International Journal of Computer Networks and Applications (IJCNA)
DOI: 10.22247/ijcna/2021/207980 Volume 8, Issue 1, January – February (2021)

RESEARCH ARTICLE

Towards Improved Detection of Intrusions with Constraint-Based Clustering (CBC)

J. Rene Beulah
Department of Computer Science and Engineering, College of Engineering and Technology, SRM Institute of Science and Technology, SRM Nagar, Kattankulathur, Kanchipuram, Chennai, Tamil Nadu, India.
renebeulah@gmail.com

C. Pretty Diana Cyril
Department of Computer Science and Engineering, College of Engineering and Technology, SRM Institute of Science and Technology, SRM Nagar, Kattankulathur, Kanchipuram, Chennai, Tamil Nadu, India.
prettydianacyril@gmail.com

S. Geetha
Department of Information Science and Engineering, CMR Institute of Technology, Bangalore, India.
geetha2016research@gmail.com

D. Shiny Irene
Department of Computer Science and Engineering, SRM Institute of Science and Technology, SRM Nagar, Chennai, Tamil Nadu, India.
dshinyirene@gmail.com

Received: 23 December 2020 / Revised: 18 January 2021 / Accepted: 23 January 2021 / Published: 23 February 2021

Abstract – The modern society is greatly benefited by the advancement of the Internet. The quick surge in the number of connections and the ease of access to the Internet have given rise to tremendous security threat to individuals and organizations. In addition to intrusion prevention techniques like firewalls, intrusion detection systems (IDS) are an obligatory level of safety for establishments to identify insiders and outsiders with malicious intentions. Anomaly-based IDS is in the literature for the last few decades, but still the existing methods lack in three main aspects – difficulty in handling mixed attribute types, more dependence on input parameters and incompetence in maintaining a good balance between detection rate (DR) and false alarm rate (FAR). The research work proposed in this paper proposes a semi supervised IDS based on outlier detection which first selects the important features that help in identifying intrusive events and then applies a constraint-based clustering algorithm to closely learn the properties of normal connections. The proposed method can handle data with mixed attribute types efficiently, requires less number of parameters and maintains a good balance between DR and FAR. The standard NSL-KDD benchmark dataset is used for performance evaluation and the experimental results yielded an overall DR of 99.52% and FAR of 1.15%. It is successful in identifying 99.81% of DoS attacks, 99.71% of Probe attacks, 98.73% of R2L attacks and 96.50% of U2R attacks.

Index Terms – Anomaly, Classification, Feature Extraction, NSL-KDD Dataset, Outlier, Intrusion Detection.

1. INTRODUCTION

A computer network, also known as data network can be defined as telecommunications network that gives accessibility to computers to exchange data. Even though there is a boom in the utilization of Internet and e-Commerce, the insecurity of online domain has an adverse effect on users and enterprises. The network traffic keeps growing in parallel with the growing rate of the number of users linked to the Internet on account of usage of various applications such as chatting, searching, video conferencing, streaming, downloading and uploading of images. The availability of various applications leads to a wide variety of cyber threats, viruses and bot attacks. Securing networks from intruders is paramount in the era of growing cyber-attacks. Network security is a process of *defense* from all forms of internal and external threats in order to guarantee the safety of communications network and information. Intrusion Detection System (IDS) is a part of network security design.

1.1. Intrusion Detection System

Intrusion prevention mechanisms like firewall, data encryption and user authentication are used by organizations as a first level of defense. Still, guaranteed prevention of all kinds of attacks is impractical. IDS can be thought of as a

RESEARCH ARTICLE

second level of defense and is not a substitute for other security services. IDS collects information from a variety of system and network sources, examines all inbound and outbound traffic, performs local analysis of that traffic to identify suspicious patterns and takes action by alerting operators. It is a special purpose device or software used to detect anomalies and attacks in networks. An IDS that is appropriately designed and deployed will help in identifying and blocking intruders.

The role of IDS in a network in shown in Figure 1. IDS operates along with other modules used for securing the networked computer systems. A firewall looks only externally and restricts entry to networks thereby preventing outside intruders to an extent, whereas an IDS also keeps watch for attacks that emerge from within a network. An IDS can sense when a system or a network is misused or is under attack. Also, IDS automates the process of detecting malicious activities thereby helping the network administrators in monitoring the network. Thus, IDS extends the level of protection of the target system, resources or information and thereby plays a vital role in securing networks from intruders.

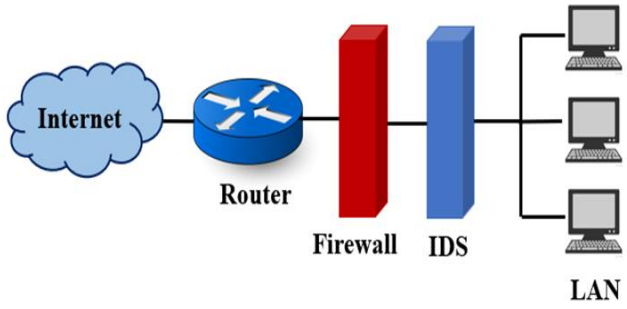

Figure 1 Role of Intrusion Detection System

IDS is categorized depending on the deployment and based on the working methodology and is depicted pictorially in Figure 2. Depending on the way in which IDS is deployed, IDS can be classified into three approaches. HIDS runs on individual devices in a network and monitors traffic to and from that particular device alone. NIDS is located at a place within a network from whence it can examine packets that pass through all the nodes in the network. Hybrid approaches use both network-based and host-based tools. Then based on the working methodology, IDS is categorized into two types: Misuse / Signature Detection and Anomaly Detection. Signature-based IDS maintains a database of signatures/patterns of known malicious activities. The packets on the network are compared with this database. If a match occurs it will be signaled as an intrusion. These systems are very effective in detecting attacks without triggering an enormous number of false alarms. But they will identify the types of intrusions that are known beforehand and they will not identify new intrusions or variations of old intrusions. Also, the library of signatures must be updated constantly. In addition, it is more difficult to identify insiders who misuse their privileges. The next type, anomaly-based IDS establishes a baseline of the normal behaviour of the network and if the traffic is different from the baseline, an alarm will be raised. Unlike signature-based methods, anomaly-based methods can detect any unusual conduct and is able to detect previously unseen intrusions. It can also detect insiders who misuse their privileges and has high detection rate. But its False Alarm Rate (FAR) is generally high because a suspicious activity do not always mean that an intrusion is happening. Hence anomaly-based IDS should be prudently adjusted to evade high FAR.

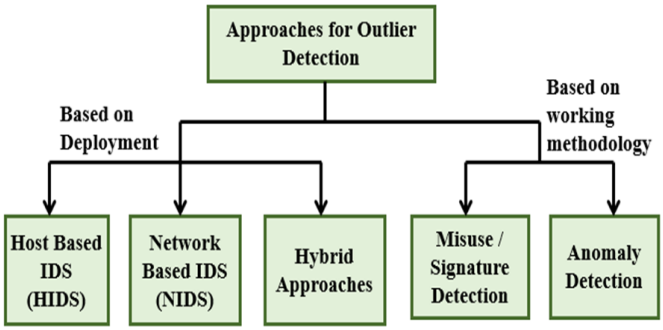

Figure 2 Classification of Intrusion Detection System

Many researchers are working on IDS for over twenty-five years. Still, the networks all over the world experience some kind of attacks by hackers every day. When the attacks are successful, the intruders obtain access to email addresses and other sensitive information. Most of the commercial IDS are signature-based. Anomaly-based IDS is not much prevalent in the market as it leads to many false alarms getting triggered every day. But because of its promising capabilities of identifying unseen attacks, anomaly-based IDS is the major emphasis of investigation nowadays. It is an inspiring mission to increase Detection Rate yonder a specific limit while maintaining FAR at a less value. The real difficulty is in configuring the boundary so as to maintain a good balance between DR and FAR.

1.2. Why Outlier Detection for IDS?

Intrusion detection can be thought of as a data analysis process. The data about the network packets is enormous and high dimensional and analyzing the data is infeasible without the support of suitable tools. Outlier analysis or outlier detection is a data mining notion that deals with finding patterns that differ pointedly form the rest of the data. Patterns or forms in data that do not adapt to the usual conduct are called as outliers [1]. The concept of outlier detection is much suitable for detecting anomalous events as

RESEARCH ARTICLE

they differ from that of normal user behavior. It is used in an assortment of applications like detection of fraudulent activities in a network, behavioural analysis, fraudulent credit card transactions, medical diagnosis and detecting failures in complex systems. In Intrusion Detection, an outlier may denote an intruder in a network with malicious intentions [2]. Hence, network intrusion detection is a major application area of outlier analysis.

1.3. Objectives of the Research Work

The objective of this research work is to apply outlier analysis to design an effective anomaly-based IDS with the following properties:

1. Should be able to handle datasets with mixed attributes, that is, numerical and categorical attributes, easily and effectively
2. Should be less dependent on parameters
3. Should maintain a good balance between DR and FAR.

1.4. Organization of the Paper

Section 2 presents a detailed survey of research works that apply various clustering methods for outlier detection. The dataset used is elaborated in Section 3. Section 4 explains in detail about outlier identification with Constraint-Based Clustering. Section 5 highlights the interesting findings of this research work and deliberates the performance evaluation of the proposed methods. Section 6 concludes the work and briefly puts forth the promising new research directions.

2. RELATED WORK

Outlier analysis deals with isolating unusual and suspicious information. Identification of outliers is an exceedingly important area of research as it finds application in fields like IDS, identifying fraudulent activities, identifying faults, military surveillance for enemy activities, event detection in sensor networks and the list goes on because abnormal or unusual activities can be found in almost all areas of life. From the characteristics of intrusions, it can be seen that outlier detection is more suitable for anomaly-based IDS.

2.1. Approaches for Outlier Detection

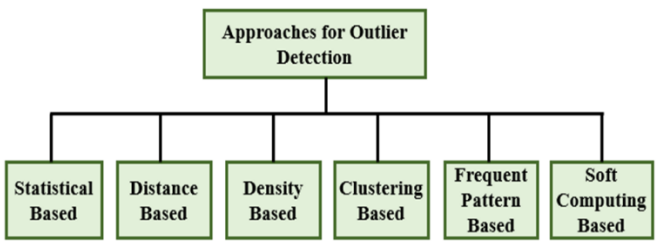

Figure 3 Classification of Outlier Detection Methods

Methods for identification of outliers are broadly classified into six categories as shown in Figure 3.

Statistical-based methods depend on the guess that usual data objects follow a producing mechanism and unusual objects diverge from this mechanism. For the dataset under consideration, a distribution is assumed and if the model data is not conforming to the model, it is an outlier. Statistical-based methods are efficient and they give meaningful interpretation of the result. Also, they are mathematically justified [3]. A major issue of statistical-based approach is that the sample is not always guaranteed to match the presumed distribution law. Also, for applications with many attributes, determining multidimensional distributions is extremely complex.

Knorr & Ng have first suggested the idea of distance-based outliers [4]. In distance-based outlier detection methods, the distances among data points are calculated and the data points that are positioned at a long distance from most of the data points are treated as outliers based on some threshold. These methods rely on the concept of the neighborhood of a data point, usually the k nearest neighbors [5]. Basically, it is assumed that the usual data points have a condensed neighborhood and outliers are at more distance from the neighbors. In general, distance-based measures are highly accurate in terms of distance measure and an effective use of threshold will yield better accuracy. But a major issue is that it takes more computation time.

Density-based outlier detection approaches apply additional intricate mechanisms to simulate the outlierness of objects than that of distance-based methods. It normally includes estimating the local density of the data object under study and that of the data objects surrounding it [6]. A data point that lies in a less dense neighborhood is considered as an outlier, whereas a data point that is present in a denser region is considered as a usual activity. This concept was first introduced by Breunig et al., who defined the degree of outlierness of a data object using Local Outlier Factor (LOF) [7]. LOF specifies how lonesome an object is with regard to the objects in the neighborhood. Efficiency of these density-based techniques is generally based on how well the input parameters are fixed. Also, if the data contains segments of varying densities, outliers cannot be found.

In Clustering-based outlier detection methods, the dataset is divided into many groups by applying some kind of clustering algorithm. Single point outliers are defined as objects which cannot be mapped to any cluster. Cluster-based outliers are defined as some very small clusters which are considered as outliers because of their very small size. Ramaswamy et. al. have first introduced the concept of finding outliers by first partitioning the dataset using some clustering algorithm [8].

Frequent pattern-based approaches find all frequent itemsets

RESEARCH ARTICLE

by some mining algorithm. Frequent itemsets are also called frequent patterns and they usually reveal the common patterns in the dataset. Data objects containing more frequent patterns carry more common features of the dataset and so they cannot be outliers. He et. al. first introduced the idea of discovering outlier patterns by means of frequent itemset mining so that a factor called Frequent Pattern Outlier Factor (FPOF) is defined [9].

Soft Computing is an emerging field for learning in unpredictable and imprecise domains. Soft computing-based outlier detection approaches use neural networks or rough set theory or fuzzy techniques or Support Vector Machines (SVMs). Rough set techniques can provide good results when the data objects have uncertainties, fuzzy rough techniques can be applied.

2.2. Clustering-Based Outlier Detection Methods

Among the varied approaches for outlier detection, clustering-based methods have yielded promising results as clustering is a very easy method which yields outliers as by-products. Some argue that the process of clustering is opposite to identifying outliers [10]. But in the process of clustering, there will be certain objects which are not falling under any cluster and thereby denotes outliers. The ultimate aim of clustering is to group data objects that are similar in nature. But the aim of outlier detection is to search for data objects that are not similar. Generally, we can say that in the processing of forming clusters, outliers are natural outcomes and in general, clustering-based approaches are more effective with respect to all aspects of parameters. They can achieve stable performance and are easily adaptable to incremental mode [11].

Ramaswamy et. al. have first introduced the concept of finding outliers by first partitioning the dataset using some clustering algorithm [8]. Cherednichenko has applied simple k-means clustering approach for grouping the objects [12]. Outlier detection using k-means clustering is found to be successful in maximizing the detection rate but it suffers from problems like high false alarm rate, data overload and k determination which causes outlier detection to be unreliable and inefficient.

Teixeira et al. have suggested a scalable method of outlier identification for large high dimensional databases [13]. The technique works in two steps: (1) the dataset is partitioned and (2) outliers are identified. To partition or divide the dataset, a protracted sort of a divisive hierarchical clustering algorithm is applied.

Singh et al. have suggested a scheme named Common Outlier Detection (COD) that identifies outliers shared by two or more partners in a collaborative IDS [14]. A sequence of characters is treated as a data point. For computing the resemblance of two objects the length of the longest common subsequence is used. A score called Maximum Dissimilarity (MD) is used in the scheme and an agglomerative clustering algorithm is used. This method is applied on web access log files to find common outliers.

An Unsupervised Network Anomaly Detection Algorithm (UNADA) is suggested by Casas et al [15]. The dataset is divided into a number of sub-spaces and DBSCAN (Density Based Spatial Clustering of Applications with Noise) method is used to divide each subspace into partitions. In each subspace, outliers are identified.

Bhuyan et al. have suggested NADO (Network Anomaly Detection using Outlier approach) which is a partitioning-based method [16]. In the training stage, a variant of k-means clustering technique is applied to cluster the normal data and for each cluster a reference point is calculated. In the testing phase, for every data point, an outlierness value is computed based on the similarity measure and the probability for that data object to belong to a particular class. Data points whose outlier grade is more than the threshold value fixed by the user will be treated as anomalies or outliers.

Tahir *et al* [17] proposed an ensemble approach which combines K-means partitioning and SVM for detection of attacks. The results show a DR of 96.26% and a FAR of 3.7%.

Bhuyan et al. have proposed a Multi-step Outlier-based approach for Anomaly Detection (MOAD) for network intrusion detection [18]. An attribute choosing method that works on the basis of mutual information and entropy is used in picking the most important features. Then a tree-based subspace clustering approach is used to partition the data objects into partitions. A reference point is calculated for every partition and outliers are identified based on outlier rank computed with regard to the reference points.

2.3. Other Anomaly Detection Methods for IDS

Singh et. al. have proposed an IDS based on Online Sequential Extreme Learning Machine (OS-ELM), a feed forward neural network with one hidden layer, to increase the learning speed of the system. Repeating, identical and alike traffic features are grouped using DBSCAN clustering algorithm. Finally, OS-ELM is used for classification [19].

De la Hoz et. al. porposed a novel method for identifying attacks in network traffic data by combining statistical methods and neural networks. Before classification, attribute selection methods are applied to choose eight attributes as more influential. Finally a statistical based neural network is used to segregate normal and attack events [20]. Bamakan et. al. have proposed a framework called Time-Varying Chaos Particle Swarm Optimization (TVCPSO) to simultaneously perform feature selection and parameter setting for classifiers. SVM classifier is used [21].

RESEARCH ARTICLE

Enache & Sgarcru have proposed an anomaly-based IDS model named IG-BAL-SVM. Information Gain (IG) is used for feature selection. The input parameters are randomized by Bat Algorithm with Levy flights (BAL) and finally SVM classifier is used [22].

Viegas & Santin have proposed a new method for creating intrusion datasets which can be easily updated and reproduced with real and valid traffic. A new evaluation scheme is also presented which allows all the general assumptions in the existing systems to be validated. A multi-objective attribute selection method is also proposed [23].

Aljawarneh et. al. have developed a new hybrid model that combines seven classifiers in WEKA. First, data is filtered using Vote algorithm and Information gain. Then the hybrid classifier is applied. NSL-KDD dataset is used for analysis [24].

An adaptive IDS is introduced by Resende et. al. based on genetic algorithms and profiling. Genetic algorithm is used to optimize the chosen attributes for profiling. Two techniques are used to identify intrusions: distribution fitting and subspace clustering [25].

Min et. al. have adopted word inserting and text convolutional neural network to excerpt attributes from the payloads in the packets transmitted. A tree-based approach is used for final organization and ISCX2012 intrusion dataset is used for performance analysis [26].

A two-level classifier collective model based on rotation forest and bagging is proposed by Tama et. al. where an ensemble attribute choosing method which applies particle swarm optimization, ant colony algorithm and genetic algorithm is used to extract important attributes [27].

An adaptive learning approach called Adaptive Grasshopper Optimisation Algorithm (AGOA) is suggested by Dwivedi et. al. to identify network intrusions. An ensemble attribute choosing method is used to find out the more important attributes and is tested in ISCX2012 intrusion dataset [28].

Sultan et. al. have focused on detection of intrusions from data which is extracted from real time packets using unsupervised deep learning techniques with semi-supervised techniques. The identification capacity of Autoencoder AE and Variational Autoencoder VAE deep learning methods together with OCSVM are examined [29].

Zhiqiang et. al. have developed a supple background depending on Deep neural network in which various attribute selection methods and activation functions are applied to get better efficiency. The framework is evaluated using ISCX2012 and CICDS2017 intrusion data sets [30].

An IDS model which functions on the basis of LeNet-5 convolutional NN is introduced by Cui et. al. This paper claims that the network self-learning ability is strengthened by the two convolution layers and one pooling layer. An attribute choosing technique which works on random forest algorithm is also used [31].

Table 1 tabulates the research works discussed in this section with the dataset used, metrics used, advantages and disadvantages.

Ref.	Author & Year	Method Used	Dataset Used	Metrics used	Advantage	Disadvantage
[12]	Cherednichenko 2005	k-means clustering	KDDCup99	DR, FAR	Successful in increasing DR	High FAR, Dependence on parameter k (number of clusters)
[13]	Teixeira et. al. 2008	Hierarchical clustering	KDDCup99	Ranking heuristics	Can handle mixed attribute datasets	DR and FAR are not analyzed
[14]	Singh et. al. 2009	Agglomerative clustering	Web access log files	No. of true intrusions	Guarantees low FAR	Dependence on parameter (number of alarms)
[15]	Casas et. al. 2011	Sub-space density clustering	MAWI, METROSEC	ROC Plot	Successful in increasing DR	Dependence on parameter k (number of clusters)
[16]	Bhuyan et al. 2011	Variant of k-means clustering	KDDCup99	Precision, Recall, F-Measure	Good F-Measure	Handling mixed attribute

RESEARCH ARTICLE

						datasets (Needs type conversion)
[17]	Tahir et. al. 2015	K-Means Clustering, SVM	NSL-KDD	DR, FAR	Good DR	Very high FAR, Can handle only numeric attributes
[19]	Singh et. al. 2015	OS-ELM	NSL-KDD	DR, FAR, Accuracy	Good DR	Can handle only continuous features
[20]	De la Hoz et. al. 2015	PCA Filtering, Probabilistic SOM	NSL-KDD	Accuracy, Sensitivity, Specificity	Good DR	Very high FAR, Can handle only numeric attributes
[22]	Enache & Sgarciu, 2015	SVM, Bat algorithm	NSL-KDD	DR, FAR, Accuracy	Good DR	Dependence on parameters C and σ
[18]	Bhuyan et. al. 2016	Tree based clustering	NSL-KDD	DR, FAR, Precision, Recall, F-Measure	Low FAR	Dependence on parameter τ
[21]	Bamakarn et. al. 2016	PSO, SVM	NSL-KDD	DR, FAR	Low FAR	Low DR for R2L and U2R attacks
[23]	Viegas et. al. 2017	Naïve Bayes, Decision Tree	TRAbID	Accuracy, FAR	Created a new dataset	New detection method is not proposed
[24]	Ajawarneh et. al. 2018	Hybrid of J48, Meta Pagging, RandomTree, REPTree, AdaBoostM1, DecisionStump, NaiveBayes	NSL-KDD	DR, FAR	Good DR and FAR	No novelty. Just an ensemble of existing methods
[25]	Resende et. al. 2018	Genetic Algorithm	CICIDS2017	DR, FAR	Low FAR	Low DR
[26]	Min et. al. 2018	Convolutional Neural Network, Random Forest	ISCX2012	DR, FAR, Accuracy	Good DR and low FAR	-
[27]	Tama et. al. 2019	Rotation Forest, Bagging	NSL-KDD, UNSW-NB15	Accuracy, FAR, Sensitivity, Precision	-	Very high FAR
[28]	Dwivedi et. al. 2020	SVM with different kernel functions	ISCX2012	Accuracy, DR, Precision, F-Measure	Can handle mixed attribute datasets	Less Accuracy
[29]	Zavrak et. al. 2020	Unsupervised deep learning methods	CICIDS2017	ROC Curve	Good DR	High false alarms

RESEARCH ARTICLE

[30]	Zhiqiang et. al. 2021	Deep Neural Network	ISCX2012, CICIDS2017	DR, FAR	Good DR and low FAR	-
[31]	Sonali & Jadav, 2021	Linear Discriminant Analysis	UNSW-NB15	Accuracy, FAR, ROC	Good Accuracy	The balance between DR and FAR is not so good

Table 1 A review of Intrusion Detection Methods in the Literature

2.4. Drawbacks of the Previous Models

- Very few methods in the literature were designed to handle mixed attribute types. Some methods have used discretization method to convert continuous values to categorical values while some others have converted categorical attributes to continuous.
- Most of the approaches in the literature are parameterized and optimized parameter selection has a vital part. In some methods, the number of outliers to be found out is an input parameter.
- Some methods have succeeded in increasing the DR while others in decreasing the FAR and they fail to maintain a good balance between DR and FAR which is indeed a difficult task.

2.5. Problem Identification

- Designing a classification method that can easily handle mixed attribute types is very much necessary.
- To make the intrusion detection robust, the proposed method should be less dependent on parameters
- A very strict definition of normal behaviour will help in increasing DR, but will lead to increased false alarms. On the other hand, if the definition of normal behaviour is relaxed, false alarms will get reduced at the the cost of reduced DR. Hence the boundary between normal and intrusive behaviour must be precisely configured so as to maintain a good balance between DR and FAR.
- Among the varied approaches for outlier detection, clustering is the most natural way for identifying outliers. Hence it is decided that clustering the normal traffic connection data with respect to the properties of each connections & analyzing behaviour of each cluster is the right way to learn normal network behaviour.

To fill the research gap in the earlier research, a well-defined clustering-based outlier detection scheme is proposed in this paper that can handle data with mixed attribute types efficiently and that maintains a good balance between DR and FAR. The proposed constraint-based clustering method makes outlier detection meaningful and accurate.

3. CONSTRAINT-BASED CLUSTERING TO IDENTIFY OUTLIERS

In the literature of identifying fraudulent activities, data mining plays an important role and most of such methods apply outlier detection. As we have discussed earlier, anomalies can be pinpointed in many different ways. But in the literature, it can be clearly seen that those methods that apply clustering to find outliers are more effective and produces accurate results naturally. Network traffic data is naturally voluminous and definitely needs techniques that can handle voluminous data of mixed data types effectively. It is proved that clustering methods can be developed to cater to these needs. If we look at the perspective of a clustering algorithm, it is prominent that outliers are data points which are not a part of any cluster or group [32]. The work presented in this paper can be used to identify outliers in any application domain where fraudulent activities need to be monitored and it works on the basis of clustering. This method employs constraint-based clustering to identify outliers.

3.1. Introduction to Constraint-Based Clustering (CBC)

Constraint-Based Clustering was proposed by Tung et. al in which the data points are clustered based on some constraints [33]. The data objects satisfying a specific set of constraints are grouped to form a cluster. This way of clustering creates partitions in a more comprehensive way and is applicable for mixed attribute types. Normally, in traditional clustering algorithms, a random set of objects are fixed as cluster centers. Then the data points that are similar to a cluster center are grouped into a cluster based on some similarity measure or distance measure. In the constraint-based clustering approach which is used in this paper, during training phase, clusters are formed based on the known properties of the data points and the properties of each cluster are learned. The properties of clusters include the center of the cluster, maximum distance between the cluster center and the boundary of the cluster and so on. In the testing phase, if a data point does not fall inside the boundary of a cluster, then it will be treated as an outlier. So, clusters are formed first and then the properties of the clusters are derived which is different from the traditional clustering process. All the intrusion datasets have different types of data like numerical

RESEARCH ARTICLE

data and categorical data. Almost all the existing approaches perform well for arithmetic values. The proposed method works well for datasets having mixed attribute types. The outline of the proposed scheme is shown in Figure 4.

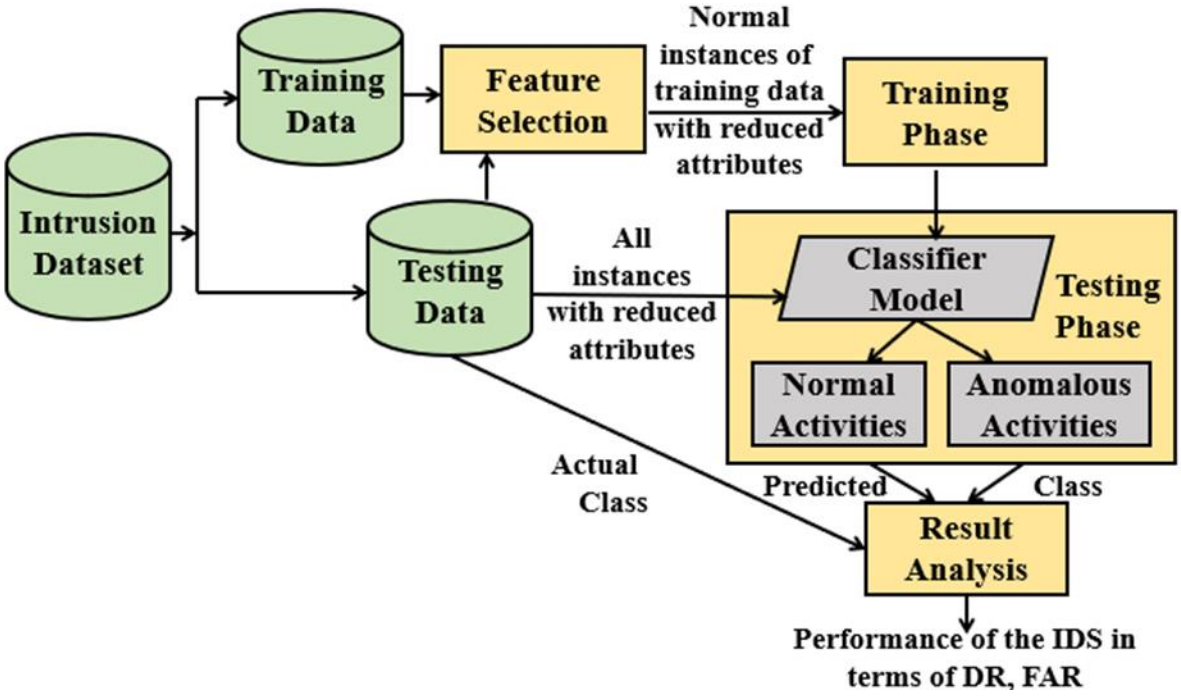

Figure 4 Outline of Anomaly based IDS

3.2. Steps in Outlier Detection with CBC: Training Phase

The input to the clustering is the set of normal instances having m categorical and n numerical attributes. The output will be clusters represented with a few parameters to denote the center of the cluster and the maximum distance of the cluster from the center. There are 6 steps involved in the training phase as described below:

3.2.1. Creating Profiles with Categorical Attributes

In this first step, the categorical attributes are taken and all possible values of the categorical values are analyzed. Let $F_1, F_2, \dots, F_m$ represent the categorical attributes. Categorical features represent distinct discrete values. Let $x_1, x_2, \dots, x_n$ signify the no. of discrete values represented by the features $F_1, F_2, \dots, F_m$ Combinations of different categorical attributes are formed and each such combination is called a profile. Every profile is given a number for easy look-up.

For example, consider a dataset describing fruits having two categorical attributes – size (big, small) and colour (red, green, yellow). There are 6 possible combinations of the attributes size and colour like {big& red, big&green, big&yellow, small&red, small&green, small&yellow}. All such combinations may not be present in the normal instances. Each combination that is present in the normal instances is called a profile and every profile is given a number for easy look-up.

3.2.2. Indexing the Training Set

Each instance in the training set is read from the beginning one by one and is compared with each profile that is formed in the first step. If a match is found, the training instance is assigned with the corresponding profile number. A table called index is formed in which the training record number and the equivalent profile number are noted down which can be used for future reference and for quick access in a random manner. It is also made sure that this index table is stored in a sorted order.

3.2.3. Clustering

All the training instances that belong to the same profile are combined together to form a cluster. If there are n categorical attributes, we get clusters with n-dimensional space. It is good to have reduced number of dimensions so as to improve computational cost. Hence it is always advised to use some attribute selection scheme in the pre-processing step to select the important features in the dataset. After forming clusters, the relationship of the attributes within each cluster, that is, inter-cluster attribute relationship has to be studied. Thus, the

RESEARCH ARTICLE

main idea behind this process is to explore the association of the features with respect to each profile.

3.2.4. Frequency Analysis

The number of occurrences of a data point in a cluster is called the frequency of the element. The frequency of each element is computed in all the clusters and is investigated or analyzed. This frequency analysis is done to eliminate duplicate entries because it is enough to represent a data point in a n-dimensional space once with a note on its frequency. Also this investigation on frequency will give us an idea of which data point has more significance and this will aid us in deciding the parameters of a cluster like the cluster center.

3.2.5. Identifying Outlier Clusters

In the context of intrusion detection, an outlier represents an anomalous or a rare event. Objects that do not fit into any of the groupings can be considered as outliers. In the terms of Intrusion Detection, those traffic patterns that do not conform with the regular pattern can be considered as anomalous patterns. Also, smaller clusters, that is, clusters with lesser objects can also be considered as anomalies as they represent rare or suspicious activities. If the no. of elements in a cluster is less than a threshold θ, the cluster is treated as an outlier as shown in equation (1)

Let $n(C_i)$ be the number of data points in the i^{th} cluster.

$$C_i = \begin{cases} normal\ cluster, when\ n(C_i) > \theta \\ outlier\ cluster, when\ n(C_i) < \theta \end{cases} \quad (1)$$

3.2.6. Calculating the Parameters of the Clusters

In traditional clustering algorithms, the cluster parameters are fixed in the beginning and based on the parameters clusters are formed. For example the no. of clusters to be created and the centers of each cluster are given as input to the clustering algorithm. Then the distance between each object and the all the cluster centers is computed and the object is associated to the cluster whose center is the closest. As the proposed method is a constraint-based clustering method, the clusters are formed first from the profiles. Then it is an important task in the training process to fix the cluster parameters which define the nature and property of each cluster. Each cluster is represented using two parameters *center* and *dist.* Let k be the no. of data objects in a cluster and let D_i be the i^{th} data point. A data object that occurs repeatedly is fixed as the *center* as defined in equation (2).

$$center = \begin{cases} \frac{\sum_{i=1}^{d} D_i}{k}, when\ \forall i, frequency(D_i) = 1 \\ D_i, where\ frequency(D_i) is\ maximum \end{cases} \quad (2)$$

The maximum distance *dist* can be set based on the density of data points around the *center*. Let $(x_1, x_2, \dots, x_n)$ be the *center*. The distance of a data point $D_i\ (y_1, y_2, \dots, y_n)$ from the *center* is computed as given in equation (3).

$$\text{dist} = \text{abs}(x_1 - y_{i_1}) + \text{abs}(x_2 - y_{i_2}) + \cdots + \text{abs}(x_n - y_{i_n}) \quad (3)$$

Apart from these two parameters, depending on the specific application and the attributes, more number of parameters can be added if it will be helpful in deciding a data point as a normal one or an outlier one. These parameters should be wisely calculated as they have more impact in categorizing a network traffic as usual or anomalous.

3.3. Steps in Outlier Detection with CBC: Testing Phase

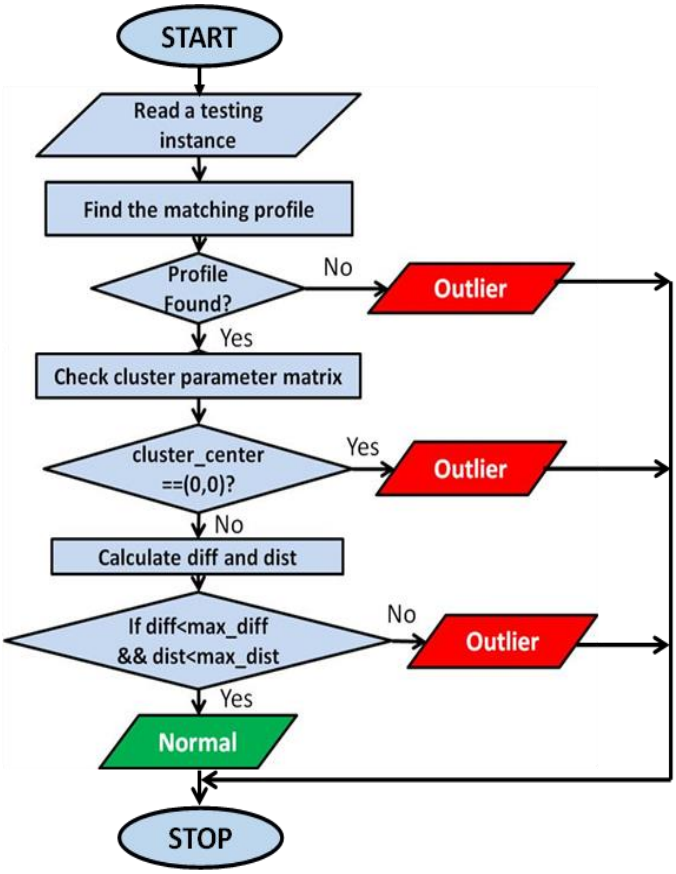

Figure 5 Steps in Testing Phase Outlier Detection with CBC

Before the testing phase, the testing set should be taken and the labels should be removed. The instances in such unlabelled testing set are given one by one as input to the testing phase. The output of the testing phase tells whether the instance is normal or outlier. The steps in the testing phase are as follows:

1. Form a combination of the categorical attributes in the testing instance and check whether it matches with any

RESEARCH ARTICLE

of the profiles formed in the testing phase. If a match is found, go to step 2; else, mark it as an outlier

2. Check the parameters of the cluster. If the value of every attribute is zero, mark it as an outlier; else go to step 3
3. The values of the numerical attributes together represent a data point. Compute d, the distance between the data point and the center. If $d < dist$, mark it as normal; else an outlier

This procedure of outlier identification with constraing-based clustering during the testing stage is shown in Figure 5.

4. DATASET DESCRIPTION AND EXPERIMENTAL SETUP

4.1. Dataset Description

The major application area of outlier analysis is Intrusion Detection System (IDS). Gogoi et. al. say the task of detecting anomalies is analogous to the task of outlier detection, especially in intrusion detection systems. [34]. An attacker within a network having a malicious determination can be identified apparently by an outlier [35]. The proposed outlier detection with CBC method is tried for Intrusion Detection System applied for network intrusion detection to check its effectiveness in identifying outliers. We applied it on the benchmark intrusion detection dataset NSL-KDD.

There are thousands of traffic connections represented in the training and testing datasets. Every traffic connection is described with 41 features as in KDD_Cup'99 dataset. All the records have a label which denote usual and intrusion connections. [36] Gives an elaborated account on the features and the different types of attacks. For attribute selection we have used Improved Hybrid Feature Selection (IHFS) which is our previous work [37]. This method identifies the attributes that have more influence on the intrusion detection.

4.2. Experimental Setup and Evaluation Method

All the experimentations were done on a computer with Intel Core i7-2600 processor @ 3.40 GHz with 2 GB RAM running Windows 7 Professional. The proposed outlier detection with CBC method is executed in MATLAB R2011.

Generally, the accuracy of an IDS is measured based on the degree of detection of attacks and the absence of false alarms. As an IDS is analogous to a classifier, the common metrics and methods used for evaluating the accuracy of classifiers can be used for IDS. The two common methods of evaluation are cross validation and hold out. Cross validation is not used in any of our experiments. Training set is used in the training phase to train the classifier and testing set, which is particularly constructed with 17 additional attacks is employed in the testing phase. This helps us to check whether the proposed system is capable of detecting new attacks.

5. PERFORMANCE ANALYSIS OF OUTLIER DETECTION WITH CBC

To evaluate the performance of the proposed system, three parameters namely Detection Rate (DR), False Alarm Rate (FAR) and Classification accuracy (ACC) are used. The proposed outlier detection with CBC method was able to yield 97.84% DR, 1.88% FAR and 97.96% ACC. The proposed method classifies if a network traffic data is usual or an attack. The confusion matrix has four components – True Positive (TP), False Negative (FN), False Positive (FP) and True Negative (TN). TP represents the number of intrusions that are correctly detected as intrusions. FN represents the number of intrusions that are missed to get detected that is, intrusions that are wrongly classified as normal patterns. FP represents the number of normal patterns that are misclassified as intrusions. TN is the number of normal events that are correctly classified.

To confirm the accuracy and efficacy of the proposed approach, eight outlier detection methods in the literature were chosen. Those eight methods have also tested their efficiency in the same NSL-KDD dataset. The TP, FN, FP and TN values obtained for the proposed method and that of the existing methods are tabulated in Table 2.

Method	TP	FN	FP	TN
PCA & SOM [20]	12448	385	68	9643
K-Means & SVM [17]	12353	480	359	9352
OS-ELM [19]	12706	127	169	9542
IG-BAL-SVM [22]	12724	109	185	9526
MOAD [18]	12478	355	95	9616
TVCPSO [21]	12452	381	84	9627
ENSEMBLE [24]	12718	115	199	9512
TSE-IDS [27]	12140	693	787	8924
Proposed CBC	**12771**	**62**	**112**	**9599**

Table 2 Results of Proposed Vs Existing Methods

The detection rate and false positive rate are defined as given in the following equations (4) and (5).

$$DR = \frac{TP}{Total\ number\ of\ attacks} X100 \quad (4)$$

$$FAR = \frac{FP}{Total\ number\ of\ normal\ events} X100 \quad (5)$$

In NSL-KDD dataset, the testing set contains 22544 instances, of which 12833 instances represent attacks and 9711 instances represent normal traffic. The attack connections in NSL-KDD dataset is broadly classified into four different types such as Dos (Denial of Service) Attacks, Probe Attacks, R2L

RESEARCH ARTICLE

(Remote to Local) Attacks and U2R (User to Root) Attacks. Table 3 projects the efficiency of the proposed method to detect different types of attacks.

The DR and FAR of the proposed method are compared to that of the existing methods in Figures 6 and 7. From the table and the graphs, it is clear that the proposed method has produced the highest DR compared to the existing methods. The False Alarm Rate is also considerably low.

Attack Type	Actual Present	Detected	Missed	Detection Rate %
Dos Attacks	7460	7446	14	99.81%
Probe Attacks	2421	2415	6	99.71%
R2L Attacks	2752	2727	35	98.73%
U2R Attacks	200	193	7	96.50%

Table 3 Detection Rate of Different Types of Attacks in the Proposed Method

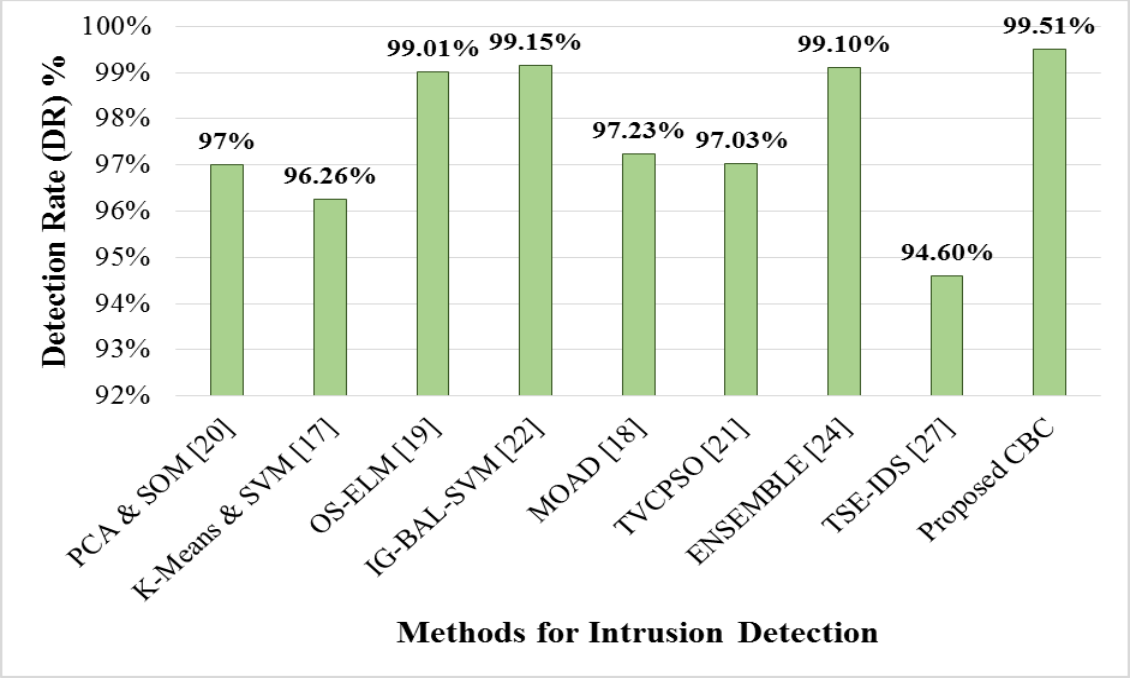

Figure 6 Comparison of Detection Rate with Existing Approaches

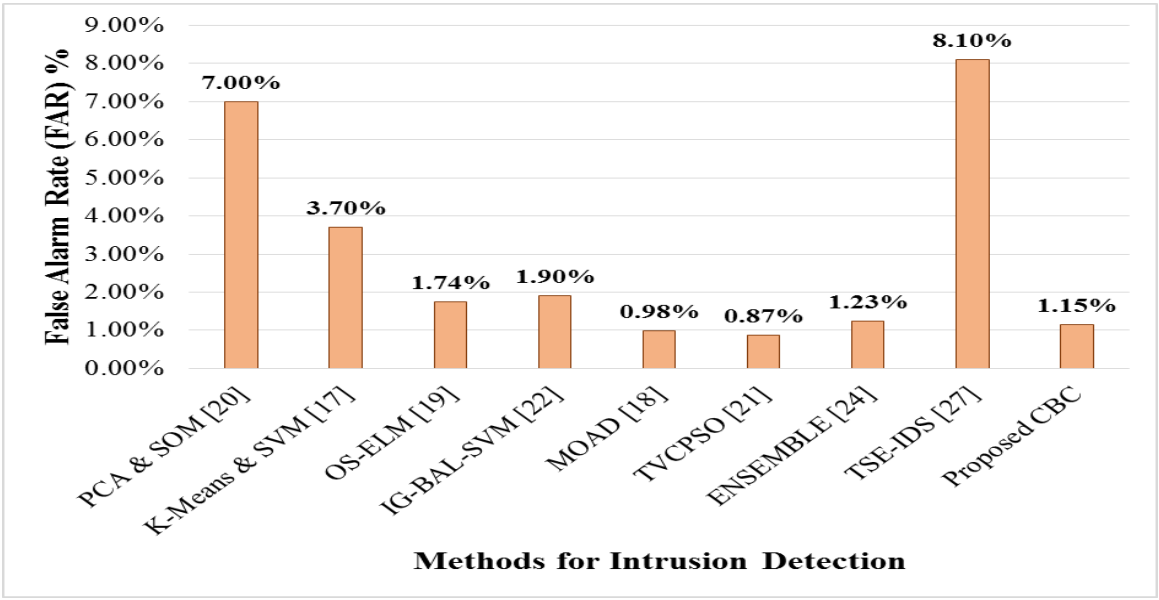

Figure 7 Comparison of False Alarm Rate with Existing Approaches

RESEARCH ARTICLE

Generally, in an intrusion detection system, the detection rate should be high and the false alarm rate must be low. In anomaly detection systems, when we try to increase the detection rate by strictly defining what is normal traffic, if any event slightly deviates the definition of normal traffic, it will be alarmed as in intrusion. So many normal events will be alarmed as intrusions which lead to an increase in the number of false alarms. So, whenever we try to increase the detection rate, the false alarm rate also increases. Hence, it is very tough to increase the DR while decreasing FAR. In Table 1, we can see that two methods MOAD and TVCPSO have achieved a very less FAR of less than 1%. But in both the methods, DR is low which is not satisfactory. Our proposed outlier detection with CBC method has achieved a good trade-off between DR and FAR.

The performance of the proposed method is analyzed with respect to Sensitivity/Recall, Specificity, Accuracy, Precision and F-Measure as defined in equations (6), (7), (8), (9) and (10).

$$\frac{\text{Sensitivity}}{\text{Recall}} = \frac{\text{TP}}{\text{TP+FN}} \quad (6)$$

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

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FN+FP+TN}} \quad (8)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (9)$$

$$\text{F} - \text{Measure} = \frac{2\text{TP}}{(2\text{TP+FP+FN})} \quad (10)$$

Sensitivity indicates how efficiently intrusions are identified. Specificity shows how efficiently the normal instances are identified. Accuracy signifies how well normal events are classified as normal and anomalous events are classified as intrusions. Precision is the fraction of alarms that are true. As the FAR increases, precision decreases. F-Measure reveals the balance between precision and recall, which in turn conveys the tradeoff between DR and FAR.

The performance of the proposed method is analyzed with the parameters Sensitivity, Specificity, Accuracy, Precision and F-Measure as shown in Figures 8, 9, 10, 11 and 12 respectively.

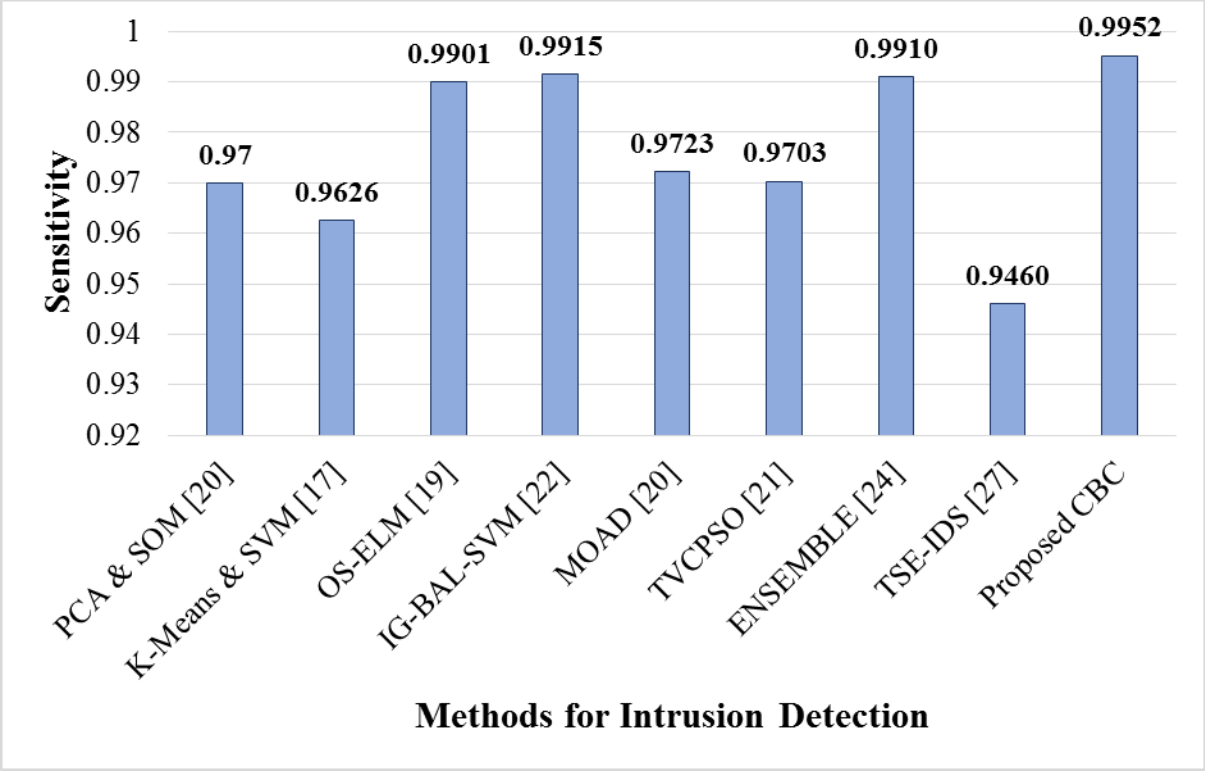

Figure 8 Comparison of Sensitivity with Existing Approaches

RESEARCH ARTICLE

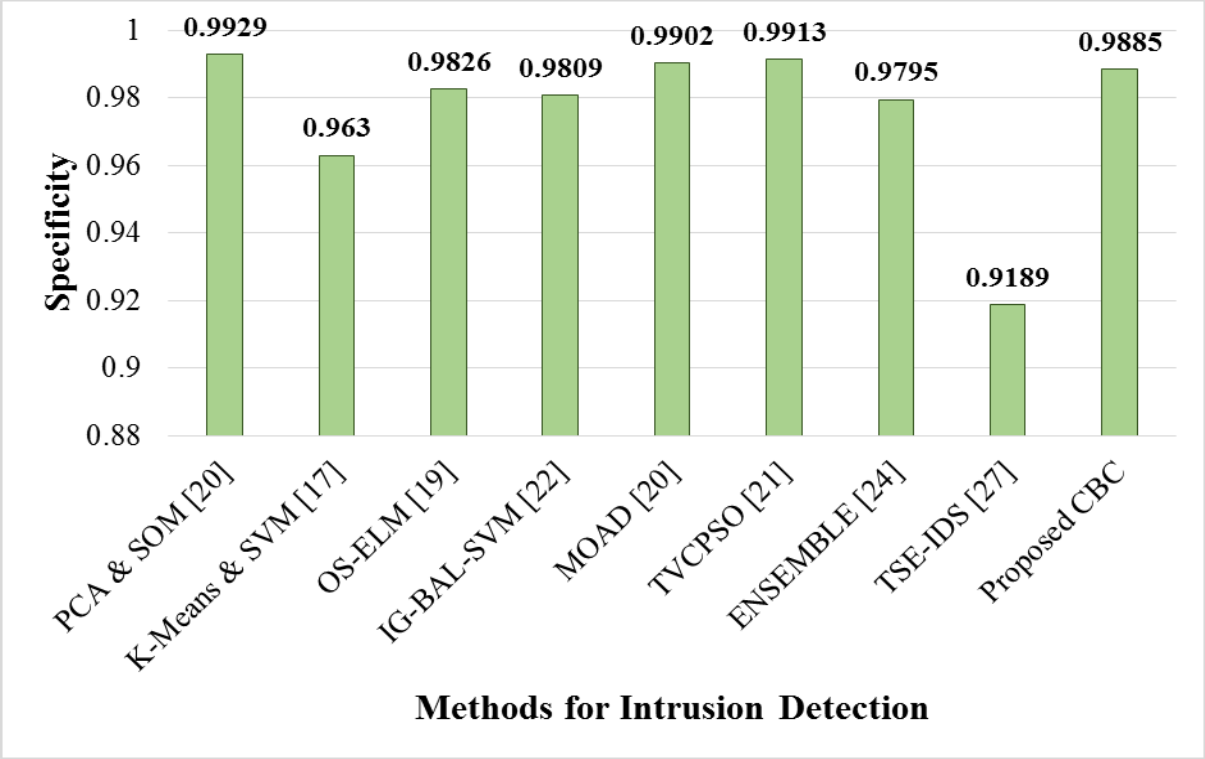

Figure 9 Comparison of Specificity with Existing Approaches

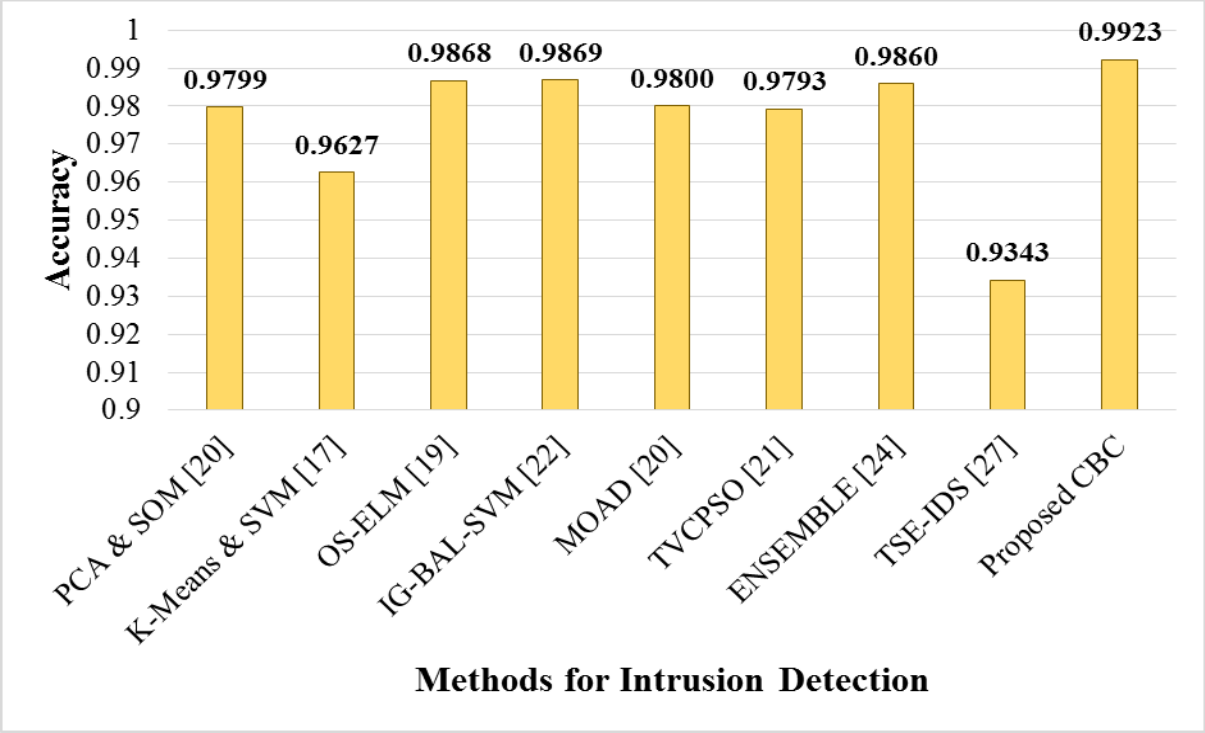

Figure 10 Comparison of Accuracy with Existing Approaches

RESEARCH ARTICLE

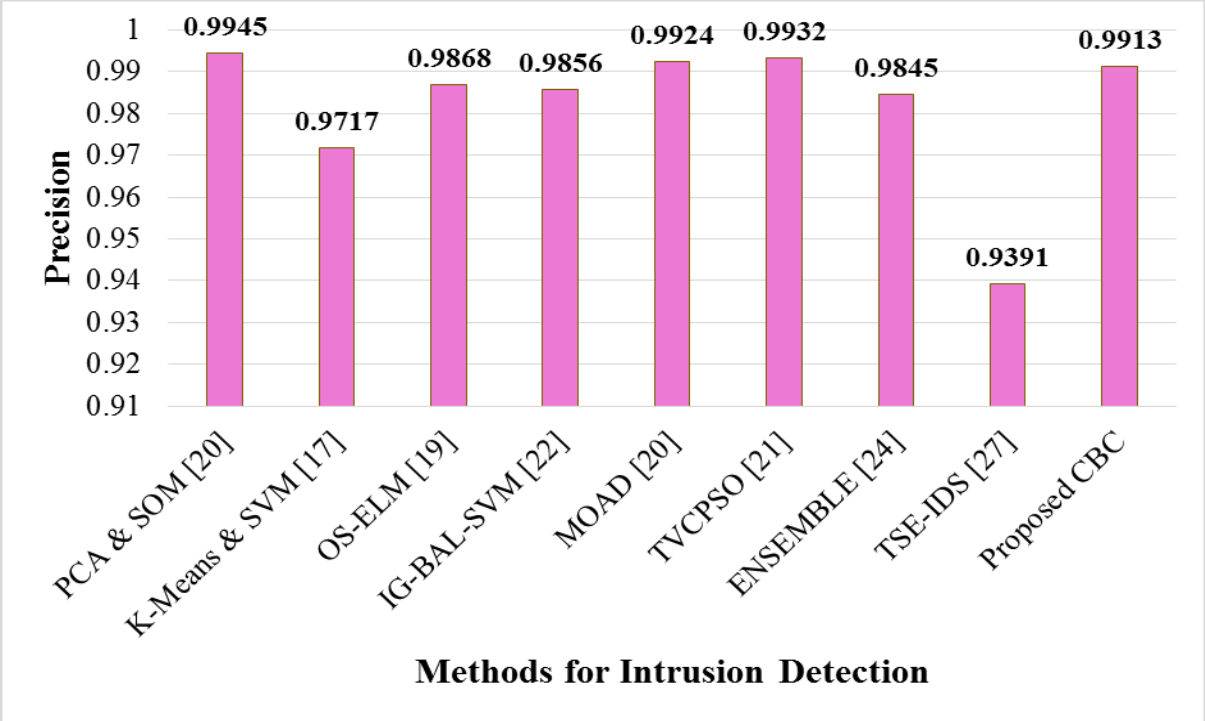

Figure 11 Comparison of Precision with Existing Approaches

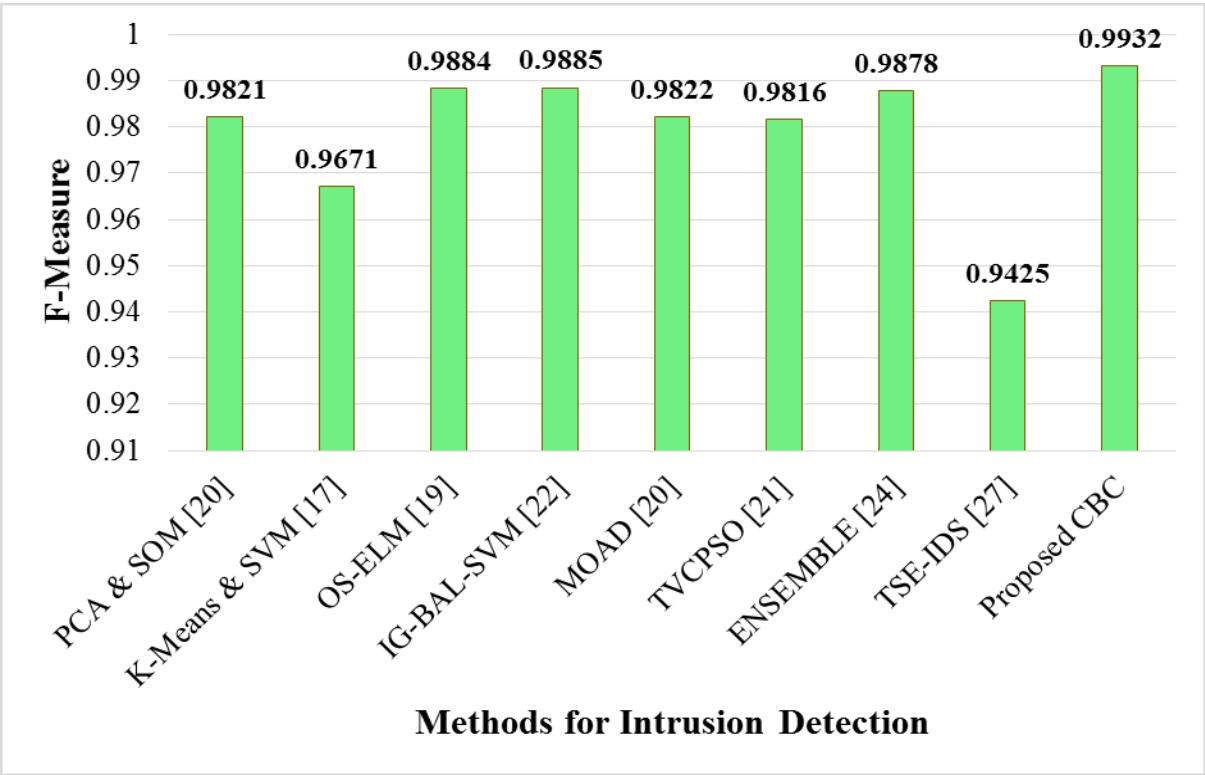

Figure 12 Comparison of F-Measure with Existing Approaches

RESEARCH ARTICLE

From the graphs, it can be clearly visualized that the proposed method has the highest values for sensitivity, accuracy and F-Measure. The methods MOAD and TVCPSO have projected slightly higher specificity and precision as the number of false alarms is slightly lesser. But they show lesser F-Measure than the proposed method. The proposed Constraint-based Clustering helped to form meaningful clusters which characterize the normal behavior of the network. This helps in clearly marking the boundary between normal and abnormal behavior in the network. Thus, the proposed method has gained higher F-Measure, which signifies the best tradeoff between DR and FAR, which confirms its supremacy.

6. CONCLUSION

The challenge of reducing False Alarm Rate while maintaining high DR in Intrusion Detection Systems based on Anomaly Detection is represented in this paper. The proposed outlier detection with Constraint-Based Clustering approach was found to be successful in detecting all types of attacks thereby producing high DR. The design of outlier detection with CBC depends mainly on clustering and frequency analysis. The multiple stages of clustering helped to learn the normal patterns of connections thoroughly and precisely. An interesting observation is that six features were sufficient to identify almost all the attacks except two, which needed four more features. From the experimental results it is very clear that the DR of R2L attacks improved significantly. Also, there is a remarkable decrease in FAR and the balance between DR and FAR is appreciable. We can extend this work to exactly identify the type of intrusion and also can be applied in real-time scenarios. This work can also be applied for identifying outliers in other domains also.

REFERENCES

[1] V. Chandola, A. Banerjee and V. Kumar, "Anomaly Detection: A Survey", in ACM Computing Surveys (CSUR), ACM, Vol. 41, No.3, pp. 1-58, 2009.

[2] V. Hodge and J. Austin, "A Survey of Outlier Detection Methodologies", Artificial Intelligence Review, Vol. 22, No. 2, pp. 85-126, 2004.

[3] M. I. Petrovskiy, "Outlier Detection Algorithms in Data Mining Systems", Programming and Computer Software, Vol. 24, No. 4, pp. 228-237, 2003;

[4] E. M. Knorr and T. Ng. Raymond, Finding intensional knowledge of distance-based outliers, in VLDB, vol. 99, pp. 211-222, 1999.

[5] F. Angiulli, S. Basta and C. Pizzuti, "Distance-based detection and prediction of outliers", IEEE Transactions on Knowledge and Data Engineering, Vol.18, No. 2, 2005, pp. 145-160.

[6] J. Zhang, "Advancements of outlier detection: A Survey", ICST Transactions on Scalable Information Systems, Vol. 13, No. 1, 2013, pp. 1-26.

[7] M. M. Breunig, H. P. Kriegel, R. T. Ng and J. Sander, "LOF: Identifying Density-based Local Outliers" in Proceedings of the 2000 ACM SIGMOD International Conference on Management of data, 2000, pp. 93-104.

[8] S. Ramaswamy, R. Rastogi and K. Shim, "Efficient algorithms for mining outliers from large data sets", in Proceedings of the 2000 ACM SIGMOD International conference on Management of Data, 2000, pp. 427-438.

[9] Z. He, X. Xu, Z.J. Huang and S. Deng, "FP-outlier: Frequent pattern based outlier detection", Computer Science and Information Systems, Vol. 2, No. 1, 2015, pp. 103-118.

[10] C.C. Aggarwal, "Data Mining: Text Book", Springer International Publishing, Switzerland, 2015, pp. 246-248.

[11] P. Murugavel and M. Punithavalli, "Improved Hybrid Clustering and Distance-based Technique for outlier removal", International Journal of Computer Science and Engineering (IJCSE), Vol. 3, No. 1, 2011, pp. 333-339.

[12] S. Cherednichenko, "Outlier Detection in Clustering", University of Joensuu, Department of Computer Science (Doctoral dissertation, Master's Thesis), 2005.

[13] C. H. Teixeira, G. H. Orair, W. Meira Jr and S. Parthasarathy, "An efficient algorithm for outlier detection in high dimensional real databases" in Technical report, University of Minas Gerais, 2008.

[14] G. Singh, F. Masseglia, C. Fiot, A. Marascu and P. Poncelet, "Data mining for intrusion detection: from outliers to true intrusions", In Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, Berlin, Heidelberg, 2009, pp. 891-898.

[15] P. Casas, J. Mazel and P. Owezarski, "UNADA: Unsupervised Network Anomaly Detection using Sub-space Outliers Ranking", in International Conference on Research in Networking, Springer, Berlin, Heidelberg, 2011, pp. 40-51.

[16] M. H. Bhuyan, D. K. Bhattacharyya and J. K. Kalita, "NADO: Network Anomaly Detection using Outlier approach", in Proceedings of the 2011 International Conference on Communication, Computing & Security, 2011, pp. 531-536.

[17] E. De la Hoz, E. De la Hoz, A. Ortiz, J. Ortega, and B. Prie, "PCA filtering and probabilistic SOM for network anomaly detection", Neurocomputing, Vol. 164, pp. 71-81, 2015.

[18] H. Mohamad Tahir, W. Hasan, A. Md Said, N.H. Zakaria, N. Katuk, N.F. Kabir, M.H. Omar, O. Ghazali, and N.I. Yahaya, "Hybrid machine learning technique for intrusion detection system", in Proc. ICOCI, 2015, pp. 464-472.

[19] R. Singh, H. Kumar, and R.K. Singla, "An intrusion detection system using network traffic profiling and online sequential extreme learning machine", Expert Systems with Applications, Vol.42, No.22, 2015, pp. 8609-8624.

[20] M.H. Bhuyan, D.K. Bhattacharyya, and J.K. Kalita, "A multi-step outlier-based anomaly detection approach to network-wide traffic", Information Science, Vol. 348, 2016, pp. 243-271.

[21] S.M.H. Bamakan, H. Wang, T. Yingjie, and Y. Shi, "An effective intrusion detection framework based on MCLP/SVM optimized by time-varying chaos particle swarm optimization", Neurocomputing, vol. 199, 2016, pp. 90-102.

[22] A.C. Enache, and V. Sgarciu, "Anomaly intrusions detection based on support vector machines with an improved bat algorithm", in Proc. CSCS, 2015, pp. 317-321.

[23] E.K. Viegas, A.O. Santin and L.S. Oliveira, "Toward a reliable anomaly-based intrusion detection in real-world environments", Computer Networks, Vol. 127, 2017, pp. 200-216.

[24] S. Aljawarneh, M. Aldwairi and M.B. Yassein, "Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model", Journal of Computational Science, vol. 25, 2018, pp. 152-160.

[25] P.A.A. Resende and A.C.Drummond, "Adaptive anomaly-based intrusion detection system using genetic algorithm and profiling", Security and Privacy, Vol. 1, No. 4, 2018, p.e36.

[26] E. Min, J. Long, Q. Liu, J. Cui and W. Chen, "TR-IDS: Anomaly-based intrusion detection through text-convolutional neural network and random forest", Security and Communication Networks, 2018.

[27] B.A. Tama, M. Comuzzi and K.H. Rhee, "TSE-IDS: A two-stage classifier ensemble for intelligent anomaly-based intrusion detection system", IEEE Access, Vol. 7, 2019, pp. 94497-94507.

RESEARCH ARTICLE

[28] S. Dwivedi, M. Vardhan, S. Tripathi and A.K. Shukla, "Implementation of adaptive scheme in evolutionary technique for anomaly-based intrusion detection", Evolutionary Intelligence, Vol. 13, No. 1, 2020, pp. 103-117.
[29] S. Zavrak, M. Iskefiyeli, Anomaly-based intrusion detection from network flow features using variational autoencoder, IEEE Access, Vol. 8, 2020, pp. 108346-108358.
[30] L. Zhiqiang, L. Zhijun, G. Ting and S. Yucheng, "A Three-Layer Architecture for Intelligent Intrusion Detection using Deep Learning", In Proceedings of Fifth International Congress on Information and Communication Technology, Springer, Singapore, 2021, pp. 245-255.
[31] W. Cui, Q. Lu, A.M. Qureshi, W. Li and K. Wu, "An adaptive LeNet-5 model for anomaly detection", Information Security Journal: A Global Perspective, Vol. 30, No. 1, 2021, pp. 19-29.
[32] Z.A. Bakar, R. Mohemad, A. Ahmad and M.M. Deris, "A comparative study for outlier detection techniques in data mining", in 2006 IEEE Conference on Cybernetics and Intelligent Systems, 2006, pp. 1-6, IEEE.
[33] A.K. Tung, J. Han, L.V. Laskhmanan and R.T. Ng, "Constraint-based clustering in large databases", in International Conference on Database Theory, Springer, 2001, pp. 405-419.
[34] P. Gogoi, D.K. Bhattacharyya, B. Borah and J.K. Kalita, "A survey of outlier detection methods in network anomaly identification", The Computer Journal, Vol. 54, No. 4, 2011, pp. 570-588.
[35] S. Ganapathy, N. Jaisankar, P. Yogesh and A. Kannan, "Intelligent agent-based intrusion detection system using enhanced multiclass SVM", Computational Intelligence and Neuroscience, vol. 2012, 10 pages.
[36] J.R. Beulah and D.S. Punithavathani, "Simple Hybrid Feature Selection (SHFS) for enhancing network intrusion detection with NSL-KDD dataset", International Journal of Applied Engineering Research, Vol. 10, No. 19, 2015, pp. 40498-40505.
[37] J.R. Beulah and D.S. Punithavathani, "A hybrid feature selection method for improved detection of wired/wireless network intrusions", Wireless Personal Communications, Vol. 98, No. 2, 2018, pp. 1853-1869.

Authors

Dr. J. Rene Beulah is an Assistant Professor in the Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, India. She received her B.E. in Computer Science and Engineering from Manonmaniam Sundaranar University, Tirunelveli, India in 2004 and M.E. in Computer Science and Engineering from Anna University, Chennai, India in 2006. She was awarded Ph.D. in the Faculty of Information and Communication Engineering from Anna University, Chennai in 2018. She has 7 years of teaching and 4 years of research experience. Her research interest includes Network Security and Data Mining.

Dr. C. Pretty Diana Cyril is an Assistant professor in the Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, India. She received her B.Tech in Information Technology from Francis Xavier Engineering College, Anna University, India, in 2005 and M.E in Computer Science and Engineering from Sathyabama University, India, in 2010. She awarded Ph.D in Computer Science and Engineering from St.Peter's Institute of Higher Education and Research, Avadi, Chennai, India in 2019. She has 14 years of teaching experience and one year of research experience. Her research interest includes Image processing, Cloud computing, Data Mining & Internet of Things.

Dr. S. Geetha is an Associate Professor in the Department of Information Science and Engineering at CMR Institute of Technology, Bangalore. She obtained her undergraduate, B.Tech (Information Technology) degree from Francis Xavier Engineering College (affiliated to Anna University, Chennai), Tirunelveli in 2005. She had completed her postgraduate, M.Tech (Information Technology) from Satyabama University, Chennai in 2009. She completed her doctorate in the Department of Computer Science and Engineering at Kalasalingam Academy of Research and Education in 2019. She worked as a faculty in different engineering colleges affiliated with Anna University, Tamilnadu as well as Visvesvaraya Technological University, Karnataka. Her research interest includes Wireless sensor networks, Data Mining and Warehousing, and Sensor cloud. She has published her research in many Scopus Indexed Journals, Conferences, and Different Patents. She was also awarded as Young Woman Educator & Scholar Award, Best Researcher Award, and Best Woman Performer in Blockchain Award.

Mrs. D. Shiny Irene is an Assistant Professor in SRM Institute of Science and Technology, Chennai, India. She received her B.E. in Computer Science and Engineering from Anna University, Chennai, India in 2011 and M.E. in Computer Science and Engineering from Anna University, Chennai, India in 2013. She has 7 years teaching experience. She is pursuing her doctorate in Anna University Chennai, India. Her area of interests include Data Mining, Machine Learning, Big Data Analytics and Mobile Computing. She has published many articles in reputed Journals.