WINNER PREDICTION IN ONE DAY INTERNATIONAL CRICKET MATCHES USING MACHINE LEARNING FRAMEWORK: AN ENSEMBLE APPROACH

*Manoj Ishi

*Research Scholar, M. Tech (Computer Science and Engineering),
Department of Computer Engineering, R. C. Institute of Technology, Shirpur, India
ishimanoj41@gmail.com

Dr. Jayantrao Patil

Professor, Ph. D. (Computer Engineering), Department of Computer Engineering, R. C. Institute of Technology, Shirpur, India
jbpatil@hotmail.com

Dr. Nitin Patil

Associate Professor, Ph. D. (Computer Engineering), Department of Computer Engineering, R. C. Institute of Technology, Shirpur, India
er.nitinpatil@gmail.com

Dr. Vaishali Patil

Professor, Ph. D. (Computer Engineering), RCPET's Institute of Management Research and Development, Shirpur, India
vaishali.imrd@gmail.com

Abstract

Forecasting output of any sports match is in massive demand for the sports industry. Avoiding defeat is the eventual aim of any sports game. Cricket is one of the well-known sports in the world. Cricket is a sport of uncertainty; the state of match gets changed on each ball of the game. Due to that, winner prediction in cricket is becoming a challenging task. In this paper, the focus is given to the prediction of victory in one day international game of cricket using machine learning. In this work, 128 features are considered for the implementation. With these feature sets, three models are proposed based on batting-bowling strength of the team, run-scoring pattern for the team, and overall team strength. The concept of ensemble algorithm using voting and stacking classifier is used for prediction with machine learning algorithms. The feature selection methods are used for this work to remove irrelevant or redundant features. The investigation of the prediction model is performed using accuracy, precision, F1- score, recall value. The Logistic Regression and Support Vector Machine give better results than other models with an accuracy of 96.30% for predicting the winner of the ODI match.

***Keywords*: Winner prediction; Machine learning; Feature selection; Discriminant analysis; Ensemble algorithm; Voting Classifier; Stacking Classifier.**

1. Introduction

Sports forecasting is a recent growing area of interest involving high predictive accuracy. Sports analytics is a technique used to explore essential data and investigate past data to get fundamental knowledge for proper decision-making. The extraction of comprehensive statistical information in sport has allowed it to apply data mining (DM) techniques for manufacturing predictive models on large datasets [1-3]. Cricket is a top-rated game. Cricket is reasonably significant within the statistical science community, but this game's unpredictable and unstable nature makes it more challenging to use in standard probability models [1,4,5]. Uncertainty is the feature of cricket, and it increases as the game becomes shorter. Mainly the shortest format takes maximum complexity, where a single over will change the game's momentum. The suspense generated during the game attracts a vast number of spectators to watch cricket. With millions watching cricket, building a model for forecasting the outcome of cricket matches is a real-world task. Sports Analytics is used to determine a cricket match outcome, whether the game is in progress or well before the game begins [3,4].

Across the three formats of cricket, one day international is one of the popular formats. The popularity increased due to the day night format of the game, umpire decision review system, strict bowling rules, and concept of the powerplay. As the popularity of one-day international (ODI) games increases, it is crucial to understand the potential predictors of the game's result. In this work, the winner prediction is made before the start of the match. A match outcome depends greatly on each player's skill and their success on a given day. However, the game's outcome is also determined by external parameters like toss, home field advantage, pitch, etc., and numerous environmental parameters [6-8].

Machine learning is used for designing a winner prediction model using historical data. The basis of machine learning is computational statistics for making predictions. As an outcome, having both live and noteworthy data makes machine learning widespread in sports analytics [9-10]. The main goal is to define the key factors that influence the game’s outcome and pick the best machine learning model that matches this data with the best results. It is essential to keep in mind that all the primary factors that may influence the match outcome need to be chosen to design the best model [1,11,12].

The outcome prediction of a cricket match is helpful for team management, coach, and captain of the team. Coaches and sports managers need models that take external and internal variables into account and provide them with vital analytics to assess their team and opposition team to formulate the tactics. These prediction models are also helpful for sports analysis shows broadcasted on sports and news channels. The sports analysis show is broadcasted before the match, during innings break, and after the completion of match for the data analysis. They try to predict the possible result of cricket matches using historical information [1,13,14].

Classification based prediction model is designed using independent features to forecast the result of cricket matches. This research aims to design a simple effective model based on machine learning to predict the result of cricket matches using significant features [15-16].

The significant contributions of this research work are:

- Three algorithms are proposed based on the features defining batting and bowling strength of the team, a run-scoring pattern for the team, and overall team strength.
- The first algorithm is used to evaluate the strength of batsmen using a weighted combination of features like a milestone reached, batting average, strike rate, etc.
- In the second algorithm bowling strength of the team is measured with features like bowling strike rate, bowling average, number of 5 wickets haul, etc.
- The third algorithm is used to reflect the overall strength of the team. In this algorithm, both the first and second algorithm output is provided as input with proper weight. This algorithm uses pressure index, winning consistency, run-scoring pattern, and wicket loss pattern as input parameters for the different game phases.
- Based on the above algorithms, three models are proposed to predict the winner of an ODI match.
- The first model predicts the winner of the match using the batting and bowling strength of the team calculated from the first two algorithms.
- In the second model, the run-scoring pattern of the team is studied phase-wise to forecast the output of cricket matches. The cricket match is divided into three phases:
  - First phase of overs 1 to 10
  - Middle overs from 20 to 40 as phase 2
  - The final phase of the last ten overs i.e., 41-50
- The third model is a combination of the first two models. It predicts the match winner using all the features from the first two models consisting of batting strength, bowling strength, and run-scoring pattern with extra features reflecting the team's winning consistency and pressure index.
- For all three models, the accuracy is enhanced using voting and stacking classifiers as ensemble techniques to combine results from more than one machine learning model having maximum accuracy.

This paper is further organized as follows. Literature review in section 2, section 3 provides information about the dataset and proposed methodology. Feature construction and model formulation are part of section 4. Section 5 conveys the result. Section 6 is related to discussion. Section 7 summarizes conclusion and future scope of work.

2. Related Work

Several researchers have worked on winner prediction in cricket. A Few of them is highlighted below.

Kumash Kapadia et al. have applied Naïve Bayes, Random Forest, K-Nearest Neighbors, and Model tree algorithms for winner prediction in the IPL T20 tournament [17]. The significant features from available data are found with feature selection techniques like Correlation, Information Gain, Relief, and Wrapper method. In the home-based feature model, the naïve Bayes has maximum accuracy of 57%. KNN algorithm achieved maximum accuracy of 62% for the toss-based model.

Dibyojyoti Bhattacharjee and Priyanka Talukdar have used the perception of pressure index with predictive discriminant analysis to predict twenty 20 matches results [18]. During the second inning of the game, the pressure index is calculated using the required run rate and wicket lost. The discriminate analysis-based prediction was made with an accuracy of 90.7%.

Wei Gu et al. have predicted the result of the National Hockey League with Principal Component Analysis, Nonparametric Statistical Analysis, and Support Vector Machine [19]. The ensemble method improves the classifier's result with an accuracy of more than 90% for SVM.

Sandesh Bananki Jayanth et al. have proposed a supervised learning model SVM with linear, poly, and RBF kernel for winner prediction in cricket based on team composition [20]. The SVM model with a nonlinear RBF kernel outperforms linear and polynomial kernel with an accuracy of 75%.

M. Asif and I.G. McHale have designed the second innings remaining run prediction model using a generalized nonlinear forecasting model (GNLM) [21]. The function of remaining overs and wickets lost is used for this forecasting model.

Stylianos Kampakis and William Thomas have investigated the prediction of results for twenty over English county cricket cup with team and player features [22]. The feature selection is performed with the Pearson Correlation, Chi-Square test, Principal Component Analysis, and Recursive Feature Elimination technique. These selected features are provided for Logistic Regression, Naive Bayes, Random Forest, and Gradient Boosting Decision tree algorithm. They finally conclude that the Naïve Bayes algorithm is better with an accuracy of 64.5%.

Neeraj Pathak and Hardik Wadhwa predicted the ODI match result using Naïve Bayesian, Support Vector Machine, and Random Forest algorithm [23]. The accuracy, sensitivity, and specificity value define SVM is better having an accuracy of 61.67%. If data is imbalanced, then Naïve Bayesian provides promising results.

Rabindra Lamsal and Ayesha Choudhary have used regression solutions of the multi-variate variables for IPL match result prediction [24]. The recursive feature selection method was used for selecting the relevant features from the given dataset. The MLP classifier is superior concerning the weighted mean accuracy of 71.66 %.

3. Methodology

Three models have been proposed, each with the input of three different feature sets. Input to the first model was features related to the batting and bowling strength of the team. The team's run-scoring pattern is input to the second model, and overall team strength-related features are input for the third model. These features work their way through to dynamically enhance and improve prediction accuracy with the benefit of the feature selection algorithm. Feature selection and dimensionality reduction algorithms reduce the number of input variables. The selected features from the feature selection algorithm are provided as input to nine classifiers: Logistic Regression, Naive Bayes, K Nearest Neighbors, Support Vector Machine, Gradient Boosting Algorithm, Decision Tree, Random Forest, XGBoost, and CatBoost algorithm. Voting and stacking classifiers are used to ensemble the result from more than one machine learning model. These two algorithms take the output of machine learning models as an input and provide the desired output with maximum accuracy. An artificial neural network is also implemented for predicting the winner of the match. It is designed with ReLU linear activation function. The sigmoid function reduces loss during data training and converts output between 0 and 1 [11]. The machine learning framework steps for predicting the result of a cricket match for this work are shown in Figure 1.

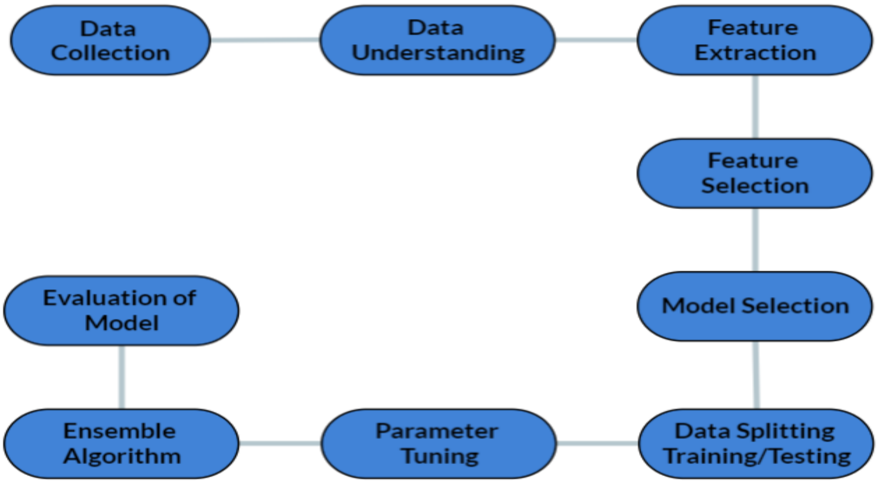

Fig. 1. Machine Learning Steps

3.1. *Data Collection and Understanding*

Sports data is obtained from freely accessible, accurate online sources. The data for this work is obtained from cricinfo.com [25]. It is a genuine website that includes all data for analysis from 1971. There is no benchmark dataset available for this work. The data of 1693 one-day international matches played from 2006 to 2019 is considered. The data is pre-processed to remove records of incomplete matches and results obtained with the Duckworth-Lewis method. Binary classification of class win/loss is performed for winner prediction. One hundred twenty-eight features are considered for this work. These features are classified into three sets:

- Batting/bowling strength-based feature
- Run-scoring pattern-based features
- Overall team strength-based feature.

3.2. *Feature Extraction*

The domain knowledge is required for extracting the necessary features. The features for predicting the results of cricket matches are primarily categorized into match and external features. The match features are having a direct impact on the result of the match. This set of features includes run scored and wicket lost during the game. The external features are toss, venue, time of play, and host country impacting cricket matches. Some external features are favoring the team during the cricket match. In this work, features like team statistics, time, toss, venue, city, runs scored and wickets loss in various game phases, team performance in past matches, etc., are extracted. The following Table 1 describes some essential features used for this work.

Feature Name	**Description**
run_pattern_team 1 and 2	weighted average of run scoring pattern for team 1 and 2 in each phase
wicket_pattern_team 1 and 2	weighted average of wicket fall pattern for team 1 and 2 in each phase
bat_strenth_team 1 and 2	weighted average of batting average and strike rate for team 1 and 2
bowl_strength_team 1 and 2	weighted average of bowling average and economy for team 1 and 2
no_30+score_team 1 and 2	summation of no of 100's, 50's and 30 plus run scored by team 1 and 2
milestone_team 1 and 2	weighted average of milestone reached by batsmen for team 1 and 2
home_runs_4s_6s_team1 and 2	runs scored in the forms of 4's and 6's by team 1 and 2
non_home_runs_1s_2s_3s_team1 and 2	runs scored in the forms of 1's, 2's, and 3's by team 1 and 2

no_30+partnership_team 1 and 2	summation of no of 100's, 50's and 30 plus partnership for team 1 and 2
overall_partnership_team 1 and 2	weighted average of partnership scored by team 1 and 2
battingstrength_team 1 and 2	weighted average of overall batting strength for team 1 and 2 in terms of batting average, milestone reached etc.
discipline_factor_team 1 and 2	weightage feature of extra runs and maiden overs for team 1 and 2
bowling_strength_team 1 and 2	weighted average of overall bowling strength for team 1 and 2 in terms of bowling average, discipline factor etc.
winning_consistency_team 1 and 2	weightage of previous 10 matches result and previous match result for team 1 and 2
overall_team 1 and 2_strenth	reflects overall team strength with weighted average of batting and bowling strength for team 1 and 2
team1_result	team 1 result (Win/Loss).

Table 1. Features with description

3.3. *Feature Selection*

Feature selection methods are applied to solve overfitting, accuracy improvement, and training time of algorithms. In this work, feature selection methods like chi-square, Mutual Information, Recursive Feature Elimination (RFE), Analysis of variance (ANOVA), Automatic Recursive Feature Elimination (ARFE), Embedded Method, Principal Component Analysis (PCA), and Linear discriminant analysis (LDA) are used. The testing of all feature selection algorithms is performed based on accuracy. The investigation for a better combination of feature selection methods with machine learning algorithms is completed to get maximum accuracy [9,17,19,22].

3.4. *Learning Algorithms/ Model Selection*

In this work, machine learning algorithms like Logistic regression, Naïve Bayes, K-Nearest Neighbors, Support Vector Machine, Decision Tree, Random Forest, Gradient Boosting algorithm, XGBoost, and CatBoost algorithm are used. Artificial Neural Network is also designed. The article studied these algorithms due to the binary nature of the classification problem. The winner prediction classification problem is y=f(x), where y indicates the dependent variable and x is a single or set of independent variables [6-10,18,23,26].

3.5. *Training and Testing*

In this work, the data for forecasting the result of cricket matches is split into training and testing size of 70-30% ratio. This is because it is necessary to conserve the chronological ordering of the results forecasts data when aiming to predict match results based on historical match records. The cross-validation approach does not apply to sports prediction since it acts with the shuffling of data that can alter the sequence of the event. So, it is more appropriate to do the split manually in training and testing data.

3.6. *Parameter Tuning*

Models might have several parameters, and identifying the right combination of parameters can be considered a search problem. The hyperparameters control the learning process of the algorithm. Hyperparameter tuning is an essential factor in improving the accuracy of the models. The parameter tuning for this work is performed with the GridSearchCV algorithm. Using the Grid Search method of parameter tuning, the parameters with the best performance are selected after getting accuracy/loss for each combination of hyperparameters [11].

3.7. *Ensemble Algorithm*

Voting and stacking methods are used to ensemble results from machine learning models for maximum accuracy. These algorithms take input from multiple machine learning models and provide final results. The first step in the voting ensemble method is to create standalone models using given training data. It wraps models in the second step and calculates the average of the sub-model's predictions to make new predictions. The stacking architecture consists of level 0 for two or more base models whose predictions are combined during a fitting on training data, and level 1 for the Meta model to combine predictions from multiple base models [2-3].

3.8. *Model Evaluation*

The result prediction is performed as a binary classification problem with a class win or loss. The model evaluation is performed with accuracy, precision, recall, F1-score and mean square value. This is also called a classification report or evaluation metrics for the classifier [15,17].

4. Model Formulation and Feature Construction

There are four sets of features built: batting and bowling strength derived features, run-scoring pattern-based features, team strength derived features and external features. Batting and bowling strength derived features define the resilience of batters and bowlers. In ODI matches, players' statistics are changed from one year to another. The performance of batters and bowlers is crucial as they play an essential role in deciding the outcome of a match. These two features are included in determining the match winner in the first model. In run-based features, the run-scoring pattern of the team is analyzed. The run scored in the first ten overs, in the middle phase of the match, and during the last ten overs are studied, and a model is built to forecast the output of the cricket match. A particular set of features is required to highlight a specific team's type or ability level. The team strength features set are built for this model. The fundamental aspect is to put out the team's strength and demonstrate their anticipated success in the coming match. Since the real success of the squad is not understood until the match is over. As a result, their average or expected performance is determined using the last ten matches performance. External features may not directly impact the player's results but can indirectly affect the match due to circumstances favoring one side over the other [8]. The classification of the feature set is shown in Figure 2.

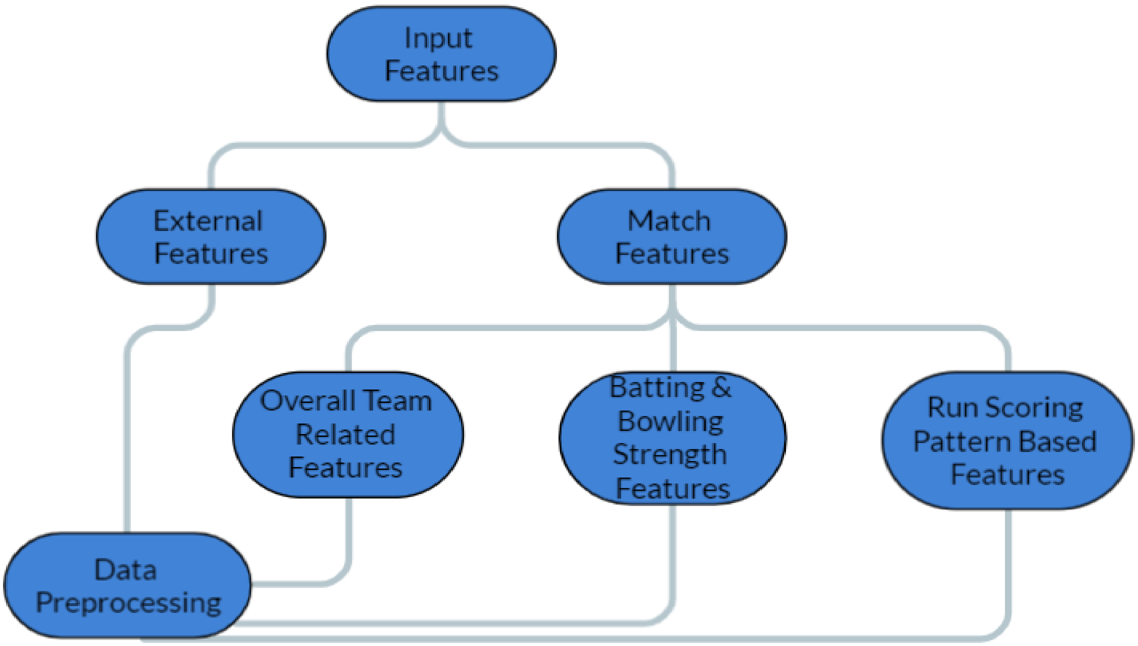
Input Features
External Features
Match Features
Overall Team Related Features
Batting & Bowling Strength Features
Run Scoring Pattern Based Features
Data Preprocessing

Fig. 2. Feature set classification

The batting and bowling strength of teams are measured with their past performances. The first proposed algorithm measures the team's batting strength ('u') with all player's average batting average ('Φ') and batting strike rate ('ψ'). It also uses the player's runs scored ('v') and partnerships ('w') made in the form of 100's, 50's, and 30 plus runs with appropriate weights assigned to each value. The team's batting average, batting strike rate, runs scored, and partnership made having very much impact on the overall batting strength of the team. So, summation and average of all parameters reflecting batting strength are performed, then the appropriate weight is assigned after studying every parameter's effect on the output of cricket match. In proposed algorithm 1 to calculate the team's batting strength, 20% weight is assigned to batting strength calculated using batting average and strike rate, 30% for a run scored, 40% for partnership, and 10% for highest score made by batters. The 'batting_strength' variable represents the overall strength of batsmen. If the team's batting strength is better, it increases the chances of winning. On the other hand, if batting strength is less, winning a game for a particular team decreases.

4.1. *Algorithm 1 Batting strength of team*

```
$\Phi(bat\_avg) = batting\ average\ of\ team$
$\psi\ (bat\_sr) = batting\ strike\ rate\ of\ team$
for all team T do
$u(bat\_strength) = 0.40 * \Phi(bat\_avg) + 0.60 * \psi\ (bat\_sr)$
$v(run\_scored) = 0.20 * no\ of\ 100's + 0.40 * no\ of\ 50's + 0.40 * no\ of\ 30\ plus\ scores$
```

```
    $w(partnership) = 0.40 * \text{no of 100's partnership} + 0.40 * \text{no of 50's partnership} + 0.20 * \text{no of 30 plus partnership}$
    $batting\_strength = 0.20 * u + 0.30 * v + 0.40 * w + 0.10 * HIRS$
end for
```

The team's bowling strength ('u') is the weighted average of bowling average ('Φ'), bowling strike rate ('ψ'), no of wickets taken ('v'), and discipline factor ('w') for the team. Second proposed algorithm calculates the bowling capacity of the team with summation and average bowling average and a bowling strike rate of the team. The weights are assigned to these parameters as per their impact on the match's final result. This article also considered the total number of wickets taken by bowlers and the five-wicket haul of the team. Based on the number of extra runs given and the number of maiden overs bowled by the team's bowlers, this article proposed a parameter known as the discipline factor. The discipline factor describes the number of runs saved by the team with a quality bowling attack. Appropriate weights are assigned to extra run's given, and maiden overs bowled to derive discipline factor parameters. In Algorithm 2, 50% weightage is given to bowling strength derived from bowling average and strike rate of team, 30% for wicket taken, and 20% for discipline in bowling attack to represent the team's overall bowling strength. The quality bowling attack of teams helps to improve the team's winning chances.

4.2. *Algorithm 2 Bowling strength of team*

```
$\Phi(bowl\_avg) = \text{bowling average of team}$
$\psi\ (bowl\_sr) = \text{bowling strike rate of team}$
for all team T do
    $u(bowl\_strength) = 0.30 * \Phi(bowl\_avg) + 0.70 * \psi\ (bowl\_sr)$
    $v(wickets\_taken) = 0.40 * \text{5 wicket haul} + 0.60 * \text{no of wickets taken}$
    $w(discipline\_factor) = 0.60 * \text{no of extra runs given} + 0.40 * \text{no of maiden overs}$
    $bowling\_strength = 0.50 * u + 0.30 * v + 0.20 * w$
end for
```

The team's overall strength is computed using pressure index, winning consistency, run-scoring pattern, wicket losing pattern, batting strength, and bowling strength. The pressure index value ('u') is assigned based on the match type; 0 represents bilateral series, 1 for triangular series, and 2 for multination series, semifinal or final match of the series. The winning consistency ('v') parameter is derived from the recent match result and the previous ten match results of the team. The reason for doing this is due to the possibility that the team won the last match with weak or low-rank team. So, the previous ten match results are also used to calculate the winning consistency of the team, whether they are winning matches with a good percentage or they come with a losing matches streak. The team's run score ('w') and wicket loss ('x') are also considered in three different game phases, with the proper weight assigned to each stage. The first phase (overs 1 to 10) and last phase (overs 41 to 50) are given the same weight (20%) as they demand run-scoring with a reasonable run rate and fewer wickets. The middle phase (overs 21 to 40) is assigned with extra weights (30%) as innings build-up overs. These overs require a steady run rate while maintaining wicket resources. The overall team strength algorithm also takes algorithm 1 and 2 output as input with 20% weight to calculate team strength. All these variables are combined to calculate the overall strength of the team using proper weights for each variable.

4.3. *Algorithm 3 Overall Team Strength*

```
for all team T do
    $u(pressure\_index) = \{0\text{: Bilateral Match, } 1\text{: Triangular Match, } 2\text{: Multinational/ semifinal/final match}\}$
    $v(winning\_consitency) = 0.30 * \text{previous match result} + 0.70 * winning\_streak$
    $w(runscoring\_pattern) = 0.30 * team1\_score + 0.20 * phase1\_runs + 0.30 * phase2\_runs + 0.20 * phase3\_runs$
    $x(wicketloss\_pattern) = 0.30 * phase1\_wicket + 0.20 * phase2\_wicket + 0.30 * phase3\_wicket + 0.20 * total\_wicket$
    $team\_strength = 0.30 * v + 0.20 * batting\_strength + 0.20 * bowling\_strength + 0.20 * w + 0.10 * x$
end for
```

The last set of features consists of external features. Toss is identified as the most important external feature from the collection of external features that impact the result of a cricket match. After winning the toss, the team can bat or field first after studying the other parameters like pitch conditions, opposite team strength, dew factor, etc. The impact of the dew factor is observed during the day-night games. Out of these external features, pitch value is also used for the prediction model. The pitch value is categorized into two types as green and dusty pitches. The pitches available in SENA (South Africa, England, New Zealand, and Australia) countries are considered green pitches and dusty in other countries. The difference between these two types of pitches is that the green pitches are bowlers friendly. While the dusty pitches favor spinners, or they are batsmen friendly. The team1 result in a class win or loss is used as the target variable for this work.

The dataset features are divided into three sets of features. The first set consists of 50 features representing the batting and bowling strength of the team for Model 1. In Model 2, 36 features are used to define run-scoring patterns for the team, and the final model consists of 48 features to reflect the team's overall strength (Model 3).

5. Results and Discussion

Various binary and categorical features are used in this research to design a binary prediction model for winner prediction in one-day international cricket matches. The data is converted into a uniform format for experimentation. Some features are derived from the existing features with a weighted combination of primary features. The reason to propose three different models is that no previous models were available to evaluate the proposed work due to different matches or variables used for the dataset. The algorithms are first trained on the dataset with three sets of features like batting-bowling strength-based features (Model 1), phase-wise run scoring features (Model 2), overall team strength-based features (Model 3), and then feature selection methods are applied to these three models. The voting and stacking classifiers combine the result from more than one model to provide a final result. An artificial neural network is also designed for prediction. The result comparison is performed with or without feature selection methods. Combining the machine learning model with the feature selection method is identified to get the best prediction accuracy. The result for all models is discussed in this section based on evaluation metrics of classifiers.

5.1. *Model 1: Based on Batting and Bowling Strength of Team*

This model consists of 50 input features and one binary output feature. The comparison of accuracy for Model 1 is shown in Table 2.

Classifier	**Accuracy**								
	Without Feature Selection	**Chi**	**M I**	**RFE**	**ANOVA**	**ARFE**	**E M**	**PCA**	**LDA**
Logistic Regression	94.91	95.38	95.61	95.61	95.15	96.30	95.15	95.15	94.68
Naïve Bayes	93.53	94.22	93.76	94.91	93.76	94.45	93.07	91.68	94.68
KNN	87.29	87.99	89.14	87.99	88.22	94.22	91.45	87.99	94.22
SVM	94.91	95.38	95.94	95.38	94.91	95.84	95.15	96.07	94.68
Decision Tree	87.75	90.53	90.76	91.22	91.68	91.45	89.83	86.37	93.07
Random Forest	93.76	94.45	94.45	94.45	93.99	93.76	93.30	89.37	93.30
GBM	94.91	94.68	95.38	95.38	95.38	95.15	95.84	92.84	93.07
XGBoost	94.22	95.38	95.15	95.15	95.38	94.91	95.84	92.6	94.68
CatBoost	95.38	94.91	95.61	95.84	95.61	95.61	95.15	93.07	94.68

Table 2. Accuracy Model 1

From Table 2 following results are highlighted.

- CatBoost achieved a high accuracy of 95.38% without feature selection.
- KNN achieved the lowest accuracy of 87.29%.
- Feature selection methods are applied to the algorithms, and due to that accuracy is improved for all algorithms.
- Logistic Regression has achieved maximum accuracy of 96.30% with Automatic Recursive Feature selection to predict a winner for the model based on the batting and bowling strength of the team.

The classification report consists of the precision, recall, and F1-score shown in Figure 3 for Model 1. The Logistic Regression, SVM, GBM, and XGBoost achieved maximum precision and F1-Score of 95% for prediction. The maximum value of 96% is achieved for recall with the CatBoost.

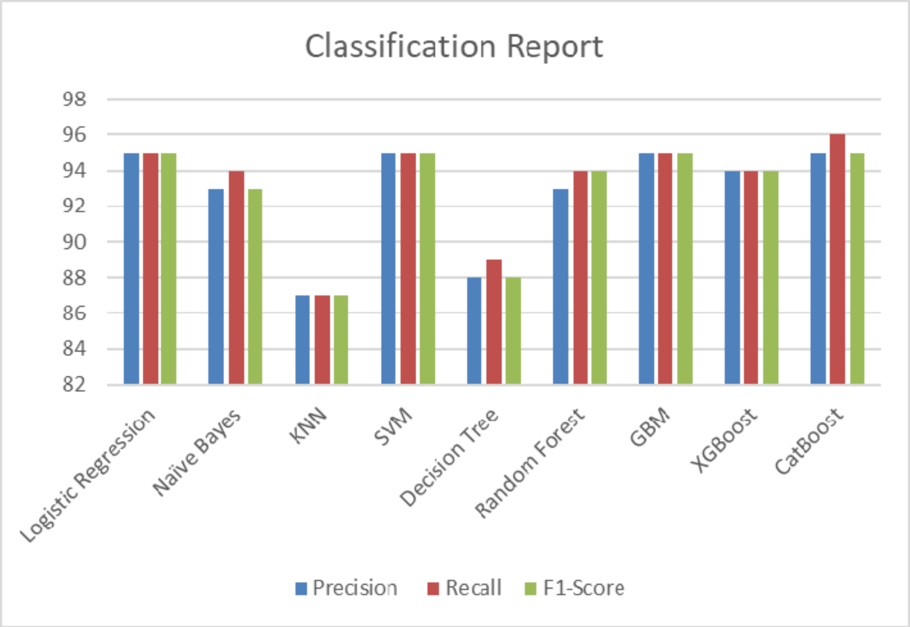

Fig. 3. Classification Report Model 1

The Mean Square Error for prediction is shown in Figure 4. The Mean Square Error for Logistic regression is less, having a value of 0.04. It means that prediction accuracy is more for these algorithms. The KNN and Decision Tree algorithm have a maximum value for Mean Square Error of 0.12 for prediction of the winner.

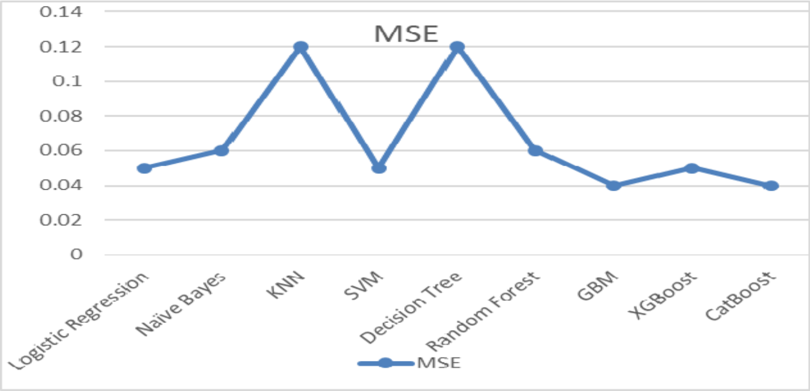

Fig. 4. Mean Square Error Model 1

5.2. *Model 2: Based on Run Scoring Pattern for Team*

The run scoring base model consists of 36 features as input and one feature as an output. The result comparison is shown in Table 3.

Classifier	**Accuracy**								
	Without Feature Selection	**Chi**	**M I**	**RFE**	**ANOVA**	**ARFE**	**E M**	**PCA**	**LDA**
Logistic Regression	95.38	94.22	96.07	95.38	95.38	95.61	92.60	94.68	94.45
Naïve Bayes	90.53	90.30	91.22	91.91	90.76	88.91	88.91	92.37	94.68
KNN	81.52	84.06	84.98	85.21	85.68	84.06	89.37	81.52	93.30
SVM	95.84	95.84	96.30	96.07	96.07	95.84	92.60	96.30	94.68
Decision Tree	81.75	85.21	84.06	84.52	84.75	86.60	80.53	85.68	91.91
Random Forest	89.14	90.76	90.3	90.53	90.76	89.60	87.29	89.37	92.37
GBM	94.68	93.53	92.84	95.15	93.99	94.68	87.75	90.30	91.91
XGBoost	93.76	93.53	93.76	93.76	93.76	94.45	89.14	93.76	94.68
CatBoost	94.68	94.91	95.15	94.91	94.91	94.22	90.3	94.22	94.91

Table 3. Accuracy Model 2

From the Table 3 following observations are made

- The accuracy of 95.84% is achieved with the SVM without using a feature selection process.
- The KNN has achieved the lowest accuracy of 81.52%.
- The combination of SVM with mutual information and PCA obtained maximum accuracy of 96.30% for model 2.

The classification report for Model 2 is shown in Figures 5. The Maximum value of Precision, Recall and F1-Score is 96% for the SVM algorithm.

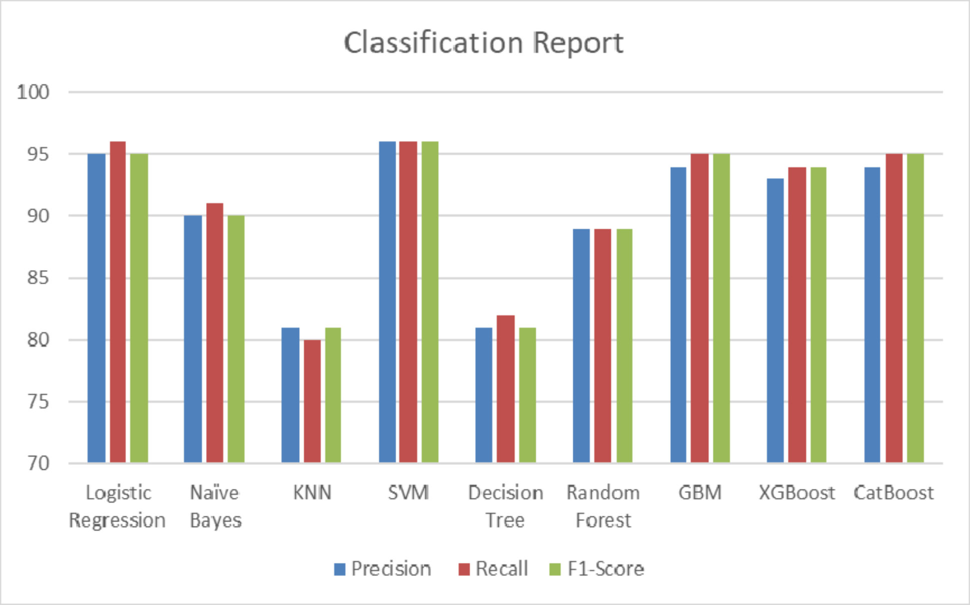

Fig. 5. Classification Report Model 2

The Mean Square Error for Model 2 is shown in Figure 6. The Mean Square Error for Logistic Regression and CatBoost algorithm is less, having a value of 0.04, while KNN and Decision tree algorithm gain the maximum value of Mean Square Error 0.18. It means that KNN and decision prediction error is more in prediction.

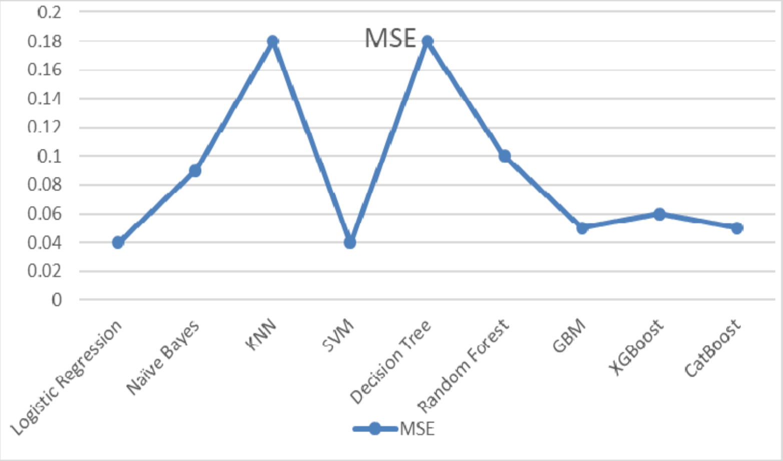

Fig. 6. Mean Square Error Model 2

5.3. *Model 3: Overall team strength-based model*

This model has 48 input features and one output feature. The comparison of accuracy for Model 3 is shown in Table 4.

Classifier	Accuracy								
	Without Feature Selection	Chi	M I	RFE	ANOVA	ARFE	E M	PCA	LDA
Logistic Regression	95.38	94.68	94.45	95.61	94.68	95.84	96.07	94.68	93.3
Naïve Bayes	93.99	93.53	93.99	93.99	93.99	93.99	93.07	92.37	93.3
KNN	88.22	88.45	86.83	88.22	88.68	88.91	93.07	88.91	93.53
SVM	95.15	95.38	94.45	95.15	96.07	94.68	96.07	95.38	93.76
Decision Tree	85.91	90.53	90.3	87.75	89.37	88.91	90.76	89.83	90.76
Random Forest	93.07	93.3	93.07	93.07	93.76	91.68	90.99	91.45	91.22
GBM	93.99	94.22	94.45	94.22	93.76	94.22	94.68	94.68	90.76
XGBoost	93.07	93.76	93.76	93.99	93.3	94.22	94.45	93.76	94.45
CatBoost	93.76	93.99	94.22	94.22	94.22	93.99	95.15	92.84	94.22

Table 4. Accuracy Model 3

The result of model 3 are described with Table 4 as follows

- Logistic Regression obtains the highest accuracy of 95.38%, and the Decision Tree has the lowest accuracy of 85.91%.
- The combination of Logistic Regression and Embedded Method has an accuracy of 96.07%.
- SVM has a maximum accuracy of 96.07% for ANOVA and the Embedded method.

The classification report for Model 3 is shown in Figure 7. The maximum value of 95 % for precision, Recall and F1- score is achieved with Logistic Regression. The Mean Square Error for Model 3 is shown in Figure 8.

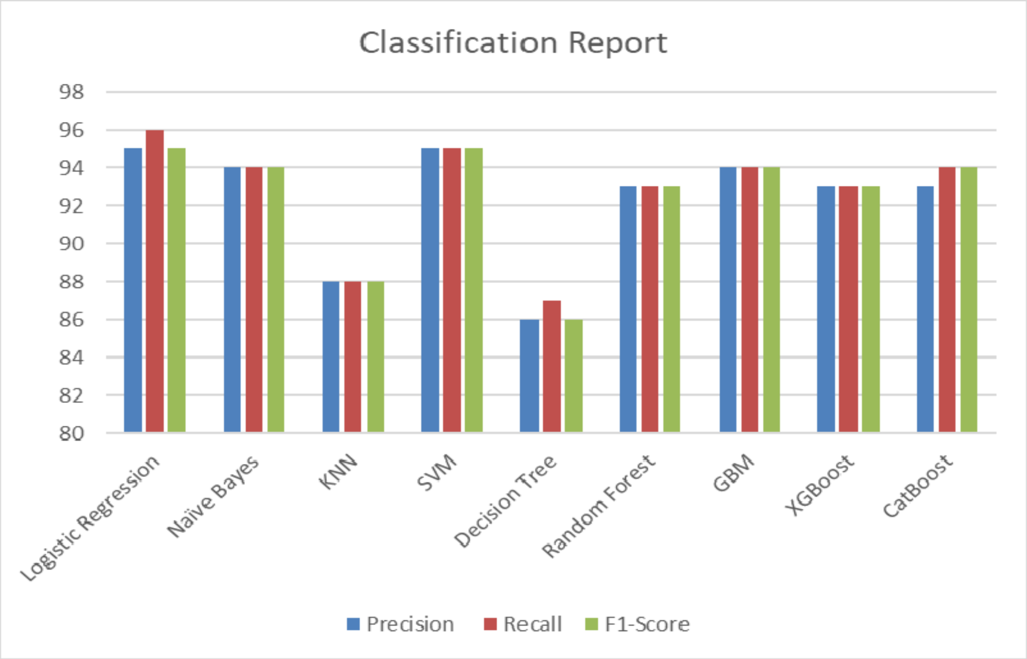

Fig.7. Classification Report Model 3

The Mean Square Error for Logistic Regression and SVM algorithm is less, having a value of 0.04. Prediction is more accurate with these models. The Decision tree algorithm gains the maximum value of Mean Square Error 0.14. It indicates that decision tree prediction error is high.

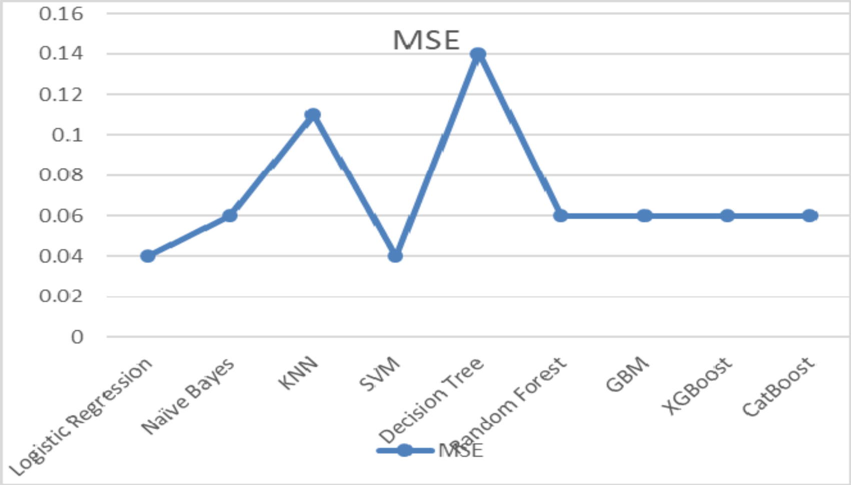

Fig. 8. Mean Square Error Model 3

5.4. *Ensemble method and Artificial Neural Network*

The result of voting and stacking ensemble is discussed in this part. An artificial neural network is designed for all three models, and their results are investigated. The algorithms with maximum accuracy are provided as input to the voting and stacking ensemble methods. The results of ensemble methods are shown in table 5. Voting classifiers for Model 1&3 have a maximum accuracy of 95.84% and 94.91% respectively. It remains unchanged after applying feature selection. For Model 2, voting classifier accuracy is 95.38, improving to 95.84% after using feature selection. Stacking classifiers accuracy for Model 1 & 2 is 95.61%. Stacking classifier with Automatic Recursive Feature elimination has achieved an accuracy of 96.3% for Model 1, and in Model 2, it improved to 95.84%. For Model 3, the accuracy of the stacking classifier is 95.15 % without feature selection and its upgrades to 95.38% after applying the feature selection. An Artificial Neural Network is designed with ReLU and sigmoid function to get improved performance for predicting the result of cricket matches [26]. ANN model achieved an accuracy of 96.31%, 94.93%, and 95.39% for Model 1, 2, and 3 respectively for prediction of the winner.

Classifier	Accuracy								
	Without Feature Selection	Chi	M I	RFE	ANOVA	ARFE	E M	PCA	LDA
Voting Model 1	95.84	95.84	95.84	95.84	95.84	95.84	95.84	95.84	95.84
Voting Model 2	95.38	95.84	95.84	95.84	95.84	95.84	95.84	95.84	95.84
Voting Model 3	94.91	94.91	94.91	94.91	94.91	94.91	94.91	94.91	94.91
Stacking Model 1	95.61	95.15	95.15	95.15	95.15	96.3	95.15	95.15	95.15
Stacking Model 2	95.61	95.84	95.84	95.84	95.84	95.84	95.84	95.84	95.84
Stacking Model 3	95.15	95.38	95.38	95.38	95.38	95.38	95.38	95.38	95.38

Table 5. Accuracy Ensemble method

As discussed earlier, the baseline models are not available to compare this proposed work due to different variables and datasets. However, variables used to predict the winner of a cricket match is nearly identical in cricket. So, the criterion for this article is diluted, and the proposed work is compared with the literature. The accuracy obtained for this work is far better than the existing work. The maximum accuracy obtained in the literature survey is 75% for ODI match [20]. This article model's maximum accuracy of 96.3 % is obtained to predict the winner of a cricket match in models 1 and 2. The right ensemble of machine learning algorithms is performed with a feature selection algorithm to get good results. The ensemble classifiers are also used to find maximum accuracy for the models. Hence, this approach can achieve maximum accuracy compared to previous work to predict a winner in one-day international matches.

In this research work, feature selection methods are used to produce good accuracy with a smaller number of features set as input for the training of machine learning models. The features which impact more on the result of cricket matches are selected using feature selection methods. The features like teams score, run rate of teams, no of wickets lost, no of 30 plus runs scored by batters, and no of 30 plus partnerships for teams have more impact on prediction of result for cricket matches in Model 1. In Model 2, team scoring runs and wickets lost in phases 2 and 3 (20 to 50 overs), run scored in 4's and 6's is given more weightage by feature selection methods. For Model 3, the features like batting and bowling strength of teams, run-scoring and wicket lost pattern for the team, milestone reaching ability of batters, ranking of teams, overall partnership for teams are identified as more important features with feature selection methods. The right combination of machine learning model with feature selection method is found after repeatedly running the model on training and testing data from the dataset. In a few cases, the model's accuracy remains unchanged or decreased after applying feature selection methods. Overall, the redundant and irrelevant features are removed with feature selection methods. The benefits of feature selection methods are accuracy of models is improved with less training time [1,9,15,17].

6. Conclusion

The prediction of the winner in any sports match is a difficult task. This work addressed the area of winner prediction in one day international matches. Three models are implemented based on the batting and bowling strength of the team, run-scoring pattern for the team, and overall strength of the team. The selective machine learning methods are applied and evaluated. The accuracy of selected models is compared with or without feature selection methods. The implementation outcome shows that model accuracy is improved after using feature selection methods. With an ensemble voting and stacking classifier method, promising results are achieved with or without feature selection methods. Neural networks are designed for these three models, and they get good accuracy for the result prediction of a cricket match. The maximum accuracy for Model 1 is 96.3 % with a combination of Logistic regression and Automatic Recursive Feature Elimination method. The combination of SVM with Mutual Information and PCA gets maximum accuracy of 96.3% for Model 2. 96.07% of accuracy is achieved by Model 3 with the blend of Logistic regression with Embedded Method and Support Vector Machine with ANOVA or Embedded method for prediction of a winner in one day international matches.

For future scope, this work can be extended with more parameters impacting the result of cricket matches. The ultimate focus will be on improving the accuracy of the classification model after adding new features to the data.

Conflicts of interest

"The authors have no conflicts of interest to declare"

References

[1] Ishi M and Patil J (2020) A Study on Impact of Team Composition and Optimal Parameters Required to Predict Result of Cricket Match. Lecture Notes Networks System, **100**, pp. 389–399. doi: 10.1007/978-981-15-2071-6_32.

[2] Beal R, Norman TJ, Ramchurn SD (2019) Artificial intelligence for team sports: A survey. The Knowledge Engineering Review, **34**, pp. 1–37. https://doi.org/10.1017/S0269888919000225.

[3] Balafoutas L, Chowdhury SM, Plessner H (2019) Applications of sports data to study decision making. Journal of Economic Psychology, **75**. https://doi.org/10.1016/j.joep.2019.02.009.

[4] Wickramasinghe I (2020) Naive Bayes approach to predict the winner of an ODI cricket game. Journal of Sports Analytics, **6**, pp. 75–84. https://doi.org/10.3233/jsa-200436.

[5] Saikia H (2020) Quantifying the Current Form of Cricket Teams and Predicting the Match Winner. Management and Labour Studies, **45**, pp. 151–158. https://doi.org/10.1177/0258042x20912603.

[6] Sivaramaraju Vetukuri V, Rajender R, Sethi N (2019) A multi-aspect analysis and prediction scheme for cricket matches in standard T-20 format. International Journal of Knowledge-Based and Intelligent Engineering Systems, **23,** pp. 149–154. https://doi.org/10.3233/KES-190407.

[7] Viswanadha S, Sivalenka K, Jhawar MG, Pudi V (2017) Dynamic winner prediction in twenty20 cricket: Based on relative team strengths. CEUR Workshop Proceedings, **1971**, pp. 41–50.

[8] Jayalath KP (2017) A machine learning approach to analyze ODI cricket predictors. Journal of Sports Analytics, **4**, pp. 73–84. https://doi.org/10.3233/jsa-17175.

[9] Passi K, Pandey N (2018) Increased Prediction Accuracy in the Game of Cricket Using Machine Learning. International Journal of Data Mining & Knowledge Management Process, **8**, pp. 19–36. https://doi.org/10.5121/ijdkp.2018.8203.

[10] Kaluarachchi A and Varde A (2010) CricAI: A classification based tool to predict the outcome in ODI cricket. Proceedings of the 2010 5th International Conference on Information and Automation for Sustainability, ICIAFS 2010, pp. 250–255. doi: 10.1109/ICIAFS.2010.5715668.

[11] Bunker RP, Thabtah F (2019) A machine learning framework for sport result prediction. Journal of Applied Computing and Informatics, **15**, pp. 27–33. https://doi.org/10.1016/j.aci.2017.09.005.

[12] Keshtkar Langaroudi M, Yamaghani M (2019) Sports Result Prediction Based on Machine Learning and Computational Intelligence Approaches: A Surveys. Journal of Advances in Computer Engineering and Technology. Science and Research Branch, Islamic Azad University [Online], **5**, pp. 27–36. Available: http://jacet.srbiau.ac.ir/article_13599.html.

[13] Agrawal S, Singh SP, Sharma JK (2018) Predicting results of Indian premier league T-20 matches using machine learning. Proceedings - 2018 8th International Conference on Communication Systems and Network Technologies, CSNT 2018, pp. 67–71. https://doi.org/10.1109/CSNT.2018.8820235.

[14] Vistro DM, Rasheed F, David LG (2019) The cricket winner prediction with application of machine learning and data analytics. International Journal of Scientific and Technology Research, **8**, pp. 985–990.

[15] Baboota R, Kaur H (2019) Predictive analysis and modelling football results using machine learning approach for English Premier League. International Journal of Forecasting, **35**, pp. 741–755. https://doi.org/10.1016/j.ijforecast.2018.01.003.

[16] Abedin M, Urmi SR, Mozumder TI (2019) Forecasting the Outcome of the Next ODI Cricket Matches to Be Played. International Journal of Recent Technology Engineering, **8**, pp. 10269–10273. https://doi.org/10.35940/ijrte.d4505.118419.

[17] Kapadia K, Abdel-Jaber H, Thabtah F, Hadi W (2019) Sport analytics for cricket game results using machine learning: An experimental study. Journal of Applied Computing and Informatics. https://doi.org/10.1016/j.aci.2019.11.006.

[18] Bhattacharjee D, Talukdar P (2019) Predicting outcome of matches using pressure index: evidence from Twenty20 cricket. Journal of Communications in Statistics: Simulation and Computation, pp. 1–13. https://doi.org/10.1080/03610918.2018.1532003.

[19] Gu W, Foster K, Shang J, Wei L (2019) A game-predicting expert system using big data and machine learning. Journal of Expert Systems with Applications, **130**, pp. 293–305. https://doi.org/10.1016/j.eswa.2019.04.025.

[20] Jayanth SB, Anthony A, Abhilasha G, et al (2018) A team recommendation system and outcome prediction for the game of cricket. Journal of Sports Analytics, **4**, pp. 263–273. https://doi.org/10.3233/jsa-170196.

[21] Asif M, McHale IG (2019) A generalized non-linear forecasting model for limited overs international cricket. International Journal of Forecasting, **35**, pp. 634–640. https://doi.org/10.1016/j.ijforecast.2018.12.003.

[22] Kampakis S, Thomas W (2015) Using Machine Learning to Predict the Outcome of English County twenty over Cricket Matches. pp. 1–17. Available: http://arxiv.org/abs/1511.05837.

[23] Pathak N, Wadhwa H (2016) Applications of Modern Classification Techniques to Predict the Outcome of ODI Cricket. Procedia of International Conference on Computational Science Applications, **87**, pp. 55–60. https://doi.org/10.1016/j.procs.2016.05.126.

[24] Lamsal R, Choudhary A (2018) Predicting Outcome of Indian Premier League (IPL) Matches Using Machine Learning. Available: http://arxiv.org/abs/1809.09813.

[25] https://www.espncricinfo.com/

[26] Hudnurkar S, Rayavarapu N (2022) Binary classification of rainfall time-series using machine learning algorithms. International Journal Electrical and Computer Engineering, **12**, pp. 1945–1954. https://doi.org/10.11591/ijece.v12i2.pp1945-1954.

Authors Profile

Manoj Ishi received the B.E. degree in Computer engineering from R. C. Patel Institute of Technology, Shirpur, in 2010 and the M. Tech degrees in Computer Engineering from Rajiv Gandhi Technical Univeristy, Bhopal, in 2014. Currently, he is a Research Scholar and Assistant Professor at the Department of Computer Engineering, R. C. Patel Institute of Technology, Shirpur. His research interests include machine learning, sports analytics, deep learning. He can be contacted at email: ishimanoj41@gmail.com.

Dr. Jayantrao Patil holds a PhD in Computer Engineering from North Maharashtra University, Jalgaon. He is currently Director at the R. C. Patel Institute of Technology, Shirpur. He achived best principal award from North Maharashtra University. He received best principal of the year by CSI, Mumabi at Technext India. His research area is web mining, machine learning, data mining. He can be contacted at email: jbpatil@hotmail.com

Dr. Nitin Patil Completed a PhD in Computer Engineering from North Maharashtra University, Jalgaon. He is currently Head of Department at the R. C. Patel Institute of Technology, Shirpur. His research area is watermarking, machine learning, data mining. He can be contacted at email: er.nitinpatil@gmail.com

Dr. Vaishali Patil received Ph. D degrees in Computer Science from North Maharashtra University, Jalgaon. She is currently the Director RCPET's Institute of Management Research and Development, Shirpur, India. Her research interests include artificial intelligence and Natural Language Processing. she can be contacted at email: vaishali.imrd@gmail.com.