Annals of Operations Research
https://doi.org/10.1007/s10479-022-04541-6

ORIGINAL RESEARCH

Best strategy to win a match: an analytical approach using hybrid machine learning-clustering-association rule framework

Praveen Ranjan Srivastava[1] · Prajwal Eachempati[1] · Ajay Kumar[2] · Ashish Kumar Jha[3] · Lalitha Dhamotharan[4]

Accepted: 18 January 2022

Abstract
One of the significant challenges in the sports industry is identifying the factors influencing match results and their respective weightage. For appropriate recommendations to the team management and the team players, there is a need to predict the match and quantify the important factors for which prediction models need to be developed. The second thing required is identifying talented and emerging players and performing an associative analysis of the important factors to the match-winning outcome. This paper formulates a hybrid machine learning-clustering-associative rules model. This paper also implements the framework for cricket matches, one of the most popular sports globally watched by billions around the world. We predict the match outcome for One day Internationals (ODIs) and Twenty 20 s (T20s) (two formats of Cricket representing fifty over and twenty over versions respectively) adopting state-of-the-art machine learning algorithms, Random Forest, Gradient Boosting, and Deep neural networks. The variable importance is computed using machine-learning techniques and further statistically validated through the regression model. The emerging talented players are identified by clustering. Association rules are generated for determining the best possible winning outcome. The results show that environmental conditions are equally crucial for determining a match result, as are internal quantitative factors. The model

✉ Ajay Kumar
akumar@em-lyon.com

Praveen Ranjan Srivastava
praveen.ranjan@iimrohtak.ac.in; fpm03.007@iimrohtak.ac.in

Ashish Kumar Jha
AKJHA@tcd.ie

Lalitha Dhamotharan
L.Dhamotharan@exeter.ac.uk

1 Indian Institute of Management (IIM) Rohtak, Rohtak, India

2 Emlyon Business School, Écully, France

3 Trinity Business School, Trinity College Dublin, Dublin, Ireland

4 University of Exeter Business School, Exeter, UK

Published online: 07 March 2022

is thus helpful for both team management and for players to improve their winning strategy and also for discovering emerging players to form an unbeatable team.

Keywords Machine learning · Sports analytics · Neural network · Random forest · Ensemble gradient boost predictive model · Match result prediction · Clustering · Apriori algorithm

1 Introduction

Cricket is one of the most-watched and most-followed sports with a presence worldwide (Saha, 2020; Thomson, Reyers, & Swartz, 2021; Thomson et al., 2021). Statistical analysis is one of the critical aspects of the game, as demonstrated by using mathematical techniques to resolve weather-affected games (Stern, 2016). It is a game that has penetrated all strata of society irrespective of age and demographics. The interest in the game has drawn many enthusiasts to watch and analyze the decisions of the game and express opinions about players, team composition, and match-winning strategies. However, opinions backed by data and statistical analysis would add weight to the observation and increase the validity.

For this purpose, data in sports are being collected and analyzed, made possible by the availability and integration of physical and digital sources. Data is available on websites like ESPN Cricinfo, CricBuzz, Cricingif, and Howstat, etc., catering to the special interests of sports experts, analysts, and enthusiasts across the world. The availability of sports records in different formats (player-wise, team-wise, and matches-wise) can enhance decision-making capabilities related to the players and team's performance, mental health, and safety. Further, this encourages fan engagement and helps in formulating marketing strategies.

Players and the team management are confronted with the impending problem of the team's performance and country rankings. All the sports managers strive to groom an ideal group of cricket players who are considered a formidable winning competition. Players, on the other hand, strive to be a part of this winning combination. They are interested to understand the rationale or strategies required to accomplish a win. There is a need, therefore, to first identify the different factors impacting favorable match results. Existing studies conducted in this regard have focused primarily on internal factors influencing match decisions.

These include runs, wickets, bowling and batting averages, strike rates, and catches for match decisions. However, the impact of player-specific factors and environmental conditions like match time, match venue (home or away), batting position (whether the team bats first or second), and toss decisions are understated in current studies. There is a need to consider these factors as they are pertinent in the long run for determining match results. There is also a need to identify the emerging talented cricket players by predicting the result of the match based on their parameters. This gains prominence in the study context of India as the sports leagues, with billions of dollars of investments, rely on the identification of talent to make it successful (Kamath et al., 2020). Such data-backed identification would help understand how their performance impacts the winning outcome and identify associative rules that determine which combination of factors leads to a favorable match outcome.

Further, while different statistical learning and operations research methods are adopted in current literature, a more customized player-wise and captain-wise analysis for selection are not considered. Cricket is not a game of any individual but a team. It is not driven by the captain's performance only but all eleven players and the impact of various categorical factors like the toss, whether the team is batting first or second, to impact match result. However,

the impact of these factors is not statistically analyzed. It is required to identify if there are patterns in deciding the match result.

While current research primarily consists of statistical and operations research-based techniques, there is a proliferation of machine learning algorithms (Deval et al., 2021) that can predict the match result and quantify and identify the most critical factors influencing the results. Given the nature of Cricket as a sport with multiple formats, this becomes even more pertinent as these results will vary according to the match format, i.e., different for One Day Internationals (ODI) match and a different set of important factors for T20 match results. Like one team leading in ODI, but the same team is placed 6th in the T20.

Hence, a more comprehensive model backed by machine learning algorithms would help in making appropriate team management, toss decisions. Player retention strategies can be formulated by identifying the top contributing cricket players leading the team for victory.

To successfully implement the model, the study is divided modularly into the following research objectives (research questions):

RQ1 What is the impact of player-specific factors and environmental conditions like match time and toss on match outcome?
RQ2 How do categorical factors like the toss, match time, and batting position impact match result?
RQ3 Which ML approach predicts the most accurate match outcome?
RQ4 What are the most significant factors (both individual and interaction) impacting ODI match and T20 match result?
RQ5 Who are the emerging talented cricket players for India?

We use a hybrid machine learning-clustering-association rule model in the paper to solve the above research questions. It enables us to predict match results, identify important factors, cluster. It also enables us to identify emerging players, and formulate association rule patterns. The state-of-the-art machine learning algorithms (Random Forest, Gradient Boosting, and Deep Neural Networks) are implemented to predict results. Association rule mining is also performed using the Apriori algorithm. This algorithm is chosen due to its simple and easy-to-understand characteristics among association rule learning algorithms. The resulting rules are intuitive and easy to communicate to an end-user. This study's data is collected from ESPN CricInfo Statsguru, a structured cricket database website that uses filters and provides customized data in terms of the player, team, and match-wise statistics.

The paper also elaborates: The literature detailing the existing studies in match result prediction and the rationale for adopting the above predictors is reviewed in Sect. 2. Section 3 shows the data collection (web scraping) and research methodology of this research. Section 4 presents the analysis and results obtained. Section 5 highlights the theoretical contribution, practical utility, limitations, and scope for future research. Section 6 concludes the paper.

2 Literature review

Prior studies conducted in the domain of sports analytics have analyzed the factors impacting match outcome and showed that the factors considered could be broadly categorized into (1) Quantitative factors and (2) Categorical factors. Quantitative metrics include the Runs scored, Wickets taken, Bowling average, and number of catches (Bose et al., 2021; Deval et al., 2021; Kamble, 2021). Other quantitative factors like player ratings and partnerships

data were not consistently available for all matches and teams on the cricket statistics websites and hence were excluded. There are other decision variables, categorical in nature, that are important in the state of the game but have been rarely included in such analysis. These are Toss decision (won/lost), Match location (home/away), Match time (day or day and night), and Batting position (batting first or batting second). These have been incorporated separately as categorical factors, as depicted in Fig. 1 below.

The factors that emerge from the existing literature and the rationale for their inclusion in the study are described below:

- *Runs scored* The number of runs scored by a batsman is one of the most critical factors deciding the match irrespective of the pitch and playing conditions. It sets a formidable benchmark and target for the batsman and is vital for match victory (Kamble, 2021).
- *Wickets taken* From a bowler's perspective, the number of wickets taken is an equally vital factor for deciding the match outcome. The factor is antithetical to the number of runs scored since the fall of an important batsman can mold the outcome of the match accordingly in the direction of the bowling team (Thorley, 2021).
- *Bowling average* Similarly, for a bowler, bowling average is also important since for a bowler, other than taking wickets, constraining the flow of runs (economy) is also critical for match outcome (Deval et al., 2021).
- *The number of catches* The number of catches, particularly, the catches taken of important players can swing the match outcome in favor of the fielding team (Bose et al., 2021).
- *Batting position* The extent to which a team is comfortable batting first or fielding first is also a deciding factor since some teams have a demonstrated history of being successful more as chasers or in batting first. This depends on the team composition and potential of the team and is hence included as a factor in categorical form for the study (1 for batting first and 0 for fielding first) (Weeraddana & Premaratne, 2021).

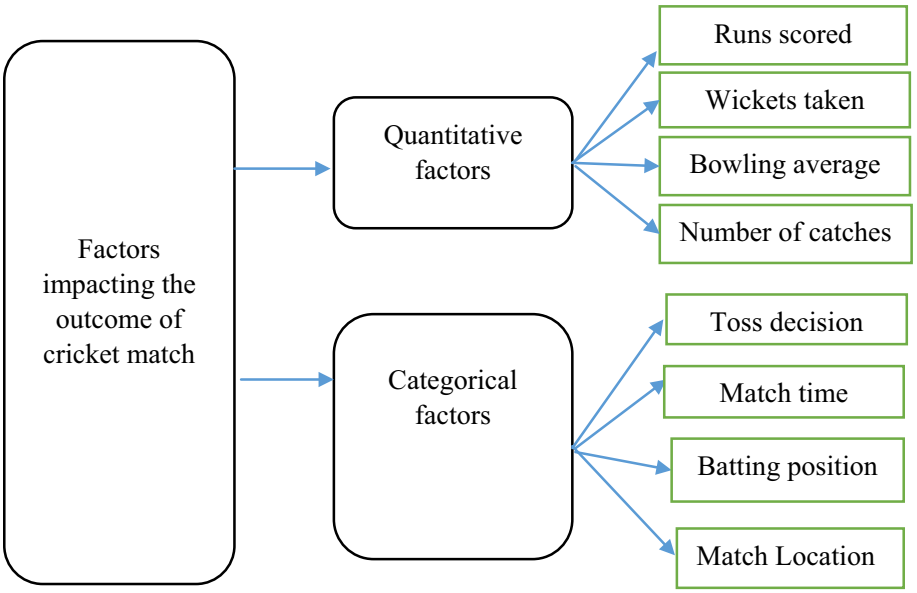

Fig. 1 The emergence of factors from prior studies

- *Match location* The venue of the match is also an important determinant of the match outcome (classified in terms of home or away i.e., whether the match is played on home ground of the team or in a different country). Different teams have different records that vary based on the venue since some teams may be more successful at their local venue than a different country due to familiarity with the pitch conditions (Bliss et al., 2021).
- *Match time* The duration of play of the match and the time at which the match starts and ends (whether day match or day and night match) is a critical external factor determining the outcome. The same match pitch can have different conditions at different times like the presence of dew on the ground and may indirectly influence the match outcome (Mondal et al., 2021)
- *Toss decision* One of the main external factors is the toss. It is believed that some matches are won just by winning the toss due to demonstrated history of some teams being successful with winning the toss. This also depends on the pitch conditions and if a team is available to get a favorable toss outcome, this may influence the match outcome probabilistically and hence is included as a factor in the study (Sahu, 2021).

2.1 Prior studies on match result prediction

Prior works adopting machine learning (ML) techniques in match result prediction are tabulated in Table 1.

The Table 1 therefore summarizes the existing research in the domain of match result prediction. The studies have currently adopted more baseline techniques like statistical analysis, optimization for team allocation and analyzing team performance. However, a more granular player-wise analysis identifying the key factors influencing match result need to be explored. There is a need to develop match result prediction models from more state-of-the-art machine learning techniques for higher accuracy and identify a weightage (importance score) to each factor influencing match outcome. We discuss the limitations of the existing studies reviewed in Sect. 2.2.

2.2 Limitations of prior studies

The limitations of the existing studies of match result prediction are illustrated below in Fig. 2:

Firstly, while existing studies identify internal factors for match outcome, the impact of player-specific factors and environmental conditions like match time and toss is not factored in for predicting match outcome.

Second, none of the existing studies stated above has considered a player-wise and captain-wise analysis for selection. However, teams are driven by the performance of the captains and individual players, and the impact of various categorical factors like the toss, match time, and whether the team is batting first or second have an impact on match outcome. However, the impact of these factors is not statistically analyzed. This is needed to identify if there are patterns in deciding the match outcome.

Third, a comparative analysis of ML approaches for match result prediction is not done, which is needed to identify and predict the match outcome. A reliable data source like ESPN CricInfo can be utilized by sourcing player-wise statistics and categorical variables like the toss, match time, and batting position to perform the prediction.

Fourth, there is a need to identify the most significant factors (both individual and interaction) impacting ODI match and for T20 match outcomes. This would help make appropriate

Table 1 Prior studies on match result prediction

S. no	Study	Research objective	Methodology/technique	Implications of the study
1	Deval et al. (2021)	The paper investigates the prediction problem of when to declare the third innings of a test match	Machine learning techniques like Support Vector Machine, logistic regression, and Artificial Neural Network are used to predict the outcome of a test match at different stages of the match. This will aid the captain to decide when to declare the innings	Support vector machine is found to be the most appropriate in terms of predictive accuracy of 88.8%
2	Cea et al, (2020)	This paper analyzes the procedure used by FIFA to screen the players and select the initial teams for the World Cup finals	A mixed-integer linear programming assignment model is developed for choosing the best teams	The environmental factors are found important in match outcome
3	Goossens et al. (2012)	In this paper, we compare the current league format and three other formats that have been considered by the Royal Belgian Football Association	Simulation strategies are evolved for predicting match outcome	The importance of match time is emphasized
4	Thorley (2021)	The impact of age on the bowler's performance is examined	Multivariate correlation analysis is performed on the dataset for mining the patterns	It is found that there is a negative correlation between age and wicket-taking ability
5	Bose et al., (2021)	A model is developed to segregate players into categories according to their performances in the T-20 tournaments	Deep neural networks are used to classify the player into the batsman, bowler, and allrounder	It is found that deep neural networks is the most accurate
6	Weeraddana and Premaratne (2021)	A novel approach to predict the winning team and next-over score is illustrated in the paper	Extra-Gradient Boosting algorithms are implemented for the problem	XGBoost is found to be 84% accurate

Table 1 (continued)

S. no	Study	Research objective	Methodology/technique	Implications of the study
7	Mondal et al. (2021)	This study analyses competitive balance (CB) in all formats of men's international Cricket	The results display a mixed picture in respect of competitive balance across the various formats of Cricket	No significant changes have been observed in CB scores in test and ODI while it is improved for T20
8	Sahu (2021)	Predictive analytics of cricket matches is performed	Machine learning in Google Colab tool is performed for analytics	And there are many factors such as individual performance, team performance and environmental factors that need to be considered in planning a game strategy
9	Jain et al. (2021)	Match outcome prediction is aimed to be improved with better mining techniques	ML techniques are used	The best prediction accuracy was found to be 70.58%
10	Nikolaidis (2015)	This paper aims to perform an analysis for better management of the Greek basketball team and strategy formulation	Statistical analysis is performed	A game plan strategy for maximizing the win ratio is provided

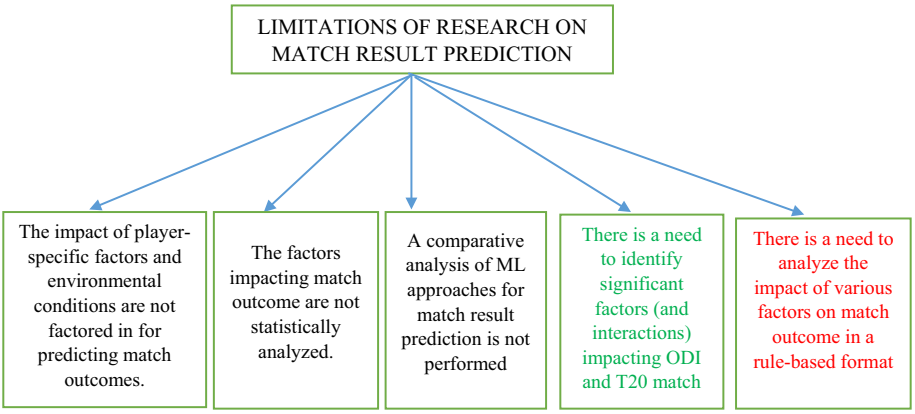

Fig. 2 Research gap

team management, toss decisions, and player retention strategies by identifying top contributing cricket players for victory of the team.

Fifth, there is a need to identify the emerging talented cricket players by predicting the outcome of the match based on their parameters. This would help in understanding how their performance impacts the win ratio of the team. This would enable clustering the players into different grades and forming effective match-winning team combinations. Further, no study associates the impact of various factors on match outcome in a rule-based format. This would help in analyzing which parameters can be tuned to achieve the desired outcome.

For overcoming the above limitations, a hybrid machine learning-clustering-association rule model is adopted in the paper for predicting match outcomes, identifying important factors, clustering and identifying emerging players, and formulating association rule patterns.

The data collection procedure and research methodology adopted in this paper are discussed next, in Sect. 3.

3 Materials and methods

3.1 Data collection

The dataset for this research is sourced from ESPN CricInfo Statsguru[1] dataset, a compendium of all cricket statistics worldwide drawn across all the three formats of the game namely, Test Cricket, One Day Internationals (ODI), and T20s. The aggregated statistics of four current and renowned international cricket team captains of Australia, India, New Zealand, and England (namely, Aaron Finch, Virat Kohli, Kane Williamson, and Joe Root) for all the three above formats were collected in terms of the number of matches played, runs scored and the trending batting average respectively. The data is aggregated according to the following quantitative and categorical attributes. The data stored in table format on the web page of ESPN Cricinfo is scraped by a built-in package in R 'rvest'. The predictors incorporated in the model are discussed below:

3.1.1 Extracting the predictors of the model

The following appropriate factors, namely, Runs scored, wickets taken, Bowling Average, number of catches taken, Toss decision, Match venue/location, Match time, and Batting Position are considered for the predictive model.

The following variables are extracted from ESPN CricInfo Statsguru:

[1] *Runs scored* operationalized as 'Runs'.
[2] *Wickets taken* operationalized as 'Wkts'.
[3] *Bowling Average* operationalized as 'Bowl Av'.
[4] *Number of catches taken* The number of catches taken in a particular match is operationalized as 'Catches'.
[5] *Toss decision* The decision taken at the toss (either the toss is won by the team which provides them an opportunity to decide whether to bat first or field first) is denoted by 'Toss decision'. In the dataset, 1 denotes that the toss is won while 0 represents the lost toss.
[6] *Match venue/location* The location of the match with respect to the playing teams (either the match venue is in one of the host countries of the playing teams or in a different

[1] Retrieved from: https://stats.espncricinfo.com/ci/engine/stats/index.html.

country) is denoted by 'Match location'. In the dataset, 1 denotes home country while 0 implies the match is "away" i.e., played in a different country.

[7] *Match time* The time of the day when the match is played is also extracted and denoted by 'Match time' representing whether the match is played as a 'Day' match or a 'Day and night' match. 'Day' match is denoted by 1 and 'Day and night' match by 0.

[8] *Batting position* The variable denotes whether the team under consideration has batted first and enforced a target to the opposition or batted second and chased a target of runs. 'Batting first' is denoted by 1 and 'Fielding first' ('Batting second') is denoted by 0.

The interactive web scraping is performed using the 'rvest' package in R. The data is present in the website links as a "HTML Table" format. The web page is first read into R and the appropriate HTML Table is scraped by inspecting the HTML code and tags used for designing the tables in the webpage. The parameters are scraped as follows illustrated in Fig. 3.

For instance, to retrieve the statistics of the Indian captain Virat Kohli, there is a name-wise search where we can enter the name of the player or team and retrieve the year-wise records aggregated according to the attributes entailed above.

The webpage showing all the links to test match, ODI, and twenty-twenty (T20) records is illustrated as in Fig. 4.

From this page, we need to retrieve the all-round records of Virat Kohli for one day internationals, test matches and Twenty-20 matches.

The page redirects to the webpage containing the data for all the formats to be web scraped in HTML tables format as illustrated in Fig. 5.

The R package 'rvest' is now deployed to scrape the third HTML table (as inspected in HTML webpage) into a data frame which is exported in Comma Delimited File (.csv) format for further analysis.

Further, for the machine learning models implementation, the other parameters defined like Toss decision(won/lost), Match time (day/day and night) and Batting Position(first/second) need to be extracted however, it is found that these parameters are not displayed directly in the HTML table as features however they are in the form of advanced search filters as illustrated below in Fig. 6.

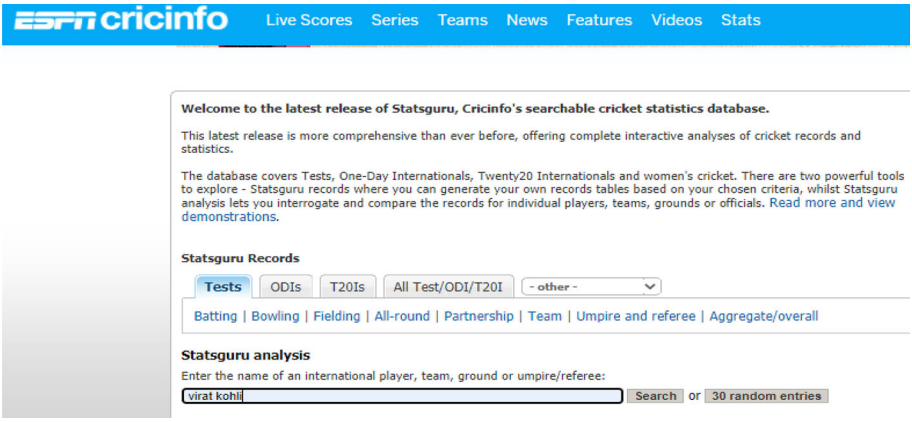

Fig. 3 Player-wise search for statistics *Source*: https://stats.espncricinfo.com/ci/engine/stats/index.html

Fig. 4 Sample webpage of Statsguru records *Source*: https://stats.espncricinfo.com/

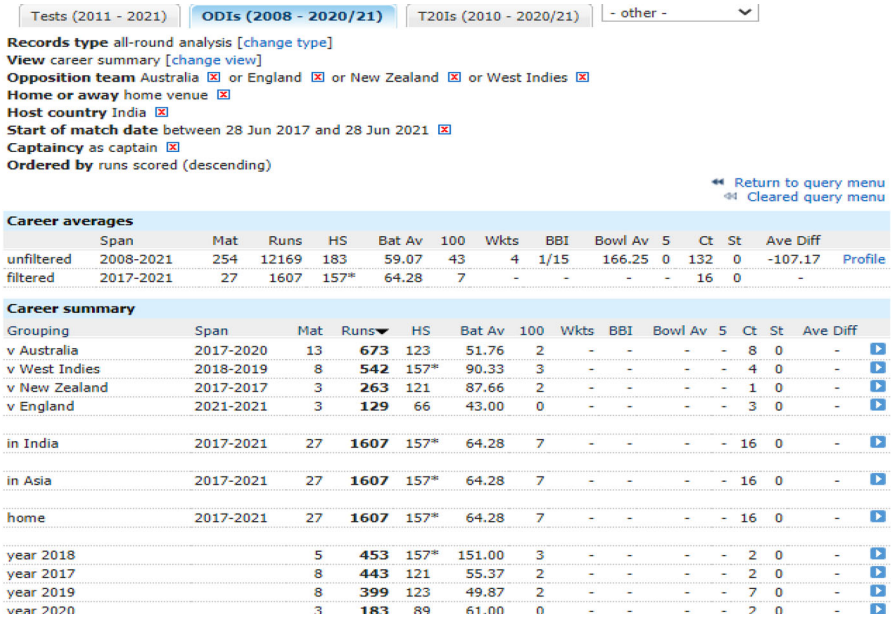

Fig. 5 Data in HTML Table format *Source*: https://stats.espncricinfo.com/

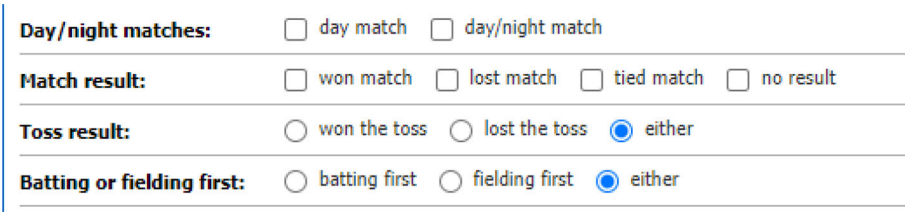

Fig. 6 Features for machine learning model as filter parameters *Source* https://stats.espncricinfo.com/

For the machine learning model, the overall aggregate records of all player matches from 2016–2020 for each match format (Test, ODI, and T20) are web scraped to form the training set illustrated below in Table 2. Since, the additional features illustrated above are filters, the filters are applied to retrieve different datasets sorted by runs scored and subsequently merged into a single dataset creating dummy variables for the above filters.

The Outcome is considered as the dependent variable in line with the objective of the paper which is to predict the match outcome from the above parameters, hence data for the above variables are collected interactively during this period. Outcomes pertaining to draw or tie are excluded from analysis for lack of relevance. The statistics are restricted to the ESPN Cricinfo website due to data availability in a structured format (both player-wise and match-wise extraction of data possible through this data source, satisfying the objectives of the paper).

Thus, the cricket dataset is constructed with 3000 ESPN match records; a snippet of the dataset is illustrated in Table 2.

Table 2 presents a snippet of the scraped player dataset. For instance, the Player Tim Southee (TG Southee) [Row number 7] has played two matches in the year of 2020 scored an aggregate of 1 run from both matches, has taken total of 5 wickets, maintained a Bowling Average of 62, took 1 important catch. New Zealand won the toss (Southee is NZ player), played at home venue (Match location = 1) during the day (Match time = 1), batting first (Batting position = 1) and won the match (Outcome = 1).

The outcome variable in this study is 'Outcome' used to quantify the team performance. The outcome of the match result is predicted for a completely new validation set of chosen players and based on respective parameters. The methodology adopted in the paper is illustrated below:

3.2 Methodology adopted

This paper examines the role of factors determining the match outcome from different match-wise parameters scraped from ESPN Cricinfo Statsguru. The factor variables are selected based on their decision-making significance in the game of Cricket. This objective is accomplished in two steps. The first step is to perform predictive modeling using ML on the dataset considering Outcome (match outcome) as the independent variable (to be forecasted) and the above-mentioned quantitative and categorical factors, as illustrated in Table 3. The models are compared in terms of accuracy and the importance of each variable, also known as relative importance. This relative importance of a variable generated from the predictive models computes the extent to which each predictor variable is significant and generated by these predictive models and thus, defines a method to compute 'Outcome' from the above predictors. Predictive modeling techniques can be broadly categorized into supervised and unsupervised techniques (Loureiro et al., 2018). In this paper, the predictive techniques, i.e., Random Forest model, Gradient Boosting, and more complex ML-based models, like deep neural networks, have been adopted to derive the relative significance of factors and improve the extent to which prediction can be made in terms of accuracy.

Further, the outcome variable predicted from the above ML techniques is tuned to different user-generated scenarios, and match result is predicted based on the parameters defined in Table 3.

Table 2 A snippet of the match ODI dataset scraped from ESPN CricInfo

Player	Mat	Runs	Wkts	Bowl Av	Catches	Toss decision	Match venue	Match time	Batting Position	Outcome
HK Bennett	1	20	1	20	0	1	0	1	1	1
MS Chapman	1	27	0	0	0	1	1	1	1	1
LH Ferguson	2	53	4	12.5	0	1	1	1	1	1
DJ Mitchell	1	55	0	0	0	1	0	1	1	1
IS Sodhi	2	68	5	11.6	2	1	1	1	1	1
BM Tickner	1	92	2	12.5	0	1	1	1	1	1
TG Southee	2	1	5	62	1	1	1	1	1	1
C Munro	1	6	0	0	2	1	1	1	1	1
MJ Santner	1	15	1	41	0	1	1	1	1	1

Table 3 Predictive Accuracy of the Random Forest Model for ODI and T20 Match Result Prediction

mtry	ntree	R-Squared for ODI	R-squared for T20
2	200	0.864	0.8926
4	200	0.865	0.8934
6	200	0.863	0.8941

3.2.1 Building the predictive models

Considering existing studies (Bose et al., 2021; Deval et al., 2021; Kamble, 2021) that build match outcome prediction models and provide a consistent output that can handle multiple features with high predictive accuracy, ML techniques are adopted paper.

The model building phase starts by simulating highly sophisticated and layered ML models, such as random forest, gradient boost, and deep neural network to compare the prediction accuracy of match outcome across different ML models, one simple random forest, an ensemble gradient boost algorithm, and multiple layered deep neural networks.

The open-source data analytics tool R (Jiang & Chen, 2019) was adopted to build the above three models is built. The methodology undertaken in the paper is illustrated in Fig. 7 and outlined by the steps given below:

Step 1 Exploratory data analysis is performed on the dataset initially on Aaron Finch for performance in Tests and further, by comparing the performance of four captains Aaron Finch, Virat Kohli, Kane Williamson, and Joe Root in terms of runs and batting average under different criteria like the toss, match time conditions, batting position and nature of the tournament. The comparison is performed for ODIs and T20.
Step 2 Perform tenfold cross-validation (Schneider and Gupta, 2016) for implementing ML using the 'trControl' function under predefined R package 'caret'.
Step 2.1 Build ML techniques (Random Forest, Gradient Boosting and Deep Neural networks) with the seven input variables for prediction. The prediction is performed separately for ODI matches and T20 matches data. The R tool is used for machine learning, which has a predefined package 'caret' for implementing the models.
Step 2.2 The prediction accuracy and performance on 5% of the real data-points (150 data-points) from the ESPN Cricinfo website for the Indian players Hardik Pandya, Shreyas Iyer, Ravindra Jadeja, Rishabh Pant, and Virat Kohli are then compared to determine which of the three models (Random Forest, Gradient Boosting and Deep Neural networks) outperforms the others in match outcome prediction. The objective of this exercise is to identify the most emerging Indian players based on different parameters. The relative importance of the variables is also compared.
Step 2.3 Further, to validate the relative importance of variables and their interactions, a multiple linear regression model is formulated from the four quantitative predictors and four external condition-based variables, and the interplay between variables is further factored in to test for interaction effects (27 new pair-wise interaction and 17 triplet interaction variables). The total number of predictors is, therefore, 52 (= 27 + 17+8) however, the most significant predictors are considered for the final regression model 3.
Step 3 The players' performance is quantified in terms of the win ratio (number of matches won to number of matches in total) and are clustered into four grades (A + , A, B, and C recoded as 1,2,3 and 4). The grades are compared with the actual grades of the players in real-time to validate the efficacy of the model.

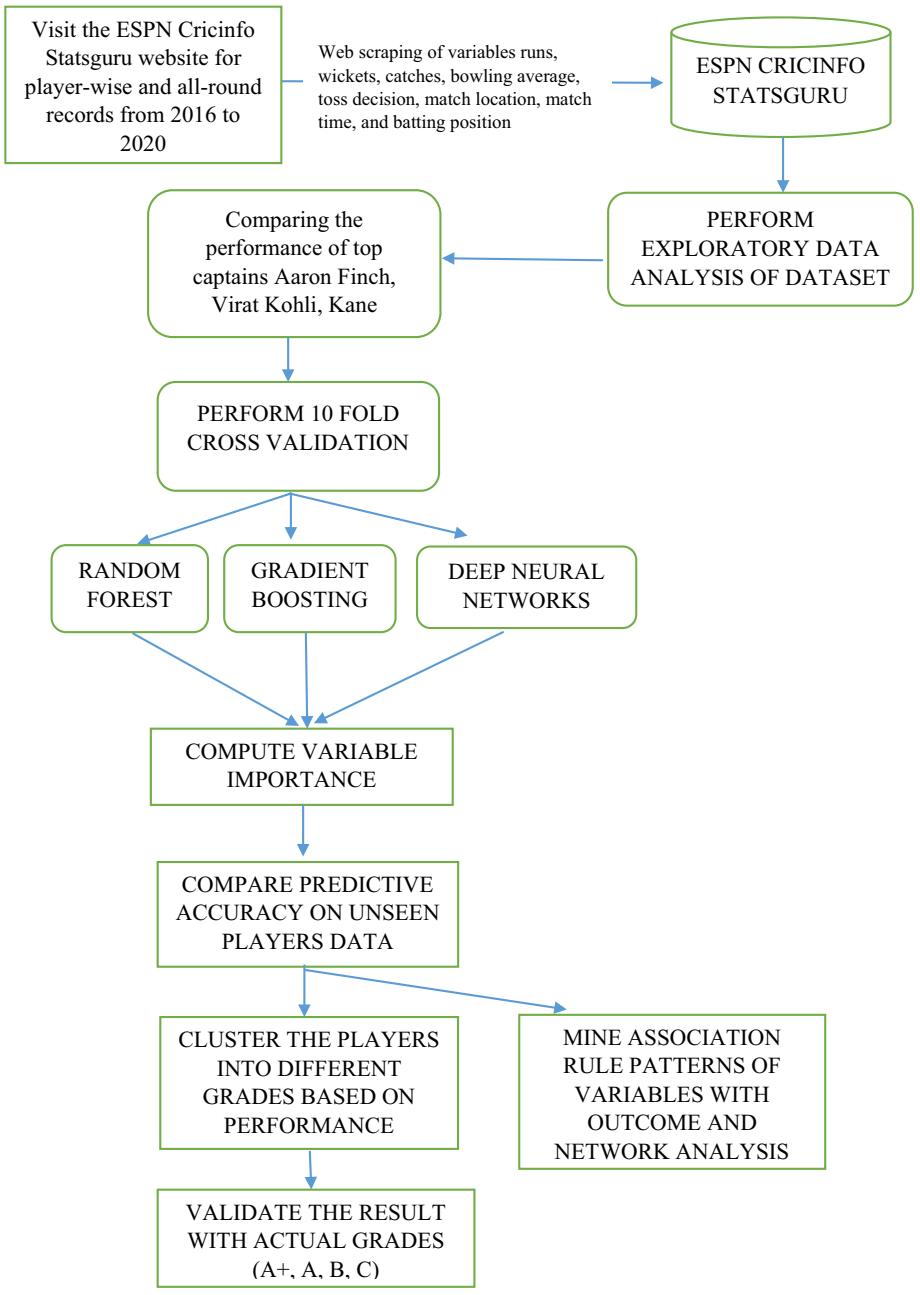

Fig. 7 The methodology undertaken for the study

Step 4 Association rule patterns are generated from the dataset for match result prediction. Further, network analysis of player performance in different countries for T20s and ODIs are also illustrated. For measuring the connectedness, the mutual spillover effects of the major cricket-playing countries Australia, New Zealand, England, and India were estimated using the Vector Auto-regressive model (VAR) for both the different periods.

Based on the spillover values, the connectedness graph or adjacency graph (minimum spanning tree) was plotted between the different countries for ODIs and T20s using the 'frequencyConnectedness' package in R. The results of the spillover matrix and the minimum spanning tree graphs are thus illustrated below in subsection 4.4.

Figure 8 illustrates the working procedure of the ML predictive models.

Random Forest-based predictive model (Bendazzoli et al., 2019): A random forest is a supervised machine learning classifier that combines the output of several Decision trees using a voting algorithm and predicts the outcome resulting from the aggregated outcomes. They are easy to implement with cleaner output and fit on a large set of data. For this model,

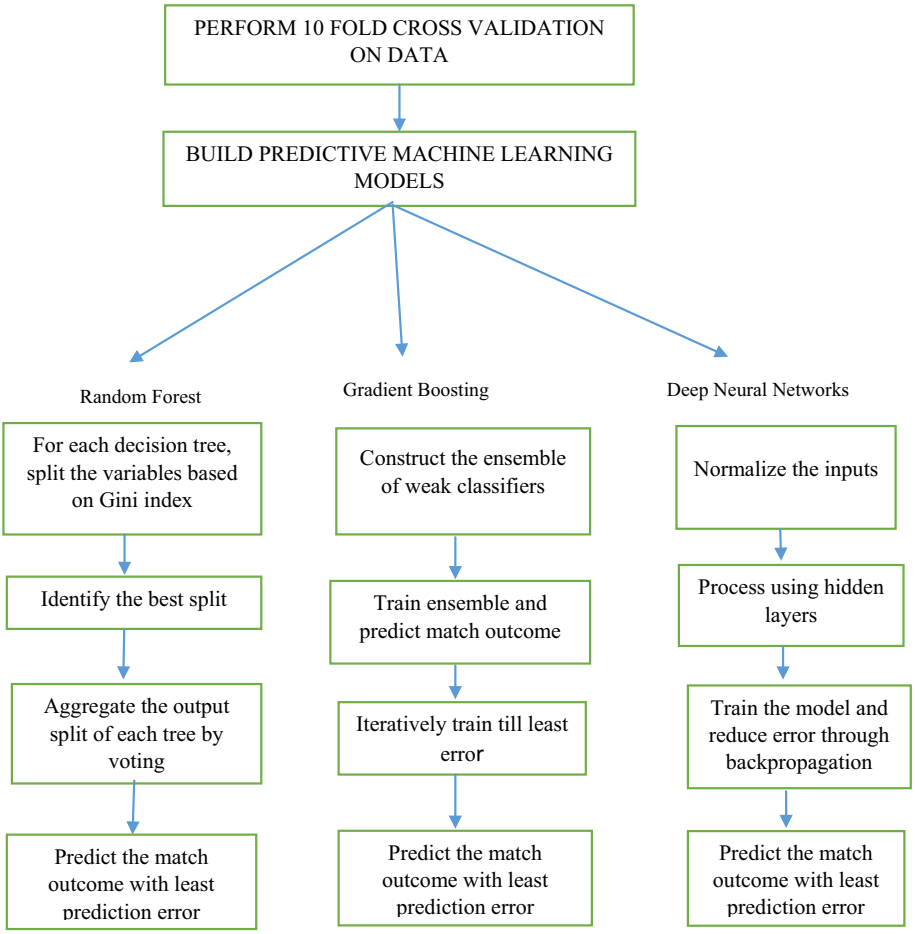

Fig. 8 Working of the ML predictive models

all predictors are converted to numeric, and the target variable, *i.e.*, Outcome, is predicted. Random Forest can be implemented by randomForest () predefined library in R tool with the following syntax:

$$model_{RF} = random\ Forest(out \sim x_1 + x_2 + \ldots + x_n, data = trset, ntree) \quad (1)$$

where out is the outcome variable; $x_1, x_2, \ldots, x_n$. represent the '*n*' number of inputs/predictors considered for the model, *trset* is the training set input to the model and '*ntree*' represents size of regression/decision tree.

Gradient Boosting predictive model (Hubáček et al., 2019) The gradient boost ensemble model is also run to predict 'Outcome' to boost the predictive accuracy and interpretability. The boosting technique can be modeled as an optimization problem where the objective is to gradually and iteratively minimize the ensemble model's eor rate and iteratively using a gradient descent-like procedure. A weak algorithm like a decision tree is combined with a robust predictive model to form an ensemble that boosts the predictive auracy. They help to deal with an unbalanced and large set of data to provide accurate results. To boost the predictive accuracy and interpretability, the gradient boost ensemble model is also run for the prediction of 'Outcome'. All predictors are converted to numeric for this model, and the output variable, *i.e.*, 'Outcome', is predicted.

The caret package in R implements gradient boost on the dataset with syntax:

$$train\big(out \sim x_1 + x_2 + \ldots + x_n,\ data = trset,\ method = ''gbm'',\ trControl = n(folds)\big) \quad (2)$$

where; out is the outcome variable; $x_1, x_2, \ldots, x_n$ represent the '*n*' number of inputs/predictors considered for the model, *trset* is the training set input to the mode; '*gbm*' stands for Gradient Boosting Machines methodology in this case and t*rControl* is the control parameter used to set number of cross-validation folds [*n*(*folds*)] (mostly 10).

Deep neural network-based model A Deep Neural Network (DNN) model is simulated based on the human brain's working. A typical architecture ilayer-wise: the input layer takes the normalized input data, and the last layer provides the output. In between, the processing of inputs is performed in hidden layers (one or more), which process the input values and compute an activation function (preferably sigmoid) based on the importance of variables. A DNN is trained to learn the input weights and consequently generate the output incrementally. DNNs effectively handle a large number of multiple input variables, for example, in big data scenarios.

The caret package in R implements deep neural networks on the dataset with syntax:

$$train\big(out \sim x_1 + x_2 + \ldots + x_n;,\ data = trset\ ,\ method = ''nnet'',\ trControl = n(fold)\ \big) \quad (3)$$

where; out is the outcome variable; $x_1, x_2, \ldots, x_n$. represent the '*n*' number of inputs/predictors considered for the model, trset is the training set input to the mode; '*nnet*' stands for Artificial Neural Networks in this case and *trControl* is the control parameter used to set number of cross-validation folds [*n*(*folds*)] (mostly 10).

Further, the variable importance of the predictors for the above machine learning models is computed by the R tool using the *varImp*() function in the predefed 'caret' package is used to compute variable importance with the syntax: $varImp(model\_trained)$; where *model_trained* represents the machine learning algorithm adopted.

In this paper, the DNN has been constructed with five hidden layers (as illustrated in Figure Eleven) due to the minimum root mean square error (RMSE) of 0.25, which also minimizes

the probability of over-fit of the DNN model, *i.e*., the model fitting only on some data points and under-performing on other data points in the dataset (Loureiro et al., 2018). The number of input nodes is 8, considering eight individual predictors.

Further, the variable importance of the predictors for the above machine learning models is computed by the R tool using the *varImp*() function in the predefined 'caret' package is used to compute variable importance with the syntax: *varImp(model_trained)*; where model_trained represents the machine learning algorithm adopted.

Multiple linear regression-based predictive model (Kong et al., 2019; Cappelli et al., 2019): A multiple regression model that factors in all the eight predictors (four quantitative and four external) and twenty-six interactions is formulated below.

Three regression models were implemented (first is baseline model with variables, second includes pair-wise interactions and the third model also incorporates triplet interactions).

The equation for implementing the baseline multiple linear regression is formulated as follows:

$$out_1 = \beta_1 z_1 + \beta_2 z_2 + \ldots + \beta_m z_m + \varepsilon\prime_1 \tag{4}$$

where out_1 is the outcome variable [match outcome]; $z_1, z_2, \ldots, z_m$. represent the 'm' number of inputs/predictors considered for the model representing the main (effect) variables Runs scored, wickets taken, Bowling Average, number of catches taken, Toss decision, Match venue/location, Match time, and Batting Position. $\beta_1, \beta_2, \ldots, \beta_m$. represent the regression coefficients which signify the sensitivity of each variable to the overall output. $\dot{\varepsilon}_1$. is the residual of the regression, which implies the component of the match result outcome variable which is unexplained by the predictors.

Further, the second regression model implementing the pair-wisinteractions is constructed as follows:

$$out_2 = \beta_1 z_1 + \beta_2 z_2 + \ldots + \beta_m z_m + \beta_1' z_1 z_2 + \beta_2' z_2 z_3. + \cdots + \beta_{m-1}' z_{m-1} z_m. + \cdots + \dot{\varepsilon}_1 \tag{5}$$

where $\beta_1' + \beta_2', \ldots, \beta_{m-1}'$ represent the regression coefficients, which signify the sensitivity of each interaction variable to the overall output. $\dot{\varepsilon_3}$ is the residual of the regression, which implies the component of the match result outcome variable that is unexplained by the predictors (main and pair-wise interaction terms).

The interaction terms $z_1 z_2, z_2 z_3$ $\ldots\ldots z_{m-1} z_m$ represent the pair-wise interactions like Runs scored* wickets taken, Bowling Average*number of catches taken. Interaction variables between Bowling Average and wickets taken are not considered due to multi-collinearity (similar variables). The interactions of quantitative factors like runs, wickets and bowling average with external match conditions like Toss decision, Match venue/location, Match time, and Batting Position are factored in the model. For instance, Runs*Toss decision and Wickets*Match time are also considered in the model.

The third regression model incorporating the triplet interactions is constructed as follows:

$$\begin{aligned} out_3 = {} & \beta_1 z_1 + \beta_2 z_2 + \ldots + \beta_m z_m + \beta_1^{'} z_1 z_2 + \beta_2^{'} z_2 z_3 + \ldots \beta_{m-1}^{'} z_{m-1} z_m \\ & + \beta_1^{''} z_1 z_2 z_3 + \beta_2^{''} z_2 z_3 z_4 + \ldots + \beta_{m-2}^{''} z_{m-2} z_{m-1} z_m \ldots + \dot{\varepsilon}_3 \end{aligned} \tag{6}$$

where $\beta_1^{''}, \beta_2^{''}, \ldots, \beta_{m-2}^{''}$ represent the regression coefficients which signify the sensitivity of each triplet interaction variable to the overall output. $\dot{\varepsilon_3}$ is the residual of the regression, which implies the component of the match result outcome variable which is unexplained by the predictors (main, pair-wise interaction and triplet interaction terms).

The interaction terms $z_1 z_2 z_3 \ldots\ldots z_{m-2} z_{m-1} z_m$ represent the triplet interactions like Number of runs x Bowling Average x Toss decision derived from grouping the most significant pair-wise interactions.

Clustering (D'Urso et al., 2019, 2020, 2021; de Zepeda et al., 2021): Clustering is used to validate the efficacy of the machine learning algorithms prediction result by clustering the players into different grades based on their contribution to the win ratio of the teams. The unseen validation data points for each of the players based on different parameter combinations is input to the machine learning algorithms to predict the match outcome in each case. The match outcome is analyzed and aggregated for each player in terms of the win ratio (number of matches won to total number of matches). This win ratio is used for clustering of the players into different grades. The predicted grades are then compared with the real grades allocated to the players by the respective cricket board. The closer the actual grades are with the predicted, the more accurate and valid is the machine learning prediction model.

Association mining (Huang et al., 2021) Association rules were formulated just to derive some patterns from the data and are not incorporated into the final model.

The measures adopted in association rules are:

- *Support* This is a measure that defines the probability of occurrence of a pattern in a transaction mathematically represented by: $p(XUY)/p(TotalTransactions)$.; where X is the LHS predictor variable and Y is the RHS outcome variable for pattern analysis.
- *Confidence* The conditional probability of Y. occurring subject to X. is confidence represented by $p(Y|X)$..
- *Lift* The ratio of confidence to expected confidence is lift, values of lift > 1 generate meaningful association patterns mathematically represented $p(Y|X)/E(p(Y|X))$.

Tain purpose of generating association rules is to determine a suitable combination of the different factors (optimal values) at which a match winning result can be achieved. The association rules generated for both ODI and T20s present an insight into how parameters can be tuned to achieve the desired outcome.

4 Analysis and results

The results of the initial exploratory data analysis performed on the four captains Aaron Finch, Kane Williamson, Joe Root and Virat Kohli are first illustrated in Sect. 4.1. Further, the machine learning model results (Random Forest, Gradient Boost and Deep Neural Network) comparing the predictive accuracy and variable importance are illustrated in Sect. 4.2. The clustering of the emerging Indian players into different grades and comparison with real-time allocated grades for model efficacy is illustrated in 4.3. The association rules and network analysis results are illustrated in Sect. 4.4.

4.1 Exploratory data analysis and comparison of best captains

The data was collected and scraped from the ESPN Cricinfo Statguru website as illustrated above in 3.1.1. Initially, an exploratory comparative analysis of Australian captain Aaron Finch with three other prominent international captains Kohli, Williamson and Root are considered for the exploratory data analysis, and their performance is compared for ODIs and T20s.

4.1.1 Exploratory comparative analysis of top four captains in ODIs and T20s

The top four captains are compared in terms of performance first in one-day internationals (ODIs) and then in twenty-twenty match formats (T20s). The four captains Aaron Finch, Kane Williamson, Joe Root, and Virat Kohli are denoted in the below graphs as AF, KW, JR, and VK for readability of labels, and their batting average and runs scored are compared under different conditions as follows:

In Fig. 9, with respect to the batting average, Virat Kohli is found to be the highest of the four batsmen with batting average of 80–140.

Virat Kohli is successful the most in matches where India lost the toss and was put in to bat first illustrating the chasing capability of the Indian team. India winning the toss and fielding first was least effective. Similar results are found also for Aaron Finch. For Kane Williamson, however, the decision to win the toss and bat first was found to be the second most effective strategy after "lost toss and batted first" decision.

Aaron Finch is found effective when fielding first while Virat Kohli when batting first in Fig. 10.

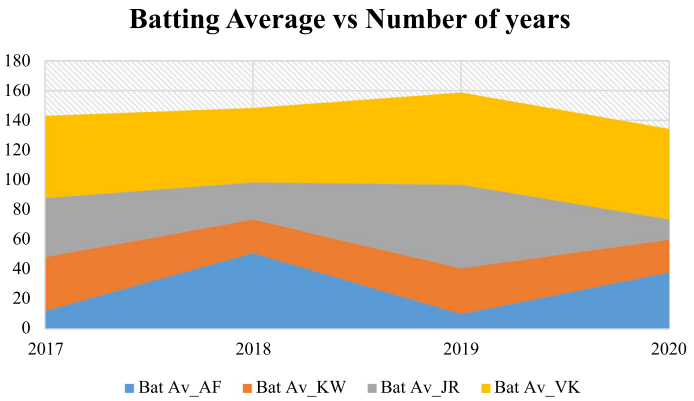

Fig. 9 Comparison of batting average scored over the years

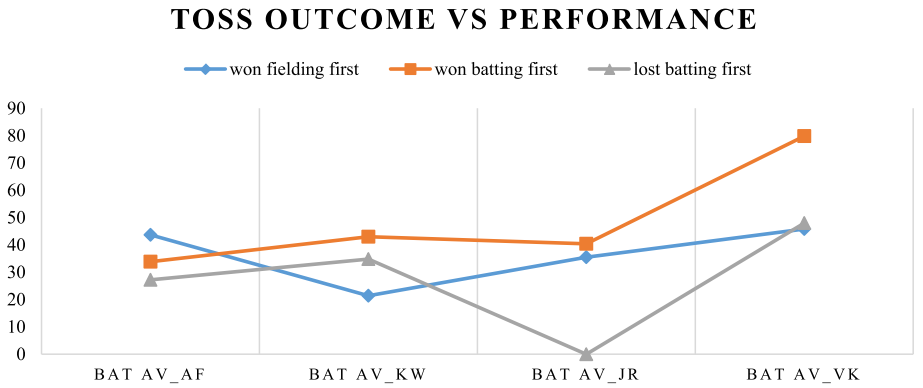

Fig. 10 Comparison of batting average for toss outcome

Virat Kohli performs the best in the third ODI match of the ODI series while Aaron Finch is most effective in the first match and Williamson and Root in the second match in Fig. 11.

Similarly, for T20 matches, the following insights were found:

Kohli again is found to be successful in scoring runs as illustrated in Fig. 12 when India loses the toss and bats first in T20s similar to ODI matches while Finch is effective when Australia wins the toss and fields first. Williamson succeeds when New Zealand wins the toss and fields first.

For Kohli, Finch, and Root, batting average is directly proportional to match victory except for Williamson who demonstrated the highest average > 80 in a drawn match(tied) result in Fig. 13.

The results of the machine learning prediction and variable importance results are illustrated below in Sect. 4.2:

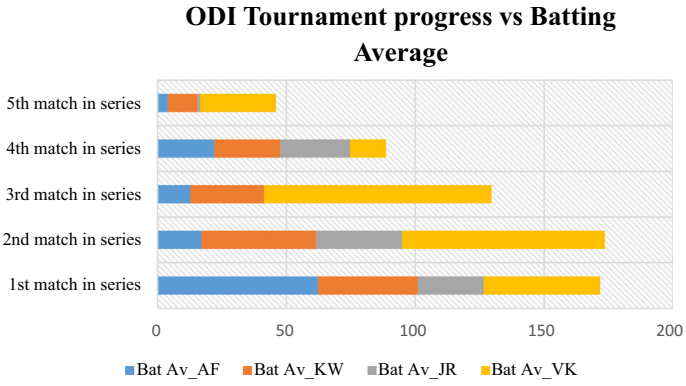

Fig. 11 Comparison of tournament progress with batting average

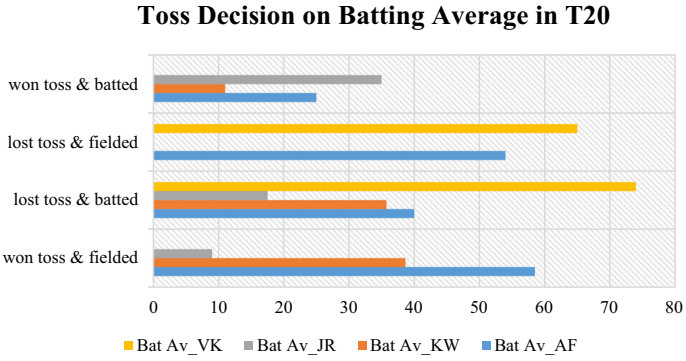

Fig. 12 Comparison of toss decision on batting average

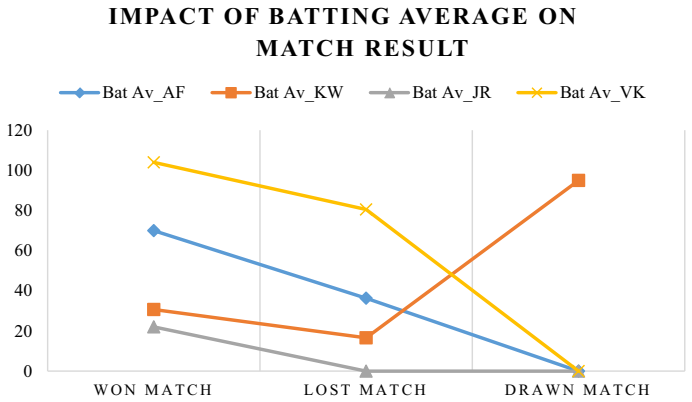

Fig. 13 Comparison of batting average on match outcome

4.2 Results of machine learning prediction models for ODI and T20

4.2.1 Analysis of random forest-based model for ODI and T20

Outcomes of the random forest-based model generated using the 'random Forest' package of the R tool are displayed in Tables 3 and 4. The training set comprises 2400 data points, while the test set had 600 data points (total 3000 instances). The parameters, i.e., the number of predictors 'mtry' and the number of trees 'ntree' are tuned to choose the best model. A sample of the dataset is illustrated in Table 4:

The number of optimal predictors is considered to be $n/3.$, where $n.$ is the number of variables considered in the model (Adam et al., 2014) while mtry values vary from 1 to $n-1$, *i.e.*, 6 in this case for seven predictors. This is to be tested for which the RMSE values are plotted against the number of predictors 'mtry' and tuned by changing the number of trees 'ntree'.

Table 4 shows that the optimal accuracy of this model's prediction for ODI matches is 86.5% at mtry $= 2$ and ntree $= 200$ while for T20 matches, the accuracy i89.41%, which implies that 89.41% of the test set's data points were accurately classified.

Fther, the weightage assigned to the predictors for the Random Forest model is tabulated in Fig. 14:

It is observed from Fig. 14 that in ODI matches, the reputation of the player is the most significant predictor, followed by runs scored, toss decision, match location(home/away), the number of catches, and bowling average. Further, match time (day or day/night) and wickets are less important.

Table 4 Predictive Accuracy of the Gradient Boosting Model for ODI and T20 Result

n_trees	R-squared for ODI	R-Squared for T20
50	0.8941	0.887
100	0.8922	0.895
150	0.8921	0.897

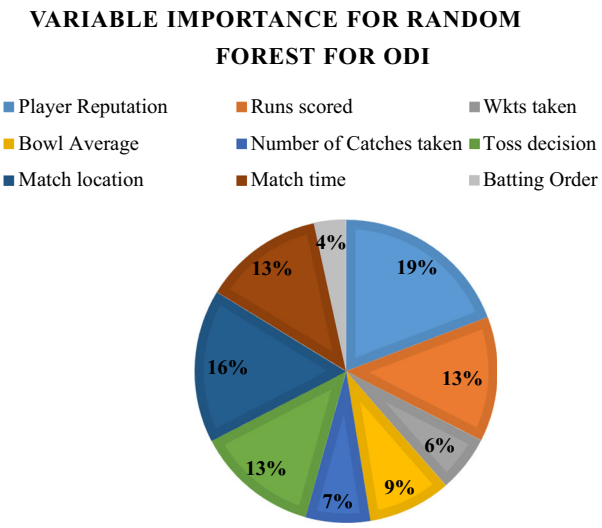

Fig. 14 Variable Importance in the Random Forest model for ODI Match prediction

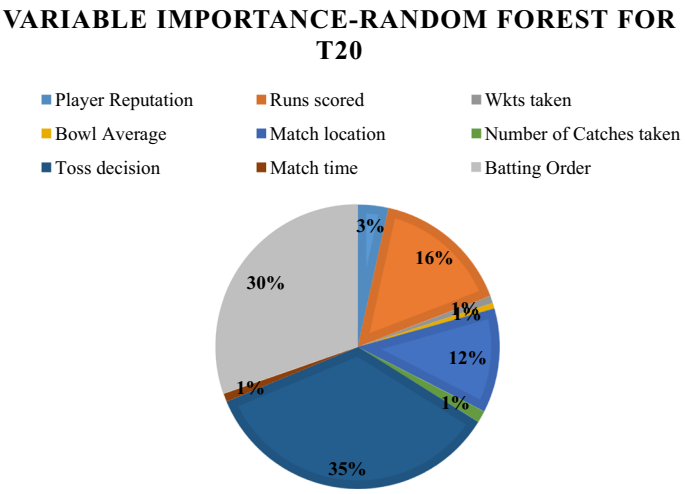

Fig. 15 Variable Importance in the Random Forest model for T20 Match prediction

It is observed from Fig. 15 that in T20 matches, the Batting order (first or second) by a team is the most significant predictor, followed by toss decision, runs scored, match location(home/away), number of catches and bowling average. Further, match time (day or day/night) and wickets are less important.

4.2.2 Analysis of gradient boosting model for ODI and T20

From Table 4, the optimal accuracy of the prediction of this model for ODI matches is 89.41%, which implies that 89.41% of the data points of the test set were accurately classified. The ideal size of the classification tree is n_tree = 50. For T20 matches, the optimal predictive

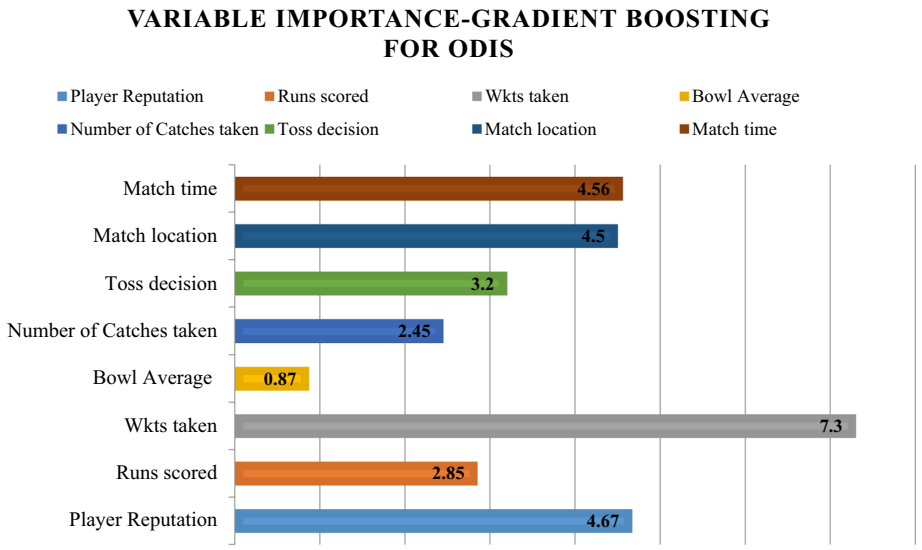

Fig. 16 Relative Importance of the variables in the Gradient Boosting Model for ODI Match Result

accuracy is 89.7% for optimal tree size n_tree = 150. This is an improvement over the Random Forest model due to an ensemble of techniques and aggregation of output from multiple decision trees.

Further, the weightage assigned to the predictors for the Gradient Boost model is illustrated in Fig. 16.

It is observed from Fig. 16 that Wickets is the most significant predictor, followed by the player reputation, match location (home/away) and match time (day or day/night). Toss decision and number of catches are the next most important factors in determining match outcome.

It is observed from Fig. 17 that Batting order (first or second) is the most significant predictor, followed by the match location (home/away), toss decision and number of runs scored. Number of catches are the next most important factors in determining match outcome.

4.2.3 Analysis of deep neural network-based model for ODI and T20

The DNN based model performance is illustrated in Fig. 18 below. A simple model showed 70.9% accuracy, but an ensemble of 100 such simple DNN models boosted the accuracy to 96% at tenfold cross validation. Cross-validation is performed to efficiently validate the performance of the designed model. It is a statistical procedure to estimate the classification ability of learning models.

Figure 18 plots the variation of the deep neural network model's performance with a number of hidden layers for processing adopted. Based on Fig. 18, the optimal weight decay of 0.1 and the optimal cross-validation RMSE is attained at hidden layer = 3. Three hidden layers have been adopted in the deep neural network model. Further, the weightage assigned to the predictors is shown in Fig. 19.

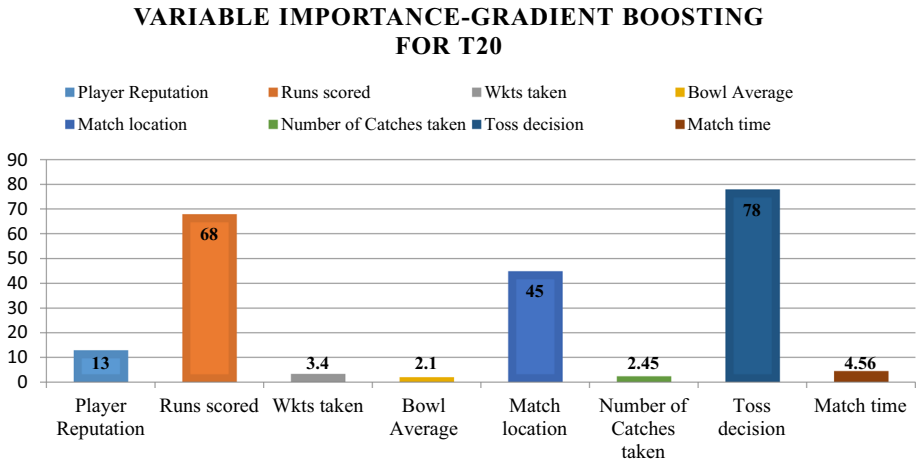

Fig. 17 Relative Importance of the variables in the Gradient Boosting Model for T20 Match Result

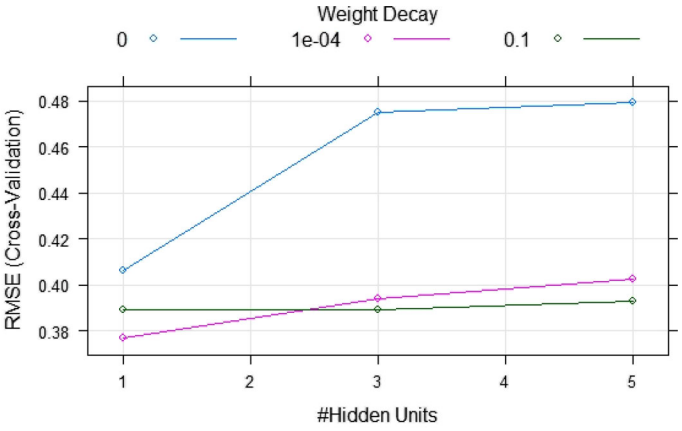

Fig. 18 Neural Network RMSE v/s Weight Decay Curve for ODI Results

It is observed from Fig. 19 that Player reputation, runs scored, match location (home/away) and Match time (day and day/night) are the most significant predictors, followed by number of catches and runs scored. Similarly, for T20 matches, the results are as follows in Fig. 20:

Figure 21 plots the variation of the deep neural network model's performance with a number of hidden layers for processing adopted for T20 results. Based on Fig. 21, the optimal weight decay of 0.1 and the minimum cross-validation RMSE is attained at hidden layer = 5. Five hidden layers have been adopted in the deep neural network model.

It is observed from Fig. 22 that Batting order (first/second), Match location (home/away), toss decision, and runs scored are the most significant predictors, followed by player reputation.

Therefore, across all the machine learning models, the most significant factors impacting ODI match outcome are Player reputation and Match time (day and day/night), followed by the number of catches and runs scored. The vital factors for T20 match outcome are

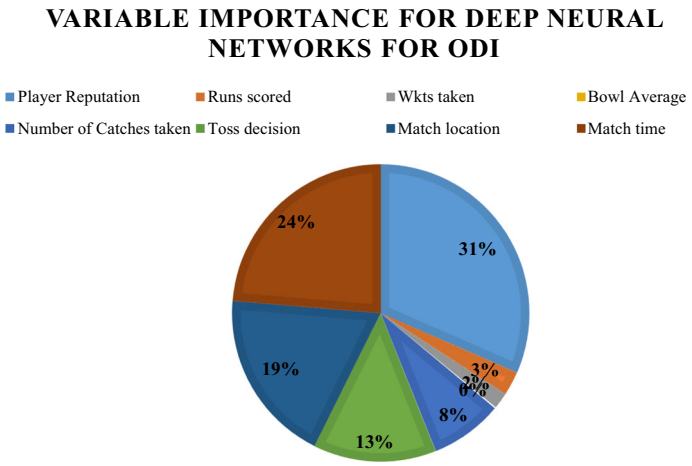

Fig. 19 Relative importance plot of Deep Neural Networks model for ODI Match

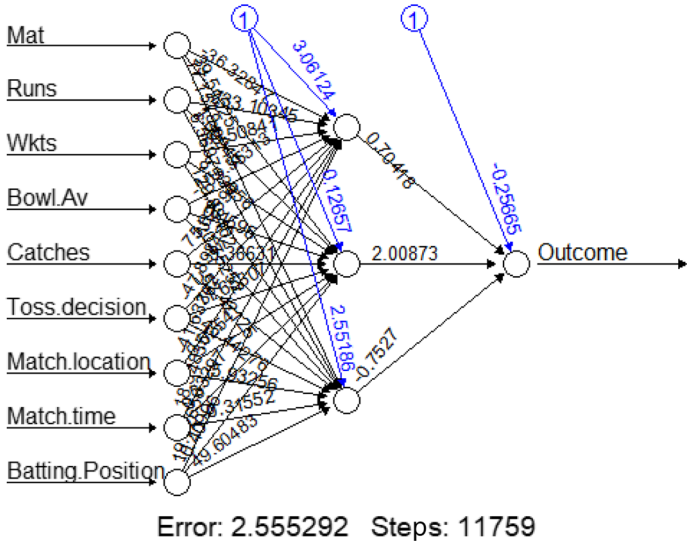

Fig. 20 Neural Network model output for T20 Match Result Prediction

Batting order(first/second), toss decision, and runs scored are the most significant predictors, followed by player reputation.

Then, 5% of the dataset (i.e., 150) with unseen new real-life data points considered from the ESPN Cricinfo Statsguru website has been considered for comparing the models' performance on these new data points for validation, and the result is illustrated in Fig. 23.

Deep Neural networks are the most accurate predictor of ODI match result (95%) [green color], followed by Gradient Boosting algorithm (78%) [yellow color] and Random Forest (60%) [red color].

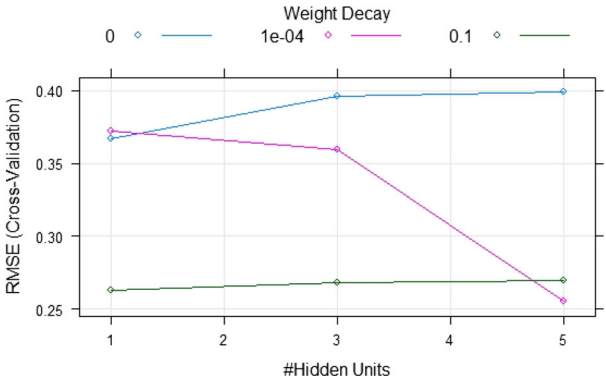

Fig. 21 Neural Network RMSE v/s Weight Decay Curve for T20 Results

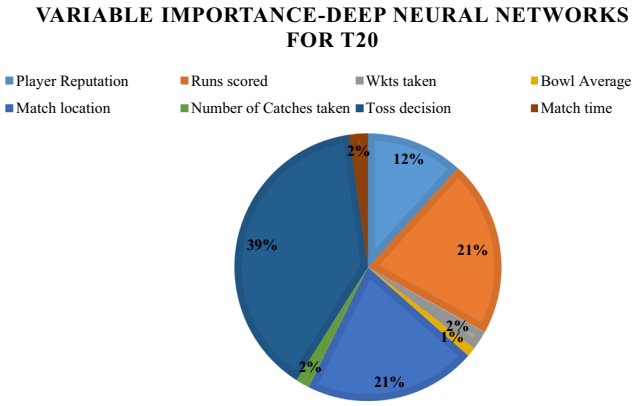

Fig. 22 Relative importance plot of Deep Neural Networks model for T20 Match

From Fig. 24, it can be inferred that the Deep Neural Network model (90%) [yellow] predicts the rank closest to the actual Outcome, followed by Gradient Boost (70%) [blue] and then the Random Forest model (50%) [green].

4.3 Feature interaction effects using regression

We discovered the feature's effect through the random forest, gradient boost, and Artificial Neural Networks (ANN) summarized in Figs. 14, 15, 16, 17, 19 and 22. Next, we attempt to understand the impact of interactions of the features like external match factors. Hence, multiple regression model including single, pairwise, triplet and quadruplet interactions [Regression Model 3] has been performed to predict the match outcome based on the features and their interactions under study.

The regression models formulated in Tables 5 and 6 ensure that all the robustness tests for the assumption of linear regression, namely multi-collinearity, linearity, auto-correlation, and homoskedasticity are validated (Abadie et al., 2020). For the match instances, the robustness tests, namely Durbin-Watson, Langrange Multiplier (LM Coefficient), and Variance Inflation Factor (VIF) are run to ensure the reliability of the model variable significance.

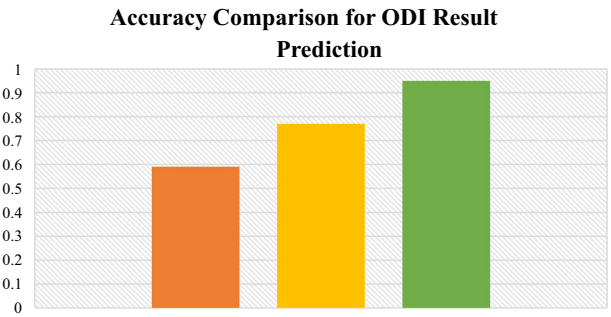

Fig. 23 Performance Comparison for ODI Match result prediction

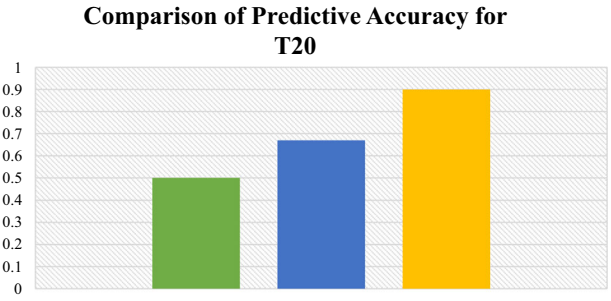

Fig. 24 Performance Comparison for T20 Match result prediction

According to the Durbin Watson test (Lumbantobing et al., 2020), the value of the Durbin Watson statistic (DW) must lie between 2 and 4 with a value tending closer to 2 implying that auto-correlation is not present in the dataset. Moreover, the significance value rho must be closer to 0. Similarly, for Langrange Multiplier (LM) test (Chauhan et al., 2020), if the p-value statistic is greater than level of significance 'alpha', the null hypothesis of homoscedasticity is validated. The VIF (Variance Inflation Factor) (Vörösmarty & Dobos, 2020) for all the predictors is expected to be < 10 to indicate that there is no multi-collinearity in the data.

Further, the tests are summarized across all the three models implemented each for ODI and T20 match instances, namely **Model 1**(which only contains individual variables), **Model 2**(which includes pair-wise interactions), and **Model 3**(which includes even triplet interactions and significant quadruplet interactions).

The unstandardized coefficients (original) and p-value statistics (in parenthesis) and above robustness statistics (DW, LM and VIF) for ODI and T20 instances are reported. *** indicates a 1% statistical significance level.

In Table 5, three regression models are implemented: Model 1 is implemented by regressing the outcome variable (Match outcome) on the individual predictors (direct effects) considered in the machine learning model. At 95% significance level, Runs, Wickets, Toss decision, Match location and Batting position are the most significant factors considered by for winning a match. The value of adjusted R-squared is 0.65 i.e., 65% is explained. Further, in Model 1, the Durbin Watson Statistic (DW) is reported to be 2.42 with a *p*-value of 0.15. The

Table 5 Summary of regression model results for ODI Dataset

Variables	Model 1	Model 2	Model3	VIF
Runs	0.13^{***}(0)	0.13^{***}(0)	0.13^{***}(0)	1.13
Wickets	0.06^{***}(0.02)	0.06^{***}(0.02)	0.06^{***}(0.02)	1.02
Catches	−0.03 (0.8)	−0.03 (0.8)	−0.03 (0.8)	1.52
Bowling Average	0.09^{***}(0.03)	0.09^{***}(0.03)	0.09^{***}(0.03)	1.07
Toss decision	0.19^{***}(0)	0.19^{***}(0)	0.19^{***}(0)	1.73
Match location	0.04^{***}(0.005)	0.04^{***}(0.005)	0.04^{***}(0.005)	1.96
Match time	-0.10^{***}(0.02)	-0.10^{***}(0.02)	-0.10^{***}(0.02)	1.01
Batting position	0.07^{***}(0)	0.07^{***}(0)	0.07^{***}(0)	1.25
Runs*Wickets		0.024^{***}(0)	0.024^{***}(0)	1.03
Runs*Catches		0.043 (0.75)	0.043 (0.75)	2.56
Runs*Bowling Average		0.034 (0.57)	0.034 (0.57)	2.08
Runs*Toss decision		0.056 (0.134)	0.056 (0.134)	5.69
Runs*Match location		-0.065^{***}(0)	-0.065^{***}(0)	4.29
Runs*Match time		0.065 (0.52)	0.065 (0.52)	1.2
Runs*Batting position		-0.074^{***}(0.003)	-0.074^{***}(0.003)	1.65
Wickets*Catches		-0.012^{***}(0)	-0.012^{***}(0)	1.44
Wickets*Toss decision		−0.032 (0.75)	−0.032 (0.75)	2.34
Wickets *Match location		0.034 (0.57)	0.034 (0.57)	1.03
Wickets *Match time		0.062 (0.134)	0.062 (0.134)	2.56
Wickets *Batting position		0.046^{***}(0)	0.046^{***}(0)	2.08
Catches* Bowling Average		0.065 (0.52)	0.065 (0.52)	5.69
Catches *Toss decision		0.004^{***}(0.003)	0.004^{***}(0.003)	4.29
Catches *Match location		0.04^{***}(0)	0.04^{***}(0)	1.2
Catches *Match time		−0.146 (0.524)	−0.146 (0.524)	1.65
Catches *Batting position		-0.092^{***} (0.005)	-0.092^{***} (0.005)	1.44
Bowling Average* Toss decision		0.153 (0.229)	0.153 (0.229)	5.41
Bowling Average* Match location		-0.012^{***}(0)	-0.012^{***}(0)	5.8
Bowling Average* Match time		−0.032 (0.75)	−0.032 (0.75)	1.81
Bowling Average* Batting position		0.034 (0.57)	0.034 (0.57)	2.02
Toss decision* Match location		0.062 (0.134)	0.062 (0.134)	5.63
Toss decision * Match time		0.046^{***}(0)	0.046^{***}(0)	4.23
Toss decision * Batting position		0.003^{***}(0)	0.003^{***}(0)	1.14
Runs*Wickets* Catches			0.145^{***} (0)	5.74

Table 5 (continued)

Variables	Model 1	Model 2	Model3	VIF
Runs*Bowling Average*Catches			0.0026*** (0.003)	3.73
Runs* Catches* Toss decision			0.024*** (0.002)	0.64
Runs* Catches* Match location			0.006*** (0.005)	1.09
Runs*Catches*Match time			0.0015 (0.485)	2.52
Runs* Match time* Batting position			0.014 (0.348)	0.88
Wickets* Catches*Toss decision			−0.011*** (0)	1.73
Wickets* Catches* Match location			−0.016 (0.163)	1.23
Wickets* Catches* Match time			0.164*** (0.01)	1.03
Wickets* Catches* Batting position			−0.005 (0.932)	2.56
Catches* Bowling Average* Toss decision			0.001 (0.930)	2.08
Catches* Bowling Average* Match location			0.0006 (0.468)	5.69
Catches* Bowling Average* Match time			−0.0015 (0.485)	4.29
Catches* Toss decision* Match location			−0.011 (0.348)	1.2
Runs*Wickets*Catches* Toss decision			0.023***(0)	1.65
Runs*Wickets*Bowling Average*Match location			0.004***(0)	1.44
Adj. R-Square	0.65	0.76	0.82	
DW Statistic	2.42 (0.15)	2.35(0.02)	2.67(0.3)	
LM Statistic	3.5(0.2)	2.66(0.47)	2.65(0.54)	

significance value rho is 0.002. Both these statistics imply that there is no presence of autocorrelation in the dataset. The Langrange Multiplier (LM) is reported to be 3.5 with a p-value of 0.2, which is greater than the level of significance alpha = 0.05 (5% significance). This implies that the dataset is homoscedastic. The above model relies on the assumption that only one factor at a time impacts match outcome. However, in real-time, multiple factors simultaneously impact the result. In light of this scenario, the regression model can be augmented with variable interactions. This implies that individual variables like 'Runs', 'Bowling Average' can be multiplied pair-wise to form a new interaction term 'Runs*Bowling Average'. Similarly, all the eight predictors are taken two at a time and assuming that no two predictor variables are multiplied twice, i.e., 27 pair-wise interaction terms are initially factored in the regression models. Thus, a new regression Model 2 is implemented augmenting Model 1 with pair-wise interaction effects.

Table 6 Summary of regression model results for T20 Dataset

Variables	Model 1	Model 2	Model3	VIF
Runs	0.223^{***}(0.02)	0.223^{***}(0.02)	0.223^{***}(0.02)	1.13
Wickets	0.485^{**}(0.01)	0.485^{**}(0.01)	0.485^{**}(0.01)	1.02
Catches	0.004 (0.16)	0.004 (0.16)	0.004 (0.16)	1.52
Bowling Average	-0.004^{***}(0)	-0.004^{***}(0)	-0.004^{***}(0)	1.07
Toss decision	0.427^{***}(0)	0.427^{***}(0)	0.427^{***}(0)	1.73
Match location	0.046^{***}(0.003)	0.046^{***}(0.003)	0.046^{***}(0.003)	1.96
Match time	0.147^{***}(0)	0.147^{***}(0)	0.147^{***}(0)	1.01
Batting position	-0.4^{***}(0)	-0.4^{***}(0)	-0.4^{***}(0)	1.25
Runs*Wickets		0.812^{***}(0.003)	0.812^{***}(0.003)	1.03
Runs*Match location		0.359^{***}(0.008)	0.359^{***}(0.008)	2.56
Runs*Batting position		0.294(0.045)	0.294(0.045)	2.08
Wickets*Catches		0.163(0.971)	0.163(0.971)	5.69
Wickets *Batting position		0.069^{***}(0.006)	0.069^{***}(0.006)	4.29
Catches *Toss decision		0.005 (0.271)	0.005 (0.271)	1.2
Catches *Match location		0.001^{***}(0.009)	0.001^{***}(0.009)	1.65
Catches *Batting position		0.048 (0.279)	0.048 (0.279)	1.44
Bowling Average* Match location				
0.003(0.156)				
0.003(0.156)	2.34			
Toss decision * Match time		0.008(0.946)	0.008(0.946)	1.03
Toss decision * Batting position		0.314^{***}(0)	0.314^{***}(0)	2.56
Runs*Wickets* Catches			0.518^{***}(0)	2.08
Runs*Bowling Average*Catches			0.043^{***}(0)	5.69
Runs* Catches* Toss decision			0.037(0.007)	4.29
Runs* Catches* Match location			$0.062^{***(}$0)	1.2
Wickets* Catches*Toss decision			0.096(0.186)	1.65
Wickets* Catches* Match time			0.003(0.93)	1.44
Runs*Wickets*Catches*Toss decision			0.052^{***}(0)	5.41
Runs*Wickets*Bowling Average*Match location			$0.008^{***(}$0)	1.44
Adj. R-Square	0.67	0.80	0.85	
DW Statistic	1.97 (0.4)	1.85(0.2)	2.25(0.3)	
LM Statistic	2.15(0.5)	2.67(0.3)	2.53(0.5)	

From Model 2, it is found that interaction variables 'Runs*Wickets', 'Runs*Match location', 'Runs*Batting position', 'Wickets *Batting position', 'Catches *Toss decision', 'Catches *Match location', 'Catches *Batting position', 'Bowling Average* Match location' and 'Toss decision * Match time' and 'Toss decision * Batting position' are significant. This corroborates the result in Model 1 that customers simultaneously consider the above factor combinations for deciding. The value of adjusted R-squared is 0.76 i.e., 76% is explained. For Model 2, the Durbin Watson Statistic (DW) is reported to be 2.35 with a p-value of 0.02, showing no autocorrelation presence in the dataset. Similarly, the Langrange Multiplier (LM) is reported to be 2.66 with a p-value of 0.47 (> alpha = 0.05), which implies that the dataset is homoscedastic.

Model 3 is the extension of Model 2, with 17 triplet interaction variables and significant quadruplet interaction variables. This is done to factor in the real-time team decisions. Other triplet interactions are not formulated since they are not significant variables and cannot form significant interactions. The triple interaction term 'Runs*Wickets*Catches' and the quadruple interaction among 'Runs*Wickets*Catches*Toss decision' are the most significant.

The variables included in Models 1 and 2 are still significant, along with the additional variables. Overall model fit, as evident Adj.R-Squared shows an increase from Model 2 with a value of 0.82. This shows that the included triplet interaction variable as anticipated. For Model 3, the Durbin Watson Statistic (DW) is reported to be 2.67 with a p-value of 0.53, showing no autocorrelation presence in the dataset. Similarly, the Langrange Multiplier (LM) is reported to be 2.65, with a p-value of 0.54(> alpha = 0.05), which implies that the dataset is homoscedastic. The Variance Inflation Factor (VIF) is found to be < 10 implying no multicollinearity.

Using the step-wise method, a significant regression equation was thus found from Model 3 with an R^2 of 0.82 and adjusted R^2 of 0.821, and the standard error of the estimate is 0.35. The mean of standard residual was found to be zero, with the standard deviation 0.87. Figure 25 depicts the contribution of the features (feature importance) and their interaction to predict the match outcome.

For the feature contribution chart, the significant variables' standardized coefficients (both direct effect and interaction effect) are considered by normalizing the original coefficient. This normalization is performed by dividing the original coefficient by the sum of all the significant variable model coefficients (denoted by *** before p-value in Table 6) and multiplying by 100. For instance, the original coefficient of 'Toss decision' is 0.19. The sum of all significant variable coefficients (all direct effect variables + all the interaction effect variables in Table 6) = 0.13 + 0.06 + 0.09 + 0.19 + 0.04–0.1 + 0.07–0.009 + 0.042–0.08–0.009 + 0.042 + 0.008 + 0.037–0.097–0.009 + 0.042 + 0.0024 + 0.145 + 0.0026 + 0.024 + 0.006–0.011 + 0.164 + 0.003 + 0.004 = 0.787. Thus, the standardized feature coefficient for variable 'Toss decision' is normalized and plotted above in Fig. 25 as = 0.19/0.787 = 0.241. Similarly, triplet interaction variable 'Wickets*Catches*Match time' which is originally 0.164 is now normalized to the value = 0.164/0.787 = 0.208 in the Figure.

The above results are corroborated in the feature importance graph where Toss decision, 'Wickets*Batting position' and the triplet interaction variables 'Runs*Wickets*Catches' and 'Wickets*Catches*Match time' are the largest positive drivers of ODI match outcome.

Further, individual variable 'Batting position' and pair-wise interaction variable 'Bowling Average*Match location' are negative drivers.

In Table 6, three regression models are implemented: Model 1 is implemented by regressing the outcome variable (match outcome) on the individual predictors (direct effects) considered in the machine learning model. At 95% significance level, Runs, Wickets, Bowling

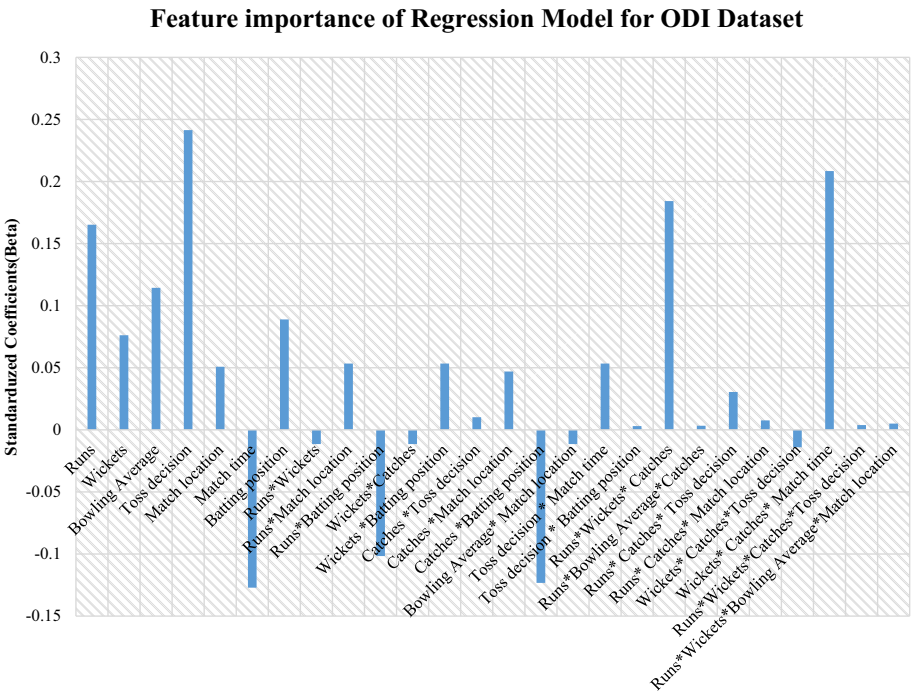

Fig. 25 Feature contribution and importance chart from the Multiple Linear Regression Model for ODI Dataset

Average, Toss decision, Match location, Match time and Batting position are the most significant factors considered for determining match outcome. The value of adjusted R-squared is 0.67 i.e., 67% is explained. Further, in Model 1, the Durbin Watson Statistic (DW) is reported to be 1.97 with a p-value of 0.4. The significance value rho is 0.002. Both these statistics imply that there is no presence of autocorrelation in the dataset. The Langrange Multiplier (LM) is reported to be 2.15 with a p-value of 0.5, which is greater than the level of significance alpha = 0.05 (5% significance). This implies that the dataset is homoscedastic (Table 6).

From Model 2, it is found that interaction variables 'Runs*Wickets', 'Runs*Match location', 'Wickets *Batting position', 'Catches *Match location' and 'Toss decision * Batting position' are significant implying and corroborating the result in Model 1. The value of adjusted R-squared is 0.80 i.e., 80% is explained. For Model 2, the Durbin Watson Statistic (DW) is reported to be 1.85 with a p-value of 0.2, showing no autocorrelation presence in the dataset. Similarly, the Langrange Multiplier (LM) is reported to be 2.67 with a p-value of 0.3 (> alpha = 0.05), which implies that the dataset is homoscedastic.

The triple interaction term 'Runs*Wickets* Catches', 'Runs*Bowling Average*Catches', 'Runs* Catches*Match location' and the quadruple interaction among 'Runs*Wickets*Catches *Toss decision' are the most significant.

The variables included in Models 1 and 2 are still significant, along with the additional variables. Overall model fit, as evident Adj.R-Squared shows an increase from Model 2 with a value of 0.85. This shows that the included triplet interaction variable as anticipated. For Model 3, the Durbin Watson Statistic (DW) is reported to be 2.25 with a p-value of 0.03, showing no autocorrelation presence in the dataset. Similarly, the Langrange Multiplier (LM)

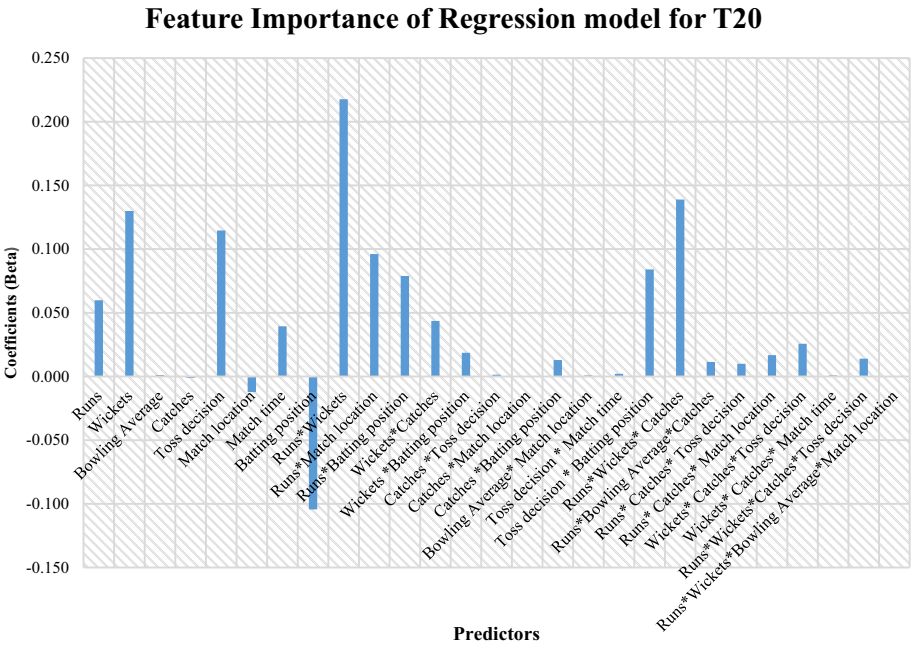

Fig. 26 Feature contribution and importance chart from the Multiple Linear Regression Model for T20 Dataset

is reported to be 2.53, with a p-value of 0.5(> alpha = 0.05), which implies that the dataset is homoscedastic. The Variance Inflation Factor (VIF) is found to be < 10 implying no multicollinearity.

Figure 26 depicts the contribution of the features (feature importance) and their interaction to predict the T20 match outcome.

For the feature contribution chart, the significant variables' standardized coefficients (both direct effect and interaction effect) are considered by normalizing the original coefficient. This normalization is performed by dividing the original coefficient by the sum of all the significant variable model coefficients (denoted by *** in Table 6) and multiplying by 100. For instance, the original coefficient of 'Batting position' is -0.4 in Table 6. The sum of all significant variable coefficients (all direct effect variables + all the interaction effect variables in Table 6)

$= 0.22324898 + 0.484620369 + 0.004078276–0.004297746 + 0.427463407–0.045603454 + 0.146893442–0.389278136 + 0.812 + 0.358546556 + 0.294270929 + 0.162731298 + 0.069206258 + 0.005201239 + 0.000618909 + 0.048063411 + 0.003083849 + 0.007749245 + 0.313611866 + 0.518307433 + 0.042724935 + 0.037069386 + 0.062479593 + 0.095869497 + 0.003183311 + 0.052044257 + 0.000259921 = 3.73.$

Thus, the standardized feature coefficient for variable 'Batting position' is normalized and plotted above in Fig. 11 as $= (-0.4/3.73) = (-0.105)$. Similarly, other coefficients are standardized in Fig. 26.

The above results are corroborated in the feature importance graph where Wickets, 'Wickets*Runs' and the triplet interaction variables 'Runs*Wickets*Catches' are the largest positive drivers of T20 match outcome. Further, individual variable 'Batting position' and pair-wise interaction variable 'Bowling Average*Match location' are negative drivers.

Further, it is to be noted that the unstandardized coefficients of the same interaction variables across Models 2 and 3 (pair-wise) are the same since Model 3 builds upon pair-wise interactions in Model 2. However, the standardized coefficients computed above for Model 2 and 3 will differ. This is due to the fact that unstandardized coefficients initially reported in Tables 5 and 6 are now divided by the sum of the coefficients of baseline Model 1 and the corresponding Model. This implies that standardized coefficients for Model 2 are the unstandardized coefficients depicted for Model 2 in both the Tables divided by sum of all the coefficients of Model 2 (including baseline variables of Model 1). Similarly, the standardized coefficient for Model 3 would be the ratio of unstandardized coefficients in Model 3 divided by sum of all the variables in Model 3(including Model 1 and Model 2 terms). Since Model 3 incrementally builds upon Model 2 with triplet and quadruplet interaction terms, the denominator (sum of coefficients) will differ. This leads to different standardized coefficients for Model 2 and Model 3.

For instance, while the pair-wise interaction term 'Runs*Wickets' have the same unstandardized coefficient value of 0.812 across Model 2 and 3 as depicted in Table 6 Row 10, the standardized coefficient for Model 2 is computed as the ratio of unstandardized value of 0.812 divided by sum of all coefficients for Model 2(incl Model 1 variables) in Table 6 =

0.812/ (0.223 + 0.485 + 0.004–0.004 + 0.427–0.046 + 0.147–0.389 + 0.812 + 0.359 + 0.294 + 0.163 + 0.069 + 0.005 + 0.001 + 0.048 + 0.003 + 0.008 + 0.314) = **0.812/2.922** = **0.277**.

On the other hand, the standardized coefficient for 'Runs*Wickets' computed for Model 3 is the same standardized coefficient value but divided by sum of coefficients for Model 3(incl variables in Models 1 and 2) in Table 7 =

0.812/ (0.223 + 0.485 + 0.004–0.004 + 0.427–0.046 + 0.147–0.389 + 0.812 + 0.359 + 0.294 + 0.163 + 0.069 + 0.005 + 0.001 + 0.048 + 0.003 + 0.008 + 0.314 + 0.518 + 0.043 + 0.037 + 0.062 + 0.096 + 0.003 + 0.052 + 0.001) = **0.812/3.73** = **0.217** as depicted in Fig. 26.

But considering that feature importance is to be represented for all the significant coefficients (including direct, pair-wise interaction and triplet interaction) in a single Fig. 26, the standardized coefficients computed from Model 3 (most accurate model with highest R-square) have been considered for the feature importance chart above due to better explainability of feature significance. This would help in revealing more accurate insights about the significant features for a more robust validation of the variable importance computed across the machine learning models.

The overall results of variable importance across all the models (machine learning, regression, and multi-criteria) are illustrated below in Table 7:

Therefore, across all the machine learning and regression models, the most significant factors impacting ODI match outcome are Player reputation and Match time (day and day/night), followed by the number of catches and runs scored. The vital factors for T20 match outcome are Batting order(first/second), toss decision, and runs scored are the most significant predictors, followed by player reputation.

As part of the unseen test set, the top five players Virat Kohli, Rishabh Pant, Ravindra Jadeja, Shreyas Iyer, and Hardik Pandya are considered by inputting their parameters and predicting the match outcome based on external match conditions and player performance. Different instances of each player are considered in the test set. Further, based on the actual match outcome, the win ratio of each player in the sample is computed as illustrated in Table 8.

For instance, for Hardik Pandya, of the three instances, the match outcome was found to be positive (1) in two of the instances and negative (0) in one of the instances. In such a scenario, the win ratio is computed as ratio of number of matches with positive outcome(won) to total

Table 7 Summary of Variable Importance across all Models and match formats

Key factors	Random forest	Gradient boost	Neural network	Regression model
ODI Matches	Player reputation and Match time (day and day/night), followed by the number of catches and runs scored	Player reputation and Match time (day and day/night), followed by the number of catches and runs scored	Player reputation and Match time (day and day/night), followed by the number of catches and runs scored	Player reputation and Match time (day and day/night), followed by the number of catches and runs scored
T20 Matches	Batting order(first/second), toss decision, and runs scored and player reputation	Batting order(first/second), toss decision, and runs scored and player reputation	Batting order(first/second), toss decision, and runs scored and player reputation	Batting order(first/second), toss decision, and runs scored and player reputation

Table 8 Clustering players based on win ratio

Player in Test Set	Win Ratio	Actual Outcome	Outcome predicted by Random Forest	Outcome predicted by Gradient Boosting	Outcome predicted by Neural Network
Hardik Pandya		1	0	1	1
Hardik Pandya	0.66	0	0	0	0
Hardik Pandya		1	1	1	1
Shreyas Iyer		0	0	0	0
Shreyas Iyer	0.25	1	1	1	1
Shreyas Iyer		0	1	0	1
Shreyas Iyer		0	0	0	0
Ravindra Jadeja	0.75	1	0	1	1
Ravindra Jadeja		0	0	0	0
Ravindra Jadeja		1	1	0	1
Ravindra Jadeja		1	0	1	1
Ravindra Jadeja		0	1	0	0
Rishabh Pant	0.6	0	1	0	0
Rishabh Pant		1	0	1	1
Rishabh Pant		1	1	0	1
Rishabh Pant		0	1	1	0
Rishabh Pant		1	1	0	1
Virat Kohli	0.8	1	1	1	1
Virat Kohli		1	0	1	1
Virat Kohli		1	1	0	1
Virat Kohli		0	0	0	0
Virat Kohli		1	1	1	1

number of matches (implying win ratio = 2/3 = 66%). Similarly, win ratio is computed for other four players. The players are then clustered into four different grades based on win ratio (players with a win ratio of greater than 70% are clustered under grade A + or 1, those with win ratio of 50–70% in grade A or 2, 30–50% in grade B or 3 and players under 30% are under grade C or 4. This is performed to identify emerging players for formulating teams with optimal winning combination. This helps in better talent acquisition and will boost the chances of Team India to win the game. The clustering results of the Indian players are compared with the actual player grades computed and stated in the Board of Cricket Control of India (BCCI) list and results are illustrated below:

4.4 Clustering results and validation of model efficacy

The predicted clusters/grades of the players are compared with the actually allocated clusters by BCCI in Fig. 27 as follows:

It can be inferred that for Virat Kohli, Rishabh Pant, Ravindra Jadeja, and Shreyas Iyer, the cluster predicted for them and actually assigned to them by BCCI in real-time are the same (1 for Virat Kohli, 2 for Pant and Jadeja, 4 for Shreyas Iyer). On the other hand, in the case of Hardik Pandya, while the predicted cluster according to the machine learning model based on winning ratio is 3, the actual grade allocated by BCCI is 2. In the case of Pandya, the grade was promoted to 2 (grade A) based on the performance in the latest match season (April 2021) which was not yet reflected in the ESPN CricInfo Statsguru website and hence predicted to be previous grade 3(B). The accuracy rate overall is found to be 95% (Figs. 28 and 29).

From the grades allocated to each of the players, it can be implied that the top five emerging players for India are: Kohli, Pant, Jadeja, Pandya, and then Iyer (lower the grade, higher is the propensity of the player to be considered emerging and chosen in the team for future matches).

Further, having identified the important factors influencing a match winning outcome, there is also a need to determine a suitable combination of the different factors (optimal values) at which a match winning result can be achieved. A pattern between different factor parameters and the ultimate match outcome needs to be derived in a rule-based method which

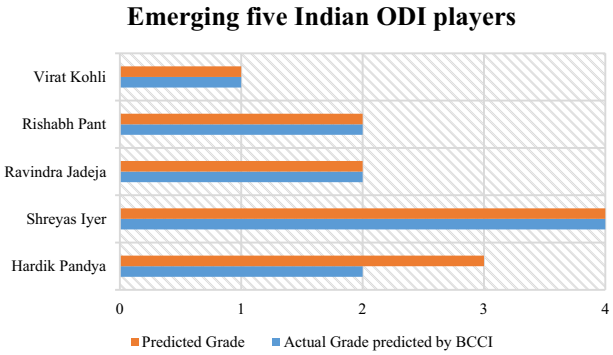

Fig. 27 Validation of model efficacy

Top 10 association rules	support	confidence	lift
{runs>50, catches>1}==> {outcome=1}	0.0103	0.8754	2.85
{bowl_avg>10}==>{outcome=1}	0.01	0.8734	2.73
{wkt>5, bowl_avg>5)==>{outcome=1}	0.007	0.8714	2.65
{match_time=1}==>{outcome=1}	0.004	0.8694	2.34
{batting_position=1, runs>20}==> {outcome=1}	0.001	0.8717	1.94
{runs<10, wkts<2,catches=0}==>{outcome=0}	0.0007	0.8697	1.86
{Catches<2, toss decision=0, match time=0}==>{outcome=0}	0.0006	0.8677	1.54
{catches<2, bowl_avg<5}==>{outcome=0}	0.0005	0.8657	1.30
{runs>50, catches>5,bowl_avg>10, match_time=0, toss decision=1}==>{outcome=1}	0.0003	0.8637	1.23
{runs<10, catches<2, match_time=0, toss decision=0}==>{outcome=0}	0.0001	0.8617	1.00

Fig. 28 Top 10 Association Rules for ODI Match Result

Top 10 Association rules	support	confidence	lift
{runs>50, catches>5}==> {outcome=1}	0.0103	0.75	2.00
{bowl_avg>10}==>{outcome=1}	0.01	0.75	2.00
{wkt>5, bowl_avg>5)==>{outcome=1}	0.007	0.74	1.00
{match_time=1}==>{outcome=1}	0.004	0.73	2.06
{batting_position=1, runs>20}==> {outcome=1}	0.001	0.73	2.00
{runs<10, wkts<2,catches=0}==>{outcome=0}	0.0007	0.73	2.00
{Catches<2, toss decision=0, match time=0}==>{outcome=0}	0.0063	0.73	1.00
{catches<2, bowl_avg<5}==>{outcome=0}	0.0005	0.72	1.00
{runs>50, catches>5,bowl_avg>10, match_time=0, toss decision=1}==>{outcome=1}	0.0003	0.72	1.00
{runs<10, catches<2, match_time=0, toss decision=0}==>{outcome=0}	0.0001	0.72	2.00

Fig. 29 Top 10 Association Rules for T20 Match Result

would give an insight into how parameters can be tuned to achieve the desired outcome separately for ODIs and T20s.

The association rule mining and network analysis results are illustrated.

4.5 Association rule mining and network analysis results

From the final dataset used for the modeling, variables such as runs, catches, bowling average, wickets, toss decision (win/loss), and match time (day/day and night) are considered. Using these variables, association rule mining is used to draw some useful rules to predict the match result (win or loss). Rules are generated such that all the above-listed features except outcome will participate in the antecedent and the outcome of the match will be the consequent. The support value at 0.001 was set with a filter condition stating a lift value greater than 1, having expected some meaningful rules and gaining some insight into the data. With the above structure in place, more than 70% of the confidence value is filtered to draw the top ten association rules.

Further, individual variable 'Batting position' and pair-wise interaction variable 'Bowling Average*Match location' are negative drivers.

The association rules are generated using the Apriori algorithm for support $= 0.001$ and confidence $= 0.7$ for ODI matches result outcome prediction, the top 10 of which are:

For instance, in Rule 9, the probability of a match won with the top player scoring more than 50 runs, more than 5 catches, bowling average > 10 of best bowler, day and night match

(match time = 0), and won the toss (toss decision = 1) has support of 0.0003 and confidence of 0.86 and lift of 1.23.

Similarly, for T20 matches win, the following association rules are observed:

For instance, the probability of won T20 match with a number of runs greater than 50 scored by top player and number of catches greater than 5 has a support of 0.01, confidence of 75% and lift of 2.0 depicting the most probable winning combination.

Thus, from a team point of view, it can be concluded that for winning an ODI match, the top performing batsman needs to score a half century (or more runs), more than five catches need to be taken and the best bowler must maintain an average of 10. Further, a day and night match with toss won will boost the match winning propensity.

Similarly, for T20 matches, a top-notch performance by the best batsman and more than five catches by the best fielder in a single innings will lead to a most probable favorable outcome for the team.

Therefore, the Association rules are found to corroborate the important factors and interactions derived from the machine learning and regression models.

Further, to analyze the influence of a player's performance with a country on how he performs with another country, network analysis is performed for both T20 and ODIs for the top four captains considered in exploratory data analysis.

The 'frequencyConnectedness' package in R computes the influence of one country over the other by constructing a mutual influence table termed as "Spillover matrix" for all the countries considered in the analysis. The "Spillover matrix" in Table 9 is obtained from the predefined function 'spilloverDY09'. This provides the variance error spillover matrix for each captain with other four countries (Australia, New Zealand, England, and India) during the pre-lockdown phase:

The respective country captains are not spilled over to their own country and indicated by zero in the matrix. For instance, Virat Kohli for India, Finch for Australia, Williamson for New Zealand, and Root for England are marked by zero since it is their respective host country. The captains are compared with the other three countries for performance indicated by the numeric spillover values.

For instance, in the first row, of a total of say 100 units (percentage), of Virat Kohli's overall performance, he is found to be 71% successful with Australia, 28% with New Zealand and 1% with England, similarly, for other captains, the spillover table is prepared.

The positive spillover values indicate a positive influence to another country in terms of performance.

Further, from the net spillover computed above to determine the connectedness of one country performance with the other, two minimum spanning trees are plotted for both ODI and T20 matches. The purpose of the minimum spanning trees is to represent the mutual influence of the countries graphically.

Table 9 Total spillover for ODIs

Country	Australia	New Zealand	England	India
Virat Kohli	0.71	0.28	0.01	0
Aaron Finch	0	0.35	0.37	0.28
Kane Williamson	0.22	0	0.37	0.41
Joe Root	0.37	0.32	0	0.31

In the minimum spanning tree, each country is represented by a small node with country name.

The topology follows connecting each country with lines. Countries connected by shorter lines indicate stronger influence while those connected by longer lines indicates relatively less mutual influence. This implies that if captains of two countries (other than those connected in the graph) perform well with one of the countries connected in the graph, they have a higher probability to perform well with other country connected by the shorter line to this country while less probability to perform well with country connected by longer line.

Following this notation, the minimum spanning tree (Zhang et al. 2020) is plotted for the ODI matches in Fig. 30.

From the minimum spanning tree above, it is observed that the England and India have a larger mutual influence while India and Australia have the least influence. This implies that if Aaron Finch and Kane Williamson (two other country captains i.e., Australia and New Zealand) perform well with England, they have a higher probability to perform well with India. If Williamson and Root (captains of New Zealand and England) perform well with India, however, they do not need to perform well with Australia due to the long line between Australia and India.

Similarly, for the T20 results, the net spillover matrix is illustrated in Table 10 as follows:

From the minimum spanning tree in Fig. 31, it is found that for T20 matches, Australia and India have a larger mutual influence while India and New Zealand have the least influence. This implies that if Joe Root and Kane Williamson (two other country captains i.e., England and New Zealand) perform well with Australia, they have a higher probability to perform well

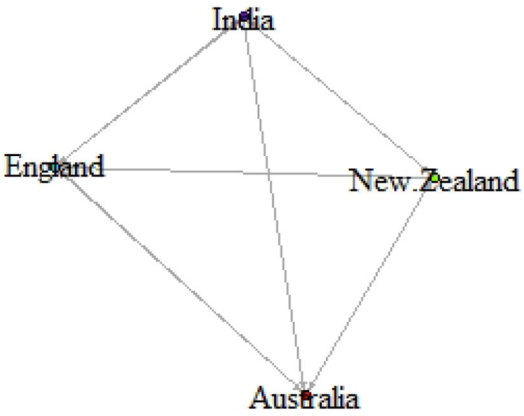

Fig. 30 Minimum spanning tree (MST) of ODI network analysis

Table 10 Total spill over-rate for T20

Country	Australia	New Zealand	England	India
Virat Kohli	0.74	0.08	0.18	0
Aaron Finch	0	0.55	0.24	0.21
Kane Williamson	0.18	0	0.43	0.39
Joe Root	0.25	0.46	0	0.29

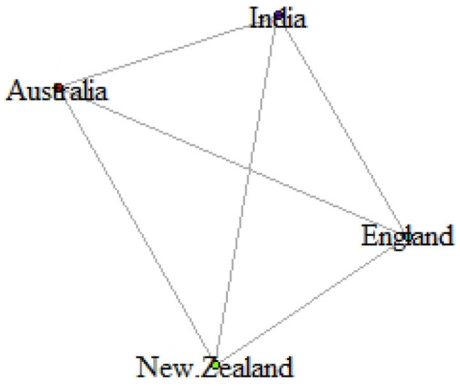

Fig. 31 Minimum spanning tree (MST) of T20 network analysis

with India in T20 matches. If Finch and Root (captains of Australia and England) perform well with India, however, they do not need to perform well with New Zealand due to longer line between New Zealand and India.

5 Discussion

5.1 Theoretical implications of the study

This study makes the following four contributions to the literature. First, this study extends the literature in the domain of match outcome prediction by factoring in categorical factors along with the traditionally used quantitative factors. Categorical environmental factors like toss result, match venue and batting order (whether the team is batting first or second) prove to be equally important as player specific statistics (quantitative factors) in determining match outcome.

Second, a comparative analysis of multiple predictive algorithms has been performed to identify the technique most suitable for prediction in context of sports outcomes, specifically cricket match outcomes. It was found that DNN outperforms the other two algorithms (random forest and gradient boosting). Thus, DNN can be recalibrated for different datasets to predict the match results in real-time. The predictors' feature importance is also computed and compared to identify the significant drivers of outcome prediction.

Third, this study validated the feature importance of predictors. It is an important step towards establishing the explainability of blackbox ML methods. We found that the most significant factors impacting ODI match outcome as predicted by the DNN model are Player reputation and Match time (day and day/night), followed by the number of catches and runs scored. The vital factors for T20 match outcome are Batting order(first/second), toss decision, and runs scored are the most significant predictors, followed by player reputation. The results of the deep learning model are validated for efficacy by clustering and association rule results.

Fourth, we introduce the nuance of the complex interaction of different parameters in the predictive models. We find that the pair-wise interaction 'Wickets*Batting position' and the triplet interaction variables 'Runs*Wickets*Catches' and 'Wickets*Catches*Match time' are the largest positive drivers of match outcome for one day international matches

(ODI). For T20s, the pair-wise interaction 'Wickets*Runs' and the triplet interaction variables 'Runs*Wickets*Catches' are the most significant drivers of match outcome.

5.2 Implications for practice

In addition to enhancing theoretical contributions in sports predictions using hybrid approaches, the results of this study have important implications for multiple stakeholders. The implications are of extreme relevance for both, sports management as well as cricket players, as discussed below.

5.2.1 Implications for sports management

The match result prediction model enables the management to identify the various environmental factors and player parameters that influence the winning rate of a sports team. The management can utilize these findings to formulate strategies to improve the performance. The cluster analysis model, which clusters and identifies emerging players, enables the management to assign a grade to the players and optimally utilize the talented players in the team. Such analysis can enable combinatorial optimization by choosing a winning combination of players to optimize team performance. It will also have the knock-on effect of reducing the incidence of idle talent on bench.

Data driven and optimized team management, in turn, would improve the reputation of the team and the national board council by uplifting the team rankings. Therefore, board councils and team management of respective countries can recalibrate the derived match result prediction model and tune the sensitive parameters to generate customized recommendations. Thus, from a team point of view, it can be concluded that for winning an ODI match, the top performing batsman needs to score a half century (or more runs), more than five catches need to be taken and the best bowler must maintain an average of 10. Similarly, for T20 matches, a top notch performance by the best batsman and more than five catches by the best fielder in a single innings will lead to a most probable favorable outcome for the team.

5.2.2 Implications for players

The match outcome prediction model enables the captains of the team to make appropriate decisions during toss and based on the condition of the pitch to maximize the winning rate. Further, captains can identify talented players in their teams, train them in a customized manner, and position them based on their strengths to win games. New teams can be groomed accordingly to know in real-time the order of priority of choosing players based on personalized preferences. Individual players can benefit in analyzing their shortcomings and making an individual game plan to maximize the winning rate of the team and their individual grades in the process. For instance, a player can gauge based on the match conditions and toss, how many runs to score, wickets to take, and catches to win the match.

6 Conclusion

The paper has attempted to compare the predictive performance of random forest, gradient boost, and deep neural network-based models and present the significance of the factors in predicting the outcome of match instances scraped from ESPN CricInfo website. The deep

neural network model is observed to outperform the other two machine learning models in terms of predictive accuracy and performance on unseen new players' data. The clustering results of the emerging players into different grades and comparison with actually allocated grades by BCCI are useful to recommend new talent and validate model efficacy. The association rules generated for both ODI and T20s present an insight into how parameters can be tuned to achieve the desired outcome. The rules generated corroborate the most significant variables and their interactions identified by machine learning and regression models.

However, the study is not without its limitations that can be worked upon in future research to generate additional insights in the area of sports outcome prediction. First, we utilize data from the ESPN CricInfo website only. Future studies could compare and aggregate the statistics with other cricket statistics websites like CricBuzz, Cricket World, HowSTAT!, and so on, to generate richer prediction. Second, the data sample considered for model building is limited to 3000 instances. This sample size can be varied, and other parameters like the significance of partnerships, match-winning batting combinations can be considered predictors in the models. The opinions about the players (player rating) (Xia et al., 2019) can also be considered as a potential predictor of match outcome. Third, the analysis can be performed during different seasons and even outside the home country to understand the efficacy of the model and the validity of the findings. The data about coaches for each team could be also scraped and included as a predictor variable of match outcome since the team composition and game strategies are defined by the coach of the team and h/she is a major driving factor of the match outcome. The results can be extended to examining the performance of bowlers and non-captain players for more varied insights.

Therefore, the paper attempts to compare and contrast techniques to solve research questions in sports and bring out insights from sports analytics in an international cricket context.

References

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. M. (2020). Sampling-based versus design-based uncertainty in regression analysis. *Econometrica, 88*(1), 265–296.

Adam, E., Mutanga, O., Abdel-Rahman, E. M., & Ismail, R. (2014). Estimating standing biomass in papyrus (Cyperus papyrus L) swamp: Exploratory of in situ hyper-spectral indices and random forest regression. *International Journal of Remote Sensing, 35*(2), 693–714.

Bendazzoli, S., Brusini, I., Damberg, P., Smedby, Ö., Andersson, L., & Wang, C. (2019). Automatic rat brain segmentation from MRI using statistical shape models and random forest. In Medical Imaging 2019: Image Processing (Vol. 10949, p. 109492O). International Society for Optics and Photonics.

Bose, A., Mitra, S., Ghosh, S., Ghosh, R., Patra, T., & Chakrabarti, S. (2021). Unsupervised learning based evaluation of player performances. *Innovations in Systems and Software Engineering, 17*(2), 121–130.

Bliss, A., Ahmun, R., Jowitt, H., Scott, P., Jones, T. W., & Tallent, J. (2021). Variability and physical demands of international seam bowlers in one-day and Twenty20 international matches across five years. *Journal of Science and Medicine in Sport, 24*(5), 505–510.

Cappelli, C., Di Iorio, F., Maddaloni, A., & D'Urso, P. (2019). Atheoretical regression trees for classifying risky financial institutions. Annals of Operations Research, 1–21.

Cea, S., Durán, G., Guajardo, M., Sauré, D., Siebert, J., & Zamorano, G. (2020). An analytics approach to the FIFA ranking procedure and the World Cup final draw. *Annals of Operations Research, 286*(1), 119–146.

Chauhan, S., Pande, R., & Sharma, S. (2020). The causal relationship between Indian energy consumption and the GDP: A shift from conservation to feedback hypothesis post economic liberalisation. *Theoretical & Applied Economics, 27*(3), 203–212.

D'Urso, P., De Giovanni, L., & Massari, R. (2019). Trimmed fuzzy clustering of financial time series based on dynamic time warping. Annals of Operations Research, 1–17.

D'Urso, P., De Giovanni, L., Massari, R., D'Ecclesia, R. L., & Maharaj, E. A. (2020). Cepstral-based clustering of financial time series. *Expert Systems with Applications, 161*, 113705.

D'Urso, P., De Giovanni, L., & Vitale, V. (2021). Spatial robust fuzzy clustering of COVID 19 time series based on B-splines. Spatial Statistics, 100518.

Deval, G., Hamid, F., & Goel, M. (2021). When to declare the third innings of a test cricket match?. Annals of Operations Research, 1–19.
de Zepeda, M. V. N., Meng, F., Su, J., Zeng, X. J., & Wang, Q. (2021). Dynamic clustering analysis for driving styles identification. *Engineering Applications of Artificial Intelligence, 97*, 104096.
Goossens, D. R., Beliën, J., & Spieksma, F. C. (2012). Comparing league formats with respect to match importance in Belgian football. *Annals of Operations Research, 194*(1), 223–240.
Hubáček, O., Šourek, G., & Železný, F. (2019). Learning to predict soccer results from relational data with gradient boosted trees. *Machine Learning, 108*(1), 29–47.
Huang, J., Tan, J., & Hua, D. (2021). Data mining of association between hyperuricemia and common chronic diseases based on evolutionary apriori algorithm (EAA). In 2021 IEEE 6th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA) (pp. 73–77). IEEE.
Jain, P. K., Quamer, W., & Pamula, R. (2021). Sports result prediction using data mining techniques in comparison with base line model. *Opsearch, 58*(1), 54–70.
Jiang, Y., & Chen, N. C. (2019). Event attendance motives, host city evaluation, and behavioral intentions. International Journal of Contemporary Hospitality Management.
Kamath, G. B., Ganguli, S., & George, S. (2020). Attachment points, team identification and sponsorship outcomes: evidence from the Indian Premier League. International Journal of Sports Marketing and Sponsorship.
Kamble, R. R. (2021). Cricket score prediction using machine learning. *Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12*(1S), 23–28.
Kong, Y. S., Abdullah, S., Schramm, D., Omar, M. Z., & Haris, S. M. (2019). Development of multiple linear regression-based models for fatigue life evaluation of automotive coil springs. *Mechanical Systems and Signal Processing, 118*, 675–695.
Lumbantobing, I. P., Sulivyo, L., Sukmayuda, D. N., & Riski, A. D. (2020). The effect of debt to asset ratio and debt to equity ratio on return on assets in hotel, restaurant, and tourism sub sectors listed on Indonesia stock exchange for the 2014–2018 period. *International Journal of Multicultural and Multireligious Understanding, 7*(9), 176–186.
Loureiro, A. L., Miguéis, V. L., & da Silva, L. F. (2018). Exploring the use of deep neural networks for sales forecasting in fashion retail. *Decision Support Systems, 114*, 81–93.
Mondal, S., Plumley, D., & Wilson, R. (2021). The evolution of competitive balance in men's international Cricket. Managing Sport and Leisure, 1–20.
Nikolaidis, Y. (2015). Building a basketball game strategy through statistical analysis of data. *Annals of Operations Research, 227*(1), 137–159.
Reyers, M., & Swartz, T. B. (2021). Quarterback evaluation in the national football league using tracking data. AStA Advances in Statistical Analysis, 1–16.
Saha, D., (2020). 10 Reasons why cricket is the most famous sport In India. Retrieved from: https://sportzwiki.com/cricket/why-cricket-most-famous-sport-india
Sahu, A. (2021). Predictive analysis of cricket. *Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12*(6), 5111–5124.
Schneider, M. J., & Sachin, G. (2016). Forecasting sales of new and existing products using consumer reviews: A random projections approach. *International Journal of Forecasting, 32*(2), 243–256.
Stern, S. E. (2016). The Duckworth-Lewis-Stern method: Extending the Duckworth-Lewis methodology to deal with modern scoring rates. *Journal of the Operational Research Society, 67*(12), 1469–1480.
Thomson, J., Perera, H., & Swartz, T. B. (2021). Contextual batting and bowling in limited overs Cricket. *South African Statistical Journal, 55*(1), 73–86.
Thorley, J. (2021). Age-related changes in the performance of bowlers in Test match cricket. International Journal of Sports Science & Coaching, 17479541211001726.
Vörösmarty, G., & Dobos, I. (2020). Green purchasing frameworks considering firm size: A multicollinearity analysis using variance inflation factor. *Supply Chain Forum: an International Journal, 21*(4), 290–301.
Weeraddana, N., & Premaratne, S. (2021). Unique approach for cricket match outcome prediction using Xgboost algorithms. *Journal of Theoretical and Applied Information Technology, 99*(9), 2162–2173.
Xia, H., Yang, Y., Pan, X., Zhang, Z., & An, W. (2019). Sentiment analysis for online reviews using conditional random fields and support vector machines. Electronic Commerce Research, 1–18.
Zhang, B., Guan, X., & Zhang, Q. (2020). Inverse optimal value problem on minimum spanning tree under unit l∞ norm. *Optimization Letters*, *14*(8), 2301–2322.