CorrAUC: A Malicious Bot-IoT Traffic Detection Method in IoT Network Using Machine-Learning Techniques

Muhammad Shafiq, Zhihong Tian, *Member, IEEE*,
Ali Kashif Bashir, *Senior Member, IEEE*, Xiaojiang Du, *Fellow, IEEE*, and Mohsen Guizani, *Fellow, IEEE*

***Abstract*—Identification of anomaly and malicious traffic in the Internet-of-Things (IoT) network is essential for the IoT security to keep eyes and block unwanted traffic flows in the IoT network. For this purpose, numerous machine-learning (ML) technique models are presented by many researchers to block malicious traffic flows in the IoT network. However, due to the inappropriate feature selection, several ML models prone misclassify mostly malicious traffic flows. Nevertheless, the significant problem still needs to be studied more in-depth that is how to select effective features for accurate malicious traffic detection in the IoT network. To address the problem, a new framework model is proposed. First, a novel feature selection metric approach named CorrAUC is proposed, and then based on CorrAUC, a new feature selection algorithm named CorrAUC is developed and designed, which is based on the wrapper technique to filter the features accurately and select effective features for the selected ML algorithm by using the area under the curve (AUC) metric. Then, we applied the integrated TOPSIS and Shannon entropy based on a bijective soft set to validate selected features for malicious traffic identification in the IoT network. We evaluate our proposed approach by using the Bot-IoT data set and four different ML algorithms. The experimental results analysis showed that our proposed method is efficient and can achieve >96% results on average.**

***Index Terms*—Attacks, detection, identification, Internet of Things (IoT), intrusion, machine learning (ML), malicious.**

Manuscript received April 19, 2020; revised May 21, 2020; accepted June 10, 2020. Date of publication June 15, 2020; date of current version February 19, 2021. This work was supported in part by the National Key research and Development Plan under Grant 2018YFB0803504, in part by the Guangdong Province Key Research and Development Plan under Grant 2019B010137004, in part by the National Natural Science Foundation of China under Grant 61871140, and in part by the Guangdong Province Universities and Colleges Pearl River Scholar Funded Scheme (2019). *(Corresponding author: Zhihong Tian.)*

Muhammad Shafiq and Zhihong Tian are with the Department of Cyberspace Institute of Advanced Technology, Guangzhou University, Guangzhou 510006, China (e-mail: srsshafiq@gmail.com; tianzhihong@gzhu.edu.cn).

Ali Kashif Bashir is with the Department of Computing and Mathematics, Manchester Metropolitan University, Manchester M15 6BH, U.K.

Xiaojiang Du is with the Department of Computer and Information Sciences, Temple University, Philadelphia, PA 19122 USA.

Mohsen Guizani is with the Department of Computer Science and Engineering, Qatar University, Doha, Qatar.

Digital Object Identifier 10.1109/JIOT.2020.3002255

I. Introduction

NOWADAYS, the Internet-of-Things (IoT) technology is growing up more day by day [1], and in every minute, numerous devices are getting connected with this technology. By using this technology, daily life becomes more convenient and well organized. For instance, initially, the IoT technology was limited to small offices and homes, but nowadays, the IoT technology integrated into industries for more reliability and saving time. However, the IoT technology is becoming an essential part of our daily life. In 2021, the IoT technology will grow up, and more than 27 million IoT devices will connect, which will be a tremendous change in the IoT technology world [2]. Though the IoT technology is growing day by day, but on the other hand, the cyberattacks are also becoming challenging and increasing. For this purpose, numerous researchers in the IoT technology field proposed several different cybersecurity systems and widely applied the proposed cybersecurity system to protect their information from cyberattacks and unauthorized access. Recently, IoT security becomes a hot topic and gained much attention in IoT cybersecurity. To overcome the problem of IoT cyberattacks, researchers try their best and proposed numerous cybersecurity systems. Similarly, numerous cybersecurity systems in IoT networks are presented and utilized for the protection of critical information and secure from unauthorized access in the IoT network. For example, in 2017, the IoT attacks such as Distributed Denial of Service (DDoS) become very spread and grow up to 172%, which gain much interest in the IoT network [2].

According to the Kaspersky Laboratory report in 2019 [3], the malware attacks in the IoT network environment increased in 2017 as compared in 2013 malware traffic attacks in the IoT network. However, in these numerous attacks, most attacks are very harmful attacks, such as Botnet attacks, etc. [4]. For the intrusion detection (ID), the first ID system (IDS) was discussed and introduced by Anderson [5]. Denning [6] proposed a new model for the detection of intrusion based on real-time ID. Their proposed intrusion-detection expert system has the capability to detect break-ins, penetrations, Trojan horses, and as well as other computer-related intrusions that lead to the damage of the computer system, etc. However, their proposed model was based on the hypothesis, mean any security violation can be detected by using the monitoring audit records and discussed that it is possible to detect the abnormal attacks or operation in a network by using the user behavior. Nowadays, in the IoT, the most dangerous and challenging widespread hazardous threats are man-in-the-middle (MITM) dangerous threats with DDoS [7]–[9]. However, numerous researchers in the research community tried their best to find out and proposed an effective system to overcome these widespread hazardous threats in the IoT network environment.

Recently, Alharbi *et al.* [10] proposed a new system for the detection of malware cyberattacks and to protect IoT against from cyberattacks named fog-computing-based security (FOCUS) system. Their proposed ID system [11] used the virtual private network (VPN) for secure communication between the IoT devices, and the system was able to use challenge-response authentication for the protection of the VPN server and keep protect from hazardous DDoS attacks in IoT. However, their proposed system is effective from the protection of potential cyberattacks and able to secure the IoT system. Furthermore, they implemented the proposed system in fog computing and got effective results. They showed that their proposed system is effective for the detection of malicious cyberattack with short response time and bandwidth. However, for the best results and accurate identification, machine learning (ML) and artificial intelligence (AI) are effective and widely applied techniques. ML and AL methods are widely applied for the detection of cyberattacks in the IoT network environment [12]–[14]. IoT traffic identification is vital for IoT security monitoring and IoT traffic management. Recently, the ML technique gains much importance and becomes very popular in numerous fields because of its accurate results.

For effective identification, the significant feature set is very important for the ML model. Effective features indicate accurate feature or attributes which keep significance information for the ML technique, and these effective feature sets include on the training and testing sets. It is impossible to evaluate the ML model without the training set and testing set. Thus, the useful feature set of training and testing sets is compulsory for the evaluation of the ML model. The ML technique is widely applied in computer science, especially in network traffic identification [15], [16]. ML methods are very useful for identifying or classifying malicious, intrusion, and cyberattacks in IoT networks. Though applying the ML method for the detection for the classification of malicious traffic of cyberattacks is effective, but as compared to other computing tools, the ML technique tool is very complex. Though using an ML method is very effective in the area of identification or classification. Still, it is also some disadvantages in the IoT malicious and ID like computation time and energy consumption problems. Currently, these two problems are a hot topic in the IoT field by using ML methods, and numerous researchers try their best to overcome these crop-up problems. To overcome the above-mentioned problems, a detection ML model should be accurate for input data sets for better performance results. It is possible to get high-performance results and apply the ML technique accurately for the detection of cyberattacks in the IoT network environment.

For the significant identification performance results, the input data set keeps an important role by using the ML technique [17]. Therefore, for the accurate and effective detection of anomalies and intrusion in IoT by using ML techniques, it is essential to select an effective input feature set and remove the unwanted feature, which does not give accurate identification information. For this purpose, feature selection methods give sufficient identification information and can remove unwanted features from the given feature data set. Thus, it is important to focus on the effective feature selection and select the essential feature set for ML detection in IoT for accurate detection of anomaly and intrusion problems. Similarly, to overcome the problem of effective feature selection, Zhang *et al.* [18] introduced two new techniques and proposed two different algorithms. Their proposed methods are able to select the effective feature set from an imbalance high-dimensional data set. For the evaluation of their experimental results, they applied three different ML algorithms by using the trace traffic completely different network environment. However, they showed in their study that their proposed methods are effective for the feature selection in high-dimensional data sets, especially for imbalanced data sets. Similarly, Koroniotis *et al.* [19] introduced a new data set named Bot-IoT for the identification of cyberattacks in the IoT network. In their study, they focused on malicious attacks in IoT networks. The developed data set includes different types of hazards attacks, especially botnets cyberattacks. More in-depth, the data set is developed in a realistic testbed with a defined feature set, which consists of normal traffic and cyberattacks traffic flow. For the experimental analysis, statistical analysis is performed to find out which feature carries accurate information for the detection of cyberattacks in the IoT network. However, they selected the ten best feature sets from the extracted feature set. They then used a well-known ML classifier for the performance analysis of the selected feature set. More in-depth for the performance analysis to find out which feature gives the most effective results, four different types of metrics are used.

In our previous study [15], [20]–[26], feature selection problems are studied and selected robust features by proposing different types of approaches for instant messages (IMs) traffic identification and attacks traffic detection. Similarly, in [15] and [16], to overcome the problem of the feature selection problem, different feature selection techniques are proposed for the accurate network traffic classification using ML algorithms. However, from the above study, we concluded that selecting more feature set is not efficient for the accurate identification by using ML techniques and showed that selecting more than 50 feature set can lower the ML classifier accuracy and can increase the computational complexity. However, in the IoT network cyberattacks traffic identification, no effective ML model is proposed yet. Therefore, it is important to study the effective feature selection problem for anomaly and malicious traffic in the IoT network and introduce a new technique that overcomes this problem.

In this article, a new effective feature selection technique is proposed for the problem of effective feature selection for cyberattacks in IoT network traffic by using the Bot-IoT data set and to improve the performance of ML techniques. However, the main contributions in this article are as follows.

1) In order to deal with the effective feature selection problem in IoT cyberattacks identification in the IoT network. First, a novel feature selection metric approach named CorrAUC is proposed to deal with the issue of the effective feature selection for cyberattacks identification in IoT networks. However, it is the first time to put forward combine the correlation attribute evaluation (CAE) metric and specific ML area under the curve (AUC) results for the effective feature selection in IoT Bot-IoT attack detection.
2) Then, based on CorrAUC, a new feature selection algorithm named CorrAUC is developed and designed for the problem of feature selection for malicious Bot-IoT traffic identification in the IoT network, which is based on the wrapper technique to filter the feature set accurately and select the feature set that carries enough information for the selected ML algorithm by using the AUC metric for the detection of Bot-IoT cyberattacks in the IoT network environment. However, the proposed algorithm includes two steps metrics for the optimum feature selection. CAE metric and a specific used the ML algorithm AUC metric.
3) Afterward, we applied integrated TOPSIS and Shannon entropy based on the bijective soft set method for the validation of selected features for malicious traffic identification in the IoT network. It is based on the selection of proper attributes mean feature set for better detection of vicious attacks in the IoT network. Furthermore, we compare the results of TOPSIS and Shannon entropy based on the bijective soft set with results achieved by the proposed approach.
4) Then, we concluded and put forward the optimum selected feature set selected by our proposed technique that carries enough information for the detection of malicious Bot-IoT traffic in the IoT network. The experimental results showed that five optimum feature sets carry enough information and have discriminative power for the detection of malicious attacks in the IoT network by using ML.

The remainder of this article is arranged as follows. Section II includes related works. In Section III, we demonstrate the proposed techniques. While in Section IV, we explain in detail the methodology, experimental work, and applied data set. Similarly, Section V includes analysis and discussion. Finally, Section VI concludes this article conclusion and future directions.

II. Related Works

From the last decade, security and trust problems become a scorching topic, and many researcher endeavors are hard to overcome this problem and proposed numerous effective models along with the future Internet [27], [28], IoV [29], wireless sensor network (WSN) [30], [31], and IoT. However, some most viewed and cited studies related to feature selection for malicious Bot-IoT in IoT networks are discussed in this section. In our recent study work [16], for the optimum feature selection problem in IM applications traffic classification, a feature selection technique is proposed based on the mutual information (MI) analysis technique. However, from the experimental results analysis, the proposed approach achieves beneficial performance results by using the selected feature set for the IM application traffic identification. More in-depth, the study only limited to feature selection for several different applications and as well as to minimize the applied ML algorithms computational complexity. The proposed approach is able to apply on an imbalance or high-dimensional data set. The experimental result showed that the proposed approach could achieve auspicious performance results for the identification of IM application traffic classification. The technique of feature selection is handy for enhancing the ML performance. However, the feature selection is a process to select the optimum feature set from several feature set and removed the features that do not carry enough identification information for the identification or removing the redundant feature. Egea *et al.* [32] studied mostly cited research studies related to the feature selection technique, especially the correlation coefficient technique, and proposed a new feature selection technique named fast-based correlation features (FCBFs) algorithm for the improvement of the performance of the IoT network in the industrial environment. The main contribution of their study is to split the feature space into several equal parts with equal size. Using the proposed approach, they showed enhanced results of correlation ML of every running node in the IoT network. They showed that their experimental results are effective, and the proposed approach is able to achieve effective performance results in terms of accuracy and execution time, which is very important for accurate identification. Similarly, Meidan *et al.* [33] studied the detection of attacks in the IoT network and proposed a new technique to overcome the problem of attacks which is initiated by the IoT devices and then for the identification of anomalies in IoT traffic, they used autoencoder. The data set that they used in their study for the evaluation of their proposed approach is botnet attacks, Bashlite and Mirai based on the IoT. However, the utilized data sets are also included on several infected devices in the IoT network. They showed in their study that the proposed approach is able to detect cyberattacks in IoT network devices with high-performance results.

Similarly, for the IoT devices, performance improvement, and detection of an anomaly, Su *et al.* [34] studied the most cited feature selection technique and introduced a feature selection method. For their study, they initially group the IoT sensors as a group for the identification of deployed sensors. After that, for anomaly detection, they control the correlation variation of data for the selection of sensors. More in-depth, for the clustering of sensors, they utilized the curve alignment technique, and for the data, the calculation window size is discussed. Then, the multicluster attributes selection (MCFS) method is conducted for the selection of features. In their experimental analysis, they showed that their proposed technique is effective for IoT performance enhancement and anomaly detection in IoT networks. More in-depth, numerous IoT security technique can be applied for the accurate cybersecurity purpose in the IoT security environment, for

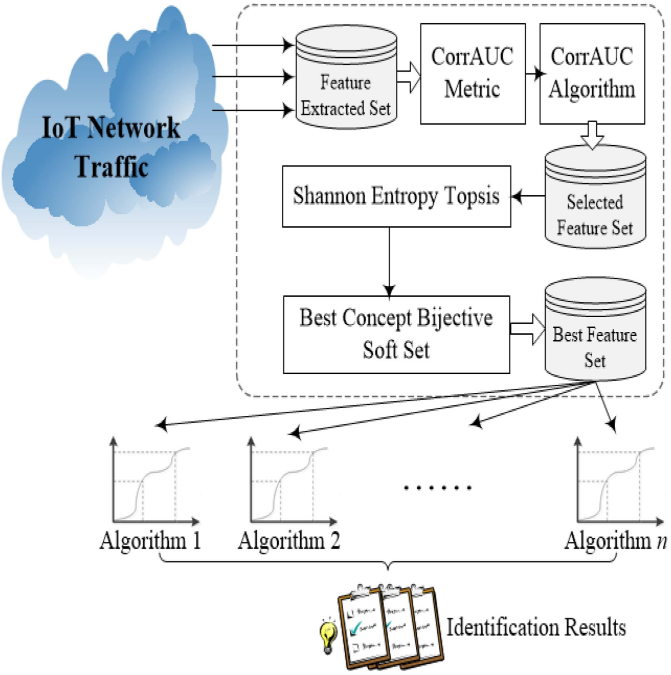


Fig. 1. Proposed framework for feature selection.

instances, cyberattacks identification in [35] and [36], effective management scheme [37], [38], evidence framework, etc. However, the above numerous techniques proposed by many researchers are effective, but it is important to select the most effective feature set that carries accurate information for the Bot-IoT attack detection in the IoT environment. The necessary key process of the feature selection technique includes on different important steps such as trace traffic, to trace the original traffic and subset generation, to generate a feature set from the trace traffic and subset evaluation, and to evaluate the generated feature set for the next phase, the decision maker takes some decision for the effective feature selection and then in the subset, evaluation gives the final decision and validates the feature set [39]–[41].

III. Proposed Method

In this section, we explain the proposed technique with details step-by-step process. For the effective selection in the IoT network, our proposed method includes four steps, as shown in Fig. 1. First, a novel feature selection metric approach named CorrAUC is proposed and applied, which selects features that carry enough information and then based on CorrAUC, a new feature selection algorithm name CorrAUC is developed and designed, which is based on the wrapper technique to filter the feature accurately and select effective features for the selected ML algorithm by using the AUC metric and Bot-IoT data set. The proposed algorithm consists of CAE and combines with the AUC metric to overcome the problem of the effective feature selection for Bot-IoT detection by using a specific ML algorithm. Then, we applied integrated TOPSIS and Shannon entropy based on a bijective soft set for the validation of selected features for Bot-IoT attacks traffic identification in the IoT network. More in-depth, the bijective soft set is a mathematical technique used for the selection in different areas. This technique produces very effective results in terms of effective feature selection for Bot-IoT attack detection in the IoT network environment. To the best of our knowledge, in this study, Corr and AUC are combined and conducted for the first time for the identification of the Bot-IoT attack in the IoT network using ML algorithms. Moreover, our proposed method selects the feature set that carries enough identification information for the Bot-IoT attacks in the IoT network. For a clear understanding, the details of methodologies are discussed in the next section for the effective feature selection in the IoT network, considering Bot-IoT malicious attacks detection.

A. Feature Selection Metrics

In this section, the conducted feature selection metrics are discussed in detail. First, the correlation-based metric is presented and then AUC metrics. However, the details are given in the following section.

1) Correlation-Based Metric: To overcome the problem of the effective feature selection, for BoT-IoT malicious attack detection in the IoT network, the Pearson moment correlation technique is adopted. This technique is used to study more in-depth and identify the relationship between independent and target class features. Fancher [42] proposed the basic idea of Pearson moment correlation in 1880s. Similarly, after 16 years in 1896, Pearson make changes in Pearson moment correlation and named it Pearson product-moment correlation. This technique is utilized for the identification of relationships among different features or attributes. However, the modified technique is based on statistical analysis operations. For the correlation coefficient, the following given formula can use. For the case of two different M and N attributes, the following given formula can use to find out the Pearson correlation coefficient between M and N attributes:

$$C_{X,Y} = \frac{\text{Covariance}(A,B)}{\sigma_x \sigma_y}. \tag{1}$$

In (1), the correlation coefficient is $C_{A,B}$, and (A, B) indicates the covariance. Similarly, ab is the standard deviation for A and B attributes $\sigma_A \sigma_B$. More in-depth, for the two sets of feature (2) can be used to calculate the correlation coefficient

$$C = \frac{\sum_{i=1}^{n}(a_i - \overline{a})(b_i - \overline{b})}{\sqrt{\sum_{i=1}^{n}(a_i - \overline{a})^2}\sqrt{\sum_{i=1}^{n}(b_i - \overline{b})^2}}. \tag{2}$$

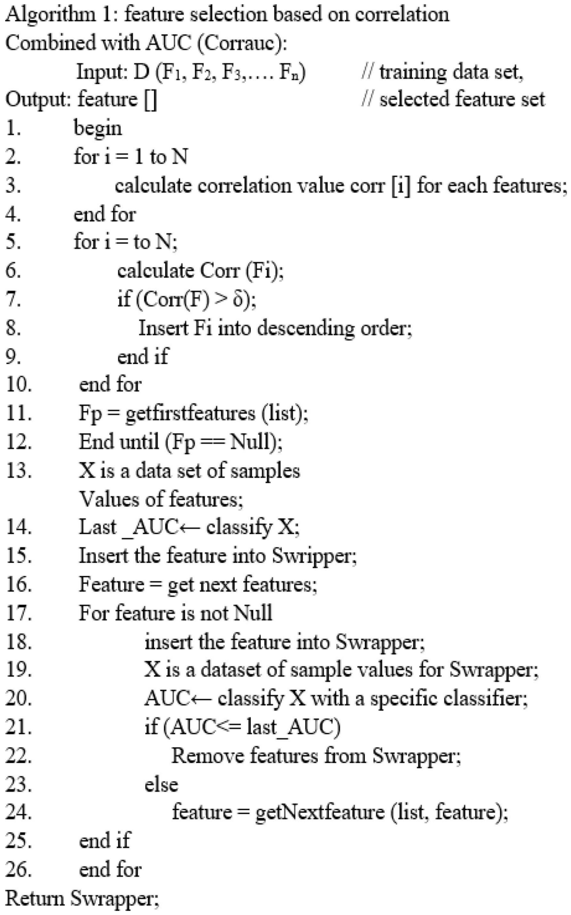
For instance, two set of features A and B with respective to its features can be indicated as $a_1, a_2, a_3, \ldots, a_n$ and B can be $b_1, b_2, \ldots, b_n$. Similarly, n indicated the number of instances of size. Where a_i and b_i are the values of data. While a bar and b bar are the mean values in (2), similarly, if the values of the C coefficient is reached to plus one $+1$ and minus one -1. It means that if the coefficient values are plus one, then it means the relationship between the features is powerful, and zero means there is no relationship between features. In contrast, if the coefficient values are minus, one means the relationship between features is very weak. The Pearson correlation technique is very effective for ranking and accurate feature selection. Therefore, to overcome the problem of effective and robust feature selection for the Bot-IoT malicious attacks detection in the IoT network, the CAE technique is adopted and applied to rank the effectiveness of features in several given feature set. The basic concept of using this ranking attribute is to find out the significance of the feature set of a data set by using the correlation between features. Nevertheless, for the Bot-IoT malicious attack detection in the IoT network using ML, a feature will be effective if the relationship between feature and class is strong, not correlated to feature. Similarly, in this way, feature effectiveness can be calculated and analyzed for accurate detection as follows:

$$\mathrm{Corr} = \frac{k\mathrm{avg}\left(\mathrm{corr}_{fc}\right)}{\sqrt{k + k(k-1)\mathrm{avg}\left(\mathrm{corr}_{ff}\right)}}. \tag{3}$$

In the above equation, Corr indicates the correlation between features and $k\mathrm{avgcorr}_{f}c$ indicates to find the average of correlation between features and its class. Similarly, $\mathrm{Avg}(\mathrm{corr}_{f}f)$ indicates the average correlation between features and while k is the number of features. However, applying the above given equation for the identification correlation relationship between attributes, the main factors are: if the correlation between the feature set is strong then it indicates that the correlation between feature set and features class is weak. If the correlation between the feature set and reliant class strong it indicate the strong correlation among the feature set and class while if there is more attributes, then it indicates a strong correlation between features and reliant class.

2) Area Under the Curve-Based Metric: After using the Corr metric, it is essential to find out the most robust features which carry accurate information for the Bot-IoT attack detection in the IoT network. Considering this case, a technique named wrapper is applied based on the AUC metric [24]. Though, for the classification of network traffic by using an ML technique, the accuracy metric is the most optimal. But, here we are interested in finding the most significant feature set for the detection of Bot-IoT attacks in the IoT network environment. Thus, the AUC metric is a significant metric for the detection of malicious attacks in IoT networks and a beneficial metric to rank features in several feature sets. However, applying the AUC metric in this research study in two different facts are as follows. If the AUC metric values are strong and high, then the model will give effective performance results. If the AUC metric values are weak and not high enough, then the model will not provide effective performance results in terms of the detection of Bot-IoT attacks in the IoT network. More in-depth, the AUC metric is also very effective for performance evaluation and ranking features. Thus, we applied in this research study the AUC metric to rank effective features and choose those attributes that carry enough information and have strong high metric values for the detection of malicious Bot-IoT attacks in the IoT network.

3) Proposed Algorithm: In this section, the proposed algorithms named CorrAUC are described with details step by step, as shown in Fig. 2. The proposed CorrAUC includes two steps. In the first step, the algorithm uses a correlation technique to filter the feature set and find out the correlation between features and class. Then, the algorithm goes to the next step to filter the features with high AUC metric values by using a specific ML algorithm. Similarly, the proposed algorithm selects the useful features which carry enough information for Bot-IoT detection in the IoT network. However, the details of step-by-step phases are discussed as follows. As discussed in the above lines, the proposed CorrAUC algorithm, a hybrid feature selection algorithm based on the correlation technique and AUC metric, used to select feature which has enough information for detection of Bot-IoT attacks in the IoT network. However, the proposed algorithm first goes in to calculate the correlation between features and select features that are a high correlation relationship. More in-depth, the algorithm will first calculate the correlation among features and placed in the ascending order with respective correlation values. Then, the algorithm compared the correlation among each feature. Afterward, a threshold value is assigned, if feature correlation values are higher than the specified threshold assign value, mean the feature is effective and puts forward in the descending order. In more detail, the higher the threshold value, the higher the proposed model speed, but it is not effective for the ML algorithm, because high threshold values decrease the identification and performance of ML algorithms [43]. Then, after calculating the correlation and filtering with threshold values, the proposed algorithm filters



```
Algorithm 1: feature selection based on correlation
Combined with AUC (Corrauc):
        Input: D ($F_1$, $F_2$, $F_3$,…. $F_n$)        // training data set,
Output: feature []                                // selected feature set
        begin
        for i = 1 to N
            calculate correlation value corr [i] for each features;
        end for
        for i = to N;
            calculate Corr (Fi);
            if (Corr(F) > δ);
               Insert Fi into descending order;
            end if
        end for
        Fp = getfirstfeatures (list);
        End until (Fp == Null);
        X is a data set of samples
        Values of features;
        Last _AUC← classify X;
        Insert the feature into Swripper;
        Feature = get next features;
        For feature is not Null
                insert the feature into Swrapper;
                X is a dataset of sample values for Swrapper;
                AUC← classify X with a specific classifier;
                if (AUC<= last_AUC)
                   Remove features from Swrapper;
                else
                   feature = getNextfeature (list, feature);
        end if
        end for
Return Swrapper;
```

Fig. 2. Proposed CorrAUC algorithm.

each feature by using the AUC metric of a specific ML algorithm. However, the algorithm filters each feature one by one by using the AUC metric and selects those features which give high AUC metric values for the detection of Bot-IoT attacks in the IoT network as well as if the AUC values of the feature are low then the algorithm will remove from the list and the algorithm will go to the next step forward to the swapper.

B. Shannon Entropy TOPSIS

For the effective feature selection, Shannon entropy TOPSIS, based on the bijective soft set technique, is applied to detect Bot-IoT attacks in IoT network environments. For better understanding, first, motivations are discussed, and then preliminary definitions for effective feature selection and its mathematical operation. Nowadays, decision making is becoming the most challenging problem in the field of operational research and numerous researchers' endeavor hard to overcome the problem of the decision-making problem and proposed effective decision-making models such as Molodtsov [44] proposed the soft set for the decision making and selection attributes from a multiple criteria attributes and then followed by Gong and proposed the bijective soft set [45]. Similarly, type-2 soft set [46] is proposed to overcome the problem of the decision-making problem. From the above literature study, it is evident that the soft set is a useful technique for the selection of effective attributes from several given attributes. However, to overcome the problem of the effective feature selection, the decision-making technique is applied after the proposed feature selection technique. It is important to verify the proposed feature selection technique. Therefore, we use the conceptual decision-making technique to select a robust feature set for Bot-IoT attack detection in the IoT network environment. Similarly, Tiwari *et al.* [47] applied the Shannon entropy weight technique, motivated by this study, we use the same method for the selection of effective features from numerous features.

1) *Introductory Definitions:* In this section, the introductory definition and basic operations of the soft set are discussed in detail.
   a) *Soft Set [48]:* If U is the universal set and S is its parameter, then U be $P(U)$ and X will be the subset of S, for example, $X \subset S$. At that point, pair (F, X) will be the soft set over U, and the function F will be $F : X \rightarrow P(U)$.
   b) *Bijective Soft Set:* If (F, S) is a soft set and U is the universal set and its parameter is S, respectively, then (F, S) is known as the bijective soft set if the below two given conditions are true.
      i) $\bigcup_{\beta \in s} F(\beta) = U$.
      ii) For two features

$$\beta_i, \beta_j, \beta_j \in S, \beta_i = \beta_j, F(\beta_i) \bigcap F\big(\beta_j =\big) \oslash .$$

2) *Method:* Input: the set of features of the data set. Output: desired selected effective feature set.
   a) Identify a feature set based on Bot-IoT attacks and normal traffic in the IoT network environment.
   b) The soft set will be developed from the identified set of features from each feature, which is the most effective and discard others. However, these function concepts are a theoretical concept that is effective for a better understanding.
   c) After the second step completion, feature set values are represented in the soft set and bijective soft set, respectively, for the decision making.
   d) Generate feature preference for the expert and then make a decision matrix as $\mathcal{E}PDM = [\rho_{ij}]_{a \times b}$, where $i = 1, \ldots, a$ and $j = 1, \ldots, b$; ρ_{ij}. M indicates the number of experts, while n indicated the numbers of features.
   e) In step e), the value of projection ($\mathfrak{p}v$), entropy ($\mathcal{E}nt$), divergence ($\mathcal{D}iv$), and weight ($\mathcal{W}gt$) of each data set feature $\mathcal{Y}_{ij}$ is calculated, respectively, [49]. $pv_{ij} = [p_{ij}/(\sum_{i=1}^{a} p_{ij})]$ and $\mathfrak{E}nt = -\kappa \sum_{i=1}^{a} pv_{ij} \ln(pv_{ij})$, where κ is a constant implied as, $\kappa = (\ln(a))^{-1}$, then $\mathfrak{D}iv = 1 - (\mathfrak{E}nt)$ and $\mathfrak{W}gt(\gamma_{ij}) = \sum_{\kappa=1}^{n} (\mathfrak{D}iv_j/\mathfrak{D}iv_\kappa)$.
   f) Taking the desired requirement from the network security expert $\mathcal{N}ER$ that may give informative feature selection suggestion.
   g) In this step, the Shannon entropy weight is calculated in the soft set form also calculated weight choice value $\mathcal{W}CV$ with respective feature as: $\mathcal{W}CV_{i\kappa} = \sum_j \mathfrak{D}iv_{ij}$, where $\mathfrak{D}iv_{ij} = \mathfrak{W}gt(\gamma_{ij}) \times q_{ij}$. Here, q_{ij} is selection concepts.
   h) In this step, the ideal (IS) and nonideal solution (NIS) as γ_i^* and $\breve{\gamma}_i$ are calculated for each network expert using TOPSIS as follows:

$$\gamma_i^* = \mathcal{M}ax(\mathcal{W}CV_{i\kappa}); \quad \breve{\gamma}_i = \mathcal{M}in(\mathcal{W}CV_{i\kappa}).$$

   i) Computer the separation measure $(\Delta_{i\kappa}^*, \breve{\Delta}_{i\kappa})$ from the IS and NIS using the n-dimensional Euclidean distance for each network expert by using the relation

$$\Delta_{i\kappa}^* = \left(\gamma_{ij} - \gamma_i^*\right)^2, \quad \breve{\Delta}_{i\kappa} = \left(\gamma_{ij} - \breve{\gamma}_i\right)^2.$$

   Then, the combined separation measure for each concept will be as $(\Delta_\kappa^*, \breve{\Delta}_\kappa)$; $(\Delta_\kappa^*, \breve{\Delta}_\kappa)$ as follows:

$$\Delta_\kappa^* = \sqrt{\sum_{i=1}^{i=m} \Delta_{i\kappa}^*}, \quad \breve{\Delta}_\kappa = \sqrt{\sum_{i=1}^{i=m} \breve{\Delta}_{i\kappa}}.$$

   j) Calculate the closeness of each feature. $\mathcal{F}\zeta_\kappa$ to IS as: $\zeta_\kappa^* = [\breve{\Delta}_\kappa/(\breve{\Delta}_\kappa + \Delta_\kappa^*)]$.

The most closer measure will be an effective feature.

C. Implementation

The Shannon entropy TOPSIS technique based on the soft set method can be applied effective feature selection problem as follows.

1) For the effective feature selection, Bot-IoT attack detection in IoT, five different features are described to develop a set of effective feature selection $\mathcal{EFS}$ attributes as:

$\mathcal{EFS} = [\mathcal{EFS}_1, \mathcal{EFS}_2, \mathcal{EFS}_3, \mathcal{EFS}_4, \mathcal{EFS}_5]$, where these selected attributes can be as:

$$\mathcal{EFS}_1 = \mathcal{M}ean, \quad \mathcal{EFS}_2 = \mathcal{S}tddev$$
$$\mathcal{EFS}_3 = \mathcal{A}r\_p\_DdtIp, \quad \mathcal{EFS}_4 = \mathcal{P}k\_Src\_IP$$
$$\mathcal{EFS}_5 = \mathcal{P}k\_Dst\_IP.$$

We give the following values to attributes with respect to the effective feature by ourself identification based on the above given metrics as we denoted by as:

$$\mathcal{EFS}_1 = \{\mathcal{Y}_{11}, \mathcal{Y}_{12}, \mathcal{Y}_{13}\} = \{Low, Medium, High\}$$
$$\mathcal{EFS}_2 = \{\mathcal{Y}_{21}, \mathcal{Y}_{22}, \mathcal{Y}_{23}\} = \{Poor, Good, V.Good\}$$
$$\mathcal{SFA}_3 = \{\mathcal{Y}_{31}, \mathcal{Y}_{32}, \mathcal{Y}_{33}\} = \{V.Good, Acceptable, Low\}$$
$$\mathcal{EFS}_4 = \{\mathcal{Y}_{41}, \mathcal{Y}_{42}, \mathcal{Y}_{43}\} = \{V.Good, Acceptable, Low\}$$
$$\mathcal{EFS}_5 = \{\mathcal{Y}_{51}, \mathcal{Y}_{52}\} = \{\text{Minimum, Maximum}\}.$$

2) In this step, the concept for the effective feature set is generated to form useful combination from $\mathcal{EFS}$ as per two given set as: $\bigcup = f\mathcal{C}_1 + f\mathcal{C}_2 + f\mathcal{C}_3 + f\mathcal{C}_4 + f\mathcal{C}_5$. Generated feature selection concept sets are given as

$$f\mathcal{C}_1 = \{\mathcal{Y}_{11}, \mathcal{Y}_{21}, \mathcal{Y}_{31}, \mathcal{Y}_{42}, \mathcal{Y}_{52}\}$$
$$f\mathcal{C}_2 = \{\mathcal{Y}_{11}, \mathcal{Y}_{23}, \mathcal{Y}_{33}, \mathcal{Y}_{43}, \mathcal{Y}_{52}\}$$
$$f\mathcal{C}_3 = \{\mathcal{Y}_{12}, \mathcal{Y}_{21}, \mathcal{Y}_{31}, \mathcal{Y}_{43}, \mathcal{Y}_{51}\}$$
$$f\mathcal{C}_4 = \{\mathcal{Y}_{13}, \mathcal{Y}_{21}, \mathcal{Y}_{32}, \mathcal{Y}_{42}, \mathcal{Y}_{51}\}$$
$$f\mathcal{C}_5 = \{\mathcal{Y}_{13}, \mathcal{Y}_{21}, \mathcal{Y}_{31}, \mathcal{Y}_{41}, \mathcal{Y}_{51}\}.$$

3) For more in-depth, we form a soft set through which we can present by using the selection concept the feature specification as given

$$(\mathcal{GG}_1, \mathcal{EFS}_1) = \{\mathcal{GG}_1(\mathcal{Y}_{11}), \mathcal{GG}_1(\mathcal{Y}_{12}), \mathcal{GG}_1(\mathcal{Y}_{13})\}$$
$$(\mathcal{GG}_2, \mathcal{EFS}_2) = \{\mathcal{GG}_2(\mathcal{Y}_{21}), \mathcal{GG}_2(\mathcal{Y}_{22}), \mathcal{GG}_2(\mathcal{Y}_{23})\}$$
$$(\mathcal{GG}_3, \mathcal{EFS}_3) = \{\mathcal{GG}_3(\mathcal{Y}_{31}), \mathcal{GG}_3(\mathcal{Y}_{32}), \mathcal{GG}_3(\mathcal{Y}_{33})\}$$
$$(\mathcal{GG}_4, \mathcal{EFS}_4) = \{\mathcal{GG}_4(\mathcal{Y}_{31}), \mathcal{GG}_4(\mathcal{Y}_{42}), \mathcal{GG}_4(\mathcal{Y}_{43})\}$$
$$(\mathcal{GG}_5, \mathcal{EFS}_5) = \{\mathcal{GG}_5(\mathcal{Y}_{41}), \mathcal{GG}_5(\mathcal{Y}_{32})\}.$$

Now, the bijective sot set can be further demonstrate as per 3) with details given as follows:

$$\mathcal{GG}_1(\mathcal{Y}_{11}) = \{f\mathcal{C}_1, f\mathcal{C}_2\}; \quad \mathcal{GG}_1(\mathcal{Y}_{12}) = \{f\mathcal{C}_3\}$$
$$\mathcal{GG}_1(\mathcal{Y}_{13}) = \{f\mathcal{C}_4, f\mathcal{C}_5\}; \quad \mathcal{GG}_2(\mathcal{Y}_{21}) = \{f\mathcal{C}_1, f\mathcal{C}_4, f\mathcal{C}_5\}$$
$$\mathcal{GG}_2(\mathcal{Y}_{22}) = \{f\mathcal{C}_3\}; \quad \mathcal{GG}_2(\mathcal{Y}_{23}) = \{f\mathcal{C}_2\}$$
$$\mathcal{GG}_3(\mathcal{Y}_{31}) = \{f\mathcal{C}_5\}; \quad \mathcal{GG}_3(\mathcal{Y}_{32}) = \{f\mathcal{C}_3, f\mathcal{C}_4\}$$
$$\mathcal{GG}_3(\mathcal{Y}_{33}) = \{f\mathcal{C}_1, f\mathcal{C}_2\}; \quad \mathcal{GG}_4(\mathcal{Y}_{41}) = \{f\mathcal{C}_5\}$$
$$\mathcal{GG}_4(\mathcal{Y}_{42}) = \{f\mathcal{C}_45\}; \quad \mathcal{GG}_4(\mathcal{Y}_{23}) = \{f\mathcal{C}_1, f\mathcal{C}_2, f\mathcal{C}_3\}$$
$$\mathcal{GG}_5(\mathcal{Y}_{51}) = \{f\mathcal{C}_4, f\mathcal{C}_5\}; \quad \mathcal{GG}_5(\mathcal{Y}_{52}) = \{f\mathcal{C}_1, f\mathcal{C}_2, f\mathcal{C}_3\}.$$

The above relations are true and satisfy the bijective soft set, thus consider that $(\mathcal{GG}_1, \mathcal{EFS}_1)$, then union soft sets of $(\mathcal{GG}_1, \mathcal{EFS}_1)$ concept sources, which is a universal set U or $\bigcup_{\mathcal{Y}_{ij} \in \mathcal{EFS}_i} \mathcal{GG}(\mathcal{Y}_{1j}) = U$. More in-depth two ($\mathcal{EFS}$) values, $\mathcal{Y}_{11}, \mathcal{Y}_{12} \in \mathcal{EFS}_1$, $\mathcal{Y}_{11} \neq \mathcal{Y}_{12}$, $\mathcal{GG}_1(\mathcal{Y}_{11}) \bigcap \mathcal{GG}_1(\mathcal{Y}_{12}) = \varnothing$.

4) After applying the bijective soft set, preference values are captured as per 4). Network security specialist $\mathcal{N}ER$ assign preference values as shown in Table I, where Low = 0.2; Medium = 0.5; High = 0.7; Very high = 0.9.

5) The projection value ($\mathfrak{p}v$), entropy ($\mathcal{E}nt$), divergence ($\mathcal{D}iv$), and weight ($\mathcal{W}gt$) of each data set feature values $\mathcal{Y}_{ij}$ are calculated as per 5) and shown in Table II.

6) After step number 5), the requirement from the network security expert for the effective feature selection we calculate the basic abstract as

$$\mathcal{N}ER_1 = \{\mathcal{Y}_{13}, \mathcal{Y}_{21}, \mathcal{Y}_{31}, \mathcal{Y}_{41}, \mathcal{Y}_{51}\}$$
$$\mathcal{N}ER_2 = \{\mathcal{Y}_{12}, \mathcal{Y}_{23}, \mathcal{Y}_{31}, \mathcal{Y}_{41}, \mathcal{Y}_{52}\}$$
$$\mathcal{N}ER_3 = \{\mathcal{Y}_{11}, \mathcal{Y}_{22}, \mathcal{Y}_{32}, \mathcal{Y}_{41}, \mathcal{Y}_{52}\}.$$

7) The network security experts the tabular soft set representation and can be shown in Tables III–V, respectively.

8) The main part of TOPSIS is conducted in this step, such as $\mathcal{NIS}$ and $\mathcal{IS}$ are calculated as shown in Table VI.

9) Then, the separation measures are computed of each $\mathfrak{N}ER$ from $\mathcal{IS}$ and $\mathcal{NIS}$ as per 9) as shown in Table VII. While combined and afore-computed separations are shown in Table VIII, respectively.

10) After the calculation of combined and separate measurements, in this final step, the closeness of $\mathcal{F}\zeta$ is calculated as shown in Table IX, which is the effective feature selection result. From the table, it clear that $\mathcal{F}\zeta_5$ gives the functional concept result as given $f\mathcal{C}_5 = \{\mathcal{Y}_{13}, \mathcal{Y}_{21}, \mathcal{Y}_{31}, \mathcal{Y}_{41}, \mathcal{Y}_{51}\} = \mathcal{EFS}$ are $f\mathcal{C}$ = High, Poor, V.Good, V.Good, Minimum. Thus, it is clear that for the effective feature selection, the above function concept should be considered for accurate identification as our proposed method selects the effective feature set.

IV. Evaluation Methodology

In this section, the evaluation criteria and selected data set for the proposed method are discussed in detail. First, the data set and then evaluation criteria are discussed for the detection of Bot-IoT attacks in the IoT network environment.

A. Bot-IoT Data Set

For the effective feature selection and accurate Bot-IoT attack identification in the IoT network environment, a new developed data set [19], [50] is used. The data set includes on the IoT, and normal traffic flows as well as several numerous cyberattacks traffic flows of botnets attacks. To trace the accurate traffic and develop an effective data set, the realistic testbed is used for the development of this data set with effective information features. Similarly, for the improvement of the ML model performance and effective prediction model, more features were extracted and added with the extracted feature set. However, for better performance results, the extracted features are labeled, such as attack flow, categories, and subcategories. The utilized testbed is categorized into three subcomponents as IoT services, which are simulated, network

TABLE I
PREFERENCE DECISION MATRIX

	$\mathcal{Y}_{11}$	$\mathcal{Y}_{12}$	$\mathcal{Y}_{13}$	$\mathcal{Y}_{21}$	$\mathcal{Y}_{22}$	$\mathcal{Y}_{23}$	$\mathcal{Y}_{31}$	$\mathcal{Y}_{32}$	$\mathcal{Y}_{33}$	$\mathcal{Y}_{41}$	$\mathcal{Y}_{42}$	$\mathcal{Y}_{43}$	$\mathcal{Y}_{51}$	$\mathcal{Y}_{52}$
$\mathfrak{R}ER_1$	0.2	0.5	0.9	0.7	0.5	0.2	0.9	0.7	0.2	0.9	0.7	0.5	0.9	0.5
$\mathfrak{R}ER_2$	0.5	0.9	0.7	0.5	0.7	0.9	0.7	0.5	0.5	0.5	0.5	0.2	0.5	0.7
$\mathfrak{R}ER_3$	0.9	0.7	0.2	0.2	0.9	0.7	0.5	0.5	0.2	0.7	0.5	0.2	0.2	0.9

TABLE II
PROJECTION, ENTROPY, DIVERGENCE, AND WEIGHT

	$\mathcal{Y}_{11}$	$\mathcal{Y}_{12}$	$\mathcal{Y}_{13}$	$\mathcal{Y}_{21}$	$\mathcal{Y}_{22}$	$\mathcal{Y}_{23}$	$\mathcal{Y}_{31}$	$\mathcal{Y}_{32}$	$\mathcal{Y}_{33}$	$\mathcal{Y}_{41}$	$\mathcal{Y}_{42}$	$\mathcal{Y}_{43}$	$\mathcal{Y}_{51}$	$\mathcal{Y}_{52}$
$\mathfrak{R}ER_1$	0.125	0.2380	0.5	0.5	0.25	0.117	0.5	0.411	0.4166	0.428	0.416	0.5	0.562	0.238
$\mathfrak{R}ER_2$	0.3125	0.4285	1.3888	0.3571	0.3	0.470	0.3888	0.2941	0.416	0.333	0.416	0.357	0.312	0.333
$\mathfrak{R}ER_3$	0.5625	0.333	0.111	0.1428	0.45	0.411	0.111	0.2941	0.166	0.238	0.166	0.142	0.125	0.428
$\mathcal{E}nt$	0.862	0.974	0.1233	0.902	0.971	0.887	0.123	0.987	0.935	0.974	0.935	0.902	0.862	0.974
$\mathcal{D}iv$	0.138	0.026	0.877	0.098	0.029	0.116	0.877	0.013	0.065	0.026	0.065	0.098	0.138	0.026
$\mathcal{W}gt$	0.053	0.010	0.338	0.0378	0.0111	0.044	0.338	0.005	0.025	0.010	0.025	0.037	0.053	0.010

TABLE III
SOFT SET REPRESENTATION OF $\mathcal{N}ER_1$

	$\mathcal{Y}_{13}$	$\mathcal{Y}_{21}$	$\mathcal{Y}_{31}$	$\mathcal{Y}_{41}$	$\mathcal{Y}_{51}$	$\mathcal{W}CV$
$\mathcal{F}\zeta_1$	0	1	0	0	0	0.0378
$\mathcal{F}\zeta_2$	0	0	0	0	0	0
$\mathcal{F}\zeta_3$	0	0	0	0	0	0
$\mathcal{F}\zeta_4$	1	1	0	0	1	0.4288
$\mathcal{F}\zeta_5$	1	1	1	1	1	0.7768
$\mathcal{W}gt$	0.338	0.0378	0.338	0.010	0.053	

TABLE IV
SOFT SET REPRESENTATION OF $\mathcal{N}ER_2$

	$\mathcal{Y}_{13}$	$\mathcal{Y}_{21}$	$\mathcal{Y}_{31}$	$\mathcal{Y}_{41}$	$\mathcal{Y}_{51}$	$\mathcal{W}CV$
$\mathcal{F}\zeta_1$	0	0	0	0	1	0.010
$\mathcal{F}\zeta_2$	0	1	0	0	1	0.054
$\mathcal{F}\zeta_3$	1	0	0	0	1	0.02
$\mathcal{F}\zeta_4$	0	0	0	0	0	0
$\mathcal{F}\zeta_5$	0	0	1	1	0	0.348
$\mathcal{W}gt$	0.10	0.044	0.338	0.010	0.010	

TABLE V
SOFT SET REPRESENTATION OF $\mathcal{N}ER_3$

	$\mathcal{Y}_{13}$	$\mathcal{Y}_{21}$	$\mathcal{Y}_{31}$	$\mathcal{Y}_{41}$	$\mathcal{Y}_{51}$	$\mathcal{W}CV$
$\mathcal{F}\zeta_1$	1	0	0	0	1	0.063
$\mathcal{F}\zeta_2$	1	0	0	0	1	0.063
$\mathcal{F}\zeta_3$	0	1	1	0	1	0.052
$\mathcal{F}\zeta_4$	0	0	1	0	1	0.015
$\mathcal{F}\zeta_5$	0	0	0	0	1	0.010
$\mathcal{W}gt$	0.053	0.0378	0.005	0.010	0.010	

TABLE VI
$\mathcal{N}IS$ AND $\mathcal{I}S$ FOR EACH $\mathfrak{R}ER$

$Network \quad Expert$	$\mathcal{I}S(\mathcal{Y}_i^*)$	$\mathcal{N}IS(\mathcal{Y}_i^{\circ})$
$\mathfrak{R}ER_1$	0.7768	0
$\mathfrak{R}ER_2$	0.348	0
$\mathfrak{R}ER_3$	0.063	0.010

platform, feature extraction, and forensics analytics. Similarly, to simulate IoT devices, five IoT devices are applied, such as an IoT device that generates weather information after every minute, such as to know about the current temperature, humidity, and atmospheric pressure. The second one is the smart cooling fridge, which gives information about cooling or current temperature information to adjust the smart IoT fridge temperature when necessary. The third one is the smart lights. These lights are a motion detector-based pseudorandom general signal. When motion is detected, the light automatically turns on, and when there is no motion, the light will remain turned off while the fourth one is a smart IoT door. Smart IoT doors are based on the probabilistic input. The fifth and the final one are an intelligent thermostat device used in houses for automatically adjusting and controlling a house temperature.

B. Performance Measurements

For the measurement of detection or identification performance of an ML model result, confusion metrics are widely used, which is based on the measurement of performance. However, the details of the graphical performance measurement presentation of a confusion matrix are shown in Fig. 3. In the graphical presentation of the confusion matrix, rows indicate the instances of classes, while the column shows identified class instances. Nevertheless, the widely used measurement for the evaluation of an ML model is discussed as follows.

1) *True Positive (TP):* In attack detection, the TP indicates that Class A is correctly identified as belonging to Class A.
2) *True Negative (TN):* This matrix indicates that Class A is correctly identified as not belonging to Class A.
3) *False Positive (FP):* It indicates that Class A is not correctly identified as belonging to Class A.
4) *False Negative (FN):* It indicates that Class A is not correctly identified as not belonging to Class A.

However, using the above-described metrics, different measurement metrics can be made to evaluate an ML model better. For accurate detection, ML classifiers minimize FP and FN metrics' values. However, the selected metrics that are used in this article explained in detail as follows.

1) *Accuracy:* In attack detection, it can be described as the correctly identified samples' of traffic in overall identified samples traffic. However, using performance measurement metrics, the accuracy can be defined mathematically as

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})}. \tag{4}$$

TABLE VII
SEPERATION MEASURE OF $\mathfrak{N}ER_1$, $\mathfrak{N}ER_2$, AND $\mathfrak{N}ER_3$ FROM $\mathcal{IS}$ AND $\mathcal{NIS}$

$\mathcal{F}unctional \quad Concepts$	$\Delta^*_{1\kappa}$	$\breve{\Delta}_{1\kappa}$	$\Delta^*_{2\kappa}$	$\breve{\Delta}_{2\kappa}$	$\Delta^*_{3\kappa}$	$\breve{\Delta}_{3\kappa}$
$\mathcal{F}\zeta_1$	0.546	0.001	0.0.114	0.000	0	0.000
$\mathcal{F}\zeta_2$	0.603	0	0.086	0.000	0	0.000
$\mathcal{F}\zeta_3$	0.603	0	0.1075	0.000	0.000	0.000
$\mathcal{F}\zeta_4$	0.121	0.183	0.121	0	0.0.000	0.000
$\mathcal{F}\zeta_5$	0	0.603	0	0.121	0.000	0

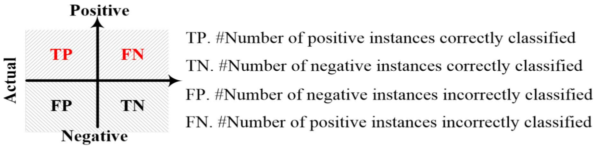

Positive
Actual
TP
FN
FP
TN
Negative
TP. #Number of positive instances correctly classified
TN. #Number of negative instances correctly classified
FP. #Number of negative instances incorrectly classified
FN. #Number of positive instances incorrectly classified

Fig. 3. Confusion matrix.

TABLE VIII
COMBINED SEPERATION MEASURE

$\mathcal{F}unctional\ Concepts$	Δ^*_{κ}	$\breve{\Delta}_{\kappa}$
$\mathcal{F}\zeta_1$	0.81240	0.031622
$\mathcal{F}\zeta_2$	0.83006	0
$\mathcal{F}\zeta_3$	0.842911	0
$\mathcal{F}\zeta_4$	0.491934	0.42778
$\mathcal{F}\zeta_5$	0	0.850881

TABLE IX
RELATIVE CLOSENESS OF $\mathcal{F}\zeta$

$\mathcal{F}unctional\ Concepts$	ζ^*_{κ}
$\mathcal{F}\zeta_1$	0.02801802
$\mathcal{F}\zeta_2$	0
$\mathcal{F}\zeta_3$	0
$\mathcal{F}\zeta_4$	0.46512285
$\mathcal{F}\zeta_5$	1.0

In our study, we used (4) for the ML classifier's performance evaluation. Using these metrics, the effectiveness of an ML classifier can be identified.

2) *Precision:* It can be defined as the correctly identified sample in the percentage of Class A in all those were identified in Class A. The mathematical formula used in this research study is shown as follows:

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})}. \tag{5}$$

3) *Sensitivity:* It can be described as the correctly detected traffic sample divided by the overall data set traffic sample. However, this metric can be used as a recall metric in Bot-IoT detection in the IoT environment. We used the following given mathematical formula for the sensitivity metric as follows:

$$\text{Sensitivity} = \frac{\text{TP}}{(\text{TP} + \text{FN})}. \tag{6}$$

4) *Specificity:* In this research study, we used the specificity metrics which can be defined as the ability of an ML classifier to detect negative results. The mathematical equation of specificity is shown as follows:

$$\text{Specificity} = \frac{\text{TN}}{(\text{FP} + \text{TN})}. \tag{7}$$

However, we used the above given metrics for the proposed technique performance evaluation.

V. RESULTS AND ANALYSIS

In this section, the detailed results and analysis of the proposed method are discussed. In this study, we proposed a new technique for the detection of Bot-IoT attacks in the IoT network environment. For the effective feature selection, our proposed method selected only five effective features, which carry enough information for the Bot-IoT attack detection in the IoT network environment. For this aim to select effective features, four different ML algorithms are applied for the proposed technique performance evaluation, such as decision tree (C4.5), support vector machine (SVM), Naive Bayes, and random forest ML algorithms. Though, all the applied four ML algorithms' performances are effective for the Bot-IoT attack's detection in the IoT network environment by using the feature set selected by our proposed technique with respective accuracy, precision, sensitivity, and specificity. However, the Naive Bayes performance result is low as compared to other ML classifiers by using the selected feature set concerning the accuracy metric for the Bot-IoT attack detection. Similarly, the performance result of SVM is slightly higher with respective accuracy as compared to Naive Bayes ML classifiers, as shown in Fig. 4. However, the C4.5 decision tree and random forest ML algorithm give beneficial performance results regarding accuracy. However, the overall applied ML classifier performance results of the C4.5 decision tree give effective results compared to other applied ML classifiers. Therefore, the C4.5 ML algorithm performs better by using the selected feature set for the detection of Bot-IoT attacks as 99.9%, which is very effective performance results. However, the detailed results' chart for accuracy is shown in Fig. 4.

In Fig. 5, the detailed precision result is shown. From the figure, it is evident that the C4.5 decision tree and random forest ML algorithm achieve effective performance results as compared to SVM and Naive Bayes ML algorithms. However, normal traffic and KeylogingTheft attacks are detected effectively, but the performance is low as compare to UDPDoS and other attacks with respective precision metric. However, taking an average of all the applied ML classifier's performance results. It has been seen that only KeylogingTheft traffics are low detected compared to other normal and other attacks using the selected features with respective precision metrics. All the applied four ML classifiers achieve very effective performance results with respective sensitivity metrics. However, random forest and C4.5 decision

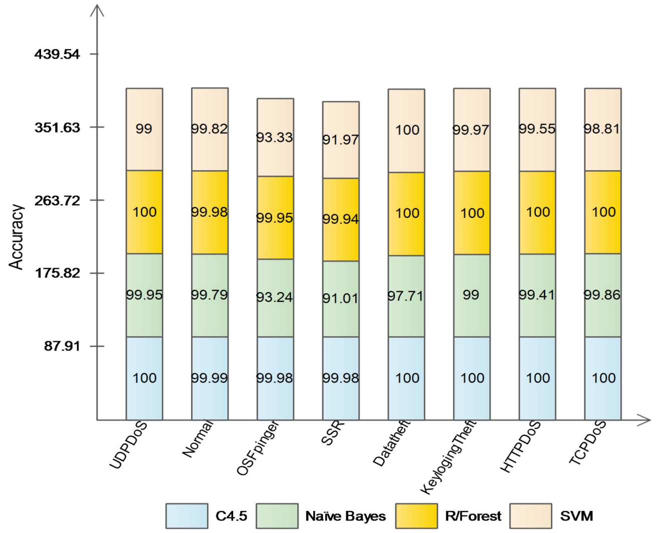


Fig. 4. Accuracy results.

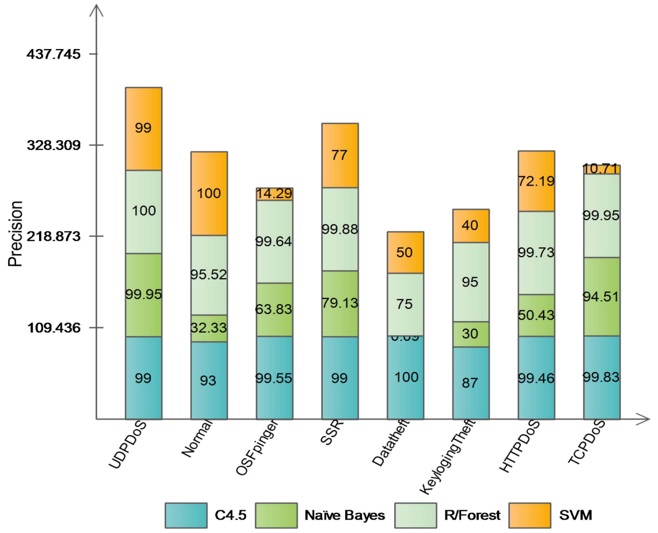


Fig. 5. Precision results.

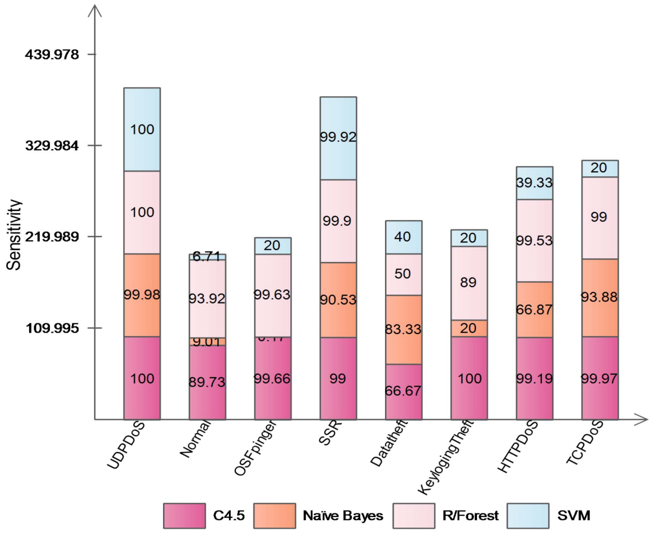


Fig. 6. Sensitivity results.

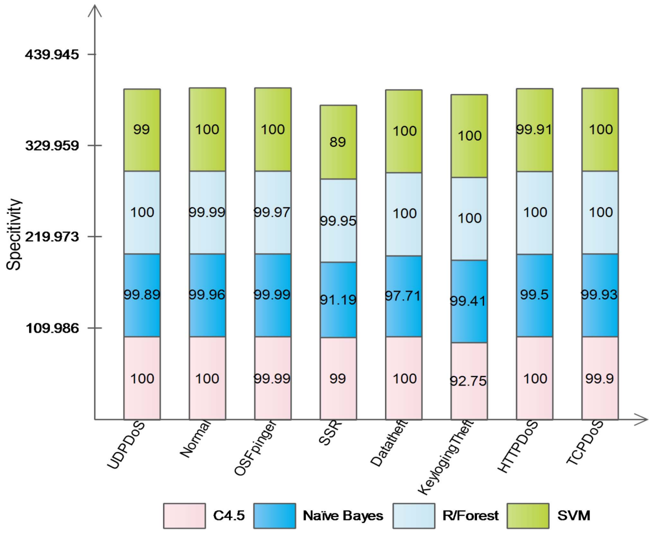


Fig. 7. Specificity results.

tree ML algorithms achieve very high-performance results by using the selected feature set as compared to other applied ML classifiers for the Bot-IoT attack detection in the IoT network environment. For the sensitivity metric, the same as accuracy and precision, the SVM and Naive Bayes ML classifier's performance results are low as compared to the C4.5 decision tree and random forest ML classifiers, as shown in Fig. 6. The specificity results of the applied ML classifiers are shown in Fig. 7 by using our proposed method selected feature set for the identification of Bot-IoT attacks in the IoT network environment. All the applied ML algorithms' performance results are very effective regarding specificity as C4.5 decision tree, and random forest are 98.95% and 99.99% while Naive Bayes and SVM are 98.44% and 98.48%, which are very effective performance results with respective specificity metric. Similarly, all the attacks and normal traffics are very effectively detected by using the selected feature set. It is clear from the analysis of the above results that our proposed feature selection technique is effective for the selection of features for the Bot-IoT detection in the IoT network environment.

VI. Analysis and Discussion

Even though the results of our proposed technique for Bot-IoT attack detection in the IoT network environment are auspicious by using the selected four different ML algorithms with accuracy, precision, sensitivity, and specificity and by utilizing the newly developed Bot-IoT data set. However, some useful information that we learned after the experimental analysis is given as follows.

1) In this study, it is clear and evident that the proposed technique is effective for the selection of optimum features in Bot-IoT attack detection in the IoT network environment by using the newly developed Bot-IoT data set. For the results' analysis and evaluation accuracy, precision, sensitivity, and specificity metrics are used

to evaluate the performance of the proposed method accurately.

2) It is also evident and seen from this study that the proposed method select optimum feature which carries enough detection information knowledge for the cyber-attacks detection in the IoT network environment.
3) In this, it is noticed that applied ML algorithms' performance is auspicious with respective accuracy, precision, sensitivity, and specificity. However, all the attacks are very precisely detected, but only KeylogingTheft attacks are poorly detected as compared to the rest of the attacks.
4) In the analysis of the experimental results, the applied ML algorithm's performance is very effective for the detection of Bot-IoT attacks. However, the C4.5 decision tree and random forest ML algorithms are very promising by using the Bot-IoT data set. SVM and Naive Bayes ML algorithms' performance are also effective, but compared to C4.5 decision tree and Random Forest algorithms, the performance is slightly weak.

VII. Conclusion

The detection of attacks in the IoT network is essential for IoT security to keep eyes and block unwanted traffic flows. Numerous ML technique models are presented by many researchers to block attack traffic flows in the IoT network. However, due to the inappropriate feature selection, several ML models prone misclassify mostly malicious traffic flows. Nevertheless, the noteworthy problem still needs to be studied more in-depth, that is how to select effective features for accurate malicious traffic detection in IoT networks. For this purpose, a new framework model was proposed. First, a novel feature selection metric approach named CorrAUC was proposed, and then based on CorrAUC, a new feature selection algorithm named CorrAUC was developed and designed, which was based on the wrapper technique to filter the feature accurately and select effective features for the selected ML algorithm by using the AUC metric. We then applied integrated TOPSIS and Shannon entropy based on a bijective soft set to validate selected features for malicious traffic identification in IoT networks. We evaluate our proposed approach by using the Bot-IoT data set and four different ML algorithms. The experimental results analysis showed that our proposed method is efficient and can achieve >96% results on average.

References

[1] J. Qiu, Z. Tian, C. Du, Q. Zuo, S. Su, and B. Fang, "A survey on access control in the age of Internet of Things," *IEEE Internet Things J.*, vol. 7, no. 6, pp. 4682–4696, Jun. 2020.

[2] Y. N. Soe, Y. Feng, P. I. Santosa, R. Hartanto, and K. Sakurai, "Implementing lightweight IoT-IDS on Raspberry Pi using correlation-based feature selection and its performance evaluation," in *Proc. Int. Conf. Adv. Inf. Netw. Appl.*, 2019, pp. 458–469.

[3] K. Lab. (2019). *Amount of Malware Targeting Smart Devices More Than Doubled in*. [Online]. Available: https://www.kaspersky.com/about/press-releases/2017_amount-of-malware

[4] J. Qiu, L. Du, D. Zhang, S. Su, and Z. Tian, "Nei-TTE: Intelligent traffic time estimation based on fine-grained time derivation of road segments for smart city," *IEEE Trans. Ind. Informat.*, vol. 16, no. 4, pp. 2659–2666, Apr. 2020.

[5] J. P. Anderson. (1980). *Computer Security Threat Monitoring and Surveillance*. Accessed: Nov. 30, 2008. [Online]. Available: https://csrc.nist.gov/csrc/media/publications/conference-paper/1998/10/08/proceedings-of-the-21st-nissc-1998/documents/early-cs-papers/ande80.pdf

[6] D. E. Denning, "An intrusion-detection model," *IEEE Trans. Softw. Eng.*, vol. SE-13, no. 2, pp. 222–232, Feb. 1987.

[7] X. Du, M. Guizani, Y. Xiao, and H.-H. Chen, "Defending DoS attacks on broadcast authentication in wireless sensor networks," in *Proc. IEEE Int. Conf. Commun.*, 2008, pp. 1653–1657.

[8] L. Wu, X. Du, W. Wang, and B. Lin, "An out-of-band authentication scheme for Internet of Things using blockchain technology," in *Proc. IEEE Int. Conf. Comput. Netw. Commun. (ICNC)*, 2018, pp. 769–773.

[9] Z. Tian, X. Gao, S. Su, and J. Qiu, "VCash: A novel reputation framework for identifying denial of traffic service in Internet of connected vehicles," *IEEE Internet Things J.*, vol. 7, no. 5, pp. 3901–3909, May 2020.

[10] S. Alharbi, P. Rodriguez, R. Maharaja, P. Iyer, N. Bose, and Z. Ye, "Focus: A fog computing-based security system for the Internet of Things," in *Proc. 15th IEEE Annu. Consum. Commun. Netw. Conf. (CCNC)*, 2018, pp. 1–5.

[11] Z. Tian, C. Luo, J. Qiu, X. Du, and M. Guizani, "A distributed deep learning system for Web attack detection on edge devices," *IEEE Trans. Ind. Informat.*, vol. 16, no. 3, pp. 1963–1971, Mar. 2020.

[12] D. Ventura *et al.*, "ARIIMA: A real IoT implementation of a machine-learning architecture for reducing energy consumption," in *Proc. Int. Conf. Ubiquitous Comput. Ambient Intell.*, 2014, pp. 444–451.

[13] R. Xue, L. Wang, and J. Chen, "Using the IoT to construct ubiquitous learning environment," in *Proc. IEEE 2nd Int. Conf. Mech. Autom. Control Eng.*, 2011, pp. 7878–7880.

[14] M. A. Alsheikh, S. Lin, D. Niyato, and H.-P. Tan, "Machine learning in wireless sensor networks: Algorithms, strategies, and applications," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 4, pp. 1996–2018, 4th Quart., 2014.

[15] M. Shafiq, X. Yu, A. A. Laghari, and D. Wang, "Effective feature selection for 5G IM applications traffic classification," *Mobile Inf. Syst.*, vol. 2017, May 2017, Art. no. 6805056.

[16] M. Shafiq, X. Yu, A. K. Bashir, H. N. Chaudhry, and D. Wang, "A machine learning approach for feature selection traffic classification using security analysis," *J. Supercomput.*, vol. 74, no. 10, pp. 4867–4892, 2018.

[17] M. Dash and H. Liu, "Feature selection for classification," *Intell. Data Anal.*, vol. 1, nos. 1–4, pp. 131–156, 1997.

[18] H. Zhang, G. Lu, M. T. Qassrawi, Y. Zhang, and X. Yu, "Feature selection for optimizing traffic classification," *Comput. Commun.*, vol. 35, no. 12, pp. 1457–1471, 2012.

[19] N. Koroniotis, N. Moustafa, E. Sitnikova, and B. Turnbull, "Towards the development of realistic botnet dataset in the Internet of Things for network forensic analytics: Bot-IoT dataset," 2018. [Online]. Available: arXiv:1811.00701.

[20] M. Shafiq and X. Yu, "Effective packet number for 5G IM wechat application at early stage traffic classification," *Mobile Inf. Syst.*, vol. 2017, Feb. 2017, Art. no. 3146868.

[21] M. Shafiq, X. Yu, A. A. Laghari, L. Yao, N. K. Karn, and F. Abdessamia, "Network traffic classification techniques and comparative analysis using machine learning algorithms," in *Proc. 2nd IEEE Int. Conf. Comput. Commun. (ICCC)*, 2016, pp. 2451–2455.

[22] M. Shafiq, X. Yu, and A. A. Laghari, "Wechat text messages service flow traffic classification using machine learning technique," in *Proc. 6th Int. Conf. IT Converg. Security (ICITCS)*, 2016, pp. 1–5.

[23] M. Shafiq *et al.*, "Wechat text and picture messages service flow traffic classification using machine learning technique," in *Proc. IEEE 18th Int. Conf. High Perform. Comput. Commun. IEEE 14th Int. Conf. Smart City IEEE 2nd Int. Conf. Data Sci. Syst. (HPCC/SmartCity/DSS)*, 2016, pp. 58–62.

[24] M. Shafiq, Z. Tian, A. K. Bashir, X. Du, and M. Guizani, "IoT malicious traffic identification using wrapper-based feature selection mechanisms," *Comput. Security*, vol. 94, Jul. 2020, Art. no. 101863.

[25] 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 Gener. Comput. Syst.*, vol. 107, pp. 433–442, Jun. 2020.

[26] M. Shafiq, Z. Tian, A. K. Bashir, A. R. Jolfaei, and X. Yu, "Data mining and machine learning methods for sustainable smart cities traffic classification: A survey," *Sustain. Cities Soc.*, vol. 60, Sep. 2020, Art. no. 102177.

[27] Y. Xiao, X. Du, J. Zhang, F. Hu, and S. Guizani, "Internet protocol television (IPTV): The killer application for the next-generation Internet," *IEEE Commun. Mag.*, vol. 45, no. 11, pp. 126–134, Nov. 2007.

[28] Z. Tian, S. Su, W. Shi, X. Du, M. Guizani, and X. Yu, "A data-driven method for future Internet route decision modeling," *Future Gener. Comput. Syst.*, vol. 95, pp. 212–220, Jun. 2019.

[29] Z. Tian, X. Gao, S. Su, J. Qiu, X. Du, and M. Guizani, "Evaluating reputation management schemes of Internet of Vehicles based on evolutionary game theory," *IEEE Trans. Veh. Technol.*, vol. 68, no. 6, pp. 5971–5980, Jun. 2019.

[30] Y. Xiao, V. K. Rayi, B. Sun, X. Du, F. Hu, and M. Galloway, "A survey of key management schemes in wireless sensor networks," *Comput. Commun.*, vol. 30, nos. 11–12, pp. 2314–2341, 2007.

[31] X. Du and H.-H. Chen, "Security in wireless sensor networks," *IEEE Wireless Commun.*, vol. 15, no. 4, pp. 60–66, Jan. 2008.

[32] S. Egea, A. R. Mañez, B. Carro, A. Sánchez-Esguevillas, and J. Lloret, "Intelligent IoT traffic classification using novel search strategy for fast-based-correlation feature selection in industrial environments," *IEEE Internet Things J.*, vol. 5, no. 3, pp. 1616–1624, Jun. 2018.

[33] Y. Meidan *et al.*, "N-BAIoT—Network-based detection of IoT Botnet attacks using deep autoencoders," *IEEE Pervasive Comput.*, vol. 17, no. 3, pp. 12–22, May 2018.

[34] S. Su, Y. Sun, X. Gao, J. Qiu, and Z. Tian, "A correlation-change based feature selection method for IoT equipment anomaly detection," *Appl. Sci.*, vol. 9, no. 3, p. 437, 2019.

[35] Q. Tan, Y. Gao, J. Shi, X. Wang, B. Fang, and Z. H. Tian, "Towards a comprehensive insight into the eclipse attacks of TOR hidden services," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 1584–1593, Apr. 2019.

[36] Z. Tian *et al.*, "Real time lateral movement detection based on evidence reasoning network for edge computing environment," *IEEE Trans. Ind. Informat.*, vol. 15, no. 7, pp. 4285–4294, May 2019.

[37] X. Du, Y. Xiao, M. Guizani, and H.-H. Chen, "An effective key management scheme for heterogeneous sensor networks," *Ad Hoc Netw.*, vol. 5, no. 1, pp. 24–34, 2007.

[38] X. Du, M. Guizani, Y. Xiao, and H. Chen, "A routing-driven elliptic curve cryptography based key management scheme for heterogeneous sensor networks," *IEEE Trans. Wireless Commun.*, vol. 8, no. 3, pp. 1223–1229, Mar. 2009.

[39] X. Du, M. Zhang, K. E. Nygard, S. Guizani, and H.-H. Chen, "Self-healing sensor networks with distributed decision making," *Int. J. Sensor Netw.*, vol. 2, nos. 5–6, pp. 289–298, 2007.

[40] X. Du, M. Shayman, and M. Rozenblit, "Implementation and performance analysis of SNMP on a TLS/TCP base," in *Proc. IEEE/IFIP Int. Symp. Integr. Netw. Manag. VII*, 2001, pp. 453–466.

[41] X. Huang and X. Du, "Achieving big data privacy via hybrid cloud," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, 2014, pp. 512–517.

[42] R. E. Fancher, "Galton on examinations: An unpublished step in the invention of correlation," *ISIS*, vol. 80, no. 3, pp. 446–455, 1989.

[43] L. Peng, B. Yang, Y. Chen, and Z. Chen, "Effectiveness of statistical features for early stage Internet traffic identification," *Int. J. Parallel Program.*, vol. 44, no. 1, pp. 181–197, 2016.

[44] D. Molodtsov, "Soft set theory—First results," *Comput. Math. Appl.*, vol. 37, nos. 4–5, pp. 19–31, 1999.

[45] K. Gong, Z. Xiao, and X. Zhang, "The bijective soft set with its operations," *Comput. Math. Appl.*, vol. 60, no. 8, pp. 2270–2278, 2010.

[46] K. Hayat, M. I. Ali, B.-Y. Cao, and X.-P. Yang, "A new type-2 soft set: Type-2 soft graphs and their applications," *Adv. Fuzzy Syst.*, vol. 2017, Oct. 2017, Art. no. 6162753.

[47] V. Tiwari, P. K. Jain, and P. Tandon, "An integrated Shannon entropy and TOPSIS for product design concept evaluation based on bijective soft set," *J. Intell. Manuf.*, vol. 30, no. 4, pp. 1645–1658, 2019.

[48] A. R. Roy and P. Maji, "A fuzzy soft set theoretic approach to decision making problems," *J. Comput. Appl. Math.*, vol. 203, no. 2, pp. 412–418, 2007.

[49] T.-C. Wang and H.-D. Lee, "Developing a fuzzy TOPSIS approach based on subjective weights and objective weights," *Expert Syst. Appl.*, vol. 36, no. 5, pp. 8980–8985, 2009.

[50] I. Van der Elzen and J. van Heugten, "Techniques for detecting compromised IoT devices," Univ. Amsterdam, Amsterdam, The Netherlands, 2017.

Muhammad Shafiq was born in Pakistan. He received the B.S. degree (Hons.) rank in computer science from the Faculty of Computer Science, Malakand University, Chakdara, Pakistan, in 2009, the M.S. degree in computer science from the Faculty of Computer Science, Malakand University in 2011, and the Ph.D. degree from the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China, in 2018. He is currently pursuing the Postdoctoral degree with the Cyberspace Institute of Advance Technology, Guangzhou University, Guangzhou, China.

His current research areas of interests include IoT security, IoT anomaly and intrusion traffic classification, IoT management, network traffic classification and network security, and cloud computing.

Zhihong Tian (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in computer science and technology from Harbin Industrial University, Harbin, China, in 2001, 2003, and 2006, respectively.

He is currently a Professor and the Dean of the Cyberspace Institute of Advanced Technology, Guangzhou University, Guangzhou, China. He is also a part-time Professor with Carlton University, Ottawa, ON, Canada. He served in different academic and administrative positions with the Harbin Institute of Technology. He has authored over 200 journal and conference papers in these areas. His research has been supported in part by the National Natural Science Foundation of China, National Key research and Development Plan of China, and National High-tech Research and Development Program of China (863 Program). His research interests include computer networks and cyberspace security.

Dr. Tian also served as a member, the chair, and the general chair of a number of international conferences. He is a Senior Member of the China Computer Federation. The Distinguished Professor of Guangdong Province Universities and Colleges Pearl River Scholar.

Ali Kashif Bashir (Senior Member, IEEE) received the B.S. degree from the University of Management and Technology, Lahore, Pakistan, the M.S. degree from Ajou University, Suwon, South Korea, and the Ph.D. degree in computer science and engineering from Korea University, Seoul, South Korea.

He is a Senior Lecturer with the Department of Computing and Mathematics, Manchester Metropolitan University, Manchester, U.K. From 2006 to 2007, he was a Project Engineer with Consistel Telecom, Lahore, Pakistan. From 2011 to 2012, he was a Lecturer with Korea Southern Power Company Ltd., Busan, South Korea. From 2012 to 2013, he was a Postdoctoral Researcher with National Fusion Research Institute, Daejeon, South Korea. From 2013 to 2016, he was a Specially Appointed Researcher with the Graduate School of Information Science and Technology, Osaka University, Osaka, Japan. From 2014 to 2016, he was a Visiting Assistant Professor with the National Institute of Technology, Nara, Japan. From 2017 to 2018, he was an Associate Professor with the Faculty of Science and Technology, University of the Faroe Islands, Faroe Islands, Denmark. He has authored over 80 peer-reviewed articles. He has delivered over 20 invited and keynote talks in seven countries. He is supervising/co-supervising several graduate (M.S. and Ph.D.) students. His research interests include Internet of Things, wireless networks, distributed systems, network/cybersecurity, and network function virtualization.

Dr. Bashir is a Distinguished Speaker of ACM and a member of ACM, IEEE Young Professionals, and the International Association of Educators and Researchers, U.K. He has served as the chair (program, publicity, and track) on top conferences and workshops. He is serving as the Editor-in-Chief for the IEEE FUTURE DIRECTIONS NEWSLETTER. He is advising several startups in the field of STEM-based education, robotics, Internet of Things, and blockchain.

Xiaojiang Du (Fellow, IEEE) received the B.S. and M.S. degrees in electrical engineering (Automation Department) from Tsinghua University, Beijing, China, in 1996 and 1998, respectively, and the M.S. and Ph.D. degrees in electrical engineering from the University of Maryland at College Park, College Park, MD, USA, in 2002 and 2003, respectively.

He is a tenured Full Professor and the Director of the Security and Networking Laboratory, Department of Computer and Information Sciences, Temple University, Philadelphia, PA, USA. His research interests are security, wireless networks, and systems. He has authored over 400 journal and conference papers in the above areas, as well as a book published by Springer.

Prof. Du has been awarded more than six million U.S. Dollars research grants from the U.S. National Science Foundation, Army Research Office, Air Force Research Laboratory, NASA, the State of Pennsylvania, and Amazon. He won the best paper award at IEEE GLOBECOM 2014 and the Best Poster Runner-Up Award at ACM MobiHoc 2014. He serves on the editorial boards of two international journals. He served as the Lead Chair of the Communication and Information Security Symposium of the IEEE International Communication Conference 2015, and the Co-Chair of Mobile and Wireless Networks Track of IEEE Wireless Communications and Networking Conference 2015. He is (was) a Technical Program Committee Member of several premier ACM/IEEE conferences, such as INFOCOM from 2007 to 2020, IM, NOMS, ICC, GLOBECOM, WCNC, BroadNet, and IPCCC. He is a Life Member of ACM.

Mohsen Guizani (Fellow, IEEE) received the B.S. (Distinction) and M.S. degrees in electrical engineering, the M.S. and Ph.D. degrees in computer engineering from Syracuse University, Syracuse, NY, USA, in 1984, 1986, 1987, and 1990, respectively.

He is currently a Professor with the Computer Science and Engineering Department, Qatar University, Qatar. His research interests include wireless communications and mobile computing, computer networks, mobile cloud computing, security, and smart grid.

Prof. Guizani received three teaching awards and four research awards. He also received the 2017 IEEE Communications Society WTC Recognition Award and the 2018 AdHoc Technical Committee Recognition Award for his contribution to outstanding research in wireless communications and Ad-Hoc Sensor networks. He was the Chair of the IEEE Communications Society Wireless Technical Committee and the TAOS Technical Committee. He served as the IEEE Computer Society Distinguished Speaker and is currently the IEEE ComSoc Distinguished Lecturer. He is a Senior Member of ACM.