Journal of Research Administration
Society of Research Administrators International

IPL MATCH PREDICTION USING MACHINE LEARNING TECHNIQUES

M Avinash[1], Dr. Harshalata Vishwakarma[2]

School Of Computer Science & Engineering, VIT Bhopal University, Kothrikalan, Sehore
Madhya Pradesh - 466114

ABSTRACT

The nation of India is enthralled with cricket, encompassing diverse formats including Test matches, ODI, and T20. The Premier League of India (IPL) exemplifies this fervor, drawing players from regional, national, and international teams. Factors like live streaming, radio coverage, and televised broadcasts buoy the league's widespread popularity among cricket enthusiasts. Predicting IPL match outcomes holds paramount significance for online traders and sponsors. This paper proposes a model for forecasting IPL match results, leveraging K-Nearest Neighbor, Logistic Regression, Random Forest Classifier (RFC), Support Vector Machines, and other machine learning methods and a voting regressor (a fusion of linear regression and support vector regressor). Our approach integrates various factors such as team composition, player batting, and bowling averages, team performance history, toss outcomes, venue dynamics, and the likelihood of winning by batting first against a certain team at a particular location. This research aims to enhance predictive accuracy and facilitate informed decision-making in IPL match analysis.

Keywords – Voting Regressor, SVR, Random Forest Classifier (RFC), K-Nearest Neighbors (KNN), and Logistic Regression (LR).

1. INTRODUCTION

Two teams of eleven players each play the outdoor bat-and-ball team sport of cricket. It spans three primary formats and ranks among the worldwide sports that are most popular. Just as in any other sport, a range of, a variety of factors affect how matches turn out. including player performance, pitch conditions, team composition, and venue dynamics, rendering cricket match analysis challenging.

The sport encompasses three formats: Tests, One Day International (ODI), Twenty-Twenty (T20), enjoying both national and international acclaim. In cricket, the significance of every ball cannot be understated, because it could change how the entire match plays out.

Indian Premier League (IPL) is a prominent cricket tournament in India that features teams from the country's national, regional and international teams. The Board of Control for Cricket in India (BCCI) is responsible for overseeing the IPL. is owned by celebrities, businesses, and others and is based on the 20-20 format. Eight teams will compete in the 2021. The Indian Premier League's squads are Delhi Capitals (DC), Kolkata Knight Riders (KKR), Chennai Super Kings (CSK),

SunRisers Hyderabad (SRH), Mumbai Indians (MI), Rajasthan Royals (RR), Royal Challengers Bangalore (RCB) and Punjab Kings (PK).

One of the main issues this paper seeks to answer is, "What is the probability of scoring 200 runs after 20 overs? “For instance, if a team has scored 48 runs for the loss of 3 wickets after 6 overs, what are the chances of reaching the milestone of 200 runs within the allotted 20 overs? Through the use of machine learning methods like We attempt to provide insights into forecasting the possibility of using SVM, Random Forest, and Logistic Regression. achieving specific run totals in IPL matches based on various game scenarios.

2. PROBLEM STATEMENT

The objective of this study revolves around the dataset named Ipl.csv (https://www.kaggle.com/datasets/yuvrajdagur/ipl-dataset-season-2008-to-2017matches)containing IPL matches data from 2008 to 2017. The primary aim is to predict the scoring range for 20 overs in cricket matches by inputting variables such the team that bats, the team that bowls, the number of overs (5.0 or more), the total wickets claimed the total amount of runs achieved, including the number of runs in the last 5 overs, and so forth.

To achieve this objective, the dataset will undergo thorough analysis and preprocessing. During preprocessing, various techniques will be applied to prepare the data for analysis by purifying, changing, and arranging it. The data that has been processed will next be used to create many machine learning models.

These machine learning models are developed with the intention of forecasting the scoring range for the full 20 overs based on the input variables provided. The outcomes from different models will be compared and evaluated to determine the most accurate and reliable prediction methodology.

3. LITERATURE REVIEW

People of all ages are drawn to cricket because of its thrill and charm, and it has grown to be a billion-dollar industry for many who gamble financially in the hopes of making big profits. However, concerns about spot-fixing loom large, intensifying the demand for predictive models to ascertain match outcomes. In this paper [1], we address the challenge of predicting the winner of upcoming IPL matches by leveraging individual player competency, team coordination, and historical match data. Our proposed model utilizes machine learning algorithms, including Accuracy rates for the Decision Tree, Naive Bayes and Support Vector Machine's performance classifiers were 95.96%, 97.98%, and 98.99%.

Forecasting match outcomes in cricket, a sport characterized by uncertainty, presents a formidable challenge. In this paper [2], we focus on predicting victory Machine learning techniques in one-

day international cricket matches were employed to evaluate the performance of classifiers. Employing 128 features encompassing batting, bowling strength, and overall team performance, we propose ensemble algorithms and feature selection methods to enhance prediction accuracy. Logistic Regression and Support Vector Machine emerge as superior models, yielding an accuracy of 96.30% in predicting ODI match winners.

In our study [3], we delve into techniques for predicting the outcomes of T20 cricket matches by considering player performance statistics, player ratings, player clustering, and an ELO-based rating system. We employ various machine learning algorithms such as In our study, we delve into techniques for predicting the outcomes of T20 cricket matches by considering player performance statistics, player ratings, player clustering, and an ELO-based rating system. We employ various machine learning algorithms to assess and forecast match results, including random forests The, Naive Bayes, Support Vector Machine's, Decision Trees

Cricket, like any sport, hinges on various factors that influence winning outcomes. Home crowd advantage, past performances, experience, venue, opposition team, and current form all contribute to a team's success. This article [4], provides insights into the factors influencing cricket match outcomes and discusses research papers that delve into predictive modeling for cricket matches.

As technology advances, predicting match outcomes has become increasingly crucial in sports, particularly cricket. This review paper [5], examines the utilization of Using methods based on machine learning, one-day cricket world match champions may be identified and has garnered considerable attention. By leveraging career statistics and team performances, supervised learning algorithms offer valuable insights for coaches to identify areas for improvement. The paper evaluates four types of machine learning algorithms and compares their effectiveness in predicting match outcomes.

4. Methods

In our methodology, we initiate by importing fundamental libraries like pandas, Seaborn, matplotlib, pickle, and metrics. These libraries will aid in data manipulation, visualization, and the calculation of evaluation metrics, which encompass the metrics three types of errors: r squared (R2), mean absolute error (MAE), and mean squared error (MSE).

Furthermore, we will import various types of regression models, including Logistic Regression, Decision Tree Regressor, Naïve Bayes, K Neighbors Regression, Linear Regression, Support Vector Regressor (SVR), and Voting Regressor.

The Voting Regressor represents a unique ensemble technique that combines Linear Regression with Support Vector Regressor to harness the strengths of both models. This amalgamation is designed to R-squared error is maximized while mean squared and mean absolute errors are also measured.

We aim to conduct an evaluation relative to the predictive model's R Squared (The coefficient of determination), mean absolute(unchanging), and mean squared (errors in squares) error. This analysis is intended to provide insights into the performance of each regression technique and aid in identifying the most effective model for our predictive task.

Based on our evaluation, the Voting Regressor, along with Linear Regression, demonstrates the lowest error rates in terms of determining the greatest r-squared error, least mean squared error and mean absolute error are taken into account. of all regression approaches. This underscores the efficacy of the ensemble approach in minimizing predictive errors and maximizing explanatory power.

Advanced predictive models offer increased prediction accuracy and more effective outcomes. However, they can suffer from the use of less accurate models and may not always provide effective predictions due to various factors such as data quality and model limitations.

5. Proposed System:

Our main goals are to pinpoint the critical elements affecting match results and ascertain which machine learning model is most suited for the job. Finding the most important variables that affect match outcomes and choosing the best machine learning model are our key objectives. capable of thoroughly analyzing this data and generating accurate predictions. While numerous studies have explored predicting cricket match outcomes, many struggle with accuracy, often due to insufficient consideration of key factors or the use of inappropriate machine learning models.

In our system, we aim to address these limitations by incorporating all relevant factors that can impact match outcomes and selecting the best-performing model for data training and testing. By doing so, we anticipate a significant improvement in prediction accuracy.

Our approach involves prioritizing important factors that influence match results by assigning them balanced strengths using intelligent formulas such as Euler's Strength calculation formula. We have conducted research on the intricacies of T20 cricket, which is highly dynamic and sensitive to small changes, such as a single over altering the course of the entire match.

Through highlighting the importance of machine learning in prediction, our aim is to instill confidence in the predicted winner. Moreover, our system incorporates Eager Learning techniques, which assist database administrators in maintaining the database efficiently. It offers flexibility in training additional data at both user and server levels, enhancing adaptability and scalability.

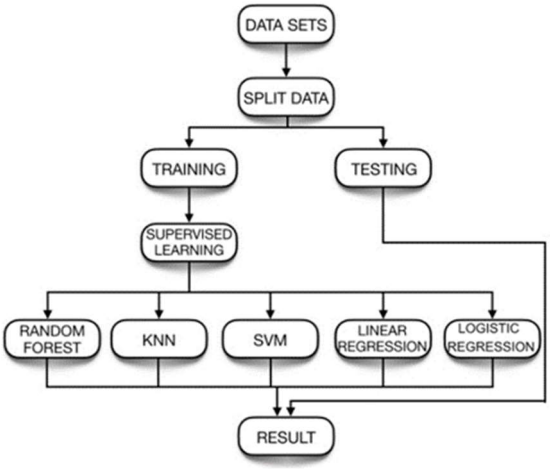

Fig.1: Design of the model [21]

The Data Flow Diagram (DFD) is a short graphical depiction of a system that is frequently called a bubble chart. It delineates the flow of input data, the processing stages within the system, and the resulting output data.

As a pivotal modeling tool, the DFD captures essential system components, including system processes, the pertinent data manipulated by these processes, the system's information routes and external entities interacting with it.

The DFD visually illustrates the trajectory of information within the system and highlights the transformations undergone as data progresses from input to output. This graphical technique effectively portrays information flow and the sequential modifications applied throughout the process.

Recognized interchangeably as a bubble chart, the DFD is adaptable to represent systems at different abstraction levels. The structure can be divided into tiers, with each tier denoting a higher degree of information flow and functional complexity.

MODULES:

To successfully carry out this project, we have created the modules shown in Given Fig. 1.

Importing Libraries:

Importing essential libraries for the analysis, machine learning, and manipulation of data. NumPy, pandas, matplotlib, seaborn, and scikit-learn are among the frequently used libraries.

Importing Dataset (IPL Dataset):

Obtain a dataset containing historical IPL match data. The dataset named "Ipl.csv," sourced from the Kaggle Repository, encompasses IPL matches data spanning from 2008 to 2017. This dataset comprises attributes such as match details (matched), date, venue, batting side, bowling side, Runs(scores) and wickets(outs) in the last 5 overs, striker, and non-striker. The primary objective is to utilize various independent variables within the dataset to predict the dependent variable, namely the total runs scored by the team after completing 20 overs. The dataset can be accessed through the following link [22].

Analysing preliminary information :

Utilize Analysing preliminary information (EDA) to learn more regarding the dataset's attributes. Key steps in EDA include Checking for Null Values: Identify and handle missing data appropriately through imputation or deletion.
Data Visualization: Create a variety of visualizations such as histograms, scatter plots, and box plots to comprehend feature distributions and relationships. Transform categorical values into numerical representations through encoding techniques like label encoding or one-hot encoding.

Data Preprocessing:

Prepare the data for model training by segmenting it into features (X) and the target variable (y). Here, the target variable corresponds to the match score, while the features encapsulate the various attributes used for prediction. Normalize or scale numerical features if necessary to ensure uniformity in the data.

Dividing the Dataset:

separating the ones used for learning and evaluating sets of the dataset in order to precisely assess how well the trained models perform.

Model Selection and Training:

Examine a range of suitable machine learning algorithms for classification tasks. Additionally, utilize ensemble methods such as Voting Regressor to combine the predictions of multiple models for reducing error rate.
By systematically implementing these modules, we aim to build a robust predictive model for forecasting IPL match scores accurately, minimizing error rates, and enhancing predictive performance.

6. ALGORITHM:

Logistic regression predicts the likelihood of a binary result based on input factors; this technique is frequently used in binary classification. A decision tree represents a structure akin to a flowchart, in which each internal node represents an attribute test and each leaf node represents a class label. This flexible approach is applied to regression as well as classification applications.
The supervised learning algorithm SVR is used for tasks involving regression and classification. It creates a barrier for decisions to differentiate classes in multi-dimensional space, striving to maximize the margin between them.
Naive Bayes operates as a probabilistic classifier based on the Bayesian hypothesis. It presupposes that the characteristics and computes the likelihood of a class-given input features**.**
By projecting one variable's value depending on another's value, variables that are both separate from one another are produced by linear regression. To create estimates, a combination of learning techniques known as Random Forest integrates the output of a number of decision trees. It’s versatile and effective for both regression and classification assignments. KNN, an instance-based, non-parametric learning algorithm, is applied in classification and regression tasks. Its predictions rely on the majority vote of its nearest neighbors**.**

The Voting Regressor, an ensemble meta-estimator, amalgamates the predictions from various base regressors to generate a final prediction. It leverages the aggregation of individual predictions to enhance overall performance.

These algorithms provide varied approaches to addressing classification and regression challenges in machine learning. Depending on the dataset's characteristics and the specific task, one or more of these algorithms might be appropriate for constructing predictive models with differing levels of complexity and error rates.

7. RESULTS AND DISCUSSION

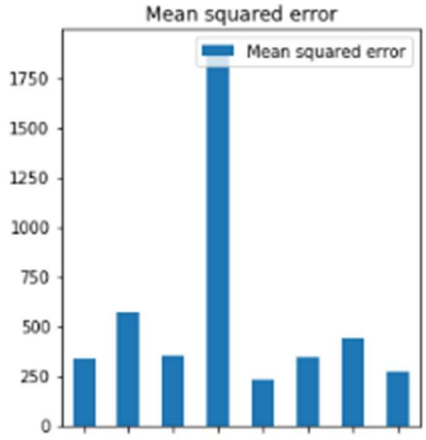

Fig.2: Comparison of Mean Squared Error

The comparative analysis of regression algorithms, as depicted in the bar graph, reveals distinct performances based on Mean Squared Error (MSE). Naive Bayes Regression emerges with the highest MSE, indicating its comparatively lower predictive accuracy, while Linear Regression stands out as the algorithm with the lowest MSE, signifying superior performance in minimizing error. Voting Regression, SVM Regression, Logistic Regression, Random Forest Regression, and KNN Regression demonstrate moderate MSE values, positioning them as average performers in the context of predictive accuracy. This succinct summary encapsulates the algorithmic landscape, providing a clear understanding of their relative effectiveness without delving into specific numerical values.

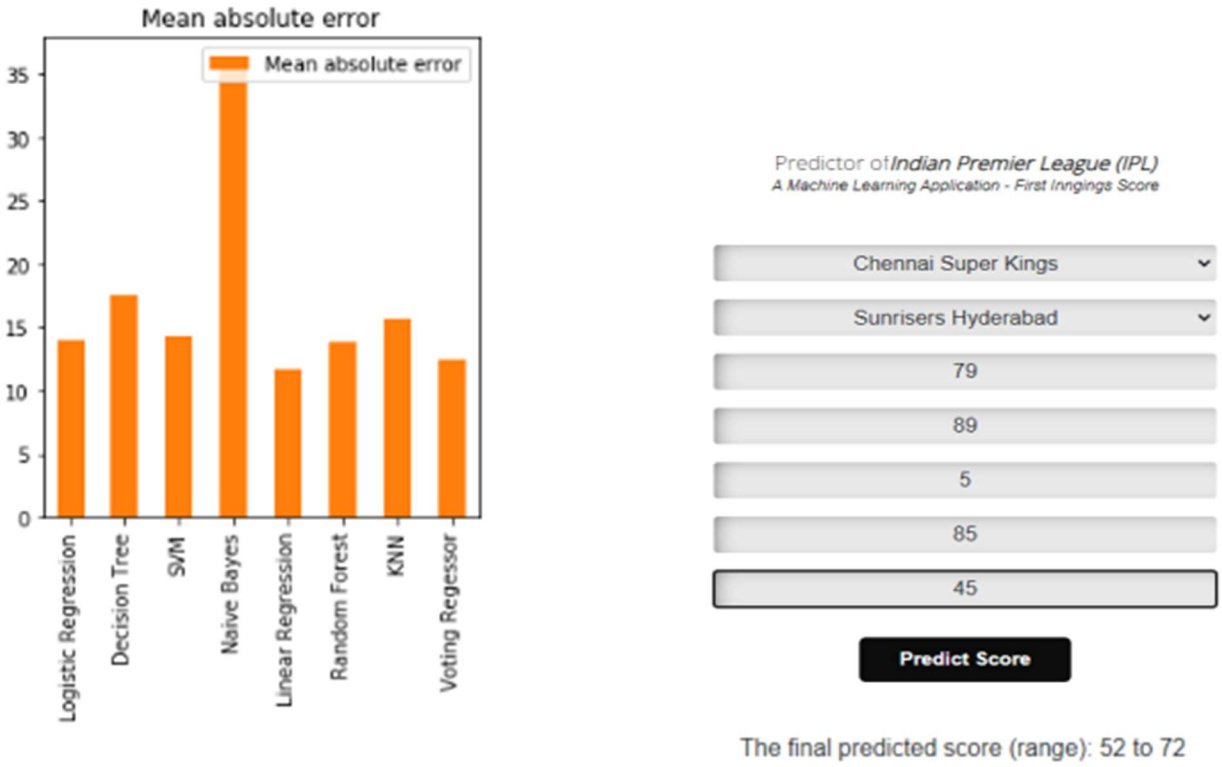

Fig.3: Comparison of Mean Absolute Error

The bar graph illustrates the performance of various regression algorithms with Mean Absolute Error (MAE) values on the y-axis, ranging from 0 to 40. Notable results include Voting Regression with the lowest MAE of 13, followed by Linear Regression and Logistic Regression with MAE values of 12 and 14, respectively. SVM Regression and REGRESSION also exhibit competitive performance, each having an MAE of 14.5 and 14. Naive Bayes Regression and Decision Tree Regression show relatively higher MAE values of 37 and 18, respectively. This summary provides a quick overview of the algorithms' relative performance in minimizing absolute errors, offering insights into their predictive values.

Fig.6: Output screen

S.No	Algorithm	Error Rate
1	Mean Squared Error	268.57
2	Mean Absolute Error	12.42
3	R2 Squared Error	0.73

Table.1: Results of Voting Regressor

8. CONCLUSIONS

The model was trained on specific characteristics and then applied to anticipate the data from tests. The data applied in this experiment was split into sets for training first. Then processing the data is done with Testing next. The efficiency of the system was then assessed. In this study, a range of classification models was applied, including Decision tree models, random forest classification, support vector machine, Voting Regressor Algorithm, Gaussian Naïve Bayes Classifier, KNN (K Nearest Neighbor) technique, and Logistic Regression.

Among these models, both Linear Regression and the Voting Regressor (a combination of Linear Regression and Support Vector Regressor) demonstrated promising results. For the Voting Regressor, the following performance metrics were obtained:

The Voting Regressor, a fusion of Linear Regression and Support Vector Regressor, yields the performance measures exhibit promise: 268.57 for Mean Squared Error, 12.42 for Mean Absolute Error, and 0.73 for R2 Squared Error. These metrics indicate the model's ability to predict outcomes effectively, showcasing its potential for accurate forecasting in regression tasks. With its robust performance, the Voting Regressor presents a compelling approach for predictive modeling in diverse applications, offering valuable insights and enhancing decision-making processes.

Looking forward, prospects in cricket analytics could involve analyzing individual player performance consistently throughout the season. This could entail predicting ratings for both bowling and batting abilities. Furthermore, there's potential for developing models to forecast who will be the "Man of the Match" in two-team games. Such models would require a thorough analysis of player statistics, match dynamics, and performance trends to identify standout performers in each game accurately. By delving deeper into player analytics and match predictions, cricket analytics can evolve to provide more insightful and precise assessments of player and team performances.

DECLARATIONS

LIST OF ABBREVATIONS

MSE – Mean squared error.
MAE – Mean Absolute error.
R2 – Coefficient of determination
LR – Logistic Regression
SVR – Support Vector Regressor
KNN – K-Nearest Neighbors
RFC – Random Forest Classifier
IPL – Indian Premier League
EDA – Exploratory Data Analysis

References:

[1] Haseeb Ahmad, Ali Daud, Licheng Wang, Haibo Hong, Hussain Dawood, and Yixian Yang, Prediction of Rising Stars in the Game of Cricket, IEEE Access, Volume 5, PP. 4104 – 4124, 14 March 2017.

[2] Haryong Song, Vladimir Shin and Moongu Jeon, Mobile Node Localization Using Fusion Prediction-Based Interacting Multiple Model in Cricket Sensor Network, IEEE Transactions on Industrial Electronics, Volume: 59, Issue: 11, November 2012.

[3] Sarbani Roy, Paramita Dey and Debajyoti Kundu, Social Network Analysis of Cricket Community Using a Composite Distributed Framework: From Implementation Viewpoint, IEEE Transactions on Computational Social Systems, Volume: 5, Issue: 1, PP. 64-81, March 2018.

[4] Priyanka S, Vysali K, K B PriyaIyer, Score Prediction of Indian Premier League- IPL 2020 using Data Mining Algorithms, International Journal for Research in Applied Science & Engineering Technology (IJRASET), Volume 8, Issue II, PP. 790-795.

[5] Prince Kansal, Pankaj Kumar, Himanshu Arya, Aditya Methaila, Player valuation in Indian premier league auction using data mining technique, International Conference on Contemporary Computing and Informatics (IC3I), 27-29 Nov 2014

[6] Shilpi Agrawal, Suraj Pal Singh, Jayash Kumar Sharma, predicting results of IPL T-20 Match using Machine Learning, 2018 8th International Conference on Communication Systems and Network Technologies (CSNT), 24-26 Nov. 2018.

[7] Harshit Barot, Arya Kothari, Pramod Bide, Bhavya Ahir, Romit Kankaria, Analysis and Prediction of Indian Premier League, 2020 International Conference for Emerging Technology (INCET), 5-7 June 2020.

[8] Amal Kaluarachchi, S. Varde Aparna, CricAI: A classification-based tool to predict the outcome in ODI cricket, 2010 Fifth International Conference on Information and Automation for Sustainability, 17-19 Dec. 2010.

[9] Kalpdrum Passi and Nirav Kumar Pandey, predicting player's performance in one day international cricket matches using machine learning, Volume 8, Computer Science & Information Technology, 2018.

[10] Nigel Rodrigues, Nelson Sequeira, Stephen Rodrigues, Varsha Shrivastava, Cricket Squad Analysis Using Multiple Random Forest Regression, IEEE Xplore, 1st International Conference on Advances in Information Technology, 2019.

[11] M. B. Wright, Scheduling fixtures for New Zealand Cricket, IMA Journal of Management Mathematics 16, PP. 99−112, 2005.

[12] Manuka Maduranga, Hatharasinghe, Guhanathan Poravi, Data Mining and Machine Learning in Cricket Match Outcome Prediction: Missing Links, 5th International Conference for Convergence in Technology (I2CT), Mar 29- 31, 2019.

[13] Monali Shetty, Sankalp Rane, Chaitanya Pandita, Suyash Salvi, Machine learning-based Selection of Optimal sports Team based on the Players Performance, Proceedings of the Fifth International Conference on Communication and Electronics Systems (ICCES 2020), IEEE Conference Record, ISBN: 978-1-7281-5371-1, 2020.

[14] Balasundaram A, Ashokkkumar S, Jayashree D, Magesh Kumar S, Data Mining based Classification of Players in Game of Cricket, proceedings of the International Conference on Smart Electronics and Communication (ICOSEC 2020), IEEE Xplore Part Number: CFP20V90-ART; ISBN: 978-1-7281-5461-9.

[15] Jalaz Kumar, Rajeev Kumar, Pushpender Kumar, Outcome Prediction of ODI Cricket Matches Using Decision Trees and MLP Networks, 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC).

[16] Prabu, S., Balamurugan Velan, F. V. Jayasudha, P. Visu, and K. Janarthanan. "Mobile technologies for contact tracing and prevention of COVID-19 positive cases: a cross-sectional study." International Journal of Pervasive Computing and Communications (2020).

[17] Subramani, Prabu, K. Srinivas, R. Sujatha, and B. D. Parameshachari. "Prediction of muscular paralysis disease based on hybrid feature extraction with machine learning technique for COVID-19 and post-COVID-19 patients." Personal and Ubiquitous Computing (2021): 1-14.

[18] Do, Dinh-Thuan, Tu Anh Le, Tu N. Nguyen, Xingwang Li, and Khaled M. Rabie. "Joint impacts of imperfect CSI and imperfect SIC in cognitive radio-assisted NOMA-V2X communications." IEEE Access 8 (2020): 128629-128645.

[19] Le, Ngoc Tuyen, Jing-Wein Wang, Duc Huy Le, Chih-Chiang Wang, and Tu N. Nguyen. "Fingerprint enhancement based on tensor of wavelet subbands for classification." IEEE Access 8 (2020): 6602-6615.

[20] Nguyen, Tu N., Bing-Hong Liu, Nam P. Nguyen, and Jung-Te Chou. "Cyber security of smart grid: attacks and defenses." In ICC 2020-2020 IEEE International Conference on Communications (ICC), pp. 1-6. IEEE, 2020.

[21] Srikantaiah, K. C., Khetan, A., Kumar, B., Tolani, D., & Patel, H. (2021, September). Prediction of IPL match outcome using machine learning techniques. In 3rd International Conference on Integrated Intelligent Computing Communication & Security (ICIIC 2021) (pp. 399-406). Atlantis Press.

[22] https://www.kaggle.com/datasets/yuvrajdagur/ipl-dataset-season-2008-to-2017