Journal of King Saud University – Computer and Information Sciences 31 (2019) 541–553

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

# Anomaly network-based intrusion detection system using a reliable hybrid artificial bee colony and AdaBoost algorithms

Mehrnaz Mazini [a], Babak Shirazi [b,*], Iraj Mahdavi [b]

[a] Department of Information Technology, Mazandaran University of Science and Technology, Babol, Iran
[b] Department of Industrial Engineering, Mazandaran University of Science and Technology, Babol, Iran

ARTICLE INFO

*Article history:*
Received 14 November 2017
Revised 9 March 2018
Accepted 18 March 2018
Available online 27 March 2018

*Keywords:*
Anomaly network-based
Intrusion detection system
Feature selection
Artificial bee colony
AdaBoost

ABSTRACT

Intrusion detection systems (IDSs) has been considered as the main component of a safe network. One of the problems of these security systems is false alarm report of intrusion to the network and intrusion detection accuracy that happens due to the high volume of network data. This paper proposes a new reliable hybrid method for an anomaly network-based IDS (A-NIDS) using artificial bee colony (ABC) and AdaBoost algorithms in order to gain a high detection rate (DR) with low false positive rate (FPR). ABC algorithm is used to feature selection and AdaBoost are used to evaluate and classify the features. Results of the simulation on NSL-KDD and ISCXIDS2012 datasets confirm that this reliable hybrid method has a significant difference from other IDS, which are accomplished according to the same dataset. It has demonstrated differently better performance in different attacks-based scenarios. The accuracy and detection rate of this method has been improved in comparison with legendary methods.


## 1. Introduction

Due to increase in Internet attacks which cause numerous damages, security of network activities highly considered in the computer networks. These networks use various security systems such as IDS to deal with attacks. IDSs are usually used along with firewalls and act as a supplement for them. This security system has been used for observation and analysis of incidents that seriously violate or threaten computer security policies in computers and networks (Gauthama Raman et al., 2017). In general, IDSs purpose is for detecting attacks and security bugs and announcing it to the administrator.

### 1.1. Problem statement

An IDS should be able to identify all abnormal patterns and traffic using monitoring, detecting and responding to unauthorized activities within the system. However, regarding its huge and unbalanced datasets, IDS encounters total data processing problem (Singh et al., 2015). Thus, different techniques have been presented which can handle this problem.

### 1.2. Contribution

Given the importance of developing a new method to deal with the illustrated problem, this paper presents the development of new approach. This purpose can be achieved by answering the following research questions:

RQ1) Which aspects of IDS classification have been considered for developing the new approach?

RQ2) What is the best algorithm for developing the new approach?

RQ3) What are the main selected criteria in approach evaluation?

As Table 1 shows, in terms of collection information, the information is divided into two main categories such as network-based and host-based. In the network-based IDS, all packets in the network are checked. This detection system can monitor traffic only on a specific part of the network and is independent of installed operating systems. Host-based IDS is deployed on a local machine or system to collect information about machine host activities. Also, in terms of detection methods, it is categorized as a general

* Corresponding author.
*E-mail address:* shirazi_b@icloud.com (B. Shirazi).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2018.03.011
1319-1578/

**Table 1**
IDS classification.

| IDS classification | Information source | N-IDS | H-IDS | Misuse detection | Anomaly detection | Passive response | Active response |
|---|---|---|---|---|---|---|---|
| Information source | ■ | ■ | ■ | | | | |
| Detection methods | | | | ■ | ■ | | |
| Reaction on intrusion | | | | | | ■ | ■ |

method (misuse detection and anomaly detection) (Guo et al., 2016). In misuse detection method, pre-built intrusion patterns as law are kept on a database that each pattern represents a specific type of attack. Therefore, the detection tries to find patterns similar to patterns which keep in the database and is able to detect the known intrusions. In this case, new attacks cannot detect in the network (because of no patterns exist in a database) (Kim et al., 2014). As a result, this method contains high-accuracy rate and low false alarm rate for detecting attacks which their patterns exist within the database.

In anomaly detection method, decisions are made based on network normal behavior or features. Therefore, a model of network normal behavior is generated and each traffic event or stream that significantly violates this model is considered an intrusion (Qassim et al., 2016). This type of IDS classification is able to detect new and unknown attacks, but since it is difficult to distinguish the boundary between normal and abnormal behavior, it has high false alarm rate. Consequently, the first research question focuses on one of the important aspects of IDS classification which called anomaly detection method. Moreover, the second research question serves to investigate the different algorithms for proposing the best option. The main contribution of this paper is an improvement a number of false alarms and two main criteria in IDS evaluation which they are the third research question aims. Also, Fig. 1 shows the proposed taxonomy to describe anomaly based intrusion detection system (Gyanchandani et al., 2012).

In order to develop the anomaly network-based IDS, a novel approach is defined Fig. 2).

As Fig. 2 shows, this new approach consists of the following main steps:

- Different attacks selection and definition,
- Hybrid computer network topology design,
- Anomaly based IDS technique selection,
- Appropriate algorithm selection and definition to improve anomaly network-based IDS behavior,
- Appropriate dataset selection.

This paper has been structured as follows: the next section provides the literature review. Section 3 defines the proposed approach. Section 4 evaluates the performance of the anomaly network-based IDS using proposed approach. Finally, Section 5 presents the conclusions.

## 2. Literature review

IDS are security management system which used to identify anomalous activities and incomplete signatures within computers or networks. The number of methods and frameworks has been proposed and many systems have been built to detect intrusions (Sujitha and Kavitha, 2015). Table 2 shows an overview of IDSs design.

In Hasani et al. (2014), work is based on the enhancement of the highest Detection Rate (DR) algorithm which is Linear Genetic Programming (LGP) reducing the False Alarm Rate (FAR) incorporates with Bees Algorithm. Finally, Support Vector Machine (SVM) is one of the best candidate solutions to settle IDSs problems. In this study, four sample dataset containing 4000 random records are excluded randomly from this dataset for training and testing purposes. Experimental results show that the LGP_BA method improves the accuracy and efficiency compared with the previous related research and the feature subcategory offered by LGP_BA gives a superior representation of data (Hasani et al., 2014). In Gupta and Shrivastava (2015); the authors propose a new approach of combining SVM and Bee Colony to achieve high quality performance of IDS. Their algorithm is implemented and evaluated using a standard benchmark KDD99 dataset. Experimental results show that SVM with Bee colony achieves an average accuracy is 88.46% (Gupta and Shrivastava, 2015). In Kim et al. (2014), a new hybrid intrusion detection method that hierarchically integrates a misuse detection model and an anomaly detection model in a decomposition structure is proposed. First, a misuse detection model is built based on the C4.5 decision tree algorithm and then

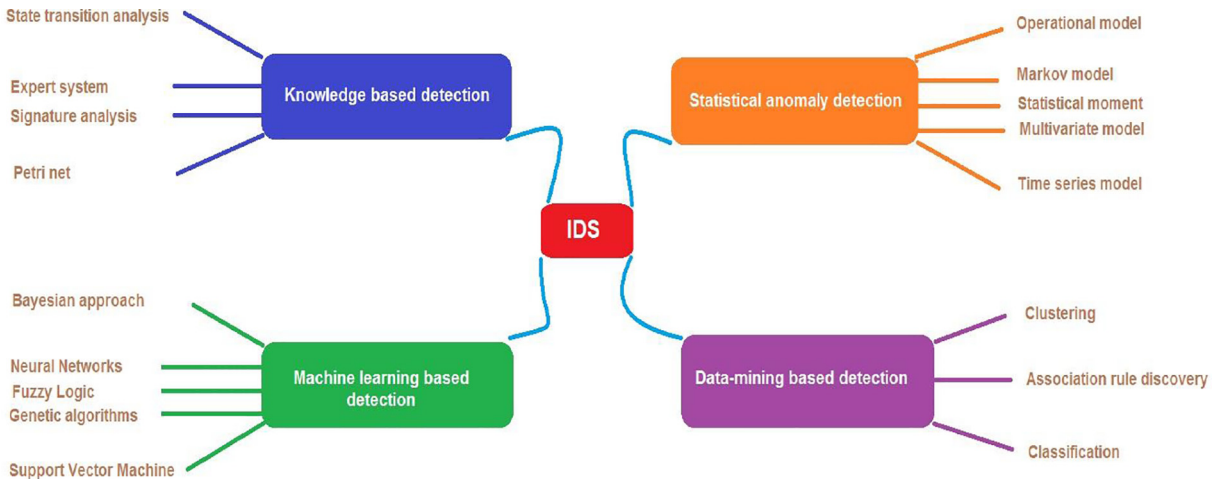


**Fig. 1.** Proposed taxonomy of anomaly based IDS.

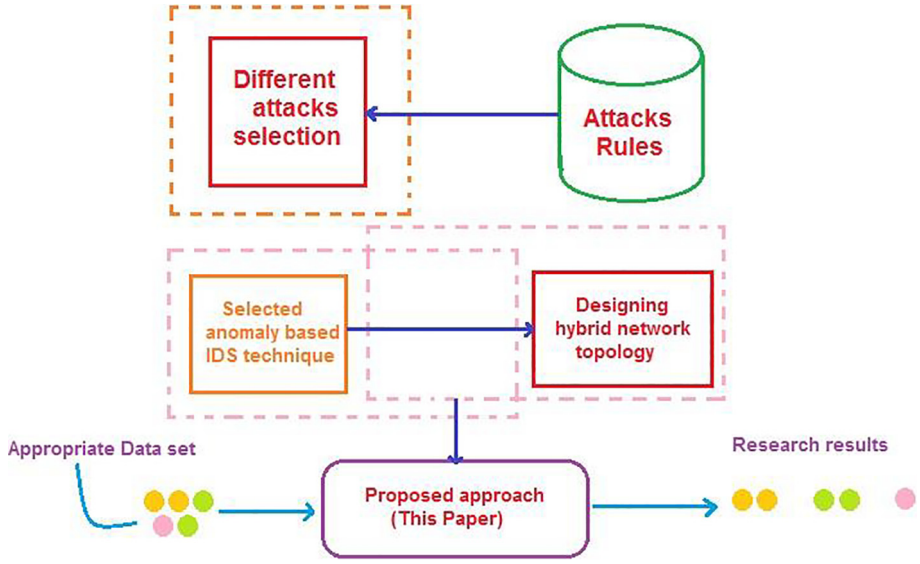


**Fig. 2.** Research stages.

**Table 2**
An overview of researches of IDS design.

| Authors (Year) | Used Algorithms | Method | Results |
|---|---|---|---|
| Hasani et al. (2014) | LGP-BA, SVM | The LGP-BA algorithm was used to feature selection and the SVM was used to categorize the acquired features. | Increased accuracy and efficiency. |
| Gupta and Shrivastava (2015) | BC, SVM | SVM was used to classify normal attacks and BC to enhance performance improvements in IDS. | Increased accuracy. |
| Kim et al. (2014) | SVM, C4.5 | The combination of misuse detection and anomaly detection methods is used to detect intrusions. | Reduces the high time complexity of the training and testing processes. |
| Guo et al. (2016) | k-NN, k-means | K-NN and K-means algorithms are used to reduce FPR and FNR. | Detect network anomalies effectively with a low false positive rate. |
| Hu et al. (2008) | AdaBoost | An Intrusion Detection Algorithm is introduced based on the AdaBoost algorithm. | Low computational complexity and error rates. |
| Mazraeh et al. (2016) | AdaBoost, j48, IG | The main learning algorithms, SVM, Bayes Naive, and J48, have been used for feature categorization. | The superiority of the efficiency of the proposed method. |
| Singh et al. (2015) | OS-ELM | A technique based on the Online Sequential Extreme Learning Machine (OS-ELM) is presented. | Outperforms other published techniques in terms of accuracy, false positive rate and detection time. |
| Al-Yaseen et al. (2016) | SVM, ELM, k-means | A multi-level hybrid intrusion detection model is presented using a combination of K-means, SVM, and ELM algorithms. | High efficiency in attack detection and its accuracy is the best performance thus far. |
| Sujitha and Kavitha (2015) | Layered MPSO | A multi-objective particle swarm optimization algorithm is used for feature selection. | The system is highly robust and efficient and can deal with real-time attacks and detect them a fast and quick response. |
| Horng et al. (2011) | SVM, BIRCH | An SVM-based intrusion detection system with BIRCH algorithm is proposed. | Detects DoS and Probe attacks better than previous methods. |

the normal training data is decomposed into smaller subsets using the model. Next, multiple one-class SVM models are created for the decomposed subsets. As a result, each anomaly detection model does not only use the known attack information indirectly but also builds the profiles of normal behavior very precisely. The proposed hybrid intrusion detection method was evaluated by conducting experiments with the NSL-KDD data set. The experimental results demonstrate that the proposed method is better than the conventional methods in terms of the detection rate for both unknown and known attacks while it maintains a low false positive rate. In addition, the proposed method significantly reduces the high time complexity of the training and testing processes. Experimentally, the training and testing time of the anomaly detection model is shown to be only 50% and 60%, respectively, of the time required for the conventional models (Kim et al., 2014). In Guo et al. (2016), the authors proposed a hybrid approach toward achieving a high detection rate with a low false positive rate. The approach is a two level hybrid solution consisting of two anomaly detection components and a misuse detection component. In stage 1, an anomaly detection method with low computing complexity is developed and employed to build the detection component. The k-nearest neighbor's algorithm becomes crucial in building the two detection components for stage 2. In this hybrid approach, all of the detection components are well-coordinated. The detection component of stage 1 becomes involved in the course of building the two detection components of stage 2 that reduce the false positives and false negatives generated by the detection component of stage 1. Experimental results on the KDD99 dataset and the Kyoto University Benchmark dataset confirm that the proposed hybrid approach can effectively detect network anomalies with a low false positive rate (Guo et al., 2016). In Hu et al. (2008), an intrusion detection algorithm based on the AdaBoost algorithm was proposed. In the algorithm, decisions are used as weak classifiers. The decision rules are provided for both categorical

and continuous features. By combining the weak classifiers for continuous features and the weak classifiers for categorical features into a strong classifier, the relations between these two different types of features are handled naturally, without any forced conversions between continuous features. Adaptable initial weights and a simple strategy for avoiding overfitting are adopted to improve the performance of the algorithm. Experimental results show that this algorithm has low computational complexity and error rates, as compared with algorithms of higher computational complexity, as tested on the benchmark sample data (Hu et al., 2008). In Mazraeh et al. (2016), the proposed method uses a training set of KDD-Cup99. The proposed method uses three main learning algorithms, SVM, Naive Bayes and the J48 decision tree is implemented and evaluated separately. These algorithms are also implemented and evaluated individually as well. The results show the superiority of the proposed method with 97% efficiency using J48 learning algorithm and AdaBoost classification by reducing the dimension IG method (Mazraeh et al., 2016). In Singh et al. (2015), a technique based on the Online Sequential Extreme Learning Machine (OS-ELM) is presented for intrusion detection. The proposed technique uses alpha profiling to reduce the time complexity while irrelevant features are discarded using an ensemble of filtered, correlation and consistency based feature selection techniques. Instead of sampling, beta profiling is used to reduce the size of the training dataset. For performance evaluation of proposed technique, the standard NSL-KDD 2009 (Network Security Laboratory-Knowledge Discovery and DataMining) dataset is used. In this paper time and space complexity of the proposed technique is also discussed. The experimental results yielded an accuracy of 98.66% with a false positive rate of 1.74% and a detection time of 2.43 s for binary class NSL-KDD dataset. The proposed IDS achieve 97.67% of accuracy with 1.74% of the false positive rate in 2.65 s of detection time for the multi-class NSL-KDD dataset. The Kyoto University benchmark dataset is also used to test the proposed IDS. The accuracy of 96.37% with a false positive rate of 5.76% is yielded by the proposed technique. The proposed technique outperforms other published techniques in terms of accuracy, false positive rate and detection time. Based on the results, it is concluded that the proposed method is an efficient method for detecting network intrusion (Singh et al., 2015). In Al-Yaseen et al. (2016), the study aims to design a model that deals with real intrusion detection problems in data analysis and classify network data into normal and abnormal behaviors. It proposes a multi-level hybrid intrusion detection model that uses support vector machine and extreme learning machine to improve the efficiency of detecting known and unknown attacks. A modified K-means algorithm is also proposed to build a high-quality training dataset that contributes significantly to improving the performance of classifiers. The modified K-means is used to build new small training datasets representing the entire original training dataset, significantly reduce the training time of classifiers, and improve the performance of intrusion detection system. The popular KDD Cup 1999 dataset is used to evaluate the proposed model. Compared with other methods based on the same dataset, the proposed model shows high efficiency in attack detection, and its accuracy (95.75%) is the best performance thus far (Al-Yaseen et al., 2016). In Sujitha and Kavitha (2015), The proposed work focuses on accuracy and efficiency. One way to improve performance is to use a minimal number of features to define a model in a way that it can be used to accurately discriminate normal from anomalous behavior. So the new system uses an optimized feature selection algorithm to produce the reduced set of features and high attack detection accuracy can be achieved by using a layered approach. The feature selection algorithm used in the proposed system is a multi-objective particle swarm optimization algorithm which does the feature selection effectively. The layered approach is effectively applicable to detect anomaly attack. The proposed system is tested with the benchmark KDD '99 intrusion dataset as well real-time captured data set, which outperforms other well-known methods such as the decision trees, naive Bayes and Ant Colony optimization. The system is highly robust and efficient. It can deal with real-time attacks and detect them fast and quick response (Sujitha and Kavitha, 2015). Some of the researches have specifically did on Denial of Service (DoS) attacks detection which have been shown in Table 3.

In Horng et al. (2011); the authors proposed an SVM-based intrusion detection system, which combines a hierarchical clustering algorithm, a simple feature selection procedure, and the SVM technique. The hierarchical clustering algorithm provided the SVM with fewer, abstracted, and higher-qualified training instances that are derived from the KDD Cup 1999 training set. It was able to greatly shorten the training time, but also improve the performance of resultant SVM. The simple feature selection procedure was applied to eliminate unimportant features from the training set so the obtained SVM model could classify the network traffic data more accurately. The famous KDD Cup 1999 dataset was used to evaluate the proposed system. Compared with other intrusion detection systems that are based on the same dataset, this system has been shown better performance in the detection of DoS and Probe attacks, and the best performance in overall accuracy (Horng et al., 2011). In Munivara et al. (2017), the authors proposed a method for detecting DDOS attacks in real time at the level of the application layer. They devised a new set of tools for teaching and testing their model, derived from the absolute time interval criterion. Their method uses cuckoo, bat, and shark algorithms. As a result, the Cuckoo's Binary Clustering strategy minimizes the cost overhead of the two algorithms and increases the accuracy of prediction (Munivara et al., 2017). In

**Table 3**
An overview of researches on DoS attacks detection.

| Authors (Year) | Method | Results |
|---|---|---|
| Horng et al. (2011) | An SVM-based intrusion detection system with BIRCH algorithm is proposed. | Detects DoS and Probe attacks better than previous methods. |
| Munivara et al. (2017) | Using cuckoo, bat and shark algorithms to detect a DDOS attack in real time at the application layer level. | Cuckoo search strategy reduces the overhead of computational cost and increases the predicted accuracy. |
| Alfantookh(2006) | An intrusion detection system called Denial of Service Intelligent Detection (DoSID) is developed. | The system was successful in identifying known and unknown attacks and caused a significant increase in false negatives. |
| Singaravelan et al. (2017) | Proposed an Inner Interruption Discovery and Defense System (IIDDS) at the System Call (SC) level using data mining and forensic techniques. | Has had a better response time, intrusion accuracy, and alarm accuracy. |
| Kaur et al. (2017) | Measured the impact of different variants of pulsating distributed denial of service attacks on the self-similar nature of the network traffic. | Pulse lengths of DoS attacks play a pivotal role in deciding the force of the PDDoS attack. |
| Venkatesan et al. (2013) | proposed a cookie based accounting model for accounting the client history. | Achieves more efficiency compared to the existing model. |

Alfantookh (2006); an intrusion detection system called Denial of Service Intelligent Detection (DoSID) is developed. The author proposed model is the feed-forward neural network and related improvements, such as the gray area that uses the concept of distribution. The neural network is composed of two layers. The dataset used is the DARPA suite, which uses its 18 features to identify attacks. The results showed that the feed-forward neural network was successful in identifying known and unknown attacks, and the improvement of the gray area caused a significant increase in false negatives. In Singaravelan et al. (2017), the authors proposed an IIDDS at the System Call (SC) level using data-mining and forensic techniques. By maintaining user profiles and based on system call patterns, the system is prevented from attack and intruders. The proposed model, in contrast to the TF-IDF model, has had a better response time, accuracy, and alarm accuracy (Singaravelan et al., 2017). In Kaur et al. (2017), The authors have examined whether the change in the H index can affect the destruction of the PDDOS attack. The NS2 simulator was used to generate network traffic in various scenarios using the change in traffic, pulse length and loading speed. A congestion control algorithm has been used to model PDDOS attacks. Self-similarity is used as a feature of traffic flows that are measured using the hurst parameter to detect a variety of PDDOS attacks. Self-similarity correction refers to certain features of objects that do not change with scale. They showed that in the event of an attack, the level of self-similarity increases, which results in an increment in the H parameter. Increasing self-similarity depends on the length, duration, and frequency of PDDOS attack pulses (Kaur et al., 2017). Given that the DoS attack on critical infrastructures such as government websites is a major issue, the authors proposed a cookie based accounting model in Venkatesan et al. (2013). Also, they have analyzed all the accounting models including the proposed accounting model based on qualitative and quantitative results to prove the effectiveness of the proposed model. The results show that the proposed model achieves more efficiency compared to the existing model (Venkatesan et al., 2013). Many studies have been conducted on intrusion detection and used from the multiple machine learning algorithms, such as classification and clustering, or combining them with different ideas along with a feature selection approach. Considering the researches done in the past, the issue of reducing false alarms, as it should not have been considered, and there seems to be a gap in this area. So, in this article, we have been proposed the novel approach to resolve the gap and provide the proposed method as followings.

## 3. Anomaly network-based IDS using proposed approach

Given the previous selection, we have introduced a novel approach to improve anomaly network-based IDS behavior. This approach has been implemented on the hybrid computer network topology (Rodas and Antonio, 2015; Carabas et al., 2016; Yao, 2016) as Fig. 3 shows.

The proposed approach will discuss in the following subsections.

### 3.1. The proposed approach

The proposed approach consists of three main phases: preprocessing, feature selection, and classification. As Fig. 1 has been shown, data-mining based detection is one of the anomaly network-based IDS technique. The classification of this category used for developing a proposed approach in this paper. As illustrated earlier, detection is a classification task which includes a prediction model for detecting attacks. In general, classification is one of the main usages in data mining domain. The main goal of it is to predict a class label of each sample according to explained information in features (Ghanem and Jantan, 2016). Unfortunately, an unbalanced data set is an important problem for classification algorithms (Moayedikia et al., 2017). When a dataset is unbalanced, it causes low performance in data-mining algorithms. Thus, an appropriate model is seriously required to overcome this problem. Furthermore, there are plenty of features with high complexity and irrelevant or redundancy states. These states may reduce classification performance due to a big search space. In this case, feature selection methods could be used to prevent these irrelevant redundant without reducing IDS performance. Detecting features related to the learner's classifier makes it easy to categorize, reduces operation time and improves classification performance and accuracy rate. Further, IDS dataset size is huge and takes time to be classified. If a dataset has a lot of items and features, consumes resources and takes a lot of memory to run. Therefore,

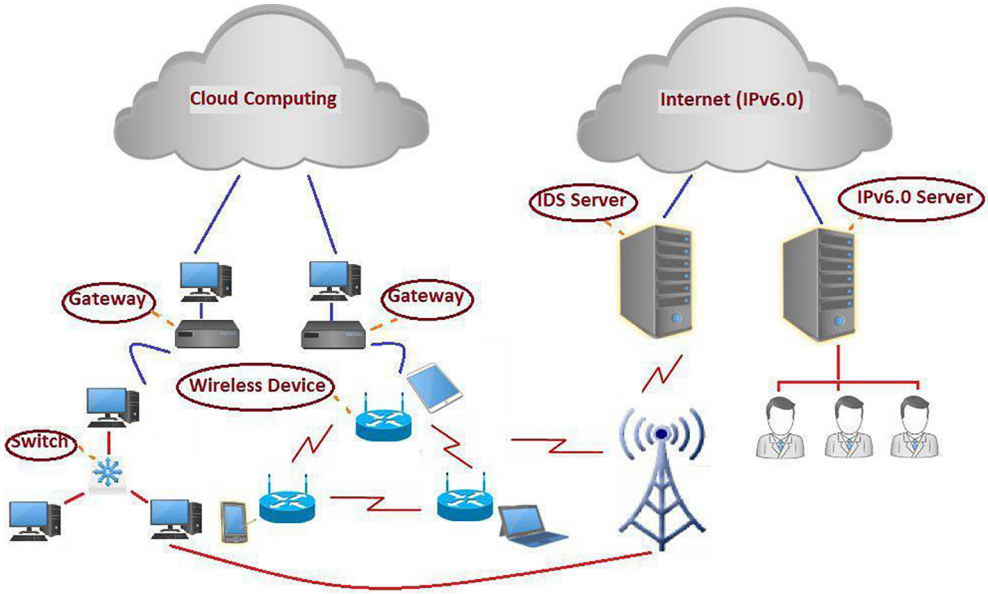


**Fig. 3.** Hybrid computer network topology.

feature selection for IDS dataset is highly essential because they usually include a great number of features and samples.

Feature selection is one of the steps of the data classification process. Feature selection that is known as dimensions reduction is a way to select a new optimal subset of features that represents a set of main features having the least error in learning the classification model. (Sujitha and Kavitha, 2015; Zorarpaco and Ozel, 2016). Feature selection algorithms are evaluated according to two major procedures that are generation procedure and evaluation function (Ghanem and Jantan, 2016). The first step produces the subset of features and the second step evaluates this produced features subset.

The evaluation of a feature subset can be done using one of the following methods: filter method or wrapper method. For depending evaluation function on the machine learning algorithm, i.e. a classifier is used to evaluate the subset of the generated feature, this method is called wrapper feature selection. When the subset of the feature is evaluated according to their information content or statistical methods without using machine learning algorithms, feature selection is known as the filter method (Zorarpaco and Ozel, 2016). Since filter method has lower computational cost, it is usually faster than wrapper method. However, wrapper method often performs better and more accurate than filter method because of more features selection than main feature set (Zorarpaco and Ozel, 2016). Additionally, IDSs should efficiently and accurately detect network attacks. Data-mining based intrusion-detection models attempt to increase detection rate. The problem unbalanced data set often happen in network attacks data, which can reduce the accuracy of IDS. Boosting is a meta-algorithm in the data-mining field that is employed to reduce unbalance and variance. Although the boosting is not in the framework of the algorithm, most of the algorithms are designed based on the boosting, train weak learners repeatedly and add them to the previous set to finally achieve a strong classifier. AdaBoost that is an abbreviation of Adaptive Boosting is a boosting algorithm and one of the most important multiple classification methods because of benefits strong theoretical foundations benefits, highly accurate computations, and simplicity.

As discussed earlier, ABC algorithm is used to select features and AdaBoost is employed for feature evaluation and classification. Fig. 4 shows the proposed approach block diagram in details. As Fig. 4 shows, since dataset includes numerical and non-numerical features or strings, they should be homogenized. Therefore, dataset must be, firstly, preprocessed. Here, preprocessing consists of two stages: first, converting dataset non-numerical features into a numerical amount; then, data normalization. Because NSL-KDD dataset features consist of discrete or continuous amounts, features amount are located at different intervals; so, these features are incomparable. Therefore, normalization is used to normalize features, and all numbers are limited to an interval [0,1].

The proposed approach increases the accuracy and the speed up detection time by ABC feature selection technique. The purpose of this approach is to find the best features for IDS classification. In a colony, each type has its own tasks (Shi et al., 2016). According to this algorithm, a food source indicates a solution (for instance, food resource location) related to the problem and the food source nectar amount shows the quality of the solution (for instance, fitness). Fig. 4 displays ABC algorithm. ABC process includes four phases: initialization phase, employee bee phase, onlooker bee phase, and scout bee phase (Bansal et al., 2013). In ABC algorithm, a group of bee populations is created that half of them employ bees, and the other half are onlooker bees. In this algorithm, first, a population of NF solutions (number of solution) is generated, where each solution $X_i (i = 1, 2, \ldots, NF)$ is a D-dimensional vector. Here, $X_i$ indicates $i$th food source in the colony. There is only one employee bee for each food source (solution). In other words, a number of employee or onlooker bees is equal to the number of solutions (Zorarpaco and Ozel, 2016). Each solution is generated randomly according to Eq. (1) (Ghanem and Jantan, 2016). Then, its fitness is calculated and the best solution is saved. In this paper, since we used the wrapper method to select the feature, hence AdaBoost has been used to evaluate fitness.

$$x_i^j = x_{min}^j + r \; and \; (0,1)(x_{max}^j - x_{min}^j), \; \forall j = 1, 2, \ldots, D \tag{1}$$

where $D$ is dimension or number of issue parameters, $x_{min}^j$ and $x_{max}^j$ are respectively low limit and high limit of $j$th dimension. Then, each employee bee to change the position, randomly selects a food source (another employee bee) in its neighborhood and moves toward it. Here, regarding Eq. (2) it makes changes to the position existed in its memory (Ghanem and Jantan, 2016). If the new position is better than the current position in terms of its honey or fitness, maintains the new position and forgets the previous position, otherwise, holds the previous position and adds a unit to the trial. The trial is the number of consecutive bees without improvement.

$$v_{i,j} = x_{i,j} + \varnothing_{i,j}(x_{i,j} - x_{k,j}) \tag{2}$$

where $X_{i,j}$ is the previous position, $\varnothing_{i,j}$ is the random amount in the interval $[-1, +1]$, and $x_k$. a neighborhood from $\boldsymbol{x_i}$. The fitness of each solution is calculated by the following equation:

$$fit_i = \begin{cases} \frac{1}{(1+f(X_i))} & if \quad f(x_i) \geqslant 0 \\ 1 + |f(x_i)| & if \quad f(x_i) < 0 \end{cases} \tag{3}$$

where $f(x_i)$. objective function value of $i$th solution.

After exploration processes all of the employee bees be completed, they share fitness value of updated solutions with onlooker

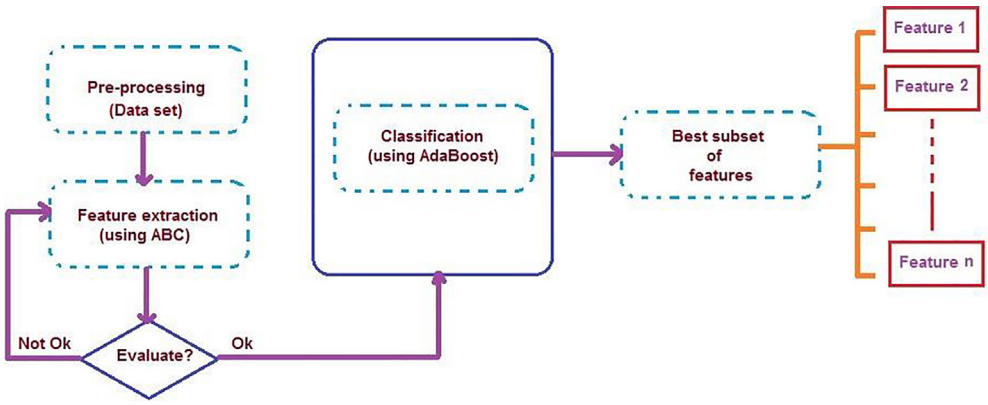


**Fig. 4.** Proposed approach block diagram.

bees. These bees choose a food source based on the probability of that food source. The probability of selecting any solution is calculated using Eq. (4) (Ghanem and Jantan, 2016):

$$p_i = \frac{fit_i}{\sum_{i=1}^{NF} fit_i} \tag{4}$$

where $fit_i$ is fitness value of $i$th solution that has been calculated by employee bee. Then the neighborhood of solutions chosen is randomly found by Eq. (4), and using Eq. (2), paid to search for a better solution around it. In this way, the variation decreases in the position $x_{i,j}$. Therefore, a search for the optimal solution is near in the search space. After each exploration using Eq. (2), if the position of previous food source is not better than the new one, a unit will be added to trial. Finally, to prevent being trapped in local optimum if the trial of a food source after several consecutive repetitions, where this number is predetermined (limit) is not updated, then the given food source will be left, and the employee bee related to being changed the abandoned source to scout. Then, a new food source is selected by use of random exploration according to Eq. (1) and replaced the abandoned source, after the trial value of the abandoned source is zero. Finally, the best food source is calculated from this repetition, and if it is better than the best food source of a whole algorithm, it will replace. In the proposed approach, the number of available solutions randomly selected a number of features, and the best solution is recorded. Fig. 5 has been shown the pseudo-code of ABC algorithm.

To evaluate the attributes, they have been sent to the AdaBoost function to be checked by the test dataset. Different types of AdaBoost algorithm have been provided. Standard AdaBoost is designed for binary classification issues, so it cannot be used for multiclass issues. Since intrusion-detection dataset contains multiclass, AdaBoost.M2 has been used specifically for multiclass issues. This algorithm utilized the concept of pseudo-loss that measure the goodness of the weak hypothesis. The pseudo-code of AdaBoost.M2 is shown in Algorithm 1. According to this Figure, decision trees with 5 leaves are used as base learners. Firstly, all samples have equal weight and change sample's weight in each T. The aim of weak learners is to minimize pseudo-loss in step 3. If a weak learner can constantly produce to generate a weak hypothesis with a pseudo-loss less than ½, it will be strengthened.

In Eqs. (5) and (6), respectively, the AdaBoost function by these attributes creates a behavior model of network users. This model predicts that the test datasets belong to which types of normal or attack classes.

$$[MDL] = adaboost(X, t) \tag{5}$$

$$pre = predict(MDL, XTest). \tag{6}$$

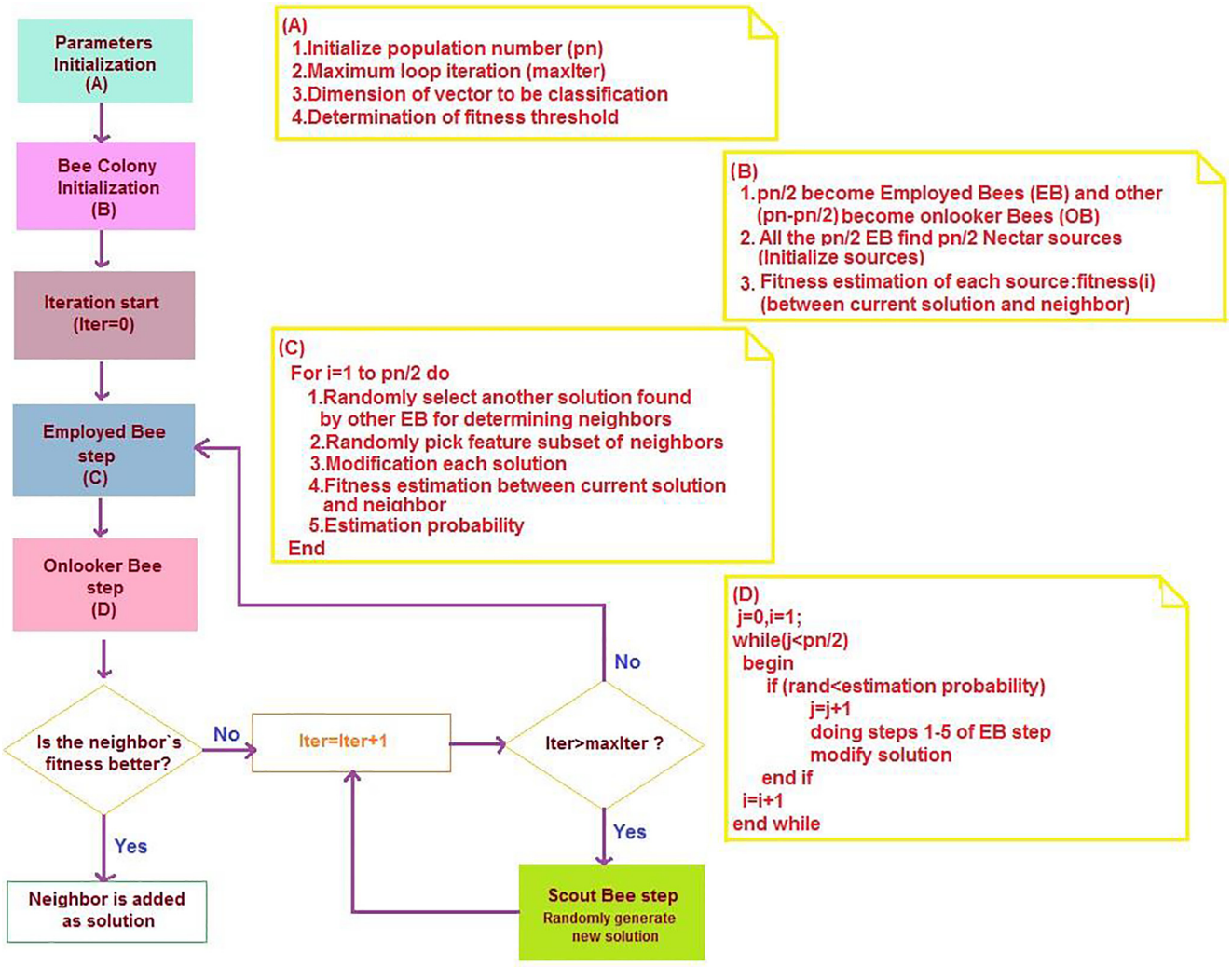


**Fig. 5.** The pseudo-code of applying ABC algorithm in proposed approach.

where X and t represent the training dataset and the labels of the samples, respectively, and XTest is a test dataset. By the Eq. (7), the amount of model error with reality is determined. This is repeated in the maximum number of searches. Ultimately, the best features that affect the performance of the IDS are identified.

$$L = loss(MDL, XTest, GroupTest) \tag{7}$$

where GroupTest is the test set label. In Section 4, the best features that are chosen by the proposed approach are given. The fundamental and important point is to set the parameters according to them the detection operation takes place. These factors have a great influence on operation time and problem efficiency

Algorithm 1: The Pseudo-code of applying AdaBoost.M2 algorithm in proposed approach

*Input: sequence of m examples* $\langle (x_1, y_1), \ldots, (x_m, y_m) \rangle$ *with labels* $y_i \in Y = \{1, \ldots, k\}$.

*Weak learning algorithm DT*

*T (number of iterations)*

*Initialize* $\boldsymbol{D}_1(\boldsymbol{i}) = 1/\boldsymbol{m}\ \boldsymbol{i} = \boldsymbol{1}, \ldots, \boldsymbol{m}$ *and* $\boldsymbol{w}^{\boldsymbol{1}}_{\boldsymbol{i,y}} = \boldsymbol{D}(\boldsymbol{i})/(\boldsymbol{k} - \boldsymbol{1});\ \boldsymbol{y} \neq \boldsymbol{y_i}$.

*For* $\boldsymbol{t} = \boldsymbol{1}, \boldsymbol{2}, \ldots, \boldsymbol{T}$ *do*

1. $\boldsymbol{W^t_i} = \sum_{y \neq y_i} \boldsymbol{W^t_{i,y}}$, $\boldsymbol{q_t(i,y)} = \frac{\boldsymbol{W^t_{i,y}}}{\boldsymbol{W^t_i}}, \forall \boldsymbol{y}$, $\boldsymbol{D_t(i)} = \frac{\boldsymbol{W^t_i}}{\sum_{i=1}^{m} \boldsymbol{W^t_i}}$
2. *DT (Providing it with distribution* $\boldsymbol{D_t}$ *label weighting function* $\boldsymbol{q_t}$
3. $\boldsymbol{h_t} : \varepsilon_t = \frac{1}{2} \sum_{i=1}^{m} \boldsymbol{D_t}(1 - \boldsymbol{h_t}(\boldsymbol{x_i}, \boldsymbol{y_i}) + \sum_{y \neq y_i} \boldsymbol{q_t}(\boldsymbol{i}, \boldsymbol{y}) \boldsymbol{h_t}(\boldsymbol{x_i}, \boldsymbol{y})), \boldsymbol{h_t} : \boldsymbol{X} \times \boldsymbol{Y} \rightarrow [\boldsymbol{0}, \boldsymbol{1}]$
4. $\boldsymbol{\beta_t} = \frac{\varepsilon_t}{(1 - \varepsilon_t)}$
5. *Update* $\boldsymbol{W^{t+1}_{i,y}} = \boldsymbol{W^t_{i,y}} \boldsymbol{\beta_t^{\left(\frac{1}{2}\right)(1 + \boldsymbol{h_t}(\boldsymbol{x_i}, \boldsymbol{y_i}) - \boldsymbol{h_t}(\boldsymbol{x_i}, \boldsymbol{y}))}}$
6. *Output:* $\boldsymbol{h_{fin}(x)} = \boldsymbol{argmax}_{\boldsymbol{y} \in \boldsymbol{Y}} \sum_{t=1}^{T} (\boldsymbol{log}\ \frac{1}{\boldsymbol{\beta_t}}) \boldsymbol{h_t(x,y)}$

**Table 4**
Simulation parameters settings.

| Parameters | Value |
|---|---|
| T as a number of boosting iteration (AdaBoost algorithm) | 200 |
| LR as a learning rate (AdaBoost algorithm) | 0.1 |
| NB as a number of Bees (ABC algorithm) | 50 |
| NF as a number of food sources (ABC algorithm) | 25 |
| Maxcycle as a maximum number of the search processes (ABC algorithm) | 50 |
| Limit as a maximum number of cycles which food source has been kept before releasing it. (ABC algorithm) | 30 |

## 4. Proposed approach validation

### 4.1. NSL-KDD and ISCXIDS2012 (Shiravi et al., 2012) datasets

In order to evaluate the effectiveness and performance of the anomaly network-based IDS using proposed approach, it has been performed the simulation through NSL-KDD (KDD99 modified version) and ISCXIDS2012 datasets. For evaluating, the authors defined different attack-based scenarios into mentioned data sets using designed network topology Fig. 3. Since there are several criteria for IDSs performance, different criteria such as Detection Rate $\left(DR = \frac{TP}{TP+FN}\right)$. False Positive Rate $\left(FPR = \frac{FP}{FP+TN}\right)$, and accuracy $\left(Ac = \frac{TN+TP}{TN+FN+TP+FP}\right)$. ve been calculated to prove the proposed approach. These criteria are calculated according to four main criteria which are True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) as follows:

- True Positive (TP): Indicates when an alarm is generated and there is an intrusion.
- False Negative (FN): Indicates when an alarm is not generated but there is an intrusion.
- False Positive (FP): Indicates when an alarm is generated but there is no intrusion.
- True Negative (TN): Indicates when an alarm is no generated and there is no intrusion.

### 4.2. Simulation results

The proposed approach performance for different scenarios (such as normal and attacks) using NSL-KDD and ISCXIDS2012 has presented as following: (These simulation has been down in MATLAB using simulation parameters which shown in Table 4. Also, the training and testing time of the proposed approach are 2.35 s and 32.55 s respectively. Also, the number of flows seen per second during the capturing period has been shown in Fig. 6. This Figure has been developed based on traffic visual analytic (Zhang et al., 2013) using Wireshark tool. Fig. 7))

- Scenario 1): In this scenario, an attacker can attack to the network using different attacks such as:
  - DOS (Denial of Service): An attacker use computational or memory resources to handle valid requests. Thus, it will

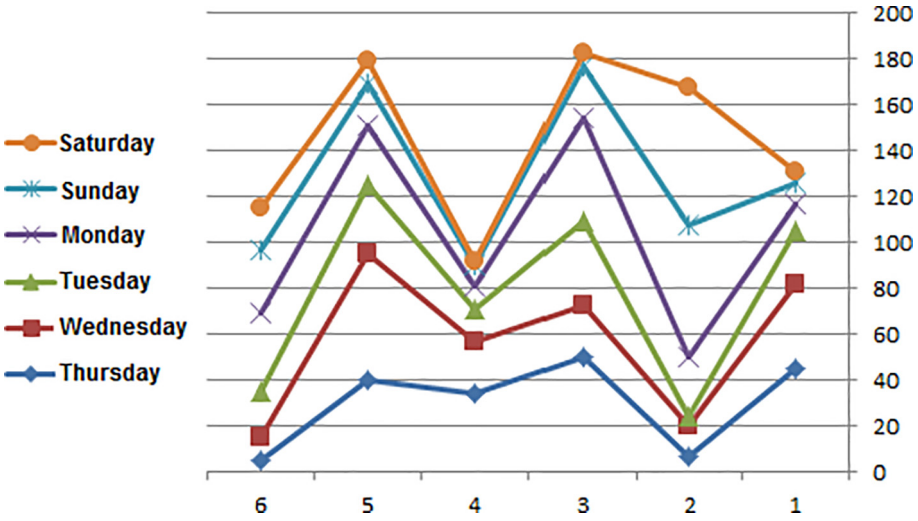


**Fig. 6.** Number of Flows/Second.

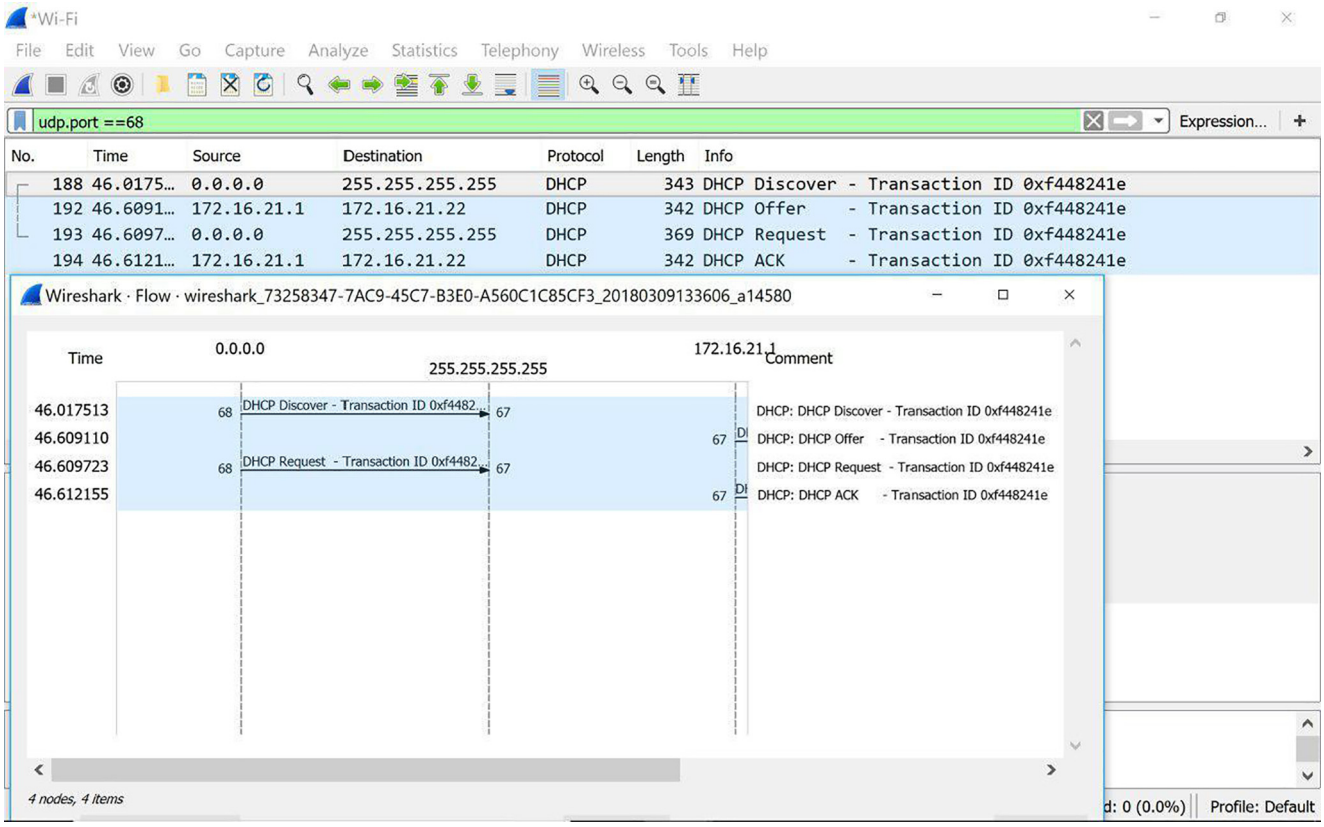


**Fig. 7.** Traffic visual analytic using Wireshark tool.

cause the system in problem which cannot make valid user requests (Aljawarneh et al., 2017).

- R2L (Remote to Local): The attacker has an unauthorized remote access to the system. This attack uses a valid user account.
- U2R (User to Root): The attacker has an unauthorized remote access to the root. This attack uses a valid user account too.
- Probe: The attacker tries for gathering information about computer networks.

Table 5 shows the operation results of the anomaly network-based IDS proposed approach using NSL-KDD dataset for scenario 1.

As Table 5 shows, the proposed approach has been succeeded in attacks detection. A remarkable point is that the features of discussed attacks are very similar to normal traffic and identifying these attacks are difficult. Table 6 shows the proposed approach operation for scenario 1 in details too.

- Scenario 2): In this scenario, an attacker can attack to the network using different attacks methods such as:
  - Infiltering: This attack scenario starts by gathering information about the target including network IP ranges, nameserver and so on. This is achieved by querying for resource records using network administrative tools (Shiravi et al., 2012).
  - Bruteforce SSH: This attack is very common against networks as they tend to break into accounts with weak username and password combinations. This scenario has been designed with the goal of acquiring an SSH account by running a bruteforce attacks against the mail server (Shiravi et al., 2012).
  - Botnet: Botnets combine previous behaviors into a single platform, essentially simplifying and assisting the user of such platform to form sophisticated attacks against workstations or networks around the world. Such behaviors include scanning, Distributed Denial of Service (DDoS) activities and so on (Shiravi et al., 2012).

Table 7 shows the operation results of the anomaly network-based IDS proposed approach using ISCXIDS2012 dataset for scenario 2 in details.

**Table 5**
Proposed approach results for detecting scenario 1 attacks (Part 1).

| | Normal | DOS | R2L | U2R | Probe |
|---|---|---|---|---|---|
| DR% | 98.99 | 99.91 | 99.37 | 99.89 | 99.89 |
| FPR | 0.015 | 0.001 | 0.034 | 0.021 | 0.015 |
| AC% | 98.71 | 99.86 | 97.90 | 98.86 | 99.18 |

**Table 6**
Proposed approach results for detecting scenario 1 attacks (Part 2).

| Traffic type | Number of total samples | Number of true detected samples | Number of false detected samples |
|---|---|---|---|
| Normal | 13,449 | 13,362 | 87 |
| Attack | 9234 | 8501 | 733 |
| Total | 22,683 | 21,863 | 820 |

**Table 7**
Proposed approach results for detecting scenario 2 attacks.

| Traffic type | Number of total samples | Number of true detected samples | Number of false detected samples |
|---|---|---|---|
| Normal | 9711 | 8798 | 913 |
| Attack | 7458 | 5432 | 2026 |
| Total | 17,169 | 14,230 | 2932 |
| True Alarm = 86% | | Accuracy rate = 83% | |
| False Alarm = 14% | | Detection rate = 73% | |
| False Alarm (positive) $\leq$ 5% | | Error rate = 27% | |
| False Alarm (Negative)=3% | | | |

The results obtained by the proposed approach have been compared with other methods. The comparison results are shown in Table 8.

As Table 8 shows, the proposed approach has significant performance improvement in all three important criteria. The ABC algorithm has been able to select the best-related features due to the ability to escape from the optimal local area and as a result; a much better performance is shown in comparison to other methods too.

### 4.3. Sensitivity analysis

Table 9 and Fig. 8 shows the results of separate implementation using the proposed approach in each feature subset for DR, AC and FPR criteria

According to Table 9, it can be presented that amount of T parameter must be at least regulated on 200. When T = 200, it rises by an increase in Max, DR, and AC, but FPR reduces. Also, when maxcycle = 10, the result is improved with increasing T.

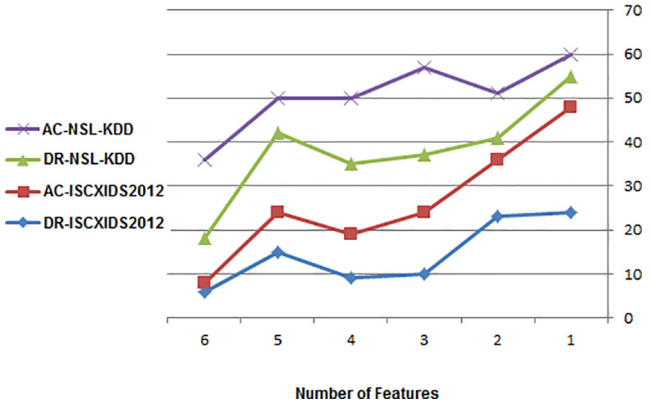


**Fig. 8.** DR and AC with various features using the ABC feature selection algorithm.

Since related features directly affect classification accuracy, the larger number of bee cycle raises the probability of selecting more appropriate features for classification; that is why it has a positive influence on detection optimum. On the other hand, since AdaBoost designation causes error reduction in classification, in case number of its repetition is regulated at a high level it would provide a desirable result even with a low maxcycle. performance in their combination if parameters are correctly regulated.

To investigate the effect of NF, the program run with maxcycle = 10, T = 200 that for the first time used NF = 25 and the second time NF = 50. As a result, obtained with 25 solutions: DR = 98.81, AC = 93.67 and FPR = 0.093, and it's also obtained with 50 solutions: DR = 99.4, AC = 97.5 and FPR = 0.04. Results indicate the positive influence of NF increase. Fig. 9 depicts classification error graph in four different performances.

As it can be seen it has an error below 0.006 that is the best performance than other performances.

The remarkable point is that the proposed approach shows high efficiency in attack detections. As illustrated earlier, simulation results on the selected datasets confirm that the proposed hybrid approach can effectively detect network anomalies with an appropriate time and space complexity rather than legendary methods Table 10).

**Table 8**
Comparing the proposed approach with other methods.

| Classification algorithms | DR% | FPR | AC% | Feature selection method |
|---|---|---|---|---|
| K-NN+K-means (Guo et al., 2016) | 91.86 | 0.78 | 93.29 | All Feature |
| DT (Eesa et al., 2014) | 91.500 | 3.372 | 92.500 | CAT |
| SVM (Gauthama Raman et al., 2017 | 97.14 | 0.83 | N/A | HG – GA |
| AdaBoost [This Paper] | 99.61 | 0.01 | 98.90 | ABC |

**Table 9**
Sensitivity analysis by ABC-AdaBoost.

| No. of Features | DR% | FPR | AC% | Parameters |
|---|---|---|---|---|
| 25 | 99.61 | 0.017 | 98.90 | T = 200, maxcycle = 50 |
| 23 | 99.55 | 0.023 | 98.56 | T = 200, maxcycle = 25 |
| 22 | 98.81 | 0.093 | 93.67 | T = 200, maxcycle = 10 |
| 19 | 99.49 | 0.030 | 98.15 | T = 500, maxcycle = 10 |
| 16 | 97.67 | 0.112 | 92.25 | T = 150, maxcycle = 45 |
| 10 | 76 | 0.4 | 89.5 | T = 250, maxcycle = 45 |
| 25 | 97 | 0.47 | 90 | T = 150, maxcycle = 45 |
| 23 | 83.8 | 0.04 | 91 | T = 200, maxcycle = 45 |

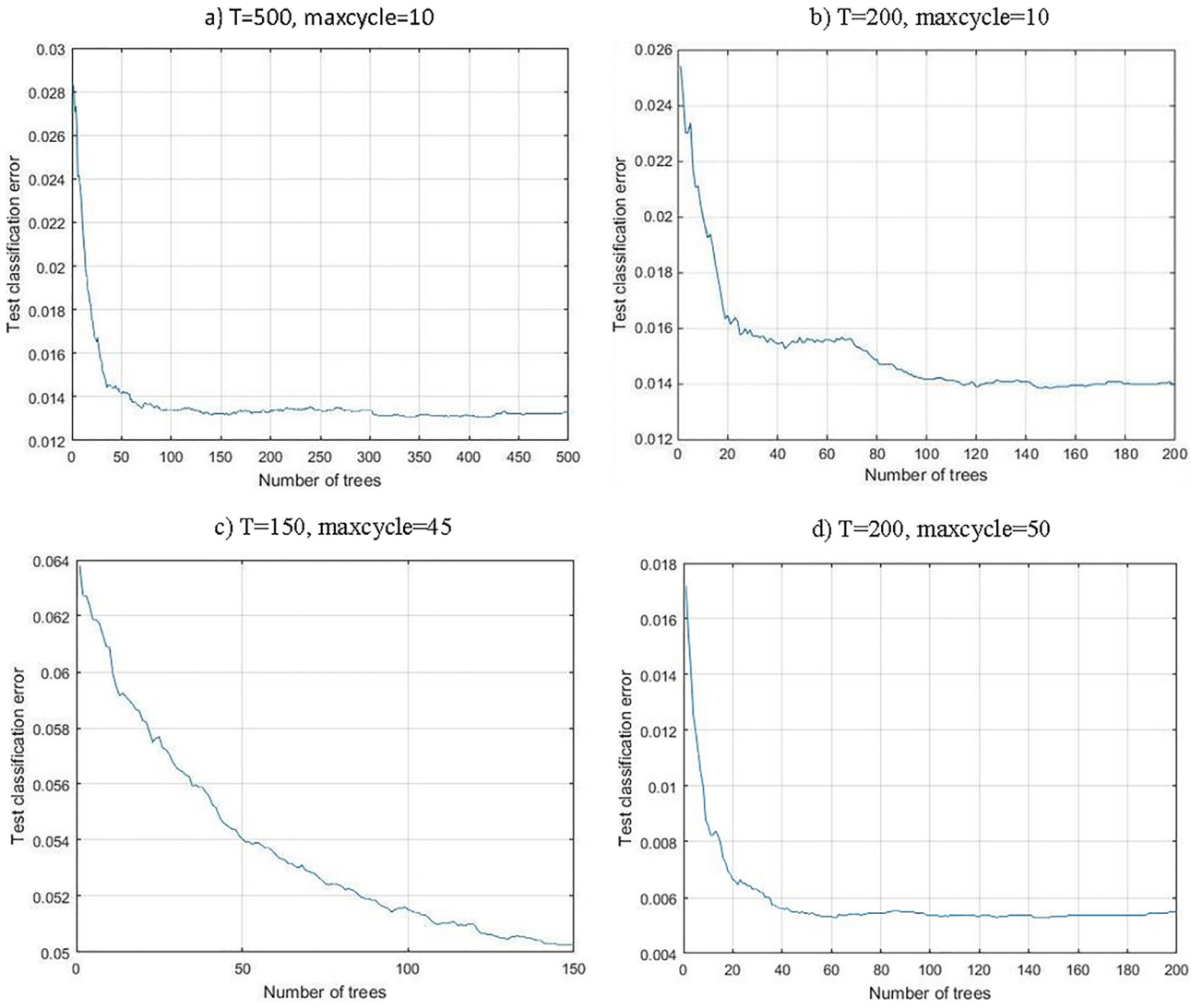


**Fig. 9.** Test classification error in different experiments.

**Table 10**
Time and space complexity results.

| Methods | Time Complexity | Space Complexity |
|---|---|---|
| AdaBoost(2008) (Hu et al., 2008) | O(n) | O(n) |
| SVM + BRICH(2011) (Horng et al., 2011) | $O(n^2)$ | O(n) |
| LGP-BA + SVM(2014) (Hasani et al., 2014) | O(nlogn) | O(n!) |
| SVM + C4.5(2014) (Kim et al., 2014) | O(nlogn) | $O(n^3)$ |
| Layered MPSO (2015) (Sujitha and Kavitha, 2015) | $O(\log n^2)$ | O(n) |
| OS-ELM(2015) (Singh et al., 2015) | $O(\log n^2)$ | O(n) |
| BC + SVM(2015) (Gupta and Shrivastava, 2015) | O(n) | O(n) |
| AdaBoost + j48 + IG(2016) (Mazraeh et al. 2016) | O(n) | O(n) |
| k-NN + k-means(2016) (Guo et al., 2016) | O(n) | O(n) |
| SVM + ELM + k-means(2016) (Al-Yaseen et al., 2016) | O(n) | O(n/2) |
| Proposed approach (This Paper) | O(n/2) | O(n/2) |

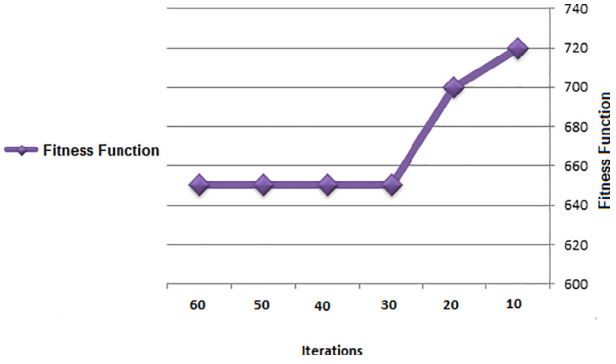


**Fig. 10.** Convergence characteristics of ABC algorithm.

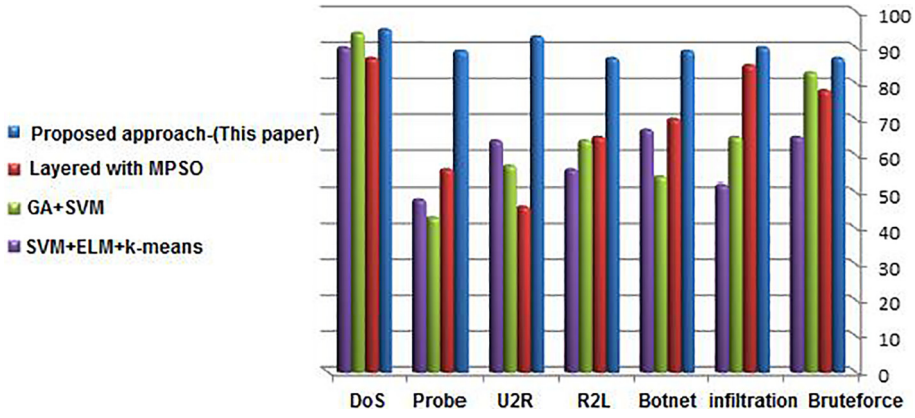


**Fig. 11.** Anomaly detection using different datasets.

Also, Fig. 10 shows the convergence characteristic of the ABC algorithm. As this Figure presented, there is no change in the fitness function after 30 iterations.

The obtained computational time from the ABC algorithm for the anomaly network-based proposed approach show better fitness function than the others algorithm. Thus, the proposed ABC is effective for the anomaly detection problem in the proposed approach.

## 5. Conclusion

In this paper, a reliable approach for anomaly network-based IDS is proposed. Simulations were performed to evaluate the performance of the proposed model on the NSL-KDD and ISCXIDS2012 data sets. The network traffic datasets are large and unbalanced, thus affects the performance of the IDS. The imbalance causes the minority class not to be properly detected by conventional data-mining algorithms. By ignoring the instance of this class, they try to increase overall accuracy, while the correct instance of minority class protocols is also important. So in the proposed approach, the AdaBoost meta-algorithm has been used for unbalanced data according to the proper design. The reason for this assertion is the high precision of the proposed approach to classifying different attacks classes. On the other hand, ABC is a useful meta-algorithm which can be used in optimization IDS problems. The proposed algorithm has been used to select the best subset of related features to detect network connections and because of the high ability of these algorithms. The parameters regulation method is also influential on problem efficiency. Considering the high number of records and data properties, the number of replicates or searches for solutions in the search space by the bee is of great importance. The accuracy and detection rate of this proposed approach has been improved in comparison with legendary methods as shown in Fig. 11.

## References

Alfantookh, A. Abdulkader, 2006. DoS attacks intelligent detection using neural networks. J. King Saud Univ. Comput. Inf. Sci. 18, 31–51.
Aljawarneh, S., Aldwairi, M., Yasin, M.B., 2017. Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model. J. Comput. Sci. https://doi.org/10.1016/j.jocs.2017.03.006.
Al-Yaseen, W.L., Othman, Z.A., Nazri, M.Z.A., 2016. Multi-level hybrid support vector machine and extreme learning machine based on modifed K-means for intrusion detection system. Expert Syst. Appl. 67, 296–303.
Bansal, J.C.H., Sharma, H., Jadon, S.H.S., 2013. Artificial bee colony algorithm: a survey. Int. J. Adv. Intell. Paradigms 5, 123–159.
Carabas, M., Carabas, C., Gheorghe, L., Deaconescu, R., Tapus, N., 2016. Monitoring and auditing mobile operating systems. Int. J. Space-Based Situated Comput. 6, 54–63.
Eesa, A.S., Orman, Z., Brifcani, A.M., 2014. A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems. Expert Syst. Appl. 42, 2670–2679.
Gauthama Raman, M.R., Somu, N., Kirthivasan, K., Lisano, R., 2017. V.S. Shankar Sriram, "An efficient intrusion detection system based on hyper graph Genetic algorithm for parameter optimization and feature selection in support vector machine". Knowl.-Based Syst. https://doi.org/10.1016/j.knosys.2017.07.005.
Ghanem, W.A.H.M., Jantan, A., 2016. Novel multi-objective artificial bee colony optimization for wrapper based feature selection in intrusion detection. Int. J. Adv. Soft Comput. Appl. 8, 70–81.
Guo, C., Ping, Y., Liu, N., Luo, S.S., 2016. A two level hybrid approach for intrusion detection. Neurocomputing 214, 391–400.
Gupta, M., Shrivastava, S.K., 2015. Intrusion detection system based on SVM and bee colony. Int. J. Comput. Appl. 111, 27–32.
Gyanchandani, M., Rana, J.L., Yadav, R.N., 2012. Taxonomy of anomaly based intrusion detection system: a review. Int. J. Sci. Res. Publ. 2, 1–13.
Hasani, S.R., Othman, Z.H., Mousavi Kahaki, S.M., 2014. Hybrid feature selection algorithm for intrusion detection system". J. Comput. Sci. 10, 1015–1025.
Horng, S.J., Su, M.Y., Chen, Y.H., Kao, T.W., Chen, R.J., Lai, J.L., Perkasa, C.D., 2011. A novel intrusion detection system based on hierarchical clustering and support vector machines. Expert Syst. Appl. 38, 306–313.
Hu, W., Hu, W., Maybank, S., 2008. AdaBoost-Based algorithm for network intrusion detection. IEEE Trans. Syst. Man Cybern. B Cybern. 38, 577–583.
Kaur, G., Saxena, V., Gupta, J.P., 2017. Detection of TCP targeted high bandwidth attacks using self-similarity. J. King Saud Univ. Comput. Inf. Sci. https://doi.org/10.1016/j.jksuci.2017.05.004.
Kim, G., Lee, S., Kim, S., 2014. A novel hybrid intrusion detection method integrating anomaly detection with misuse detection. Expert Syst. Appl. 41, 1690–1700.
Mazraeh, S., Ghanavati, M., Neysi, S.H.N., 2016. Intrusion detection system with decision tree and combine method algorithm. Int. Acad. J. Sci. Eng. 3, 21–31.
Moayedikia, A., Ong, K.L., Boo, Y.L., Yeoh, W.G., Jensen, R., 2017. Feature selection for high dimensional imbalanced class data using harmony search. Eng. Appl. Artif. Intell. 57, 38–49.
Munivara Prasad, K., Rama Mohan Reddy, A., Venugopal Rao, K., 2017. BARTD: bio-inspired anomaly based real time detection of under rated App-DDoS attack on web. J. King Saud Univ. Comput. Inf. Sci. https://doi.org/10.1016/j.jksuci.2017.07.004.
Qassim, Q.S., Zin, A.M., Aziz, M.J.A., 2016. Anomalies classification approach for network-based intrusion detection system. Int. J. Netw. Secur. 18, 1159–1172.
Rodas, O., Antonio, M., 2015. A study on network security monitoring for the hybrid classification-based intrusion prevention systems. Int. J. Space-Based Situated Comput. 5, 115–125.
Shi, Y., Pun, C., Hu, H., Gao, H., 2016. An improved artificial Bee colony and its application. Knowledge-Based System 107, 14–31.
Shiravi, A., Shiravi, H., Tavallaee, M., Ghorbani, A.A., 2012. Toward developing a systematic approach to generate benchmark datasets for intrusion detection. Comput. Secur. 31, 357–374.
Singaravelan, S., Arun, R., Arunshunmugam, D., Jerina Catherine Joy, S., Murugan, D., 2017. Inner interruption discovery and defense system by using data mining. J. King Saud Univ. Comput. Inf. Sci. https://doi.org/10.1016/j.jksuci.2017.09.009.
Singh, R., Kumar, H., Singla, R.K., 2015. An intrusion detection system using network traffic profiling and online sequential extreme learning machine. Expert Syst. Appl. 42, 8609–8624.
Sujitha, B., Kavitha, V., 2015. Layered approach for intrusion detection using multiobjective particle swarm optimization. Int. J. Appl. Eng. Res. 10, 31999–32014.

Venkatesan, S., Saleem Basha, M.S., Chellappan, C., Vaish, A., Dhavachelvan, P., 2013. Analysis of accounting models for the detection of duplicate requests in web services. J. King Saud Univ. Comput. Inf. Sci. 25, 7–24.
Yao, X., 2016. The realization of goal-driven airport enclosures intrusion alarm system. Int. J. Grid Util. Comput. 7, 1–6.
Zhang, J., Huang, M.L., Hoang, D., 2013. Visual analytics for intrusion detection in spam emails. Int. J. Grid Util. Comput. 4, 178–186.
Zorarpaco, E., Ozel, S.A., 2016. A hybrid approach of differential evolution and artificial bee colony for feature selection. Expert Syst. Appl. 62, 91–103.