IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 25, 2019, accepted April 16, 2019, date of publication April 18, 2019, date of current version April 30, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2912115

Network Anomaly Intrusion Detection Using a Nonparametric Bayesian Approach and Feature Selection

WAJDI ALHAKAMI[1], ABDULLAH ALHARBI[1], SAMI BOUROUIS[1,2], ROOBAEA ALROOBAEA[1], AND NIZAR BOUGUILA[3], (Senior Member, IEEE)

[1]College of Computers and Information Technology, Taif University, Taif 21431, Saudi Arabia
[2]LR-SITI Laboratoire Signal Image et Technologies de l'Information, Université de Tunis El Manar, Tunis 1002, Tunisia
[3]Concordia Institute for Information Systems Engineering (CIISE), Concordia University, Montreal, QC H3G 1T7, Canada

Corresponding author: Nizar Bouguila (nizar.bouguila@concordia.ca)

This work was supported by the Deanship of Scientific Research at Taif University, Saudi Arabia, under Grant 1-439-6115.

ABSTRACT Anomaly-based intrusion detection systems (IDSs) have been deployed to monitor network activity and to protect systems and the Internet of Things (IoT) devices from attacks (or intrusions). The problem with these systems is that they generate a huge amount of inappropriate false alarms whenever abnormal activities are detected and they are not too flexible for a complex environment. The high-level rate of the generated false alarms reduces the performance of IDS against cyber-attacks and makes the tasks of the security analyst particularly difficult and the management of intrusion detection process computationally expensive. We study here one of the challenging aspects of computer and network security and we propose to build a detection model for both known and unknown intrusions (or anomaly detection) via a novel nonparametric Bayesian model. The design of our framework can be extended easily to be adequate for IoT technology and notably for intelligent smart city web-based applications. In our method, we learn the patterns of the activities (both normal and anomalous) through a Bayesian-based MCMC inference for infinite bounded generalized Gaussian mixture models. Contrary to classic clustering methods, our approach does not need to specify the number of clusters, takes into consideration the uncertainty via the introduction of prior knowledge for the parameters of the model, and permits to solve problems related to over- and under-fitting. In order to get better clustering performance, feature weights, model's parameters, and the number of clusters are estimated simultaneously and automatically. The developed approach was evaluated using popular data sets. The obtained results demonstrate the efficiency of our approach in detecting various attacks.

INDEX TERMS Intrusion detection systems (IDS), anomaly intrusion detection, infinite mixture models, bounded generalized Gaussian models, Bayesian inference, Markov chain Monte Carlo (MCMC).

I. INTRODUCTION

Cyber-security systems are broadly used to protect information and computers from attack, destruction, and unauthorized access. In particular, intrusion detection systems (IDS) have been proposed as an effective tool to monitor network activity, to help in determining unauthorized use, to identify information systems destruction, and to protect systems from internal and external intrusions (intrusions from within or from outside the firm). On the other hand, IDS can be considered as one of the most significant security solutions for new online web-based applications related to smart city and Internet of Things (IoT) environment. Indeed, IDSs have attracted recently the attention of security specialists to protect IoT networks, devices application domains such as smart homes/cities, and health monitoring [1]. Old IDS-based solutions are usually operated by external providers which can cause difficulties in term of security management for smart city managers, since these solutions are implemented by different technologies (protocols, devices, etc.) which may result in an extremely varied environment. However, IDS-based systems generate generally a huge amount of inappropriate and false alarms whenever abnormal activities are detected. The high level

The associate editor coordinating the review of this manuscript and approving it for publication was Jorge Parra.

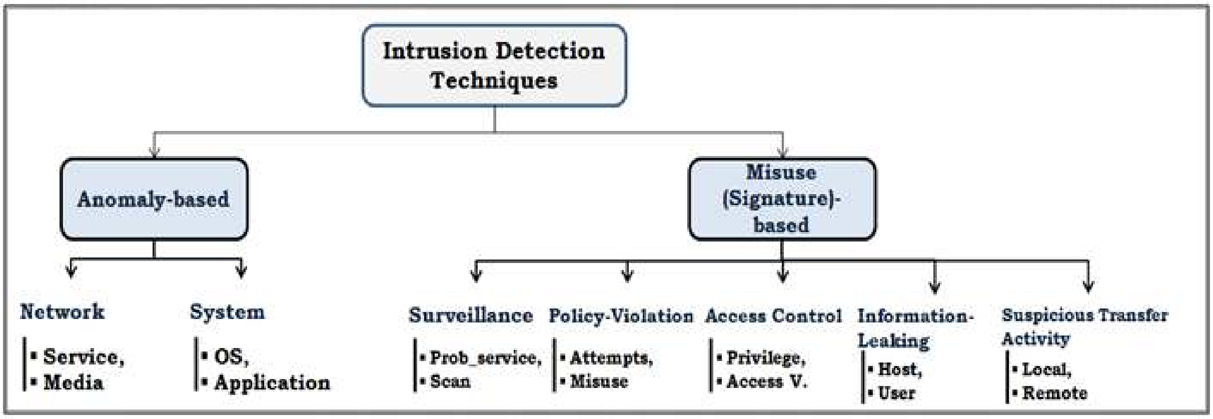

FIGURE 1. Classification of intrusion detection techniques.

rate of the generated false alarms reduces the performance of IDS against cyber-attacks and makes the tasks of the security analyst particularly difficult and the management of intrusion detection process computationally expensive. Moreover, using traditional IDS-based methods to IoT can fail and will not authorize the rapid expansion of smart city applications given the specifications and constraints related of IoT devices and networks such as the specific used protocol stacks, standards, and constrained-resource devices. These problems have received considerable attention from researchers in computer and network security community and so the designing of modern and smart IDS-based solutions, which represents an important challenge, is of high priority for users, security researchers, manufacturers, and IoT infrastructures. Therefore, the design of more robust information security systems will authorize especially the rapid expansion of smart city applications and IoT technology. New systems should be the basis for providing resilient services and reinforcing smart city applications.

A possible division of existing IDS is based often on which cyber analytic technique is used: misuse-based (signature-based) or anomaly-based [2]. Figure 1 shows an overview of main existing approaches. Misuse (Signature)-based detection [3] has been proven effective to detect only known attacks and they are used in the several commercial tools. In this case, signatures (patterns) are created for known attacks and stored as a prior knowledge into specific databases. These databases are then used to verify if the current activity matches a known pattern, indicating the presence of an anomaly. The problem with this approach is it cannot identify unknown attacks since they are not saved into the datasets; for that reason they must be regularly updated with new attack's signatures.

Anomaly-based detection [4] use generally models (e.g., statistical profiles) for supervising normal activities. If an activity deviates from the normal behavior, the administrator will be informed about this anomalous traffic to take suitable actions for them. On the other hand, if the constructed model is not well-defined, we can have a lot of false alerts. The challenge for such approach is related to the specification of the normal network behavior and a threshold that can prevent false alerts. The main advantages of such approach are the ability to identify unusual behavior and to detect attacks without any prior knowledge. But, in practice, a large number of false alerts are produced due to small "training sets" that characterize normal behavior. Techniques driven from this approach are essentially thresholding-based methods, statistical-based methods (parametric, non-parametric), rule-based methods, and machine learning-based methods (neural networks, genetic algorithms, hidden Markov model, etc.).

II. RELATED WORKS

Due to the importance of this area of research, complete and extensive surveys have been provided in [5]–[8] where interested readers can found sophisticated descriptions of various techniques for cyber intrusion detection. In particular, some promising soft computing, data mining and machine learning-based techniques have been proposed in the literature for intrusion detection [2]. For performance evaluation, most researchers investigate some well-known recorded benchmarks such as the KDDCup'99 dataset.[1] Among the most relevant approaches from the state of the art that can achieve good results in term of low false alarm rate and high accuracy, we can cite for example the recent work published in [9]. In this work, soft computing techniques like fuzzy logic and genetic algorithm are considered to deal with imprecision and uncertainties and are employed to make a decision if an instance contains an anomaly or not. A hybrid approach is developed in [10], where some algorithms are combined into the same multi-level formalism including a modified k-means as a pre-processing step (for training the dataset), SVM classifier and an extreme learning machine method. In [11], an unsupervised alternative of the k-means clustering method to detect anomalous behaviors is also proposed. In fact, a data point is considered as an anomaly if and only if it is located so far from the cluster's center. SVM is also used as an online discriminative learning classifier to deal with high-speed

[1] *http : //kdd.ics.uci.edu/databases/kddcup99/kddcup99.html*

network's requirements and to reduce the huge number of false alerts in [12]. The proposed online self-trained classifier, which is based on different modes of learning of large scale datasets, is validated on three different datasets: virtualized (KDDCup'99), realistic (Kyoto 2006+), and synthesized (ISCX) datasets. Another clustering approach based on fuzzy association rules is developed to determine natural relationship patterns [13]. The mechanism of alert correlation is also adopted in this context via Naive Bayes classifier to detect complex attacks [14]. Indeed, correlation between attack plans (anomalies) is performed via expert knowledge, clustering techniques and similarity measures. Other issues in anomaly detection problem are also addressed in order to diminish the high volume of false alerts generated by systems, to uncover complex patterns of possible attacks, and to reduce computational issues. One of these issues is the online and real-time attacks detection which is investigated via some clustering approaches like in [15], [16]. In particular, anomalies can be identified in an online fashion using a genetic weighted K-nearest-neighbor (K-NN) based classifier [16]. In [17], authors studied and compared the performances of various anomaly detection-based methods such as the k-medoids, k-means, EM and distance-based techniques. Another important issue is determining the boundaries (borders) between known and unknown intrusions classes which is tackled by [18]. Indeed, authors designed an artificial algorithm allowing generating anomalies so as to find a precise border between normal known clusters and anomalies ones. This strategy has the advantage to not provide in advance the abnormal data type but to discover boundaries between the two classes, and also to increase the detection performance.

According to the literature review, some pertinent intrusion detection methods (most significant and related to our focus in this work) have been based on clustering paradigm given that they are based on totally unlabeled traffic data. A lot of research is going on for improving the output of anomaly intrusion detection methodologies while the research on this hot area needs more flexible and powerful models to achieve accurate results. Most of existing instance-based learning techniques can only be applied to identify known cyber attacks and rarely discover new ones because they are not learned before. Clustering is an essential problem used to accurately model data in order to help analysts for taking appropriate and automatic actions (decisions). In particular, unsupervised clustering is a very useful tool for attack's patterns discovery and high-dimensional data grouping on the basis of some criteria. Its application for intrusion detection is highly desirable in order to achieve good performance in term of detection rate, false positive rate, and accuracy.

It is noteworthy that there are various popular clustering methods such as: k-means, k-nearest neighbors, hierarchical clustering, mixture models, density-based models, expectation maximization (EM) algorithm, graph models, etc. Some of these algorithms have been applied with success for cyber security [19]–[21] but others are not well-defined and so fail in achieving high accuracy. Thus, it is important to develop more powerful clustering-based method able to detect accurately abnormal activities, to represent them in a compact form, and to well-describe normal/abnormal attack patterns in order to reduce as much as possible false positive alarms and to provide early warnings against cyber-intrusions.

III. MOTIVATIONS

In recent years some attractive machine learning-based techniques have been proposed to address the aforementioned issues and especially to deal with complex patterns in order to take correct decisions while considering into account the observed data. In particular, the named "finite mixture models (FMM)" [22], [23] have been developed as a powerful machine learning tool to solve the problem of complex data clustering and modeling in a formal way [24]–[26]. Even though finite unbounded mixtures (e.g. unbounded Gaussian) have been extensively used in data analysis thanks to their approximation properties, other mixtures such as the bounded generalized Gaussian (BGG) mixture have been revealed to offer more flexibility in modeling data and can be an interesting alternative for data clustering and especially anomaly detection. It should be noted also that, within finite mixture models, the most difficult problem is selecting the optimal number of data clusters in order to avoid under and overfitting. This complication can be solved by extending finite models to infinite mixtures [27]–[29]. On the other hand, data-features can be either informative (relevant) or uninformative (irrelevant). Considering all possible features will augment the computational cost and becomes an obstacle against high performance as cited in [30]–[34]. In fact, the presence of irrelevant features can form new false clusters and this issue may lead to raise the false positive intrusion detection rate and make the overall process time consuming. It should be noted that several anomaly detection techniques ignored this step of feature selection or solved it independently as a pre-processing step. However, it is extremely imperative to found an intelligent way to enable the removal of uninformative features during the clustering process. Here, we are motivated by the issue of selecting and weighting automatically and simultaneously the most informative data-features. This step is very important since it will increase the flexibility and capability of the developed algorithm and also reduce the computational cost.

Based on all these assumptions, we are mainly motivated by developing a new fully Bayesian-based approach for infinite bounded Generalized Gaussian mixture model (InBGG) for anomaly-based IDS detection problem in general. Then, in the future, we plan to adapt and extend the developed framework to be useful for specific problems related to IoT and smart cities-based security. The main advantages of using Bayesian statistical methodology are to avoid under and overfitting, to formalize our prior knowledge and to express our uncertainty through probability as cited in [35], [36]. On the other hand, the major benefit in using the infinite assumption instead of finite one is that the problem of determining the correct number of mixture components can

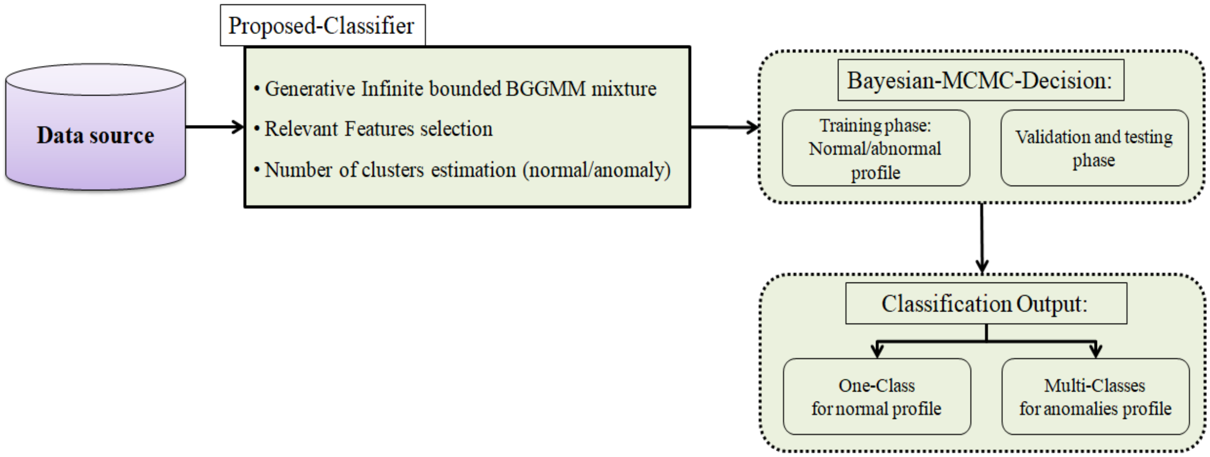
Data source
Proposed-Classifier
• Generative Infinite bounded BGGMM mixture
• Relevant Features selection
• Number of clusters estimation (normal/anomaly)
Bayesian-MCMC-Decision:
Training phase: Normal/abnormal profile
Validation and testing phase
Classification Output:
One-Class for normal profile
Multi-Classes for anomalies profile

FIGURE 2. Proposed Bayesian-based framework for intrusion data classification.

be handled and solved. The proposed framework has also the advantage to take into account estimating -at the same time- feature weights with model's parameters in a closed-form which can definitely increase the performance of the anomaly intrusion detection especially in the context of smart city web-based applications. The efficiency of our proposed method is confirmed by testing it on anomaly intrusions detection application, while comparing it to other published methods from the literature. To the best of our knowledge, very little works have been implemented to date while considering flexible statistical models and unsupervised feature selection mechanism simultaneously.

In the next section we present the infinite mixture model with feature selection as well as a fully Bayesian learning approach. Section 4 is dedicated to the experimental results. Finally, we end this work with a conclusion in section 5.

IV. INFINITE BOUNDED MIXTURE MODEL WITH FEATURES SELECTION

In this work, we address the classification problem of intrusion anomaly by constructing a statistical Bayesian model that allows grouping network traffic behaviors into a Multi-class anomaly (i.e., a set of categories: one for normal and others for different type of attacks) and not as One-class anomaly (i.e., two categories: normal or attack). It is noteworthy that the Multi-class anomaly approaches are more interesting than One-class anomaly methods especially if we want to recognize more different attack types. We solve this problem by developing an infinite mixture model with feature selection as well as a fully Bayesian learning approach - as depicted in figure 2- that we will present in the next subsections.

A. THE FINITE BOUNDED MIXTURE WITH FEATURES SELECTION: (FIGG-FS)

The bounded mixture models are proposed in order to solve the problem of unbounded distributions with support range $(-\infty, +\infty)$. In particular, the finite bounded generalized Gaussian mixture model (FiBGG) can be seen as an extensible model to the unbounded (FiGG) which is able to fit compactly supported data. Let $\mathcal{Y} = \{\vec{Y}_1, \vec{Y}_2, \ldots, \vec{Y}_N\}$, be a set of D-dimensional vectors where $\vec{Y}_n = (\vec{Y}_{n1}, \ldots, \vec{Y}_{nD})$, $n = 1, \ldots, N$, be the observed data from a M-component mixture distribution. A crucial problem when deploying finite mixture models is the choice of the per-components distributions. In this work, we assume that each of these vectors is generated from a FiBGG with M components which gives the following likelihood:

$$p(\mathcal{Y}|\Theta) = \prod_{n=1}^{N} \sum_{j=1}^{M} \pi_j BGG(\vec{Y}_n|\theta_j) \tag{1}$$

where $\Theta = (\vec{\pi}, \theta)$, the $\{\pi_j\}$'s are the mixing parameters. $\pi_j > 0$ and $\sum_j \pi_j = 1$. $\vec{\pi} = (\pi_1, \ldots, \pi_M)$, and the $\theta = \{\theta_j\} = \{\mu_1, ..., \mu_M, \sigma_1, ..., \sigma_M, \lambda_1, ..., \lambda_M\}$ are vectors containing the model's parameters (mean, variance and shape). Each $\vec{Y}_n$ is supposed to be drawn from one of the M-components (clusters), but the cluster memberships are not known and they should be determined. Indeed, $\pi_j = p(Z_n = j)$, where Z_i indicates from which cluster each vector $\vec{Y}_n$ arose. In our case, $Z_n = j$ means that $\vec{Y}_n$ comes from component j and according to Bayes's theorem, we have

$$p(Z_n = j|\vec{Y}_n) \propto \pi_j BGG(\vec{Y}_n|\theta_j) \tag{2}$$

The BGG is defined by:

$$BGG(\vec{Y}_i|\theta_j) = \prod_{l=1}^{D} BGG(\vec{Y}_{il}|\theta_{jl}) = \prod_{l=1}^{D} \frac{p(\vec{Y}_{il}|\theta_{jl})H(\vec{Y}_{il}|\Omega_j)}{\int_{\delta_j} p(\vec{Y}_{il}|\theta_{jl})dy} \tag{3}$$

where $p(\vec{Y}_{il}|\theta_{jl})$ represents the generalized Gaussian distribution of the lth feature in the component j. It is defined as follows:

$$\begin{aligned} p(\vec{Y}_{il}|\theta_{jl}) &= p(\vec{Y}_{il}|\vec{\mu}_{jl}, \vec{\sigma}_{jl}, \vec{\lambda}_{jl}) \\ &= A(\lambda_{jl})exp\left[-B(\lambda_{jl})\left|\frac{X_{il} - \mu_{jl}}{\sigma_{jl}}\right|^{\lambda_{jl}}\right] \end{aligned} \tag{4}$$

where $A(\lambda_{jl}) = \frac{\lambda_{jl}\left[\frac{\Gamma(3/\lambda_{jl})}{\Gamma(1/\lambda_{jl})}\right]^{1/2}}{2\sigma_{jl}\Gamma(1/\lambda_{jl})}$; $B(\lambda_{jl}) = \left[\frac{\Gamma(3/\lambda_{jl})}{\Gamma(1/\lambda_{jl})}\right]^{\lambda_{jl}/2}$; $\Gamma(.)$ is the Gamma function defined as: $\Gamma(t) = \int_0^\infty u^{t-1}e^{-u}dt$.

B. THE INFINITE BOUNDED MIXTURE WITH FEATURES SELECTION: (INBGG-FS)

So far, we have assumed that M has a fixed value. In this section, we present the infinite bounded generalized Gaussian (BGG) mixture model with feature selection by considering that $M \to \infty$ which provides us:

$$p(\vec{Y}_i|\Theta) = \sum_{j=1}^{\infty} \pi_j BGG(\vec{Y}_i|\theta_j) = \sum_{j=1}^{\infty} \pi_j \prod_{l=1}^{D} BGG(\vec{Y}_{il}|\theta_{jl}) \tag{5}$$

In our case, we adopt also an unsupervised learning scheme for feature selection as cited in [37]. Indeed, the lth feature is considered irrelevant if it follows a common density, that is, if its distribution is independent of the class labels and follows a BGG distribution: $BGG(\vec{Y}_{il}|\varphi_l)$, where $\varphi_l = (\mu_l^{irr}, \sigma_l^{irr}, \lambda_l^{irr})$. Let $\vec{\phi} = (\phi_1, \dots, \phi_D)$ be a set of binary parameters and known as the feature relevance indicator, where: $\phi_l = \begin{cases} 0 & \text{when then } l\text{th feature is irrelevant (i.e. noise)} \\ 1 & \text{Otherwise} \end{cases}$

Based on the assumption given in [38], our distribution can be rewritten as:

$$BGG(\vec{Y}_{il}|\theta_{jl}, \phi_l, \varphi_l) \simeq \left[BGG(\vec{Y}_{il}|\theta_{jl})\right]^{\phi_l}\left[BGG(\vec{Y}_{il}|\varphi_l)\right]^{1-\phi_l} \tag{6}$$

And in this case the mixture density will be expressed as:

$$p(\vec{Y}_i|\xi) = \sum_{j=1}^{\infty} \pi_j \prod_{l=1}^{D} [\rho_l BGG(\vec{Y}_{il}|\theta_{jl}) + (1-\rho_l)BGG(\vec{Y}_{il}|\varphi_l)] \tag{7}$$

where $\rho_l = p(\phi_l = 1)$ represents the probability that the lth feature is relevant for the clustering task, and ξ represents the set of all parameters $\xi = \{\Theta, \{\phi_l\}\}$.

C. BAYESIAN LEARNING

In this section, we develop a Bayesian inference framework for learning (i.e. model selection, feature selection and parameters estimation) the parameters of our infinite mixture model IBGGM. It is noteworthy that for Bayesian learning, prior distributions are defined over all the model's parameters, and the posteriors are used for the inference. By adopting a Dirichlet prior, with equal parameters η, over the mixing weights and letting the number of components goes to infinity, we obtain the following conditional posterior [39]

$$p(Z_n = j|\eta, Z_{-n}) = \begin{cases} \frac{a_{-n,j}}{N-1+\eta} & \text{if } a_{-n,j} > 0 \text{ (cluster } j \in \mathcal{R}) \\ \frac{\eta}{N-1+\eta} & \text{if } a_{-n,j} = 0 \text{ (cluster } j \in \mathcal{U}) \end{cases} \tag{8}$$

where $\mathcal{R}$ and $\mathcal{U}$ are the sets of represented and unrepresented clusters, respectively, $Z_{-n} = \{Z_1, \dots, Z_{n-1}, Z_{n+1}, \dots, Z_N\}$, $a_{-n,j}$ is the number of observations, excluding $\vec{Y}_n$, in cluster j. The previous equation describe in fact a Dirichlet process and is very important in our intrusion detection problem especially in the case of unknown attacks. Indeed, if a given attack nature is unknown it creates a new cluster.

Our learning Bayesian inference framework is based on estimating the posterior distribution of the mixture model using Markov Chain Monte Carlo (MCMC) techniques. In particular, we consider Gibbs sampling which allows to update each model's parameter, using its posterior, in turn given the rest of all parameters in the model.

In order to obtain our posteriors, we consider the following priors for the parameters of the distributions representing the relevant features and irrelevant features

$$\mu_{jl} \sim \mathcal{N}(\mu_0, \sigma_0^2), \quad \sigma_{jl} \sim \mathcal{G}(\alpha_\sigma, \beta_\sigma), \ \lambda_{jl} \sim \mathcal{G}(\alpha_\lambda, \beta_\lambda)$$
$$\mu_l^{irr} \sim \mathcal{N}(\mu_0, \sigma_0^2), \quad \sigma_l^{irr} \sim \mathcal{G}(\alpha_\sigma, \beta_\sigma), \ \lambda_l^{irr} \sim \mathcal{G}(\alpha_\lambda, \beta_\lambda)$$

where $\mathcal{N}(\mu_0, \sigma_0^2)$ is a bounded normal distribution, defined in the same bounded region as the mixture model, with mean μ_0 and variance σ_0^2, $\mathcal{G}(\alpha_\sigma, \beta_\sigma)$ is a Gamma distribution with shape parameter α_σ and rate parameter β_σ, and $\mathcal{G}(\alpha_\lambda, \beta_\lambda)$ is a Gamma distribution with shape parameter α_λ and rate parameter β_λ. It is noteworthy that $\mu_0, \sigma_0^2, \alpha_\sigma, \beta_\sigma, \alpha_\lambda$, and β_λ are called the model's hyperparameters. Then, it is straightforward to obtain all parameters posteriors by multiplying the chosen priors by the complete model's likelihood.

Concerning the weights ρ_l, we know that they are defined in [0,1], then a good prior would be a Beta distribution with parameters δ_1 and δ_2, common to all dimensions:

$$p(\rho_l|\delta_1, \delta_2) = \left[\frac{\Gamma(\delta_1)\Gamma(\delta_2)}{\Gamma(\delta_1+\delta_2)}\right]\rho_l^{\delta_1-1}(1-\rho_l)^{\delta_2-1} \tag{9}$$

We know that $\rho_l = p(\phi_l = 1)$ and $1 - \rho_l = p(\phi_l = 0)$, $l = 1, \dots, D$, thus ϕ_l follows a Bernoulli distribution and we have

$$p(\phi_l|\rho_l) = \rho_l^{\phi_l}(1-\rho_l)^{1-\phi_l} \tag{10}$$

Then, the posterior for ρ_l is given by

$$p(\rho_l|\dots) \propto p(\rho_l|\delta_1, \delta_2)p(\phi_l|\rho_l) \propto \rho_l^{\delta_1+\phi_l-1}(1-\rho_l)^{\delta_2-\phi_l} \tag{11}$$

An important part when dealing with infinite mixture models is the posterior of the membership variables Z_n which will allow to assign a new observed vector to an existing cluster or to force the creation of a new cluster. Having the conditional priors in Eq. 8, the conditional posteriors are obtained by combining these priors with the likelihood of the data [40], [41]

$$p(Z_n = j|\dots) = \begin{cases} \frac{a_{-n,j}}{N-1+\eta} p(\vec{Y}_n|\theta_j, \{\varphi_l\}, Z_n) & \text{if } j \in \mathcal{R} \\ \int \frac{\eta p(\vec{Y}_n|\theta_j, \{\varphi_l\}, Z_n)p(\theta_j, \{\varphi_l\})}{N-1+\eta} d\theta_j d\varphi_l & \text{if } j \in \mathcal{U} \end{cases} \tag{12}$$

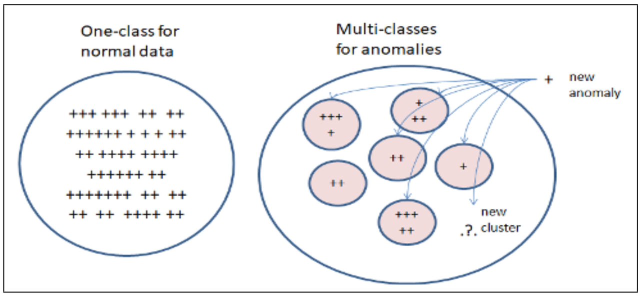

FIGURE 3. Proposed classification strategy of intrusion data.

Having all the posteriors, we can employ a Gibbs sampler, as the one developed in [27], and each iteration can be summarized by the following steps:

- Generate Z_n from Eq. 12 and then update a_j, $j = 1, \ldots, M, n = 1, \ldots, N$.
- Update the number of represented components M.
- $p_j = \frac{a_j}{N+\eta}, j = 1, \ldots, M$ and the mixing parameters of unrepresented components $p_U = \frac{\eta}{\eta+N}$.
- Generate ρ_l from Eq. 11, $l = 1, \ldots, D$.
- Generate θ_j and $\varphi_l, j = 1, \ldots, M, l = 1, \ldots, D$, from their respective posteriors.

V. EXPERIMENTAL RESULTS

The objective of this section is to conduct experiments in order to investigate the performance of our statistical framework applied to the anomaly intrusion detection challenging problem and test it against comparable approaches. We adopt here the multi-class anomaly with a different set of classes: one for normal intrusion and several others for anomalies. The output of the proposed classification strategy is illustrated as in 3. Indeed, any new incoming data will be classified as a normal attack or anomalous one. If it is an anomaly attack, it will be assigned to one of the known existing classes or to a new cluster. For evaluation, we have considered different old and new challenging publicly available datasets such as the KDD Cup'99,[2] the Kyoto 2006+,[3] and the ISCX that we describe in the following subsection.

A. DATASETS

1) KDDCup'99 DATASET [42]

KDDCup'99 is one of the most widely employed benchmark for the performance evaluation of network-based intrusion detection systems. It was produced by simulation over virtual network and is built on the basis of a subset data taken from DARPA'98 program. The KDDCup'99 contains 4,898,431 records, and each consists of 41 features which can be classified as normal or malicious attacks. In total the data set has 5 classes, one "Normal" and four attack classes which are: "Denial of Service attack (DoS)", "User to root attack (U2R)", "Probing attack (Probe)", and "Remote to local (R2L)". From the whole KDDCup'99 dataset, only 10% are designed for training purpose, and the rest for testing. It is noteworthy that KDDCup'99 dataset presents many issues such as the fact that its data cannot reflect real traffic and it cannot represent up to date network traffic since it was produced by simulation. Moreover, the normal and attack's behaviors are too different from up to date network traffic and it contains a lot of redundant records which could lead to erroneous results. To deal with the drawbacks of KDDCup'99, some researchers proposed new ones such as the ISCX developed in [43].

2) KYOTO 2006+ DATASET [44]

The Kyoto 2006+ dataset is a set of real traffic data obtained from different honeypots which can directly capture and analyze the network traffic and this process was done from 2006 to 2009 by Kyoto University. It was created without any deletion or modification using the IDS named "BRO". It consists of 24 features where 14 are extracted from the KDDCup'99 dataset and 10 additional features that can be used to investigate more efficiently the network's characteristics and especially examine the new attacks in the network. In our case, this dataset contains 784,000 21-dimensional records where 388,632 are attacks and 395,368 are normal records. Nevertheless, its major limitation is that there are no measures for labeled traffic and it is limited to attacks generated only from honeypots and not from other systems. Moreover, there is no legitimate (normal) traffic to be evaluated since the output of honeypots is considered only attack traffic. For this reason, we propose to evaluate our framework on the basis of another challenging dataset which is the "ISCX dataset".

3) ISCX DATASET [43]

The ISCX (Information Security Centre of Excellence) is a benchmark intrusion detection dataset that was designed to be used for one week for malware prevention and security testing in 2011. Each record is taken from simulation for one week and consists of 11 features. This dataset contains descriptions of both synthetic attack and legitimate network traffic. The major benefit of this data set is that it contains both captured traffics and description of them.

B. RESULTS

For performance investigation, we run the developed Infinite bounded generalized Gaussian mixture with feature selection (InBGG-Fs) for the three data sets described above. For comparison purposes, we have applied also the following mixture-based methods: finite Gaussian mixture (FiG), finite bounded Gaussian (FiBG), finite generalized Gaussian (FiGG), finite bounded generalized Gaussian (FiBGG), finite Gaussian mixture with feature selection (FiG-Fs), finite bounded Gaussian with feature selection (FiBG-Fs), finite generalized Gaussian with feature selection (FiGG-Fs), finite bounded generalized Gaussian with Feature selection (FiBGG-Fs), infinite Gaussian (InG), infinite bounded Gaussian (InBG), infinite generalized Gaussian (InGG),

[2] *http : //kdd.ics.uci.edu/databases/kddcup99/kddcup99.html*

[3] *http : //www.takakura.com/Kyoto_data/*

TABLE 1. Accuracy when deploying the different mixture models without feature selection to the different datasets.

	KDDCup'99	Kyoto 2006+	ISCX
FiG	81.34%	85.77%	89.21%
FiBG	81.86%	86.29%	89.43%
FiGG	82.52%	86.68%	89.51%
FiBGG	82.77%	86.93%	89.62%
InG	81.97%	86.88%	90.02%
InBG	82.11%	87.09%	90.09%
InGG	83.04%	87.21%	90.27%
InBGG	83.49%	87.41%	90.40%

TABLE 2. FPR when deploying the different mixture models without feature selection to the different datasets.

	KDDCup'99	Kyoto 2006+	ISCX
FiG	18.85%	15.81%	10.33%
FiBG	18.71%	15.65%	10.28%
FiGG	17.88%	15.38%	10.26%
FiBGG	17.75%	15.01%	10.08%
InG	17.70%	14.55%	10.03%
InBG	17.09%	14.47%	9.95%
InGG	16.90%	14.32%	9.88%
InBGG	16.84%	14.24%	9.79%

infinite bounded generalized Gaussian (InBGG), infinite Gaussian with feature selection (InG-Fs), infinite bounded Gaussian with feature selection (InBG-Fs), infinite generalized Gaussian with Feature selection (InGG-Fs). All these approaches have been implemented using Bayesian-based MCMC inference.

The average classification accuracy rate (Accuracy) and the false positive rate (FPR) have computed when applying each approach for quantitative comparison:

- Accuracy metric: represents the percentage of correctly classified instances compared to the total number of instances. In other word, it computes the overall detection's percentages which indicates the success rate of any intrusion detection method. It is given by: $Accuracy = (TN + TP)(TP + FP + TN + FN)$.
- False Positive Rate (FPR): it represents the percentage of normal instances wrongly categorized as malware attacks compared with the total number of normal instances. It is given by: $FPR = FP(FP + TN)$.

where FP, TN are the number of false positives and true negatives, respectively. A perfect intrusion detection method should have a 100% accuracy while a 0% false positive rate (FPR) which indicate that it can detect all possible attacks without any error (misclassification) which is very difficult and may be impossible in real environments.

Tables 1 and 2 summarize the accuracy and FPR results, respectively, when deploying the different finite and infinite mixture models without feature selection. According to the results, it is clear that the proposed infinite InBGG model provides the best detection results. Indeed, if we see at the accuracies for the KDDCup'99 dataset, it is about 83.49% for our infinite bounded mixture model (InBGG), but for finite Gaussian mixture (FiG) it is equal to 81.34% and for finite generalized Gaussian mixture (FiGG) it is equal to 82.52% and for the finite bounded FiBGG model is about 82.77%. Similarly, we can obtain the best values with our method as compared to other finite models, for Kyoto 2006+ and ISCX datasets. These outcomes confirm evidently that our choice for the infinite formalism allows improving expected detection accuracy since the determining of the optimal number of classes becomes more precise. On the other hand, we can see also that the bounded generalized Gaussian outperforms the other distributions when deployed in both finite and infinite mixture models. We can notice also that bounded support distribution improves the results slightly as compared to their unbounded counterparts. In particular, the best values for infinite models is found for the infinite bounded generalized Gaussian InBGG with 83.49%, 87.41%, and 90.40% for KDDCup'99, Kyoto 2006+ and ISCX datasets, respectively. In the same way, for the case of finite models, the bounded one outperforms also all the rest finite Gaussian-based models.

A comparative study was also carried out for showing the merits of integrating a feature selection mechanism with the statistical model. Tables 3 and 4 display the accuracy and FPR results, respectively, when integrating feature selection within the different models. The results show that feature selection improves the detection in terms of accuracy and FPR. According to these results, the best accuracies are obtained with our method and are equal to 84.06%, 88.13%, and 91.82% for KDDCup'99, Kyoto 2006+ and ISCX datasets, respectively. Similarly, the minimum percentage of normal instances wrongly categorized as malware attacks compared with the total number of normal instances (i.e. FPR) are obtained with our model with feature selection (InBGG-Fs).

TABLE 3. Accuracy when deploying the different mixture models with feature selection to the different datasets.

	KDDCup'99	Kyoto 2006+	ISCX
FiG-Fs	82.57%	86.65%	90.39%
FiBG-Fs	82.80%	87.02%	90.50%
FiGG-Fs	82.98%	87.14%	90.66%
FiBGG-Fs	83.08%	87.28%	91.08%
InG-Fs	83.02%	87.40%	91.13%
InBG-Fs	83.10%	87.69%	91.25%
InGG-Fs	83.63%	88.04%	91.38%
InBGG-Fs	84.06%	88.13%	91.82%

TABLE 4. FPR when deploying the different mixture models with feature selection to the different datasets.

	KDDCup'99	Kyoto 2006+	ISCX
FiG-Fs	18.22%	14.65%	9.15%
FiBG-Fs	18.08%	14.58%	9.11%
FiGG-Fs	17.74%	14.38%	8.88%
FiBGG-Fs	17.19%	14.19%	8.64%
InG-Fs	16.94%	13.90%	8.57%
InBG-Fs	16.83%	13.84%	8.48%
InGG-Fs	16.26%	13.72%	8.42%
InBGG-Fs	16.02%	13.39%	8.37%

These results are justified and they are due to the importance of taking into account only most relevant features. This is actually expected since this step allows to eliminate features that may compromise the detection process. Moreover, we can notice again that the generalized Gaussian-based models outperform the Gaussian-based ones which is due to the flexibility that the generalized Gaussian add to data modeling.

VI. CONCLUSION AND FUTURE WORK

In this paper, we have addressed the problem of anomaly-based intrusion detection and we have developed a novel fully Bayesian-based approach for infinite bounded Generalized Gaussian mixture model. An important characteristic of the developed model is that it integrates a feature selection mechanism to prevent irrelevant features from compromising the modeling process. Our choice of Bayesian inference methodology is motivated by the fact that it permits to avoid under and over-fitting, to formalize our prior knowledge and to express our uncertainty through probability distributions. In addition, the main goal of using the infinite assumption instead of finite one is its capability in learning simultaneously (i.e. parameters estimation and model selection) the model's parameters and number of components. On the other hand, the integration of a feature selection mechanism aims at eliminating irrelevant features and considering only most relevant ones and then increasing the performance in term of accuracy. The effectiveness of our framework is confirmed by testing it on the challenging application namely anomaly intrusion detection, while comparing it to other comparable published methods from the literature. Future works could be devoted to adapt and extend the developed framework to be useful for specific problems related to IoT and smart cities-based security. We are trying also to build our own data sets for a real IoT environment. To achieve these objectives, we plan to implement a generative discriminative framework based on the bounded models in order to avoid drawback of generative methods alone and to enhance expected results when taking simultaneously the advantages of both discriminative/generative approaches. Another future work could be the handling of more large scale IDS-based datasets to offer a deep comprehensive analysis and detection system.

REFERENCES

[1] B. B. Zarpelão, R. S Miani, C. T. Kawakani, and S. C. de Alvarenga, "A survey of intrusion detection in Internet of Things," *J. Netw. Comput. Appl.*, vol. 84, pp. 25–37, Apr. 2017.

[2] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1153–1176, 2nd Quart., 2016.

[3] H.-J. Liao, C.-H. R. Lin, Y.-C. Lin, and K.-Y. Tung, "Intrusion detection system: A comprehensive review," *J. Netw. Comput. Appl.*, vol. 36, no. 1, pp. 16–24, 2013.

[4] W. Chimphlee, A. H. Abdullah, M. N. M. Sap, S. Srinoy, and S. Chimphlee, "Anomaly-based intrusion detection using fuzzy rough clustering," in *Proc. Int. Conf. Hybrid Inf. Technol.*, 2006, pp. 329–334.

[5] M. Ahmed, A. N. Mahmood, and J. Hu, "A survey of network anomaly detection techniques," *J. Netw. Comput. Appl.*, vol. 60, pp. 19–31, Jan. 2016.

[6] M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, "Network anomaly detection: Methods, systems and tools," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 303–336, 1st Quart., 2014.

[7] P. Garcia-Teodoro, J. E. Díaz-Verdejo, G. Maciá-Fernández, and E. Vázquez, "Anomaly-based network intrusion detection: Techniques, systems and challenges," *Comput. Secur.*, vol. 28, nos. 1–2, pp. 18–28, 2009.

[8] S. X. Wu and W. Banzhaf, "The use of computational intelligence in intrusion detection systems: A review," *Appl. Soft Comput.*, vol. 10, no. 1, pp. 1–35, 2010.

[9] A. H. Hamamoto, L. F. Carvalho, L. D. H. Sampaio, T. Abrão, and M. L. Proença, Jr., "Network anomaly detection system using genetic algorithm and fuzzy logic," *Expert Syst. Appl.*, vol. 92, pp. 390–402, Feb. 2018.

[10] W. L. Al-Yaseen, Z. A. Othman, and M. Z. A. Nazri, ''Multi-level hybrid support vector machine and extreme learning machine based on modified k-means for intrusion detection system,'' *Expert Syst. Appl.*, vol. 67, pp. 296–303, Jan. 2017.

[11] S. Zhong, T. M. Khoshgoftaar, and S. V. Nath, ''A clustering approach to wireless network intrusion detection,'' in *Proc. 17th IEEE Int. Conf. Tools Artif. Intell. (ICTAI)*, Hong Kong, Nov. 2005, pp. 190–196.

[12] H. Sallay and S. Bourouis, ''Intrusion detection alert management for high-speed networks: Current researches and applications,'' *Secur. Commun. Netw.*, vol. 8, no. 18, pp. 4362–4372, 2015.

[13] A. Tajbakhsh, M. Rahmati, and A. Mirzaei, ''Intrusion detection using fuzzy association rules,'' *Appl. Soft Comput.*, vol. 9, no. 2, pp. 462–469, 2009.

[14] S. Benferhat, T. Kenaza, and A. Mokhtari, ''A Naïve Bayes approach for detecting coordinated attacks,'' in *Proc. 32nd Annu. IEEE Int. Comput. Softw. Appl. Conf. (COMPSAC)*, Turku, Finland, Jul./Aug. 2008, pp. 704–709.

[15] H. Sallay, A. Ammar, M. B. Saad, and S. Bourouis, ''A real time adaptive intrusion detection alert classifier for high speed networks,'' in *Proc. IEEE 12th Int. Symp. Netw. Comput. Appl.*, Cambridge, MA, USA, Aug. 2013, pp. 73–80.

[16] M.-Y. Su, ''Using clustering to improve the KNN-based classifiers for online anomaly network traffic identification,'' *J. Netw. Comput. Appl.*, vol. 34, no. 2, pp. 722–730, 2011.

[17] I. Syarif, A. Prugel-Bennett, and G. Wills, ''Unsupervised clustering approach for network anomaly detection,'' in *Proc. 4th Int. Conf. Netw. Digit. Technol. (NDT)*, Dubai, UAE, Apr. 2012, pp. 135–145.

[18] W. Fan, M. Miller, S. Stolfo, W. Lee, and P. Chan, ''Using artificial anomalies to detect unknown and known network intrusions,'' *Knowl. Inf. Syst.*, vol. 6, no. 5, pp. 507–527, 2004.

[19] K. Leung and C. Leckie, ''Unsupervised anomaly detection in network intrusion detection using clusters,'' in *Proc. 28th Australas. Comput. Sci. Conf. (ACSC)*, Newcastle, NSW, Australia, Jan./Feb. 2005, pp. 333–342.

[20] W. Fan, N. Bouguila, and D. Ziou, ''Unsupervised anomaly intrusion detection via localized Bayesian feature selection,'' in *Proc. 11th IEEE Int. Conf. Data Mining (ICDM)*, Vancouver, BC, Canada, Dec. 2011, pp. 1032–1037.

[21] T. Bdiri and N. Bouguila, ''Positive vectors clustering using inverted Dirichlet finite mixture models,'' *Expert Syst. Appl.*, vol. 39, no. 2, pp. 1869–1882, 2012.

[22] G. McLachlan and D. Peel, *Finite Mixture Models* (Wiley Series in Probability and Statistics). New York, NY, USA: Wiley, 2000.

[23] S. Bourouis, A. Zaguia, N. Bouguila, and R. Alroobaea, ''Deriving probabilistic SVM kernels from flexible statistical mixture models and its application to retinal images classification,'' *IEEE Access*, vol. 7, pp. 1107–1117, 2019.

[24] F. Najar, S. Bourouis, N. Bouguila, and S. Belghith, ''A fixed-point estimation algorithm for learning the multivariate GGMM: Application to human action recognition,'' in *Proc. IEEE Can. Conf. Elect. Comput. Eng. (CCECE)*, May 2018, pp. 1–4.

[25] F. Najar, S. Bourouis, N. Bouguila, and S. Belghith, ''Unsupervised learning of finite full covariance multivariate generalized Gaussian mixture models for human activity recognition,'' *Multimedia Tools Appl.*, to be published.

[26] I. Channoufi, S. Bourouis, N. Bouguila, and K. Hamrouni, ''Image and video denoising by combining unsupervised bounded generalized gaussian mixture modeling and spatial information,'' *Multimedia Tools Appl.*, vol. 77, no. 19, pp. 25591–25606, 2018.

[27] N. Bouguila and D. Ziou, ''A countably infinite mixture model for clustering and feature selection,'' *Knowl. Inf. Syst.*, vol. 33, no. 2, pp. 351–370, 2012.

[28] W. Fan, N. Bouguila, and H. Sallay, ''Anomaly intrusion detection using incremental learning of an infinite mixture model with feature selection,'' in *Proc. 8th Int. Conf., Rough Sets Knowl. Technol. (RSKT)*, Halifax, NS, Canada, Oct. 2013, pp. 364–373.

[29] N. Bouguila, ''Infinite Liouville mixture models with application to text and texture categorization,'' *Pattern Recognit. Lett.*, vol. 33, pp. 103–110, Jan. 2012.

[30] Y. Li and L. Guo, ''TCM-KNN scheme for network anomaly detection using feature-based optimizations,'' in *Proc. ACM Symp. Appl. Comput. (SAC)*, Fortaleza, Brazil, Mar. 2008, pp. 2103–2109.

[31] P. Helman and J. Bhangoo, ''A statistically based system for prioritizing information exploration under uncertainty,'' *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 27, no. 4, pp. 449–466, Jul. 1997.

[32] I. Channoufi, S. Bourouis, N. Bouguila, and K. Hamrouni, ''Color image segmentation with bounded generalized Gaussian mixture model and feature selection,'' in *Proc. 4th Int. Conf. Adv. Technol. Signal Image Process. (ATSIP)*, Sousse, Tunisia, Mar. 2018, pp. 1–6.

[33] N. Bouguila, ''A model-based approach for discrete data clustering and feature weighting using map and stochastic complexity,'' *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 12, pp. 1649–1664, Dec. 2009.

[34] I. Channoufi, S. Bourouis, N. Bouguila, and K. Hamrouni, ''Spatially constrained mixture model with feature selection for image and video segmentation,'' in *Proc. 8th Int. Conf. Image Signal Process. (ICISP)*, Cherbourg, France, Jul. 2018, pp. 36–44.

[35] S. Bourouis, Y. Laalaoui, and N. Bouguila, ''Bayesian frameworks for traffic scenes monitoring via view-based 3D cars models recognition,'' *Multimedia Tools Appl.*, to be published.

[36] S. Bourouis, F. R. Al-Osaimi, N. Bouguila, H. Sallay, F. Aldosari, and M. Al Mashrgy, ''Bayesian inference by reversible jump MCMC for clustering based on finite generalized inverted Dirichlet mixtures,'' *Soft Comput.*, to be published.

[37] C. Constantinopoulos, M. K. Titsia, and A. Likas, ''Bayesian feature and model selection for Gaussian mixture models,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 6, pp. 1013–1018, Jun. 2006.

[38] M. H. C. Law, M. A. T. Figueiredo, and A. K. Jain, ''Simultaneous feature selection and clustering using mixture models,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1154–1166, Sep. 2004.

[39] N. Bouguila and D. Ziou, ''A Dirichlet process mixture of Dirichlet distributions for classification and prediction,'' in *Proc. IEEE Workshop Mach. Learn. Signal Process.*, Oct. 2008, pp. 297–302.

[40] R. M. Neal, ''Markov chain sampling methods for Dirichlet process mixture models,'' *J. Comput. Graph. Statist.*, vol. 9, no. 2, pp. 249–265, Jun. 2000.

[41] C. E. Rasmussen, ''The infinite Gaussian mixture model,'' in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2000, pp. 554–560.

[42] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, ''A detailed analysis of the KDD CUP 99 data set,'' in *Proc. IEEE Symp. Comput. Intell. Secur. Defense Appl. (CISDA)*, Ottawa, ON, Canada, Jul. 2009, pp. 1–6.

[43] A. Shiravi, H. Shiravi, M. Tavallaee, and A. A. Ghorbani, ''Toward developing a systematic approach to generate benchmark datasets for intrusion detection,'' *Comput. Secur.*, vol. 31, no. 3, pp. 357–374, 2012.

[44] J. Song, H. Takakura, Y. Okabe, M. Eto, D. Inoue, and K. Nakao, ''Statistical analysis of honeypot data and building of Kyoto 2006+ dataset for NIDS evaluation,'' in *Proc. 1st Workshop Building Anal. Datasets Gathering Exper. Returns Secur. (BADGERS@EuroSys)*, Salzburg, Germany, Apr. 2011, pp. 29–36.

WAJDI ALHAKAMI received the B.Sc. degree in computer science from Jeddah University, Saudi Arabia, the M.Sc. degree in computer network, and the Ph.D. degree in network security from the University of Bedfordshire, U.K. His research interests include the Internet of Things, cyber security, and computer networking.

ABDULLAH ALHARBI received the Ph.D. degree in information technology from the University of Technology Sydney, Australia. He is currently an Assistant Professor with Taif University. His research interests include the Internet of Things, human computer interaction, information systems and security, and big data analytics.

SAMI BOUROUIS received the Engineering and M.Sc. degrees in computer science from National School for Computer Science (ENSI), Tunisia, in 2003 and 2005, respectively, and the Ph.D. degree from the National Engineering School of Tunis (ENIT), Tunisia, in 2011. He is currently an Assistant Professor with the College of Computers and Information Technology, Taif University, Saudi Arabia. His research interests include image processing, machine learning, computer vision, and pattern recognition.

ROOBAEA ALROOBAEA received the bachelor's degree (Hons.) in computer science from King Abdulaziz University (KAU), Saudi Arabia, in 2008, and the master's degree in information system and the Ph.D. degree in computer science from the University of East Anglia, U.K., in 2012 and 2016, respectively. He is currently an Assistant Professor with the College of Computers and Information Technology, Taif University, Saudi Arabia. His research interests include human computer interaction, cloud computing, and machine learning.

NIZAR BOUGUILA received the Engineering degree from the University of Tunis, Tunis, Tunisia, in 2000, and the M.Sc. and Ph.D. degrees from Sherbrooke University, Sherbrooke, QC, Canada, in 2002 and 2006, respectively, all in computer science. He is currently a Professor with the Concordia Institute for Information Systems Engineering (CIISE), Concordia University, Montréal, QC, Canada. His research interests include image processing, machine learning, data mining, 3-D graphics, computer vision, and pattern recognition.