A Systematic Literature Review of Machine Learning Applications for Team Formation Problems

Literatür Makalesi/Literature Article

Soner KARATAŞ[1], Hüseyin ÇAKIR[2]

[1]Gazi University, Graduate Schools of Informatics, Ankara, Turkey
[2]Gazi University, Gazi Faculty of Education, Ankara, Turkey
soner.karatas@gazi.edu.tr, hcakir@gazi.edu.tr
(Geliş/Received:04.01.2024; Kabul/Accepted:16.05.2024)
DOI: 10.17671/gazibtd. 1414527

Abstract **—** With the development of technology, the variety and number of data held for any process has increased exponentially. By processing and analyzing this data, it is possible to solve many problems. Selection of the most appropriate team member and correct team formation in the activities carried out by the team are the factors that affect the success and result of teamwork. For this reason, the problem of team member selection and team formation has become one of the increasing research topics in recent years. Researchers from different disciplines are trying to develop tools, techniques and methodologies to ensure a successful team building process. Machine Learning (ML) methods have become one of the methods that have started to be used in team formation and team member selection problems in recent years. The successful outcome of this problem depends on the correct collection and processing of data and the selection of appropriate machine learning methods. The aim of this article is to present a systematic literature review of machine learning methods applied in team formation and team member selection problems, and to show which machine learning methods are applied in this field and their performance. Articles on the subject were searched in six scientific databases. In addition to providing fundamental information about ML methods, this review also supports new research efforts on team formation problems.

Keywords**—** team formation, player selection, machine learning, systematic literature review

Ekip Oluşturma Sorunlarına Yönelik Makine Öğrenimi Uygulamalarına İlişkin Sistematik Bir Literatür Taraması

Özet**—** Teknolojinin gelişmesiyle birlikte herhangi bir sürece ait tutulan veri çeşitliliği, veri sayısı katlanarak arttı. Bu verilerin işlenmesi ve analiz edilmesiyle bir çok problemin çözümü mümkün olabilmektedir. Ekip tarafından gerçekleştirilen faaliyetlerde en uygun ekip üyesinin seçimi ve doğru ekip oluşumu ekip çalışması başarısını ve sonucunu etkileyen unsurdur. Bu nedenle ekip üyesi seçimi, takım oluşturma problemi son yıllarda artan araştırma konularında biri olmuştur. Farklı disiplinlerden araştırmacılar, başarılı bir ekip oluşturma sürecini sağlayabilmek için araçlar, teknikler ve metodolojiler geliştirmeye çalışmaktadırlar. Makine Öğrenmesi (ML) yöntemleri takım oluşumu, ekip üyesi seçimi problemlerinde son yıllarda kullanılmaya başlayan yöntemlerden biri olmuştur. Bu problemin başarılı sonucu verilerin doğru bir şekilde toplanması, işlenmesi ve uygun makine öğrenme yöntemlerinin seçimine bağlıdır. Bu makalenin amacı takım oluşumu, ekip üyesi seçimi problemlerinde uygulanan makine öğrenme yöntemlerinin sistematik bir literatür taramasını sunmak, bu alanda hangi makine öğrenme metodlarının uygulandığını ve bunların performansını göstermektir. Altı bilimsel veri tabanında konuyla ilgili makaleler araştırılmıştır. Bu inceleme ML yöntemleri hakkında temel bilgiler sağlamanın yanısıra takım oluşumu problemlerinde yeni araştırma çalışmalarını da desteklemektedir.

Anahtar Kelimeler **—** takım oluşumu, oyuncu seçimi, makine öğrenimi, sistematik literatür taraması

1. INTRODUCTION

In addition to the activities that can be carried out individually, there are activities planned to be carried out by the team. When it comes to choosing the team members who will carry out this activity among the alternatives, the success and performance of the team will vary depending on the choice. Team formation is crucial because team success depends largely on the appropriate assignment of team members to the teams [1]. Some teams could be more effective than others, only because of the composition of the characteristics of its members [2]. Attributes such as communication skills, teamwork experience, and personality traits, are criteria that affect team effectiveness and have an impact on team collaboration, efficiency and productivity [1]. The assignment will need to be made according to the characteristics of the candidate to be selected for the task. Therefore, it is important to implement and apply an effective technique to ensure (to some extent) the optimal team composition [1]. Many academic personnel have studied on to research novel techniques and methodologies to solve this problem. Team formation, which is not based on any basis and is based solely on the selection of team members based on human instinct, is an issue that is criticized for many reasons such as loss of time, accuracy and effort. Therefore, demand and orientation for methodology, techniques and tools that enable team formation by identifying the best team member is increasing recently.

Data production is increasing exponentially every day. Using machine learning to extract information from this data in different fields has been widely used in recent years. Machine Learning algorithms and associated Artificial Intelligence technologies are helpful in various fields such as prediction and decision making [3]. ML approaches have the ability to handle high dimensional and multivariate data, and to extract hidden relationships within data in complex and dynamic environments (such as, industrial environments) [4]. Using the machine learning approach in selecting the most suitable team member and team building problems has become one of the popular areas of study in recent years. The results show that machine learning algorithms can be used for player selection and team formation process [5]. However, the success of these applications depends on which ML technique is applied to which data and under which conditions.

The purpose of this article is to present a Systematic Literature Review (SLR) of articles whose topic is team formation and team member/player selection using machine learning techniques. This article serves as an important resource on machine learning techniques, the characteristics of the data involved in their use, problems in implementation, and recent advances, inspiring new research efforts on team formation and team member selection.

The contents of other chapters of the article are given below. Chapter 2 explains how SLR is implemented. Chapter 3 provides an overview of machine learning. Chapter 4 includes considerations regarding the research questions and brief information of the literature reviewed. Finally, in chapter 5, the conclusions and contributions of this article are presented.

2. LITERATURE REVIEW PLANNING PROTOCOL

2.1. Research Methodology

SLR is a process that allows the collection of relevant evidence that meets predetermined eligibility criteria on a particular topic and answers to formulated research questions [6]. Although they are similar in general terms, methodologies of systematic literature reviews may differ. For this purpose, different methodologies have been included and applied in many sources. In this article, the five-stage steps for the research methodology suggested by Petersen et al. [7] were applied. This systematic mapping method proposed by Peterson et al. aims to provide an overview of the area of interest, reduce systematic errors, and increase the legitimacy of the analyzed data for more reliable results [1], [7].

2.2. Search Strings

Keywords for search strings were selected based on words commonly taking part in the literature. For the systematic literature study, searches were made in 6 different online databases containing publications in this field. These are Scopus, IEEE, Web of Science, Ebsco (Academic Search Ultimate), Science Direct, Springer Link databases. Since database search functions work differently, search strings have determined as follows.

• Scopus: Article title-abstract-keywords (“team formation” AND “machine learning”) OR (“player selection” and “machine learning”) OR (“player ranking” and “machine learning”)

• IEEE: Abstract (“team formation” AND “machine learning”) OR (“player selection” and “machine learning”) OR (“player ranking” and “machine learning”)

• Web of Science: Abstract (“team formation” AND “machine learning”) OR (“player selection” and “machine learning”) OR (“player ranking” and “machine learning”)

• EBSCO (Academic Search Ultimate): Abstract or author-supplied abstract (“team formation” AND “machine learning”) OR (“player selection” and “machine learning”) OR (“player ranking” and “machine learning”)

• Science Direct: Title, abstract or author-specified keywords (“team formation” AND “machine learning”) OR (“player selection” and “machine learning”) OR (“player ranking” and “machine learning”)

• Springer Link: (“team formation” AND “machine learning”) OR (“player selection” and “machine learning”) OR (“player ranking” and “machine learning”)

6 databases were searched. As a result of this search, 376 articles that fell within the search criteria were found. The exclusion criteria specified in item 2.4 were applied to these articles. It was determined that 28 articles were appropriate for the subject of this article.

2.3. Research Questions

Research questions prepared for machine learning applications in team formation problems are given below. While determining the research questions, similar literature studies on machine learning and the needs of new research studies in this field were taken into considerations.

Q1) What are the current studies and research on team formation problems using ML techniques?

Q2) What are the machine learning methods that are being used on team formation/player selection problems?

Q3) What are the data/datasets used on machine learning and what are their properties?

2.4. Screening of Papers for Inclusion & Exclusion

After examining publications obtained through the search string search, only 28 articles were included in the systematic literature review. Different criteria have been created to determine articles suitable for the research topic and objective. Criteria for inclusion and exclusion are set out below.

Inclusion Criteria:

- Include papers related to team formation studies using machine learning methods
- Include papers from 2000 (year)
- Include papers only in English language

Exclusion Criteria:

- Papers containing team formation studies using different methods other than machine learning
- Not relevant to the research questions
- Papers in languages other than English
- Only one of the same publications in different databases is taken into account and the others are excluded.

2.4. Data Extraction & Mapping

All articles were reviewed and the information listed below was extracted; title, author, year, keywords, country, publication type, publisher, research type, used machine learning methods, data source and type, study summary, contribution and future work proposal. These extracted data were analyzed for the research questions given in the previous sections.

3. MACHINE LEARNING FUNDAMENTALS

In daily life, people make many decisions. In which areas they invest in different periods, which study subjects they pursue, which school they prefer, etc. When making these decisions, it may be difficult for them to make logical decisions due to reasons such as not having access to the correct information, not evaluating the correct information in an appropriate and methodological way, being influenced by their emotions, and missing details due to the mass of data. At this stage, machine learning, which models human thought structure and decision-making ability, evaluates all situations, examines millions of data very quickly and enables rational decisions to be made in a very short time, is the solution for such problems. For machine learning, it can be said that it is a developing branch of computational algorithms designed to imitate human intelligence [8], and it is the modeling of a problem with an algorithm suitable for data [9], [10].

Machine learning involves many steps, starting from data to applying the model. Below is the figure for this.

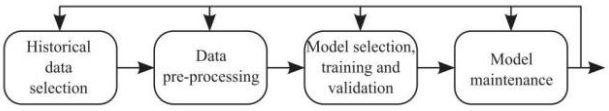

Figure 1. Machine learning steps [4]

One of the main factors affecting the results of machine learning studies is working with the most appropriate and accurate data set. Obtaining this data is the first step in machine learning. The second step is the data preprocessing step, where the data is converted and processed into a suitable form for the ML model. Below are some of these processes with examples.

- Data transformation,
- Data cleaning
- Data reduction

The third step consists of three main components. In model selection, the appropriate model that can solve the problem is selected. The next component is training the model based on the data. In the last component, the model's potential to represent the system is evaluated. The final step of machine learning is model maintenance and includes checks and

improvements to application performance that may change over time.

4. RESULTS OF SLR

4.1. What is the latest studies and research on team formation problems using ML techniques? (Q1)

Totally 28 articles were found in the search made in 6 large databases according to the criteria specified in Chapter 2. The table including general information about these articles is given below.

In the investigation of the field in which the studies were carried out, 10 articles on the selection-ranking of football players, 7 articles on the selection of cricket players, 4 articles on the selection of software team members, 3 articles on the selection of expert team members, 3 article on the selection-ranking of basketball players and 1 article on the selection-ranking of ice hockey players were identified. From these results, it has been seen that studies mainly focus on creating sports and software teams. It has been concluded that machine learning methods have not been studied for other topics in team formation problems and that there are areas where they can be conducted.

In the investigation, it was seen that although studies on team formation problems took place before 2015, studies using machine learning techniques were not available. Figure 2 shows the number of articles published after 2000 (using this article's extraction criteria). It appears that the first article was published in 2016. This research shows that articles on ML techniques in team formation and team member selection have become more intense, especially after 2020. Therefore, it is possible to say that the application of machine learning for team formation problems is a new application area. It can be said that interest in this field of research has increased after 2020. On the other hand, when the studies on publication type are examined, the number of articles is 18 and the number of conference papers is 10.

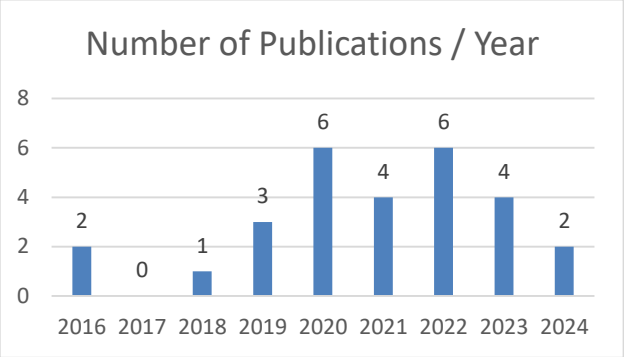

Figure 2. Number of publications / Year

It was examined in which countries the articles were prepared. India has been the country where the most intensive studies in this field have been carried out, with 9 articles. Thailand, where similar articles were published on the same subject, came in second place with 3 articles. Turkey, Bangladesh and Iran are the countries that have published 2 publications in this field. Taiwan, Ireland, Palestine, Canada, Australia, Austria, Greece, Pakistan, Saudi Arabia, USA, Sweden, Italy are among the countries that publish in this field with 1 article each. Considering that the subject of machine learning is a developing field in recent years, it is seen that researchers in India tend to use machine learning techniques in solving team formation problems in the 2020s.

Table 1. Information about studies

Ref. No	In which field?	Publication Type	Database	Country	Year
3	Cricket player selection	Article	Scopus	India	2022
5	Football player selection	Article	Scopus	Turkey	2021
11	Football player selection	Conference Paper	Scopus	India	2022
12	Football player selection	Conference Paper	Scopus	Bangladesh	2019
13	Cricket player selection	Conference Paper	Scopus	India	2023
14	Cricket player selection	Conference Paper	IEEE	India	2020
15	Cricket player selection	Conference Paper	IEEE	India	2020

Ref. No	In which field?	Publication Type	Database	Country	Year
16	Software team member selection	Article	Scopus	Thailand	2022
17	Expert team member selection	Article	Scopus	Taiwan	2022
18	Software team member selection	Conference Paper	Scopus	Palestine	2022
19	Software team member selection	Article	Scopus	Thailand-Pakistan	2021
20	Football player selection	Conference Paper	Scopus	India	2021
21	Expert team member selection	Article	Scopus	Ireland	2020
22	Software team member selection	Conference Paper	Scopus	Thailand	2020
24	Expert team member selection	Article	Scopus	India	2016
25	Football player selection	Article	Academic Search Ultimate (EBSCO)	Canada	2018
26	Football player selection	Article	Academic Search Ultimate (EBSCO)	Turkey	2023
27	Football player selection	Article	Academic Search Ultimate (EBSCO)	Iran	2019
28	Cricket player selection	Article	Science Direct	India	2023
29	Basketball player selection	Article	Science Direct	Australia	2024
30	Cricket player selection	Article	Science Direct	Bangladesh	2022
31	Football player selection	Article	Springer link	Iran-Austria	2023
32	Basketball player ranking	Article	Springer link	Greece	2024
33	Basketball player ranking	Article	Science Direct	Pakistan-Saudi Arabia	2021
34	Football player ranking	Conference Paper	Web of Science	USA	2016
35	Ice hockey player ranking	Conference Paper	Scopus	Sweden	2020
36	Football player ranking	Article	Scopus	Italy	2019
37	Cricket player ranking	Article	Scopus	India	2020

4.2. What are the Machine learning methods that are being used on team formation/player selection problems? (Q-2)

One of the search questions of the article was to determine which ML methods are used in team formation problems. In the review conducted for this purpose, it was seen that more than one machine learning algorithm was used in many articles. It has been stated that the purpose of this is to achieve the most efficient result by using different machine learning algorithms and to compare these algorithms. The figure showing the number of machine learning algorithms used is given below in Fig.3. Accordingly, it was seen that the most used machine learning algorithms were Random forest (16), Decision Trees (13), Support Vector Machines (13). Since these results were obtained in 28 article reviews, it is seen that these algorithms were used in %57,1 - %46,4 of the articles. On the other hand, the least used machine learning algorithms were Polynomial Regression, Q Learning with 1 article each (Polynomial Regression was stated in the others groups). It has been observed that the use of boosting algorithms such as XGBoost, Catboost, etc. has increased in recent years. When the machine learning classes were examined, it was seen that classification class algorithms were used in 21 articles, regression class algorithms were used in 16 articles, and the clustering class algorithm was used in 3 articles (Figure 4). From these results, it has been determined that ML methods are used extensively for classification and regression purposes in team formation problems. Another important data was obtained from machine learning types (Fig 5). Accordingly, supervised learning methods were used in 25 articles (%89,2). Unsupervised learning method was used in only 3 articles (%10,7). It was determined that the reinforcement method was used in 1 article (%3,5). As a result, it has been observed that supervised learning methods are used extensively in team formation problems.

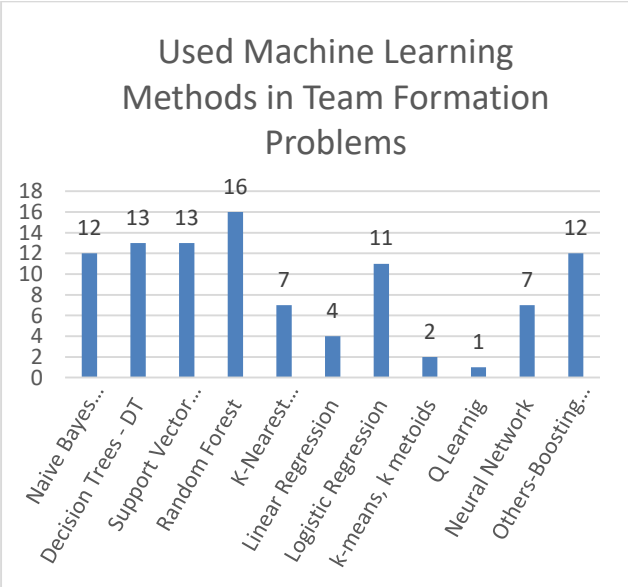

Figure 3 Used machine learning methods in Team formation problems

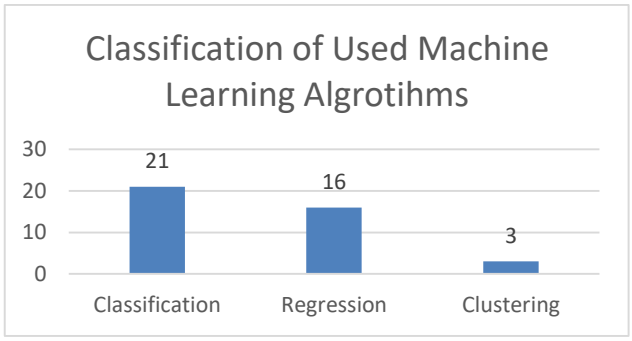

Figure 4. Classification of used machine learning algorithms

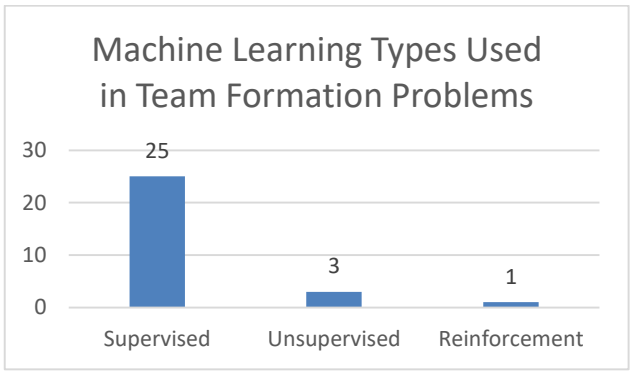

Figure 5. Machine Learning Types

In the study of Datta and Rudra [11], the researchers evaluated the performance of players in football clubs. For performance evaluation of players, they used five machine learning models and compared their efficiency values. These machine learning methods are Decision Tree (DT), Support Vector Machine (SVM), Linear Regression, Random Forest (RF) and XGBoost. The XGBoost method is superior to other methods with low error values. When the article is examined, it is seen that the technique with the lowest performance was obtained in support vector machine.

Shahriar and his colleagues [12] studied to develop a player classification and selection method for a football team. The players were classified based on their performances using multiple ML techniques. Five algorithms were used for the classification. These are: Support Vector Machine (SVM), Decision Tree, Naïve Bayes Classifier (NB), Random Forest Classifier and KNN Classifier. A small comparison was made between them after the classification. Five factors were included in the comparison. As a result of the comparison, the best performance was obtained from the Decision Tree algorithm, and the weakest performance was obtained from the Support Vector Machine algorithm.

Sumathi, Prabu and Rajkamal [13] developed a system to evaluate the performance of cricket players. Linear Regression technique was used to predict the players' performance. K-means algorithm was used to find players with similar performance. Random Forest technique was applied to generate the ranking list.

Shetty and her colleagues [14] developed the selection of the best 11 players in the Indian cricket team with their proposed model. Different techniques such as Random Forest, SVM, Decision Tree and Logistic Regression were used. Random Forest were the most successful techniques obtained results.

Santra and colleagues [15] developed a technique to predict the batsman's (cricket player) order by evaluating the batsman's past performance. The novelty of the technique lies in designing the algorithm and taking into account multiple cricket parameters instead of a single parameter in predicting the best batsman. Polynomial Regression and Linear Regression models were created for prediction. Polynomial Regression was found to outperform linear regression in every aspect.

Random Forest, Decision Tree, Second Order Discriminant Analysis (QDA), Neural Network, Naive Bayes, k-Nearest Neighbors (kNN), Logistic Regression machine learning algorithms were used and compared for the team effectiveness scoring function for software development team in Assavakamhaenghan et al. [16] study. In this study, the most successful results were obtained with the Random Forest model.

Chang et al. [17] proposed an approach using the Reinforcement Learning (RL) method to create a solution to the team formation problem.

In Ishi and his colleagues' paper [3], the hybrid approach of CS-PSO with machine learning models was used to find the right set of team combinations from the group of players. 9 machine learning methods, including K Nearest Neighbors, Random Forest, Decision Tree, Gradient Boosting Algorithm, Logistic Regression, Naive Bayes, XGBoost, CatBoost and Support Vector Machine were applied and the most successful results were obtained from SVM.

Tanbour et al. [18] developed a ML technique that can better match software development experts with software project tasks. Three different models were tested in their studies: Random Forest Classifier, Decision Tree Classifier and Logistic Regression. Random Forest Classifier has been the most successful model.

In the study of Tuarob's and his colleagues [19], they proposed a machine learning model that can recommend suitable software team members for software development tasks. Seven classification algorithms were tested and Random Forest received the best score for all datasets.

Ghar et al. [20] designed a ML model that predicts the future performance (VAEP value) by examining the past performances of football players.

Abidin [5] applied ML algorithms for team formation and player selection for the players of a football team. 7 machine learning algorithms were used and their performances were compared. These algorithms were Logistic regression, Random Forest (RF), Classification and Regression Tree (CART), Artificial Neural Networks (ANN), Decision Trees (DT), Support Vector Machines (SVM) and Bayes Theorem.

In their proposed framework, Keane et al. [21] incorporated both the individual's attributes and the topological features of the individual's network into a ML link prediction task. They developed models such as LR, SVM and CART Using Python's XGBoost module.

Assavakamhaenghan et al. [22] studied on possibility to adopt the team recommendation algorithm proposed by Liu et. al. [23] to develop a software team recommendation. They proposed the logistic regression approach to recommend suitable software team members. The approach take both individual strength and collaborative efficiency among team members into account to give a recommendation.

Three machine learning techniques were applied to select the best employees for effective team formation and their results were compared in Krishankumar and Ravichandran's [24] article. The three methods were Ensemble Decision Tree (EDT), Artificial Neural Network (ANN) and Decision Tree. The results reveal that the EDT approach performs better.

In Tosato and Wu's article [25], Projective Adaptive Resonance Theory (PART) was used to provide data-based sports decisions. PART is the machine learning projective clustering algorithm and based on neural network. It was seen that PART provides a purely data-driven analysis to identify attributes for a group of players in an unsupervised way.

Buyrukoglu and Savas [26] studied on a technique to classify footballer positions using a stacked ensemble ML model. Firstly they used 4 filter based feature selection methods to choose optimal feature subsets. Then 2 level stacked ensemble machine learning algorithms were used to determine the position of the football player. In Level-0, Gradient Boosting, Random Forest and Deep Neural Network algorithms were used as based models and then in

Level-1, the Logistic Regression algorithm was used as a meta-model.

Machine learning approaches were employed to establish a ranking for players in Maanijou and Mirroshande's article [27]. Different algorithms (SVM, Logistic Regression, PART, Naive Bayes, etc.) were used to classification and ranking. Experiments were done in the Persian soccer league. The study showed promising results for ranking.

Manju and Philip [28] studied on a novel framework that ranks batsmen (cricket players) on their performances. New performance index was created using different machine learning algorithms (Logistic Regression, Support Vector Machine, RF, XGB Classifier, CART and Naive Bayes). Then players were clustered with k-means clustering algorithm to identify the best players.

Ke, Bian and Chandra [29] developed a unified framework to categorize the players and built the optimal team model. Principal Components Analysis (PCA) was used to reduce the number of features describing player performance in the unsupervised phase. Then simple neural network was used to build a team model in the supervised learning phase.

Tirtho, Rahman and Mahbub [30] studied on to forecast player performance for future cricket competitions. Different machine learning methods were used for this purpose (Random Forest, Decision Trees, K-nearest Neighbors, Support Vector Machine and Naive Bayes). It was determined that Random Forest produced the best accurate prediction models for cricket players.

Nouraie, Eslahchi and Baca [31] developed deep learning models to obtain right scores for players' positions. They designed a procedure that can identify the best players for each position.

To forecast NBA player performance several ML models (Decision tree, Linear regression, Random forest, etc.) were employed in the study of Papageorgiou et al. [32]. It was seen that the approach of blending data from the last ten seasons and last three seasons increased the prediction accuracy and model stability. They emphasized that the use of standard statistics, advanced statistics datasets, long-term and short-term data, is crucial for obtaining right predictions.

Mahmood et al. [33] presented prediction mechanism to explore rising stars in basketball using machine learning methods. Different machine learning methods like Maximum Entropy Markov Model (MEMM), Classification and Regression Trees (CART), Support Vector Machines (SVM) and Naïve Bayes (NB) were applied to find a function which can assign class label to a feature set. It was seen that the Maximum Entropy Markov Model was dominant in all data sets in terms of F-measure score.

Brooks et al. [34] proposed the Pass Shot Value (PSV) which is a metric to predict the importance of a pass resulting with a shot. They used support vector machine model to estimate whether or not a given pass generates a shot. When they ranked players in La Liga with more than 200 passes with stated metric, they saw some of the outstanding players at the top.

Persson et al. [35] compared which features best predict the success of ice hockey players in different positions. 6 different machine learning methods were used (Logistic regression, Bayesian network, k-Nearest neighbor, Naive Bayes, Decision tree and Random forest) to predict players' ranking tier for 3 player positions. They compared the result of the models and they concluded that two Bayesian classifiers had best performance and sensitivity.

PlayeRank which is a data-driven framework that offers evaluation of the performance of football players has been presented by Papalodios et al. [36]. They used Linear Support Vector Machine (SVM) to solve classification problem. They compared PlayeRank with known algorithms for performance evaluation in soccer and they saw that PlayeRank outperforms the competitors.

Kaviya, Mishra and Valarmathi [37] ranked the players, based on the Player Ranking Index using machine learning techniques. They used various algorithms for ranking which include Decision Tree, JRIP reduced error pruning algorithm, Support Vector Machine, Random Forest, and Naïve Bayes classifier, etc. JRIP was seemed the most promising amongst all the algorithms.

In order to compare the results of this study, other literature studies on machine learning were also examined. Carvalho et al. [38] explored a systematic literature review, covering the main papers of Predictive Maintenance (PdM) using ML techniques. In this study, the most used ML methods to perform PdM were found to be Random Forest (RF), Artificial Neural Network (ANN) and Support Vector Machine (SVM). Another literature study is the work of Xames et al. [39] on recent trends of machine learning applications in additive manufacturing. The most used ML methods in this literature study were ANN, SVM, Ensemble (including RF) based ML methods. Similarly, in our literature review in the field of team formation, RF, SVM and Decision Trees were among the most frequently used ML techniques. It is considered that the reason for this choice is that RF and SVM can be used in classification and regression problems and provide versatile flexible solutions. However, in the literature studies examined, it is seen that the use of the Boosting algorithm family

(XGBoost, Catboost etc.) has increased, especially in recent years.

4.3. What are the data/datasets used on machine learning and what are their properties?

To answer question Q3, all articles were examined on this axis and the results were shared below.

Table 2. Properties of Data

Ref. No.	Data Type	Source (from)	Volume of data	Data cleaning and Pre-processing	Train/Test Data
3	Real Data	ESPN Cricinfo website.	101*4 players. The number of columns could not be determined.	Pre-processing was carried out.	%70 training-%30 testing
5	Real and Synthetic data	From Hit/it Assistant and the coach evaluations. Synthetically generated data also.	Football team has only 21 players. Synthetic data generation is done as 10 synthetic instances for one player. So, the size of the dataset becomes 231 instances in total. Feature : 28	Preprocessing and standardization of two data sources	Not mentioned
11	Real Data	SoFIFA.com	89 columns (attributes) and 18.207 rows (players).	Filter and clean data	%80 training - %20 testing
12	Real Data	www.footystats.org	The number of rows could not be determined. Eight main attributes.	Not mentioned	Multiple training datasets.
13	Real Data	https://www.kaggle.com	304 records, 13 attributes.	Remove noisy and missing data.	Not mentioned
14	Real Data	espncricinfo.com	The number of rows could not be determined.	It is stated that preprocessing is applied.	%80 of the dataset was used for training, %20 of the dataset was used to test.
15	Real Data	IPL website.	The number of rows and columns could not be determined.	Not mentioned	Train regression models to predict batsman positions. Test data from ball by ball data of 2019
16	Real Data	Jira (Apache and Atlassian datasets)	507.319 and 238.322 Rows	Cleaning of data	%80 training and %20 testing
17	Real Data	SNAP and SIoT	The number of rows and columns could not be determined.	Not mentioned clearly	Not mentioned

Ref. No.	Data Type	Source (from)	Volume of data	Data cleaning and Pre-processing	Train/Test Data
18	Real Data	https://data.world website.	1.000 computer development-related experts (Rows). Columns: could not be determined.	Not mentioned	%60 training and %40 testing
19	Real Data	Jira(Atlassian, Apache, and Moodle)	The number of rows could not be determined. There are about 20 features.	The collected data are then prepossessed (e.g. removing incomplete issues).	Each dataset are separated into %80 training and %20 testing sets.
20	Real Data	EA Sports, FIFA, FBref, Wyscout	511 players. 104 attributes	Clean not mentioned, preprocess yes	%80 training and %20 testing.
21	Real Data	USPTO data set.	Not mentioned	Not mentioned	New datasets were created for the purpose of training machine learning models. The model was tested on unseen dataset.
22	Real Data	Jira (Moodle).	Total number of projects about 26.800. Feature number: 6	Not mentioned clearly	%80 training, %20 testing
24	Real Data	Several websites like Odesk, Elance etc.	474 recipients (row), eight predictors (Column)	Normalizing the data using max-min normalization.	The training and testing data are split equally for training and validating methods.
25	Real Data	Football Manager 2018	Forty seven attributes for twenty four soccer player	Not mentioned	Not mentioned
26	Real Data	https://www.kaggle.com/ karangadiya/fifa19	89 columns and 18.207 rows	Missing/null values are removed from the dataset.	%80 training and %20 testing
27	Real Data	Online freely sources	495 players, 20 features	Missing and noisy data are removed from the dataset	Not mentioned
28	Real Data	Websites ESPNcricinfo and IPLT20	283 players	Data extraction and data cleaning are performed	%70 training and %30 testing
29	Real Data	5 data sets from the NBA official website	8.511 entries and 71 features	Dimensional reduction and visualization	Not mentioned

Ref. No.	Data Type	Source (from)	Volume of data	Data cleaning and Pre-processing	Train/Test Data
30	Real Data	www.howstat.com, www. cricmetric.com	152 players	Dimensional reduction	%80 training and %20 testing
31	Real Data	Sofifa dataset	18.034 players and 48 attributes	Not mentioned	%80 training data, %20 testing
32	Real Data	National Basketball Association: www.nba.com	79.036 instances, 67 features, 203 players	Cleaning and pre-processing are performed	%70 training data, %30 testing data
33	Real Data	A sports website www.espn.com/nba/	100 players	Not mentioned clearly	Not mentioned clearly
34	Real Data	2012-13 La Liga season	Not mentioned clearly	Normalization is performed	%80 training data, %20 testing data
35	Real Data	https://www.hockey-reference.com/	Not mentioned clearly	Some data was removed. Numerical data was normalized.	%80 training data, %20 testing data
36	Real Data	https://wyscout.com	31.496.332 events, capturing 19.619 matches, 296 clubs and 21.361 players.	Normalization is performed	%80 training data, %20 testing data
37	Real Data	Dataset obtained from a cricsheet website	Approximately 550 rows, 22 features	It states that the data sets are processed, but does not clearly state the processes.	Not mentioned clearly

When the data sources are examined, it is seen that the majority of them are collected from public websites. Many websites that keep sports player statistics were used in the studies. Open source software systems hosted on the Jira platform, such as Apache, Atlassian and Moddle, were preferred in software team member selection. When the data sets were examined, it was evaluated that the most distinctive source was in Abidin's work [5]. In this study, datasets were obtained from 2 main sources. The first source is the Hit/it device, an electronic sports system that can measure a player's skills. The coach's evaluations are the second source.

When the data type issue was examined, it was seen that real data was used in all articles. Only in Abidin's article [5] synthetic data were used as well as real data. Since the number of players was 21 and it was thought that this number was not enough for classification, 10 synthetic data were produced for each player. According to the results of these two examinations, we can state that it is preferred to use statistical real data obtained from open sources in team building problems.

When the articles were evaluated according to their data size, it was determined that some of them did not include this issue clearly. In article no. [16], a total of around 700.000 lines of data were studied. However, the number of features cannot be clearly seen in this article. In article no [11], a large data set consisting of 18.207 rows (18.207 players) and 89 columns (attributes and skills of the players) was used. In Abidin's article [5], where the lowest data set was used and therefore synthetic data addition was needed, 231 players and 28 features were included.

When the majority of the articles were examined, there were statements that data clean and data preprocessing processes were used. For example, in the study of Sumathi et al. [13] before analysis the dataset is preprocessed to remove noisy and missing data. In this article, each numeric column is normalized because it has a different representation.

The last issue examined regarding the data was the ratio of training and testing data. While this rate information was not clearly shared in 11 articles, the amount of this rate was given in 17 articles. Accordingly, the number of articles using the training test ratio %80-%20 is 12, the number of articles using %70%30 is 3, the number of articles using %60-%40 is 1 and the number of articles using %50-%50

was 1. According to these values, we can say that the %80-%20 scale is commonly used for the training test data usage rate in team formation problems.

When other literature studies on machine learning were examined, it was determined that in the studies of Carvalho et al. [38], %89 real data was used and %11 synthetic data was used in the literature in the field of PdM. In this study, it was observed that real data was used in all studies on team formation and synthetic data was additionally used in one study. It was evaluated that this similarity resulted from the availability of historical real data in both study areas.

5. CONCLUSION

In this article, a systematic literature review on team formation problems was conducted using machine learning techniques. It was aimed to answer the research questions specified in the planning protocol. As a result, it has been seen in many articles that different ML techniques are used to evaluate performance among them. The majority of articles stated that the best performance result was obtained from Random Forest (RF) technique. It has been seen that Supervised Machine Learning methods are dominant in practice. However, it has been determined that ML algorithms are used extensively for classification and regression purposes in team formation problems.

It has been revealed in the reviewed articles that the application of ML techniques to team formation problems is mostly in sports and software fields. Looking at the article publication momentum, articles subject on team formation using ML techniques have an increasing momentum after 2020. For this reason, it is anticipated that the volume of articles in this scope will continue to increase.

Within the scope of this study, the data/datasets used in machine learning and their characteristics were also investigated. Data in machine learning was mostly obtained from public websites and was real. In terms of data size, there were examples where a large amount of data was used, as well as articles where a small amount of data was used. Similar to the results of this literature study, it was observed that real data was used instead of synthetic data in other literature studies on machine learning. Since the success of machine learning depends on obtaining accurate data, the data acquisition processes were also examined and it was determined that the data preprocessing stages was explained clearly and in detail in many articles. It was considered that not explaining this issue clearly in some articles was a deficiency, and it was evaluated that these preprocessing stages should be given in detail in articles related to machine learning. In the articles reviewed, examples of data cleaning, data reduction and data normalization were mostly included in the preprocessing stages. When the separation of training and test data sets was examined, it was seen that the %80-%20 scale is commonly used in team formation problems.

It can be stated that machine learning techniques such as RF, SVM and Decision Tree have been implemented extensively to team formation problems. When other literature review studies on machine learning were examined, it was seen that RF, SVM, ANN methods were most commonly used in solving other problems. It has been evaluated that the reason for this similarity is that these methods provide versatile and flexible solutions. Similar to the findings of this literature study, it has been observed in other literature studies that the use of ensemble algorithms has increased.

However, it is considered that there are still some aspects that can be further researched in solving team formation problems with machine learning. In this regard, suggestions for future research are as follows;

• develop works that apply machine learning methods other than sports and software teams, such as quality assessment teams, and comparing the results

•comparison of performance rates by applying different data sets for the same problem.

In summary, this comprehensive review highlights the applicability of machine learning techniques to team formation problems and sets the stage for future research efforts.

REFERENCES

[1] G. Stavrou, P. Adamidis, J. Papathanasiou, K. Tarabanis “Team Formation: A Systematic Literature Review”, *Int. Journal of Business Science and Applied Management*, 18(2), 2023.

[2] J. Juárez, C. Santos, F. A. A. M. N. Soares, R. Vita, R. P. Francisco, J. P. Basto, S. G. S. Alcalá, “A systematic literature review of machine learning methods applied to predictive maintenance”, ACM Computing Surveys, 54(7), 2021.

[3] M. Ishi, J. Patil, J. Jhang, V. Patil, “An efficient team prediction for one day international matches using a hybrid approach of CS-PSO and machine learning algorithms”, Array 14, 2022.

[4] T. P. Carvalho, C. Santos, C. A. Brizuela, “A Comprehensive Review and a Taxonomy Proposal of Team Formation Problems”, Computers & Industrial Engineer, 2019.

[5] D. Abidin, “A case study on player selection and team formation in football with machine learning”, *Turkish Journal of Electrical Engineering & Computer Sciences*, 29, 1672 – 1691, 2021.

[6] W. Mengist, T. Soromessa, G. Legese, "Method for conducting systematic literature review and meta-analysis for environmental science research", MethodsX 7, 2020.

[7] K. Petersen, S. Vakkalanka, L. Kuzniarz, "Guidelines for conducting systematic mapping studies in software engineering: An update" Information and Software Technology, 64, 1–18, 2015

[8] I. El Naqa, M. J. Murphy, "What is machine learning?" In Machine Learning in Radiation Oncology, 3-11, 2015

[9] M. Atalay, E. Çelik, "Artificial Intelligence and Machine Learning Applications in Big Data Analysis", *Mehmet Akif Ersoy Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 9(22), 155-172, 2017

[10] E. Tosunoğlu, R. Yılmaz, E. Özeren, Z. Sağlam, "Eğitimde Makine Öğrenmesi: Araştırmalardaki Güncel Eğilimler Üzerine Inceleme", *Ahmet Keleşoğlu Eğitim Fakültesi Dergisi*, 3(2), 178-199. 2021

[11] M. Datta, B. Rudra, N. Mead, C. Rolland, "An Intelligent Decision Support System for Bid Prediction of Undervalued Football Players", 2nd International Conference on Intelligent Technologies (CONIT), 2022.

[12] T. Shahriar, Y. Islam, N. Amin, "Player Classification Technique Based on Performance for a Soccer Team Using Machine Learning Algorithms", 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), 2019.

[13] M. Sumathi, S. Prabu, M. Rajkamal, "Cricket Players Performance Prediction and Evaluation Using Machine Learning Algorithms", 2023 International Conference on Networking and Communications (ICNWC), 2023.

[14] M. Shetty, S. Rane, C. Pandita, S. Salvi, "Machine learning-based Selection of Optimal sports Team based on the Players Performance", Proceedings of the Fifth International Conference on Communication and Electronics Systems (ICCES 2020), 2020.

[15] A. Santra, A. Sinha, P. Saha, A. K. Das, "A Novel Regression based Technique for Batsman Evaluation in the Indian Premier League", 2020 IEEE International Conference for Convergence in Engineering, 2020.

[16] N. Assavakamhaenghan, W. Tanaphantaruk, P. Suwanworaboon, M. Choetkiertikul, S. Tuarob, "Quantifying efectiveness of team recommendation for collaborative software development", Automated Software Engineering, 2022.

[17] C. Chang, M. Chang, J. Jhang, L. Yeh, C. Shen "Learning to Extract Expert Teams in Social Networks", IEEE Transactions On Computational Social Systems, 9(5), 2022.

[18] Z. Tanbour, D. Khudarieh, H. Abuodeh, A. Hawash, "Forming Software Development Team: Machine-Learning Approach", 2022 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS), 2022.

[19] S. Tuarob, N. Assavakamhaenghan, W. Tanaphantaruk, P. Suwanworaboon, S. Hassan, M. Choetkiertikul, "Automatic team recommendation for collaborative software development", Empirical Software Engineering, 2021.

[20] S. Ghar, S. Patil, W. Tanaphantaruk, V. Arunachalam, "Data Driven football scouting assistance with simulated player performance extrapolation", 20th IEEE International Conference on Machine Learning and Applications (ICMLA), 2021.

[21] P. Keane, F. Ghaffar, D. Malone, "Using machine learning to predict links and improve Steiner tree solutions to team formation problems - a cross company study", Applied Network Science, 2020.

[22] N. Assavakamhaenghan, P. Suwanworaboon, W. Tanaphantaruk, S. Tuarob, M. Choetkiertikul, "Towards Team Formation in Software Development: A Case Study of Moodle", 2020 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2020.

[23] H. Liu, M. Qiao, D. Greenia, R. Akkiraju, S. Dill, T. Nakamura, Y. Song, and H. M. Nezhad, "A machine learning approach to combining individual strength and team features for team recommendation," Preceedings of The 13th International Conference on Machine Learning and Applications, 12, pp. 213–218, 2014.

[24] R. Krishankumar, K. S. Ravichandran, "A Novel Trio Combo Strategy For Efficient Team Formation Using Hybrid Triangulation Mechanism", *ARPN Journal of Engineering and Applied Sciences*, 11(5), 2016.

[25] M. Tosato, J.Wu, "An Application Of Part To The Football Manager Data For Players Clusters Analyses To Inform Club Team Formation", Big Data & Information Analytics. 3 (1), 43-54, 2018.

[26] S. Buyrukoglu, S. Savas, "Stacked-Based Ensemble Machine Learning Model for Positioning Footballer", *Arabian Journal for Science and Engineering*. 48, 1371-1383, 2022.

[27] R. Maanijou, S. A. Mirroshandel, "Introducing an expert system for prediction of soccer player ranking using ensemble learning", Neural Computing and Applications. 31, 9157-9174, 2019.

[28] M. K. Manju, A. O. Philip, "Novel method for ranking batsmen in Indian Premier League", Data Science and Management. 6, 158-173, 2023.

[29] Y. Ke, R. Bian, R. Chandra, "A unified machine learning framework for basketball team roster construction: NBA and WNBA", Applied Soft Computing. 153, 2024.

[30] D. Tirtho, S. Rahman, S. Mahbub, "Cricketer's tournament-wise performance prediction and squad selection using machine learning and multi-objective optimization", Applied Soft Computing. 129, 2022.

[31] M. Nouraie, C. Eslahchi, A. Baca, "Intelligent team formation and player selection: a data-driven approach for football coaches", Applied Intelligence. 53, 30250-30265, 2023.

[32] G. Papageorgiou, V. Sarlis, C. Tjortjis, "An innovative method for accurate NBA player performance forecasting and line-up optimization in daily fantasy sports", *International Journal of Data Science and Analytics*, 2024.

[33] Z. Mahmood, A. Daud, R. A. Abbasi, "Using machine learning techniques for rising star prediction in basketball", Knowledge-Based Systems, 211, 2021.

[34] J. Brooks, M. Kerr, J. Guttag, "Developing a Data-Driven Player Ranking in Soccer Using Predictive Model Weights", KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 49-55, 2016.

[35] T. L. Persson, H. Kozlica, N. Carlsson, P. Lambrix, "Prediction of Tiers in the Ranking of Ice Hockey Players", 7th International Workshop on Machine Learning and Data Mining for Sports Analytics, MLSA 2020, 89-100, 2020

[36] L. Pappalardo, P. Cintia, P. Ferragina, E. Massucco, D. Pedreschi, F. Giannotti "PlayeRank: Data-driven Performance Evaluation and Player Ranking in Soccer via a Machine Learning Approach", ACM Transactions on Intelligent Systems and Technology, 10(5), 1-27, 2019

[37] A. Kaviya, A. S. Mishra, B. Valarmathi, Comprehensive Data Analysis and Prediction on IPL using Machine Learning Algorithms", *International Journal on Emerging Technologies*, 2020.

[38] T. P. Carvalho , A. A. Fabrízzio, M. N. Soaresa, , V. Roberto, P. F. Roberto, J. P. Bastoc , S. G. S. Alcalá, "A systematic literature review of machine learning methods applied to predictive maintenance", Computers & Industrial Engineering. 137, 2019.

[39] D. Xames, F. K. Torsha, F. Sarwar, "A systematic literature review on recent trends of machine learning applications in additive manufacturing", *Journal of Intelligent Manufacturing*. 34, 2529–2555, 2023F