2023 IEEE 12th International Conference on Communication Systems and Network Technologies (CSNT) | 978-1-6654-6261-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/CSNT57126.2023.10134747

Analysis and Predictions of Winning Indian Premier League match using Machine Learning Algorithm

A.P Nirmala
Department of MCA
New Horizon College of Engineering
Bengaluru, India
nirmalasuresh.ap@gmail.com

V Asha
Department of MCA
New Horizon College of Engineering
Bengaluru, India
asha.gurudath@gmail.com

Arpana Prasad
Department of MCA
New Horizon College of Engineering
Bengaluru, India
arpana.sus@gmail.com

Biswajit Gogoi
Department of MCA
New Horizon College of Engineering
Bengaluru, India
aabiswajit0316@gmail.com

Arveti Naveen
Department of MCA
New Horizon College of Engineering
Bengaluru, India
arvetinaveen8@gmail.com

D Prathap Reddy
Department of MCA
New Horizon College of Engineering
Bengaluru, India
prathapreddya1998@gmail.com

***Abstract*— Cricket is the one of the most popular sports. A well-known domestic T20 league in the world is Indian Premier League (IPL). IPL is a very lucrative sport, hence players are in high demand, earning more than $2 million a season. Focusing on each player's performance and rating for each season is the need of the hour. A batter or bowler's performance can be evaluated more precisely using IPL data analysis. There is a huge amount of information and statistics involved in cricket and machine learning algorithms use this data to forecast the result of the match. This paper aims to predict the probability of winning IPL cricket matches in each ball of a over while running the second inning and rank the bowlers and batters based upon their performances. A number of vital factors, including team strength, players' form, possibility of draw and venue, were combined to provide the result of the game. In this paper, the match predictions are made using logistic regression techniques and its result accuracy is shown.**

Keywords—Cricket, Logistic Regression, IPL, Probability, Error, Accuracy.

I. INTRODUCTION

With billions of fans in countries like India, Australia, Africa, Pakistan, Great Britain and Great Britain, cricket is the second most popular sport in the world. It is an outdoor game between two teams of 11 players each, played on a rectangular, 22-yard long field on a cricket ground. The sport is played in Test, One Day International (ODI) and Twenty Over International (20 Over) (T20) formats. As with any sport, few factors that determines the outcome of a game such as player performance and factors like team size, pitch factor and geography, are taken into consideration while choosing a team. It is challenging to judge and predict a cricket match due to the many factors and limitations that exist. Every ball is important in this game as it has the power to change the course of a cricket match. The recognized shortest variant of the game is T20, which features 20 overs for each innings for each team. The other two versions are less "explosive" and athletic than this one. IPL features regional teams, a national squad, and an international team from India. The IPL as a whole is governed by India's Board of Control for Cricket. It follows 20-20 format and is owned by Stars, Merchants and Others (BCCI). A total of 8 teams are participating in IPL for the year (2021) including Rajasthan Royals (RR), Chennai Super Kings (CSK), Mumbai Indians (MI), Kolkata Knight Riders (KKR), Royal Challengers Bangalore (RCB), Delhi Capitals (DC), Punjab Kings (PK) and SunRisers Hyderabad (SRH). Despite the fact that Lucknow Super Giants and Gujarat Titans were two brand new teams from 2022.

The response to the queries "What are the chances of winning a match at a specific venue depending on the rival team, target" and "Most dismissals are a reason to beat," served as inspiration for this work. The Indian Premier League's most well-liked format of cricket is the subject of the investigation. The outcome of the IPL is influenced by a wide range of factors, and it can be anticipated like the outcome of any other game. The following are some of the top qualities or elements that affect how a game turns out. Used in this analysis are variables that have been shown to have a extensive impact on the game's result. Among the elements included in the analysis are: Team Past Performance: This element keeps track of every game ever played between teams on the field. This is essential since certain teams do well there and have a psychological advantage over rivals. This article mostly focuses on IPL analysis and determining IPL match winners using logistic regression based on the opposition, players' picks, and targets. Model is proposed for winning probability of the second batting team's can be predicted in each moment of the match. The paper is organized as literature review is

978-1-6654-6261-7/23/$31.00 ©2023 IEEE
DOI: 10.1109/csnt.2023.30

provided in Section II, III gives the dataset used, IV deals with the proposed model, V provides summary result and discussion and VI deals with conclusion.

II. Literature Review

Ahmad et al[1], key players were forecasted using machine learning approaches. Song et al. Roy et al.[2] stated that processing the data involves utilizing a projected ranking system. It is based on social network characteristics and their assessment in the form of a composite distributed framework using the Hadoop framework and the MapReduce programming paradigm. Barot et al.[3] based on the toss and the location, projected the result of the game. Rodrigues et al.[4], forecasted the importance of the bowlers' and batsmen's characteristics in the present game. Through the use of Multiple Random Forest Regression and prior performances of a sportsman playing opposition to particular team, this would aid for the selection of players for the upcoming matches. Shetty et al.[5], Using machine learning techniques, it was projected what each player could do depending on the field, pitch type, rival team, and many other variables. Using the Random Forest Algorithm, the model provided accuracy for batters, bowlers, and all-rounders of 76%, 67%, and 96%, respectively. They were able to forecast game outcomes and choose the top players thanks to this approach. Maduranga et al.[6] employing data mining algorithms, they forecasted the results of any cricket match and offered alternatives to other authors' methods. A predictive algorithm suggested in[7] , takes data from social network especially a feed from twitter to forecast a game's outcome. They also suggested a way to anticipate a player for a match even before it started. J. Kumar et al.[8] utilised decision tree algorithm and MLP networks to forecast the results of a match. In[9], the results of IPL games are predicted using random forest, K-nearest neighbour, Support Vector Machines, and Gaussian Naive Bayes classifier. To carry out this action, use a classifier. K. Abbas et al.[10] have examined the effects of the DLS approach on the results of ODI matches postponed because of climate using machine learning algorithms.

It is obvious from the aforementioned justifications that finding the outcomes of the match is not an easy task. The goal is to employ various approaches to build a system that can forecast cricket game outcomes, particularly in the IPL format.

III. IPL Dataset

A. Matches dataset

The 17 columns dataset, teamwise Home and Away dataset have CSV files such as matches and deliveries those have data about each match summary and ball-by-ball details. We have extracted some important features for IPL match analysis like city, team1, team2, draw_winner, result etc. from below table.

Column Name	Description
ID	Match ID
Season	IPL Season Year
City	City in which match was held
date	When the contest took place
Team1	Team1
Team2	Team2
Toss_winner	Whoever win the toss
Toss_decision	Opted to bat or field
Result	Standard, tie, or bat
Winner	Winning team in the game
Win by runs	How many runs did the team score
Win by wickets	Win by wickets in the game that was played
Player of match	Winner of the "Match of the Match" award
venue	Stadium where the game took place
Umpire 1	Umpire 1 name
Umpire 2	Umpire 2 name
Umpire 3	Umpire 3 name

Table 1. Match-wise dataset

B. Deliveries dataset

The delivery dataset has 20 columns. Table 2 lists each dataset column and its description. This dataset consists of two separate CSV files: matches and deliveries. These files contain information about summary each match and details of ball-by-ball.

Column Name	Column Description
Inning	Tells us the innings being played both team
Batting team	Tells us the batting team name
Bowling_Team	Tells us the bowling team name
Over	Tells over numbers being bowled
Ball	Tells the ball number of the over
Batsman	The batter name on strike
Non-striker	The batter name on runner
Bowler	The bowler name on bowling
Is_super_over	Tells if over is super over or not
Wide_run	If there are runs given for wide ball
Bye_run	If there are nay bye run given
Legby_run	If there are any leggy run given
No_ball_run	If the ball is a no ball
Penalty run	Penalty run due to any reason
Batsmen_runs	Runs hit by batter on the ball
Extra run	Extra runs given
Total run	Runs in the ball
Player_dismissed	If the player was given out or not
Dismissal_kind	What kind of dismissal it was
Fielder	Player who caused dismissal

Table 2. Deliveries dataset

IV. Proposed System

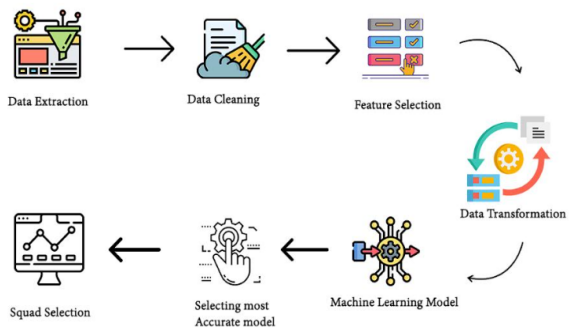

Fig. 1 Block diagram of proposed system

A. Data Extraction

Dataset collection from different sources is the initial phase in the construction of the model. The model's behavior and response are determined by the data that it is fed. We will have accurate results or forecasts if the data is correct and current [11]. On the Kaggle website, the dataset was accessible. The format was in CSV. The two datasets, delivery.csv and matches.csv, were downloaded from the website. The dataset is described in the same way as in the dataset section above.

B. Data Cleaning

The dataset may contain missing values, records with errors or corruption, or character values. As an example, the NULL values in the dataset for hitters who have not bat, have to be rectified before being used. As the data loaded as a.CSV file, the model won't be able to read commas, braces, or any other special characters, thus they will all be removed.

C. Feature Selection

Attributes that were not required for model training will be present in the extracted raw data. The dataset no longer contains attributes such as minutes, match date, match ID, and others of a similar nature.

D. Data Transformation

The data set's incomplete and inconsistent data were transformed into the proper data format. A sorted list of individual players would make it difficult to effectively train the model, so this data was generated at random. This may prevent certain players from adapting because they aren't employed in training. Data must be labelled for categorical values or in numerical format for logistic regression algorithms to work properly.

E. Machine Learning Model

Predictions can be made using a variety of machine learning methods. We must decide which algorithm will work with our dataset and predicted model the best. The model can be further optimized after choosing the algorithm [12]. Numerous machine learning methods may be used to forecast the outcome of the IPL match. We must determine which method will work best with our dataset and model. The model can be further optimized after the algorithm has been decided upon. The dataset is split into two groups in this step, one for training and the other for testing. Using supervised learning techniques, Machine learning algorithms are developed using training data. Once the trained model has been put to the test using algorithms, the outcome is anticipated. Selecting the most accurate model [13].

We use the logic regression technique to accomplish training, the most significant stage in machine learning. The system is trained by looking for trends in the training data and forecasting the results of each IPL match stage. As a result, the dataset provides the model with the knowledge it needs to complete the current task. We used logistic regression to classify the features in our model. We trained and evaluated the model by calculating all essential variables.

V. Result and Discussion

A. Analysis of IPL match.

Cricket is an exciting sport. A side may appear to be far ahead at halftime or at any other point in the game, but it only takes one player from the other team to put on an outstanding performance to swing the match's outcome in a matter of minutes. Additionally, a number of aspects, including the weather, intricate game regulations, player performance on the day, etc., have a significant impact on how the match turns out. Predicting the results of a cricket game is a difficult endeavor because of the large number of variables that play a role in it as well as its dynamic character.

We shall use logistic regression in this model since we must forecast the match's outcome in terms of percentages. A graph that compares the performance of the two specific teams competing in the match is displayed and could be used to determine the winner. Figure 2 depicts a comparison between MI and RCB throughout many seasons.

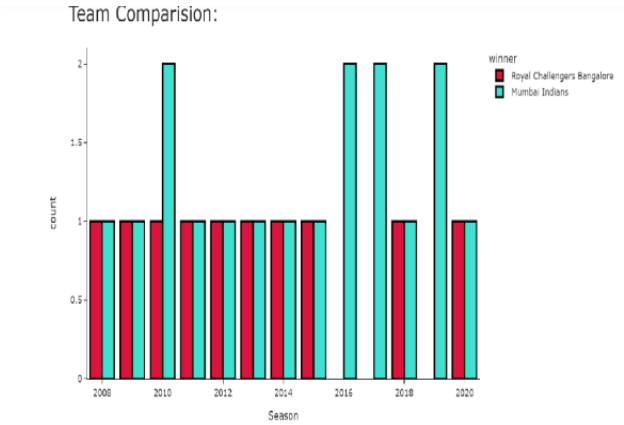

Fig. 2 Team Comparision

In the below figure it is described about each team match winning over match played by using histplot.

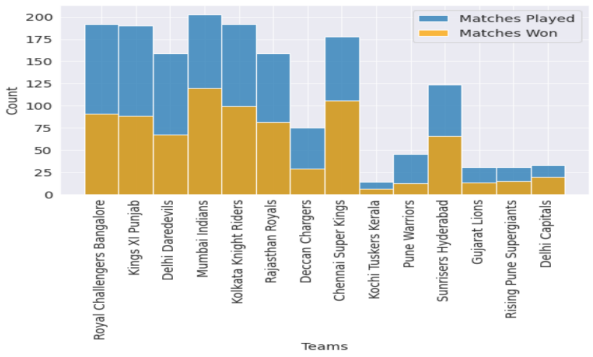

Fig. 3 Match win over match played

The top wicket falls overs and main reasons for dismissals are calculated over the previous matches, and the results are shown in the pie chart. These analysis give us the significance insight about the IPL matches.

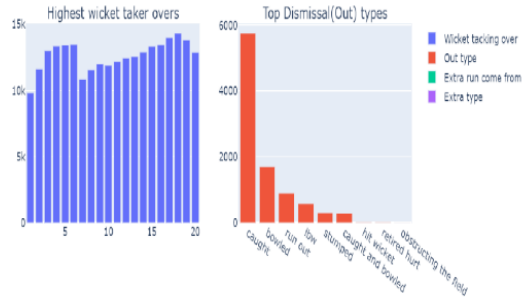

Fig. 4 Wickets falls reason on various over

In the below figure it has used bar plot to described the view of the highest run scorer and highest wickets taker in a IPL season. As a result it is viewed as Virat Kohli has scored most runs in a single IPL 2016 season and DJ bravo has taken highest wickets 32 in a single IPL 2013 season

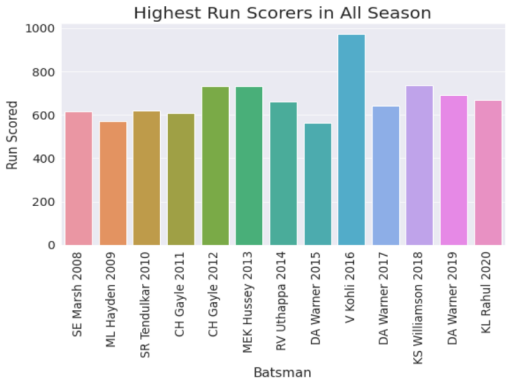

Fig. 5 Highest run scorer in a single season

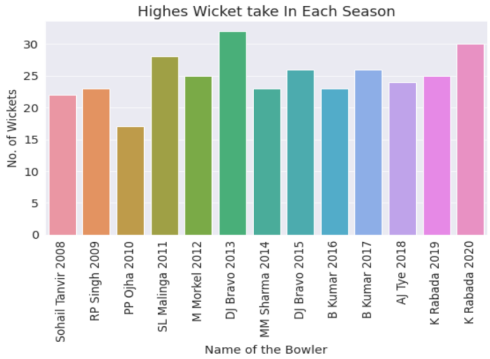

Fig. 6 Highest wicket taker in a single season

The match winning probability is illustrated in the following figure by taking into account a number of match-related factors, including first field, first ball, toss and match win, toss and match loss. The win and loss probabilities for variables like field first, first bat, toss and match win, toss and match loss, are shown in this figure.

Fig. 7 Match winning probability irrespective of toss win

B. Implementation of the model

Finding the ideal collection of measurements that would help create a squad with the best likelihood of winning is the major objective. We must evaluate a lot of data in order to reach trustworthy accuracy. As a result, gathering statistics for every IPL game was the initial stage in the implementation process. Ball-by-ball information has been gathered for every IPL game from 2008 until 2020. The required data was then retrieved and improved. Finally, the extracted data is stored in a relational database.

This model first began calculating the cumulative innings-level statistics for each game after saving the acquired data in a relational database. It was thought to use the shift level analysis framework. This is because, while building a representational model for learning, It enables you to keep track of your wins, losses, and draws by giving you a clear perspective of each inning's specifics. The intricacies of the player's level are somewhat reflected in the elements picked. They actually represent the total sum of player performances. Each player's individual contribution determines how well the team performs. The batter_and bowling_team, city, run left, wickets_left, current run rate, total runs x, required run rate, result, maximum individual

score in innings, runs in powerplay innings, etc are among the ten elements here computed and constructed. Since attributes are simply an accumulation of little details, relatively little domain knowledge is required. Since the shift level information can be labelled, In this study, the problem is modelled as a supervised learning problem. This can be categorized as a classification problem because the labels are discrete, so we decide to build this model using logistic regression. Then, using all the data we had gathered, we constructed a representative model with labels designating a win, loss, or tie for each inning. Here it is used 700 innings from 350 IPL games for the analysis.

After training the model, we independently predicted the outcomes for the test data for each conceivable combination of the produced parts. The outcomes were then contrasted with the initial outcomes. For every combination of features, accuracy was discovered [14]. Predictions for these feature combinations are more accurate and more accurate. This can be used to identify the ideal combination of variables that have a significant influence on match results [15]. With respect to many crucial match properties, such as batter_and bowling_team, run_left, city, wickets left, current run rate, total runs x, required run rate, and result, the results for each series or match are shown in the figure below.

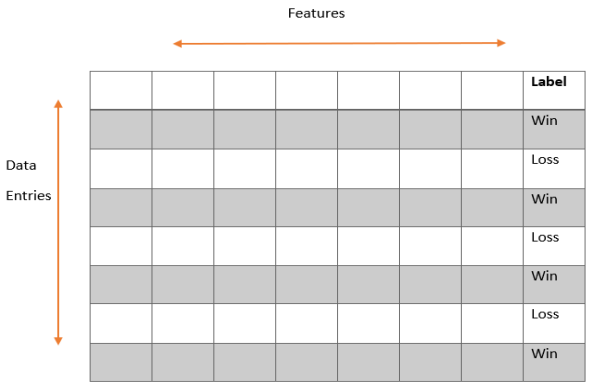

Fig. 8 Result of each match

It must discover new matches functions like run left, balls left, crr, and rrr result while taking into account both datasets. In this applied the formulas are shown in the below figure.

To find run_left

```
deliveries_df['run_left'] = delivery_df['total_runs_x'] - deliveries_df['current_score']
```

To find balls_left

```
deliveries_df['ball_left'] = 126 - (deliveries_df['over']*6 + deliveries_df['ball'])
```

To find current run rat

```
deliveries_df['crr'] = (delivery_df['current_score']*6)/(120 – deliveries_df['ball_left'])
```

To find required run_rate

```
deliveries_df['rrr']=(deliveries_df['runs_left']*6)/deliveries_df['balls_left']
```

Typically, binary classification jobs are handled via logistic regression. The SoftMax function is employed in place of the sigmoid function when multi-class classification is being done. We utilized logistic regression to categories the features in this model, and by computing all the necessary factors, we trained and tested the model and obtained a probability with good accuracy. The likelihood of winning and losing a game is depicted in the graph below at any given time. The likelihood of batting team winning is shown by the green line, the probability of losing is represented by the red curve, the wickets falling in each ball of an over are represented by the yellow curve, and the runs scored by the team batting second are represented by the blue bar. This model gives us accuracy approximate 80%, the winning and losing probability of the second batting team in each point of the match.

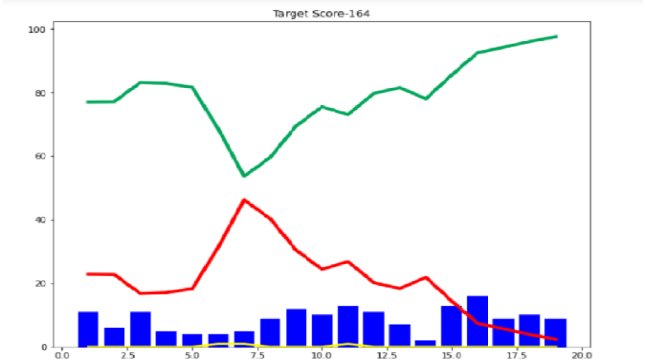

Fig. 9 Winning and losing prediction

VI. Conclusion and Future Scope

Predicting the results of cricket matches is still a young and exciting field of study. Achieving high accuracy scores is still difficult because of the game's complexity and dynamic nature, especially in the T20 format. Many researchers are motivated to model the game of cricket because of the rich and insightful insights and monetary advantages it can provide. Machine learning can be used to forecast the results of cricket matches by combining several match-affecting factors This research work discussed the impacts of pitching, positioning, and batting first or second. IPL team selectors, coaches and skippers may find this information useful for selecting the best player and for preparation before or during matches. This study has some limitations, such as the impact of the team. It is expected to use a mathematical model between team success and player performance as the future work. Also the most competent players could be found by grouping players using clustering into comparable groups depending on their performance. In future, the emphasize on each player's performance and evaluating it on a regular basis throughout the season. His bowling and batting ratings can also be predicted, which can compare player bid price and the price can be justified in auction. There may be an opportunity to find the man of the match for both teams player performance in a specific game

and series results. In this research there can have more feature added as batting opening pair, batting order and bowling order which could be done in these dynamic game to advance the result accuracy by considering these groupings.

REFERENCE

[1] H. Ahmad, A. Daud, L. Wang, H. Hong, H. Dawood, and Y. Yang, "Prediction of Rising Stars in the Game of Cricket," *IEEE Access*, vol. 5, pp. 4104–4124, 2017, doi: 10.1109/ACCESS.2017.2682162.

[2] S. Roy, P. Dey, and D. Kundu, "Social Network Analysis of Cricket Community Using a Composite Distributed Framework: From Implementation Viewpoint," *IEEE Trans. Comput. Soc. Syst.*, vol. 5, no. 1, pp. 64–81, Mar. 2018, doi: 10.1109/TCSS.2017.2762430.

[3] H. Barot, A. Kothari, P. Bide, B. Ahir, and R. Kankaria, "Analysis and Prediction for the Indian Premier League," in *2020 International Conference for Emerging Technology (INCET)*, Belgaum, India, Jun. 2020, pp. 1–7. doi: 10.1109/INCET49848.2020.9153972.

[4] N. Rodrigues, N. Sequeira, S. Rodrigues, and V. Shrivastava, "Cricket Squad Analysis Using Multiple Random Forest Regression," in *2019 1st International Conference on Advances in Information Technology (ICAIT)*, Chikmagalur, India, Jul. 2019, pp. 104–108. doi: 10.1109/ICAIT47043.2019.8987367.

[5] M. Shetty, S. Rane, C. Pandita, and S. Salvi, "Machine learning-based Selection of Optimal sports Team based on the Players Performance," in *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, Coimbatore, India, Jun. 2020, pp. 1267–1272. doi: 10.1109/ICCES48766.2020.9137891.

[6] M. M. Hatharasinghe and G. Poravi, "Data Mining and Machine Learning in Cricket Match Outcome Prediction: Missing Links," in *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*, Bombay, India, Mar. 2019, pp. 1–4. doi: 10.1109/I2CT45611.2019.9033698.

[7] A. N. Wickramasinghe and R. D. Yapa, "Cricket Match Outcome Prediction Using Tweets and Prediction of the Man of the Match using Social Network Analysis: Case Study Using IPL Data," in *2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer)*, Colombo, Sri Lanka, Sep. 2018, pp. 1–1. doi: 10.1109/ICTER.2018.8615563.

[8] J. Kumar, R. Kumar, and P. Kumar, "Outcome Prediction of ODI Cricket Matches using Decision Trees and MLP Networks," in *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)*, Jalandhar, India, Dec. 2018, pp. 343–347. doi: 10.1109/ICSCCC.2018.8703301.

[9] P. Somaskandhan, G. Wijesinghe, L. B. Wijegunawardana, A. Bandaranayake, and S. Deegalla, "Identifying the optimal set of attributes that impose high impact on the end results of a cricket match using machine learning," in *2017 IEEE International Conference on Industrial and Information Systems (ICIIS)*, Peradeniya, Dec. 2017, pp. 1–6. doi: 10.1109/ICIINFS.2017.8300399.

[10] K. Abbas and S. Haider, "Duckworth-Lewis-Stern Method Comparison with Machine Learning Approach," in *2019 International Conference on Frontiers of Information Technology (FIT)*, Islamabad, Pakistan, Dec. 2019, pp. 197–1975. doi: 10.1109/FIT47737.2019.00045.

[11] K. S. V. Swarna, A. Vinayagam, M. Belsam Jeba Ananth, P. Venkatesh Kumar, V. Veerasamy, and P. Radhakrishnan, "A KNN based random subspace ensemble classifier for detection and discrimination of high impedance fault in PV integrated power network," Measurement (Lond.), vol. 187, no. 110333, p. 110333, 2022.

[12] K. S. V. Swarna, A. Vinayagam, M. Belsam Jeba Ananth, P. Venkatesh Kumar, V. Veerasamy, and P. Radhakrishnan, "A KNN based random subspace ensemble classifier for detection and discrimination of high impedance fault in PV integrated power network," *Measurement (Lond.)*, vol. 187, no. 110333, p. 110333, 2022.

[13] S. Kumar R, A. Arulanandham, S. Arumugam, G. Dinesh, R. Thirukkumaran, and R. Subashmoorthy, "Analysis of classification and clustering techniques for ambient AQI using machine learning algorithms," in *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 2022.

[14] S. T. Suganthi, A. Vinayagam, V. Veerasamy, A. Deepa, M. Abouhawwash, and M. Thirumeni, "Detection and classification of multiple power quality disturbances in Microgrid network using probabilistic based intelligent classifier," *Sustain. Energy Technol. Assessments*, vol. 47, no. 101470, p. 101470, 2021.

[15] A. Puviarasu, M. Balaji, R. Thirukkumaran, A. Siva Kumar, and M. Premkumar, "Dynamic uneven clustering protocol for efficient energy management in EH-WSNs," *Mater. Today*, 2021.