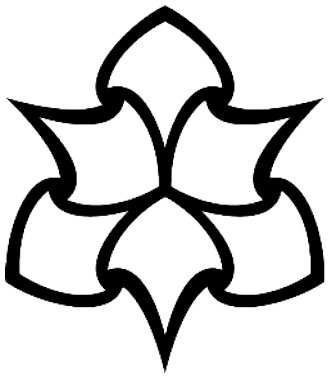

CorrAUC: a Malicious Bot-IoT Traffic Detection Method in IoT Network Using Machine Learning Techniques

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 IoT network. To address the problem, a new framework model is proposed. Firstly, a novel feature selection metric approach named CorrAUC proposed, and then based on CorrAUC, a new feature selection algorithm name Corrauc is develop and design, which is based on wrapper technique to filter the features accurately and select effective features for the selected ML algorithm by using AUC metric. Then, we applied 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 dataset and four different ML algorithms. Experimental results analysis showed that our proposed method is efficient and can achieve >96% results on average.**

***Index Terms*—Internet of Things, Malicious, Intrusion, Attacks, Detection, Idntification, Machine Learning.**

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, IoT technology was limited to small offices and homes, but nowadays, IoT technology integrated into industries for more reliability and saving time. However, 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 IoT technology world [2]. Though IoT technology is growing day by day, but on the other hand, the cyber-attacks are also becoming challenging and increase. 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 cyber-attacks and unauthorized access. Recently, IoT security becomes a hot topic and gained much attention in IoT cybersecurity. To overcome the problem of IoT cyber-attacks, 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 Internet of Things (IoT) attacks such as Denial of Service (DDoS) become very spread and grow up to 172 %, which gain much interest in IoT network [2].

According to the Kaspersky Lab 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 Intrusion Detection System (IDS) was discussed and introduced by Anderson in 1980 [5]. In 1987 Denning [6] proposed a new model for the detection of intrusion based on real-time intrusion detection. 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 damage to the computer system, etc. However, their proposed model was based on 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 user behavior. Nowadays, in the Internet of Things (IoT), the most dangerous and challenging widespread hazardous threats are man-in-the-middle (MITM) dangerous threats with distributed denial of service (DDoS) [7][8][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 S et al. in [10] proposed a new system for the detection of malware cyber-attacks and to protect IoT against from cyber-attacks named Fog Computing-based

Muhammad Shafiq and Zhihong Tian are with the Department of Cyberspace Insitute of Advanced Technology, GuangZhou University, GuangZhou,China, 510006.
E-mail: srsshafiq@gmail.com,

Ali Kashif Bashir is with the Department of Computing and Mathematics, Manchester Metropolitan University, United Kingdom

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

Mohsen Guizani is with the College of Engineering, Qatar University, Qatar

Correspondence: Zhihong Tian is with Cyberspace Institute of Advanced Technology, GuangZhou University, GuangZhou, China. (tianzhihong@gzhu.edu.cn).

Security (FOCUS) system. Their proposed intrusion detection system [11] used the virtual private network (VPN) for secure communication between the Internet of Thing 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 cyber-attacks 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 cyber-attack 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 cyber-attacks in IoT network environment [12], [13], [14]. Internet of Things (IoT) traffic identification is vital for IoT security monitoring and IoT traffic management. Recently, the ML technique gains much importance and become very popular in numerous fields because of its accurate results.

For effective identification, significant features set is very important for the ML model. Effective features indicate accurate feature or attributes which keep significance information for ML technique, and these effective features set includes on training and testing set. It's impossible to evaluate the machine learning model without training set and testing set. Thus useful features set of training and testing sets are compulsory for the evaluation of the ML model. ML technique is widely applied in computer science, especially in network traffic identification [15], [16]. Machine Learning methods are very useful for identifying or classifying malicious, intrusion, and cyber-attacks in IoT networks. Though applying the ML method for the detection for the classification of malicious traffic of cyber-attacks is effective, but as compare to other computing tools, the ML technique tool is very complex. Though using a machine learning method is very effective in the area of identification or classification. Still, it is also some disadvantages in the IoT malicious and intrusion detection 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 cyber-attacks in the IoT network environment.

For the significant identification performance results, the input dataset 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 is don't 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 features selection and select the essential features set for ML detection in IoT for accurate detection of anomaly and intrusion problems. Similarly, to overcome the problem of effective feature selection Zhang H et al. [18] introduced two new techniques and proposed two different algorithms. Their proposed methods are able to select effective features 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 feature selection in high dimensional datasets, especially for imbalanced datasets. Similarly, Doroniotis et al. in 2018 in [19] introduced a new data set named Bot-IoT for the identification of cyber-attacks 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 cyber-attacks. More in-depth, the dataset is developed in a realistic testbed with defined features set, which consist of normal traffic and cyber-attacks traffic flow. For the experimental analysis, statistical analysis is performed to find out which feature carries accurate information for the detection of cyber-attacks in the IoT network. However, they selected the ten best feature set from the extracted feature set. They then used a well-known machine learning classifier for the performance analysis of the selected feature set. More in-depth for the performance analysis to find out which feature give the most effective results four different types of metrics are used.

In our previous study [15], [20], [21], [22], [23], [24], [25] [26] feature selection problems are studied and select robust features by proposing different types of approaches for Instant Messages (IM) traffic identification and attacks traffic detection. Similarly, in [15], [16], to overcome the problem of feature selection problem, different feature selection techniques are proposed for the accurate network traffic classification using machine learning algorithms. However, from the above study, we concluded that selecting more features set is not efficient for the accurate identification by using ML techniques and showed that selecting more than 50 features set can low the ML classifier accuracy and can increase computational complexity. However, in the IoT network cyber-attacks 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 paper, a new effective feature selection technique is proposed for the problem of effective feature selection for cyber-attacks in IoT network traffic by using the Bot-IoT dataset and to improve the performance of ML techniques. However, the main contributions in this paper are:

- In order to deal with the effective feature selection problem in IoT cyber-attacks identification in the IoT network. Firstly, a novel feature selection metric approach named CorrAUC proposed to deal with the issue of effective

feature selection for cyber-attacks identification in IoT networks. However, it's the first time to put forward combine correlation attribute evaluation metric and specific machine learning AUC results for the effective features selection in IoT Bot-IoT attacks detection.

- Then based on CorrAUC a new feature selection algorithm named Corrauc is develop and design for the problem of features selection for malicious Bot-IoT traffic identification in IoT network, which is based on wrapper technique to filter the features set accurately and select the features set that carry enough information for the selected ML algorithm by using AUC metric for the detection of Bot-IoT cyber-attacks in IoT network environment. However, the proposed algorithm includes two steps metrics for the optimum feature selection. Correlation Attributes Evaluation (CAE) metric and a specific used ML algorithm Area Under roc Curve (AUC) metric.
- Afterwards, 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 IoT network. It is based on the selection of proper attributes mean features 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.
- Then, we concluded and put forward the optimum selected features set selected by our proposed technique that carry enough information for the detection of malicious Bot-IoT traffic in the Internet of Things (IoT) network. Experimental showed that five optimum feature set carry enough information and have discriminative power for the detection of malicious attacks in IoT network by using machine learning.

The remainder of this paper is arranged as follows: Section 2 includes related works. In Section 3 we demonstrate the proposed techniques. While in Section 4, we explain with details the methodology, experimental work, and applied dataset. Similarly, Section 5 includes analysis and discussion. Finally, Section 6 concludes the paper conclusion and future directions.

II. Related Works

From the last decade, security and trust problems become a scorching topic, and many researcher endeavors hard to overcome this problem and proposed numerous effective models along with the future Internet [27], [28], IoVs [29], wireless sensor network (WSN) [30], [31] and IoTs. 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 Instant Messaging (IM) applications traffic classification, a feature selection technique is proposed based on 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 ML performance. However, feature selection is a process to select the optimum features set from several features set and removed the features that don't carry enough identification information for the identification or removing the redundant feature. S Egea et al. in 2018 [32] studied mostly cited research studies related to feature selection technique, especially the correlation coefficient technique, and propose a new feature selection technique named Fast Based Correlation Features (FCBF) algorithm for the improvement of the performance of 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, in 2018 Meidan Yair at el. [33] studied the detection of attacks in IoT network and proposed a new technique to overcome the problem of attacks which is initiated by the Internet of Things (IoT) devices and then for the identification of anomalies in IoT traffic they used autoencoder. The dataset that they used in their study for the evaluation of their proposed approach is botnet attacks Bashlite and Mirai based on the Internet of Things. However, the utilized datasets are also included on several infected devices in the IoT network. They showed in their study that the proposed approach is able to detect cyber-attacks in IoT network devices with high-performance results.

Similarly, for the IoT devices, performance improvement, and detection of an anomaly, Shen Su at el. [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 Multi-Cluster 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 cyber-security purpose in IoT security environment, for instances, cyber-attacks identification in [35], [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 feature selection technique includes on different important steps such as trace traffic, to trace the original traffic, subset generation, to generate features set from the trace traffic, subset evaluation, to evaluate the generated features set for next phase, decision-maker take some decision for the effective feature selection and then in subset evaluation gives the final decision and validate the feature set [39][40][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. Firstly, a novel feature selection metric approach named CorrAUC is proposed and applied, which select features select that carry enough information and then based on CorrAUC a new feature selection algorithm name Corrauc is develop and design, which is based on wrapper technique to filter the feature accurately and select effective features for the selected ML algorithm by using AUC metric and Bot-IoT dataset. The proposed algorithm consists of Correlation Attribute Evaluation (CAE) and combines with Area Under Roc Curve (AUC) metric to overcome the problem of effective feature selection for Bot-IoT detection by using a specific machine learning (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 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 study knowledge, in this study, Corr and AUC are combined and conducted for the first time for the identification of the Bot-IoT attack in IoT network using machine learning algorithms. Moreover, our proposed method select feature set that carries enough identification information for the Bot-IoT attacks in IoT network. For a clear understanding, the details methodologies are discussed in the next section for the effective feature selection in IoT network, considering Bot-IoT malicious attacks detection.

A. Feature Selection Metrics

In this section, the conducted features selection metrics are discussed with details. Firstly correlation-based metric is presented and then AUC metrics. However, the details are given bellow subsection.

1) Correlation Based Metric: To overcome the problem of effective feature selection, for BoT-IoT malicious attack detection in 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. F Galton proposed the basic idea of Pearson moment correlation in the 1880s [42]. Similarly, after sixteen years in 1896, K Pearson make changes in Pearson moment correlation and named it Pearson Product Moment Correlation. This technique is utilized for the identification

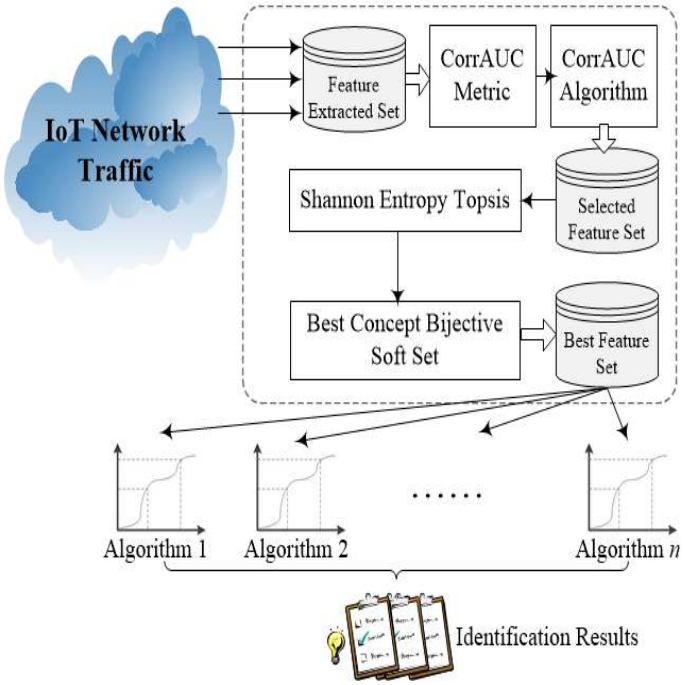

Fig. 1: Proposed Framework for Feature Selection

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.

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

In equation 1, the correlation coefficient is $C_{A,B}$, and (A,B) indicate the covariance. Similarly, ab is the standard deviation for the A and B attributes $\sigma_A \sigma_B$. More in depth, for the two sets of feature equation 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 it's features can be indicated as $a_1, a_2, a_3, ...a_n$ and B can be $b_1, b_2, ...b_n$. 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 equation 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. Pearson Correlation technique is very effective for the ranking and accurate feature selection. Therefore, to overcome the problem of effective and robust features selection for the Bot-IoT malicious attacks detection in IoT network, the correlation

attributes evaluation technique is adopted and applied to rank the effectiveness of features in several given features set. The basic concept of using this ranking attributes is to find out the significance of features set of a dataset by using the correlation between features. Nevertheless, for the Bot-IoT malicious attacks detection in IoT network using machine learning, 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 analyze for accurate detection as follows:

$$Corr = \frac{kavg(corr_{fc})}{\sqrt{k + k(k-1)avg(corr_{ff})}} \tag{3}$$

In the equation Corr indicate the correlation between features and kavg$corr_{fc}$ is indicate to find the average of correlation between features and it's class. Similarly, Avg($corr_{ff}$) indicate 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 features set are strong then it indicate that the correlation between features set and features class is weak. Similarly, if the correlation between the features set and reliant class strong it indicate the strong correlation among the features set and class while if there is more attributes then it indicate strong correlation between features and reliant class.

2) Area under the Curve (AUC) 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 attacks detection in IoT network. Considering this case, a technique named wrapper is applied based on the area under the ROC curve (AUC) metric [24]. Though, for the classification of network traffic by using a machine learning technique, the accuracy metric is the most optimal. But, here we are interested in finding the most significant features 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 features set. However, applying the AUC metric in this research study is two different facts are: If the AUC metric values strong high, then the model will give effective performance results. If the AUC metric values are weak, 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 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 IoT network.

3) Proposed Algorithm: In this section, the proposed algorithms named Corrauc is describe with details step by step, as shown in Fig.2. The proposed Corrauc includes two steps. In step, the algorithm uses a correlation technique to filter the features 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 machine learning algorithm. Similarly, the proposed algorithm selects the useful features which carry enough information for Bot-IoT detection in the IoT network. However, the details step by step phases are discussed as: As discussed in the above lines, the proposed Corrauc algorithm, a hybrid feature selection algorithm based on correlation technique and area under the roc curve metric, used to select feature which has enough information for detection of Bot-IoT attacks in 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 ascending order with respective correlation values. Then the algorithm compared the correlation among each feature. Afterward, a threshold value is assigned, if a feature correlation values are higher than the specified threshold assign value, mean the feature is effective and put forward in descending order. In more detail, the higher the threshold value, the higher the proposed model speed, but it's not effective for the machine learning 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 filter each feature by using the AUC metric of specific ML algorithm. However, the algorithm filters each feature one by one by using AUC metric and select those features which give high AUC metric values for the detection of Bot-IoT attacks in IoT network as well as if the AUC values of feature are low then the algorithm will remove from the list and algorithm will go to next step forward to 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, firstly, 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 decision-making problem and proposed effective decision-making models such as Molodsov in [44] proposed soft set for the decision making and selection attributes from a multiple criteria attributes and then followed by Gong and proposed bijective soft set [45]. Similarly, type-2 soft sets [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 effective feature selection 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, in 2019, [47] applied Shannon entropy weights technique, motivated by this study, we use the same method for the selection of effective features from numerous features.

```
Algorithm 1: feature selection based on correlation
Combined with AUC (Corrauc):
        Input: D (F1, F2, F3,.... Fn)        // training data set,
Output: feature []                           // selected feature set
1.      begin
2.      for i = 1 to N
3.          calculate correlation value corr [i] for each features;
4.      end for
5.      for i = to N;
6.          calculate Corr (Fi);
7.          if (Corr(F) > δ);
8.             Insert Fi into descending order;
9.          end if
10.     end for
11.     Fp = getfirstfeatures (list);
12.     End until (Fp == Null);
13.     X is a data set of samples
        Values of features;
14.     Last _AUC← classify X;
15.     Insert the feature into Swripper;
16.     Feature = get next features;
17.     For feature is not Null
18.             insert the feature into Swrapper;
19.             X is a dataset of sample values for Swrapper;
20.             AUC← classify X with a specific classifier;
21.             if (AUC<= last_AUC)
22.                Remove features from Swrapper;
23.             else
24.                feature = getNextfeature (list, feature);
25.     end if
26.     end for
Return Swrapper;
```

Fig. 2: Proposed Corrauc algorithm

1. Introductory definitions: In this subsection, the introductory definition and basic operations of the soft set are discussed with details.

a) Soft Set [48]: If U is the universal set and S is it's parameter then U be P(U) and X will subset of S, for example, $X \subset S$. At that point, pair (F, X) will be soft set over U, and function F will $F : X \rightarrow P(U)$.

b) Bijective Soft Set: If (F,S) is a soft set and U is the universal set and it's parameters is S respectively then (F,S) is known as Bijective soft set if the below two given condition are true:

i. $\bigcup_{\beta \in s} F(\beta) = U$

ii. For two features;

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

2. Method

Input: Set of features of dataset Output: Desired Selected Effective feature set

a) Identify a features set based on Bot-IoT attacks and normal traffic in 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 $i = 1, ..., a$ and $j = 1, ...b$; ρ_{ij}. M indicate the number of experts, while n indicated numbers of features.

e) In step e the value of projection ($\mathfrak{p}v$), entropy entropy ($\mathcal{E}nt$) , divergence ($\mathcal{D}iv$), and weight ($\mathcal{W}gt$) of each dataset feature $\mathcal{Y}_{ij}$ are calculated respectively [49].
$pv_{ij} = \frac{p_{ij}}{\sum_{i=1}^{a} p_{ij}}$, $\mathfrak{E}nt = -\kappa \sum_{i=1}^{a} pv_{ij} \ln(pv_{ij})$, where κ is a constant implied as, $\kappa = (\ln(a))^{-1}$, then
$\mathfrak{D}iv = 1 - (\mathfrak{E}nt)$, $\mathfrak{W}gt(\gamma_{ij}) = \sum_{\kappa=1}^{n} \frac{\mathfrak{D}iv_j}{\mathfrak{D}iv_\kappa}$.

f) Taking the desired requirement from the network security expert $\mathcal{NER}$ that may give informative feature selection suggestion.

g) In this step the Shannon entropy weight is calculate in soft set form also calculated weight choice value $\mathcal{WCV}$ with respective feature as; $\mathcal{WCV}_{i\kappa} = \sum_j \mathfrak{D}iv_{ij}$, 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 Non-ideal solution (NIS) as γ_i^* and $\check{\gamma_i}$ are calculated for each network expert using TOPSIS as below;
$\gamma_i^* = \mathcal{M}ax(\mathcal{WCV}_{i\kappa})$; $\check{\gamma_i} = \mathcal{M}in(\mathcal{WCV}_{i\kappa})$

i) Computer the separation measure ($\Delta_{i\kappa}^*$, $\check{\Delta}_{i\kappa}$) from the IS and NIS using n-dimensional Euclidean distance for each network expert by using the relation;
$\Delta_{i\kappa}^* = (\gamma_{ij} - \gamma_i^*)^2$, $\check{\Delta}_{i\kappa} = (\gamma_{ij} - \check{\gamma_i})^2$
Then combined separation measure fore each concept and will be as (Δ_κ^*, $\check{\Delta_\kappa}$); (Δ_κ^*, $\check{\Delta_\kappa}$) below;
$\Delta_\kappa^* = \sqrt{\sum_{i=1}^{i=m} \Delta_{i\kappa}^*}$, $\check{\Delta_\kappa} = \sqrt{\sum_{i=1}^{i=m} \check{\Delta}_{i\kappa}}$

j) Calculate the closeness of each feature. $\mathcal{F}\zeta_\kappa$ to IS as;
$\zeta_\kappa^* = \frac{\check{\Delta_\kappa}}{\check{\Delta_\kappa} + \Delta_\kappa^*}$

The most closer measure will be effective feature.

C. Implementation

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

i For the effective feature selection, Bot-IoT attacks detection in Internet of Things 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;
$\mathcal{EFS}_1 = \mathcal{M}ean$, $\mathcal{EFS}_2 = \mathcal{S}tddev$, $\mathcal{EFS}_3 = \mathcal{A}r\_p\_DdtIp$, $\mathcal{EFS}_4 = \mathcal{P}k\_Src\_IP$, $\mathcal{EFS}_5 = \mathcal{P}k\_Dst\_IP$.
We give the following values to attributes with respect to effective feature by ourself identification based on the above given metrics as we denoted by as:
$\mathcal{EFS}_1 = \{\mathcal{Y}_{11}, \mathcal{Y}_{12}, \mathcal{Y}_{13}\} = \{Low, Medium, High\}$
$\mathcal{EFS}_2 = \{\mathcal{Y}_{21}, \mathcal{Y}_{22}, \mathcal{Y}_{23}\} = \{Poor, Good, V.Good\}$
$\mathcal{SFA}_3 = \{\mathcal{Y}_{31}, \mathcal{Y}_{32}, \mathcal{Y}_{33}\} = \{V.Good, Acceptable, Low\}$

$\mathcal{EFS}_4 = \{\mathcal{Y}_{41}, \mathcal{Y}_{42}, \mathcal{Y}_{43}\} = \{V.Good, Acceptable, Low\}$
$\mathcal{EFS}_5 = \{\mathcal{Y}_{51}, \mathcal{Y}_{52}\} = \{Minimum, Maximum\}$

ii In this step the concept for effective features set are generated to form useful combination from $\mathcal{EFS}$ as per 2 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;
$f\mathcal{C}_1 = \{\mathcal{Y}_{11}, \mathcal{Y}_{21}, \mathcal{Y}_{31}, \mathcal{Y}_{42}, \mathcal{Y}_{52}\}$
$f\mathcal{C}_2 = \{\mathcal{Y}_{11}, \mathcal{Y}_{23}, \mathcal{Y}_{33}, \mathcal{Y}_{43}, \mathcal{Y}_{52}\}$
$f\mathcal{C}_3 = \{\mathcal{Y}_{12}, \mathcal{Y}_{21}, \mathcal{Y}_{31}, \mathcal{Y}_{43}, \mathcal{Y}_{51}\}$
$f\mathcal{C}_4 = \{\mathcal{Y}_{13}, \mathcal{Y}_{21}, \mathcal{Y}_{32}, \mathcal{Y}_{42}, \mathcal{Y}_{51}\}$
$f\mathcal{C}_5 = \{\mathcal{Y}_{13}, \mathcal{Y}_{21}, \mathcal{Y}_{31}, \mathcal{Y}_{41}, \mathcal{Y}_{51}\}$

iii For more in depth, we form a soft set through which we can present by using 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 below;
$\mathcal{GG}_1(\mathcal{Y}_{11}) = \{f\mathcal{C}_1, f\mathcal{C}_2\}$; $\mathcal{GG}_1(\mathcal{Y}_{12}) = \{f\mathcal{C}_3\}$;
$\mathcal{GG}_1(\mathcal{Y}_{13}) = \{f\mathcal{C}_4, f\mathcal{C}_5\}$; $\mathcal{GG}_2(\mathcal{Y}_{21}) = \{f\mathcal{C}_1, f\mathcal{C}_4, f\mathcal{C}_5\}$;
$\mathcal{GG}_2(\mathcal{Y}_{22}) = \{f\mathcal{C}_3\}$; $\mathcal{GG}_2(\mathcal{Y}_{23}) = \{f\mathcal{C}_2\}$;
$\mathcal{GG}_3(\mathcal{Y}_{31}) = \{f\mathcal{C}_5\}$; $\mathcal{GG}_3(\mathcal{Y}_{32}) = \{f\mathcal{C}_3, f\mathcal{C}_4\}$;
$\mathcal{GG}_3(\mathcal{Y}_{33}) = \{f\mathcal{C}_1, f\mathcal{C}_2\}$; $\mathcal{GG}_4(\mathcal{Y}_{41}) = \{f\mathcal{C}_5\}$;
$\mathcal{GG}_4(\mathcal{Y}_{42}) = \{f\mathcal{C}_45\}$; $\mathcal{GG}_4(\mathcal{Y}_{23}) = \{f\mathcal{C}_1, f\mathcal{C}_2, f\mathcal{C}_3\}$;
$\mathcal{GG}_5(\mathcal{Y}_{51}) = \{f\mathcal{C}_4, f\mathcal{C}_5\}$; $\mathcal{GG}_5(\mathcal{Y}_{52}) = \{f\mathcal{C}_1, f\mathcal{C}_2, f\mathcal{C}_3\}$;
The above relations are true and satisfy 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 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, $\mathcal{Y}_{11}$, $\mathcal{Y}_{12} \in \mathcal{EFS}_1$, $\mathcal{Y}_{11} \neq \mathcal{Y}_{12}$, $\mathcal{GG}_1(\mathcal{Y}_{11}) \bigcap \mathcal{GG}_1(\mathcal{Y}_{12}) = \varnothing$

iv After applying bijective soft set, preference values are captures as per 4. For the Shannon weight and to show it as $\mathcal{E}PDM$. Network security specialist $\mathcal{N}ER$ assign preference values as shown in Table 1, where;
Low = 0.2; Medium = 0.5; High = 0.7; Very high = 0.9

v The Projection value ($\wp v$), entropy ($\mathcal{E}nt$), divergence ($\mathcal{D}iv$), and weight ($\mathcal{W}gt$) of each dataset feature values $\mathcal{Y}_{ij}$ are calculated as per 5 and shown in Table 2.

vi After step number 5 the requirement from network security expert for effective feature selection we calculate the basic abstract as;
$\mathcal{N}ER_1 = \{\mathcal{Y}_{13}, \mathcal{Y}_{21}, \mathcal{Y}_{31}, \mathcal{Y}_{41}, \mathcal{Y}_{51}\}$
$\mathcal{N}ER_2 = \{\mathcal{Y}_{12}, \mathcal{Y}_{23}, \mathcal{Y}_{31}, \mathcal{Y}_{41}, \mathcal{Y}_{52}\}$
$\mathcal{N}ER_3 = \{\mathcal{Y}_{11}, \mathcal{Y}_{22}, \mathcal{Y}_{32}, \mathcal{Y}_{41}, \mathcal{Y}_{52}\}$

vii The network security experts tabular soft set representation can be shown in in table 3,4 and 5 respectively.

viii The main part of TOPSIS is conducted in this step such as $\mathcal{N}IS$ and $\mathcal{I}S$ are calculated as shown in Table 6.

ix Then the separation measures are computed of each $\Re ER$ from $\mathcal{I}S$ and $\mathcal{N}IS$ as per 9 as shown in Table 7. While combined and afore-computed separations are shown in Table 8 respectively.

x After the calculation of Combined and separate measurements, in this final step the closeness of $\mathcal{F}\zeta$ is calculated as shown in Table 9, 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 effective feature selection the above function concept should be consider for accurate identification as our proposed select the effective feature set.

IV. Evaluation Methodology

In this section, the evaluation criteria and selected dataset for the proposed method are discussed with details. Firstly dataset 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 attacks identification in IoT network environment a new develop dataset [19],[50] is used. The dataset includes on the Internet of Things, and normal traffic flows as well as several numerous cyber-attacks traffic flows of botnets attacks. To trace the accurate traffic and develop effective dataset, the realistic testbed is used for the development of this dataset with effective information features. Similarly, for the improvement of machine learning model performance and effective prediction model, more features were extracted and added with extracted features 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 sub-components as Internet of Things (IoT) services, which are simulated, network 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 current temperature, humidity, and atmospheric pressure. IoT devices or Weather stations. 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 turn off while the fourth one is a smart IoT door. Smart IoT doors are based on probabilistic input. The fifth and final one is 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 a machine learning model result, confusion metrics are widely used, which is base on the measurement of performance. However, the details graphical performance measurement presentation of a confusion matrix is shown in Fig.3. In the graphical presentation of confusion matrix rows indicate the instances of classes, while column shows identified class

TABLE I: Preference Decision Matrix

	$\mathcal{Y}_{11}$	$\mathcal{Y}_{12}$	$\mathcal{Y}_{13}$	$\mathcal{Y}_{21}$	$\mathcal{Y}_{22}$	$\mathcal{Y}_{23}$	$\mathcal{Y}_{31}$	$\mathcal{Y}_{32}$	$\mathcal{Y}_{33}$	$\mathcal{Y}_{41}$	$\mathcal{Y}_{42}$	$\mathcal{Y}_{43}$	$\mathcal{Y}_{51}$	$\mathcal{Y}_{52}$
$\Re ER_1$	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
$\Re ER_2$	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
$\Re ER_3$	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, Weight

	$\mathcal{Y}_{11}$	$\mathcal{Y}_{12}$	$\mathcal{Y}_{13}$	$\mathcal{Y}_{21}$	$\mathcal{Y}_{22}$	$\mathcal{Y}_{23}$	$\mathcal{Y}_{31}$	$\mathcal{Y}_{32}$	$\mathcal{Y}_{33}$	$\mathcal{Y}_{41}$	$\mathcal{Y}_{42}$	$\mathcal{Y}_{43}$	$\mathcal{Y}_{51}$	$\mathcal{Y}_{52}$
$\Re ER_1$	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
$\Re ER_2$	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
$\Re ER_3$	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 $\mathcal{NER}_1$

	$\mathcal{Y}_{13}$	$\mathcal{Y}_{21}$	$\mathcal{Y}_{31}$	$\mathcal{Y}_{41}$	$\mathcal{Y}_{51}$	$\mathcal{WCV}$
$\mathcal{F}\zeta_1$	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 $\mathcal{NER}_2$

	$\mathcal{Y}_{13}$	$\mathcal{Y}_{21}$	$\mathcal{Y}_{31}$	$\mathcal{Y}_{41}$	$\mathcal{Y}_{51}$	$\mathcal{WCV}$
$\mathcal{F}\zeta_1$	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	

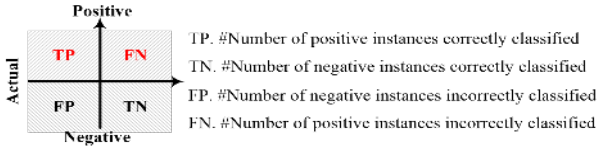

Fig. 3: Confusion Matrix

instances. Nevertheless, the widely used measurement for the evaluation of a machine learning model is discus below as:

- True Positive (TP): In attack detection the TP indicate that Class A is correctly identified as belonging to Class A.
- True Negative (TN): This matrix indicate that Class A is correctly identified as not belonging to Class A.
- False Positive (FP): It indicate that Class A is not correctly identified as belong to Class A.
- False Negative (FN): It indicate that Class A is not correctly identified as not belong to Class A.

However, using the above describe metrics, different measurement metrics can be made to evaluate a machine learning model better. For accurate detection, machine learning classifiers minimize false positive and false negative metrics values. However, the selected metrics that are used in this paper explains with details below:

- Accuracy: In attacks detection, it can be described as the

TABLE V: Soft set representation of $\mathcal{NER}_3$

	$\mathcal{Y}_{13}$	$\mathcal{Y}_{21}$	$\mathcal{Y}_{31}$	$\mathcal{Y}_{41}$	$\mathcal{Y}_{51}$	$\mathcal{WCV}$
$\mathcal{F}\zeta_1$	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: $\mathcal{NIS}$ and $\mathcal{IS}$ for each $\Re ER$

Network	*Expert*	$\mathcal{IS}(\mathcal{Y}_i^*)$	$\mathcal{NIS}(\mathcal{Y}_i^*)$
$\Re ER_1$		0.7768	0
$\Re ER_2$		0.348	0
$\Re ER_3$		0.063	0.010

correctly identified samples of traffic in overall identified samples traffic. However, using performance measurement metrics, the accuracy can be defined mathematically as:

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

In our study, we used equation 4 for the machine learning classifiers performance evaluation. Using this metrics the effectiveness of a ML classifiers can be identified.

- Precision: It can be defined as the correctly identified sample in percentage of Class A in all those were identified in Class A. The mathematically formula used in this research study is shown below.

$$Precission = \frac{TP}{(TP + FP)} \tag{5}$$

- Sensitivity: It can be describe as the correctly detected traffic sample divided by overall dataset traffic sample. However, this metric can be used as a recall metric in Bot-IoT detection in IoT environment. We used the following given mathematical formula for the sensitivity metric as below.

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

- Specificity: In this research study, we used the specificity metrics which can defined as the ability of a machine

TABLE VII: Seperation measure of $\Re ER_1$, $\Re ER_2$, $\Re ER_3$ from $\mathcal{IS}$ and $\mathcal{NIS}$

Functional Concepts	$\Delta^*_{1\kappa}$	$\Delta^{\circ}_{1\kappa}$	$\Delta^*_{2\kappa}$	$\Delta^{\circ}_{2\kappa}$	$\Delta^*_{3\kappa}$	$\Delta^{\circ}_{3\kappa}$
$\mathcal{F}\zeta_1$	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

TABLE VIII: Combined seperation measure

Functional Concepts	Δ^*_{κ}	Δ°_{κ}
$\mathcal{F}\zeta_1$	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 $\mathcal{F}\zeta$

Functional 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

learning classifiers to detect negative results. The mathematical equation of specificity is shown in Eq 7.

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

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 machine learning 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 performance are effective for the Bot-IoT attacks detection in the IoT network environment by using the features set selected by our proposed technique with respective accuracy, precision, sensitivity, and specificity. However, Naive Baye's performance result is low as compare to other machine learning classifiers by using the selected features set concerning accuracy metric for Bot-IoT attack detection. Similarly, the performance result of SVM is slightly higher with respective accuracy as compare 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 others applied ML classifiers. Therefore, the C4.5 ML algorithm performs better by using the selected features 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.

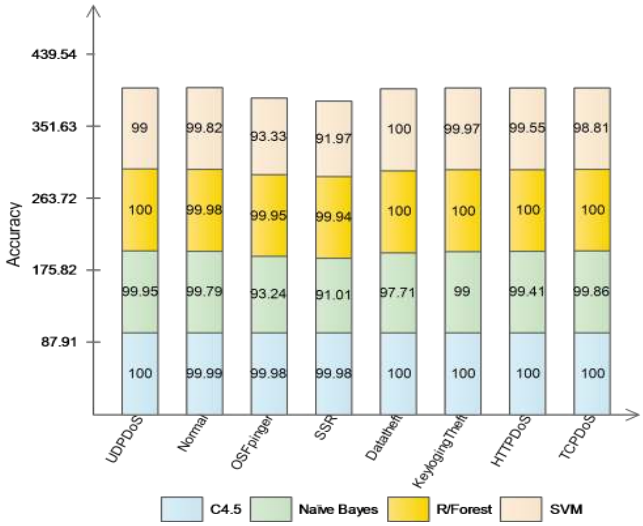

Fig. 4: Accuracy Results

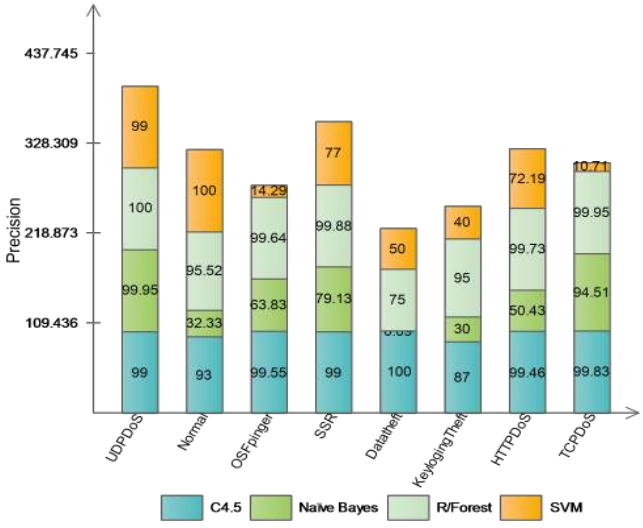

Fig. 5: Precision Results

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

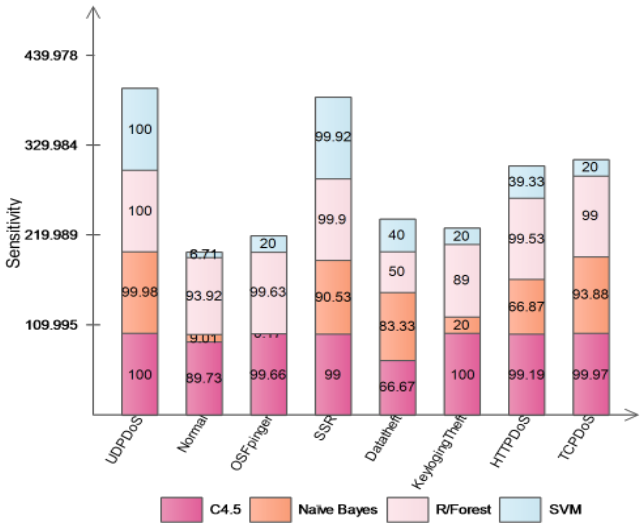

Fig. 6: Sensitivity Results

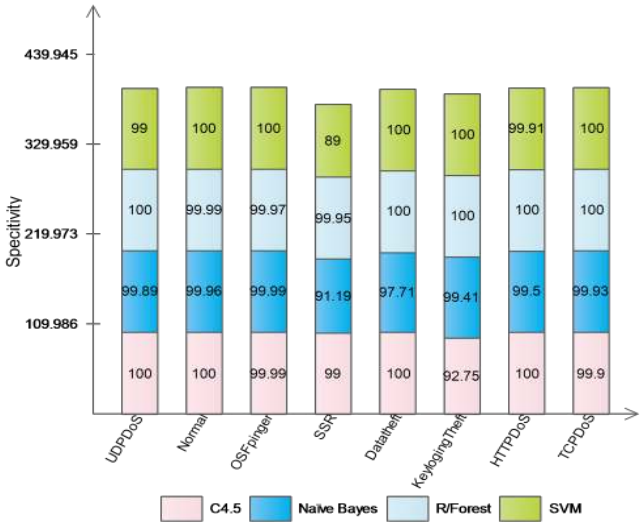

Fig. 7: Specificity Results

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 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 tree machine learning algorithms achieve very high-performance results by using the selected features set as compare 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 compare 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 features 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 features 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 IoT network environment are auspicious by using the selected four different Machine Learning (ML) algorithms with accuracy, precision, sensitivity, and specificity and by utilizing new develop Bot-IoT dataset. However, some useful information that we learned after the experimental analysis are given below.

- In this study, it is clear and evident that the proposed technique is effective for the selection of optimum features in Bot-IoT attacks detection in the IoT network environment by using the newly developed Bot-IoT dataset. For the results analysis and evaluation accuracy, precision, sensitivity, and specificity metrics are used to evaluate the performance of the proposed method accurately.
- 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 IoT network environment.
- In this, it is noticed that applied Machine Learning (ML) algorithms performance are 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.
- 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 dataset. SVM and Naive Bayes ML algorithms performance are also effective, but compared to C4.5 decision tree and Random Forest algorithms, the performance as slightly weak.

VII. Conclusion

Detection of attacks in the Internet of things (IoT) network is essential for the IoT security to keep eyes and block unwanted traffic flows. Numerous machine learning (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 is proposed. Firstly, a novel feature selection metric approach named CorrAUC proposed, and then based on

CorrAUC, a new feature selection algorithm name Corrauc is develop and design, which is based on wrapper technique to filter the feature accurately and select effective features for the selected ML algorithm by using 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 dataset and four different ML algorithms. Experimental results analysis showed that our proposed method is efficient and can achieve >96% results on average.

ACKNOWLEDGMENT

This work is supported by the National Key research and Development Plan (Grant No. 2018YFB0803504), the Guangdong Province Key Research and Development Plan (Grant No. 2019B010137004) and the National Natural Science Foundation of China under Grant No.61871140, and Guangdong Province Universities and Colleges Pearl River Scholar Funded Scheme (2019).

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 of Things Journal*, 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 *International Conference on Advanced Information Networking and Applications*. Springer, 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 Transactions on Industrial Informatics*, 2019.

[5] J. P. Anderson, "Computer security threat monitoring and surveillance, 1980. lastaccessed: Novmeber 30, 2008."

[6] D. E. Denning, "An intrusion-detection model," *IEEE Transactions on software engineering*, no. 2, pp. 222–232, 1987.

[7] X. Du, M. Guizani, Y. Xiao, and H.-H. Chen, "Defending dos attacks on broadcast authentication in wireless sensor networks," in *2008 IEEE International Conference on Communications*. IEEE, 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 *2018 International Conference on Computing, Networking and Communications (ICNC)*. IEEE, 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 of Things Journal*, 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 *2018 15th IEEE Annual Consumer Communications & Networking Conference (CCNC)*. IEEE, 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 Transactions on Industrial Informatics*, 2020. Vol 16(3): 1963-1971.

[12] D. Ventura, D. Casado-Mansilla, J. López-de Armentia, P. Garaizar, D. López-de Ipina, and V. Catania, "Ariima: a real iot implementation of a machine-learning architecture for reducing energy consumption," in *International Conference on Ubiquitous Computing and Ambient Intelligence*. Springer, 2014, pp. 444–451.

[13] R. Xue, L. Wang, and J. Chen, "Using the iot to construct ubiquitous learning environment," in *2011 Second International Conference on Mechanic Automation and Control Engineering*. IEEE, 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 Communications Surveys & Tutorials*, vol. 16, no. 4, pp. 1996–2018, 2014.

[15] M. Shafiq, X. Yu, A. A. Laghari, and D. Wang, "Effective feature selection for 5g im applications traffic classification," *Mobile Information Systems*, vol. 2017, 2017.

[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," *The Journal of Supercomputing*, vol. 74, no. 10, pp. 4867–4892, 2018.

[17] M. Dash and H. Liu, "Feature selection for classification," *Intelligent data analysis*, vol. 1, no. 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," *Computer Communications*, 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," *arXiv preprint arXiv:1811.00701*, 2018.

[20] M. Shafiq and X. Yu, "Effective packet number for 5g im wechat application at early stage traffic classification," *Mobile Information Systems*, vol. 2017, 2017.

[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 *2016 2nd IEEE International Conference on Computer and Communications (ICCC)*. IEEE, 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 *2016 6th International Conference on IT Convergence and Security (ICITCS)*. IEEE, 2016, pp. 1–5.

[23] M. Shafiq, X. Yu, A. A. Laghari, L. Yao, N. K. Karn, F. Abdesssamia *et al.*, "Wechat text and picture messages service flow traffic classification using machine learning technique," in *2016 IEEE 18th International Conference on High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*. IEEE, 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," *Computers & Security*, p. 101863, 2020.

[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 Generation Computer Systems*, vol. 107, pp. 433–442, 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," *Sustainable Cities and Society*, 2020.

[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 Communications Magazine*, vol. 45, no. 11, pp. 126–134, 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 Generation Computer Systems*, vol. 95, pp. 212–220, 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 Transactions on Vehicular Technology*, 2019. 68(6): 5971-5980.

[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," *Computer Communications*, vol. 30, no. 11-12, pp. 2314–2341, 2007.

[31] X. Du and H.-H. Chen, "Security in wireless sensor networks," *IEEE Wireless Communications*, vol. 15, no. 4, pp. 60–66, 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 of Things Journal*, vol. 5, no. 3, pp. 1616–1624, 2018.

[33] Y. Meidan, M. Bohadana, Y. Mathov, Y. Mirsky, A. Shabtai, D. Breitenbacher, and Y. Elovici, "N-baiotâĂŤnetwork-based detection of iot botnet attacks using deep autoencoders," *IEEE Pervasive Computing*, vol. 17, no. 3, pp. 12–22, 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," *Applied Sciences*, 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 of Things Journal, 2019. vol. 6, no. 2, pp. 1584-1593, April.
[36] Z. Tian, W. Shi, Y. Wang, C. Zhu, X. Du, S. Su, Y. Sun, and N. Guizani, "Real time lateral movement detection based on evidence reasoning network for edge computing environment," *IEEE Transactions on Industrial Informatics*, 2019. Vol 15(7): 4285-4294.
[37] X. Du, Y. Xiao, M. Guizani, and H. Chen, "An effective key management scheme for heterogeneous sensor networks," *Ad Hoc Networks*, 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 Communications*, vol. 8, no. 3, pp. 1223–1229, 2009.
[39] X. Du, M. Zhang, K. E. Nygard, S. Guizani, and H.-H. Chen, "Self-healing sensor networks with distributed decision making," *International Journal of Sensor Networks*, vol. 2, no. 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 *2001 IEEE/IFIP International Symposium on Integrated Network Management Proceedings. Integrated Network Management VII. Integrated Management Strategies for the New Millennium (Cat. No. 01EX470)*. IEEE, 2001, pp. 453–466.
[41] X. Huang and X. Du, "Achieving big data privacy via hybrid cloud," in *2014 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, 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," *International Journal of Parallel Programming*, vol. 44, no. 1, pp. 181–197, 2016.
[44] D. Molodtsov, "Soft set theoryâĂŤfirst results," *Computers & Mathematics with Applications*, vol. 37, no. 4-5, pp. 19–31, 1999.
[45] K. Gong, Z. Xiao, and X. Zhang, "The bijective soft set with its operations," *Computers & Mathematics with Applications*, 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," *Advances in Fuzzy Systems*, vol. 2017, 2017.
[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," *Journal of Intelligent Manufacturing*, 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," *Journal of Computational and Applied Mathematics*, 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 systems with applications*, vol. 36, no. 5, pp. 8980–8985, 2009.
[50] I. Van der Elzen and J. van Heugten, "Techniques for detecting compromised iot devices," *University of Amsterdam*, 2017.

Muhammad Shafiq was born in Pakistan. He received the B.S. degree with honor rank in computer science from the Faculty of Computer Science, Malakand University, Chakdara, Pakistan, in 2009, and the M.S. degree in computer science from Faculty of Computer Science, Malakand University, Chakdara, Pakistan, in 2011. He is received the Ph.D. degree at the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China, in 2018. He is currently pursuing the Post-Doctorate. at 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, Cloud Computing.

Zhihong Tian is currently a Professor, and Dean, with the Cyberspace Institute of Advanced Technology, Guangzhou University, Guangdong Province, China. Guangdong Province Universities and Colleges Pearl River Scholar (Distinguished Professor). He is also a part-time Professor at Carlton University, Ottawa, Canada. Previously, he served in different academic and administrative positions at the Harbin Institute of Technology. He has authored over 200 journal and conference papers in these areas. His research interests include computer networks and cyberspace security. His research has been supported in part by the National Natural Science Foundation of China, National Key research and Development Plan of China, National High-tech R&D Program of China (863 Program), and National Basic Research Program of China (973 Program). He also served as a member, Chair, and General Chair of a number of international conferences. He is a Senior Member of the China Computer Federation, and a Member of IEEE.

Ali Kashif Bashir is a Senior Lecturer at the Department of Computing and Mathematics, Manchester Metropolitan University, United Kingdom. His past assignments include Associate Professor of Information and Communication Technologies, Faculty of Science and Technology, University of the Faroe Islands, Denmark; Osaka University, Japan; Nara National College of Technology, Japan; the National Fusion Research Institute, South Korea; Southern Power Company Ltd., South Korea, and the Seoul Metropolitan Government, South Korea. He received his Ph.D. in computer science and engineering from Korea University, South Korea. MS from Ajou University, South Korea and BS from University of Management and Technology, Pakistan. He is supervising/co-supervising several graduate (MS and Ph.D.) students. His research interests include internet of things, wireless networks, distributed systems, network/cybersecurity, network function virtualization, etc. He has authored over 80 peer-reviewed articles. He has served as a chair (program, publicity, and track) on top conferences and workshops. He has delivered over 20 invited and keynote talks in seven countries. He is a Distinguished Speaker, ACM; Senior Member of IEEE; Member, ACM; Member, IEEE Young Professionals; Member, International Association of Educators and Researchers, UK. He is serving as the Editor-in-chief of 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 (S'99–M'03–SM'09–F'20) received his B.S. and M.S. degree in Electrical Engineering (Automation Department) from Tsinghua University, Beijing, China in 1996 and 1998, respectively. He received his M.S. and Ph.D. degree in Electrical Engineering from the University of Maryland, College Park in 2002 and 2003, respectively. Dr. Du is a tenured Full Professor and the Director of the Security And Networking (SAN) Lab in the Department of Computer and Information Sciences at Temple University, Philadelphia, USA. His research interests are security, wireless networks, and systems. He has authored over 400 journal and conference papers in these areas, as well as a book published by Springer. Dr. Du has been awarded more than 6 million US Dollars research grants from the US National Science Foundation (NSF), Army Research Office, Air Force Research Lab, 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 the ACM MobiHoc 2014. He serves on the editorial boards of two international journals. Dr. Du served as the lead Chair of the Communication and Information Security Symposium of the IEEE International Communication Conference (ICC) 2015, and a Co-Chair of Mobile and Wireless Networks Track of IEEE Wireless Communications and Networking Conference (WCNC) 2015. He is (was) a Technical Program Committee (TPC) member of several premier ACM/IEEE conferences such as INFOCOM (2007 - 2020), IM, NOMS, ICC, GLOBECOM, WCNC, BroadNet, and IPCCC. Dr. Du is an IEEE Fellow and a Life Member of ACM.

Mohsen Guizani received the B.S. (with 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 at the Computer Science and Engineering Department in Qatar University, Qatar. His research interests include wireless communications and mobile computing, computer networks, mobile cloud computing, security, and smart grid. Throughout his career, he received three teaching awards and four research awards. He also received the 2017 IEEE Communications Society WTC Recognition Award as well as 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 Chair of the TAOS Technical Committee. He served as the IEEE Computer Society Distinguished Speaker and is currently the IEEE ComSoc Distinguished Lecturer. He is a Fellow of IEEE and a Senior Member of ACM.