Grenze International Journal of Engineering and Technology, Jan Issue

Automated Player Selection for a Cricket Team using Machine Learning

Manasi Kasande[1] and Dr. Sunita Jahirabadkar [2]
[1-2]MKSSS's Cummins College of Engineering for Women, Pune, India
Email: manasi.kasande@cumminscollege.in, sunita.jahirabadkar@cumminscollege.in

***Abstract*—Machine learning is a field of computer science where historical data is used to predict future outcomes. Machine learning has found applications in various domains like finance, e-commerce, healthcare, sports, etc. In sports, machine learning is used for purposes such as predicting the outcome of a game, workload management and player recruitment. Cricket is a popular sport which has a huge following in commonwealth countries, especially in the subcontinent. Selecting the right team is the first step towards winning a cricket match. Machine learning algorithms can be used to select the correct team which is capable of winning the match. This paper presents a survey of various algorithms proposed in cricket team selection. We also present extensive discussions on factors governing cricket team selection and methods proposed in other sports. By reviewing the state-of-the-art cricket team selection methods, this study provides an overview of the current literature.**

***Index Terms*— cricket, machine learning, classification and regression.**

I. INTRODUCTION

Assembling a perfect team is the first step towards winning a cricket match. Typically, a cricket team's squad consists of fifteen players, comprising of batters, bowlers and wicketkeepers. All-rounders who are proficient in both batting and bowling are also a part of the squad. The playing-*11* for a cricket match is picked from these fifteen players. A batter can bat at different positions in a cricket team. The positions are - opener, top order, middle order, lower middle order and tail ender. Players who bat from position one to eight are either specialist batters or all-rounders with very good batting skills. Tailenders are bowlers with little or no batting skills. Each batting position requires a different kind of skillset. Bowlers are of two types – pacers and spinners. Pacers can be divided into fast bowlers and medium pacers based on their speeds. Spinners can be divided into off spinners and leg spinners based on the direction of spin. Generally, a cricket team consists of five batters, one wicketkeeper and five bowlers. Teams prefer to have one or two all-rounders. This adds depth to the batting and bowling. For instance, the composition of the Indian cricket team in a recent cricket match is shown in Table I [1]. The team is playing with five specialist batters, one wicketkeeper, three all-rounders and three specialist bowlers.

This effectively means that the team has eight very good batting options and six very good bowling options. The presence of three all-rounders gives the captain plenty of options in case a player has a bad day. Selecting a cricket team is a complex process. It depends on many factors like format of cricket, the past performance of a player, pitch type, weather conditions, fitness status and the player's record against that particular opposition. Selection meetings involve the opinions of a selection panel, the coach and the captain of the team. For example,

Grenze ID: 01.GIJET.9.1.320_1

TABLE I SAMPLE STRUCTURE OF A CRICKET TEAM

Sr No.	Player	Role	Position
1	Rohit Sharma	Batter	Opener, Top Order
2	Ishan Kishan	Batter	Opener, Top Order
3	Virat Kohli	Batter	Top Order
4	Rishabh Pant	Batter, Wicket keeper	Middle Order
5	Suryakumar Yadav	Batter, Wicket keeper	Middle Order
6	Venkatesh Iyer	All-rounder	Middle Order
7	Deepak Chahar	All-rounder	Lower Middle Order
8	Harshal Patel	All-rounder	Lower Middle Order
9	Bhuvaneshwar Kumar	Bowler	Tailender
10	Ravi Bishnoi	Bowler	Tailender
11	Yuzvendra Chahal	Bowler	Tailender

India's selection committee consists of five members, one from each zone. The five zones are North, South, East, West and Central. One out of the five committee member serves as the chairman of the selection committee [2]. Selection meetings are rarely unanimous. Each decision maker might have a different viewpoint on selection which makes the selection of an optimal team even harder. Some decision makers may have biased views which comes in the way of picking the perfect team. This traditional method of team selection has its own short comings. To counter these challenges, there is a need to adopt a more scientific and data-driven approach to team selection. Machine learning algorithms can be meaningfully deployed for cricket team selection.

Sports including baseball and basketball have made use of predictive modelling of historical data to analyze and take optimal decisions. Sabermetrics in baseball is one such example [3]. Cricket has also embraced the idea of using data-driven analytics to make optimal decisions. The support staff of a modern cricket team is incomplete without a video analyst and a performance analyst [4]. These analysts are tasked with making sense of historical cricket data and helping the coach and captain in formulating strategies to win a match. One such example is Pakistan Cricket Team's three-year agreement with CricHQ and CricViz. [5] CricViz is one of the leading companies in cricket analytics. Inputs provided by Cricviz's analysts was one of the key factors in Pakistan's first ever victory over India in a world cup match [6]. Developing machine learning algorithms for cricket team selection has its own set of challenges. Team selection in cricket depends on many factors. A discussion of these factors in presented below.

Cricket is unique as a sport because three different formats are played at the international level – Test cricket, One day internationals (ODIs) and Twenty20 Internationals (T20Is). Test cricket is the oldest among the three formats. It is played over five days. Both teams have to play two innings each. Teams have to bowl 90 overs in a day. ODIs are faster. One day matches typically last for about 8 hours and teams play 50 overs each. T20I is the youngest and fastest form of the game where each team has to play 20 overs [7]. Each format of cricket demands different skills from the players. Figure 1 compares the three formats based on statistics obtained from www.espncricinfo.com [8]. Batting average refers to the number of runs scored per dismissal by a batter. Batting Run rate refers to the runs scored per over by a batter. Batting Strike Rate refers to the number of runs scored in hundred balls by a batter. Bowling Average refers to the number of runs conceded by a bowler per wicket. Bowling strike rate refers to the number of balls required for a bowler to pick one wicket. Economy refers to the runs conceded per over by a bowler. High run rates of batters in T20Is shows batters take the most risks in this format. Hence, batters with good attacking skills will succeed in T20Is and ODIs but good defensive skills are a must in tests. Conversely, bowlers have to be attacking in tests and largely defensive in white-ball cricket. High strike rates of bowlers in tests shows that the bowlers have the liberty to set up a batter by bowling long spells. Batters have to defend the good balls and wait for a bad ball to arrive. The past decade has seen the emergence of specialists. There has been a gradual specialization within cricket formats (Figure 2). A player proficient in tests may not be successful in T20Is. A machine learning model built for selecting a test team may not work for the other two formats. The extant literature seems to focus on developing a solution for one particular format of the game. Hence, **the type of format** plays a crucial role in team selection. Secondly, **pitches and weather conditions** also dictate the team combination. The nature of a cricket pitch depends natural factors on the type of soil, amount of grass and also the prevailing weather conditions [10] [11]. Rain interruptions can lead to a damp pitch. Some pitches favour the batters while some favour the bowlers. This is in contrast to other team sports like baseball and football. In these sports the playing area is constant and doesn't influence player performance. Table II contains a broad summary of different pitches in cricket. Another factor, especially in white-ball cricket, is the **opposition** or **match-ups**. Teams calculate the effectiveness of a bowler or a batter against all types of opponents and choose their team accordingly. This kind of selection where a player is selected because he/she tends to perform better against a specific kind of opponent is more prevalent in Twenty20 cricket. In Twenty20 cricket teams treat every moment as a game-changing one and hence look to squeeze out the tiniest advantage [12]. It is very common to bring on an off spinner against a left handed batter and leg spinner against a right handed batter.

TABLE II TYPES OF PITCHES

Pitch Type	Description
Flat	Batting friendly pitch. No movement for pacers and spinners.
Green	Suited for pace bowlers.
Dry	Suited for spinners.
Hard	Ball comes on to the bat. Seam movement for pacers and bounce for spinners.
Wet	Ball comes slower off the pitch. Not conducive for stroke play.

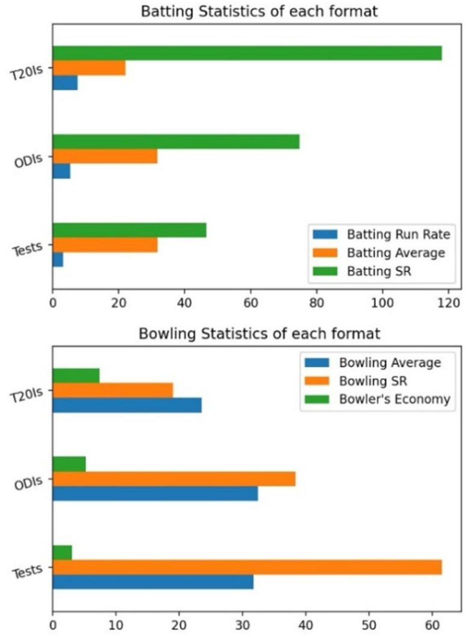

Figure 1 Batting and Bowling statistics for each format in the last 10 years [8]

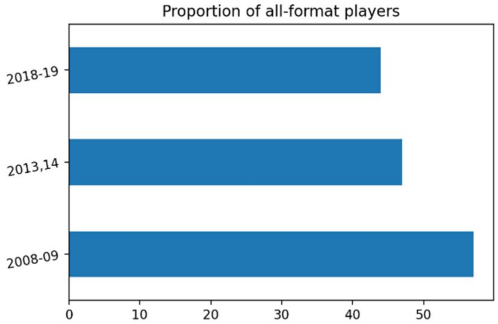

Figure 2 The proportion of all format players in international cricket has decreased from 57 in 2008-09 to 44 in 2018-19. [9]

A recent example where the opposition influenced the team selection is the IPL 2020 final. Mumbai Indians, who eventually went on to win the IPL, decided to drop their lead spinner Rahul Chahar in favour of Jayant Yadav. Yadav was selected because he is an off-spinner. The opposition had four left-handers in its batting line-up and he went on to dismiss the leading run-scorer of the opposition. Yadav only played two matches in the whole tournament, both against the same opposition [13]. Teams are countering match-ups by using left-right batting combinations. This helps them in scoring against both bowling types. Figure 3 confirms the steady rise in adoption of this strategy. Lastly, the **past performance** of a player is the most important factor. Historical cricket data is the backbone of every cricket team selection model. Ball by ball data of cricket matches is publicly available on the internet [15] [16]. In addition to this data, common cricket statistics are also available on these websites. This data is used to build machine learning models for cricket team selection. Players who have performed well in lead-up to the tournament are likely to continue their good form in the coming matches. Players who haven’t been performing well may need a break and analyse their shortcomings. However, form is something which is very hard to predict. A player who did not play well in the previous series might go on to perform very well in the next and vice versa. Such complex predictions can be simplified with the help of machine learning models.

Predictive algorithms and thorough analysis of historical data can help a sports team in making the right decisions. These algorithms help us in objective analysis and eliminate biases. Various researchers have developed algorithms which can help a cricket team in selecting the right set of players. The objective of this work is to present a survey of these algorithms. There is a gap in literature with reference to a survey on this topic. Section II contains the actual literature survey. In the first sub section, we explore how machine learning algorithms are used to predict an optimal team in sports other than cricket. Second sub section reviews different methods for selecting optimal cricket teams. We discuss the research trends and challenges in Section III. Section IV contains a summary and suggestions on future directions.

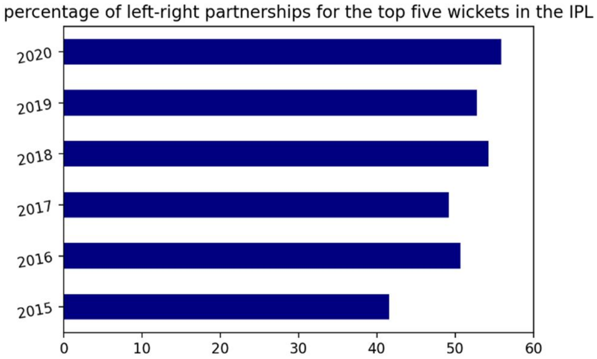

Figure 3 The percentage of right-left combinations for the top five wickets in the Indian Premier League shot up from 41.6% in 2015 to 55.9% in 2020. Data obtained from [14]

II. Literature Survey

This section surveys the existing literature on team selection using machine learning algorithms. The section is divided into two parts. Part one reviews algorithms proposed for sports like baseball and football. Part two contains the main survey – review of methods used for cricket team selection.

A. Other sports

Data driven analytics in sports has been rapidly growing in the past decade, mainly in American sports leagues and European football leagues. This section gives a brief overview of existing machine learning approaches for selecting an optimal team in other team sports like baseball and football. Baseball is a game that is perfectly suited for statistical analysis because it consists of discrete events like pitches. Baseball teams are famous for following a sabermetric approach in attempt to win championships [3]. One such example is Oakland Athletics, a team which finished first in the American League West despite the departure of three key players [17]. A key factor in Oakland Athletics' win was the selection of the pitcher. One of the most important decision to be taken in a baseball game is predicting entry point of a starting pitcher in a pitching sequence. This problem is treated as a binary classification problem ("High Performance / "Low Performance"). A predictive model for the same was developed by César, Mabel & Irvin (2017) [18]. Here, the authors used data of sequential baseball pitches as time series data. This data contains a detailed description of every pitch thrown during that particular baseball game [19]. K-NN algorithm was used for classification and performance prediction of the pitcher. The authors conducted experiments for various lengths of time series. It was observed that the accuracy of the model increased with the length of the time series. This model can be improved further by testing it on various different classification methods and tuning the hyper parameters for the same. Nevertheless, the predictive approach followed in this paper can be used for team selection in other sports, like cricket [20].

A sport like football is extremely hard to analyze due to the difficulty in deriving meaningful insights from the statistics available and the continuous nature of play involved. However, researchers have tried to develop algorithms mainly due to the availability of various public datasets. A football club's main aim to is to win the important tournaments. A team may contain brilliant players but collective performance is what matters the most. Uzochukwu C and Enyindah (2015) developed a tool using neural networks for football team selection [21]. They suggest that the four major attributes needed for a football player are "Technique", "Speed", "Physical Status" and "Resistance". Four Feed Forward Neural networks are used to calculate each of these attributes for each player. Data was collected from Player Stats Database for the popular Pro Evolution Soccer series video game [22]. Features from this dataset were used as the weights of the neural network. For example, "Response", "Explosive Power", "Dribble Speed" and "Top Speed" are used as inputs to the feed forward network used to calculate "Speed" of the player. Optimal weights were found using the feed forward algorithm. After obtaining the values of these four attributes, the average status for each player was calculated using the below formula.

$$Average\ stat\ (X) = \ (Physical\ status + technique + speed + resistance)/4 \qquad (1)$$

A player is considered as above average if the average stat is greater than 50, average if it is 50 and below average if the stat is less than 50. Above average players are selected in the team. Average players are used as substitutes. The authors also compared the online value of the players with their obtained value. It can be concluded that neural networks are a useful tool for team selection. Values obtained for each attribute can also be used to select players by position. This can be done by categorizing the players by position. For example, physical status and resistance are important attributes for a defender or a defensive midfielder. Players who rank higher in these attributes can be considered for the said positions. One drawback of this method is that it does not consider factors such as team chemistry and cohesion between teammates. Al-Shboul et al. (2017) improved on this aspect and proposed a semi-supervised learning approach [23]. They also focus on selecting a team for a specific opponent by studying the individual match-ups between players. 10 years of data from the English Premier League was used for this purpose [24]. This data was split into four categories – goalkeeper, defender, midfielder and striker. To calculate player ratings, a neural network with 11 input neurons was used. A neural network with 22 neurons was used for team predictions. By using 22 neurons, the strengths and weaknesses of the opposition are also taken into account. The first neural network gave an accuracy of 54% and the second neural network had an accuracy of 60%. Individual player ratings can be used in situations where like for like replacements for a player a required. In club football, the owners of a football club also give importance to marketing and ticket sales to remain profitable [25]. P Rajesh, Mansoor and Mansour proposed a data science approach to optimize the time taken in selecting a player for a football club [26]. The problem of team selection is viewed from two perspectives – selecting a team that can win and selecting players who have a good commercial value. FIFA 19 dataset from Kaggle [27] consisting of data for approximately 18000 players was used. K-means clustering algorithm was used to group the players based on their positions. For each cluster obtained, various classification algorithms were implemented to classify the players based on their skills. Furthermore, a correlation matrix of all the attributes was created to assign a suitable position to the player. The proposed methodology had an accuracy of 78.52%. The proposed method achieves the goal of selecting a dream team of commercially viable players. The same method can also be extended for cricket tournaments like the Indian Premier League where there is a price constraint while building a team. However, the authors could have shown some examples of dream teams predicted by their model and compared it with the current best teams in football. The algorithms and techniques discussed in this section are summarized in table VI. Current literature mainly consists of techniques like neural networks, classification and clustering for team selection in sports like baseball and football. These techniques can also be extended to predict optimal teams in cricket. The next section contains a review of methods used for selecting cricket team.

B. Cricket

This section contains a review of methods used for selecting a cricket team. As mentioned earlier, team selection in cricket is slightly more complicated due to presence of 3 formats which require different sets of skills. As of now, methods for selecting an optimal test team do not exist in current literature. Hence, our review focuses on the other two formats of cricket – ODIs and T20s. As mentioned earlier, one-day cricket comprises of a single day match. Each team plays 50 overs each. Iyer and Sharda (2009) made one of the first attempts to use neural networks for cricket team selection. Six different types of neural networks were used for the same [20]. Training data for this model was collected from two websites - espncricinfo [15] and cricket archive [28]. Only players who have been active from 1985 onwards were considered. Data from 2007 One-day cricket world cup was used as testing data. The authors conducted two experiments. In the first experiment. the obtained dataset was divided into two parts – bowlers and batters. In the second experiment, the batter and bowler datasets were further categorized to account for recent form of cricketers. One category comprised of player data from 1985 to 2002. The other category contains data from 2003-04 season to 2006-07 season. For both experiments, neural networks were trained to generate player ratings. Players were classified into 3 categories – "performer", "moderate" and "failure". Table III explains the rules used for player classification. After classifying the players into the above mentioned categories, players were selected in the team based on the following heuristics:

1. Batter is selected if he receives the following ratings:
    a. At least two "performer'', at least 1 "moderate" and no "failure"
    b. All "moderate" and no "failure"
2. Bowler is selected if he receives the following ratings:
    a. At least 2 "performer" and at least 1 "moderate"
    b. At least 1 "performer" and 4 "moderate"
    c. All "moderate" and no "failure"

Out of the five neural networks, a multi-layer perceptron with 41 input neurons and 20 hidden neurons gave the best accuracy of 86% for batters. For bowlers, a multi-layer perceptron with 40 input neurons and 21 hidden neurons gave the best accuracy of 80%. The accuracy of this study can be further improved by including fielding data and wicket-keeping data. Passi and Pandey (2018) treated cricket team selection as a classification problem.

TABLE III RULES FOR PLAYER CLASSIFICATION

Player Type	Criteria	Performer	Moderate	Failure
Batter	Runs Scored across ODIs and tests	More than or equal to 4000	More or equal to than 800 but less than 4000	Less than 800
	Matches Played across ODIs and tests	Minimum 50	More than 40 but less than 50	Less than 40
Bowler	Balls bowled across ODIs and tests	More than 7000	More than 1000 but less than 7000	Less than 1000 balls
	Wickets taken across ODIs and tests	More than and equal to 150	More than or equal to 40 but less than 150	Less than 40
	Wickets taken in each form (ODIs and Tests)	Minimum 30	Minimum 20	Less than 20

They tried to predict the number of runs scored and wickets taken into different ranges [29]. The algorithms used are naïve bayes, random forest, multiclass SVM and decision tree classifiers. The uniqueness of this model lies in the derived attributes of the dataset. Data of matches played from 2nd January 2000 to 10th July 2017 was collected from espncricinfo [15]. Apart from containing traditional cricket statistics like runs scored, average, strike rate and so on; the dataset also contains a number of derived attributes which give a more detailed insight into a cricketer's abilities and performance. These attributes are consistency of the player, recent form of the player, player's performance against a certain team and the player's performance at a certain venue. The dataset also contains a feature called pressure which accounts for the pressure on the player depending on the type of the match. Knockout matches like quarter finals, semi-finals and finals are high pressure matches while matches in bilateral series and league stages are comparatively low pressure. The above mentioned classification algorithms were trained and tested on the collected data. For batters, Random Forest algorithm gave the best accuracy of 90.27% for an 80-20 split. For bowlers, Random forest algorithm gave the best accuracy of 92.25% for a 90-10 split. This model is extremely useful in selecting the best batter and bowler for a cricket match. However, a similar model can be developed for selecting all-rounders as well. Some more factors which affect the performance of a player such as overhead conditions and pitch type can also be included in the future. Shetty, Rane, Pandita et al. (2020) also used classification algorithms to select an optimal team for India [30]. Similar to the previous study, this study also treats team selection as a classification problem. The classification algorithms used are logistic regression, SVM and random forest classifier. The model uses the data of One Day Internationals played by the Indian cricket team in the last several years. Information about weather conditions is also included in the dataset. Three different datasets were created for batters, bowlers and all-rounders [30]. These datasets were again split into training and testing sets with an 80:20 ratio. The classification algorithms mentioned previously were applied on the test dataset. Random forest algorithm gave the best result out of the three algorithms. It gave an accuracy of 76%, 67% for bowlers and 95% for all-rounders. Another useful application of machine learning is early identification of potential superstars of the future. Such players may not succeed in the initial few matches. However, they can be selected for a longer time because of the possibility of the player playing at a higher level in the future. For team selection, it is important to identify these players and make sure that they receive the right support and resources to maximize their potential. Such selection policies ensure a smooth transition between two generations of players in a team. Khot, Shinde and Magdum (2020) used support vector machine algorithm to predict rising stars in cricket [31]. Rising stars in each domain (batting, bowling, and all-rounder) were predicted. Data from [15] for the years 2006 to 2018 was collected. Players under the age of thirty who have played at least twenty matches are considered for the evaluation. The unique feature of this approach is that it defines a concept of co-players, i.e. a player is evaluated using both his performance and also the performances of other players he had played with. Performance of the team as well as the opposition is considered. For example, the feature set of batters considers features such as co-batters runs, co-batters average, co-batters strike rate, team average, team strike rate, team win loss ratio, opposition team average, opposition team strike rate and opposition team win loss ratio. RS score (Rising Star Score) is calculated for each domain. Feature analysis is performed on each feature set. The features which have a positive correlation are added to the RS score and the ones with a negative correlation are subtracted from the RS score. Finally, top 10 rising stars in each domain are predicted using SVM and the RS score. The results are compared with the highest ICC rankings of the predicted players. If the RS score and ICC ranking matches, we consider it as an accurate result. Using the said heuristic, the accuracy of the results is shown in Table IV. Having discussed the algorithms for one-day format, we now discuss it in the context of T20 cricket. Inception of T20 cricket has led to the growth of franchise based T20 Leagues. These leagues consist of players from different nationalities. Team are given a specific purse to spend while building a team. Team selection in these leagues is more complicated because of this added constraint of money Sinha, Das and Saha (2020) proposed a supervised machine learning technique that predicts the ranking of batters in the Indian Premier League based on the batter's past performances [32]. In this method, a combination of polynomial and linear regression is used. The dataset contains statistics from two sources – the orange cap data from 2008 to 2019 and ball by ball data of two editions of IPL [33]. Apart from the traditional cricket statistics, SR_R, ACVT and Boundary are derived parameters in the dataset. SR_R unifies the total impact of runs scored and strike rate of the batter. ACVT accounts for the effect of milestones like high score, 100s and 50s. Boundary integrates the effect of 4s and 6s hit by a batter. Ball by ball data of IPL 2019 season is chosen as the test dataset. Pair wise scatter plots of predictor and target are plotted, where predictor is the feature from the dataset and target is the rank of the batter. It is observed that most of the features have a polynomial relationship with rank. Features with acceptably high correlation factors are used for further analysis. Final ranking of the batter is determined by assigning certain weight factors to the outputs of linear and polynomial regression. Orange cap data, which has the list of the top run getters, was used as test data. Table V shows the comparison of original vs predicted rankings. It is clear that the rankings predicted by this method have a close resemblance to the actual rankings for the 2019 season. The highest difference between original and predicted is for Quinton De Kock. One possible explanation for this is that he was traded to a different team in the 2019 season and it took time for him to adjust to a new team environment. Machine learning models find it difficult to quantify the effect of such factors. However, there is scope to build similar models for bowler and all-rounder selection, which will make the model more useful vis-à-vis complete team selection. Nevertheless, this model can be used for any other T20 tournament in the world.

TABLE IV RESULTS FOR RISING STARS

Player Type	Accuracy
Batting	60%
Bowling	70%
All-rounders	40%

TABLE V COMPARISON OF RANKINGS

Player Name	IPL 2019 Orange Cap Rank	Predicted Rank	Difference (Original – Predicted)
David Warner	1	3	-2
KL Rahul	2	1	1
Andre Russell	3	5	-2
Shikhar Dhawan	4	4	0
Virat Kohli	5	8	-3
Quinton De Kock	6	2	4
Chris Gayle	7	6	1
Jonny Bairstow	8	10	-2

The Indian Premier League (IPL) is one such league. IPL auctions are very intense where bidding wars can inflate the price of players and teams don't always end up getting targeted players. Hence, there is a need to develop methods which consider the price constraint for IPL team selection. Rani, Kamath, Menon et al. (2020) used neural networks and clustering algorithms to predict the next batter / bowler to be sent / used for a given match condition [34]. They also predicted an ideal IPL team using clustering algorithms. The algorithm predicts an ideal IPL team that can be formed in the auctions, while taking into consideration the performance of the player over the past seasons and also the price of the player in the previous auction. The algorithms used for prediction are K-means algorithm and Hierarchical clustering. The players are clustered based on their performance, which is why players of a similar level/similar quality fall into the same cluster. In an ideal world, any team owner would like to buy the players from the best clusters. This is not always possible as the best players are available at a premium price and the price may inflate due to bidding wars between the teams. However, teams can strike the right amount of balance by trying to bid for players from the best cluster more often compared to players in the lower cluster. For team selection, the authors got a result which satisfied the winning combination of Mumbai Indians in 2013. To predict the ideal batter or bowler to be used during the match, a neural network was trained on the data of the previous five years. The neural network had three hidden layers and one output layer. This model is one of them best models when it comes to predicting an ideal team for the IPL. The model takes into account all three types of players – i.e. batters, bowlers and all-rounders and gives accurate results. In addition, the authors have also proposed a dynamic model which can help in decision making during an in-progress cricket match. A limitation of the model is that it relies on the price at which the player was bought in the previous auction. The price of a player may not be a true indicator of a player's quality. There are many factors which affect the price at which a player is bought in the auction. The popularity of the player is a key factor. Furthermore, the popularity varies across states and regions. The matter is complicated by recent performances of players. Teams tend to give more weightage to the recent performance before an IPL auction, often underplaying the consistency of a performance for a long period of time. This is possibly because of the peculiar nature of the T20 format. However, it would be interesting to account for these factors in further research.

III. DISCUSSIONS

This paper has reviewed the state-of-the-art methods and a summary of the approaches is provided in Table VI. It is noticeable that team selection is generally treated as a classification or a regression problem. When treated as a classification problem, players are generally classified into different classes where each class has a priority. Regression has been mainly used to rank players. One can observe from Table VI that all the methods mentioned in this paper focus on the shorter formats of cricket. One possible explanation for this is that it is easier to develop models for one day internationals and T20 cricket. This is because attributes like average, strike rate, boundaries, economy rate directly influence the result of the match and the quality of the player. Another promising line of research would be to build similar models for test cricket as well. Another bottleneck that cricket faces is the unavailability of ball-tracking data [35]. Baseball has highly benefitted from the availability of public data. Ball tracking data contains a lot of useful information like the speed, velocity, trajectory, line, length etc. of every delivery. This, combined with ball by ball data which is already available in public domain, can help in development of more robust and accurate algorithms.

TABLE VI SUMMARY

Sr. No	Author and Date	Sport	Objective	Technique	Key Features	Remarks
1	César Soto-Valero , Mabel González-Castellanos & Irvin Pérez-Morales (2017)	Baseball	Pitcher selection – analyze the performance of the pitcher and predict when the next player needs to enter the pitching sequence	- Time series classification using KNN algorithm. - Supervised learning algorithm	- A classification model is developed by using pitch by pitch data as time series data. - Dynamic Time Warping is used as a distance measure for KNN - Lower bound of Keogh is used to speed up Dynamic Time warping	.- Accuracy of 0.906009 was obtained when number of throws left were 5. - Accuracy of 0.701079 was obtained when number of throws left were 50. - Accuracy increases as number of throws left decrease.
2	Onwuachu Uzochukwu C and P. Enyindah (2015)	Football	Select an ideal team for a football club	Neural networks	- Neural network analyses a player's speed, technique, physical status and resistance - Average score is calculated from above four attributes - A player is selected if average score is greater than 50	Scores of players who are categorized as "above average" by the model matches with the online scores of the player
3	Rabah Al-Shboul, Tahir Syed, Jamshed Memon and Furqan Khan (2017)	Football	Select the best possible team for a football match when the opposition lineup is known. Predict the probability of victory	Feed forward neural networks	Two experiments are conducted. First experiment has 11 input neurons and second experiment has 22 input neurons. Opposition team is considered in the second experiment.	Experiment 1 (Opposition not considered) - 54.016% Experiment 2 (Opposition considered) - 60.741%

4	Dr P Rajesh Bharadwaj, Dr Mansoor Alam & Dr Mansour Tahernezhadi (2020)	Football	Form a dream team which is capable of winning football tournaments. The players in the team are also profitable for the club.	- K means clustering - Classification techniques - Naïve Bayes, Random Forest, Decision Tree, Support Vector Machine (SVM)	- Presents a novel pseudocode algorithm consisting of clustering and classification to select players - Correlation matrix to select a player's best position	Pseudocode – 78.52% Among the traditional classification methods Random forest had the best accuracy of 83.25%
5	Iyer and Sharda (2009)	Cricket (One Day Cricket)	Predict individual cricketer's performance and therefore select an optimal team for 2007 world cup	Neural Networks – Multilayer perceptron (MLP), linear, radial basis function (RBF) Two types of MLP and RBF were used	Classifies players into 3 classes – performer, moderate and failure	Neural networks were able to correctly identify 70% of the players who performed well in the world cup. Multi-layer perceptron with 40 input neurons and 21 hidden neurons had the best accuracy.
6	Passi and Pandey (2018)	Cricket (One day cricket)	Predict performance of players and select the ones who are predicted to perform better than the rest	Classification algorithms – Naïve Bayes, Random Forest, Multiclass SVM, Decision Trees	Additional features like consistency, form, venue and opposition are considered. Results of 4 different train-test splits are compared: 60-40 70-30 80-20 90-10	Random forest algorithm performed the best and Naïve Bayes performed the worst Random forest 90-10 split had the best result among the different split ratios: Batters – 90.74% Bowlers – 92.25%
7	Shetty, Rane, Pandita et al. (2020)	Cricket (One day cricket)	Select the best playing 11 for Indian cricket team in one day internationals.	Classification algorithms – Logistic regression, Support Vector Machine (SVM), Random Forest Algorithm.	Extra features like weather, pitch, opponent, home/away records are considered.	Batters – 76% Bowlers – 67% All-Rounder – 95% Random Forest Algorithm gave the best results.
8	Amruta Khot, Aditi Shinde and Anmol Magdum (2020)	Cricket (One day Cricket)	Prediction of rising star in cricket for better team selection.	Support Vector Machine (SVM)	A player is classified either as a rising star or not rising star. Introduces the idea of co-players	Batting – 60% Bowling – 70% All-rounders – 40%
9	Arnab Santra, Pritilata Saha, Abhirup Sinha and Amit Kumar Das (2020)	Cricket (T20 Cricket)	Rank the quality of a batter based on the batter's past profile	Regression – linear and polynomial regression	Derived parameters SR_R, ACVT and Boundary Compares results for original parameters vs derived parameters	At Tolerance level <= 10: Linear Regression Original parameters: 88% Derived parameters: 90% Polynomial Regression Original parameters: 89% Derived Parameters: 91%

IV. CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

Analysis of historical data has become a part and parcel of the sports ecosystem. A sport like baseball is designed for meticulous data analysis due to its discrete structure. Various statistics in baseball have a direct correlation with the outcome of the match and the performance of the players. In addition, the availability of comprehensive data has ensured that statistical analysis continues to influence the sport in a major way. A sport like football is extremely hard to analyze due to its fluid nature. It is hard to break down 90 minutes of a football match into smaller units. Hence, it becomes harder to derive meaningful information from football statistics.

Baseball and cricket has pitches and innings as discrete units. Within these units, there are further smaller events such as balls [36]. However, a few variables like weather conditions, nature of pitches and format of cricket separate cricket from baseball. These variables make it a bit harder for algorithms in cricket team selection to be perfect. Hence, there is always a room for improvement in algorithms for cricket team selection.

REFERENCES

[1] ESPNCricinfo, "India vs West Indies, 1st T20I," 2022. [Online]. Available: https://www.espncricinfo.com/series/west-indies-in-india-2021-22-1278661/india-vs-west-indies-1st-t20i-1278679/full-scorecard. [Accessed 10 March 2022].

[2] "Indian National Cricket Team Selectors," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/India_national_cricket_team_selectors. [Accessed 28 February 2022].

[3] J. M. B. James Albert, Curve Ball: Baseball,Statistics, and the Role of Chance in the Game, Springer, 2001.

[4] S. Giridhar and V. Raghunath, "Running the numbers: How data and analytics influence cricket," Times of India, 2018. [Online]. Available: https://timesofindia.indiatimes.com/sports/cricket/news/running-the-numbers-how-data-and-analytics-influence-cricket/articleshow/64252696.cms.

[5] Pakistan Cricket Board, "PCB signs three-year agreement with CricHQ and CricViz," 2021. [Online]. Available: https://www.pcb.com.pk/press-release-detail/pcb-signs-three-year-agreement-with-crichq-and-cricviz.html. [Accessed 28 February 2022].

[6] CricViz, "CricViz's Role In Pakistan's T20 World Cup Win V India," 2021. [Online]. Available: https://www.cricviz.com/cricvizs-role-in-pakistans-t20-world-cup-win-v-india/. [Accessed 03 March 2022].

[7] ICC, "The Three Formats of Cricket," [Online]. Available: https://www.icc-cricket.com/about/cricket/game-formats/the-three-formats. [Accessed 27 July 2021].

[8] "Espncricinfo Statsguru," [Online]. Available: https://stats.espncricinfo.com/ci/engine/stats/index.html. [Accessed 28 02 2022].

[9] "Will all-format players become a thing of the past?," Espncricinfo, 31 May 2020. [Online]. Available: https://www.thecricketmonthly.com/story/1223774/will-all-format-players-become-a-thing-of-the-past. [Accessed 03 March 2022].

[10] S. Singh, "Cricket Pitches – Science behind the Art of Pitch," International Journal of Science and Research (IJSR) , 2014.

[11] Swetha and Saravanan.KN, "Analysis on Attributes Deciding Cricket Winning," International Research Journal of Engineering and Technology (IRJET), vol. 04, no. 03, 2017.

[12] H. Ganjoo, "Do match-ups work in T20? The data says yes," ESPNCricinfo, 17 April 2021. [Online]. Available: https://www.espncricinfo.com/story/ipl-2021-do-match-ups-work-in-t20-the-data-says-yes-1258833.

[13] "Rohit Sharma: 'We said at the start we want to make winning a habit'," ESPNCricinfo, 11 November 2020. [Online]. Available: https://www.espncricinfo.com/story/dc-vs-mi-ipl-2020-final-rohit-sharma-says-we-said-at-the-start-we-want-to-make-winning-a-habit-1238984. [Accessed 28 February 2022].

[14] "Ten ways T20 has changed since the last World Cup," Espncricinfo, 14 October 2021. [Online]. Available: https://www.thecricketmonthly.com/story/1282046/t20-world-cup-2021---ten-ways-t20-has-changed-since-the-last-world-cup. [Accessed 28 February 2022].

[15] "Espncricinfo," [Online]. Available: www.espncricinfo.com. [Accessed 28 02 2022].

[16] "CricSheet," [Online]. Available: https://cricsheet.org/. [Accessed 10 March 2022].

[17] M. Lewis, Moneyball : the Art of Winning an Unfair Game, W. W. Norton & Company, 2003.

[18] M. G.-C. &. I. P.-M. César Soto-Valero, "A predictive model for analysing the starting pitchers' performance using time series classification methods," International Journal of Performance Analysis in Sport, 2017.

[19] "PITCH f/x," [Online]. Available: https://www.brooksbaseball.net/pfxVB/pfx.php. [Accessed 03 March 2022].

[20] S. R. Iyer and R. Sharda, "Prediction of athletes performance using neural networks : An application in cricket team selection," Expert Systems with Applications, 2009.

[21] O. U. C and P. Enyindah, "A Machine Learning Application for Football Players' Selection," International Journal of Engineering Research and Technology, vol. 4, no. 10, 2015.

[22] "PES Stats Database," [Online]. Available: https://pesstatsdatabase.com/PSD/playerClassic.php?id=157. [Accessed 03 March 2022].

[23] R. Al-Shboul, T. Syed, J. Memon and F. Khan, "Automated Player Selection for a Sports Team using Competitive Neural Networks," International Journal of Advanced Computer Science and Applications, vol. 8, no. 8, 2017.

[24] "Premier League Player Stats," [Online]. Available: https://www.premierleague.com/stats/top/players/goals?se=418. [Accessed 15 March 2022].

[25] Y. Yilancioglu, "The business of football — breaking down the finances of a football club," 13 April 2021. [Online]. Available: https://medium.datadriveninvestor.com/the-business-of-football-breaking-down-the-finances-of-a-football-club-8263614059d8.

[26] P. Rajesh, Bharadwaj, M. Alam and M. Tahernezhadi, "A Data Science Approach to Football Team Player Selection," in IEEE International Conference on Electro Information Technology (EIT), 2020.

[27] K. Gadiya, "FIFA 19 Complete dataset," Kaggle, [Online]. Available: https://www.kaggle.com/karangadiya/fifa19. [Accessed 15 March 2022].

[28] "Cricket Archive," [Online]. Available: https://www.cricketarchive.com/. [Accessed 10 March 2022].

[29] N. P. Kalpdrum Passi, "Predicting Players' Performance in One Day International Cricket Matches Using Machine Learning," in 8th International Conference on Computer Science, Engineering and Applications, 2018.

[30] M. Shetty, S. Rane, C. Pandita and S. Salvi, "Machine learning-based Selection of Optimal sports Team based on the Players Performance," in 2020 5th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, 2020.

[31] A. S. a. A. M. Amruta Khot, "Rising Star Evaluation using Statistical Analysis in Cricket," in International Conference on Advances in Distributed Computing and Machine Learning, 2020.

[32] S. P. M. P. S. Amlan Ghosh, "A Novel Regression based Technique for Batsman Evaluation in the Indian Premier League," in IEEE 1st International Conference for Convergence in Engineering (ICCE), 2020.

[33] "IPL Statistics," [Online]. Available: https://www.iplt20.com/stats. [Accessed 20 March 2022].

[34] P. J. Rani, A. V. Kamath, A. Menon, P. Dhatwalia, D. Rishabh and A. Kulkarni, "Selection of Players and Team for an Indian Premier League Cricket Match Using Ensembles of Classifiers," in 2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), Bangalore, 2020 .

[35] Ponsonby, "Why cricket is waiting for its data revolution," [Online]. Available: https://www.cricbuzz.com/cricket-news/121227/why-cricket-is-still-waiting-for-its-data-revolution.

[36] Jones and N. Laemon, Hitting against the spin : How Cricket Really Works, 2020.