Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 6274114, 15 pages
https://doi.org/10.1155/2022/6274114

Research Article

IOTA-Based Mobile Crowd Sensing: Detection of Fake Sensing Using Logit-Boosted Machine Learning Algorithms

Mazhar Hameed,[1] **Fengbao Yang**,[1] **Muhammad Imran Ghafoor**,[2] **Fawwad Hassan Jaskani**,[3] **Umar Islam**,[4] **Muhammad Fayaz**,[5] **and Gulzar Mehmood**[4]

[1]*School of Information and Communication Engineering, North University of China, China*
[2]*Department of Electrical Engineering, Superior University Lahore, Pakistan*
[3]*Department of Computer Systems Engineering, Islamia University of Bahawalpur, Pakistan*
[4]*Department of Computer Science, IQRA National University, Swat Campus, Pakistan*
[5]*Department of Computer Science, University of Central Asia, Naryn, Kyrgyzstan*

Correspondence should be addressed to Muhammad Fayaz; muhammad.fayaz@ucentralasia.org

Received 28 December 2021; Revised 2 March 2022; Accepted 3 March 2022; Published 23 April 2022

Academic Editor: Narasimhan Venkateswaran

In the Internet of Things (IoT) era, the mobile crowd sensing system (MCS) has become increasingly important. The Internet of Things Auto (IOTA) has evolved rapidly in practically every technology field over the last decade. IOTA-based mobile crowd sensing technology is being developed in this study using machine learning to detect and prevent mobile users from engaging in fake sensing activities. It has been determined through testing and evaluation that our method is effective for both quality estimation and incentive allocation. Using the IOTA Bottleneck dataset, multiple performance metrics were used to demonstrate how well logit-boosted algorithms perform. After applying logit-boosted algorithms on the dataset for the classification, Logi-XGB scored 95.7 percent accuracy, while Logi-GBC scored 90.8 percent accuracy. As a result of this, Logi-ABC had an accuracy rate of 89%. Logi-CBC, on the other hand, got the highest accuracy of 99.8%. Logi-LGBM and Logi-HGBC both scored 91.37 percent accuracy, which is identical. On the given dataset, our Logi-CBC algorithm outperforms earlier Logit-boosted algorithms in terms of accuracy. Using the new IoTA-Botnet 2020 dataset, a new proposed methodology is tested. In comparison to prior Logit-boosted algorithms, the new model Logi-CBC has a highest detection accuracy of 99.8%.

1. Introduction

IOTA is a new computing model that has rapidly expanded in nearly every technology industry in the last decade, including smart anomaly detection and intelligent security systems, intelligent banking systems, cryptocurrencies, sensor use, smart cities, and satellites [1]. These devices have sensors, actuators, storage, processing, and networking capabilities to collect and share data via the internet [2]. An IOTA network collects and processes sensitive data; therefore, it must be protected against potential threats in order to function properly [3].

In order to protect sensitive data from vulnerable mobile device security threats such as the distributed denial-of-service (DDoS) attack, firewalls, authentication systems, various types of encryption, antivirus, and other security measures are currently being implemented. Firewalls are the first line of defense against DDoS attacks and other mobile device security threats [4]. IOTA has the potential to enable software-defined networking (SDN), future network structure, deep learning (DL), artificial intelligence (AI), and machine learning, all of which are examples of data networking and cloud network computing, VoIP fiber optics, global microwave access interoperability (WiMAX), deep learning (DL), AI, and machine learning, all of which are examples of data networking (NDN). Because of the integration of a huge amount of data, a high number of new anomalies (both unique and mutations of an existing anomaly) are produced on a regular basis [5]. Because of this, a second-line defensive intrusion detection system (IDS) for an IOTA network

can provide additional security protection for the network [6]. Mobile crowd sensing (MCS) allows smartphone users to explore a new path in the Internet of Things by using their phones to gather information from others (IoT). Volunteers collect data from their surroundings by utilizing smartphone functions such as GPS and camera (camera, temperature, GPS, microphone, etc.) [7].

Attacks can be classified based on the methods of deployment and detection that are used. An intrusion detection system (IDS) can be either host-based or network-based, as well as signature-based, specifier-based, or hybrid detection, depending on the detection technique used to detect the intrusion [8]. The goal of this research is to apply the machine learning-based fake sensing detection technique to provide IOTA security at an early stage in the development process.

Machine learning (ML) and deep learning (DL) techniques have recently been investigated to improve detection accuracy and minimize NIDS false alarm rate. In research, both ML and DL techniques have been proven to extract meaningful patterns from network data to classify flows as anomalous or benign. Logit-boosted algorithms has demonstrated speed in learning valuable characteristics from raw data and has emphasized integrating IOTA networks into NIDS [9]. Logit-boosted algorithm is a machine learning approach researched by academics in data mining, data science, and network security [10]. Because of their in-depth design, Logit-boosted algorithms have done remarkably well in specific industries, providing various abstractions for using complex learning elements to predict effectively. Because of the vast amount of data produced by IOTA mobile devices, these qualities of logit-boosted algorithms have made it ideal for a fake sensing detection system designed for an IOTA network. In this research, we look at the possibility of employing different Logit-boosted algorithms to present a cost-effective IOTA mobile fake crowd sensing detection solution.

2. Literature Review

Throughout the last decade, researchers have been investigating artificial intelligence technologies such as machine learning and deep learning to provide effective NIDS solutions [11, 12]. Advancements in graphical processing unit (GPU) technology have answered the speedy calculation need for DL algorithms. According to current NIDS trends, DL methods have favored ML algorithms over three years. It has inspired scientists to apply the DL algorithms in an IOTA network to develop effective security solutions that process large numbers of raw data [13, 14]. Because of its deep structure, the DL can learn the complex pattern and aid in classifying benign and pathological traffic. Researchers in the field of NIDS commonly use machine learning techniques. Ali et al., for example, suggested IDS based on the particle swarm algorithm that uses a fast learning network. Despite being efficient enough to predict most attacks, the performance of the minority class label detection model was not encouraging. Shen et al. developed an ensemble approach methodology that included applying the Bat-optimization algorithm during the ensemble cutting step. In another fantastic piece, Yao et al. explain a multilevel semisupervised machine learning model that incorporates clustering and the random forest approach [14, 15]. Their methodology has been successful in detecting multilevel assault classes. Researchers also use ML and DL methodologies to produce successful NIDS solutions using various hybrid strategies. These methods are investigated utilizing DL algorithms for feature and complexity reduction, followed by a machine learning predictor [15, 16].

Recent research has pointed out that a bid is a private piece of information, and bidding-preserving algorithms with differential privacy have been proposed to protect against inference attacks [17]. However, all of these methods rely on a trusted platform, and they would all fail in terms of bid protection if the platform were not charged. Using innovative privacy-preserving incentive mechanisms, we can secure users' genuine bid information from the honest-but-curious platform while also reducing the societal cost of the winner selection process. Instead of uploading the genuine bid to the platform, a differentially private bid obfuscation method based on the exponential mechanism is devised, which allows each user to obfuscate bids locally before submitting the obfuscated bids to the platform.

Another hybrid concept occurs in paper [18] when sparse AE is combined with support vector machine (SVM). Using this methodology, minor anomaly labels have also proven difficult to locate. Marie et al. established another hybrid way to merge the deep-belief network (DBN) with SVM, using the ensemble approach. Researchers have also proposed effective NIDS models using stand-alone DL techniques, including AE, recurrent neural network, DBN, convolutional neural network (CNN), and Morlet neural wavelet network. For example, [19]. As a memory unit, Xiao et al. suggest a recurrent neural network- (RNN-) based NIDS that uses gated recurrent units. Also available is a CNN-based technique that uses primary component analysis and AE for functional extraction tasks, followed by CNN for prediction [18]. Using their approaches, only the class label with the most occurrences was successful [20] by merging the CNN with a bidirectional short-term memory to give another extremely complex NIDS technique, long short-term memory (LSTM) [21].

For various reasons such as crowdsourcing, new technologies and requirements have made the public available with many mobile and social network smart. Mobile crowd sensing applications must secure sensing against threats like jamming, bogus sensing attacks, and other threats during transmission. Previous research on a certain subject was done. But the new sensing paradigm requires an inventive protective solution. This research [22] investigates advanced mobile crowd sensing security using SVM (support vector machine) and ANN (artificial neural network) approaches. The author used full-blown implementation and experimental evaluation approaches, focusing on precision and false alarm rate. The accuracy and false alarm rate of artificial neural networks were compared to SVM. Using 10-fold cross-validation, the proposed ANN attained an average of 96.4% and less than 7% of the usual erroneous positive rate.

Table 1: Comparative analysis.

References	Technique	Dataset	Outcome	Efficiency
Yang et al. [26]	Data quality-aware truth estimation and surplus sharing method for Mobile crowd sensing	Real-time data for mobile sensing	Quality estimation, mobile crowd sensing	89%
Arafeh et al. [27]	Blockchain-based architecture	MCC dataset	Detection of fake sensing in Mobile crowd sensing	92%
Kucuk et al. [28]	Design with IoT technologies	IoT-based data	Crowd sensing aware disaster	80%
Mrazovic [29]	Crowd sensing-driven route optimization algorithms	Self-created	Smart urban mobility	93%
Haseeb et al. [30]	Crowd sensing IOT based	Real-time data for mobile sensing	Detection of fake sensing in mobile crowd sensing	91%
Kianoush et al. [31]	Blockchain-based fake detection	Self-created	Detection of fake sensing in mobile crowd sensing	87%
Owoh and Singh [32]	Deep learning-based fake sensing	Real-time data for mobile sensing	Detection of fake sensing in mobile crowd sensing	85%
Ali Al-Muqarm and Rabee [33]	Cloud computing/edge computing	Cloud-based dataset	Detection of fake sensing in mobile crowd sensing	82%
Zhou et al. [34]	Wifi-based route optimization and mobility crowd sensing	Wifi-based data collection	Detection of fake sensing in mobile crowd sensing	83%
Louta et al. [35]	Blockchain/federated learning	Real-time data for mobile sensing	Detection of fake sensing in mobile crowd sensing	88%
Sisi and Souri [36]	Blockchain	Real-time data for mobile sensing	Quality estimation, mobile crowd sensing	90%
Reddy et al. [22]	Machine learning, support vector machine	Data for mobile sensing	Quality estimation, mobile crowd sensing	96.4%
Feng et al. [24]	Machine learning, random forests	Data for mobile sensing	Quality estimation, mobile crowd sensing	87.5%

We strongly advise the use of artificial neural networks for mobile crowd sensing.

Liu et al. [23] present FEDXGB, a federated XGBoost system with forced aggregation. FEDXGB primarily makes two advances. A new secure aggregation strategy for FL is FEDXGB. The solution overcomes the second barrier by combining secret sharing with homomorphic encryption. Then, FEDXGB applies the secure aggregation approach to XGBoost's classification and regression tree building. We also perform rigorous theoretical and experimental evaluations of FEDXGB's security, effectiveness, and efficiency. With FEDXGB, FL model update aggregation is 23.9 percent faster and 33.3 percent less communication intensive than with the original XGBoost.

Feng et al. [24] create a platform to collect data in the real world, including user images. Combining online and offline learning improves the time complexity and accuracy of the random forest method. This technique is compared to two baselines: subsets of complete datasets and six conventional models (such as logistic and naive Bayes). Evaluators utilize six indices to assess performance: precision, recall, TPR, F-measure, and receiver operating characteristic curve area. The experimental findings suggest that our method surpasses other methods in estimation accuracy (precision: 0.875, recall: 0.872). In another study, Akhtar and Feng [25] used the Shapley value technique along with deep learning models for crowd fake sensing detection and model which achieves good accuracy in quality assessment and anomaly detection. Table 1 shows the comparative analysis of some of the state of art researches based on machine learning, blockchain, IoT, and boosting algorithms for mobile crowd sensing:

Furthermore, machine learning-based IDS research is still in its infancy on the IOTA network. Thus, there is plenty of space for additional research in this area. We describe a Logit-boosted algorithm-based intrusion detection system and solution for an IOTA network to achieve this goal. The importance of the Logit-boosted method's performance qualities for an IOTA network is discovered in particular.

The study's significant contributions are divided into four categories:

(1) To investigate the current state-of-the-art novel Logit-boosted algorithm-based crowd fake sensing through mobile devices

(2) Using Logit-boosted models, we provide an effective technique for detecting IOTA anomalies

(3) Using the IoT-Botnet 2020 [37] dataset and analyzing its effectiveness, we intended to evaluate the efficiency of our model with other deep learning models based on different Logit-boosted techniques

(4) This study is aimed at seeing how numerical and categorical factors affected the performance of Logit-boosted network-based intrusion detection system models

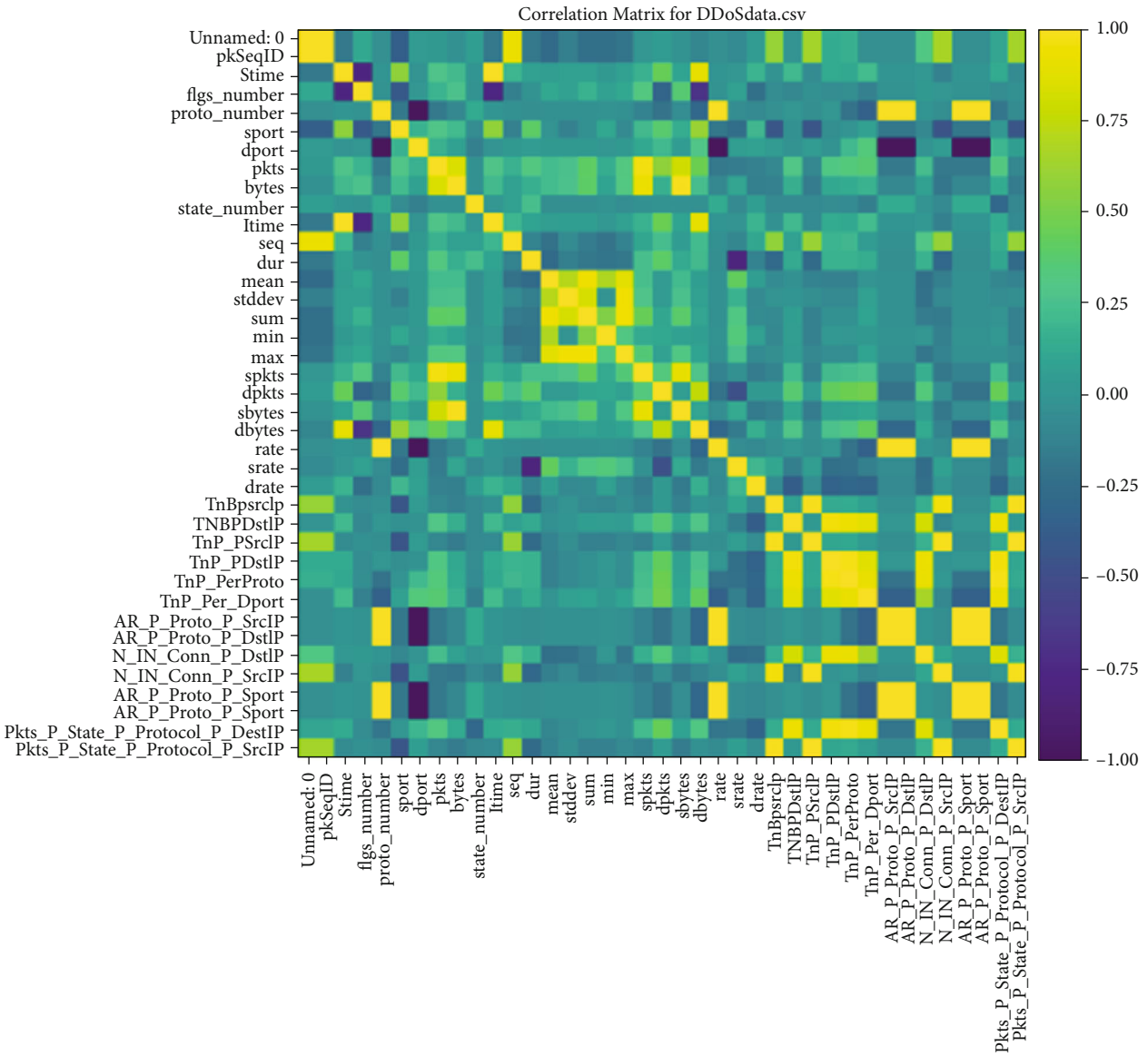

Figure 1: Feature correlated matrix.

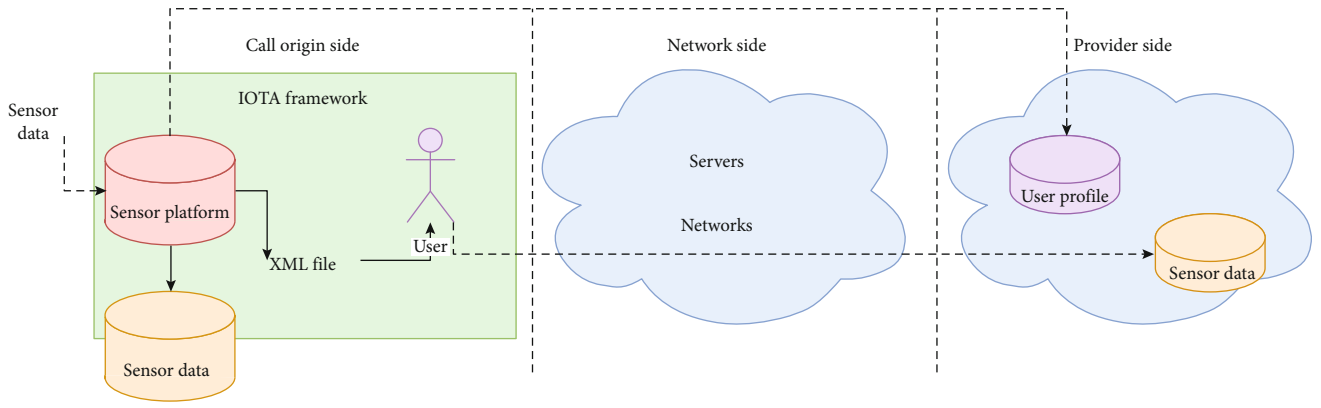

Figure 2: MCS IOTA-based architecture.

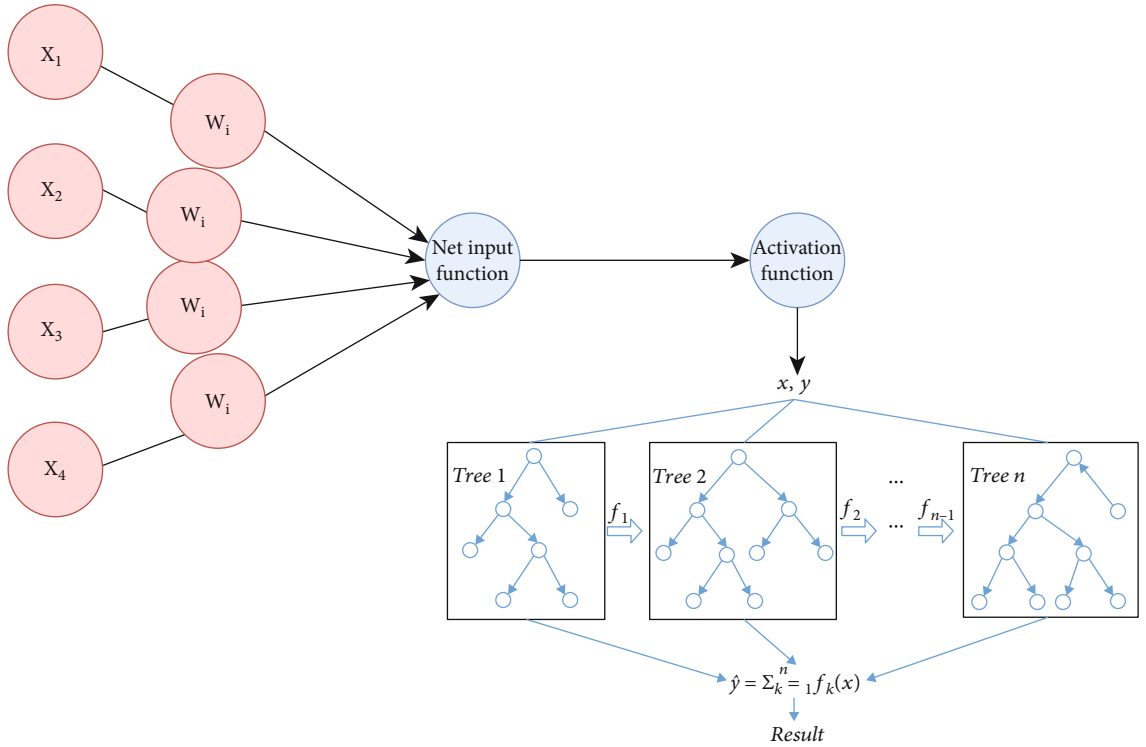

Figure 3: Hybrid classifier (Logi-XGB classification) model.

3. Material and Methods

This section shows the detailed methodology of the proposed work. The system has been divided into two platforms, i.e., IOTA and Logit-boosted models.

3.1. Framework of iMCS

3.1.1. MCS IOTA Framework in Caller Side. In iMCS, the IoTA framework is used on the caller side. As illustrated in Figure 1, iMCS architecture consists of the sensor platform, the sensor data interface, and the SIP client. Other instruments (e.g., laptops, tablets, televisions, microphones, and speakers), as well as sensors for false and abnormality detection, can be added to the calling side. The IoTA framework can be found on the caller's other end of the line. Figure 2 shows the sensor platform, the sensor data interface, and the SIP client. Additionally, numerous devices (such as laptops, tablets, TV sets, microphones, and speakers) and environmental sensors (such as humidity, smoke detectors, and motion sensors) can be installed on the caller's end.

3.1.2. MCS IOTA Framework in Network Side. The sensor data interface has been established as a user-server model to facilitate communication between the sensor platform and the user on the network side. In XML format, the interface accepts the critical data. Sensor model language has been chosen as the standard, unified data representation model. This file will also be utilized as a parameter in a Logit-boosted model on the provider side.

3.2. Data Collection and Preprocessing. The raw data was acquired from the IOTA Bottle Neck Dataset as it was previously used in [25]. As a result, the data has been cleaned using various strategies, such as removing duplicates and removing null values.

3.3. Feature Engineering. Feature engineering is a process that leverages data from a specific domain to build functions used by learning machines. It analyzes raw data and transforms it into machine learning formats by extracting the most important attributes. The correlation matrix is utilized in this study to determine the relationship between variables. The mobile device categorization model was built using individual traffic filtering and a pcap file created using the device's MAC address. The IP address supplied to a device by DHCP servers (dynamic host configuration protocol) can change over time and is therefore not a reliable feature for accurately filtering traffic to a single device over time. The traffic characteristics of each mobile device (41 in total) included in the study are monitored at the traffic flow level. To classify traffic flows, packets with the same source and destination addresses as well as communication ports and protocols (TCP (transmission control protocol) or UDP (user datagram protocol)) are grouped. According to the packet header's aggregated (statistical) data, traffic flow is chosen as the observation and analysis level best portrays communication between source and destination. Packet-level traffic analysis demands more processing power and storage capacity to store and analyze extra data. There is a correlation between the number of traffic flows and the number of packets that Google Chromecast (the device analyzed in this study) sends over 24 hours. Figure 1 shows the correlation matrix of dataset features.

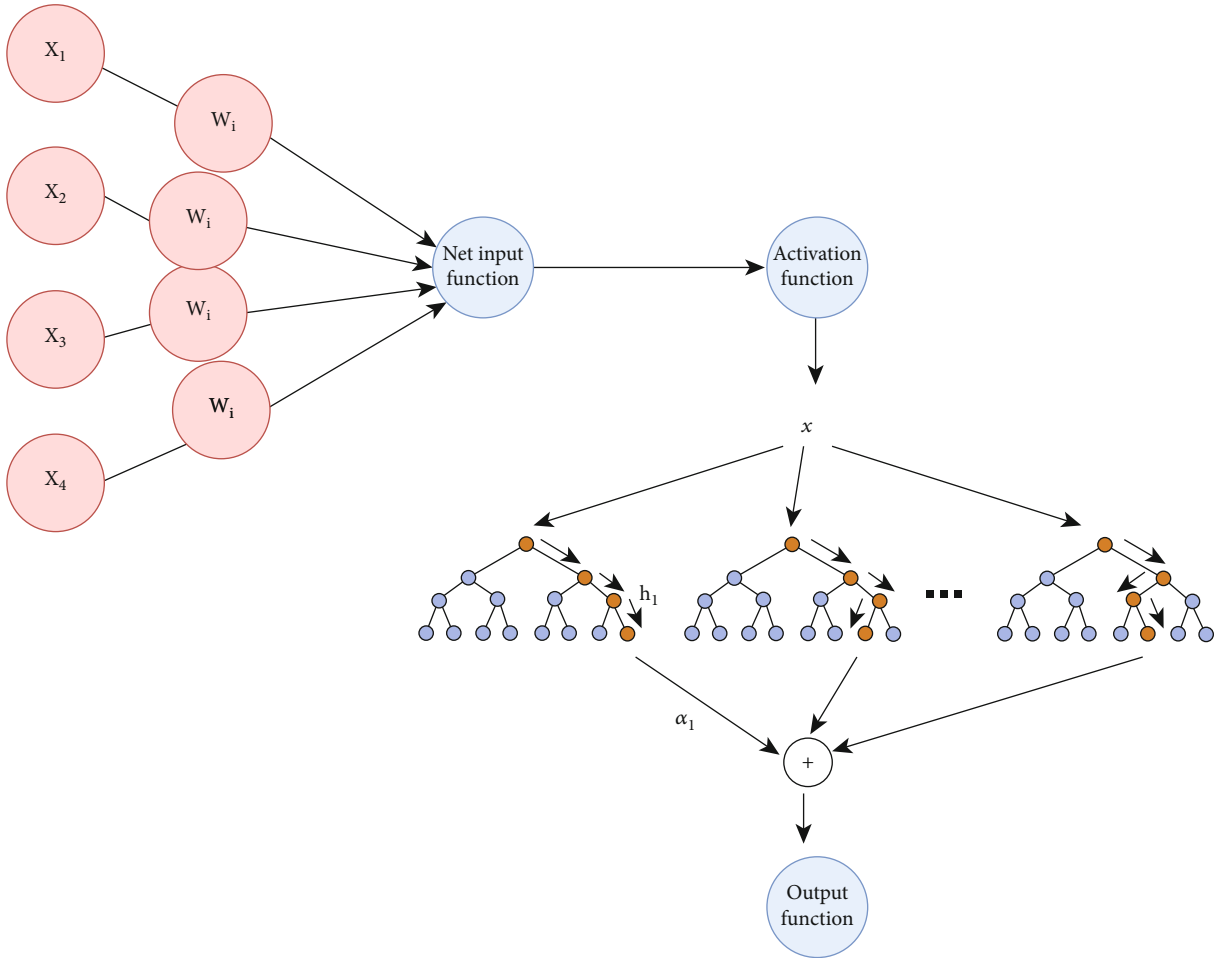

Figure 4: Hybrid classifier (Logi-GBC classification) model.

3.4. Calculation of Index Feature. Predictability in IoT device behavior is a phenomenon that has emerged as a result of research into IoT devices' communication activities. Given that mobile devices have limited capabilities, their behavior will be very predictable over time. A limited number of applications can be run on devices that are not connected to the Internet of Things (IoT). On the other hand, IoT devices rely only on their end-users for communication activities. Mobile devices, as a result, can be predicted using the index of the amount of predictability of IoT devices (Cu index) over time. The closer the index (Cu) gets to 0, the more predictable it is, and the less it differs from the quantity of data received and sent. It is possible to calculate an index feature:

$$C_u = C\,\mathrm{var}_u \frac{\sqrt{(1/(N-1))\sum_{i=1}^{N}(x_i - x_{i*})^2}}{(1/N)\sum_{i=1}^{N}x_i}. \quad (1)$$

3.5. Data Preprocessing. An important step in the data mining is transforming raw data into something usable. When it comes to some behaviors and patterns, our data is often partial, mismatched, or lacking altogether. The Cu index value was used to classify the device classes using the coefficients of variation classification approach. It assumes a normal distribution of the data. Because the derived values (Cu index) distribution is biased to the left, the data are transformed. Using the ladder of powers method, researchers were able to find the best data transformation function for the study to construct a normal distribution.

3.6. Data Normalization. Normalization is a common practice in data preparation for machine learning. You must normalize your data to a standard scale without distorting the range of numbers or surrendering any information if you want it to be consistent.

3.7. Logit-Boosted Model Development

3.7.1. Data Balancing. An obstacle to accurate predictive modeling is an unbalanced set of classifications. It is common for machine learning algorithms to use the same number of examples in each class. It results in inaccurate models, especially for minorities. It is a problem since the minority group is more important and more susceptible to mistakes in classification than the majority group. As a result, we could eliminate the outliers from the sample and bring the dataset back into line. Many more sophisticated resampling

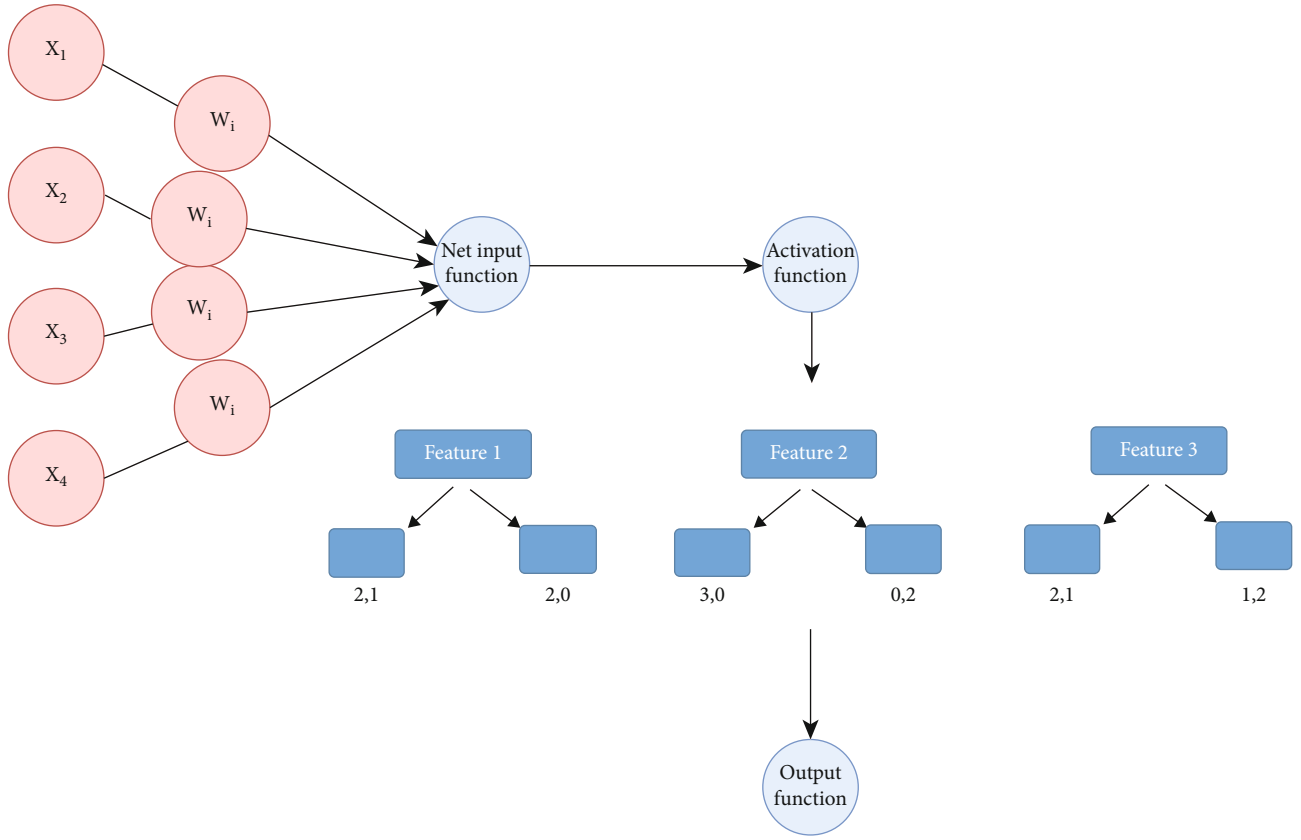

Figure 5: Hybrid classifier (Logi-ABC classification) model.

algorithms have been proposed due to this research. For example, we can aggregate most class records under sampling to conserve information by extracting records from each cluster. Instead of making exact replicas of minority class data, we can introduce small changes to these versions during the sampling process, resulting in more diversified synthetic samples. Data mining research requires a well-balanced and uniform dataset. In a dataset, "outliers" can be found. Outliers are the values in a dataset that are different from the rest. The outlier has been normalized using SMOTE technique in to order to handle imbalanced dataset.

The outliers can be produced by misreading's, faulty devices, or human mistake. It must be omitted from the data before conducting any research or statistical tests. Any information outliner can generate partial or incorrect results, affecting the analysis and subsequent processing. The IQR approach eliminates outliers when the data boxplot exceeds the specified range. The discrepancy is the difference between the upper and lower quartiles' IQRs. Statistical approaches such as IQR, Z-score, and data smoothing are used in this study to find outliers in the data. The first quartile ($Q1$) and third quartile ($Q3$) of a dataset, i.e., the 25th and 75th percentiles, are used to calculate the IQR.

$$\text{IQR} = Q3 - Q1. \quad (3.2)$$

3.7.2. Hybrid Classification Algorithms. One relies on a limited number of complementary ways to categorization. The classification conclusion is based on a single method that solves various tasks. IoT devices can be categorized by the amount of data they send and receive. Each model's explanation is provided below.

3.7.3. Logi-XGB. This model has been developed by ensembling the logistic regression model into XGBoost classifier to improve both models' accuracy. Mathematical model of Logi-XGB classification model is as follows:

$$\begin{aligned} y &= \sum_{k=1}^{n} f(x), \\ \ln \frac{P}{1-P} &= a + by, \\ \frac{P}{1-P} &= e^{a+by}, \\ P &= \frac{e^{a+by}}{1+e^{a+by}}. \end{aligned} \quad (2)$$

Here, P is the probability function of logistic regression and Y is the output of XGBoost classification model. $\sum_{k=1}^{n} f(x)$ shows the boosting function of XGB classifier. When XGB takes the output of y, it will be sent to probability function of logistic regression for classification. Figure 3 shows the hybrid model of Logi-XGB classification model.

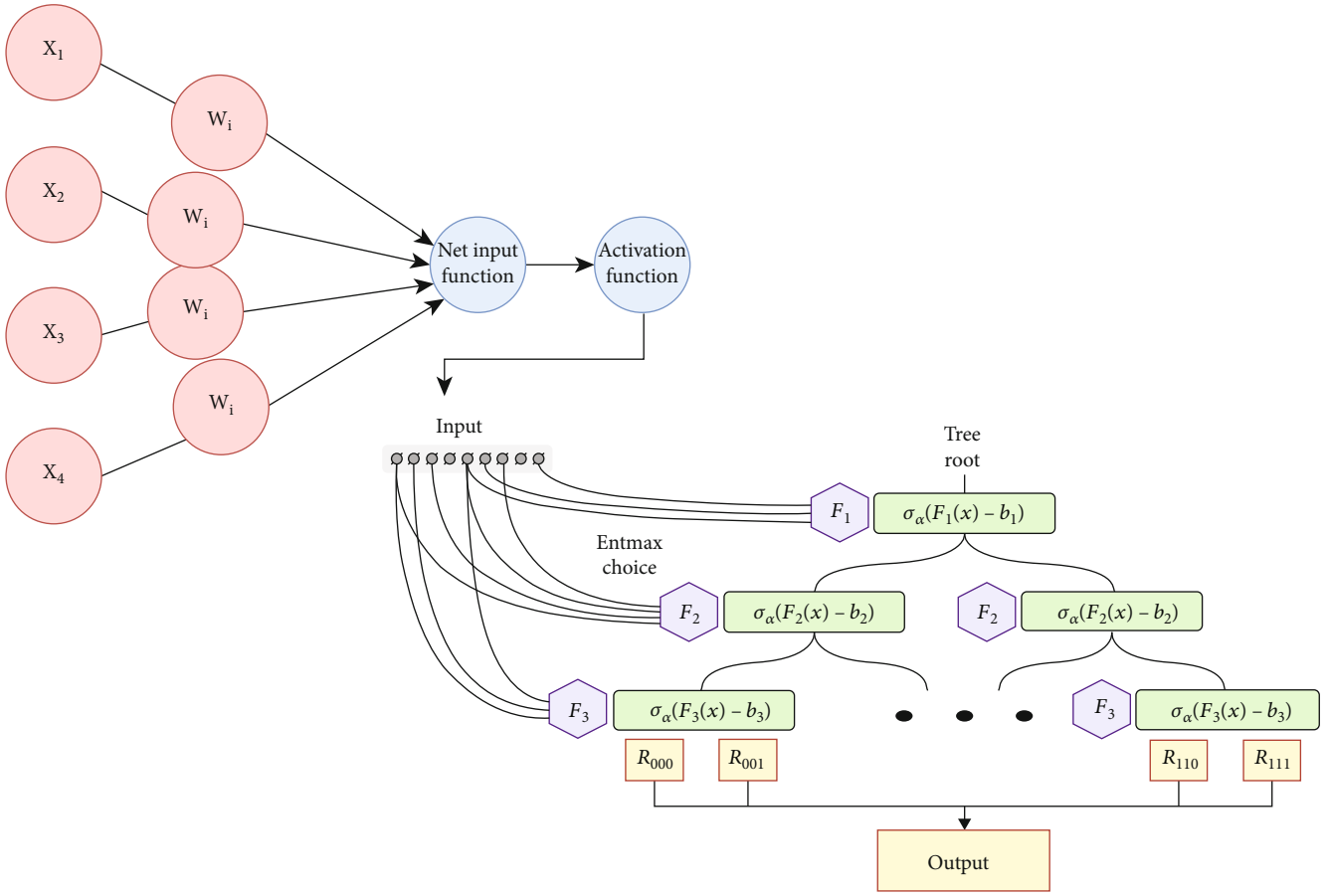

Figure 6: Hybrid classifier (Logi-CBC classification) model.

3.7.4. Logi-GBC. This model has been developed by ensembling the logistic regression model into gradient boosting classifier to improve both models' accuracy. Mathematical model of Logi-GBC classification model is as follows:

$$
\begin{aligned}
y = y^i = y^i + \alpha * \frac{\partial \sum (y_i - y_i^p)^2}{\partial y_p^i}, \\
\ln \frac{P}{1-P} = a + by, \\
\frac{P}{1-P} = e^{a+by}, \\
P = \frac{e^{a+by}}{1+e^{a+by}}.
\end{aligned}
\tag{3}
$$

Here, P is the probability function of logistic regression and y^i is the output of GBC classification model. $(\partial \sum (y_i - y_i^p)^2)/\partial y_p^i$ Shows the sum of residual in trees, and α is the learning rate of GBC. When GBC takes the output of y, it will be sent to probability function of logistic regression for classification. Figure 4 shows the hybrid model of Logi-GBC classification model.

3.7.5. Logi-ABC. This model has been developed by ensembling the logistic regression model into AdaBoost classifier to improve both models' accuracy. The mathematical model of Logi-ABC classification model is as follows:

$$
\begin{aligned}
y = \text{significance} \sum_{t=1}^{T} \alpha_t h_t(x), \\
\ln \frac{P}{1-P} = a + by, \\
\frac{P}{1-P} = e^{a+by}, \\
P = \frac{e^{a+by}}{1+e^{a+by}}.
\end{aligned}
\tag{4}
$$

Here, P is the probability function of logistic regression and y is the output of ABC classification model. $\sum_{t=1}^{T} \alpha_t h_t(x)$ shows the sum of residual in trees with significance α. When ABC takes the output of y, it will be sent to probability function of logistic regression for classification. Figure 5 below is the hybrid model of Logi-ABC classification model.

3.7.6. Logi-CBC. This model has been developed by ensembling the logistic regression model into CatBoost classifier to improve both models' accuracy. The mathematical model of Logi-CBC classification model is as follows.

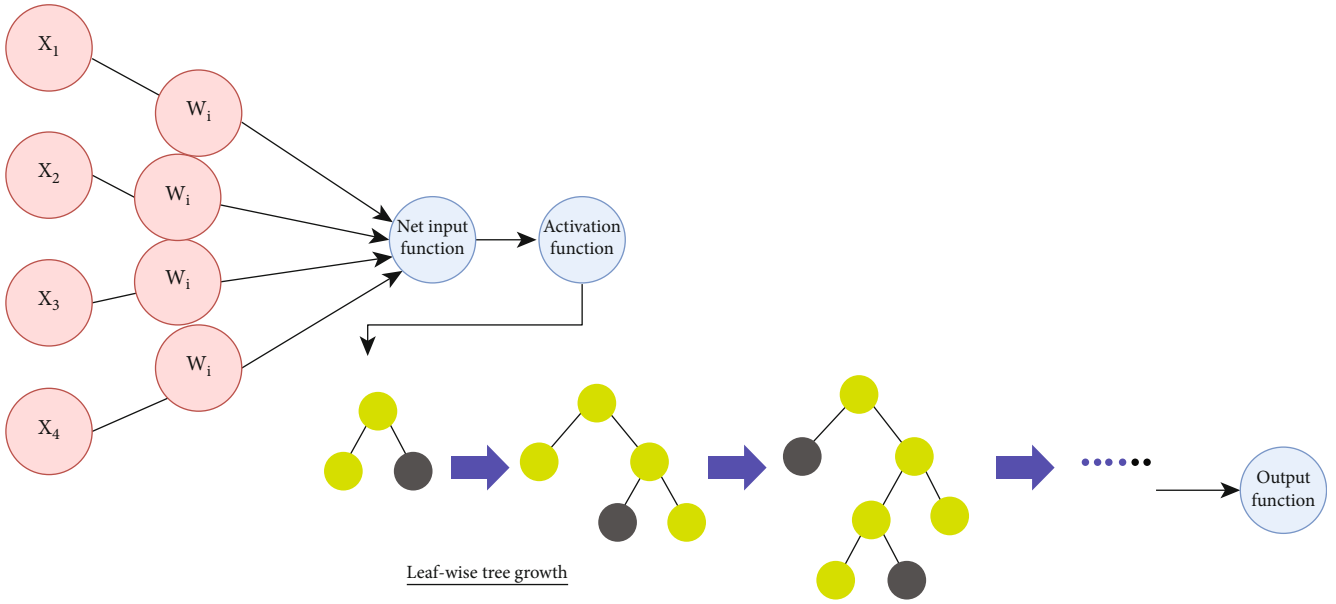

Figure 7: Hybrid classifier (Logi-LGBM classification) model.

In the first step, we will initialize the model:

$$F_o(x) = \text{argmin}_\gamma \sum_{i=1}^{n} L(y, \gamma). \tag{5}$$

For $m = 1$ to M, we will compute the residuals.

$$\gamma_{im} = -\left[\frac{\partial L[y, F(x_i)]}{\partial F x_i}\right]_{F(x)=F_{M-1}(x)}. \tag{6}$$

Then, we will fit the base leaner to compute it with pseudoresiduals:

$$\gamma_{im} = \text{argmin}_\gamma \sum_{xi}^{n} L(y, F_{M-1}(x)). \tag{7}$$

The updated model will be

$$\begin{aligned} F_m(x) &= F_{M-1}(x) + \alpha \sum_{i=1}^{n} \gamma_{im}, \\ \ln \frac{P}{1-P} &= a + bF_m(x), \\ \frac{P}{1-P} &= e^{a+bF_m(x)}, \\ P &= \frac{e^{a+bF_m(x)}}{1+e^{a+bF_m(x)}}. \end{aligned} \tag{8}$$

Here, P is the probability function of logistic regression and y is the output of CBC classification model. $[(\partial L[y, F(x_i)])/\partial F x_i]_{F(x)=F_{M-1}(x)}$ Shows the sum of residual in trees with significance α. When CBC takes the output of y as $\text{argmin}_\gamma \sum_{xi}^{n} L(y, F_{M-1}(x))$, it will be sent to probability function of logistic regression for classification. Figure 6 shows the hybrid model of the Logi-CBC classification model.

3.7.7. Logi-LGBM. This model has been developed by ensembling the logistic regression model into light-gradient boosting model classifier to improve both models' accuracy. The mathematical model of Logi-LGBM classification model is as follows:

$$\begin{aligned} y &= \alpha \sum_{t_i \epsilon \text{ Tree}}^{T} \eta^i * \text{leaf}(t_i), \\ \ln \frac{P}{1-P} &= a + by, \\ \frac{P}{1-P} &= e^{a+by}, \\ P &= \frac{e^{a+by}}{1+e^{a+by}}. \end{aligned} \tag{9}$$

Here, P is the probability function of logistic regression and y is the output of LGBM classification model. $\sum_{t_i \epsilon \text{ Tree}}^{T} \eta^i * \text{leaf}(t_i)$ shows the sum of residual in leaves with learning rate α. When LGBM takes the output of y, it will be sent to probability function of logistic regression for classification. Figure 7 shows the hybrid model of Logi-LGBM classification model.

3.7.8. Logi-HGBC. This model has been developed by ensembling the logistic regression model into histogram gradient boosting classifier to improve both models' accuracy.

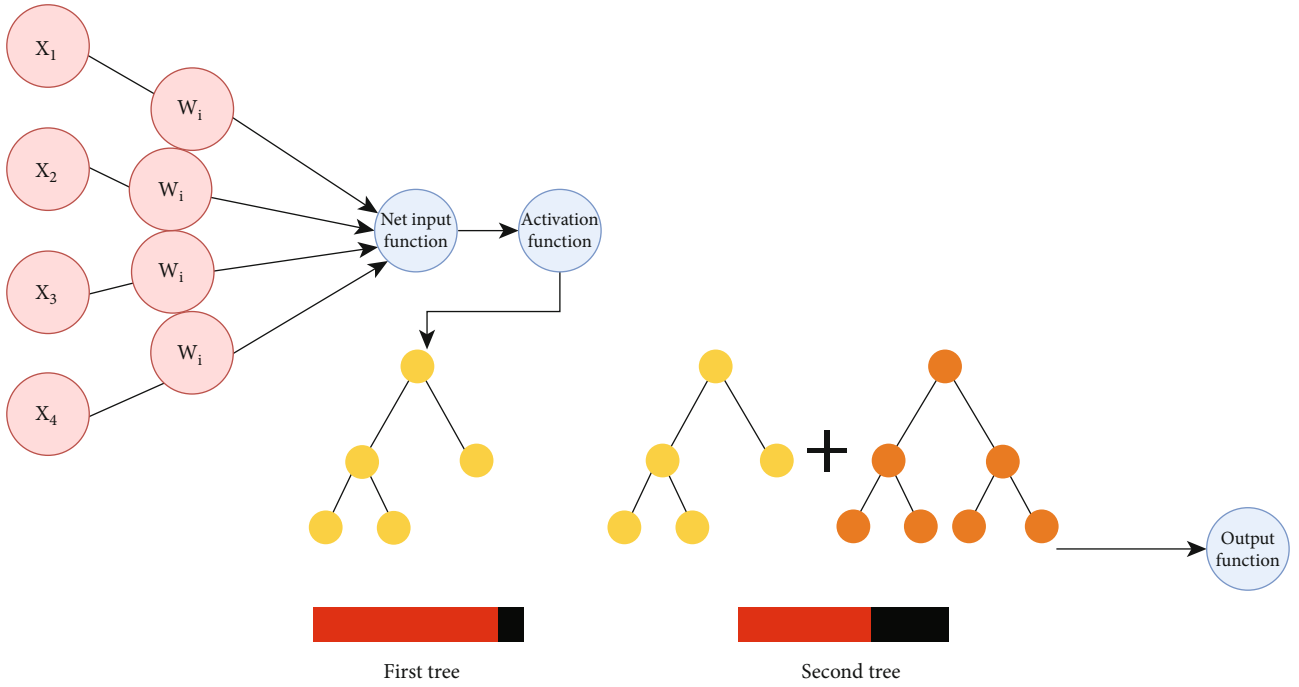

Figure 8: Hybrid classifier (Logi-HGBC classification) model.

Table 2: Description of metrics.

Metric	Description
Accuracy	$\text{Accuracy} = \frac{\text{TP}}{(\text{TP} + \text{TN}) * 100}$ True-positive (TP): the feature result is 1 and sample is present in this data file. True-negative (TN): the feature result is 0 and sample is absent in data file.
Confusion matrix	True Negative / False Positive False Negative / True Positive

The mathematical model of Logi-HGBC classification model is as follows:

$$y = \frac{\text{sum of residuals}}{\text{sum of each } (1-p) \text{for each sample in the leaf}},$$
$$\ln \frac{P}{1-P} = a + by,$$
$$\frac{P}{1-P} = e^{a+by}, \tag{10}$$
$$P = \frac{e^{a+by}}{1+e^{a+by}}.$$

Here, P is the probability function of logistic regression and y is the output of HGBC classification model. sum of residuals/sum of each $(1-p)$for each sample in the leaf shows the sum of residual in trees. When HGBC takes the output of y, it will be sent to probability function of logistic regression for classification. Figure 8 shows the hybrid model of the Logi-HGBC classification model.

3.8. Performance Parameters. $F1$ score and accuracy measures have been used to evaluate the system's accuracy. While the confusion matrix has indicated that the classified and misclassified clauses have been classified and misclassified, the metrics utilized in this investigation are shown in Table 2.

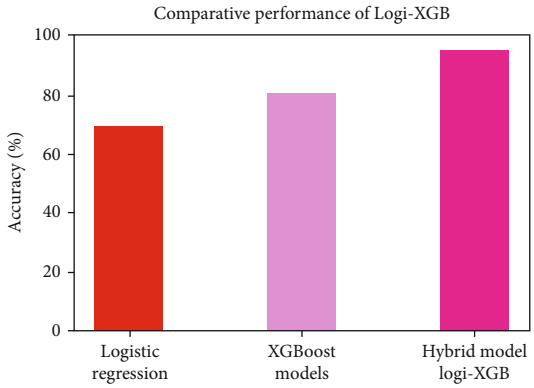

Figure 9: Logi-XGB classification model performance.

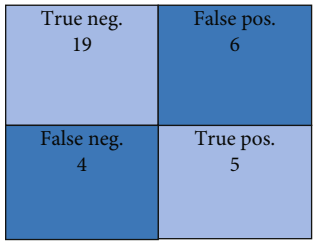

Figure 10: The confusion matrix of Logi-XGB classification model.

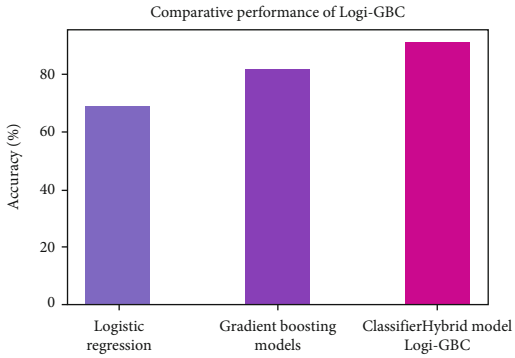

Figure 11: Logi-GBC classification model performance.

4. Results and Discussions

A device's class is determined by analyzing network flow data for ten days. Traffic flow feature vectors for mobile devices are categorized by class. The number of mobile traffic flows that a device creates in a given period depends on its mobile characteristics. A total of 681,684 feature vectors are separated into four classes for the initial dataset, as indicated above.

As a result, the majority class was undersampled in the dataset used to develop a classification model. The original dataset took into account the traffic flow of each unique device. Before creating a model, stratifying classes is essential to avoid model bias in classes with the most feature vectors. After stratification, the dataset contains 117,423 feature vectors that will further develop the classification model. The performance of log-boosted algorithms has been demonstrated in this section by displaying various performance

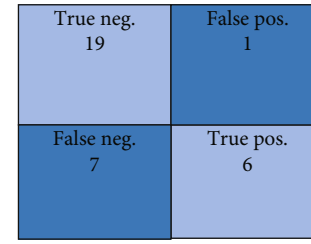

Figure 12: The confusion matrix of Logi-GBC classification model.

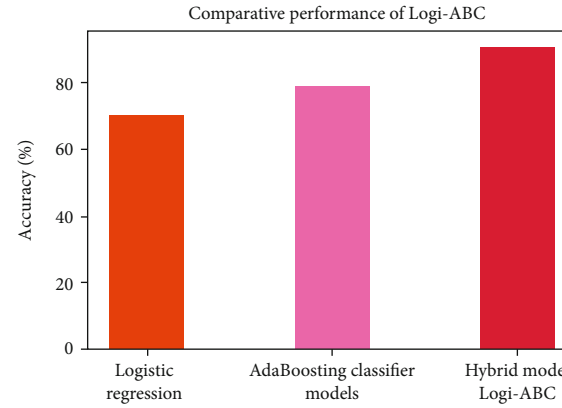

Figure 13: Logi-ABC classification model performance.

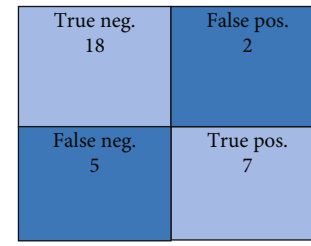

Figure 14: confusion matrix of Logi-ABC Classification Model.

metrics. On the other hand, Logi-XGB scored 95.7% accuracy, Logi-GBC with 90.8% accuracy. On the other hand, Logi-ABC scored 89.33% accuracy, while Logi-CBC scored the highest accuracy of 99.80%. Logi-LGBM and Logi-HGBC scored the same accuracy of 91.37%, respectively. Comparing with previous Logit-boosted algorithms implemented in previous studies, our proposed Logi-CBC has scored the highest accuracy on the given dataset.

4.1. Hybrid Model Logi-XGB. These two models have been combined in order to simultaneously increase their accuracy by using the XGBoost classifier. Logistic regression's probability function will be fed the data received from y by XGB. An independent logistic regression study found that 69.2 percent of the time, the hybrid classifier raised this accuracy to 95.7 percent. Figure 9 shows hybrid model of Logi-XGB classification model performance.

Figure 10 shows the confusion matrix of the Logi-XGB classification model with 19 true negative, 2 false positive, 9 false negative, and 3 true positive values.

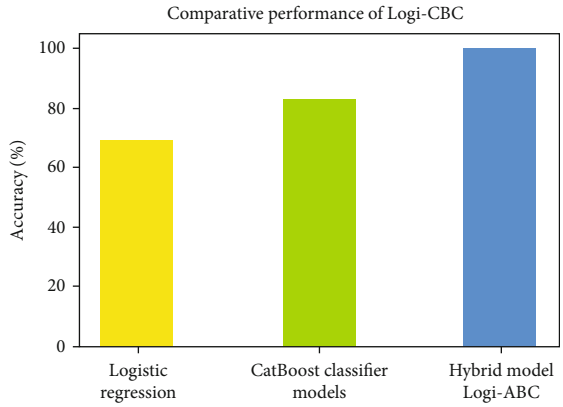

Figure 15: Logi-CBC classification model performance.

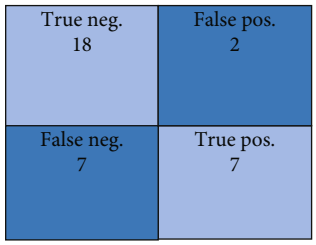

Figure 16: The confusion matrix of Logi-CBC classification model.

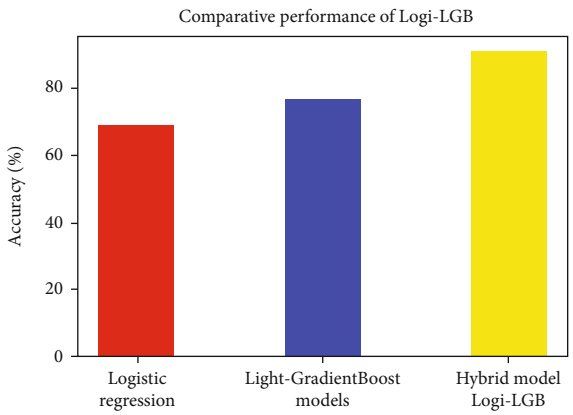

Figure 17: Logi-LGB classification model performance.

4.2. Hybrid Model Logi-GBC. Combining the logistic regression and gradient boosting classifier models was necessary to increase the accuracy of both models, which is how this model was constructed. Immediately upon receipt of *y*'s output by the GBC. Figure 11 below illustrates the performance of the hybrid model of the Logi-GBC classification model, which achieved an accuracy of 90.8 percent.

Figure 12 shows the confusion matrix of the Logi-GBC classification model with 18 true negative, 2 false positive, 9 false negative, and 7 true positive values.

4.3. Hybrid Model Logi-ABC. AdaBoost classifier was used with the logistic regression model in order to increase the accuracy of both models. It is submitted to the probability of logistic regression. With an accuracy rate of 89.33 percent,

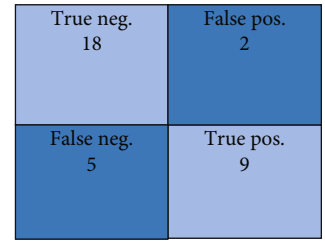

Figure 18: The confusion matrix of Logi-CBC classification model.

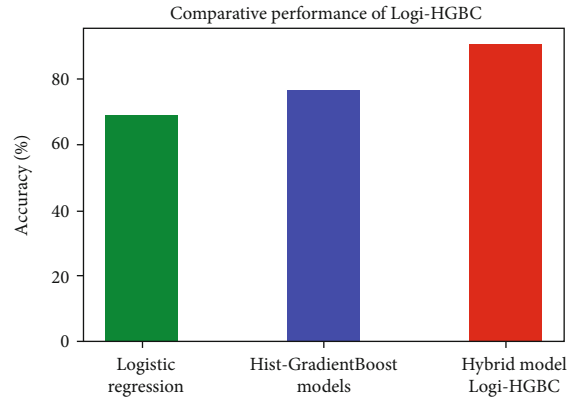

Figure 19: Logi-HGBC classification model performance.

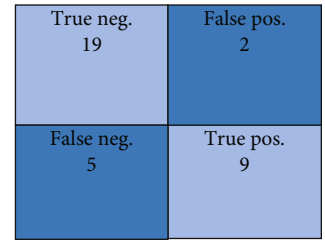

Figure 20: the confusion matrix of Logi-HGBC Classification Model.

Table 3: Comparative analysis.

Model	Accuracy
Logi-XGB	95.70%
Logi-GBC	90.80%
Logi-ABC	89.33%
Logi-CBC	99.80%
Logi-LGBM	91.37%
Logi HGBC	91.37%

the hybrid Logi-ABC classification model performance model is shown in Figure 13.

Figure 14 shows the confusion matrix of Logi-ABC classification model with 19 true negative, 8 false positive, 2 false negative, and 2 true positive values.

4.4. Hybrid Model Logi- CBC. In order to improve the accuracy of both models, the CatBoost Classifier was utilized to

Table 4: Comparative analysis with previous studies.

Reference	Dataset	Techniques	Accuracy
[38]	IoT-based SH dataset	Logit-boosted algorithms	81%
[39]	IoT network	Logit-boosted algorithms	84%.
[40]	NoIR-based IoT security system dataset	Logit-boosted algorithms	80.67%
Our proposed work	IoT dataset for smart home	Logit-boosted algorithms	85.66%

combine the logistic regression model with the CatBoost classifier. When CBC takes the output of y as $L(y, F(M-1)(x))$, it will be sent to the logistic regression's probability function for classification. As of this writing, Logi-CBC is the most accurate at 99.80%. The Logi-CBC classification model is depicted as a hybrid model in Figure 15.

Figure 16 shows the confusion matrix of the Logi-CBC classification model with 19 true negative, 2 false positive, 7 false negative, and 7 true positive values.

4.5. Hybrid Model Logi-LGB. The accuracy of both models can be improved by combining them. The LGBM will next use the likelihood function of logistic regression to classify the attacks. Figure 17 shows the hybrid Logi-LGBM classification model performance with the accuracy of 91.37%.

Figure 18 shows the confusion matrix of the Logi-LGB classification model with 18 true negative, 2 false positive, 2 false negative, and 7 true positive values.

4.6. Hybrid Model Logi- HBC. It was developed by merging the logistic regression model with the histogram gradient boosting classifier in order to improve the accuracy of both models. As soon as the probability function of logistic regression is received by HGBC, it will be evaluated in order to identify whether or not a class has changed. Figure 19 shows the Logi-HGBC classification model's hybrid model performance with the accuracy of 91.37%.

Figure 20 shows the confusion matrix of the Logi-HGBC classification model with 19 true negative, 2 false positive, 9 false negative, and 3 true positive values.

4.7. Comparative Analysis. Below table shows the accuracy percentage of each model. Comparatively, Logi-XGB scored 95.7% accuracy, Logi-GBC with 90.8% accuracy. On the other hand, Logi-ABC scored 89.33% accuracy, while Logi-CBC scored the highest accuracy of 99.80%. Logi-LGBM and Logi-HGBC scored the same accuracy of 91.37%, respectively. Comparing with previous Logit-boosted algorithms implemented in previous studies, our proposed Logi-CBC has scored the highest accuracy on the given dataset. Comparative analysis of proposed models can be seen in Table 3.

Furthermore, we have compared our model with previous Logit-boosted algorithms used in previous state-of-art models as shown in Table 4.

5. Conclusions

Machine learning algorithms can be used to detect and prevent mobile users from engaging in false sensing activities. Our research focused on a real-world scenario in which fictional users are monitored by a mobile crowd sensing system as part of a demand response program to ensure that the total number of fake users does not surpass a predetermined limit, for the network to receive a discount on quality. Consequently, in order to maximize the mobile network, a coalition of users must work together. Distributed MCS architecture systems can validate obtained data using behavioral analysis based on participant reliability ratings provided by MCS's behavioral analysis solution. According to the results of the evaluation, our method has a positive impact on quality. Each classification's investigation was done (benign and harmful). In this study, we provide a deep neural network-based anomaly detection system for the IOTA network architecture, which effectively learns valuable complicated patterns from IOTA network flows in order to classify data as either normal or abnormal. IoT has made mobile crowdsourcing systems (MCS) a must-have for any business. Some of the examples of how the Internet of Things Auto (IOTA) has grown rapidly over the past decade are shown in this list. To prohibit mobile users from engaging in false sensing activities, IOTA-based MCS (iMCS) technology is being developed, and it will leverage machine learning. For the first time, our method has been evaluated and proved to be effective in both quality estimation and incentive allocation in a distributed system with the MCS design. To achieve a 99.8% accuracy rate on the IOTA Bottleneck dataset, Logi-CBC resorted to deep learning techniques. In terms of accuracy, Logi-XGB scored 95.7 percent, while Logi-GBC scored 90.8 percent. As a result of this, Logi-ABC had an accuracy rate of 89%. Logi-CBC, on the other hand, got the highest accuracy of 99.8%. Logi-LGBM and Logi-HGBC both scored 91.37 percent accuracy, which is identical. On the given dataset, our Logi-CBC algorithm outperforms earlier Logit-boosted algorithms in terms of accuracy. Using the new IoTA-Botnet 2020 dataset, a new proposed methodology is tested. In comparison to prior Logit-boosted algorithms, the new model had a detection accuracy of 99.8%, according to the research. As an additional benefit, using only the top five category features (in terms of % accuracy, recall, and F 1) enhances detection precision even further.

Data Availability

Dear Sir/Madam, as you are concerned with the availability of my code and practical work. It is confidential to my lab and cannot be shared with anyone until it gets published. And secondly, I will publish my code and lab work according to the instructions of my supervisor (lab name: supervisor name: Regards Mazhar Hameed).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

I would like to acknowledge my indebtedness and render my warmest thanks to my supervisor, Professor Yang Fengbao, who made this work possible. His friendly guidance and expert advice have been invaluable throughout all stages of the work. I would also wish to express my gratitude to Miss Gao Min for extended discussions and valuable suggestions which have contributed greatly to the improvement of the paper. This work was supported by the National Natural Science Foundation of China (Grant Nos. 61672472 and 61972363) and Science Foundation of North University of China, Postgraduate Science and Technology Projects of North University of China (Grant No. 20181530).

References

[1] R. Wang, K. Nie, T. Wang, Y. Yang, and B. Long, "Deep learning for anomaly detection," in *Proceedings of the 13th international conference on web search and data mining*, pp. 3569-3570, Jan 2020.

[2] K. Singh, S. Rajora, D. K. Vishwakarma, G. Tripathi, S. Kumar, and G. S. Walia, "Crowd anomaly detection using aggregation of ensembles of fine-tuned ConvNets," *Neurocomputing*, vol. 371, pp. 188–198, 2020.

[3] Y. Hao, Z. J. Xu, Y. Liu, J. Wang, and J. L. Fan, "Effective crowd anomaly detection through spatio-temporal texture analysis," *International Journal of Automation and Computing*, vol. 16, no. 1, pp. 27–39, 2019.

[4] I. Alrashdi, A. Alqazzaz, E. Aloufi, R. Alharthi, M. Zohdy, and H. Ming, "AD-IoT: anomaly detection of IoT cyberattacks in smart city using machine learning," in *2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)*, pp. 305–310, Jan 2019.

[5] M. Shafiq, Z. Tian, A. K. Bashir, X. Du, and M. Guizani, "CorrAUC: a malicious bot-IoT traffic detection method in IoT network using machine-learning techniques," *IEEE Internet of Things Journal*, vol. 8, no. 5, pp. 3242–3254, 2021.

[6] M. Shafiq, Z. Tian, Y. Sun, X. Du, and M. Guizani, "Selection of effective machine learning algorithm and Bot-IoT attacks traffic identification for internet of things in smart city," *Future Generation Computer Systems*, vol. 107, pp. 433–442, 2020.

[7] J. Huang, L. Kong, H. N. Dai et al., "Blockchain-based mobile crowd sensing in industrial systems," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 10, pp. 6553–6563, 2020.

[8] B. A. Ng and S. Selvakumar, "Anomaly detection framework for Internet of things traffic using vector convolutional deep learning approach in fog environment," *Future Generation Computer Systems*, vol. 113, pp. 255–265, 2020.

[9] L. Kuang, P. Shi, C. Hua, B. Chen, and H. Zhu, "An enhanced extreme learning machine for dissolved oxygen prediction in wireless sensor networks," *IEEE Access*, vol. 8, pp. 198730–198739, 2020.

[10] G. L. Santos, P. T. Endo, D. Sadok, and J. Kelner, "When 5G meets deep learning: a systematic review," *Algorithms*, vol. 13, no. 9, pp. 208–234, 2020.

[11] M. Gupta, M. Abdelsalam, S. Khorsandroo, and S. Mittal, "Security and privacy in smart farming: challenges and opportunities," *IEEE Access*, vol. 8, pp. 34564–34584, 2020.

[12] M. Stoyanova, Y. Nikoloudakis, S. Panagiotakis, E. Pallis, and E. K. Markakis, "A survey on the internet of things (IoT) forensics: challenges, approaches, and open issues," *IEEE Communication Surveys and Tutorials*, vol. 22, no. 2, pp. 1191–1221, 2020.

[13] A. Zielonka, M. Wozniak, S. Garg, G. Kaddoum, M. J. Piran, and G. Muhammad, "Smart homes: how much will they support us? A research on recent trends and advances," *IEEE Access*, vol. 9, pp. 26388–26419, 2021.

[14] M. G. Al Zamil, M. Rawashdeh, S. Samarah, M. S. Hossain, A. Alnusair, and S. M. M. Rahman, "An annotation technique for in-home smart monitoring environments," *IEEE Access*, vol. 6, pp. 1471–1479, 2017.

[15] M. Taneja, N. Jalodia, P. Malone, J. Byabazaire, A. Davy, and C. Olariu, "Connected cows: utilizing fog and cloud analytics toward data-driven decisions for smart dairy farming," *IEEE Internet of Things Magazine*, vol. 2, no. 4, pp. 32–37, 2019.

[16] W. H. Hassan, "Current research on Internet of Things (IoT) security: a survey," *Computer Networks*, vol. 148, pp. 283–294, 2019.

[17] S. Jovanović, M. Jovanović, T. Škorić et al., "A mobile crowd sensing application for hypertensive patients," *Sensors*, vol. 19, no. 2, pp. 1–16, 2019.

[18] R. Chalapathy and S. Chawla, "Deep learning for anomaly detection: a survey," pp. 1–50, 2019, http://arxiv.org/abs/1901.03407.

[19] P. He, G. Zhang, X. Liao et al., "Sodium ion stabilized vanadium oxide nanowire cathode for high-performance zinc-ion batteries," *Advanced Energy Materials*, vol. 8, no. 10, pp. 1–6, 2018.

[20] S. Hu, M. Arellano, P. Boontheung et al., "Salivary proteomics for oral cancer biomarker discovery," *Clinical Cancer Research*, vol. 14, no. 19, pp. 6246–6252, 2008.

[21] L. Zhao, X. Chen, X. Wang et al., "One-step solvothermal synthesis of a carbon@TiO_2 dyade structure effectively promoting visible-light photocatalysis," *Advanced Materials*, vol. 22, no. 30, pp. 3317–3321, 2010.

[22] S. M. Reddy, V. Kakulapati, and A. Prince, "Advance security: anomaly detection in mobile crowd sensing using machine learning techniques," 2021.

[23] Y. Liu, Z. Ma, X. Liu, S. Ma, S. Nepal, and R. Deng, "Boosting privately: privacy-preserving federated extreme boosting for mobile crowdsensing," pp. 1–11, 2019, http://arxiv.org/abs/1907.10218.

[24] C. Feng, Y. Tian, X. Gong, X. Que, and W. Wang, "MCS-RF: mobile crowdsensing–based air quality estimation with random forest," *International Journal of Distributed Sensor Networks*, vol. 14, no. 10, 2018.

[25] M. S. Akhtar and T. Feng, "EAI endorsed transactions IOTA based anomaly detection machine learning in mobile sensing," pp. 1–10, 2020.

[26] S. Yang, F. Wu, S. Tang, X. Gao, B. Yang, and G. Chen, "On designing data quality-aware truth estimation and surplus sharing method for mobile crowdsensing," *IEEE Journal on*

Selected Areas in Communications, vol. 35, no. 4, pp. 832–847, 2017.

[27] M. Arafeh, M. El Barachi, A. Mourad, and F. Belqasmi, "A blockchain based architecture for the detection of fake sensing in mobile crowdsensing," in *2019 4th International Conference on Smart and Sustainable Technologies (SpliTech)*, Split, Croatia, June 2019.

[28] K. Kucuk, C. Bayilmis, A. F. Sonmez, and S. Kacar, "Crowd sensing aware disaster framework design with IoT technologies," *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, no. 4, pp. 1709–1725, 2020.

[29] P. Mrazovic, *Crowdsensing-Driven Route Optimisation Algorithms for Smart Urban Mobility*, Universitat Politècnica de Catalunya, Spain, 2018.

[30] K. Haseeb, N. Islam, Y. Javed, and U. Tariq, "A lightweight secure and energy-efficient fog-based routing protocol for constraint sensors network," *Energies*, vol. 14, pp. 1–14, 2021.

[31] S. Kianoush, S. Savazzi, and M. Nicoli, "Device-free crowd sensing in dense WiFi MIMO networks: Channel features and machine learning tools," in *2018 15th Workshop on Positioning, Navigation and Communications (WPNC)*, pp. 1–6, Bremen, Germany, Oct 2018.

[32] N. P. Owoh and M. M. Singh, "SenseCrypt: a security framework for mobile crowd sensing applications," *Sensors*, vol. 20, no. 11, pp. 1–23, 2020.

[33] A. M. Ali Al-Muqarm and F. Rabee, "IoT technologies for mobile crowd sensing in smart cities," *Journal of Communication*, vol. 14, no. 8, pp. 745–757, 2019.

[34] Z. Zhou, H. Liao, B. Gu, K. M. S. Huq, S. Mumtaz, and J. Rodriguez, "Robust mobile crowd sensing: when deep learning meets edge computing," *IEEE Network*, vol. 32, no. 4, pp. 54–60, 2018.

[35] M. Louta, K. Mpanti, G. Karetsos, and T. Lagkas, "Mobile crowd sensing architectural frameworks: a comprehensive survey," in *2016 7th International Conference on Information, Intelligence, Systems & Applications (IISA)*, Chalkidiki, Greece, July 2016.

[36] Z. Sisi and A. Souri, "Blockchain technology for energy-aware mobile crowd sensing approaches in Internet of Things," *Transactions on Emerging Telecommunications Technologies*, vol. 6, pp. 1–18, 2021.

[37] A. Khormali, J. Park, H. Alasmary, A. Anwar, M. Saad, and D. Mohaisen, "Domain name system security and privacy: a contemporary survey," *Computer Networks*, vol. 185, article 107699, 2021.

[38] G. Spanos, K. M. Giannoutakis, K. Votis, and D. Tzovaras, "Combining statistical and machine learning techniques in IoT anomaly detection for smart homes," in *2019 IEEE 24th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, pp. 1–6, Limassol, Cyprus, Sept 2019.

[39] N. Awan, S. Khan, M. Khalid Imam Rahmani et al., "Machine learning-enabled power scheduling in IoT-based smart cities," *Computer Materials and Continua*, vol. 67, no. 2, pp. 2449–2462, 2021.

[40] M. A. Hoque and C. Davidson, "Design and implementation of an IoT-based smart home security system," *The International Journal of Networked and Distributed Computing*, vol. 7, no. 2, pp. 85–92, 2019.