International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys.14, No.1 (Jul-23)

http://dx.doi.org/10.12785/ijcds/140122

Enhancing Data Integrity In Mobile Crowdsensing Environment With Machine Learning And Cost-Benefit Analysis

Ramesh K. Sahoo 1[1], Sateesh Kumar Pradhan 2[2], Srinivas Sethi 3[3] and Siba K. Udgata 4[4]

[1,2]*Department of Computer Science,Utkal University, Bhubaneswar, India*
[3]*Department of Computer Science Engineering and Application, IGIT Sarang, India*
[4]*WiSeCom Lab, School of Computer and Information Sciences, University of Hyderabad, India*

Received 1 Oct. 2022, Revised 6 May. 2023, Accepted 7 May. 2023, Published 1 Jul. 2023

Abstract: Mobile Crowdsensing (MCS) is a major source of a vast dataset containing heterogeneous types of data collected from various sources and stored in the local or remote server. Proper analysis of MCS data helps in better decision-making. However, MCS data suffers from data integrity issues, such as validity, accuracy, and reliability, that affect decision-making. Therefore, ensuring data integrity in the MCS environment is essential as it is a major source of a huge dataset. The proposed work considers user review data collection and analysis using a mobile application developed for the purpose. To ensure the data integrity, identification of fake and invalid reviews in the dataset need to be determined. This work proposes two approaches to solve data integrity issues. The first approach is to detect and eliminate fake/ invalid reviews from the dataset. The second is to identify the sources of fake/ invalid reviews and block them to protect the dataset from future fake reviews. Machine learning (ML) models are proposed to solve these issues and to ensure data integrity by filtering out fake reviews from real-time data sets. The proposed model uses data fuzzification over a purely mathematical model that categorizes users or customers as honest, suspicious, or malicious and their reviews/ feedback as genuine or fake using ratings provided by the user in the MCS Environment. Using the developed mobile application, user can give feedback about the desired location through various devices, which is stored in a cloud platform. The dataset can be analyzed through a fuzzy logic-based mathematical model followed by an ML algorithm and cost-benefit analysis to detect genuine reviews for maintaining data integrity. Further accuracy of the proposed models is compared with popular ML algorithms such as Naive Bayes (NB), Bayes Net(BN), Support Vector Machine(SVM), Decision Tree(J48), and Random Forest(RF). Initially, it achieves 99.79% of accuracy using the Random Forest algorithm that has been enhanced to 100% using cost-benefit analysis in cross-validation mode.

Keywords: Data Integrity, Mobile Crowdsensing(MCS), Review classification, Machine Learning, Rating, Fuzzy Model, Cost Benefit Analysis.

1. INTRODUCTION

Data integrity is defined as maintenance, assurance of completeness, consistency, safety, and accuracy throughout its life cycle. It is essential for any database or cloud system that stores, processes, and analyzes data. It must be secured and cannot be modified maliciously so that obtained information from the dataset will be reliable at any time. Various standards and rules have been designed to ensure data integrity. Data integrity ensures that information retrieved from the dataset will be reliable, complete, and accurate. It is essential to identify and eliminate invalid and fake data and also identify users who wants to temper dataset through invalid and fake data maliciously to ensure data integrity. It can be applied and studied through different real-time applications such as fake review analysis considered in the proposed work. Feedback/Review provided by the user after getting an experience on a particular thing is called a review. It may be obtained from users in online and offline modes. It may be positive or negative as per the user's experience with a specific product, place, person, etc. Positive reviews can enhance popularity, whereas; negative reviews can reduce the popularity of certain products, places, people, etc. Therefore It plays an important role in a majority of sectors. A review may be genuine or fake as per the user's intention. Some malicious users consistently provide fake reviews on specific products, places, people, etc., to compromise their popularity by increasing or decreasing. In the current era, Reviews given by the user are considered correct, and they blindly believe the feedback or reviews provided by other customers or users. So it needs to be reliable, real, accurate, and complete, but a few malicious

Email:ramesh0986@igitsarang.ac.in,sateesh1960@gmail.com,igitsethi@gmail.com,udgata@uohyd.ac.in

users who provide wrong reviews/feedback make it really difficult to achieve. Therefore, it is essential to identify and isolate fake feedback/reviews and the users who share fake reviews. Only accurate and honest reviews can provide reliable and trustable information that will maintain data integrity in the MCS environment. Data collection, data integration, and complete or detailed analysis of vast and various types of data received from a diversity of sectors or platforms, such as mobile devices, sensors, vehicles, buildings, and humans, are termed mobile crowdsensing(MCS) [1]. Data can be obtained from various users using various sources with the user's knowledge or without the user's knowledge.

Reviews/ observations provided by users or volunteers usually are considered genuine. Based on this assumption, certain information has been retrieved from this dataset and delivered to other users as per requirement. The information is obtained from the reviews delivered by different users/customers, and a user may deliver a genuine or fake review/feedback as per their purpose or experience. Therefore, it is highly required to analyze reviews/observations given by users to detect fake and accurate reviews and categorize users/customers as malicious or suspicious, or honest who give reviews. The information generated by analyzing the dataset contains reviews collected from legitimate or honest users only to be considered as accurate, reliable, and trustworthy. Detection and isolation of fake reviews are vital and essential areas for reliable online activity in the modern world. Various approaches and methods, like classification, deep learning, machine learning(ML), convolutional neural network(CNN), sentiment analysis, and feature extraction, are generally considered for the processing of data to detect and remove fake reviews for the isolation of genuine reviews. Review data given by the user and activities performed by users are also studied to identify fake and invalid reviews as well as malicious users.

Information published in online content must be authentic, complete, and accurate as it may influence society and individuals and their decision-making power positively or negatively. Unverified, incomplete, and wrong information can be considered false information, and its major contributors are easy-access, low-cost, and large-scale applications. Rumour, fake news, spam, fake reviews, misinformation, and hoax are major sources of false information [2]. False information like Fake news, rumors, hoaxes, etc., are the unverified information that has been considered accurate information on the web. It goes viral and spreads worldwide, affecting people or society's decisions, perceptions, and opinions. Different algorithms and approaches have been proposed for detecting fake reviews/ news in the recent past. A few important methods proposed in the literature are mainly based on data retrieving and labeling, finding context-based features [3], and unsupervised machine learning models for unlabelled data. Deep learning models [2], [4], different supervised, unsupervised, and semi-supervised machine learning algorithms [5], [3] are also proposed for detecting the fake reviews and to help the user in proper decision making.

A mathematical approach-based fake review detection framework has been presented in [6] to identify fake reviews for the isolation of genuine reviews. Further, it categorized the users into suspicious, honest, and malicious. But in this model, reviews given to the user has been estimated as either fake or correct review by comparing with estimated rating obtained using previous reviews. Instead of directly saying that reviews are either fake or correct, the probability of being a correct or fake review has been estimated using fuzzy logic in [7]. But the static weight factor has been used for activeness, incentive, and reliability level to determine the honesty level of the user. It might be the reason for having more malicious users than honest users. In the proposed work, dynamic weight has been taken instead of static one to resolve this.

The proposed model provides a mathematical approach to identifying and isolating fake and genuine reviews and also categorizes the users into three categories such as malicious, suspicious, or honest. Only reviews received from honest and legitimate users will be used for analysis to get information that can be given to forthcoming users per the requirement that helps in decision-making. In this work, feedback/reviews have been gathered from a variety of users through the developed smartphone-based android application or web-based application through various devices such as smartphones, laptops, tablets, desktops, etc., in the MCS environment for this purpose only. Further, it has been transmitted to the cloud database for storage and analyzed using the proposed mathematical model and fuzzy logic in the MATLAB environment for the categorization of users and the identification of genuine reviews and legitimate users, and also dataset has been generated. Further, it has been classified using various ML Algorithms to estimate the accuracy of the proposed model and cost-benefit analysis. It has been used to enhance the accuracy of the proposed model.

The rest of the paper has been presented as follows. In section 2, background and related work have been placed; then, the proposed methodology is discussed in section 3. Results have been discussed in Section 4, and finally, the conclusion is placed in section 5.

2. Background

A. Data Integrity

Data integrity is a challenge for cloud computing, data storage, security, and reliability. Cloud computing allows any user to store data in a cloud server remotely instead of on a local server in a cost-effective manner. Due to malicious activity, it may not be honest and fully trustworthy, as remote data can be corrupted at any time. Remote Data auditing methods such as private auditing, public auditing based on a third-party auditor(TPA), and Blockchain-based collaborative auditing methods [8] that eliminate the need for TPAs can be used to preserve the trustworthiness of data. Users delegate proxies to process and store data in a public cloud server as it is efficient and flexible. This model is based on proxy-oriented data uploading, and remote data integrity checking in a public cloud has been discussed in [9] to resolve security issues. The Internet of Things (IoT) relies on cloud computing for storage and computation to store and analyze vast data sensed by various sensors. Due to the limited capacity of smart products, any vulnerability, such as remote data integrity in the cloud, will affect its security and reliability. Various schemes based on RSA, BLS Signature mechanism, and ZSS signature [10] support privacy protection. Public auditing, Cryptographic-accumulator provable data possession (CAPDP) [11] based on RSA-based cryptographic accumulator that provides data dynamics and unlimited remote data integrity check with cost-effective in terms of communication, computation, and storage, a bilinear group based simple and efficient auditing service [12] has been employed to ensure remote data integrity and public verification of unreliable and outsourced storage to support dynamics of data. Similarly, Blockchain and Bilinear mapping-based Data Integrity Scheme (BB-DIS) has been discussed in [13] for large-scale IoT data without any Third-party Auditors(TPA). In [14], the authors described a stochastic blockchain-based data-checking scheme that can limit the number of cooperative nodes and distribute the load to IoT edge nodes to deal with limited computing and network resources. It also avoids network congestion and single-point failure due to the centralized architecture of IoT. Dual access control and data integrity verifiable (DCDV) scheme based on time and attribute has been discussed in [15] to provide fine-grained data access and ensure data integrity using attribute-based encryption in cloud-based industrial applications. Data integrity is essential for any distributed machine learning. Any modification, such as insertion, updation, and deletion of data by any network attacker, can affect the training and testing model to compromise the prediction and classification output. A secure and efficient Distributed machine learning-oriented data integrity verification scheme (DML-DIV) has been discussed in [16] to maintain the integrity of training data and testing data. Identity-based Remote data integrity checking (RDIC) scheme[17] uses the homomorphic verifiable tag to decrease the complexity of certificate management derived from public key infrastructure and also ensure data privacy in a cloud server.

B. *Data Integrity In Mobile Crowdsensing*

Data Integrity is essential for the isolation of real data in various environments. It is particularly highly needed for the mobile crowdsensing environment to ensure data integrity, which is our main objective. In MCS, data can be gathered from various sources, such as mobile devices, wearable devices, electronic gadgets, vehicles, buildings, human beings, etc., from different locations. They can be transmitted to the cloud platform for storage and analysis. Finally, it can be analyzed using various approaches such as mathematical models, artificial intelligence, neural network, and machine learning to get the required information to meet the requirements. Opportunistic reporting-based distributed and sustainable framework has been discussed in [18] to collect and store data in a cloud platform that minimizes sensing and reporting costs in a mobile crowdsensing environment. Various approaches for task management and incentive to preserve privacy in a mobile crowdsensing environment have been discussed in [19]. Humans can also act as social sensors[20] in a mobile crowdsensing environment due to various mobile devices; they are not only data consumers but also data producers as per their sensing needs.

C. *Standard Machine Learning Algorithms*

Machine learning is an Artificial Intelligence-based technology that enables computers to train and learn automatically from past data. Various Machine algorithms are used for building mathematical models for classification and regression using historical data or information. In this paper, five mostly used machine learning algorithms like SVM, Naïve Bayes, Bayes Net, Decision Tree(J48), and Random Forest are used for classification and to judge the accuracy of the proposed mathematical model.

a) SVM[21]

SVM is a well-known supervised machine learning(ML) algorithm implemented to solve various classification and regression problems. It contains a hyperplane that isolates datasets into different unique classes and data points known as support vectors that are used to define hyperplane, and hence it is named a support vector machine. It can handle both multiple continuous and categorical variables.

b) Naïve Bayes[21]

Naïve Bayes algorithm also comes to the supervised machine learning category that is based on the probability of the object, and it is based on one simple assumption that variables are independent of each other. A conditional probability-based Bayesian theorem has been applied, which means the likelihood that event (A) will happen when it is given that event (B) has already happened. It can analyze a huge amount of datasets easily using a Bayesian model. It is primarily used for text classification.

c) Bayes Net[22]

Bayes net, also known as Bayesian network (BN), is a machine learning(ML) algorithm implemented to solve classification and regression problems based on the probabilistic graphical model that gives focuses on knowledge regarding an uncertain domain where each node corresponds to a random variable, and each edge corresponds to the conditional probability for the corresponding random variables. It can be represented as a directed acyclic graph (DAG) where self-connection or loops are not allowed due to conditional probabilities and dependencies.

d) Decision Tree (J48)[21]

It is a supervised machine learning algorithm that can deal with both continuous and categorical variables. It generates a tree-like structure that includes nodes such as root nodes, leaf nodes, and branches, and it starts with the root node that expands on further branches until the leaf node. Features of the dataset will be represented by an interior node, whereas branches represent the decision rules, and finally, leaf nodes deliver the solution to the problem. Both classifications, as well as regression problems, can be solved using the Decision Tree(J48).

e) Random Forest[21]

Random forest is an ensemble learning technique based fastest supervised machine learning(ML) algorithm that can efficiently handle both missing and incorrect data to solve classification and regression problems. Multiple classifiers will be merged to enhance the efficiency of the model and deliver reliable predictions. It is a collection of multiple decision trees between 64 to 128 for a subset of a given dataset, and it considers the average to increase the predictive accuracy of the model. A larger number of trees leads to higher accuracy of the algorithm. Each decision tree delivers a classification result based on the majority votes; the Random Forest algorithm provides the final result to classify a new dataset.

D. Case Study Through Fake Review Analysis

Data Integrity can be studied through different real-time applications such as fake review analysis. Authors in [1] discussed a framework to detect fake reviews by using feature extraction of reviews. The Trust level of the user, activity during the review process, and social and personal behavior of the user need to be analyzed along with textual review for better results. Reviews and comments given to the product have been analyzed for detection of an outlier review in [23]. Another kind of new model for fake review detection is based on the semantic and emotional level of the reviewer as well as the density of reviews, which gives a much better performance than the traditional method, which is based on reviewer info, behavior, and textual review [24]. In [25], [26], authors described social behavior, affective, perceptual, and cognitive behavior like linguistic characteristics of the reviewer and multiple aspects of review inconsistency like content, rating sentiment, language, attitude, behavior consistency theories, and also estimate the impact of reviewer's location, distance as well as time on the reviews for fake review detection.

Blockchain-based with incentive mechanism internet of fake media things solution has been used in [27] to detect fake news and provide truthful news published in online mode by using blockchain technology, smart contracts. A weighted ranking algorithm has been used with a customized Proof-of-Authority consensus algorithm to provide incentives to motivate users. It works with multiple media types, like a hash of any text, audio, video, or image file. Fuzzy logic, Machine learning(ML), and artificial intelligence-based algorithms have been implemented on vast datasets based on social media for spam detection using neural network multi-layer perception to overcome the shortcomings of supervised ML algorithms through unsupervised approaches. Authors in [28] suggest that vast datasets can be analyzed quickly in less time using Fuzzy logic to minimize time and cost. It also eliminates the requirements of complex software for the detection of spam. A fuzzy modeling-based approach has been discussed for opinion spam detection that is based FSL detection algorithm and 81 no. of fuzzy rules. Authors in [29] discussed a fuzzy ranking evaluation algorithm that provides 80.77% accuracy for suspicious group detection, and it is based on four linguistic variables. Interval type 2 fuzzy set has been discussed in [30] that will provide better control for various categories of spam as well as personalization for detection of spam and classification of email.

Similarly Different approaches such as: Latent Dirichlet Allocation [31] on yelp dataset, ensemble model [32] based on data resampling method using meta classifier,textual based feature extarction dynamic random sampling techniques[33], multi feature feasion of features of labelled and unlabelled data [34], linguistic model [35] that extracts syntactic, grammatical, sentimental, and readability features of particular news, review grouping method [36] and Dynamic knowledge graph [37] using conditioned bidirectional long short-term memory(LSTM) algorithm, ,deep convolutional neural network [38] withensemble learning model [39] based on embedding LSTM, depth LSTM, LIWC CNN, and N-gram CNN, cognitive science [40],using replication research and sensitivity analysis [41] can be used to remove bias caused by fake learners have been used for review analysis for the detection of fake reviews in real time applications. Similarly different kinds of fake detection techniques may be placed in real-time applications used through different techniques, such as: using ensemble model based on data resampling method [32] using meta classifier, using multi-feature fusion of features using labelled and unlabeled data in [34],dynamic random sampling techniques based on textual based feature in [33], using Latent Dirichlet Allocation [31] on yelp dataset,using review grouping method [36] and Dynamic knowledge graph [37] using conditioned bidirectional long short-term memory algorithm,using linguistic model [35] that extracts syntactic, grammatical, sentimental, and readability features of particular news,using deep convolutional neural network [38] with ensemble learning model [39] based on embedding LSTM, depth LSTM, LIWC CNN, and N-gram CNN,using cognitive science [40],using replication research and sensitivity analysis [41] can be used to remove bias caused by fake learners. Similarly, Neural network algorithms can be used to establish a relationship between opinion fraud [42] and characteristics of social interaction and attention-based multilevel interactive [43]. This model integrates user, review text, product, and fine-grained aspects for fake review detection. Further, it has been attached with the emotion level of the user [44]to enhance the performance of the model. Geolocation-based account detection model [45] based on AdaBoost model supported by long short-term memory (LSTM) neural network used to find out the honesty of users/reviewers and the review delivered by users/reviewers to identify genuine reviews to maintain data integrity. XG Boost ensemble-based machine learning classifier and deep neural network model [46], [47] is used for classification in news content in a social context, whereas bidirectional encoder representations from transformers based deep learning approach [48] has been used to deal with ambiguity. The greatest challenge to natural language understanding.Convolutional neural network with a comparison of static word embedding with non-static word embedding used in [49] to remove irrelevant news.

For the isolation of genuine reviews from other online reviews, different supervised and unsupervised machine learning approaches [50] have been used for the classification and regression of real-time datasets to maintain data integrity by eliminating invalid data. A logistic regression algorithm has been used for an online spammer in [51] and achieves 88.3% accuracy.PU Learning-based classification algorithm [52] has been used to detect deceptive review-based classification done in [53], [54], [55], [56] using sentiment analysis, the behavior of reviewer and their reviewing style, drift detection, and text classification, and SVM algorithm gives better performance. Authors in [57], [58] use semantic analysis techniques using decision tables, information gain, XGBoost Model to identify and remove fake reviews for isolation of genuine and reliable reviews. Similarly, various machine learning algorithms like AdaBoost, SVM, Bagging Algorithms [59], [60], Hierarchical Attention Network (HAN), and visual image feature using image captioning and forensic analysis[61] for isolation of fake news from real news articles. Author in Hybrid deep learning model[62] a combination of convolutional and recurrent neural networks and Multi-layer Perceptron Model (MLP)[63] based on Convolutional Neural Network (CNN) and Bi-directional Long Short Term Memory (Bi-LSTM) used for classification of bad information or fake news. Local convolutional features and global semantic features [64] have been used to get semantic information from news article texts to classify it as fake or real. This also can be done using a graph-based neural network model[65] based on enhanced text representation using local and global sentence representation.CNN with generic pooling function [66] based deep profile can be used for classification of fake profiles to avoid invalid data in the online social network. Enhanced graph-based semi-supervised learning algorithm[67] that contains modules like data collection, feature extraction, classification, and decision making, and it uses a vast volume of data obtained from Twitter using scraps to detect fake users.

E. Comparative Analysis For Data Integrity

A reliable trust management scheme based on a mathematical approach has been discussed in [68] to categorize secondary users into honest, malicious, or suspicious based on spectrum sensing reputation in the cognitive radio network. Based on this concept, authors in [6] proposed a mathematical model with static weights factor for ensuring data integrity by filtering out fake and invalid reviews, isolating genuine reviews, and categorizing various users into Malicious, Suspicious, and Honest categories. But in this model, a range has been used by using the estimated rating of a location; if the review given by a user is placed in that range, the review will be considered genuine or otherwise fake for detecting and eliminating fake reviews from the dataset to ensure data integrity. But a review directly cannot be placed as genuine or fake. It should be the extent to which a review can be genuine or fake. Further, this mathematical model has been extended using Fuzzy logic with static weight factor in [7] to provide the probability of the review is genuine or fake. But due to the use of static weight factors, it is not able to efficiently categorize the users and reviews, and hence more users are considered malicious. Thus, most genuine reviews and honest users are placed in a malicious category. This will reduce valuable data used to maintain data integrity that provide reliable information. The proposed work has been influenced by these works and tried to improve it by solving the issues. It has been reflected in Table I.

Parameter	Article[6]	Article[7]	Proposed Model
Mathematical Model	Yes	Yes	Yes
Weight Factor	Static	Static	Dynamic
Fuzzy Logic	No	Yes	Yes
Machine Learning	No	No	Yes
Cost-benefit	No	No	Yes

TABLE I. Comparative Analysis for Data Integrity

F. Technical Contributions Of The Proposed Work

After observing all the above related works, we found the gaps in the research and proposed a novel method to address these gaps as discussed in the following;

- The novel model has been developed using a proposed fuzzy logic based mathematical model with a dynamic weight factor supported by machine learning and cost-benefit analysis for ensuring data integrity in the MCS environment through fake review analysis.
- Fuzzy logic has been used with the mathematical model to identify and eliminate all the fake and invalid reviews available in the dataset to have a dataset with valid, genuine, and accurate reviews. It will provide reliable information to forthcoming users whenever required during its life cycle. Thus it ensures data integrity in the MCS environment.
- The mathematical model with fuzzy logic has been used to categorize the users into honest, suspicious, and malicious categories to identify the valid and invalid sources of data. Data received from valid sources only will be accepted and added to the dataset, and the rest will be discarded without consideration to reduce complexity for ensuring data integrity.
- Further, classification is done using various machine learning algorithms such as SVM, Naive Bayes, Bayes Net, Random Forest, and Decision Tree(J48) to judge the model's efficiency.
- Cost-benefit analysis has been used to enhance the

accuracy of the model by minimizing errors. It has been used to choose the best ML algorithm for the proposed model so that it will be used to predict the category of user to identify the type of source of data. If Data is received from an invalid source, it will be rejected instead of added to the dataset. It will stop future contamination of the dataset with invalid and fake reviews that will reduce complexity for ensuring data integrity.

3. Methodology

A. Problem Statement And Motivation

Data integrity ensures that data stored in the database or dataset is complete, accurate, and valid, and it will provide exactly the same reliable information whenever retrieved. But due to the presence of invalid, noisy, and fake data, it is very difficult to ensure data integrity. This work is mainly aimed at maintaining data integrity in the MCS environment through location-based fake review analysis. The followings are the motivational factors of the proposed work.

1) Review dataset is a collection of reviews obtained from various users for any experienced place, product, or service that may be genuine or fake as per the user's intentions. Further, it will be analyzed to provide valid information about the product, service, or place to the forthcoming users. This information should be reliable, but due to the presence of invalid and fake data, this generated information is not completely valid and also reliable. It violates the principles of data integrity that should be handled efficiently to ensure data integrity.
2) Due to incomplete and unreliable information about the product, service, places, etc., Forthcoming users may not get the expected result which leads to the wrong decisions; further, the review dataset will be contaminated with the reviews based on the wrong decision. This will increase unreliable and inaccurate data in the dataset with the valid one that will enhance the complexity for ensuring data integrity.
3) Due to the presence of invalid, fake, and incomplete data in the database, the reliability of obtained information after proper analysis decreases, and it violates the integrity of data. Due to the maintenance of data integrity, data stored in the database should provide the same reliable information whenever it is retrieved during its life cycle. It is not possible due to the presence of fake and invalid reviews in the dataset.
4) It is highly essential to detect and remove fake and invalid data from datasets in order to get a dataset with valid and genuine review data that may provide reliable information whenever retrieved in order to ensure data integrity in the MCS environment.
5) It is also required to identify the valid and invalid sources of data to accept only valid reviews from valid sources, and the rest will be discarded to prevent future contamination of the dataset with invalid and fake reviews. It will help to reduce complexity by filtering out invalid and inaccurate data instead of adding to the dataset to maintain data integrity in the MCS environment.

It has been tried to solve these issues with the help of a proposed fuzzy logic based mathematical model backed up by Machine learning and cost-benefit analysis in order to ensure data integrity in the MCS environment. Two approaches have been considered to solve these issues. First, it has been tried to detect and eliminate all fake reviews from the dataset to have only genuine, valid and complete reviews that will provide reliable information throughout the life cycle that will ensure data integrity. Second, Users have been categorized using the proposed mathematical model followed by the machine learning algorithm with the cost-benefit analysis for identification of invalid sources so that all the data received from invalid sources will be discarded instead of added to the dataset in the future. It will reduce complexity for ensuring data integrity.

B. Proposed Model

A real-time dataset has been obtained for ensuring data integrity using collected review data from different users using a developed web-based application or android-based smartphone application through multiple devices, such as mobile devices, laptops, tablets, desktops, etc., in a mobile crowdsensing environment. The proposed model has the following objectives.

1) Gathering review data from users using a web-based application or android app in the on-site or off-site mode that may contain fake and genuine reviews as per the user's intention to have a real-time dataset for ensuring data integrity.
2) Storage and analysis of real-time dataset in the cloud platform using fuzzy logic based mathematical model backed up by machine learning and cost-benefit analysis to detect fake and genuine reviews and also categorize the users to detect or identify fake reviews and sources of fake reviews for ensuring data integrity in the dataset.
3) Detection and removal of fake and invalid reviews for isolation of genuine reviews in the dataset for ensuring data integrity using fuzzy logic in order to provide reliable information that came from genuine reviews only for forthcoming users.
4) Trustworthiness of users has been estimated on the basis of their reliability, incentive, and activeness levels for categorization of users as suspicious, honest, and malicious, and also associated reviews as genuine and fake to identify the fake reviews and sources of fake reviews for ensuring data integrity. Reviews collected from honest/legitimate users only will be treated as genuine reviews, and they will be added to the dataset, and the rest will be discarded to stop future contamination of the dataset with fake reviews, and it will reduce complexity for ensuring data integrity by filtering out invalid or fake data

instead of adding to the dataset. It may ensure data integrity in the MCS environment as it first filters out fake reviews from the dataset and also prevents future contamination of the dataset with fake reviews by the categorization of sources of reviews.

5) Finally efficiency of the proposed model has been judged using various standard and well-known Machine Learning(ML) Algorithms in robust cross-validation mode and also tried to enhance the model's accuracy for ensuring data integrity in the MCS environment using cost-benefit analysis that focuses on minimizing the errors available in the dataset. The best ML algorithm has been estimated to train the model and predict the type of user for the identification of valid and invalid sources of data. It is required to stop future contamination of the dataset with fake reviews for ensuring data integrity.

C. *Proposed Architecture*

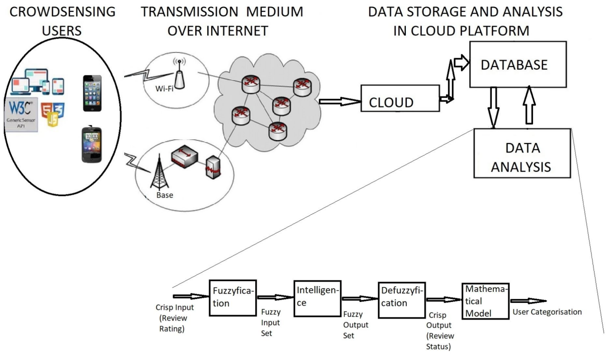


Figure 1. Architecture of the proposed system

The proposed work has been focused on ensuring data integrity in the MCS environment through location-based fake review analysis using the fuzzy logic based mathematical model backed up by machine learning and cost-benefit analysis. Its architecture is reflected in Figure 1. In the proposed work, review/feedback data has been gathered from various crowdsensing users for different tourist places through the developed android-based app and web-based application, and it has been stored and analyzed in the cloud platform for ensuring data integrity in the MCS environment. Further, it has been analyzed using the fuzzy logic based mathematical model and standard machine learning algorithms for filtering genuine reviews by removing fake reviews from the real-time dataset to ensure data integrity. Cost-benefit analysis has been used to enhance the accuracy of the proposed model to enforce data integrity by minimizing errors available in the dataset after classification. The trustworthiness of the user and his reviews has been computed to categorize the users as; suspicious, honest, and malicious and received reviews as genuine or fake for ensuring data integrity. Hence all review data received from malicious and suspicious users will be discarded to prevent future contamination of the review dataset with fake reviews and also to reduce the complexity of maintaining data integrity. Since only valid, complete, genuine, and accurate data is available in the dataset, it will generate reliable information for forthcoming visitors of the location whenever retrieved. It will ensure data integrity in the MCS environment. The architecture of the proposed work has been presented in three stages; Crowdsensing Users, Data transmission from crowdsensing users to the cloud platform for storage and analysis, and finally, data analysis for ensuring data integrity.

a) *Crowdsensing Users*

An android-based app for smartphones and a web-based application have been developed for review data collection to have a real-time dataset for ensuring data integrity in the MCS environment from various users as per their experience after visiting any tourist place or any location. It can be deployed through different electronic gadgets such as smartphones, laptops, tablets, desktops, palmtops, etc., to provide valuable feedback/review for their visited location in the MCS environment. Initially, a user will register with the developed android application or web-based application by providing information like Phone No., full name, email address, occupation, date of birth, gender, and address for review submission. During visiting a location or a tourist place, a user can provide review/feedback about the visited place as per their experience through the android app or web-based application in either onsite mode in which location has been automatically fetched through a GPS sensor from the desired location during the visit or in the offsite mode in which user will enter detail about visited location manually after the visit. Review data contains various information about the visited location, such as the ID of the user, location information in terms of longitude and latitude, rating between 1 to 5, communicative language, and any additional information.

b) *Review Data Transmission from User to Cloud Platform for Storage and Analysis*

For review data transmission to the cloud platform, a user must have dedicated internet connectivity through Bluetooth tethering, Wi-Fi, or a cellular network. After getting connected to the internet, a user may use the app or web-based application to submit review/feedback for the desired visited location, and it will be transmitted to the cloud platform for storage and analysis to have a real-time dataset for ensuring data integrity in the MCS environment.

c) *Data Analysis for Ensuring Data Integrity*

A real-time dataset that is a collection of fake and genuine reviews has been obtained from various users through the designed android app and web-based application for ensuring data integrity. Further, this real-time dataset has been organized into a comma-separated value(CSV) format for analysis using the proposed fuzzy logic based mathematical model in the MATLAB environment. Initially, the Estimated Rating(ERT) of the location has been determined to find the approximate actual rating of the location. Then it is compared with the rating delivered by the user using the fuzzy concept to categorize the review as fake or genuine. It helps to detect and remove all fake reviews from the dataset to have only genuine reviews in the dataset that will provide reliable information whenever retrieved to ensure data integrity. Based on the number of fake and genuine reviews, the user's reliability level has been estimated to find the extent to which a user is reliable. The user's activeness level has been estimated to estimate the review activity of the user. Further, some incentives, based on the less suspicious behavior of the user, have been assigned to the user to encourage him to consistently participate in the review process with valid reviews for ensuring data integrity. Users' trustworthiness has been determined on the basis of their reliability, activeness, incentive levels, and their respective dynamic weight factors. It is used to categorize users into honest, suspicious, and malicious categories for maintaining data integrity through the proper identification of valid and invalid sources of data. Reviews received from the honest user only will be considered genuine reviews, and further, they will be added to the real-time dataset, and the rest of the reviews will be discarded to stop future contamination of the real-time dataset with fake reviews and also to reduce the complexity for ensuring data integrity. Due to the presence of only genuine, valid, and complete review data in the dataset, it will provide reliable information to the forthcoming user of the location whenever asked. Hence, it will ensure data integrity in the MCS environment. A labeled dataset has been generated using the proposed fuzzy logic based mathematical model and real-time dataset to judge the efficiency of the proposed model using standard ML algorithms like SVM, Naive Bayes, Decision Tree(J48), Bayes Net, and Random Forest Algorithm in robust Cross-Validation mode. Due to the presence of errors in the dataset, accuracy decreases; therefore, a cost-benefit analysis has been done to enforce data integrity using the virtual screening technique to enhance the accuracy of the proposed model. The best ML algorithm has been estimated using machine learning and cost-benefit analysis to train the model and predict the type of users for identification of valid and invalid sources of data to stop future contamination of the dataset with fake reviews for ensuring data integrity.

D. Data Collection Framework

Review data will be received from legitimate users as information about their experiences with a product, service, or place in off-site or on-site mode to have a real-time dataset for ensuring data integrity. In on-site mode, immediate feedback will be received from the registered user just after the experience. In off-site mode, feedback will be obtained about their experience after some period, not immediately. Users' review or feedback data contain user ID, coordinates of the location in terms of longitude and latitude, current address, the communicative language used by people at that location, remark about that location, and a rating in the range of 1 to 5 obtained. Users can also give their feedback in terms of like, dislikes, averages, and rectifications required, which will be accepted as consolidated feedback regarding the desired visited place using various smart devices such as smartphones, tablets, laptops, etc. Further, it will be stored in a real-time cloud database and analyzed using the proposed mathematical model based on fuzzy logic for identification and isolation of genuine reviews from the dataset for ensuring data integrity and also compute the honesty level of users for categorization of users in the MCS environment. In the proposed work, 51 different users deliver multiple reviews for 42 different locations.

E. Data Pre-processing And Analysis

Review data available in the cloud platform will be retrieved and restructured in comma-separated value(CSV) format and stored in a text file for analysis. It is analyzed using the proposed mathematical model based on fuzzy logic in the MATLAB environment.

F. Estimation Of Average Rating(ERT) of Location

Review data will be analyzed using the maximum likelihood approach to compute the average rating of location/place that will be considered as the true rating of that location approximately. It is determined using the following equation.

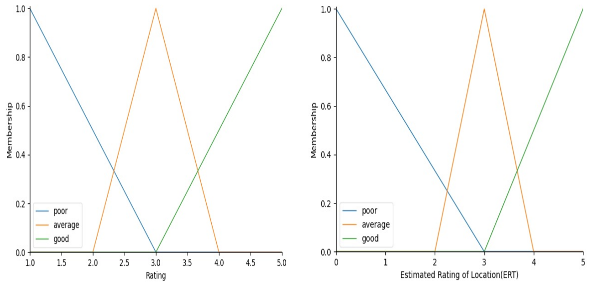
$$E_l = \frac{\sum_u^n L_u}{n} \tag{1}$$

In Eq.- 1, n represents no. of users, and u represents the unique user ID that is in between 0 to n. The rating provided by the user u for the desired location l is reflected by L_u. The ERT of location l is represented by E_l and is determined by taking the mean of all ratings delivered by all the users for the desired location l.

G. Reliability Level Of The User Based On Location

The reliability level of the user reflects the extent to which a user and his review are reliable. It is determined by comparing the rating provided by the user for the desired location with the ERT of the particular location determined in Eq.1 using fuzzy logic. If they are closer, then the review will be assumed as genuine review, and the user's reliability level will increase; otherwise, the review will be assumed as fake, and the user's reliability level will decrease.

Rating provided by the user for a certain location and ERT of that location will be compared to determine the review status of the feedback/review delivered by the user in the range of 0 to 1 using fuzzy logic. A threshold value between 0 to 1 for review status(δ) will be determined. Further, obtained review status as per review will be compared with δ; if it is higher, then the review will be genuine, the no. of correct reviews will enhance by 1, and also user's reliability level will increase; otherwise, it is fake and as a result no. of fake reviews will enhance by 1, and also user's reliability level will decrease. A user may provide multiple different reviews from the same location in different time intervals. Therefore, Initially, the user's location-wise reliability level has been determined, then the average of the user's location-wise reliability level has been estimated to determine the user's reliability level through the proposed mathematical model based on fuzzy logic in a mobile crowdsensing environment. It reflects the trustworthiness of the user and his review in the review process. It will also help to detect and eliminate fake reviews from datasets in order to have genuine reviews only for maintaining data integrity in the MCS environment.



(a) Fuzzy membership value of rating given by user for a location

(b) Fuzzy membership value of Estimated Rating of location

Figure 2. Fuzzy membership value for Input

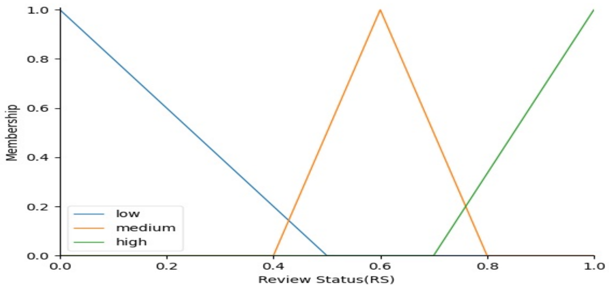


Figure 3. Fuzzy membership value of Review Status of User(Output)

Membership values of the rating provided by the user for a certain location in between 1 to 5 using fuzzy logic have been reflected in Figure 2 (a). Rating delivered by the user has been represented by L_u, and good, average, and poor are the three membership parameters that have been made for the rating provided by the user through the Triangular membership function(trimf) [69] of fuzzy logic as per Eq-2, Eq-3, Eq-4 [7].

$$L_u['poor'] = trimf(L_u, (1, 1, 3)) \tag{2}$$

$$L_u['average'] = trimf(L_u, (2, 3, 4)) \tag{3}$$

$$L_u['good'] = trimf(L_u, (3, 5, 5)) \tag{4}$$

Membership values of ERT of a certain location that is in between 0 to 5, determined using Eq.-1 using fuzzy logic, have been reflected in Figure 2 (b). Fuzzy membership function trimf has been considered to bring out three membership values, good, average, and poor, for the ERT of the location using Eq-5, Eq-6, Eq-7 [7]. E_l represents the ERT of a certain location.

$$E_l['poor'] = trimf(E_l, (0.0, 0.0, 3.0)) \tag{5}$$

$$E_l['average'] = trimf(E_l, (2.0, 3.0, 4.0)) \tag{6}$$

$$E_l['good'] = trimf(E_l, (3.0, 5.0, 5.0)) \tag{7}$$

Membership parameters of review status for the rating delivered by the user for a certain location through fuzzy concept have been reflected in Figure 3. R_s represents the review status of the review delivered by the user, and it is in between 0 to 1. Three membership parameters, high, medium, and low, have been made using the trimf fuzzy membership function of fuzzy logic for the review status of the user using Eq-8, Eq-9, and Eq-10 [7].

$$R_s['low'] = trimf(R_s, (0, 0, 0.5)) \tag{8}$$

$$R_s['medium'] = trimf(R_s, (0.4, 0.6, 0.8)) \tag{9}$$

$$R_s['high'] = trimf(R_s, (0.7, 1, 1)) \tag{10}$$

Algorithm 1 Fuzzy rule Generation for computation of Reliability level of the user

```
Require: Rating given by user for the particular location($L_u$) and estimated rating of the location($E_l$) for input and Review Status($R_s$) for output.
Ensure: For the desired particular location, $L_u$ and $E_l$ has been estimated .
  if $L_u$ is poor then
      if $E_l$ is poor then
          $R_s$ is high
      else if $E_l$ is average then
          $R_s$ is medium
      else
          $R_s$ is low
      end if
  else if $L_u$ is average then
      if $E_l$ is poor then
          $I_{ch}$ is medium
      else if $E_l$ is average then
          $I_{ch}$ is high
      else
          $I_{ch}$ is medium
      end if
  else
      if $E_l$ is good then
          $I_{ch}$ is high
      else if $E_l$ is average then
          $I_{ch}$ is medium
      else
          $I_{ch}$ is low
      end if
  end if
```

Fuzzy rules have been generated to estimate the review status that represents the probability that the review is correct. For this, two input parameters, the rating provided by the user for the location(L_u) and the estimated rating of that location(E_l), have been considered as input and Review status(R_s) as output. If L_u is closer or equal to E_l, then R_s will be high; otherwise, it will be either low or medium as per the difference between L_u and E_l. It has been reflected in Algorithm 1.

A set of 9 fuzzy rules has been used in the proposed model to compare the rating delivered by the user(L_u) with the ERT(E_l) of a certain location to determine the review status(R_s) of the delivered review. Three membership parameters for fuzzy logic are good, average, and poor for the rating (L_u) and ERT(E_l) of a certain location. Three membership parameters for review status(R_s) are high, medium, and low. Following Fuzzy Rules from R1 to R9 have been generated using membership parameters of input(L_u and E_l) and output(R_s) as per Algorithm 1.

R 1 :- L_u['poor'] and E_l['poor'] $\implies$ R$_s$['high']
R 2 :- L_u['average'] and E_l['average'] $\implies$ R$_s$['high']
R 3 :- L_u['good'] and E_l['good'] $\implies$ R$_s$['high']
R 4 :- L_u['poor'] and E_l['good'] $\implies$ R$_s$['low']
R 5 :- L_u['good'] and E_l['poor'] $\implies$ R$_s$['low']
R 6 :- L_u['good'] and E_l['average'] $\implies$ R$_s$['medium']
R 7 :- L_u['average'] and E_l['good'] $\implies$ R$_s$['medium']
R 8 :- L_u['poor'] and E_l['average'] $\implies$ R$_s$['medium']
R 9 :- L_u['average'] and E_l['poor'] $\implies$ R$_s$['medium']

The user's Rating(L_u) of a certain location will be matched with ERT(E_l) of that particular location to determine the review status(R_s) for the delivered review using fuzzy logic, and it is in between 0 to 1. The impact of review status(R_s) on the identification and isolation of genuine reviews by identification and removal of fake reviews, and also on the user categorization process, has been observed through different observations. The threshold value for review status(δ) has been estimated. The review status(R_s) for the delivered review will be compared with δ; if it is higher, then the delivered review is honest/genuine, and location-wise no. of correct reviews (NCR_L) will be incremented by one; otherwise, treated as fake, and the no. of fake review (NFR_L) will be incremented by one. The user's reliability level has been determined using NCR_L and NFR_L as per the following equations.

$$R_{uL} = \frac{NCR_L}{NCR_L + NFR_L} \tag{11}$$

In Eq.-11, Location wise Reliability level of the user is R_{uL} and it is determined using NFR_L and NCR_L.

$$R_u = \frac{\sum_{L=0}^{n} R_{uL}}{n} \tag{12}$$

In Eq.-12, Reliability level of the user is represented by R_u and computed using average of R_{uL} and total no. of users(n).

H. Activeness Level Of User

The user's Activeness level reflects the consistent involvement of the user in the review process. During the review process, if a user participates by delivering a review for the desired location, then N_p (No. of times participated) will be increased by 1; otherwise, the user doesn't participate, and N_a(No. of times not participated) will enhance by 1. The user's activeness level will be determined using N_p and N_a as per the equations Eq.-13 and Eq.-14;

$$U_p = \begin{cases} 1, & \text{if User involved in review process for various locations} \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

In Eq.-13, During the review process, the participation of the user has been represented by U_p. U=1 means the user delivered a review and participate, and 0 means the user did not provide the review.

N_p will enhance by 1 if U_p is 1, otherwise N_a enhance by 1.

N_p and N_a have been used to compute the user's Activeness level(A_u) as per the following Equation.

$$A_u = \frac{N_p}{N_p + N_a} \tag{14}$$

I. Incentive Level Of User

Incentives are nothing but the additional benefits provided to the user for active participation and consistent, reliable reviews in the review process. Users can give genuine or fake reviews as per their intention. Genuine reviews can enhance the incentive, whereas fake reviews reduce it. Review status has been computed using fuzzy logic to categorize it as genuine or fake. Finally, the user's suspicious level will be estimated using the no. of fake and genuine reviews given by the user. The incentive given to the user is inversely proportional to the suspicious behavior of the user. If the suspicious behavior of the user is more, then less incentive will be given and vice versa. Therefore incentive level of users can be computed using the suspicious level and activeness level of users using the following equations. It encourages users to provide genuine reviews in order to earn more incentives. It also helps to detect honest users and correct reviews that can be isolated from the rest of the dataset to deliver accurate information and also ensures data integrity in the MCS environment.

The user's suspicious level will be determined using no. of genuine review/feedback and fake review/feedback using fuzzy logic as per the equations mentioned below.

Rating delivered by the user has been represented by L_u, and good, average, and poor are the three membership parameters that have been made for the rating provided by the user using the trimf fuzzy membership function[69] using Eq-15, Eq-16, Eq-17 [7].

$$L_u['poor'] = trimf(L_u, (1, 1, 3)) \tag{15}$$

$$L_u['average'] = trimf(L_u, (2, 3, 4)) \tag{16}$$

$$L_u['good'] = trimf(L_u, (3, 5, 5)) \tag{17}$$

The user's Rating(L_u) of a location will be matched with the ERT(E_l) of that particular location through the fuzzy concept to determine the review status(R_s) of the user. The threshold value(δ) has been determined for review status(R_s) through various observations. Further, review status(R_s) will be compared with δ. If it is higher, then the review is genuine/honest, and the no. of correct reviews (NCR) will increase by 1; otherwise, it is fake, and as a result, the no. of fake reviews (NFR) will enhance by 1. S_u reflects user's Suspicious level, computed as per Eq-18.

$$S_u = \frac{NCR}{NCR + NFR} \tag{18}$$

In Eq.-18, the suspicious level(S_u) of users has been determined using no. of fake and genuine reviews delivered by the user.

The incentive provided to the user depends on the reliability and consistency of the user in the review process. The reliability level is reciprocal of the user's suspicious level. Therefore it will be estimated using the user's activeness and suspicious levels.W_a and W_r weight factors will be assigned to the user's activeness level and user's suspicious level, respectively, and it will be computed using Eq-19 and Eq-20.

$$W_a = \frac{A_u}{A_u + (1 - S_u)} \tag{19}$$

$$W_r = \frac{(1 - S_u)}{A_u + (1 - S_u)} \tag{20}$$

$$I_u = W_a * A_u + W_r * (1 - S_u) \tag{21}$$

In Eq.-21, I_u reflects the amount of incentives/benefits provided to the different users based on their current performance during review process.

J. *Estimation Of Honesty Level Of User*

The honesty level of users reflects the extent to which a user and his review are trustworthy. It is in the range of 0 to 1. The review given by the user will be reliable and correct if the honesty level of the user is better. Therefore it also helps to identify and isolate genuine reviews by detecting fake reviews and honest users to maintain the integrity of data in the MCS environment. In this section, dynamic values of weight factor for Incentive, Activeness, and Reliability level of the user have been considered instead of fixed values, that is considered in [6]. W_1, W_2 and W_3 are the weight factors for the Reliability level (R_u), Activeness level (A_u), and Incentive level (I_u), respectively, and it is computed using the given equation in Eq.22 to Eq.24.

$$W_1 = \frac{R_u}{R_u + A_u + I_u} \tag{22}$$

$$W_2 = \frac{A_u}{R_u + A_u + I_u} \tag{23}$$

$$W_3 = \frac{I_u}{R_u + A_u + I_u} \tag{24}$$

The user's honesty level (H_u) can be computed using dynamic weight factors and different performance metrics such as Incentive, Activeness, and Reliability level of the user placed in Eq.25 [6].

$$H_u = W_1 * R_u + W_2 * A_u + W_3 * I_u \tag{25}$$

K. *Categorization Of The User*

Various users will be categorized as Malicious, Suspicious, and Honest users on the basis of the Honesty level(H_u) of the user. The reliability and suspicious level median has been computed in Eq.(26) and Eq.(27). median of the user's reliability level will be used as a threshold for the Honest user(H_t), and the median of the suspicious level of the user will be used as a threshold for the Malicious user(S_t). If the honesty level of the user is greater than H_t, then the user will be considered as an Honest user, whereas if the honesty level of the user is less than S_t, then the user will be considered as Malicious one otherwise user will be treated as a Suspicious user.

$$H_t = \begin{cases} O_u^1[\frac{u}{2}], & \text{if u is even} \\ O_u^1[\frac{u-1}{2}] + O_u^1[\frac{u+1}{2}], & \text{u is odd .} \end{cases} \tag{26}$$

In Eq.-26, H_t represents median of ordered list of Reliability level of user. O_u^1 is the ordered list of Reliability level of user.

$$S_t = \begin{cases} O_u^2[\frac{u}{2}], & \text{if u is even} \\ O_u^2[\frac{u-1}{2}] + O_u^2[\frac{u+1}{2}], & \text{u is odd .} \end{cases} \tag{27}$$

In Eq.-27, S_t represents median of ordered list of suspicious level of user. O_u^2 is the ordered list of Suspicious level of user.

$$U_t = \begin{cases} Honest, & \text{if } H_u > H_t \\ Suspicious, & \text{if } S_t < H_u \leq H_t \\ Malicious, & \text{if } H_u \leq S_t \ . \end{cases} \tag{28}$$

In Eq.-28, Type of user will be represented by U_t. Various users will be categorized as Malicious, Suspicious, and Honest on the basis of the honesty level of the user and threshold values H_t and S_t. H_u represents the honesty level of the user, whereas H_t and S_t represent the median of the reliability level and suspicious level of users, respectively.

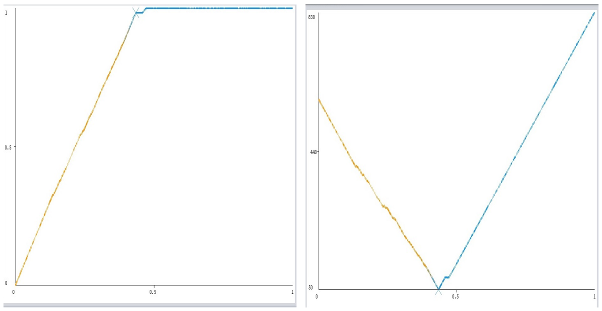
L. *Classification Of User*

Classification is a process of understanding, identifying, and grouping of objects into various identical categories in machine learning. Naïve Bayes, Bayes Net, Support Vector Machine(SVM), Decision Table, Decision Tree(J48), Random Forest, etc., are some popular and extensively used machine learning algorithms particularly used for the classification of vast datasets into respective and relevant classes/categories using these pre-categorized training datasets. The review dataset has been generated in CSV format using review data obtained from various users using the developed app and web-based application through various electronic gadgets, such as; smartphones, tablets, laptops, etc., in the mobile crowdsensing environment. Further, it has been analyzed through the proposed mathematical model based on fuzzy logic in MATLAB environment for identification and isolation of genuine reviews from the dataset by detecting fake reviews and also categorized users as honest, suspicious, and malicious to achieve reliability and accuracy of information obtained in the mobile crowdsensing environment and also to ensure the integrity of data so that information would be same for any time interval. Finally, it is classified using various machine learning algorithms such as Naïve Bayes, Bayes Net, SVM, Random Forest, and Decision Tree(J48) to estimate the accuracy of the proposed model in WEKA Environment and also cost-benefit analysis has been done to increase the accuracy level of the proposed model.

The review dataset set will be divided into the training set and testing set for classification through various modes such as full training set, cross-validation, and percentage split modes in the weka environment. In Full training set mode, the dataset will be partitioned into the training set and testing set only once by shuffling the dataset to create only one model for classification. Similarly, as per the user's input, a certain percentage of the dataset will be used for testing, and the remaining for the training set to generate only one model for classification in percentage split mode. But in cross-validation mode, multiple times, a dataset will be partitioned into the training set and testing set in all possible ways to generate multiple models for classification instead of only one. Overfitting and underfitting are the two major issues that generally occur due to the size of the training set in machine learning. Overfitting occurs due to more data being used for training than required, whereas the use of fewer data than required in the training dataset results in underfitting. Chances of getting overfitting and underfitting are more in the case of full training set mode and percentage split mode due to the consideration of only one model with one pair analysis of the training set and testing set of the dataset. But in the case of cross-validation mode, the dataset will be partitioned into the training set and testing set multiple times in all possible ways, and multiple models have been generated, and the best one has been selected in order to overcome the issues related to overfitting and underfitting. The dataset has been classified in cross-validation mode instead of the full training set, and percentage split mode as cross-validation mode is more robust, but accuracy is a little bit less than the full training set and percentage split mode in the proposed model. True Positive(TP) Rate, False Positive(FP) Rate, False Negative(FN) Rate, True Negative (TN) Rate, Precision, Recall, F Measure, Accuracy level, Mean Absolute Error(MAE), Root Mean Squared Error(RMSE), Kappa Statistics, etc. are various parameter considered for comparison of various Ml algorithms considered for analysis through classification.

M. Cost And Benefit Analysis

True Positive (TP) and True Negative (TN) reflect truly classified data, whereas False Positive (FP) and False Negative (FN) represent wrongly classified data. The cost reflects the number of FPs and FNs. The cost of each FN and FP data is considered to be Rs. 1 as well as each TP and TN are considered as Rs 0. Therefore Cost-benefit analysis is used to minimize cost, which means minimizing the number of FNs and FPs by converting them to TP and TN through virtual screening. Benefit represents no. of FP and FN records which are converted to TP and TN. Due to this Cost-Benefit Analysis, the accuracy of the model will be increased due to the reduction of wrongly classified data. Virtual screening is a collection of computational methods which is used for analyzing and evaluating vast datasets to get the desired set of records. It uses a portion of the entire population of data for training to create a pattern for classification and applies this to the entire population to identify the required record. In the case of the classification algorithm, it can be used to identify the record as TP, TN, FP, and FN.



(a) Threshold Curve

(b) Cost Benefit Curve

Figure 4. Threshold and cost benefit curve of honest user in Naïve Byes model

Figure 4(a) is termed as the threshold graph in which the Y-axis represents TP Rate, and the X-axis Represents Sample size. This graph represents that 43.4629% of the population is used for virtual screening, reflected in orange color used for minimizing wrongly classified data by converting FPs and FNs into TPs and TNs marked with orange color and achieving 98.2906% of the target. This graph also reflects the position at which the accuracy of the model is highest by minimizing wrongly classified data using the X symbol.

Figure 4(b) is termed as the cost-benefit curve in which the Y-axis represents cost, and the X-axis represents the sample size. The above figure reflects that the earlier cost was 585, but now it is minimized to 50. 585 FPs and FNs were reduced to 50 FPs and FNs, and the remaining 535 were added to TP and TN, which enhanced the accuracy of the model from 58.6572% to 96.4664%.

4. Experimental Results And Discussions

A real-time review dataset has been obtained to ensure data integrity from various crowdsensing users through a developed android application and web-based application and stored in the cloud platform for storage and analysis in the MCS environment. This real-time dataset may contain fake and genuine reviews as per the intention of the user. Further, it has been analyzed using the proposed fuzzy logic based mathematical model to categorize the reviews as genuine and fake and also to categorize the users as Malicious, Suspicious, and Honest using the estimated trustworthiness of users for ensuring data integrity. It has been done to ensure data integrity by isolating genuine reviews only in the dataset through the detection and removal of fake reviews that will provide reliable information whenever retrieved and also for the identification of valid and invalid sources of review data to reduce the complexity for ensuring data integrity by filtering out invalid and fake reviews instead of adding to the dataset. Due to the presence of genuine reviews, it will provide reliable, valid, accurate, and complete information to forthcoming users whenever retrieved throughout its life-cycle, which will ensure data integrity in the MCS environment. Review data obtained from valid sources only will be treated as genuine reviews and added to the review dataset, and the rest will be treated as fake and discarded to prevent future contamination of the dataset with fake reviews in order to reduce complexity for ensuring data integrity. It has been discussed through proper analysis in section 4.A. Further, the efficiency of the proposed model has been judged using various standard and well-known ML algorithms in cross-validation mode, and it has been discussed in section 4.B. Due to the presence of error accuracy of the model decreases, and these errors have been detected and solved using cost-benefit analysis to enhance the accuracy of the proposed model and enforce data integrity. It has been discussed in detail with data in section 4.C.

A. *Categorization Of User*

In this section, the efficiency of the proposed model for ensuring data integrity has been discussed with the proper analysis that is used for categorization of the users as malicious, suspicious, and honest and also the received reviews as genuine and fake as per the trustworthiness of users in order to have a dataset with valid and genuine reviews only that will provide reliable, accurate and complete information to forthcoming users. Also, it is used to identify the valid sources of review that will be added to the review dataset in the future, and the rest that came from invalid sources will be discarded to prevent future contamination in order to reduce complexity for ensuring data integrity. It is used for maintaining data integrity in the MCS environment. The ERT of the location has been estimated to find the approximate correct rating of the location. Further, It is compared with the review submitted by users to categorize the review as genuine or fake and also to estimate the user's reliability level. The user's activeness level has been estimated to judge their activity in the review process. Some incentives have been assigned to users to encourage them for consistent participation and provide reliable reviews. Finally, the user's honesty level has been estimated using the user's reliability, incentive, and activeness levels and the associated dynamic weight factor that reflects the trustworthiness of users.

Figure 5 Provides a visual comparison among mathematical models with static weight factor[6], Fuzzy based mathematical model With Static Weight Factor Model[7], and the proposed model in which no. of users(in %age) represented by the x-axis and user type by the y-axis. The no. of Honest, Malicious, and Suspicious users has been determined after identification using Eq.-28. Finally, this result will be compared with the number of Honest, Suspicious, and Malicious users obtained as per the Conventional mathematical model without fuzzy[6] and the Fuzzy-based mathematical model with Static Weight Factor Model[7]. It is observed that the number of honest users increases while the number of malicious and Suspicious users decreases in the proposed model as compared to the Conventional mathematical model without fuzzy[6] and the Fuzzy-based mathematical model with Static Weight Factor Model[7].

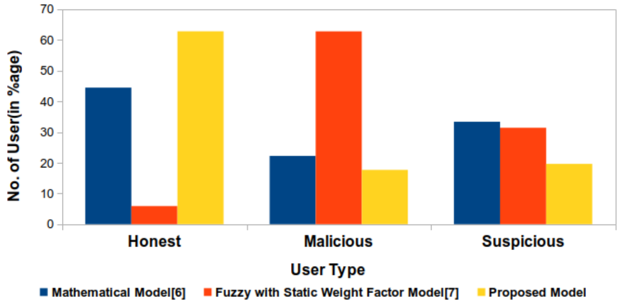
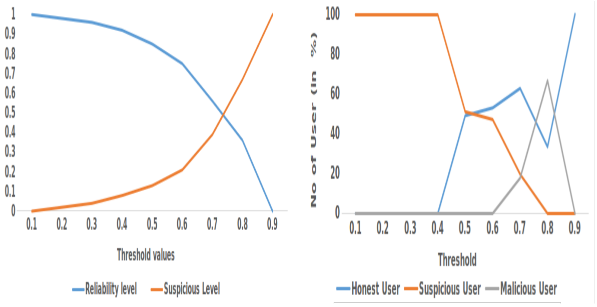


Figure 5. Compare with Existing Model



(a) Review status's impact on Reliability and Suspicious level

(b) Review status's impact on categorisation of user

Figure 6. Estimation of threshold value of review status

The impact of review status's of reviews on the user's reliability and suspicious levels and also in the categorization of the user has been reflected in Figure- 6. The output of the fuzzy model is the Review status that varies from 0 to 1. if the review status is more than 0.4, then the user's reliability level reduces, and the user's suspicious level enhances, and finally, both intersect each other at the review status of 0.75 as observed in Figure- 6 (a) and (b). Initially, all users are considered as suspicious users for the value of review status in between 0 to 0.4, but beyond 0.4 number of honest users increases. Similarly, the no. of malicious users enhances after the value of the review status more than 0.6, and it is minimized to 0 after the value of the review status more than 0.8. Since at review status=0.7, all three types of users are available. Therefore, the threshold value for the review status(δ) might be considered as 0.7. It is used to detect the reviews as genuine or fake. If the review status obtained using Eq.- 8 to 10 is higher than δ then the review will be treated as genuine otherwise fake. It plays a crucial role in the detection and removal of fake reviews from the dataset to have a dataset with genuine, complete, and valid reviews that will provide reliable information throughout its life cycle and ensure data integrity.

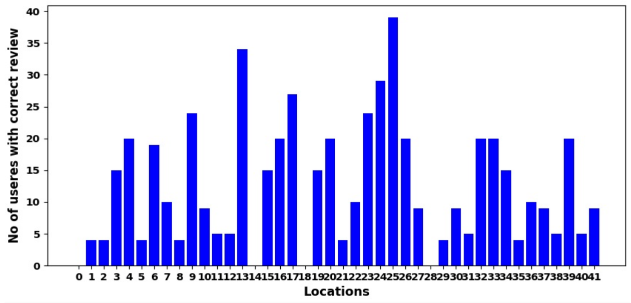


Figure 7. No of users with correct review vs. Locations

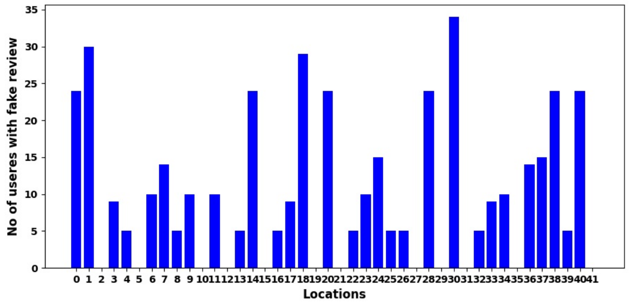


Figure 8. No of users with fake review vs. Locations

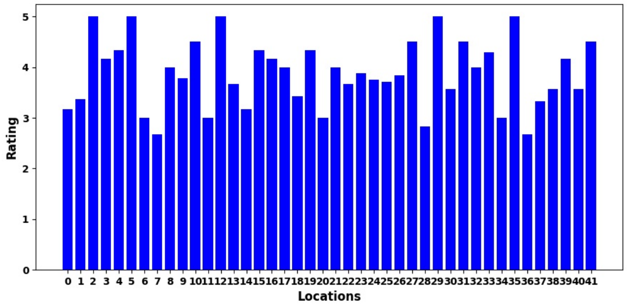


Figure 9. Estimated Rating of Locations

Location-wise no. of users who deliver genuine and fake reviews has been reflected in Figure 7 and Figure 8. The no. of users who have given genuine reviews for different locations has been reflected in Figure 7, and the number of users with fake reviews has been reflected in Figure 8. It has been observed that for some locations like location IDs 0,1,18,30 etc., the no. of fake reviews is greater than genuine reviews, and for location IDs such as 2,3,13,25, etc., the no. of correct reviews is greater than fake reviews also, for most locations mixed response has been observed that means some reviews are genuine as well as fake. As obtained real-time dataset contains both valid and invalid reviews, it is not able to provide reliable information to the forthcoming users, and it violates the principles of data integrity.

In Figure 9, the ERT of location has been represented by Rating axes in the range of 1 to 5, whereas the location axis represents unique ID of locations numerically in between 0 to 41. It is computed using Eq.-1. It may be treated as approximately the true/correct rating of the location that is determined by analyzing the real-time dataset of reviews obtained from various users.

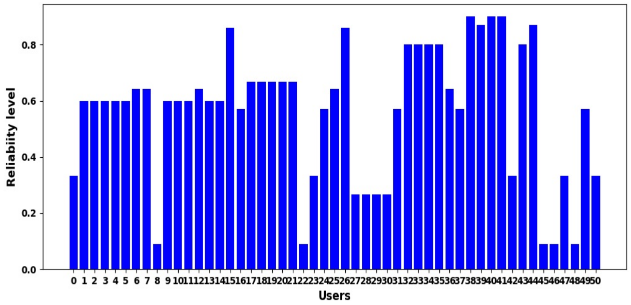


Figure 10. Reliability Level of Various Users

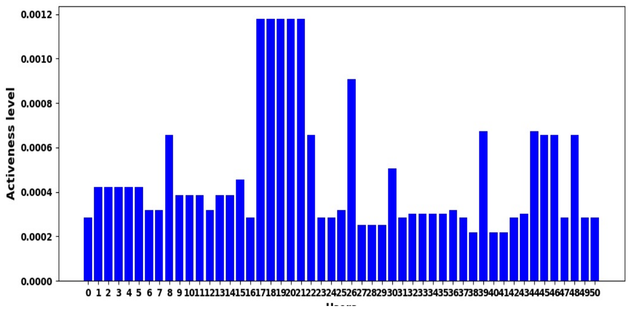


Figure 11. Activeness level of Various Users

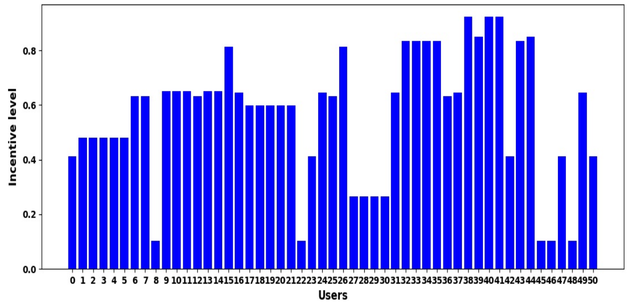


Figure 12. Incentive level of Various Users

The reliability level of users has been determined as per their given review, which has been reflected in Figure 10. A unique User ID has been assigned to the various user using numerical values from 0 to 50 in the proposed analysis. It is in the range of 0 to 1. It is computed using Eq.-12. Fuzzy logic will be used to find review status, which is the probability of being review as genuine. Further, it has been compared with the threshold for review status(δ) to discriminate review as genuine or fake. Based on these delivered reviews by users, reliability has been determined using Eq.-12. It describes how much extent a user and his review is reliable. The reliability level of the user is positively affected by the no. of genuine/honest reviews and negatively influenced by the no. of fake/invalid reviews. Therefore, due to genuine/honest reviews, the user's reliability level will enhance, whereas it will reduce when the delivered review is invalid/fake. It also helps to identify and eliminate fake/invalid reviews from the dataset for the isolation of genuine ones in order to maintain data integrity.

Figure 11 reflects the Activeness level of different users with unique and distinct user IDs between 0 to 50. It is in the range of 0 to 1. It is determined using Eq.-14. It represents the active participation of users during the review process. It also reflects the consistency of users in the review process. It represents the population of data in the dataset used for ensuring data integrity.

Figure 12 reflects the user's incentive level. It is in between 0 to 1. Unique user ID has been assigned to various users using numerical range ranges from 0 to 50. It is determined using Eq.-21. Review status has been obtained using fuzzy logic to detect it as genuine or fake by comparing it with the threshold(δ). It is used to determine the suspicious level of users as per Eq.- 18. Further activeness level and suspicious level of the user will be used with dynamic weight factors to compute incentive level as per Eq.19 to 21. In the review process, it reflects the amount of incentive given to various users for less suspicious behavior and consistent participation for ensuring data integrity. More amount of incentives will be provided to the user who constantly delivers genuine reviews that will help to ensure data integrity and provide more reliable information. Users' no. of genuine reviews positively influence Incentives; therefore, incentive increases, whereas no. of fake/invalid reviews negatively influence it; as a result, incentive decrease. It boosts users to give genuine reviews to earn more incentives and enforce data integrity with valid and reliable reviews. It also plays a vital role in ensuring data integrity through review analysis.

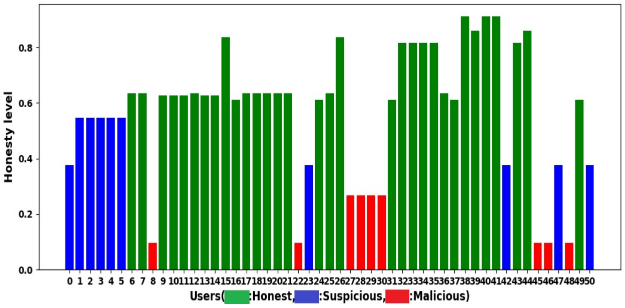


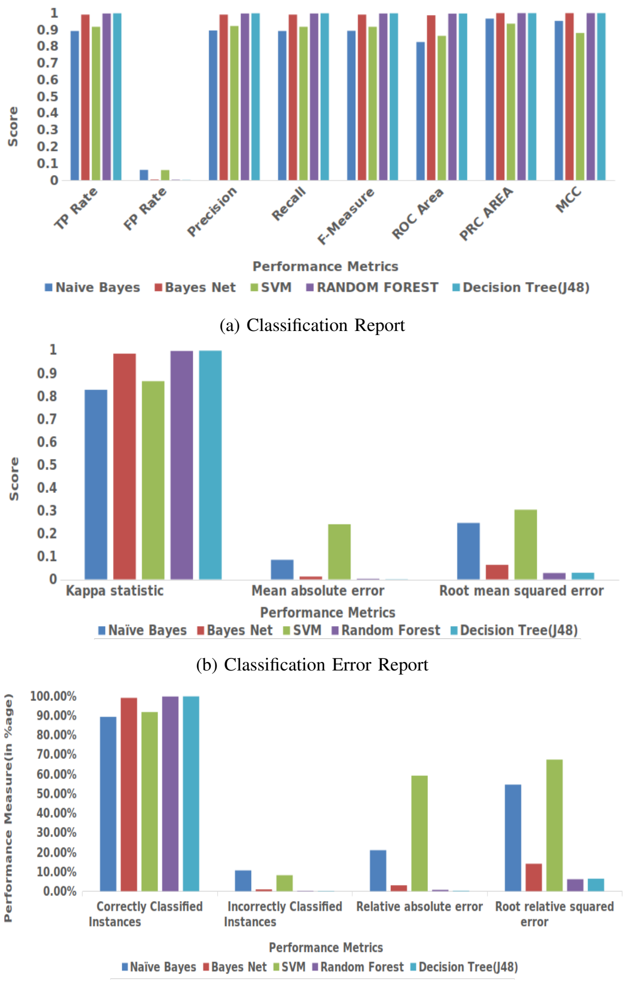
Figure 13. Honesty level of various Users

Figure 13 reflects the user's honesty level. Unique and distinct user IDs numerically in between 0 to 50 have been assigned to different users. The user's honesty level is in the range of 0 to 1. It is computed using Eq.-25. The honesty level of the user depends on the Incentive, Activeness, and Reliability level of the user with dynamic weight factors as per Eq.22 to 25. Users have been categorized as Malicious, Suspicious, and Honest users on the basis of their honesty level. In Figure 13, the red color reflects Malicious users, the green color reflects Honest users, and the blue color reflects Suspicious users. The trustworthiness of the user and his review in the review process has been reflected by the honesty level of the user, and it plays a crucial role in the identification of valid and invalid reviews and users for ensuring data integrity. Reviews received from honest users may be treated as reliable, and they may be isolated from other reviews so that information obtained from these reviews should be accurate and complete as well as the integrity of data should be ensured in the mobile crowdsensing environment. It also helps to discard all the reviews received from invalid sources like suspicious and malicious users instead of adding to the dataset, which will prevent future contamination of the dataset with fake reviews and also reduce the complexity of ensuring data integrity in the MCS environment.

B. *Classification Of User*

A labeled dataset has been generated by associating the categories of users determined using the proposed mathematical model with the reviews delivered by users. It has been classified using SVM, Naive Bayes, Bayes Net, Decision Tree(J48), and Random Forest ML algorithms in robust cross-validation mode to evaluate the efficiency of the proposed model.

Figure 14 (a), (b), and (c) demonstrate that proposed work is classified using various ML algorithms such as Naïve Bayes, Bayes Net, SVM, Random Forest and Decision Tree(J48) in Cross-validation mode. It is observed in Figure 14 (a) Decision Tree(J48), Random Forest, and Bayes Net are better than SVM and Naïve Bayes Algorithms in terms of performance metrics such as TP Rate, FP Rate, Precision, Recall, F-Measure, ROC Area, PRC Area, and MCC while FP Rate of Decision Tree(J48), Random Forest and Bayes Net are significantly less. Similarly, in Figure 14 (b), it is observed that kappa statistics of Random Forest and Decision Tree(J48) are greater than Bayes Net whereas MAE and RMSE are less than Bayes Net algorithm. The relative absolute error of the Random Forest is lowest than the Decision Tree(J48) algorithm. Decision Tree(J48) and Random Forest provide better accuracy than other ML algorithms as reflected in Figure 14 (c). The accuracy level of the Decision Tree(J48) algorithm is 99.86%, and for the Random Forest algorithm, it is 99.79%. Therefore, in cross-validation mode, the Decision tree(J48) provides a better accuracy level than other ML algorithms for the proposed model. As the accuracy level of the Decision tree(J48) is very much closer to Random Forest, a cost-benefit analysis has been performed using Random Forest and Decision Tree(J48) algorithm to enhance the accuracy of the proposed model in both algorithms by minimizing the errors to enforce data integrity. Finally, it helps to choose the best one.



(a) Classification Report

(b) Classification Error Report

(c) Classification Accuracy report

Figure 14. Classification of proposed work using different ML Algorithm in Cross validation mode

C. Cost Benefit Analysis

The accuracy level of the proposed model using the Decision Tree(J48) algorithm is 99.86%, and it is quite closer to the accuracy level of 99.79% provided by the Random Forest algorithm. It is possible due to the presence of various errors like mean absolute error, root means squared error, relative absolute error, etc. Cost-benefit has been done through virtual screening to minimize the error in order to enhance the accuracy level provided by both algorithms. It will help to select the best algorithm for the proposed work and also helps to enforce data integrity by minimizing errors through virtual screening. It has been discussed in detail with the help of data before virtual screening and after virtual screening in this section.

Classifier	Class	Accuracy	TP	FP	FN	TN	Cost	Cost (in %age)
Naïve Bayes	Honest	58.6572	0	0	585	830	585	41.34275618
Naïve Bayes	Malicious	86.2191	0	0	195	1220	195	13.78091873
Naïve Bayes	Suspicious	55.1237	0	0	635	780	635	44.87632509
Bayes Net	Honest	58.6572	0	0	585	830	585	41.34275618
Bayes Net	Malicious	86.2191	0	0	195	1220	195	13.78091873
Bayes Net	Suspicious	55.1237	0	0	635	780	635	44.87632509
SVM	Honest	58.6572	0	0	585	830	585	41.34275618
SVM	Malicious	86.2191	0	0	195	1220	195	13.78091873
SVM	Suspicious	55.1237	0	0	635	780	635	44.87632509
Random Forest	Honest	58.6572	0	0	585	830	585	41.34275618
Random Forest	Malicious	86.2191	0	0	195	1220	195	13.78091873
Random Forest	Suspicious	55.1237	0	0	635	780	635	44.87632509
Decision Tree(J48)	Honest	58.6572	0	0	585	830	585	41.34275618
Decision Tree(J48)	Malicious	86.2191	0	0	195	1220	195	13.78091873
Decision Tree(J48)	Suspicious	55.1237	0	0	635	780	635	44.87632509

TABLE II. Cost Benefit Analysis before Virtual Screening

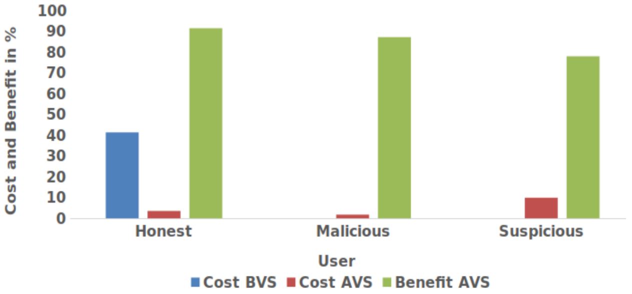


Figure 15. Cost Benefit Analysis using Naive Bayes Algorithm

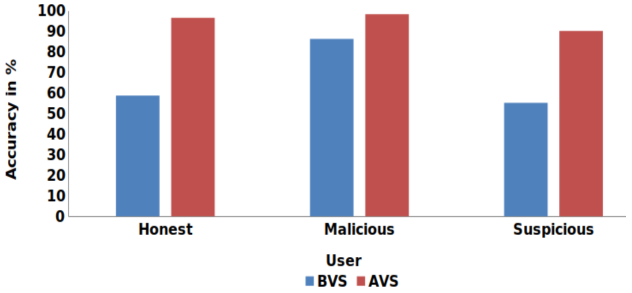


Figure 16. Accuracy level before and after Virtual Screening using Naïve Bayes Algorithm

Table II and Table III Reflects data obtained for cost-benefit analysis using Naïve Bayes Algorithm before and after the virtual screening. For Honest users, before the virtual screening, the no. of False Positive(FP) is 0, and False Negative(FN) is 585, so the cost is 585, but after the virtual screening, FP increased by 40, but FN decreased to 575 so gradually total cost has been decreased to 50. 43.46% of the entire population has been used for virtual screening, and it converts 98.2906% of FP and FN records to True Positive (TP) and True Negative (TN); therefore,

Classifier	Class	Accuracy	TP	FP	FN	TN	Cost	Cost (in %age)	Benefit	Benefit (in %age)	Population used (in %age)	Target achieved (in %age)
Naïve Bayes	Honest	96.47	575	40	10	790	50	3.53	535	91.45	43.46	98.29
Naïve Bayes	Malicious	98.23	170	0	25	1220	25	1.77	170	87.18	12.01	87.18
Naïve Bayes	Suspicious	90.11	515	20	120	760	140	9.89	495	77.95	37.81	81.1
Bayes Net	Honest	99.79	585	3	0	827	3	0.21	582	99.49	41.55	100
Bayes Net	Malicious	99.29	195	10	0	1210	10	0.71	185	94.87	14.49	100
Bayes Net	Suspicious	100.00	635	0	0	780	0	0	635	100	44.88	100
SVM	Honest	93.57	579	85	6	745	91	6.43	494	84.44	46.93	98.97
SVM	Malicious	98.23	170	0	25	1220	25	1.77	170	87.18	12.01	87.18
SVM	Suspicious	91.8	550	31	85	749	116	8.2	519	81.73	41.06	86.61
Random Forest	Honest	100	585	0	0	830	0	0	585	100	41.34	100
Random Forest	Malicious	100	195	0	0	1220	0	0	195	100	13.78	100
Random Forest	Suspicious	100	635	0	0	780	0	0	635	100	44.88	100
Decision Tree(J48)	Honest	100	585	0	0	830	0	0	585	100	41.34	100
Decision Tree(J48)	Malicious	99.86	193	0	2	1220	2	0.14	193	98.97	13.64	98.97
Decision Tree(J48)	Suspicious	99.86	635	2	0	778	2	0.14	633	99.69	45.02	100

TABLE III. Cost Benefit Analysis after Virtual Screening

it achieves 91.45% of benefit. It is also reflected in Figure 15, whereas the cost Before Virtual screening is reduced after the virtual screening, and it also reflects the benefit obtained For Honest Users. Similarly, for Malicious Users, before the virtual screening, the number of FP is 0 and FN is 195, so the cost is 195, but after the virtual screening, FP is 0, but FN is reduced to 25, so the gradually total cost is also reduced from 195 to 25. 12.014% of the entire population has been used for virtual screening, and it converts 87.18% of FP and FN records to TP and TN. Therefore, it achieves 87.18% of the benefit. It is also reflected in Figure 15. Where the cost Before Virtual screening is reduced after the virtual screening, and it also reflects the benefit obtained. In the case of Suspicious User, before the virtual screening, the number of FP is 0 and FN is 635, so the cost is 635, but after the virtual screening, FP is 20, but FN is reduced to 120, so gradually, the total cost is also reduced from 635 to 140. 37.81% of the entire population has been used for virtual screening, and it converts 81.10% of FP and FN records to TP and TN; therefore, it achieves 77.95% of benefit. It is also reflected in Figure 15. Whereas the cost Before Virtual screening is reduced after the virtual screening, reflecting the benefit obtained.

Figure 16, Reflects the accuracy level of different types of users before and after virtual scanning. For Honest users, the cost is reduced by 535, so accuracy has been increased from 41.34% to 96.47%. Similarly, for malicious users, due to cost reduction from 195 to 25, accuracy has been increased from 86.21% to 98.23%, and also for Suspicious users, after the virtual screening, accuracy level increased from 55.12% to 90.10% due to cost reduction from 635 to 140.

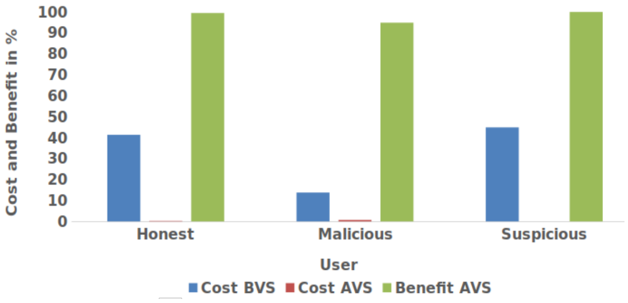


Figure 17. Cost Benefit Analysis using Bayes Net Algorithm

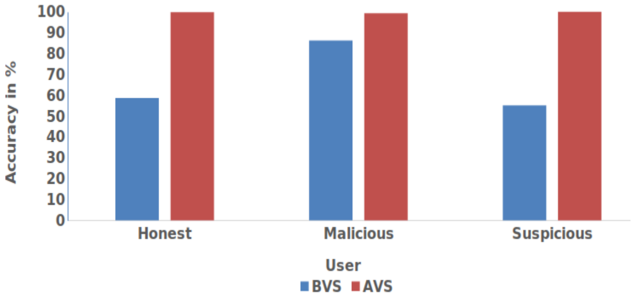


Figure 18. Accuracy level before and after Virtual Screening using Bayes Net Algorithm

Table II and Table III Reflects data obtained for cost-benefit analysis using Bayes Net Algorithm before and after the virtual screening. For Honest users, before the virtual screening, the no. of False Positive(FP) is 0, and False Negative(FN) is 585, so the cost is 585, but after the virtual screening, FP is 0 and FN is reduced to 3, so gradually total cost is decreased by 582. 41.55% of the entire population has been used for virtual screening, and it converts 100% of FP and FN records to TP and TN; therefore, it achieves 99.49% of benefit. It is also reflected in Figure 17, whereas the cost before virtual screening is reduced after the virtual screening, and it also reflects the benefit obtained For Honest User. Similarly, for Malicious Users, before the virtual screening number of FP is 0 and FN is 195, so the cost is 195, but after the virtual screening, FP is 10, but FN is reduced to 0, and TP is 195, so gradually total cost is also reduced from 195 to 10. 14.49% of the entire population has been used for virtual screening, and it converts 100% of FP and FN records to TP and TN; therefore, it achieves 94.87% of benefit. It is also reflected in Figure 17 where the cost before virtual screening is reduced after the virtual screening, and it also reflects the benefit obtained. In the case of Suspicious User, before the virtual screening, the number of FP is 0 and FN is 635, so the cost is 635, but after the virtual screening, FP is 0 and FN is reduced to 0, so gradually total cost is also reduced from 635 to 0. 44.87% of the entire population has been used for virtual screening, and it converts 100% of FP and FN records to TP and TN; therefore, it achieves 100% of benefit. It is also reflected in Figure 17 whereas the cost before virtual screening is reduced after the virtual screening, and it also reflects the benefit obtained.

Figure 18, Reflects the accuracy level of different types of users before and after virtual scanning using Bayes Net Algorithm. For Honest users, Cost is reduced by 582, so accuracy has been increased from 58.65 to 99.79%. Similarly, for malicious users, due to vast cost reduction from 195 to 10, accuracy has been increased from 86.21% to 99.29%, and also for suspicious users after the virtual screening, accuracy level increased from 55.12% to 100 % due to cost reduction from 635 to 0.

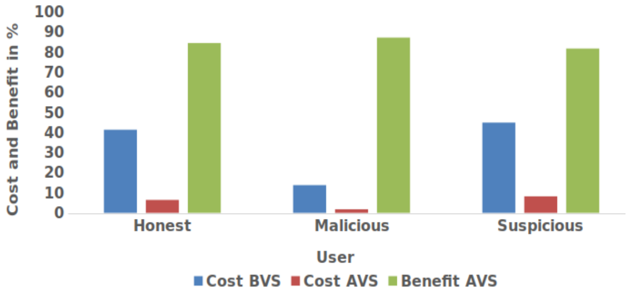


Figure 19. Cost Benefit Analysis using SVM Algorithm

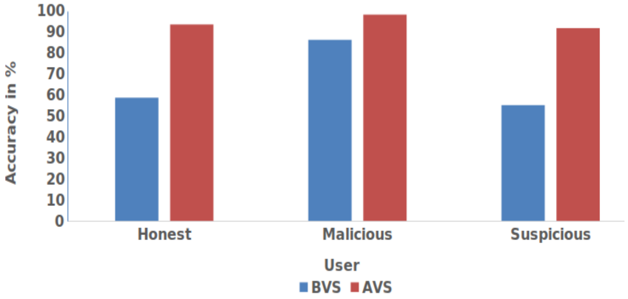


Figure 20. Accuracy level before and after Virtual Screening using SVM Algorithm

Table II and Table III Reflects data obtained for cost-benefit analysis using the SVM algorithm before and after the virtual screening. For Honest users, before the virtual screening, the no. of False Positive(FP) is 0, and False Negative(FN) is 585, so the cost is 585, but after the virtual screening, FP is 85 and FN is reduced to 6, so gradually total cost is decreased by 494. 46.92% of the entire population has been used for virtual screening, and it converts 98.97% of FP and FN records to TP and TN; therefore, it achieves 84.45% of benefit. It is also reflected in Figure 19, whereas the cost before virtual screening is reduced after the virtual screening, and it also reflects the benefit obtained For Honest User. Similarly, for Malicious Users, before the virtual screening number of FP is 0 and FN is 195, so the cost is 195, but after the virtual screening, FP is 0, and FN is reduced to 25, so gradually, the total cost is also reduced from 195 to 25. 12.01% of the entire population has been used for virtual screening, and it converts 87.18% of FP and FN records to TP and TN; therefore, it achieves 87.18% of benefit. It is also reflected in Figure 19 where the cost before virtual screening is reduced after the virtual screening, and it also reflects the benefit obtained. In the case of Suspicious User, before the virtual screening, the number of FP is 0, and FN is 635, so the cost is 635, but after the virtual screening, FP is 31 and FN is reduced to 85, so gradually, the total cost is also reduced from 635 to 116. 41.06% of the entire population has been used for virtual screening, and it converts 86.61% of FP and FN records to TP and TN; therefore, it achieves 81.73% of benefit. It is also reflected in Figure 19, where the cost before virtual screening is reduced after the virtual screening, and it also reflects the benefit obtained.

Figure 20, Reflects the accuracy level of different types of users before and after virtual scanning using the SVM Algorithm. For Honest users, Cost is reduced by 494, so accuracy has been increased from 58.65% to 93.56%. Similarly, for malicious users, due to vast cost reduction from 195 to 25, accuracy has been increased from 86.21% to 98.23%, and also for suspicious users after the virtual screening, accuracy level increased from 55.12% to 91.80% due to cost reduction from 635 to 116.

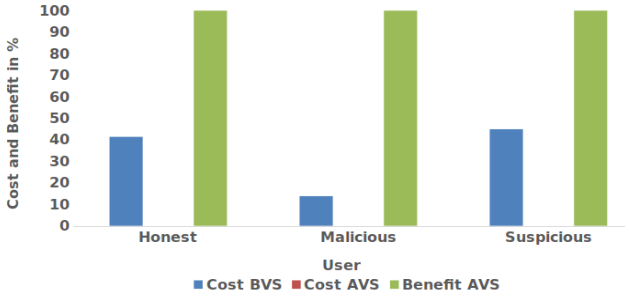


Figure 21. Cost Benefit Analysis using Random Forest Algorithm

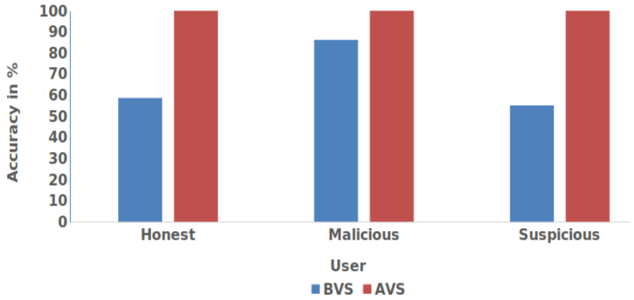


Figure 22. Accuracy level before and after Virtual Screening using Random Forest Algorithm

Table II and Table III reflects data obtained for cost-benefit analysis using Random Forest Algorithm before and after the virtual screening. For Honest user, before the virtual screening, the no. of False Positive(FP) is 0, and False Negative(FN) is 585, so the cost is 585, but after the virtual screening, FP is 0 and FN is reduced to 0, so gradually total cost is decreased by 585. 41.34% of the entire population has been used for virtual screening, and it converts 100% of FP and FN records to TP and TN; therefore, it achieves 100% of benefit. It is also reflected in Figure 21 whereas the cost Before Virtual screening is reduced after the virtual screening, and it also reflects the benefit obtained For Honest users. Similarly, for Malicious users, before the virtual screening number of FP is 0 and FN is 195, so the cost is 195, but after the virtual screening, FP is 0, and FN is reduced to 0, so gradually, the total cost is also reduced from 195 to 0. 13.78% of the entire population has been used for virtual screening, and it converts 100% of FP and FN records to TP and TN; therefore, it achieves 100% of benefit. It is also reflected in Figure 21 where the cost before virtual screening is reduced after the virtual screening, and it also reflects the benefit obtained. In the case of Suspicious user, before the virtual screening number of FP is 0 and FN is 635, so the cost is 635, but after the virtual screening, FP is 0 and FN is reduced to 0, so gradually, the total cost is also reduced from 635 to 0. 44.87% of the entire population has been used for virtual screening, and it converts 100% of FP and FN records to TP and TN; therefore, it achieves 100% of benefit. It is also reflected in Figure 21 whereas the cost Before Virtual screening is reduced after the virtual screening, and it also reflects the benefit obtained.

Figure 22, Reflects accuracy level of different types of users before and after virtual scanning using Random Forest Algorithm. For Honest users, Cost is reduced by 585, so accuracy has been increased from 58.65% to 100%. Similarly, for malicious users, due to a vast cost reduction from 195 to 0, accuracy has been increased from 86.21% to 100%, and also for suspicious users after the virtual screening, accuracy level increased from 55.12% to 100% due to cost reduction from 635 to 0. Random Forest algorithm provides 100% accuracy for Honest, Suspicious, and Malicious users.

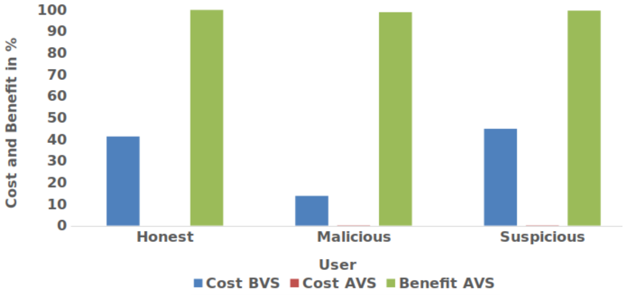


Figure 23. Cost Benefit Analysis using Decision Tree(J48) Algorithm

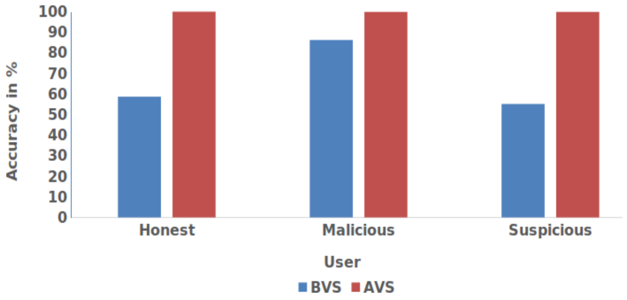


Figure 24. Accuracy level before and after Virtual Screening using Decision Tree (J48) Algorithm

Table II and Table III Reflects data obtained for cost-benefit analysis using the Decision Tree(J48) Algorithm before and after the virtual screening. For Honest users, before virtual screening no. of False Positive(FP) is 0 and False Negative(FN) is 585, so the cost is 585, but after the virtual screening, FP is 0, and FN is reduced to 0, so gradually total cost is decreased by 585. 41.34% of the entire population has been used for virtual screening, and it converts 100% of FP and FN records to TP and TN; therefore, it achieves 100% of benefit. It is also reflected in Figure 23, whereas the cost Before Virtual screening is reduced after the virtual screening, and it also reflects the benefit obtained for Honest User. Similarly, for Malicious Users, before the virtual screening number of FP is 0 and FN is 195, so the cost is 195, but after the virtual screening, FP is 0 and FN has been reduced to 2, so gradually, the total cost is also reduced from 195 to 2. 13.63% of the entire population has been used for virtual screening, and it converts 98.97% of FP and FN records to TP and TN; therefore, it achieves 98.97% of benefit. It is also reflected in Figure 23 where the cost before virtual screening is reduced after the virtual screening, and it also reflects the benefit obtained. In the case of suspicious users, before the virtual screening number of FP is 0 and FN is 635, so the cost is 635, but after the virtual screening FP is 2 and FN is reduced to 0, so gradually, the total cost is also reduced from 635 to 2. 45.01% of the entire population has been used for virtual screening, and it converts 100% of FP and FN records to TP and TN; therefore, it achieves 99.68% of benefit. It is also reflected in Figure 23 whereas the cost before virtual screening is reduced after the virtual screening, and it also reflects the benefit obtained.

Figure 24, reflects the accuracy level of different types of users before and after virtual scanning using Decision Tree (J48) Algorithm. For Honest users, Cost is reduced by 585, so accuracy has been increased from 58.65% to 100%. Similarly, for malicious users, due to vast cost reduction from 195 to 2, accuracy has been increased from 86.21% to 99.87%, and also for suspicious users after the virtual screening, accuracy level increased from 55.12% to 99.87% due to cost reduction from 635 to 2.

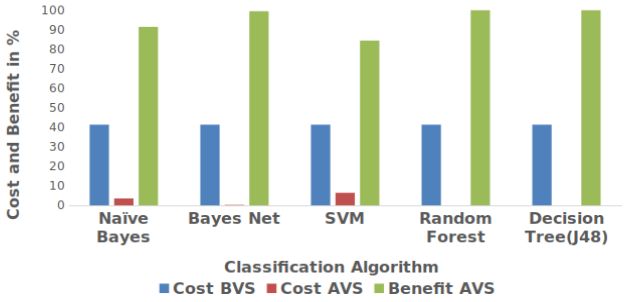


Figure 25. Cost Benefit Analysis before and after Virtual Screening using Various Algorithm of Honest User

Figure 25, reflects the cost and benefit obtained before and after Virtual Screening using various algorithms like Naïve Bayes, Bayes Net, SVM, Random Forest and Decision Tree(J48) of Honest User. In the case of the Naïve Bayes Algorithm, the cost is reduced from 585 to 50, so it provides 91.45% of benefit; also Bayes Net Algorithm reduces cost from 585 to 3, so it provides 99.49% of benefit. Similarly, the cost is reduced from 585 to 91 and provides 84.45% of benefit using the SVM algorithm. But Decision Tree(J48) and Random Forest algorithms reduce cost from 835 to 0, providing 100% benefit. Therefore for Honest users, Decision Tree(J48) and Random Forest algorithms provide better benefits than other ML algorithms.

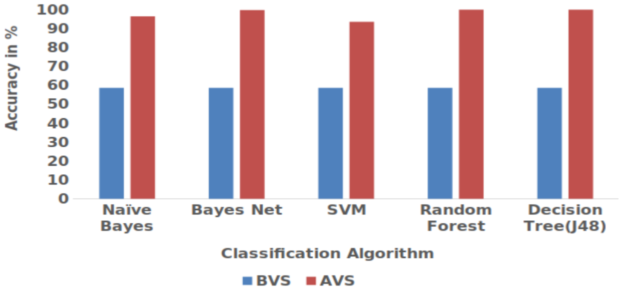


Figure 26. Accuracy level before and after Virtual Screening using Various Algorithm of Honest User

Figure 26, reflects the accuracy level obtained before and after virtual screening using various algorithms like Naïve Bayes, Bayes Net, SVM, Random Forest, and Decision Tree (J48) of Honest User. In the case of the Naïve Bayes Algorithm accuracy level increased from 58.65 % to 96.47%; also, Bayes Net Algorithm increased accuracy from 58.65% to 99.79%. Similarly Accuracy level increased from 58.65% to 93.57% using the SVM algorithm. But the Decision tree(J48) and Random Forest Algorithm increase accuracy from 58.65% to 100%. Therefore for Honest users, Decision Tree(J48) and Random Forest algorithms provide a better Accuracy level than other ML algorithms.

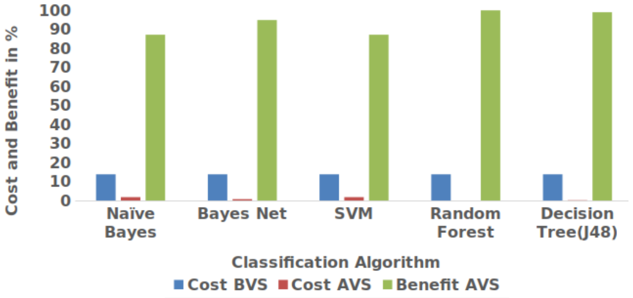


Figure 27. Cost Benefit Analysis before and after Virtual Screening using Various Algorithm of Malicious User

Figure 27, reflects the cost and benefit obtained before and after virtual screening using various algorithms like Naïve Bayes, Bayes Net, SVM, Random Forest, and Decision Tree(J48) of Malicious User. In the case of Naïve Bayes and SVM algorithms, the cost is reduced from 195 to 25, so they provide 87.18% of benefit; also, Bayes Net Algorithm reduces cost from 195 to 10, so it provides 94.87% of benefit. Similarly, the cost is reduced from 195 to 0 and provides 100% of the benefit using the Random Forest algorithm. The Decision Tree(J48) algorithm reduces cost from 195 to 2; hence it provides 98.97% of benefit. Therefore, the Random Forest algorithm provides better benefits for malicious users than other ML algorithms.

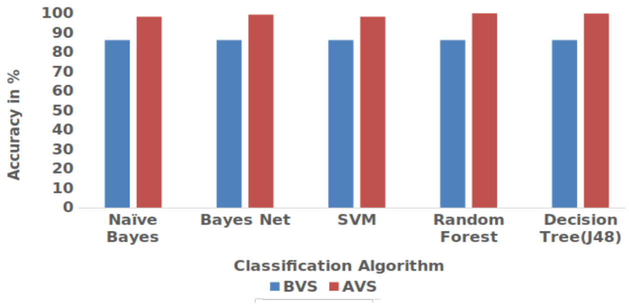


Figure 28. Accuracy level before and after Virtual Screening using Various Algorithm of Malicious User

Figure 28, reflects the accuracy level obtained before and after virtual screening using various algorithms like Naïve Bayes(NB), Bayes Net(BN), SVM, Random Forest, and Decision Tree(J48) of malicious users. In the case of Naïve Bayes(NB) and SVM Algorithms, the accuracy level increased from 86.21% to 98.23%; also, Bayes Net Algorithm increased accuracy from 86.21 % to 99.29%. Similarly Accuracy level increased from 86.21% to 100% using the Random Forest algorithm, whereas the Decision Tree(J48) algorithm improved accuracy from 86.21% to 98.97%. Therefore, for malicious users, the Random Forest algorithm provides better accuracy than other ML algorithms.

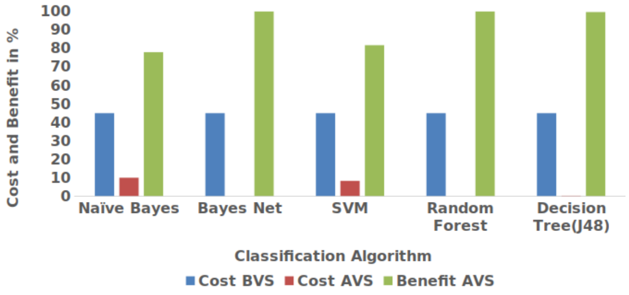


Figure 29. Cost Benefit Analysis before and after Virtual Screening using Various Algorithm of Suspicious User

Figure 29, reflects the cost and benefit obtained before and after virtual screening using various algorithms like Naïve Bayes, Bayes Net, SVM, Random Forest, and Decision Tree(J48) of suspicious users. In the case of the Naïve Bayes Algorithm, the cost is reduced from 635 to 140, providing 77.95% of benefit; Bayes Net and Random Forest Algorithms reduce cost from 635 to 0, providing 100% of benefit. Similarly, the cost is reduced from 635 to 116 and provides 81.73% of benefit using the SVM algorithm. But Decision Tree(J48) reduces cost from 635 to 2, providing 99.68% benefit. Therefore, Bayes Net and Random Forest algorithms provide better benefits for Suspicious users than other ML algorithms.

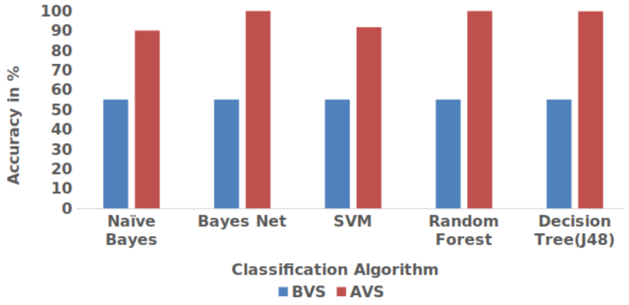


Figure 30. Accuracy level before and after Virtual Screening using Various Algorithm of Suspicious User

Figure 30, reflects the accuracy level obtained before and after virtual screening using various algorithms like Naïve Bayes, Bayes Net, SVM, Random Forest, and Decision Tree (J48) of suspicious users. In the case of the Naïve Bayes algorithm, the accuracy level has been increased from 55.12% to 90.10%; Bayes Net and Random Forest algorithms increase accuracy from 55.12% to 100%. Similarly Accuracy level increased from 55.12% to 91.80%, using the SVM algorithm. But Decision tree(J48) algorithms increase accuracy from 55.12% to 99.85%. Therefore, Bayes Net and Random Forest algorithms provide a better Accuracy level than other ML algorithms for suspicious users.

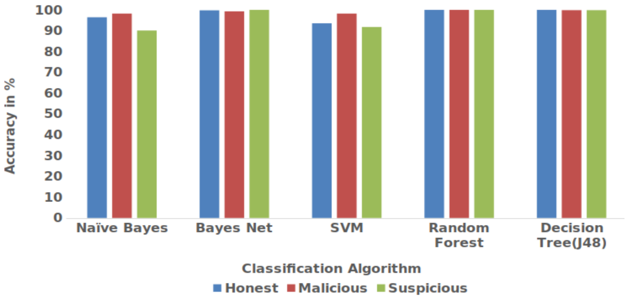


Figure 31. Accuracy level after Virtual Screening using Various Algorithm

Figure 31, reflects the accuracy level obtained after virtual screening using various algorithms like Naïve Bayes, Bayes Net, SVM, Random Forest, and Decision Tree(J48) of Honest, Malicious, and Suspicious users. In the case of the Naïve Bayes algorithm, accuracy levels of Honest, Malicious, and Suspicious users are 96.46%, 98.23%, and 90.10%, respectively. Also, for Bayes Net algorithm, the accuracy levels of Honest, Malicious, and Suspicious users are 99.78%, 99.29%, and 100%, respectively. Similarly, the accuracy level of Honest, Malicious, and Suspicious users are 93.56%, 98.23%, and 91.80%, respectively, obtained using the SVM algorithm. Decision Tree(J48) Algorithms provide accuracy levels of Honest, Malicious, and Suspicious users are 100%, 99.85%, and 99.85%, respectively. But a 100% accuracy level has been observed for Honest, Suspicious, and Malicious Users using the Random Forest algorithm. Therefore, for Honest, Malicious, and Suspicious users, the Random Forest algorithm provides better accuracy than other ML algorithms.

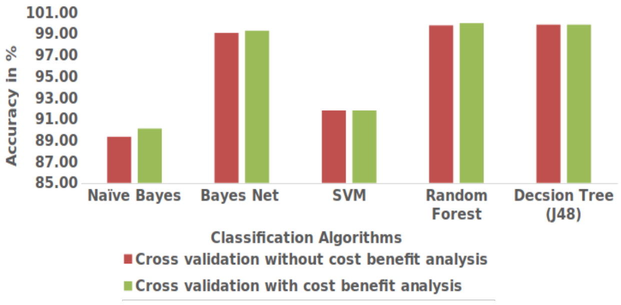


Figure 32. Accuracy level without and with Cost benefit analysis

Figure 32 Reflects the effect of cost-benefit analysis on accuracy level. Initially proposed model has been classified in cross-validation mode without cost-benefit analysis using different ML algorithms like Naïve Bayes, Bayes Net, SVM, Random Forest, and Decision Tree(J48).In the cross-validation mode, the accuracy level of Random Forest and Decision Tree(J48) are 99.79 and 99.86, respectively. Without cost-benefit analysis, the Decision Tree(J48) provides better accuracy, but it is closer to Random Forest Algorithm. It has been classified in cross-validation mode with the cost-benefit analysis to enhance the accuracy of the ML algorithms to pick the best one for the proposed work. After the cost-benefit analysis, the accuracy level of different algorithms has slightly increased in the case of Naïve Bayes, Bayes Net, SVM, and Random forest algorithms, but the accuracy level of Decision Tree(J48) has not been affected. After the cost-benefit analysis, it is concluded that Random Forest Algorithm provides better accuracy than the Decision Tree(J48) algorithm.

5. Conclusions And Future Scope

In this work, a novel concept has been described to maintain data integrity in the MCS environment by categorizing the reviews/feedback as genuine/honest and fake/invalid through a Fuzzy approach and various membership functions (MF) over the mathematical model. Further, the categorization of the users as suspicious or malicious and honest has been done for the identification of valid and invalid sources of data that will reduce complexity for ensuring data integrity. In this context, some users are classified as suspicious or malicious, which were earlier classified as honest users according to the models presented in [6], [7]. The estimated rating (ERT) of different locations is determined by analyzing the dataset that contains feedback submitted by different users. It may be treated as the true rating of the location. It is compared with the review submitted by the user using the fuzzy concept to detect fake and genuine/honest reviews for ensuring data integrity. The user's reliability level has been computed by using the number of fake and genuine reviews delivered by the user. It reflects the extent to which the user and his review are reliable. It helps to detect and eliminates all the fake reviews from the dataset to have genuine and valid reviews only in the dataset that will ensure data integrity and provide reliable information throughout its life cycle. The user's activeness is also computed to find out the user's consistency, along with the reliability level in the review process. In order to enhance consistent participation and reliable reviews for ensuring data integrity, on the basis of user's reliability and activeness levels, incentives can be provided for encouragement, as depicted in Figures 9 and 10. The user's honesty level has been computed using the user's reliability level, activeness levels, allocated incentives to users, and the attached dynamic weight factors. It represents the extent to which the user and his review are trustworthy, reliable, and genuine. Thus, this model has been used to first identify and remove all the fake reviews from the dataset to have only genuine, accurate and complete reviews in the dataset that will provide accurate information to forthcoming users whenever retrieved throughout its life-cycle. This will ensure data integrity in the MCS environment. Further, to stop future contamination of the dataset with fake reviews, users have been categorized as per their honesty level as honest, suspicious, or malicious in order to identify valid and invalid sources of review for minimization of the complexity of maintaining data integrity. Reviews/feedback gathered from valid sources or honest users only will be considered as genuine reviews, and they will be added to the dataset, and the rest of the review will be discarded to minimize complexity for ensuring data integrity. This real-time dataset has been classified using various ML algorithms to determine the accuracy of the proposed model. It is observed that in the case of cross-validation mode, the Decision tree (J48) and Random Forest algorithms provide 99.86% and 99.79% accuracy, respectively and it is nearly similar to each other. Therefore, the cost-benefit analysis has been done to eliminate the errors that reduce accuracy using the Decision Tree (J48) and Random Forest algorithm in cross-validation mode. It is observed that the accuracy level of the Random forest algorithm increases to 100%, whereas the accuracy level using the Decision Tree(J48) remains constant.

6. Declarations

A. Competing Interests

The authors have no competing interests to declare that are relevant to the content of this article.

B. Funding

No funding was received to assist with the preparation of this manuscript.

C. Authors' Contributions

All authors contributed to the study's conception and design. Material preparation, data collection, and analysis were performed by Ramesh Kumar Sahoo under the guidance of Dr. Sateesh Kumar Pradhan, Dr. Srinivas Sethi, and Dr. Siba K. Udgata. The problem is formulated by Dr. Siba K Udgata and he also contributed to draft the response to the reviewer comments. All authors have contributed to the writing, and review of the paper.

D. Acknowledgements

I, Ramesh K. Sahoo, am thankful to my supervisor Dr. Sateesh Kumar Pradhan, and also to Dr. Srinivas Sethi and Dr. Siba K. Udgata for their proper guidance.

References

[1] R. Barbado, O. Araque, and C. A. Iglesias, "A framework for fake review detection in online consumer electronics retailers," *Information Processing & Management*, vol. 56, no. 4, pp. 1234–1244, 2019.

[2] A. Habib, M. Z. Asghar, A. Khan, A. Habib, and A. Khan, "False information detection in online content and its role in decision making: a systematic literature review," *Social Network Analysis and Mining*, vol. 9, no. 1, pp. 1–20, 2019.

[3] A. Bondielli and F. Marcelloni, "A survey on fake news and rumour detection techniques," *Information Sciences*, vol. 497, pp. 38–55, 2019.

[4] J. K. Rout, A. K. Dash, and N. K. Ray, "A framework for fake review detection: issues and challenges," in *2018 International Conference on Information Technology (ICIT)*. IEEE, 2018, pp. 7–10.

[5] N. A. Patel and R. Patel, "A survey on fake review detection using machine learning techniques," in *2018 4th International Conference on Computing Communication and Automation (ICCCA)*. IEEE, 2018, pp. 1–6.

[6] R. K. Sahoo, S. Sethi, and S. K. Udgata, "A smartphone app based model for classification of users and reviews (a case study for tourism application)," in *Intelligent Systems*. Springer, 2021, pp. 337–348.

[7] R. K. Sahoo, S. K. Pradhan, and S. Sethi, "Ensuring data integrity in mobile crowdsensing environment using fuzzy logic," in *Intelligent Systems*. Springer, 2022, pp. 223–237.

[8] P. Huang, K. Fan, H. Yang, K. Zhang, H. Li, and Y. Yang, "A collaborative auditing blockchain for trustworthy data integrity in cloud storage system," *IEEE Access*, vol. 8, pp. 94 780–94 794, 2020.

[9] H. Wang, D. He, and S. Tang, "Identity-based proxy-oriented data uploading and remote data integrity checking in public cloud," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 6, pp. 1165–1176, 2016.

[10] H. Zhu, Y. Yuan, Y. Chen, Y. Zha, W. Xi, B. Jia, and Y. Xin, "A secure and efficient data integrity verification scheme for cloud-iot based on short signature," *IEEE Access*, vol. 7, pp. 90 036–90 044, 2019.

[11] W. I. Khedr, H. M. Khater, and E. R. Mohamed, "Cryptographic accumulator-based scheme for critical data integrity verification in cloud storage," *IEEE Access*, vol. 7, pp. 65 635–65 651, 2019.

[12] S. Tan, L. Tan, X. Li, and Y. Jia, "An efficient method for checking the integrity of data in the cloud," *China Communications*, vol. 11, no. 9, pp. 68–81, 2014.

[13] H. Wang and J. Zhang, "Blockchain based data integrity verification for large-scale iot data," *IEEE Access*, vol. 7, pp. 164 996–165 006, 2019.

[14] Y.-J. Chen, L.-C. Wang, and S. Wang, "Stochastic blockchain for iot data integrity," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 1, pp. 373–384, 2018.

[15] Q. Zhang, S. Wang, D. Zhang, J. Wang, and Y. Zhang, "Time and attribute based dual access control and data integrity verifiable scheme in cloud computing applications," *IEEE Access*, vol. 7, pp. 137 594–137 607, 2019.

[16] X.-P. Zhao and R. Jiang, "Distributed machine learning oriented data integrity verification scheme in cloud computing environment," *IEEE Access*, vol. 8, pp. 26 372–26 384, 2020.

[17] J. Li, H. Yan, and Y. Zhang, "Identity-based privacy preserving remote data integrity checking for cloud storage," *IEEE Systems Journal*, vol. 15, no. 1, pp. 577–585, 2020.

[18] A. Capponi, C. Fiandrino, D. Kliazovich, P. Bouvry, and S. Giordano, "A cost-effective distributed framework for data collection in cloud-based mobile crowd sensing architectures," *IEEE Transactions on Sustainable Computing*, vol. 2, no. 1, pp. 3–16, 2017.

[19] F. Khan, A. U. Rehman, J. Zheng, M. A. Jan, and M. Alam, "Mobile crowdsensing: A survey on privacy-preservation, task management, assignment models, and incentives mechanisms," *Future Generation Computer Systems*, vol. 100, pp. 456–472, 2019.

[20] Z. Xu, L. Mei, K.-K. R. Choo, Z. Lv, C. Hu, X. Luo, and Y. Liu, "Mobile crowd sensing of human-like intelligence using social sensors: A survey," *Neurocomputing*, vol. 279, pp. 3–10, 2018.

[21] N. Eligüzel, C. Çetinkaya, and T. Dereli, "Comparison of different machine learning techniques on location extraction by utilizing geo-tagged tweets: A case study," *Advanced Engineering Informatics*, vol. 46, p. 101151, 2020.

[22] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Machine learning*, vol. 29, no. 2, pp. 131–163, 1997.

[23] W. Liu, J. He, S. Han, F. Cai, Z. Yang, and N. Zhu, "A method for the detection of fake reviews based on temporal features of reviews and comments," *IEEE Engineering Management Review*, vol. 47, no. 4, pp. 67–79, 2019.

[24] Y. Li, X. Feng, and S. Zhang, "Detecting fake reviews utilizing semantic and emotion model," in *2016 3rd international conference on information science and control engineering (ICISCE)*. IEEE, 2016, pp. 317–320.

[25] L. Li, K. Y. Lee, M. Lee, and S.-B. Yang, "Unveiling the cloak of deviance: Linguistic cues for psychological processes in fake online reviews," *International Journal of Hospitality Management*, vol. 87, p. 102468, 2020.

[26] G. Shan, L. Zhou, and D. Zhang, "From conflicts and confusion to doubts: Examining review inconsistency for fake review detection," *Decision Support Systems*, vol. 144, p. 113513, 2021.

[27] Q. Chen, G. Srivastava, R. M. Parizi, M. Aloqaily, and I. Al Ridhawi,

"An incentive-aware blockchain-based solution for internet of fake media things," *Information Processing & Management*, vol. 57, no. 6, p. 102370, 2020.

[28] P. Tehlan, R. Madaan, and K. K. Bhatia, "A spam detection mechamism in social media using soft computing," in *2019 6th International Conference on Computing for Sustainable Global Development (INDIACom)*. IEEE, 2019, pp. 950–955.

[29] K. Dhingra and S. K. Yadav, "Spam analysis of big reviews dataset using fuzzy ranking evaluation algorithm and hadoop," *International Journal of Machine Learning and Cybernetics*, vol. 10, no. 8, pp. 2143–2162, 2019.

[30] R. Ariaeinejad and A. Sadeghian, "Spam detection system: A new approach based on interval type-2 fuzzy sets," in *2011 24th Canadian Conference on Electrical and Computer Engineering (CCECE)*. IEEE, 2011, pp. 000 379–000 384.

[31] S. Jia, X. Zhang, X. Wang, and Y. Liu, "Fake reviews detection based on lda," in *2018 4th International Conference on Information Management (ICIM)*. IEEE, 2018, pp. 280–283.

[32] J. Yao, Y. Zheng, and H. Jiang, "An ensemble model for fake online review detection based on data resampling, feature pruning, and parameter optimization," *IEEE Access*, vol. 9, pp. 16 914–16 927, 2021.

[33] G. S. Budhi, R. Chiong, and Z. Wang, "Resampling imbalanced data to detect fake reviews using machine learning classifiers and textual-based features," *Multimedia Tools and Applications*, vol. 80, no. 9, pp. 13 079–13 097, 2021.

[34] J. Wang, H. Kan, F. Meng, Q. Mu, G. Shi, and X. Xiao, "Fake review detection based on multiple feature fusion and rolling collaborative training," *IEEE Access*, vol. 8, pp. 182 625–182 639, 2020.

[35] A. Choudhary and A. Arora, "Linguistic feature based learning model for fake news detection and classification," *Expert Systems with Applications*, vol. 169, p. 114171, 2021.

[36] Y. Li, F. Wang, S. Zhang, and X. Niu, "Detection of fake reviews using group model," *Mobile Networks and Applications*, vol. 26, no. 1, pp. 91–103, 2021.

[37] Y. Fang, H. Wang, L. Zhao, F. Yu, and C. Wang, "Dynamic knowledge graph based fake-review detection," *Applied Intelligence*, vol. 50, no. 12, pp. 4281–4295, 2020.

[38] R. K. Kaliyar, A. Goswami, P. Narang, and S. Sinha, "Fndnet–a deep convolutional neural network for fake news detection," *Cognitive Systems Research*, vol. 61, pp. 32–44, 2020.

[39] Y.-F. Huang and P.-H. Chen, "Fake news detection using an ensemble learning model based on self-adaptive harmony search algorithms," *Expert Systems with Applications*, vol. 159, p. 113584, 2020.

[40] B. Guo, Y. Ding, Y. Sun, S. Ma, K. Li, and Z. Yu, "The mass, fake news, and cognition security," *Frontiers of Computer Science*, vol. 15, no. 3, pp. 1–13, 2021.

[41] G. Alexandron, L. Y. Yoo, J. A. Ruipérez-Valiente, S. Lee, and D. E. Pritchard, "Are mooc learning analytics results trustworthy? with fake learners, they might not be!" *International journal of artificial intelligence in education*, vol. 29, no. 4, pp. 484–506, 2019.

[42] K. Goswami, Y. Park, and C. Song, "Impact of reviewer social interaction on online consumer review fraud detection," *Journal of big data*, vol. 4, no. 1, pp. 1–19, 2017.

[43] M. Liu, Y. Shang, Q. Yue, and J. Zhou, "Detecting fake reviews using multidimensional representations with fine-grained aspects plan," *IEEE Access*, vol. 9, pp. 3765–3773, 2020.

[44] P. Hajek, A. Barushka, and M. Munk, "Fake consumer review detection using deep neural networks integrating word embeddings and emotion mining," *Neural Computing and Applications*, vol. 32, no. 23, pp. 17 259–17 274, 2020.

[45] N. Ruan, R. Deng, and C. Su, "Gadm: Manual fake review detection for o2o commercial platforms," *Computers & Security*, vol. 88, p. 101657, 2020.

[46] R. K. Kaliyar, A. Goswami, and P. Narang, "Deepfake: improving fake news detection using tensor decomposition-based deep neural network," *The Journal of Supercomputing*, vol. 77, no. 2, pp. 1015–1037, 2021.

[47] ——, "Echofaked: improving fake news detection in social media with an efficient deep neural network," *Neural computing and applications*, pp. 1–17, 2021.

[48] ——, "Fakebert: Fake news detection in social media with a bert-based deep learning approach," *Multimedia Tools and Applications*, vol. 80, no. 8, pp. 11 765–11 788, 2021.

[49] M. H. Goldani, R. Safabakhsh, and S. Momtazi, "Convolutional neural network with margin loss for fake news detection," *Information Processing & Management*, vol. 58, no. 1, p. 102418, 2021.

[50] J. K. Rout, S. Singh, S. K. Jena, and S. Bakshi, "Deceptive review detection using labeled and unlabeled data," *Multimedia Tools and Applications*, vol. 76, no. 3, pp. 3187–3211, 2017.

[51] P. K. Jain, R. Pamula, and S. Ansari, "A supervised machine learning approach for the credibility assessment of user-generated content," *Wireless Personal Communications*, vol. 118, no. 4, pp. 2469–2485, 2021.

[52] J. K. Rout, A. Dalmia, K.-K. R. Choo, S. Bakshi, and S. K. Jena, "Revisiting semi-supervised learning for online deceptive review detection," *IEEE access*, vol. 5, pp. 1319–1327, 2017.

[53] E. F. Cardoso, R. M. Silva, and T. A. Almeida, "Towards automatic filtering of fake reviews," *Neurocomputing*, vol. 309, pp. 106–116, 2018.

[54] E. Elmurngi and A. Gherbi, "Detecting fake reviews through sentiment analysis using machine learning techniques," *IARIA/data analytics*, pp. 65–72, 2017.

[55] A. Mukherjee, V. Venkataraman, B. Liu, N. Glance *et al.*, "Fake review detection: Classification and analysis of real and pseudo reviews," *UIC-CS-03-2013. Technical Report*, 2013.

[56] R. Mohawesh, S. Tran, R. Ollington, and S. Xu, "Analysis of concept drift in fake reviews detection," *Expert Systems with Applications*, vol. 169, p. 114318, 2021.

[57] K. Sanjay and A. Danti, "Detection of fake opinions on online products using decision tree and information gain," in *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*. IEEE, 2019, pp. 372–375.

[58] A. Sihombing and A. C. M. Fong, “Fake review detection on yelp dataset using classification techniques in machine learning,” in *2019 International Conference on contemporary Computing and Informatics (IC3I)*. IEEE, 2019, pp. 64–68.

[59] G. Gravanis, A. Vakali, K. Diamantaras, and P. Karadais, “Behind the cues: A benchmarking study for fake news detection,” *Expert Systems with Applications*, vol. 128, pp. 201–213, 2019.

[60] A. Mahabub, “A robust technique of fake news detection using ensemble voting classifier and comparison with other classifiers,” *SN Applied Sciences*, vol. 2, no. 4, pp. 1–9, 2020.

[61] P. Meel and D. K. Vishwakarma, “Han, image captioning, and forensics ensemble multimodal fake news detection,” *Information Sciences*, vol. 567, pp. 23–41, 2021.

[62] J. A. Nasir, O. S. Khan, and I. Varlamis, “Fake news detection: A hybrid cnn-rnn based deep learning approach,” *International Journal of Information Management Data Insights*, vol. 1, no. 1, p. 100007, 2021.

[63] A. Roy, K. Basak, A. Ekbal, and P. Bhattacharyya, “A deep ensemble framework for fake news detection and classification,” *arXiv preprint arXiv:1811.04670*, 2018.

[64] Q. Li, Q. Hu, Y. Lu, Y. Yang, and J. Cheng, “Multi-level word features based on cnn for fake news detection in cultural communication,” *Personal and Ubiquitous Computing*, vol. 24, no. 2, pp. 259–272, 2020.

[65] Y. Wang, L. Wang, Y. Yang, and T. Lian, “Semseq4fd: Integrating global semantic relationship and local sequential order to enhance text representation for fake news detection,” *Expert Systems with Applications*, vol. 166, p. 114090, 2021.

[66] P. Wanda and H. J. Jie, “Deepprofile: Finding fake profile in online social network using dynamic cnn,” *Journal of Information Security and Applications*, vol. 52, p. 102465, 2020.

[67] M. Balaanand, N. Karthikeyan, S. Karthik, R. Varatharajan, G. Manogaran, and C. Sivaparthipan, “An enhanced graph-based semi-supervised learning algorithm to detect fake users on twitter,” *The Journal of Supercomputing*, vol. 75, no. 9, pp. 6085–6105, 2019.

[68] S. Kar, S. Sethi, and R. K. Sahoo, “A multi-factor trust management scheme for secure spectrum sensing in cognitive radio networks,” *Wireless Personal Communications*, vol. 97, no. 2, pp. 2523–2540, 2017.

[69] Matlab, “Triangular membership function,” https://www.mathworks.com/help/fuzzy/trimf.html(retrieveason21/09/2021), May 2021.

Ramesh K. Sahoo Continuing Ph.D. in Computer Science, Department of Utkal University Bhubaneswar, India. He is working as Assistant Professor at the Indira Gandhi Institute of Technology, Sarang, India. He is working in the field of cognitive science, crowdsensing, data analytics, wireless sensor network, IoT, and cognitive network.

Sateesh Kumar Pradhan Professor Sateesh Kumar Pradhan completed his Ph.D. in Computer Science from Berhampur University, Berhampur, Odisha, India. The title of his Ph. D. thesis is “Neuron-Based Parallel Algorithms for Solution of Linear Systems and Digital Signal Processing Applications”. He was the Former Professor and Head of, the P.G. Department of Computer Science, and Former Dean Faculty of Engineering, at Utkal University, Bhubaneswar, Odisha, India. Twenty-two research scholars have been awarded Ph. D. degrees in the area of E-Commerce, Mobile Ad hoc Networks, Intrusion Detection, Computer Forensics, Web Services, Parallel Task Scheduling, Medical Data Mining, Software Testing, Internet of Things, Brain Image Classification, Optical Character Recognition of Odia Document, etc. under his supervision and currently eight scholars are continuing for Ph. D. degree under his supervision and guidance. Prof. Pradhan was also the Organizing Chair of the International Conference on Information Technology (ICIT-2005). There are more than forty Journal Publications and more than Sixty Conference Publications to his credit. He also served as Senior Faculty in Computer Engineering, at King Khalid University, Abha, Saudi Arabia (2006-2011) and as Faculty, of Computer Science, at Berhampur University, Odisha, India (1987-2000).

Srinivas Sethi Dr. Srinivas Sethi is Professor and has been actively involved in teaching and research in Computer Science since 1997. He did his Ph.D., in the area of Mobile Ad hoc Networks, and is also continuing his work in wireless communication, sensor network, cognitive radio network, IoT, BCI, and cloud computing. He was Volume Editor for Springer LNNS, International Conference Proceedings, Board for different journals, and Program Committee Member for different international conferences/workshops. Now he is working as Faculty in the Department of Computer Science Engineering and Application at Indira Gandhi Institute of Technology Sarang, India, and has published more than 80 research papers in International journals and conference proceedings. He completed 7 research projects funded by different Government of India funding agencies such as DST, AICTE, NPIU, and DRDO.

Siba K. Udgata Dr. Siba Kumar Udgata is presently serving as a Professor of Computer Science at the AI Lab, School of Computer and Information Sciences at the University of Hyderabad. He also served as Director at the Center for Modeling for Simulation and Design (CMSD), University of Hyderabad. He was a research fellow at International Institute for Software Technology, United Nations University (UNU/IIST), Macau. His main research interests include Wireless communication, IoT and Sensor Networks, and Intelligent Algorithms. He has authored more than 100 research papers published in reputed international journals and conferences. He has worked as principal investigator in many Government of India (and other funding agencies) funded research projects mainly for the development of wireless sensor network applications, network security related applications, and application of swarm intelligence techniques in the cognitive radio network domain. He has worked as a consultant to Tata Steel Ltd and Scientific Analysis Wing (SAG), DRDO, Govt. of India. He has been awarded with IBM SUR (Shared University Research) Award for the research project, “Mobile and Sensor Network based Disaster Management System with emphasis on rescue management” He has around eight edited book volumes published by Springer and also recently coauthored a book entitled “Internet of Things and Sensor Network for COVID-19” published by Springer Nature publication house.