Improving Information Integrity using Artificial Bee Colony based Intrusion Detection System

Qamar Rayees Khan
Department of Computer Sciences, Baba Ghulam Shah Badshah University, Rajouri (J&K), 185234, India

Mohammad Asger
Department of Computer Sciences, Baba Ghulam Shah Badshah University, Rajouri (J&K), 185234, India

Muheet Ahmed Butt
Department of Computer Sciences University of Kashmir,Hazratbal, Srinagar, (J&K),190006, India

ABSTRACT
Nowadays, Information, which is managed by Database Management System, is deemed as an asset for any of the organizations. Malicious attacks over a computer network can decrease security and trust of a system which may lead to various threats, which will be mitigated by introducing Intrusion Detection System. Emerging Intrusion detection systems (IDS) cannot improve the Integrity and in order to offer higher security for the confidential data with information integrity, an Artificial Bee Colony algorithm based Intrusion Detection system (ABC-IDS) is proposed. Clustering, Mining and Classification are the three major phases of this proposed system. The primary step is clustering the given datasets which helps to recover the quality of the datasets by partitioning the quantity in a cluster. Clustering is worked out by the clustering algorithm, Fuzzy C-Means and the second phase Mining is completed on the clustered datasets in order to get mined results by successfully mining the given datasets with the aid of Frequent Item set mining. The generated rules in the mining process get optimized by the usage of Artificial Bee Colony algorithm. After acquiring optimized mined results, the Classification phase is carried out by using Artificial Neural Network classifier, which classifies the input dataset into Intrusion or Non-Intrusion packets. This proposed method is implemented in MATLAB platform over DARPA dataset and then it is analyzed for its accuracy of Intrusion detection rate and Non-Intrusion detection rate, which will also evidently improve the consistency and reliability of the ABC-IDS system. Moreover, comparison of the proposed methodology with the state-of-art works is done to prove the improvement in information integrity in the proposed method. Hence, the optimum Integrity is increased by this research by the increase of its intrinsic attributes of accuracy, consistency and reliability.

Keywords
Data Mining, Intrusion Detection System, Fuzzy C-Means, Frequent Item Set Mining, Artificial Bee Colony Algorithm, Artificial Neural Network.

1. INTRODUCTION
Intrusion detection is defined as the process to identify the internal or external users who intend to do something unauthorized against the computer system [9]. It is one of the most useful tools for detecting suspicious traffic in different ways. The intrusion detection technology has been a new network security technology in the past 20 years [4]. For computers and networks, it acts as a security management system. Any intelligent algorithm or tool which has been designed to detect and report any unauthorized access within a network can be viewed as an Intrusion Detection System. [1] Further, it is a useful tool for detecting attacks, so that appropriate action shall be taken to ensure the confidentiality, integrity and availability of the network resources [2] [4]. Denning introduced the first intrusion detection model to find these behaviors which are different from users [2]. An intrusion detection system can be divided into two approaches which are behavior based and knowledge based. The behavior based approach is also known as anomaly based system while knowledge based approach is known as misuse based system [5].

Anomaly detection is based on statistical description of the normal behavior of users or applications. The misuse detection is based on collecting attack signatures in order to store them in a database [7]. Intrusion detection techniques are classified into a host- based and a network- based intrusion detection by a data source, and classified into misuse and abnormal detection by a detection method rules [3]. Host-based IDS employs audit logs and system calls as its data source, whereas network-based IDS employs network traffic as its data source [6]. Research on anomaly, host-based intrusion detection has adopted either of three main approaches: system-based, specification-based, and learning-based [11]. Existing anomaly detection methods are primarily categorized as statistical anomaly detection, anomaly detection based on neural network and anomaly detection based on data mining, etc [12].

Currently many IDS are rule-based system where the performances highly rely on the rules identified by security experts. To overcome the limitations of rule-based system, a number of IDSs employ data mining techniques. Data mining is the analysis of large data sets to discover understandable patterns or models [8]. Many methods have been proposed in the past few years on the design of IDSs. For example, Bridges and R.B. Vaughn [3] proposed IDS that combines both misuse and anomaly intrusion detection systems. A comprehensive IDS requires a significant amount of human expertise and time for development [10]. Also the Clustering techniques can be useful for detecting intrusions from network data [15]. These Data mining systems are also capable of generalizing new and unknown attacks [13]. Data mining systems make it possible to easily perform data summarization and visualization that help the security analysis in various areas [14]. Data mining-based IDSs require less expert knowledge yet provide good performance.

2. RELATED WORKS
Several techniques are proposed by various authors for Intrusion Detection and a few of them are explained below:

Francisco Maciá-Pérez and J. Mora-Gimeno [16] have suggested an intrusion detection system (NIDS) which operated independently as an anomaly-based NIDS, or incorporated clearly in a distributed intrusion detection system

(DIDS) and was implanted in a smart sensor-inspired device using service-oriented architecture (SOA) approach. This IDS was innovative as it combined the advantages of the smart sensor approach and the succeeding offering of the NIDS functionality as a service with the SOA to achieve their integration with other DIDS components. They explained how vast amount of management tasks inherent to this type of network services can be reduced, as well as restricting the design complexity of IDS within definite margins.

The ability of Intrusion Detection and detection of newest form of disruption can be enhanced by sharing information as well as opinions between Collaborative Intrusion Detection Network and Intrusion Detection Systems. For maintaining this objective Carol J Fung et al. [17] estimated a collaboration system that is distributed on the basis of host id's known as Host-based ID's (HIDS), essentially concentrating on friendly management of the system. Catalog is maintained by HIDS consisting of false negative and false positive for later evaluation to detect intrusion. Based on a simulated collaborative HIDS network they have evaluated their system to effectively deal with malicious nodes in wireless sensor networks (WSNs). Bo-Heung Chung et al. [18] proposed a trust management protocol which is highly scalable and is cluster-based hierarchical in nature. In contrast to previous work, they evaluated the overall trust of a sensor node by considering multidimensional trust attributes derived from communication and social networks. Using probability model, they described a heterogeneous WSN that comprises of sensor nodes that are larger large enough with immensely diverse social and the corresponding quality of service behaviors with the aim to yield node status i:e "ground truth". This was served as a foundation for validating their design of protocol by evaluating the subjective trust that is generated because of runtime protocol execution against the objective trust that is obtained from the actual node status. To exhibit the effectiveness of their hierarchical based trust management protocol, they have applied geographic routing and intrusion detection that are both trust-based. For every application, they have recognized the excellent trust composition/ formation so as to maximize the performance of the application.

One of the major challenge for security these days in wireless network is precisely and swiftly detection of intrusion. The flow of traffic can be used for detecting numerous types of intrusions. Now a day a frame known as Super frame can be used for industrial automation process automation (WIA-PA) on wireless communication networks to enforce standard.

A technique for detecting Intrusion in WIA-PA was devised by Min Wei and Keecheon Kim [19]. Using Time-Sequence technique for flow data after examining and demonstrating, and using this Time-Sequence technique a model for predicting data was created for explanatory purpose. Network traffic was swiftly and accurately foreseen by the model. 16 channel analyzer was used for measuring traffic of data with which the model was initialized. The results obtained by this model have shown productive intrusion detection attacks with improved performance in the network lifetime.

Many traditional IDSs functions in isolation and could be compromised very easily by the unknown threats. In order to overcome the above weakness and the overall improvement in the intrusion assessment accuracy, intrusion detection network was proposed that allows IDS peers to share the intrusion detection knowledge and the corresponding experience. Hence Quanyan Zhuet al. [20] have designed a GUIDEX which is based on IDN system. The system use game theory modeling technique and trust management so as to collaborate honestly and actively. They have established the uniqueness and the existence of Nash equilibrium so that peers shall be able to communicate in a mutual incentive compatible manner. An iterative algorithm have been developed converge to the stability so as to cope up with the duality of the said problem. In a WSN, intrusion detection plays a significant role in detecting the threat or malicious acts or any unpredicted intruder in many applications. The intruder can be a person or malicious code that affects the system. With the deployment of WSM uniformly, the probability of detection remains the same at any point. Hence Yun Wanget al. [21] have proposed a Gaussian-distributed WSNs as it provided the distinguished detection capabilities in diverse locations and limited related work has been noticed in this regard. In this case, they have analyzed the problem of probability of detecting the intrusion with respect to the network parameters and the corresponding application requirements under the scenario of both single as well as multiple sensing detections. Various network parameters that affect the detection probability were studied in detail. Further, the WSN like Gaussian-distributed was compared for performance evaluation with uniformly-distributed.

Existing intrusion detection systems are not able to cope with the changing environment in the network. Incorporating intrusion detection in distributed system architectures is a major challenge. In this direction, two new algorithms for Intrusion detection based on Adaboost were proposed by Weiming Hu et al [22] wherein decision stumps were employed as weak classifiers in the first algorithm whereas Gaussian mixture models were employed as weak classifiers. A universal model for detection was devised for each node by merging its local parametric models using particle swarm optimization and the algorithms based on support vector machines.

Zhenwei Yu et al. [23] have proposed an automatically tuning IDS (ATIDS), which was automatically tune the detection model on-the-fly based on the response offered by the system operator when false predictions were found. Pradhan M, Pradhan S K, Sahu S K [24] has given the anomaly detection using neural network based intrusion detection. For this, they have proposed neural based data mining technique to achieve the desired results in an well-organized mode. The technique clearly demonstrated that Neural Network based IDS has maximum categorization accuracy and minimum error rate as compared to other classifier algorithm.

3. PROBLEM DEFINITION

In the existing work [16] the network based intrusion detection system has embedded in a smart-sensor-inspired device. It was not able to reach the high level of automation and also the method used in this work could not attain the whole configuration and management task process [16]. The method was used only for an isolated device and in an Effective Acquaintance Management based Bayesian approach system [17], If provision of HIDS peer endorsements is extended it may critically affect the robustness of the acquaintance management system. In addition to this, the field devices in WIA-PA network of detection intrusion system in [19] accept beacons and launch data corresponds to the super frame cycles. The data series may not follow the normal distribution, if the system runs with various data cycles. This leads into inappropriate ARIMA (Autoregressive Integrated Moving Average). An

intrusion detection network [20] has used resistance of GUIDEX to detect the common insider attacks. But, this network could not identify other potential attacks by reverse engineering process, which changes the objectives from binary codes. These all are the problems in state-of-art works, which encourages for making this research on intrusion detection based on datasets.

4. PROPOSED ARTIFICIAL BEE COLONY BASED INTRUSION DETECTION SYSTEM

When an intrusion occurs, security and privacy of any system gets compromised. Intrusion Detection System (IDS) is a device with a software application, which observes the whole activities of the system mainly for malicious activities and then creates reports to the services under management. Information security is the main concernment regarding with IDS, in which the objective is for protecting the confidentiality, integrity and availability of data in the system. In order to provide higher security for the confidential data with information integrity, we have intended to propose an ABC based Intrusion Detection System (ABC-IDS) over IDS with information security in data mining. The phases of this ABC-IDS technique (as in fig. 1) are as follows:

(1) Clustering

(2) Mining

(3) Classification

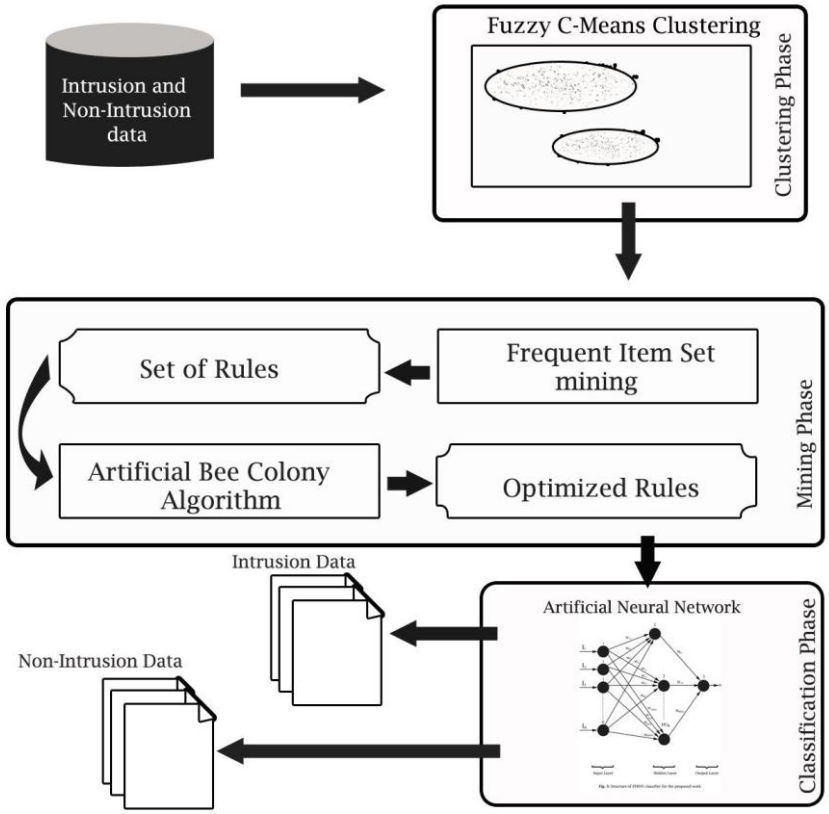

Fig. 1: Proposed ABC based Intrusion Detection System

Primarily, a dataset is given as the input for this ABC-IDS technique. Initial Phase is clustering using Fuzzy C-Means, which helps to improve the quality of the datasets by reducing (partitioning) the quantity in a cluster [25]. This initial phase helps to improve and to make easy the second phase by improving quality of the datasets, since the databases are usually very much vulnerable to noisy, missing and inconsistent. Then the second phase Mining is performed on the clustered datasets in order to get mined results by effectively mining the given datasets with the concept of Frequent Item set mining. Followed by the optimization in the rules in mining process, we move to further process. The optimization of rules is handled by Artificial Bee Colony (ABC) algorithm. After obtaining optimized mined results, the final phase Classification is carried out by using Artificial Neural Network classifier. Intrusion or Non-Intrusion is classified in this phase effectively. The detail explanation of the proposed method is given below.

4.1 Phase I – Clustering

Input for the proposed Intrusion Detection system is a dataset with both Intrusion and Non-Intrusion data. In order to decrease the quantity of the dataset, the clustering process is employed on the input dataset. Fuzzy C-Means clustering algorithm is utilized as the clustering algorithm in the initial phase for getting the input dataset to be clustered

4.1.1 Fuzzy C-Means clustering

For making two or more clusters, a set of data is used in Fuzzy C-Means clustering, where the data points to be allocated for the clusters, are not "hard" but fuzzy. So, Fuzzy clustering is also indicated as soft clustering, where each element is correlated with a set of membership levels.

FCM has a goal to divide a finite set of N data points $X = \{x_1, x_2, ..., x_N\}$ into a set of R fuzzy clusters based

on some conditions. For these set of data points, FCM makes a set of R cluster centers $C=\{c_1,c_2,...,c_R\}$ with a matrix $U=[u_{ij}] \in [0,1]$, *where,* $i=1,2,...,N$ *and* $j=1,2,...,R$ for partitioning of clusters. FCM clustering algorithm works by minimizing the objective function, which is given in eqn. (1).

$$O=\sum_{i=1}^{N}\sum_{j=1}^{R}\left[\mu^{m}{}_{ij}\,(x_i-c_j)^2\right] \tag{1}$$

here, m denotes any real number > 1, uij denotes degree of membership of xi in the cluster j, xi is the i th of d - dimensional measured data, cj is the d - dimension center of the cluster, and ‖*‖ is any norm transmitting the similarity between any calculated data and the center

Fuzzy partitioning can be done by frequently optimizing the objective function shown above, with the modification of membership uij and the cluster centers cj by:

$$\mu_{ij}=\frac{1}{\sum_{k=1}^{R}\left(\frac{\|\,x_i-c_j\,\|}{\|\,x_i-c_k\,\|}\right)^{2/m-1}} \tag{2}$$

In above eqn. (2), Fuzzy Exponent ρ is calculated as $\rho=\frac{2}{m-1}$. The value of m decides the cluster fuzziness level, where, $1<m<\infty$. Higher values of m lead to get smaller fuzzy memberships μ_{ij}. In general, the value of m is set to 2, because if $m=1$, then $\rho=\infty$; if $m=3$ then $\rho=1$ (i.e.) linear relationship and if $m=\infty$, then $\rho=0$ (i.e.) all the memberships become same. And also, the cluster centres are given below as in eqn. (3).

$$c_j=\frac{\sum_{i=1}^{N}\mu_{ij}^{m}\;x_j}{\sum_{i=1}^{N}\mu_{ij}^{m}} \tag{3}$$

This iteration will end, when $\max_{ij}\left\{\left|\mu_{ij}^{k+1}-\mu_{ij}^{k}\right|\right\}<\tau$, where τ is a termination criterion between 0 and 1, while k is the iteration step. This process unites to a local minimum or a saddle point of O.

4.1.2. Steps in Fuzzy C-Means clustering

The clustering steps for the process of Fuzzy C-means are given below.

Step 1:- Select the centre of cluster c_j, randomly for N data points.

Step 2:- Initialize the matrix as $U^{(0)}$ with $U=[u_{ij}]$ and the value of u_{ij} is computed from the eqn. (2) by calculating the distance from each point to each cluster centre. Thus, a Fuzzy membership is assigned for each of the N data points to each of the clusters.

Step 3:- With the aid of the matrix $U^{(k)}$, compute the centers vectors $C^{(k)}=[c_j]$ at step k from eqn. (3).

Step 4:- Update $U^{(k)}$, $U^{(k+1)}$, by finding new cluster centers.

Step 5:- If $\left\|U^{(k+1)}-U^{(k)}\right\|<\tau$ or minimum of objective function as in eqn. (1) is achieved, then stop the process; otherwise, Repeat the process until cluster membership no longer changes.

The clustered results of FCM provide the whole input data into two clusters – (i) cluster with intrusion data and (ii) cluster with non-intrusion data.

4.2 Phase II – Mining

The mining process is performed separately on both the clustered data to further classify the data for detecting the intrusions. One of the mining algorithms is incorporated in this mining phase to mine the rules presented in the data. The algorithm used in this second phase is Frequent Item set mining algorithm.

4.2.1 Frequent Item set Mining

Frequent sets have an important role in many data mining works that provide various patterns from databases. In general, frequent item set mining looks at sequences of actions. Both the clustered data are separately subjected to the process of frequent item set mining to obtain set of rules as the mined data.

Initially, the input data is converted into crisp values by getting the median value of the attributes in the given clustered data. This conversion of crisp data contains class 1 and class 2, which are the input for the frequent item set mining. The fundamental data taken for the mining process is a set of instances each of which has a number of attributes. In data mining, the set of instances are called the transactions. In the proposed work, the classes of different attributes are considered as the set of instances. Each class combination is represented as an item.

The process of these frequent item set algorithm then identifies all the common sets of classes for the attributes, in which these set of classes have at least a minimum support. Minimum support specifies the minimum number of times the attributes exist in the common set of classes. Then the classes for the attributes are combined with parameters such as {1-length combination, 2-length combination, (n-1)-length combination} based on the number of attributes for the classes. For example, if a set of classes is {a, b, c}, then 1-length combinations are {a}, {b}, {c} and the 2-length combinations are {a, b}, {b, c}, {a, c} taken from the common set. Likewise, according to the length of the set of classes, the combinations are taken by this frequent item set mining algorithm.

After the identification of a set of frequent item sets, association rules are produced. Every association rule has qualities such as support, confidence.

Support of a class “a” is defined as the proportion of attributes

in the given whole classes of input application.

Confidence of a rule is defined as follows,

$$Confidence(a \Rightarrow b) = \frac{\sup port(a \cup b)}{\sup port(a)} \quad (4)$$

After generating the rules, we have a number of rules to find the frequency of attributes for a particular class. To overcome this, we need to optimize these rules, for which Artificial Bee Colony (ABC) Algorithm is exploited in the proposed work.

4.2.2 Artificial Bee Colony Algorithm

If we use only the frequent item set mining for getting the mined clustered data, it is possible to produce more number of rules, in which many non-frequent item sets also be presented. To avoid these non-frequent item sets, we use ABC algorithm within the frequent item set mining for optimizing the rules.

ABC is swarm based algorithm that is inspired by searching nature of the bees and is comprised of three main components:

(i) Employed bees: That deals with the source of food in the hive region and also transfers the gathered information about the quality of nectar to onlookers.

(ii) Onlooker bees: Utilizing the information gathered from employed bees to pick a single source of food by watching their dance (employed bees).

(iii) Scout bees: The Scout bees are basically the employed bees whose food source get abandoned and try to search new source of food randomly. The amount of nectar and the corresponding food resources represents the position of probable solution. The functioning of ABC algorithm is depicted in fig. 2.

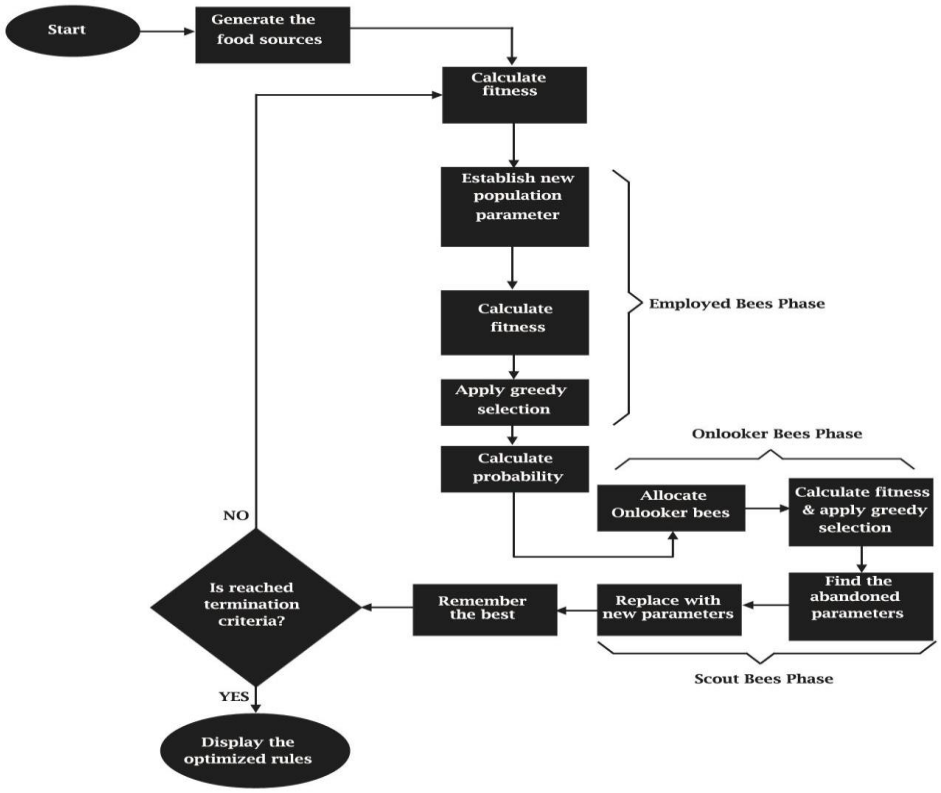

Fig. 2: ABC algorithm Flowchart

4.2.2.1 Initial Phase

In this phase, the populations of food sources $x_i, (i = 1,2,....R)$ are generated arbitrarily. R represents the population size of. This food sources as above contains the rules (R_i) generated for each attributes of the two classes 1 and 2. This process is so called as initialization process. In order to evaluate the food source that is considered as best, the value of fitness is calculated using equation (5) for generated food sources.

Fitness Function-F(j) = confidence X *(log(rand number* X *support* X *no. of values +1))* (5)

After the fitness value calculation is done, the iteration takes place and is set to '1'. And later the employed bee phase is carried out.

4.2.2.2. Employed Bee Phase

The said phase generates new population based on new parameters by using the following equation: -

$$V_{i,j} = x_{i,j} + \phi_{ij}\left(x_{i,j} - x_{k,j}\right) \quad (6)$$

here, k and j represents index selected randomly, ϕ represents the number produced randomly in the range [-1, 1] and $V_{i,j}$ represents new value at the j^{th} position. After every new population that is been generated based on parameters, fitness value is calculated for food sources. After calculating the fitness value, selection of best parameter among population is made by using the process of greedy selection. Now after the best parameter has been selected, its probability is calculated using the following equation:

$$P_j = \frac{F_j}{\sum_{j=1}^{d} F_j} \quad (7)$$

here, P_j denotes probability of j^{th} parameter.

4.2.2.3. Onlooker Bee Phase

Once the probability of calculating the selected parameter is done using equation (7), the calculation with regard to computing the quantity of onlooker bees is calculated. After this, using the solution $x_{i,j}$ which is based on probability value P_j, create new solution $V_{i,j}$ and calculate its fitness function. Then select the best parameter based on successive greedy selection.

4.2.2.4. Scout Bee Phase

Evaluate the fitness value by replacing the abandoned parameters present if any by new parameters that have been discovered using equation (7). The best parameter need to be memorized that have been achieved till now. The process is continued till certain terminating condition is met which discovers the optimized rule.

Thus, the optimized rules that obtained from frequent item set and ABC algorithms are further utilized in the classification phase by merging both the optimized rules of Intrusion and Non-Intrusion data.

4.3 Phase III – Classification

Classifying the datasets for detecting the Intrusions is worked out by the help of an Artificial Neural Network (ANN) classifier. An ANN is a mathematical model containing a number of highly interrelated processing elements sorted out into layers, geometry and functionality. The ANN may be considered as possessing learning capabilities in as much as it has a normal propensity for storing experimental knowledge and making it accessible for later use. In order to generate the wanted mapping, the neural network has been coached to fine-tune the connection weights and biases. At the training stage, the characteristic vectors are employed as input to the network and the network adjusts its changeable parameters, the weights and biases, to incarcerate the affiliation among the input patterns and outputs.

In order to detect and classify intrusion data, Feed Forward Back Propagation Neural Network is employed. This type of classifier is standard three layer neural network with I_n as input, H_n as hidden and O_n as output nodes.

The input layers $\{I_1, I_2, ..., I_n\}$ as depicted in the fig. 3 are the optimized rules that have been derived from the mining phase, with HU_a Hidden nodes and only one output node O for the proposed work. The overall structure of classifier is depicted in fig.3

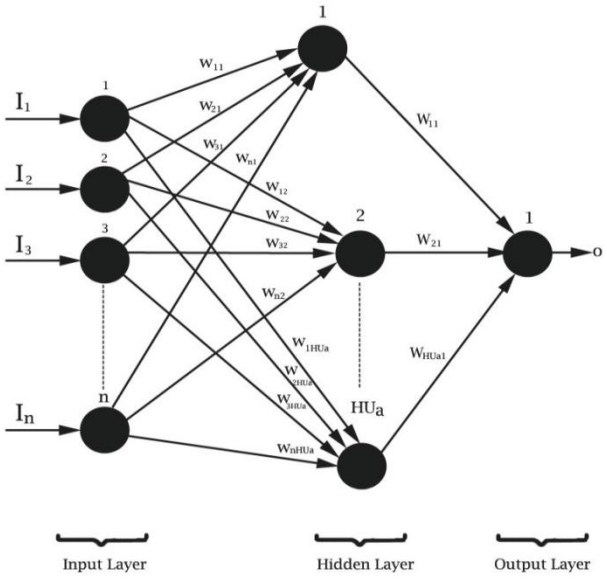

Fig. 3: Structure of FFBNN classifier for the proposed work

4.3.1 NN Function Steps

i) In the first step, weight of each neuron is to be set except the input layer neurons.

ii) Create a neural based network with $\{I_1, I_2, ..., I_n\}$ as the input, HU_a as Hidden and O as output unit whether the given data is Intrusion or Non-Intrusion.

iii) For the input layer, the bias function that have been proposed is calculated as

$$X = \beta + \sum_{n=0}^{HU_{a-1}} W_{(n')} I_1(n') + W_{(n')} I_2(n') + W_{(n')} I_3(n') + ... + W_{(n')} I_n(n') \quad (8)$$

and for the output layer, activation function is calculated as :

$$Active\,(X) = \frac{1}{1+e^{-X}} \quad (9)$$

iv) The learning error rate is calculated as.

$$LE = \frac{1}{HU_a} \sum_{n'=0}^{N_a-1} Y_{n'} - Z_{n'} \quad (10)$$

here, LE - represent rate of learning for FFBNN.

$Y_{n'}$ - represent the desired outputs.

$Z_{n'}$ - represent the actual outputs.

4.3.2 Learning Algorithm – Back Propagation Algorithm used for minimizing the error

The most often employed training algorithm in classification problems is the Back-Propagation (BP) algorithm. In FFNN, BP Algorithm is a supervised learning technique. In order to

construct training data set, a requisite data set is required for a variety of inputs. For Feed-Forward Networks, BP Algorithm is very helpful and requires activation function be differentiable for neurons. In order to reduce a cost function, learning in a neural network involves adapting the weights and biases of the network. The cost function all the time comprises a fault term a measure of how shut the network's predictions are to the class labels for the examples in the training set. Moreover, it may contain a complexity term that react a former distribution over the values that the parameters can obtain.

Steps of BP Algorithm using FFBNN

i) The hidden and output layer neurons are assigned weights randomly while it remains constant for input layer.

ii) The equation (8) and (9) are used to calculate the bias and activation function respectively for the FFBNN.

iii) The BP Error is calculated for each and every node, and new weights are calculated as

$$w_{(n')} = w_{(n')} + \Delta w_{(n')} \qquad (11)$$

iv) The change in weight $\Delta w_{(n')}$ is calculated as.

$$\Delta w_{(n')} = \delta . X_{(n')} . E^{(BP)} \qquad (12)$$

here, δ - represent rate of learning, usually ranging (0.2 to 0.5).

$E^{(BP)}$ - Back Propagation Error.

v) To minimize Back Propagation error (BP) Step (ii) and (iii) are repeated. i.e. $E^{(BP)} < 0.1.$

vi) The minimum value signifies FFBNN is fit for testing phase.

The output is generated from the classifier in the form of intrusion and non-intrusion data after well training and testing using optimizing rules accordingly.

5. RESULTS AND DISCUSSIONS

The proposed ABC based Intrusion Detection System is implemented in the working platform, Matlab and which work is effectively detected the Intrusion data from the datasets. The experimentation results and the performance evaluation results are given in this section in detail.

5.1 Dataset Description

The dataset used for the proposed ABC-IDS work is DARPA dataset [26]. DARPA dataset is used for the purpose of training as well as testing the intrusion detectors. The evaluations so made in this regard contributed extensively to the research on intrusion detection by giving direction for efforts in research and a calibration of objective of the specialized futuristic work. In the DARPA dataset used for IDS evaluation, all the network traffic including the entire payload of each packet is recorded in dump format and provided for evaluation. The fields in the datasets are examined and labeled as intrusion and normal packets. The attacks fall into five major classes namely Probe, Denial of Service (DoS), Remote to Local (R2L), User to Remote (U2R) and the Data attacks. The sample data set description is given in the following table. 1.

Table 1: Sample dataset description for Intrusion Detection System

A	B	C	D	E	F	G	H	I	J	K	L	M
184	1511	2957	0	0	3	0	2	1	0	0	1	0
305	1735	2766	0	0	3	0	2	1	0	0	1	0
79	281	1301	0	0	2	0	1	1	0	0	4	2
25	269	2333	0	0	0	0	0	1	0	2	2	1
150	1587	6707	0	0	1	0	3	0	0	1	1	0
60	2328	4551	0	0	3	0	1	1	0	0	0	0
158	1567	3095	0	0	3	0	4	1	0	0	1	0
103	302	8876	0	0	2	0	4	1	0	3	4	2
54	260	2635	0	0	0	0	0	1	0	2	2	1
0	0	5921	0	0	0	0	0	0	0	0	0	0
0	0	5014	0	0	0	0	0	0	0	0	0	0
0	0	2072	0	0	1	0	0	1	0	0	0	0
113	6274	16771	0	0	5	0	2	1	0	0	0	0
53	2628	3860	0	0	3	0	1	1	0	0	0	0
7	230	644	0	0	4	0	0	0	0	0	4	0
31	142	1278	0	0	0	0	0	0	0	0	1	0
21	135	1290	0	0	0	0	0	0	0	0	0	0
0	0	5690	0	0	0	0	0	0	0	0	0	0
0	0	5828	0	0	0	0	0	0	0	0	0	0
0	0	5020	0	0	0	0	0	0	0	0	0	0

The dataset contains totally 34 columns of attributes (A-Z and AA-AH) and 42 rows. Here, is a sample of dataset with 13 columns and 20 rows.

5.2 Experimentation Results

The proposed Artificial Bee Colony based Intrusion Detection System is implemented on the DARPA dataset and thus the experimentation results for every process is illustrated in this section.

Initially, DARPA dataset is given as the input for the proposed method. The input dataset contains both Intrusion and Normal packets. The first step is to cluster the datasets into these two kinds of packets. The clustering process is carried out using Fuzzy C-Means Clustering method. The sample clustering of packets is given in the following fig. 4.

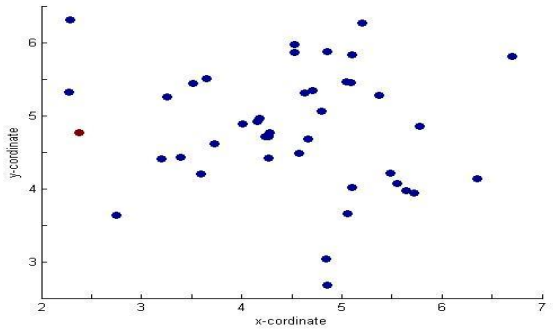

Fig. 4: Clustering with Fuzzy C-Means

After clustering the Intrusion and Non-Intrusion packets, we process on these two kinds of data separately to detect the Intrusions from the whole dataset. The clustering results are then given to the Association Rule mining process to mine rules as the frequent item sets. The results of frequent item set mining are given in the following table 2.

Table 2: Sample results of frequent item set mining process

A	B	C	D	E	F	G	H	I	J	K
2	3									
4	5									
4	5	6								
5	6	7								
3	4	5	6							
4	5	6	7							
4	5	6	7	8						
5	6	7	8	9						
3	4	5	6	7	8					
4	5	6	7	8	9					
4	5	6	7	8	9	1				
5	6	7	8	9	1	11				
4	5	6	7	8	9	1	11			
5	6	7	8	9	1	11	12			
3	4	5	6	7	8	9	1	11		
4	5	6	7	8	9	1	11	12		
2	3	4	5	6	7	8	9	1	11	
3	4	5	6	7	8	9	1	11	12	
3	4	5	6	7	8	9	1	11	12	13
4	5	6	7	8	9	1	11	12	13	14

The original results obtained from the frequent item set have totally 314 rows. In table-2 a sample of 20 rows are shown. Artificial Bee Colony algorithm is then utilized to optimize the results of frequent item set mining process. The optimized results of frequent item set mining from ABC algorithm are given in the following table 3.

Table 3: Sample results of optimized frequent item sets using ABC

A	B	C	D	E	F	G	H	I	J	K
1	11	12	13	14	15	16	17	18	19	2
7	8	9	1	11	12	13	14	15	16	17
13	14	15	16	17	18	19	2	21	22	23
9	1	11	12							
5	6	7	8							
1	11	12	13							
16	17	18	19	2	21	22	23	24	25	26
3	4	5	6	7	8	9	1	11	12	13
1	2	3	4	5	6	7	8	9	1	11
25	26	27	28	29	3	31	32	33		
6	7									
8	9	1	11	12	13	14	15	16	17	18
2	3	4	5	6	7	8	9	1	11	12
2	3	4	5	6	7	8	9	1	11	12

2	3	4	5	6	7	8	9	1	11	12
7	8	9	1	11						
8	9	1	11	12	13	14	15	16	17	18
2	21	22	23	24	25	26	27	28	29	3
2	3	4	5	6	7	8	9	1	11	12
24	25	26	27	28	29	3	31			

The unique optimized results acquired from the ABC algorithm have only 50 rows, which includes only the optimized frequent item sets of the 34 columns of attributes values. A sample of 20 rows is shown in table 3. Artificial Bee Colony algorithm is then utilized to optimize the results of frequent item set mining process. Finally, we can detect the Intrusions based on these optimized rules by using Feed Forward Back Propagation Neural Network. The results so generated give the impression that the threats that harm the integrity of information and the corresponding information system is mitigated to a greater extent. The evaluation parameters of Integrity are also evaluated based on the results.

5.3 Evaluation Metrics

An evaluation metric is used to evaluate the effectiveness of proposed Intrusion Detection system for its efficient detection of Intrusion packets in the dataset and to justify theoretical and practical developments of these systems. It consists of a set of measures that follow a common underlying evaluation methodology. The metric values are found based on True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) with the option of Intrusion detection. The performance of this method is analyzed by nine metrics namely Detection Rate (DR), False Alarm Rate (FAR), Sensitivity, Specificity, Accuracy, Reliability, Consistency Error, False Positive Rate (FPR) and False Negative Rate (FNR). The subsequent table-4 proves how the positive and negative values are explained for the Normal-Intrusion packet detection.

Table 4: Description of TP, TN, FP and FN valuesDetection Rate (DR)

Description		OUTPUT	
		Detected as Intrusion	**Detected as Non-Intrusion**
INPUT	**Intrusion**	TP	FN
	Non-Intrusion	FP	TN

The Number of Intrusion packets detected by the IDS is defined as Detection Rate.

$$DR = \frac{TP}{Total\ No.of\ Intrusion\ Packets} \quad (13)$$

False Alarm Rate (FAR)

FAR is the ratio between the number of Non-Intrusion packets that detected as attacks and the total numbers of Non-Intrusion packets.

$$FAR = \frac{FP}{Total\ No.of\ Non-Intrusion\ Packets} \quad (14)$$

Sensitivity

The ratio of actual positives that are detected properly is termed as sensitivity. Thus it is capacity to generate positive results.

$$Sensitivity = \frac{TP}{TP + FN} \times 100 \quad (15)$$

Specificity

The ratio of actual negatives that are detected properly is termed as specificity. Thus it is the capacity to generate negative results.

$$Specificity = \frac{TN}{TN + FP} \times 100 \quad (16)$$

Accuracy

In order to calculate the ratio of Intrusion as well as Non-Intrusion packets for a given data set, accuracy is measured as.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \times 100 \quad (17)$$

Reliability

Reliability is referred as the ratio between the detection difference in Intrusion and Non-Intrusion packets and the total number of Intrusion detection.

$$\text{Re}liability = \frac{(TP+TN)-(FP+FN)}{TP + TN} \quad (18)$$

Consistency Error

The Error rate of incorrect detection of Intrusion and Non-Intrusion packets in datasets is measured by Consistency Error. It is represented as,

$$Consistency\ Error = \frac{FP+FN}{TP+FP+TN+FN} \quad (19)$$

False Positive Rate (FPR)

The percentage of cases where a packet was detected as Intrusion, but in fact it did not.

$$FPR = \frac{FP}{FP+TN} \quad (20)$$

False Negative Rate (FNR*)*

The percentage of cases where a packet was detected as Non-Intrusion, but in fact it did.

$$FNR = \frac{FN}{FN+TP}$$

(21)

5.4 Performance Evaluation Of The Proposed Work

The performance of the proposed Artificial Bee Colony based Intrusion Detection System is analyzed in this section. The mining performance for the elapse time is initially evaluated in this section against the support value of association rule mining. Fig. 4 shows the graph for the mining performance against support value.

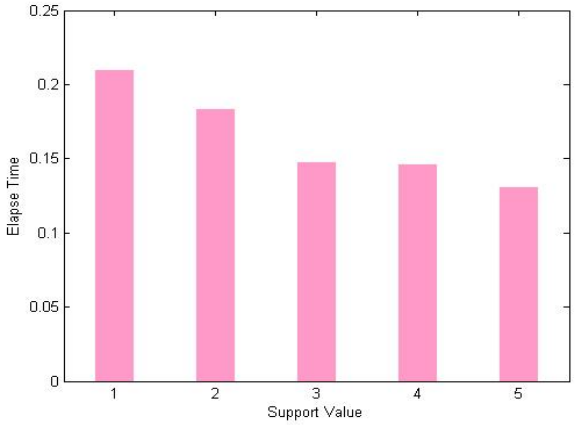

Fig. 5: Mining Performance results - Elapse time versus support value

From fig.5, analyses the mining performance for the proposed Intrusion Detection system. The elapse time is decreased, when the support value increases. For the support values from 1 to 5, the elapse time is 0.21 s, 0.18 s, 0.15 s, 0.15 s and 0.13 s, respectively. Thus, we can observe the reduction in elapse time for every increase of support values.

The Performance is evaluated by the evaluation metrics Detection Rate, False Alarm Rate, Sensitivity, Specificity, Accuracy, Reliability, Consistency Error, False Positive Rate and False Negative Rate. These evaluation measures are given in the following table 5 for the proposed work.

Table 5: Performance Analysis of the proposed IDS with various evaluation metrics

Evaluation Metrics	Measurements
True Positive	46
True Negative	50
False Positive	0
False Negative	4
Detection Rate	0.9200
False Alarm Rate	0
Sensitivity	0.92
Specificity	1
Accuracy	0.9600
Reliability	0.9583
Consistency Error	0.0400
False Positive Rate	0
False Negative Rate	0.0800

Both the total number of Intrusion and Non-Intrusion packets in the testing dataset of the proposed work is 50. The number of Intrusions correctly detected is 46 and the remaining 4 intrusions are incorrectly detected as Non-Intrusion packets. But, all the 50 Non-Intrusion Packets are correctly detected as Non-Intrusion packets. The Rate of detected Intrusion packets by the proposed IDS is 0.92, which is a good value and which gives the results of detection rate. FAR provides the ratio between the number of Non-Intrusion packets that detected as Intrusions and the total numbers of Non-Intrusion packets. The rate of false alarm for the proposed work is equal to 0, which highly improve the efficacy of the Intrusion detection rate and its accuracy. There are 0.80 percentages of cases, where a packet is detected as Non-Intrusion, but actually it did not. This value is False Negative Rate value and in the work only the correct Intrusion packet detection is attained, because the False Positive Rate of 0. 0.92% sensitivity is achieved, which shows that higher rate of actual positives are properly detected and the Non-Intrusion packets are 100% attained which yields superior specificity value. Accuracy and reliability are also highly increased with the lower value in consistency error. The Performance evaluation for the improvement in integrity of proposed IDS is given in following fig. 6

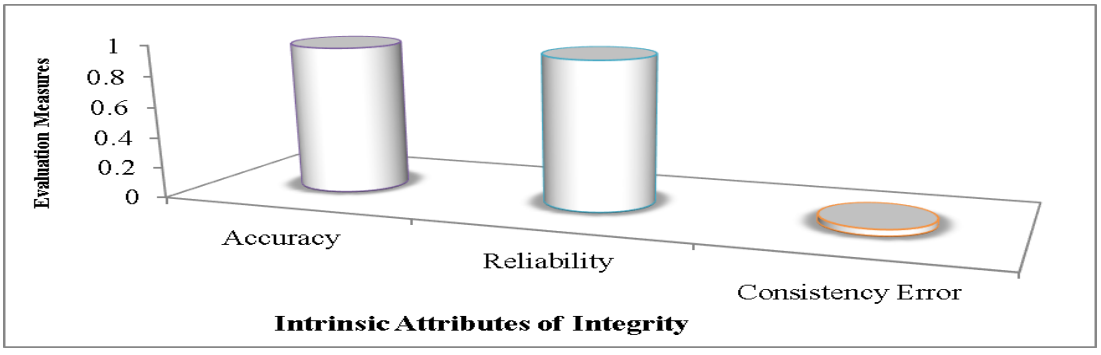

Fig. 6: Performance evaluation for Integrity of Proposed Intrusion Detection System

The integrity of the proposed ABC based Intrusion Detection system is improved by the intrinsic attributes of accuracy, reliability and consistency. First attribute is accuracy, which is 96% for the detection of Intrusion packets by the proposed method. The increase value obtained in both the sensitivity and specificity makes a way to get 96% of accuracy. The higher value of accuracy proves that the detection of both Intrusion and Non-Intrusion packets is very accurate. The

second intrinsic attribute to get improved integrity is reliability, which is 95.83% and the third intrinsic attribute is consistency. In the proposed work, we measure the consistency error by which we can say the consistency is good or bad. Here, we get 0.0400 value of consistency error, which is too low value, by which we can say that the consistency of the proposed system is also too high. Thus, we can demonstrate the proposed work and prove that the proposed Detection system is capable to detect the Intrusions accurately and it can be able to improve the integrity of the proposed IDS by gaining superior values in the intrinsic attributes accuracy, reliability and consistency.

5.5 Comparison Evaluation With The Proposed Work

The proposed method is compared with other existing papers to show and prove that the proposed Intrusion Detection system is best system for the detection of Intrusion packets. For this purpose, we consider two of the state-of-art work papers from the related work section. References [23] [24] are compared with the proposed work.

Table 6: Comparison results for the existing works with the proposed work

Methods	Accuracy (in %)
MC-SLIPPER	92.59
ATIDS	94.61
Neural Network based Anomaly Detection System	88
Proposed method	96

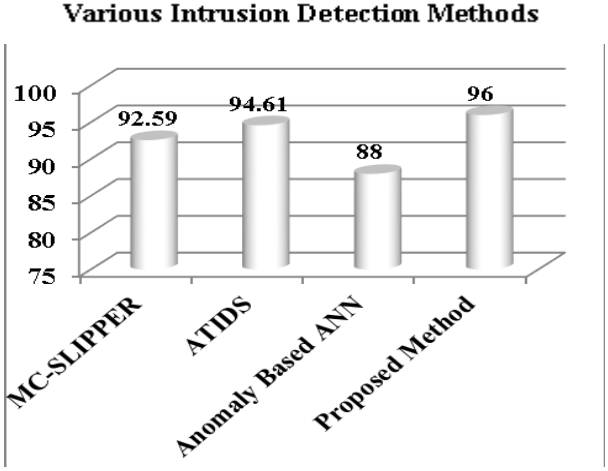

Fig. 7: Graph for comparing the existing works with the proposed IDS

The state-of-art methods MC-SLIPPER and ATIDS are found from the paper [23], which provides 92.59% and 94.61% of accuracy respectively. And also, another paper [24] has presented various neural network method based Intrusion Detection. These neural networks utilized in that existing paper were Neural Network based Anomaly Detection System, which gives 88 respectively. Among these existing methods ATIDS and MC-SLIPPER are the good methods. Even though these two are good one, these methods are not better than the proposed ABC based IDS. The reason is that the proposed ABC-IDS facilitates 96% of accuracy than the other existing methods as above, which is 1.39% and 3.41% higher than the methods ATIDS and MC-SLIPPER, respectively. Hence, we can prove that the proposed work can outperform other existing methods by giving superior accuracy values and by offering higher security for the confidential data with information integrity.

6. CONCLUSION

The proposed ABC-IDS method was implemented on Matlab platform over the dataset DARPA with the three stages – Clustering, Mining and Classification. The results were evaluated for the proposed work, which shows the improvement in the integrity of the proposed Intrusion Detection System. The Detection Rate for the input dataset was 0.92 with good detection of Intrusion packets and the False Alarm Rate was 0, which clearly states that too accurate detection of Non-Intrusion packets. Regarding the improvement in Integrity, we have computed its intrinsic attributes of Accuracy, Reliability and Consistency Error. 96% of accuracy and 95.83% of reliability were achieved, which confirms that the detection of Intrusion and normal packets were very accurate. We have obtained 0.0400 value of consistency error, which was too low, by which we can say that the consistency of the proposed method was also too high. The proposed method was also compared with existing papers and which has outperformed the state-of-art works by providing higher accuracy. By the higher values in the intrinsic attributes of accuracy and reliability and the lower value of consistency error, we can prove that the proposed Intrusion Detection System have the ability to detect and mitigate the harm caused by the Intrusions with the improvement in Integrity. Thus, we can hope that the proposed system will be helpful for the future researchers to make the research over Intrusion Detection System.

7. REFERENCES

[1] S.Janakiraman, S.Rajasoundaran and P.Narayanasamy, "The Model - Dynamic and Flexible Intrusion Detection Protocol for High Error Rate Wireless Sensor Networks Based on Data Flow", In The Proceeding of IEEE 6th International Conference of Mobile Adhoc and Sensor Systems, pp. 313-321, Oct 2009.

[2] Zhou Mingqiang, Huang Hui, Wang Qian, "A Graph-Based Clustering Algorithm for Anomaly Intrusion Detection", In Proceedings of 7th International Conference onComputer Science &Education, pp. 1311-1314, July 2012.

[3] Zhou Mingqiang, Huang Hui, Wang Qian, "A Graph-Based Clustering Algorithm for Anomaly Intrusion Detection", In Proceedings of 7th International Conference onComputer Science &Education, pp. 1311-1314, July 2012

[4] Yongquan Mo,Yizhong Ma And Liang Xu, "Design And Implementation of Intrusion Detection Based on Mobile Agents", In Proceeding of IEEE International Conference of Medical and Education, pp. 278-281, 2008.

[5] Amrita Anand, Brajesh Patel, "An Overview on Intrusion Detection System and Types of Attacks It Can Detect Considering Different Protocols ", Journal of Advanced Research in Computer Science and Software Engineering, Vol. 2, No. 8, pp. 94-98, Aug 2012.

[6] Mostaque Md. Morshedur Hassan, "Current Studies on Intrusion Detection system, Genetic Algorithm and Fuzzy Logic ", International Journal of Distributed and Parallel Systems (IJDPS), Vol.4, No.2, Mar 2013.

[7] Chunfu Jia, Deqiang Chen, "Performance Evaluation of A Collaborative Intrusion Detection System ",In

Proceeding of theFifth International Conference on Natural Computation, pp. 409-413, Aug 2009.

[8] Jiong Zhang, Mohammad Zulkernine, and Anwar Haque, "Random-Forests-Based Network Intrusion Detection Systems", In Proceeding of IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews, Vol. 38, No. 5, pp. 649-659, Sept 2008.

[9] Bin Zeng, Lu Yao, ZhiChen Chen, "A Network Intrusion Detection System with the Snooping Agents", International Conference on Computer Application and System Modelling, pp. 232-236, Oct 2013.

[10] Abebe Tesfahun, D. Lalitha Bhaskari, "Intrusion Detection using Random Forests Classifier with Smote and Feature Reduction", International Conference on Cloud & Ubiquitous Computing & Emerging Technologies, pp. 127-132, Nov 2013.

[11] Karen A. Garc´Ia, Raul Monroy, Luis A. Trejo, Carlos Mex-Perera, and Eduardo Aguirre, "Analyzing Log Files For Postmortem Intrusion Detection", In Proceeding of IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews, Vol. 42, No. 6, pp. 1690-1704, Nov 2012.

[12] Wu Yang, Wei Wan , Lin Guo, and Le-Jun Zhang, "An Efficient Intrusion Detection Model Based on Fast Inductive Learning", In Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong, pp. 3249-3254, Aug 2007.

[13] M. Moorthy, Dr. S. Sathiyabama , "A Study of Intrusion Detection using Data Mining", In Proceeding of IEEE-International Conference on Advances In Engineering, Science and Management (ICAESM -2012), pp. 8-15, March 2012.

[14] Mohammadreza Ektefa, Sara Memar, Fatimah Sidi, Lilly Suriani Affendey, "Intrusion Detection Using Data Mining Techniques", International Conference on Information retrival & Knowledge Management, (CAMP),pp. 200-203,Mar 2010.

[15] Kapil Wankhade, Sadia Patka, "An Efficient Approach for Intrusion Detection Using Data Mining Methods", International Conference on Information Retrival & Knowledge Management, pp. 200-103,Mar 2010.

[16] Francisco Maciá-Pérez, J. Mora-Gimeno, "Network Intrusion Detection System Embeddedon a Smart Sensor", In Proceeding of IEEE Transactions on Industrial Electronics, Vol. 58, No. 3,pp. 722-732, Mar 2011.

[17] Carol J Fung, Jie Zhang, "Effective Acquaintance Management based on Bayesian Learning for Distributed Intrusion Detection Networks", In Proceeding of IEEE Transactions on Network and Service Management, Vol. 9, No. 3,pp. 320-332, September 2012.

[18] Fenye Bao, Ing-Ray Chen, MoonJeong Chang, and Jin-Hee Cho, "Hierarchical Trust Management for Wireless Sensor Networks and its Applications to Trust-Based Routing and Intrusion Detection", In Proceeding of IEEE Transactions on Network and Service Management, Vol. 9, No. 2, pp. 169-183, June 2012.

[19] Min Wei and Keecheon Kim, "Intrusion Detection Scheme Using Traffic Prediction forWireless Industrial Networks", Journal of Communications and Networks, Vol. 14, No. 3, pp. 310-318, June 2012.

[20] Quanyan Zhu, Carol Fung, Raouf Boutaba, Tamer Basar, "GUIDEX: A Game-Theoretic Incentive-Based Mechanism for Intrusion Detection Networks", IEEE Journal on Selected Areas In Communications, Vol. 30, No. 11, pp. 2220-2230, Dec 2012.

[21] Yun Wang, Weihuang Fu, and Dharma P. Agrawal, "Gaussian versus Uniform Distribution for Intrusion Detection in Wireless Sensor Networks", In Proceeding of IEEE Transactions on Parallel and Distributed Systems, Vol. 24, No. 2, pp. 342-355, Feb 2013.

[22] Weiming Hu, Jun Gao, Yanguo Wang, Ou Wu, and Stephen Maybank, "Online Adaboost-Based Parameterized Methods for Dynamic Distributed Network Intrusion Detection", In Proceeding of IEEE Transactions on Cybernetics, Vol. 44, No. 1, pp. 66-82, Jan 2014.

[23] Zhenwei Yu, Jeffrey J. P. Tsai, and Thomas Weigert, "An Automatically Tuning Intrusion Detection System", IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics, Vol. 37, No. 2, pp. 373-384, April 2007.

[24] Pradhan M, Pradhan S K, Sahu S K, " Anamoly Detection using Artificial Neural Network, International Journal of Engineering Sciences & Emerging Technologies,Volume 2, Issue 1, April 2012.

[25] Rayees, Q. Khan, Butt, Muheet A & Asger, M., Zaman M, 2015, Integrity Model based Intrusion Detection System: A Practical Approach, Inter Jour. of Comp. Science-IJCA", ISSN: 2249-6645X, 4 : 1-7.

[26] www.cs/dal/ca/%7Eriyad/Dataset/DARPA99/DARPA99 Week1.zip