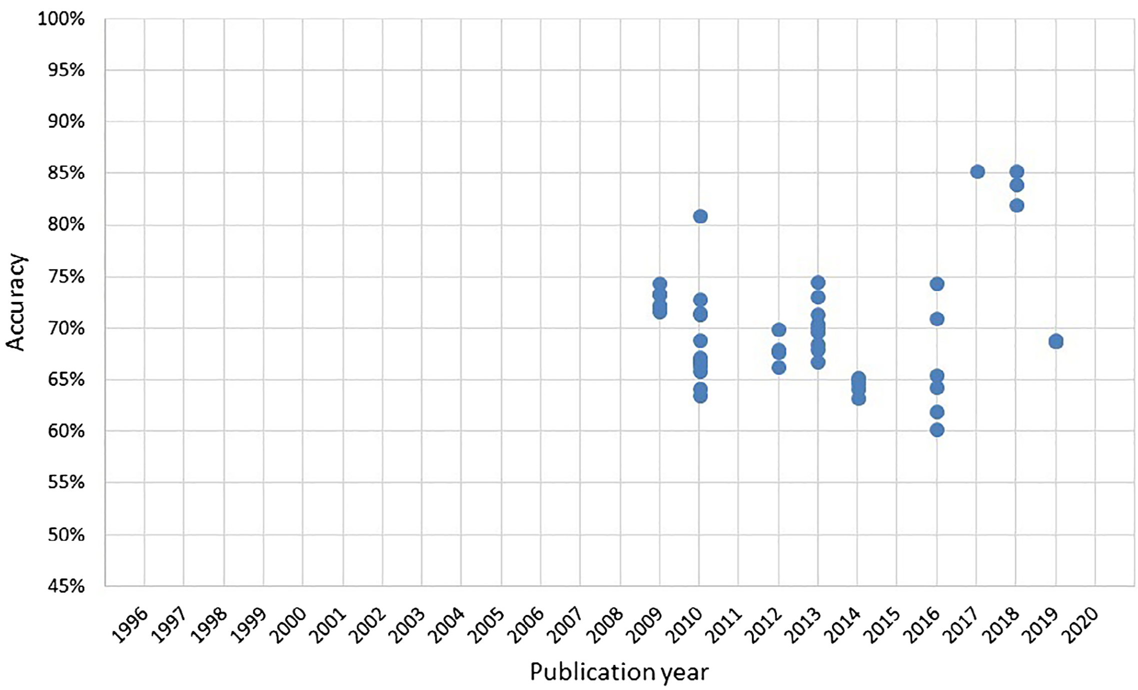

FIGURE 18 Progress of the ML models related to basketball

and the biggest problem with the result comparison is the use of different datasets and leagues of different competitiveness. Samples for sports, such as cricket or baseball, are too small to draw any conclusions. Cricket results analyze the prediction results for different leagues so drawing conclusions is even more difficult.

The analyzed papers use a different number of references and it is necessary to analyze a correlation between the number of references and the highest achieved accuracy. Figure 23 shows the correlation between the number of references and the highest achieved accuracy.

Figure 23 shows that there is no correlation between the ML algorithm accuracy and the number of references. There are also no outliers that could lead to some conclusions. Most of the papers have fewer than 20 references, which is logical given that most of these papers were published in the period from 1996 to 2013.

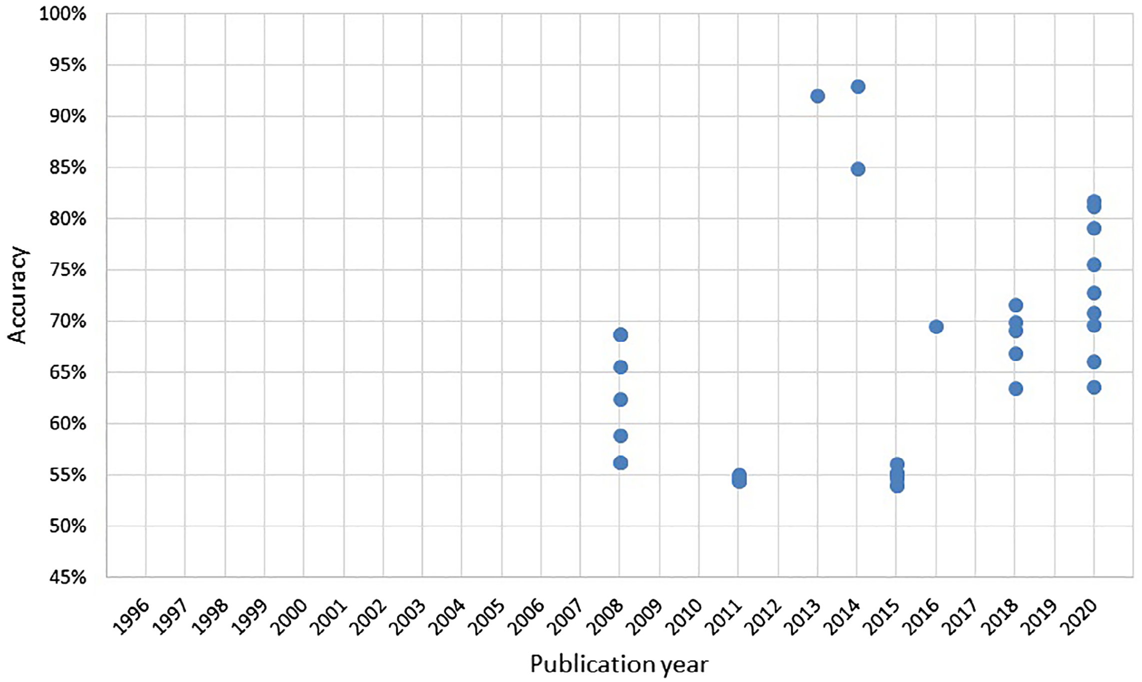

FIGURE 19 Progress of the ML models related to soccer

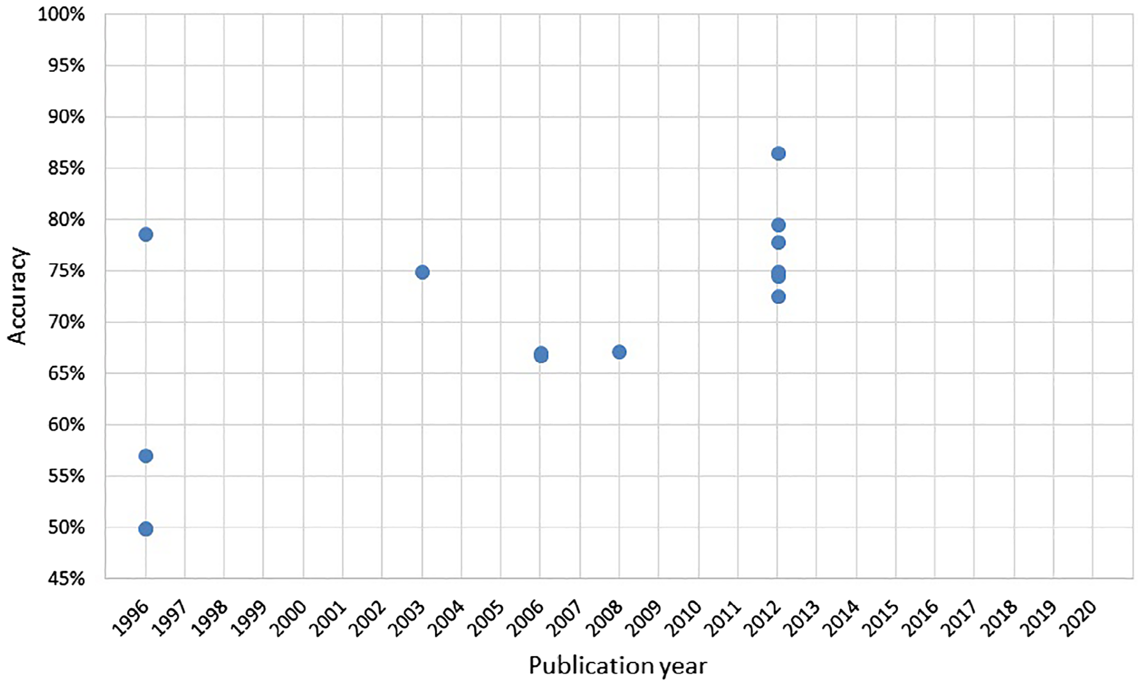

FIGURE 20 Progress of the ML models related to football

5 | CONCLUSIONS

The popularity of ML has been growing every day, not only in sport but also in other aspects of human life. Process outcome predictions can even be defined as a human need. Sports predictions are becoming more popular and thus represent a challenge not only to people involved in sports but also to researchers who use sporting data when developing

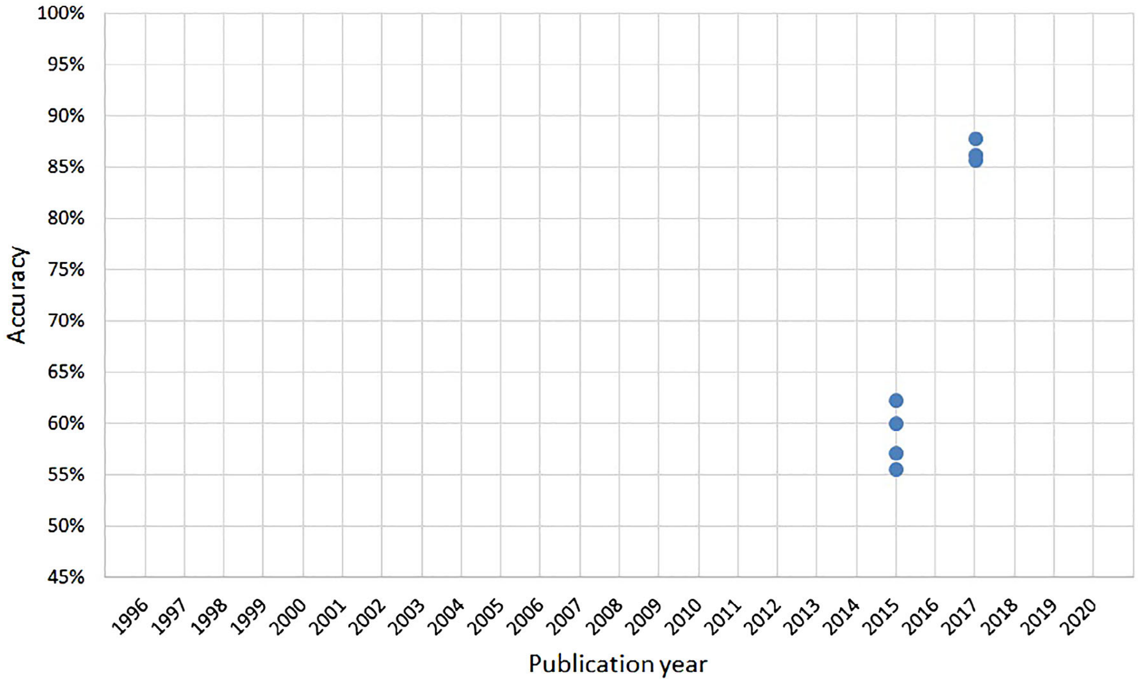

FIGURE 21 Progress of the ML models related to cricket

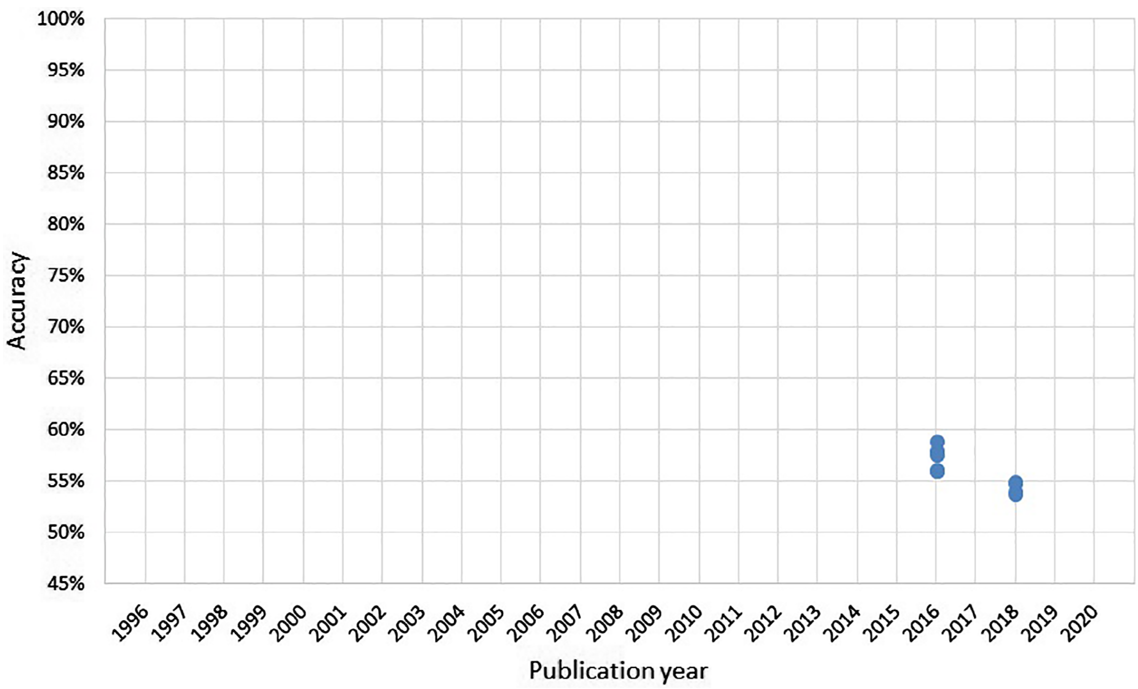

FIGURE 22 Progress of the ML models related to baseball

new algorithms and methods. Sport outcome predictions are most commonly used by supervised ML, more precisely the classification methods. There are also examples in which the outcome of a sport event is predicted by regression methods but in that case, usually a spread is calculated and based on the spread, a winner team is determined. Contrarily, there are quite a few cases in which the outcome of a sport event is predicted by using unsupervised or reinforcement ML algorithms. Unsupervised learning methods are usually used in cases where the final outcome of the

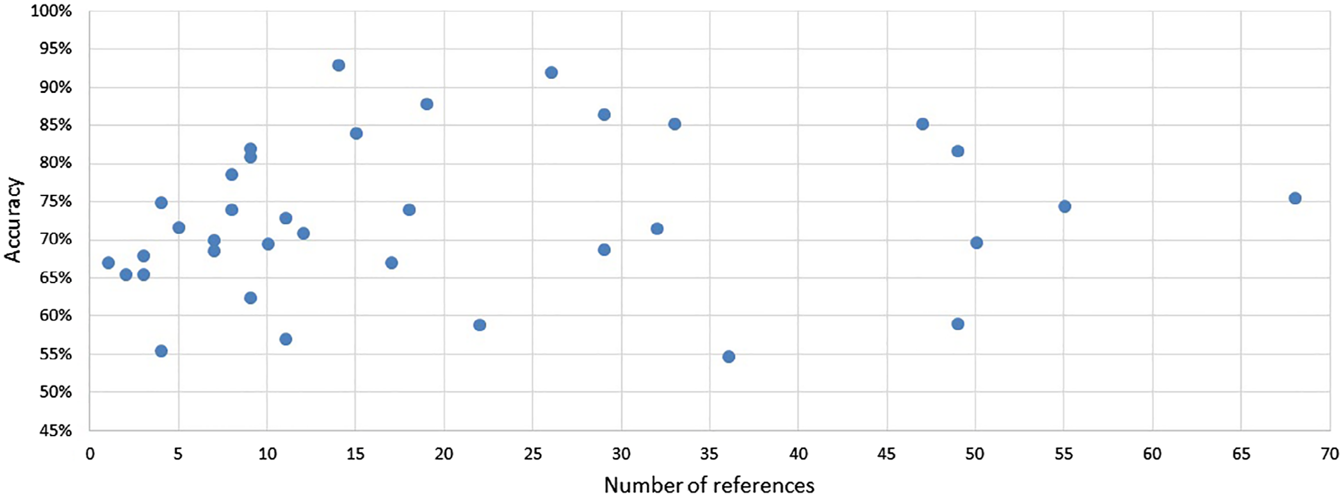

FIGURE 23 Dependence of accuracy to the number of references

process is unknown. In that case, the purpose of the unsupervised learning method is finding regularities in data or to predict the process outcome based only on input data, which represents a more complex and demanding task. The least common method of predicting outcomes in sports is reinforcement learning. In this case, neither input nor output data are known, only the process "behavior" rules are known. The proposed method works by trying each action to figure out which actions yield the best reward and, accordingly, predict process outcome. In this article, available literature has been examined and major challenges have been detected. The main problem lies in the fact that there are practically no papers using the same datasets, which is an obstacle when comparing different methods.

Table 2 shows the evaluation methods of the proposed models. The authors used two evaluation methods called dataset segmentation and *k*-fold CV. Better results were provided by dataset segmentation since the training dataset occurred prior to testing dataset. Sports events are not entirely independent events and certain trends or regularities may occur as a result of previous events. In that case, data contamination may occur as a consequence of future events when the predictive model is informed by "future knowledge" (Yuan et al., 2015). In the dataset segmentation method, most authors do not use a validation set for parameter tuning. The input dataset is usually portioned into two—the training dataset and the testing dataset which are commonly defined by the season duration. Several research has yielded prediction results based on a different number of seasons. The paper by Tran (2016) has shown that including too many seasons reduces the quality of results, which is not surprising given that in just a few years, a lot of things related to team composition, strengths, and weaknesses can change. The best prediction results were achieved by researchers who used data from a single season and data segmentation evaluation method. When using data from a single season, most of the data is used for training and a small portion for testing. It can be said that most of the games are played so it is possible to get highly predictive results. In addition to the mentioned problem, some researchers used the same dataset for training and testing yielding unrealistically high results.

Surprises in sports are not coincidences. Common surprises occur in games when a much smaller team draws a surprising result. Recently, the biggest surprise was made by Leicester City, winning the EPL in season 2015/2016. Analyzing Table 2, most sports predictions are based on the data segmentation evaluation method. Using the data segmentation evaluation method odds for Leicester City winning Premier League were 1–5,000 (Ruiz et al., 2017). Analyzing Leicester City's performance in season 2015–2016 compared to other teams, winning the championship was absolutely deserved. Surprises in sport are happening daily and are usually difficult to predict.

Predicting outcomes is certainly the most popular when it comes to sport predictions. Recently, researchers have been using ML algorithms for many other purposes that are not part of this review paper. For example, Despande and Jensen (2016) and Gramacy et al. (2013) proposed a model for assessing the contribution of individual players to the team winning probabilities or scoring. Wetzels et al. (2016) used the ML approach to analyze the streakiness rates in basketball. Pradier et al. (2016) also used the Bayesian Nonparametric ML model for assessing marathon runner's performance. These are just a few examples of the use of ML in sports without directly predicting outcomes.

The review opened up suggestions for future research that can certainly help in achieving better prediction results. Some of the suggestions are to improve the training methods, use multiple ML algorithms in finding the optimal one, improve feature selection methods, optimize ML parameters, use optimal and relevant dataset, find patterns among data, and so on. Also, research has shown that using alternative, newly proposed ML algorithms, can achieve good, in some cases, even better prediction results. Most authors use feature selection based on expert's experience or filter feature selection methods. It would be very important to explore the impact of wrapper and embedded feature selection methods, as well as the impact of hybrid feature selection methods. The prediction of sport outcomes is a challenge. Various authors have suggested different methods for predicting outcomes and achieved good results. The problem is that most of the proposed methods can only be used for one process, that is, one sport. One of the challenges in sport outcomes prediction should be achieving a unified method that will predict outcomes of more than single sport. The most important conclusion is to use unified dataset across all research to enable comparison between used methods and algorithms. The research should aim on determining the best number of seasons, used features, and interconnection of multiple leagues, or better sports.

ACKNOWLEDGMENT

This work serves as an initial stage in the development of a doctoral dissertation on the outcome prediction in sport and was supported by the University North and Faculty of Electrical Engineering, Computer Science, and Information Technology Osijek, Josip Juraj Strossmayer University of Osijek.

CONFLICT OF INTEREST

The authors have declared no conflicts of interest for this article.

AUTHOR CONTRIBUTIONS

Josip Job: Conceptualization; methodology; resources; supervision; validation; writing-review and editing.

ORCID

Tomislav Horvat https://orcid.org/0000-0002-8358-3218
Josip Job https://orcid.org/0000-0002-6998-5907

RELATED WIRES ARTICLE

Recent trends in machine learning for human activity recognition: A survey

FURTHER READING

Guyon, I., & Elisseef, A. (2006). *An introduction to feature extraction*. Berlin/Heidelberg: Springer.

Stanczyk, U., & Jain, L. C. (2015). *Feature selection for data and pattern recognition*. Cham: Springer.

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, *1*(1), 67–82.

REFERENCES

Alpydin, E. (2010). *Introduction to machine learning* (2nd ed.). Cambridge, MA, London, England: MIT Press.

Avalon, G., Balci, B., Guzman, J. (2016). Various machine learning approaches to predicting NBA score margins. Final Project.

Barto, A. G., & Sutton, R. (1997). *Introduction to reinforcement learning*. Cambridge, MA, London, England: MIT Press.

Blaikie, A. D., David, J. A., Abud, G. J., & Pasteur, R. D. (2011). NFL & NCAA Football Prediction using Artificial Neural network. Retrieved from https://www.semanticscholar.org/paper/NFL-%26-NCAA-Football-Prediction-using-Artificial-Blaikie-Abud/5207ead80e566abd29bf2c171143fa6473c28b6a.

Bunker, R., & Susnjak, T. (2019). The application of machine learning techniques for predicting results in team sport: A review. *arXiv preprint* arXiv:1912.11762.

Bunker, R. P., & Thabtah, F. (2019). A machine learning framework for sport result prediction. *Applied Computing and Informatics*, *15*(1), 27–33.

Buursma, D. (2011). *Predicting sports events from past results: Towards effective betting on football matches*. In Proceedings of the 14th Twente Student Conference on IT. Twente, Holland.

Cao, C. (2012). *Sports data mining technology used in basketball outcome prediction [Master dissertation]*. Ireland: Dublin Institute of Technology.

Cheng, G., Zhang, Z., Kyebambe, M. N., & Kimbugwe, N. (2016). Predicting the outcome of NBA playoffs based on the maximum entropy principle. *Entropy*, *18*(12), 450.

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.

Deshpande, S. K., & Jensen, S. T. (2016). Estimating an NBA player's impact on his team's chances of winning. *Journal of Quantitative Analysis in Sports*, *12*(2), 51–72.

Elfrink, T. (2018). *Predicting the outcomes of MLB games with a machine learning approach*. Business Analytics Research Paper. Retrieved from https://beta.vu.nl/nl/Images/werkstuk-elfrink_tcm235-888205.pdf.

Ganguly, S., Frank, N. (2018). *The Problem with Win Probability*. MIT Sloan Sports Analytics Conference. Retrieved from https://statsweb-wpengine.netdna-ssl.com/wp-content/uploads/2018/09/2011.pdf.

Gorges, C., Öztürk, K., & Liebich, R. (2018). Impact detection using a machine learning approach and experimental road roughness classification. *Mechanical Systems and Signal Processing*, *117*, 738–756.

Gramacy, R. B., Jensen, S. T., & Taddy, M. (2013). Estimating player contribution in hockey with regularized logistic regression. *Journal of Quantitative Analysis in Sports*, *9*(1), 97–111.

Grossberg, S. (1988). Nonlinear neural network: Principles, mechanisms and architectures. *Neural Networks*, *1*(1), 17–61.

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

Hamadani, B. (2006). *Predicting the outcome of NFL games using machine learning, Project Report*. Stanford, CA: *Stanford University*.

Han, J., & Kamber, M. (2006). *Data mining concepts and techniques* (2nd ed.). Burlington, MA: Morgan Kaufmann Publishers.

Horvat, T., & Job, J. (2019). Importance of the training dataset length in basketball game outcome prediction by using naïve classification machine learning methods. *Elektrotehniški Vestnik*, *86*, 197–202.

Horvat, T., Job, J., & Medved, V. (2018). *Prediction of Euroleague games based on supervised classification algorithm k-nearest neighbours*. In Proceedings of the 6th International Congress on Sport Sciences Research and Technology Support: K-BioS, Seville, Spain. pp. 203–207.

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

Hucaljuk, J., & Rakipović, A. (2011). *Predicting football scores using machine learning techniques*. In: Proceedings of the 34th International Convention MIPRO, Opatija, Croatia. pp. 23–27.

Igiri, C. P., & Nwachukwu, E. O. (2014). An improved prediction system for football a match result. *IOSR Journal of Engineering*, *4*(12), 12–20.

Ivanković, Z., Racković, M., Markoviski, B. (2010). Analysis of basketball games using neural networks. 11th International Symposium on Computational Intelligence and Informatics (CINTI).

Jović, A., Brkić, K., Bogunović, N. (2015). A review of feature selection methods with applications. *38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. Opatija, Croatia.

Kahn, J. (2003). *Neural Network Prediction of NFL Football Games* (pp. 1194–1197). Washington, DC: IEEE Computer Society.

Kampakis, S., & Thomas, W. (2015). Using Machine Learning to Predict the Outcome of English County twenty over Cricket Matches. Retrieved from https://arxiv.org/ftp/arxiv/papers/1511/1511.05837.pdf.

Knoll, J., & Stübinger, J. (2020). Machine-learning-based statistical arbitrage football betting. *KI-Künstliche Intelligenz*, *34*(1), 69–80.

Koseler, K., & Stephan, M. (2018). Machine learning applications in baseball: A systematic literature review. *Intelligence*, *31*(9-10), 745–763.

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Informatica*, *31*, 249–268 Retrieved from https://datajobs.com/data-science-repo/Supervised-Learning-[SB-Kotsiantis].pdf

Kravanja, A. (2013). *Napovedanje zmagovalcev košarkaških tekem. Undergraduate thesis*. Ljubljana, Slovenia: Faculty of Computer and Information Science, University of Ljubljana.

Lam, M. W. Y. (2018). One-match-ahead forecasting in two-team sports with stacked Bayesian regressions. *Journal of Artificial Intelligence and Soft Computing Research*, *8*(3), 159–171.

Lin, J., Short, L., & Sundaresan, V. (2014). Predicting National Basketball Association Winners. Final Project.

Loeffelholz, B., Bednar, E., & Bauer, K. W. (2009). Predicting NBA games using neural networks. *Journal of Quantitative Analysis in Sports*, *5*(1), 1–17.

McCabe, A., & Travathan, J. (2008). *Artificial Intelligence in Sports Prediction*. Fifth International Conference on Information Technology: New Generations.

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 and informatics, Subotica, Serbia. pp. 309–312.

Mustafa, R. U., Nawaz, S. M., Ikram Ullah Lali, M., Zia, T., & Mehmood, W. (2017). Predicting the cricket match outcome using crowd opinions on social networks: A comparative study of machine learning method. *Malaysian Journal of Computer Science*, *30*(1), 63–76.

Owramipur, F., Eskandarian, P., & Mozneb Sadat, F. (2013). Football result prediction with Bayesian network in Spanish league-Barcelona team. *International Journal of Computer Theory and Engineering*, *5*(5), 812–815 Retrieved from http://www.ijcte.org/papers/802-N30016.pdf.

Ping-Feng, P., Lan-Hung, C., & Kuo-Ping, L. (2017). Analyzing basketball games by a support vector machines with decision tree model. *Neural Computing & Applications*, *28*(12), 4159–4167.

Pradier, M. F., Ruiz, F. J., & Perez-Cruz, F. (2016). Prior design for dependent Dirichlet processes: An application to marathon modelling. *PLoS One*, *11*(1), e0147402.

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

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

Ruiz, H., Power, P., Wei, X., Lucey, P. (2017). *"The Leicester City Fairytale?": Utilizing New Soccer Analytics Tools to Compare Performance in the 15/16 & 16/17 EPL Seasons*. Proceedings of KDD Conference. El Halifax, Nova Scotia Canada.

Samuel, A. (1959). Some studies on machine learning using the game of checkers. *IBM Journal of Research and Development*, *3*(3), 210–229.

Soto Valero, C. (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.

Stübinger, J., Mangold, B., & Knoll, J. (2020). Machine learning in football betting: Prediction of match results based on player characteristics. *Applied Sciences*, *10*(1), 46.

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.

Torres, R. A. (2013). *Prediction of NBA games based on machine learning methods, computer-aided engineering*. Madison, WI: University of Wisconsin.

Tran, T. (2016). *Predicting NBA games with matrix factorization (Master dissertation)*. Cambridge, MA: Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

Trawinski, K. (2010). *A fuzzy classification system for prediction of the results of the basketball games*. In: International Conference on Fuzzy Systems.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison Wesley.

Tukey, J. W. (1990). Data-based graphics: Visual display in the decades to come. *Statistical Science*, *5*(3), 327–339.

Wetzels, R., Tutschkow, D., Dolan, C., van der Sluis, S., Dutilh, G., & Wagenmakers, E.-J. (2016). A Bayesian test for the hot hand phenomenon. *Journal of Mathematical Psychology*, *72*, 200–209.

Yu, L., & Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, *5*, 1205–1224.

Yuan, L. H., Liu, A., Yeh, A., Kaufman, A., Reece, A., Bull, P., … Bornn, L. (2015). A mixture-of-modelers approach to forecasting NCAA tournament outcomes. *Journal of Quantitative Analysis in Sports*, *11*(1), 13–27.

Zaveri, N., Shah, U., Tiwari, S., Shinde, P., & Kumar Teli, L. (2018). Prediction of football match score and decision making process. *International Journal on Recent and Innovation Trends in Computing and Communication*, *6*(2), 162–165.

Zdravevski, E., & Kulakov, A. (2010). System for prediction of the winner in a sports game. *ICT Innovations 2009*, *Part 2*, 55–63.

Zhang, P. Z. (2000). Neural network for classification: A survey. *IEEE Transactions on Systems, Man and Cybernetics (Applications and Reviews)*, *30*(4), 451–462.

Zimmermann, A., Moorthy, S., Shi, Z. (2013). Predicting college basketball match outcomes using machine learning techniques: Some results and lessons learned. Retrieved from https://arxiv.org/pdf/1310.3607.pdf.

How to cite this article: Horvat T, Job J. The use of machine learning in sport outcome prediction: A review. *WIREs Data Mining Knowl Discov*. 2020;e1380. https://doi.org/10.1002/widm.1380