DOI: 10.21917/ijct.2013.0114

PERFORMANCE COMPARISON OF INTRUSION DETECTION SYSTEM USING VARIOUS TECHNIQUES – A REVIEW

S. Devaraju[1] and S. Ramakrishnan[2]

[1]*Department of Computer Applications, Dr. Mahalingam College of Engineering and Technology, India*
E-mail: deva_sel@yahoo.com

[2]*Department of Information Technology, Dr. Mahalingam College of Engineering and Technology, India*
E-mail: ram_f77@yahoo.com

Abstract

Nowadays, the security has become a critical part of any organization or industry information systems. The Intrusion Detection System is an effective method to deal with the new kind of threats such as DoS, Porbe, R2L and U2R. In this paper, we analyze the various approaches such as Hidden Semi Markov Model, Conditional Random Fields and Layered Approach, Bayesian classification, Data Mining techniques, Clustering Algorithms such as K-Means and Fuzzy c-Means, Back Propagation Neural Network, SOM Neural Network, Rough Set Neural Network Algorithm, Genetic Algorithm, Pattern Matching, Principle Component Analysis, Linear Discriminant Analysis, Independent Component Analysis, Multivariate Statistical Analysis, SOM/PSO algorithm etc. The performance is measured for two different datasets using various approaches. The datasets are trained and tested for identifying the new attacks that will affect the hosts or networks. The well known KDD Cup 1999 or DARPA 1999 dataset has been used to improve the accuracy and performance.

The four groups of attacks are identified as Probe, DoS, U2R and R2L. The dataset used for training set is 494,021 and testing set is 311,028. The aim is to improve the detection rate and performance of the proposed system.

Keywords:

Intrusion Detection, Neural Networks, Data Mining, KDD Cup, DARPA

1. INTRODUCTION

An intrusion detection system (IDS) is a major component of the information security framework. The main goal of IDS is to develop a system which can automatically scan network activity and detect the attacks. Once an attack is detected, the system administrator can decide who can take necessary action and prevent those attacks.

In past years, there were only few intruders and so the user could manage them easily from the known or unknown attacks, but in recent years the security is the most serious problem. Because the intruders introduce a new variety of intrusions in the market, so that the user can't manage the computer systems and networks properly.

Intrusion detection attacks can be classified into two groups:

i) Misuse based or Signature based / Known Attacks

ii) Anomaly based Intrusion Detection / Unknown Attacks

The misuse based or signature based intrusion detection system detects the intrusion by comparing the existing signatures in the database. The signature based intrusions are called known attacks. The users detect the intrusion when they match with the signatures log files. The log file contains the list of known attacks which are detected from the computer system or networks. The anomaly based intrusion detection is called as unknown attacks and, this attack is observed from network and thus deviates from the normal attacks.

The intrusion detection systems are classified as Network based, Host based and Web based attacks. The network based attack may be either misuse based or anomaly based attacks. The network based attacks are caused due to interconnection of computer systems. The system communicates with each other and so the attack is sent from one computer system into another computer system by the way of routers and switches.

The host based attacks are detected in a single computer system and it is easy to prevent the attacks. This attack mainly occurs when some external devices are connected. The external devices are pen drive, CD, VCD, Floppy, etc. The web based attacks occurs, when systems are connected over the internet and so, the attacks spread into different systems through the email, chatting, downloading materials etc.

The examples of different attacks are denial-of-service (DoS), Distributed denial-of-services (DDoS), Worm based attack, port scanning, Flash crowd, Alpha flows, probe, user-to-root (U2R), remote-to-local (R2L) etc.

Different approaches and algorithms are used to detect the attack. The most widely used approaches are: Neural Network based approaches, Statistical based approaches, Data Mining based Approaches, Genetic Algorithm based approaches, and Fuzzy Logic based approaches.

In this paper we propose the techniques which can detect network based attacks using neural network classification. This method follows a pattern of normal and intrusive activities, such as DoS, U2R, Probe, R2L and Normal and classified a set of classification techniques based on deviation between current and reference behavior. Neural network is evaluated by dataset KDD99 or DARPA Dataset. We study the various neural network classification techniques to verify its feasibility and effectiveness. Experimental results show that this method can improve the performance, effectiveness and reduce the missing alarm in IDS.

The rest of the paper discusses the different approaches. The section 2 describes the datasets, section 3 discusses Network Based Approach and section 4 describes Host Based Approach. Section 5 describes comparative analysis; Section VI derives the summaries of section 3 & 4 and section 7 discusses the references. The intrusion detection can be classified into three categories and the classification is shown in Fig.1.

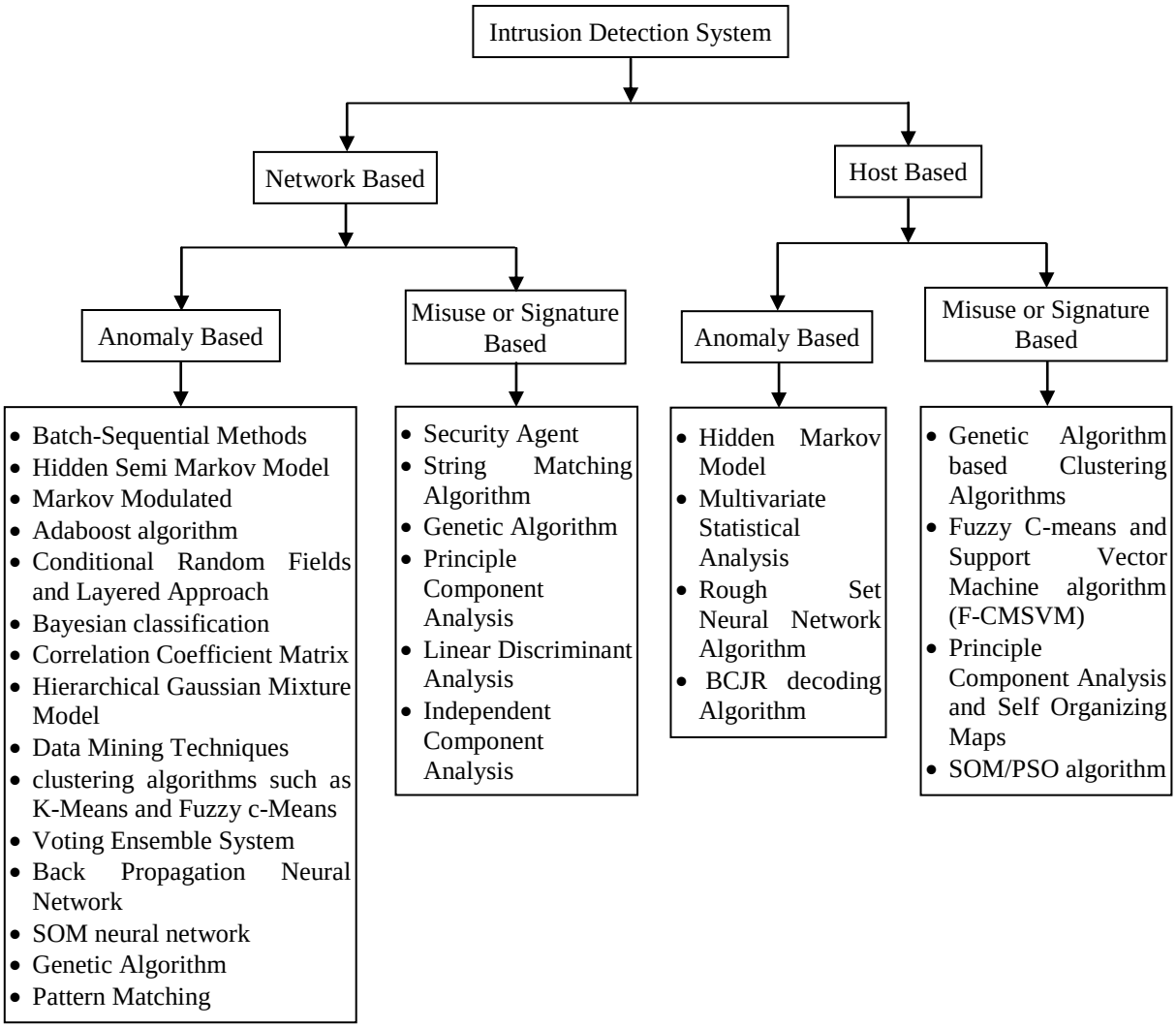

Fig.1. Classification of Intrusion Detection System

2. DATASETS

There are two well known data sets used in this area of intrusion detections. They are KDDcup99 Dataset and DARPA98 Dataset [1].

2.1 KDD CUP 1999 DATASET

The KDD Cup 1999 dataset has been used for the evaluation of anomaly detection methods. The KDD Cup 1999 training dataset consists of approximately 4,900,000 single connections vectors each of which containing 41 features and is labeled as either normal or an attack, with exactly one specific attack type. The datasets contain a total number of 24 training attack types and 14 testing attack types.

In KDD Cup 1999 dataset has different types of attacks. They are back, buffer_overflow, ftp_write, guess_passwd, imap, ipsweep, land, loadmodule, multihop, neptune, nmap, normal, perl, phf, pod, portsweep, rootkit, satan, smurf, spy, teardrop, warezclient, warezmaster. These attacks can be divided into 4 groups are denial of service attacks, attacks from a remote system to a local user, attacks from a local user to root, and surveillance or probing attacks. The Table.1 shows the list of attacks category wise.

Table.1. List of attacks - category wise

DoS	R2L	U2R	Probe
back	ftp_write	buffer_overflow	ipsweep
land	guess_passwd	loadmodule	nmap
neptune	imap	perl	portsweep
pod	multihop	rootkit	satan
smurf	phf		
teardrop	spy		
	warezclient		
	warezmaster		

- Denial of Service (DoS) attacks: deny legitimate requests to a system, e.g. flood
- User-to-Root (U2R) attacks: unauthorized access to local super user(root) privileges, e.g. various buffer overflow attacks

- Remote-to-Local (R2L) attacks: unauthorized access from a remote machine, e.g. guessing password
- Probing: surveillance and other probing, e.g. port scanning.

2.2 DARPA DATASET

The DARPA dataset was designed to work at MIT Lincoln Laboratory to support the 1998 DARPA Intrusion Detection Evaluation. This is a complex project supported by many workers. The 1998 DARPA evaluation was designed to find the strength and weakness of existing approaches leading to large performance improvements and valid assessments of intrusion detection systems. The concept was to generate a set of realistic attacks, embed them in normal data, evaluate the false alarm and detection rates of systems with these data, and then improve systems to correct the weaknesses found [2].

Two data sets are the result of the DARPA Intrusion Detection Evaluations.

- 1998 DARPA Intrusion Detection Evaluation Data Sets
- 1999 DARPA Intrusion Detection Evaluation Data Sets

This evaluation was measured on the probability of detection and probability of false alarm for each system under test. These evaluations contributed significantly to the intrusion detection. All the researchers work on the general problem of workstation and network intrusion detection. The evaluation was designed to be simple, to focus on core technology issues, and to encourage the widest possible participation by eliminating security and privacy concerns, and by providing data types that were used commonly by the majority of intrusion detection systems.

3. NETWORK BASED APPROACHES

3.1 ANOMALY BASED APPROACH

The network based anomaly is a process of monitoring the events occurring in a network and analyzing them for intrusions called unknown attacks. These attacks attempt to bypass the security mechanisms of network traffic. This attack affects the network when a user wants to access resources over the network [10], [20], and [29].

The following are the major improvements in the network based anomaly:

- Fast and accurate real-time anomaly detection
- Minimum false alarm rate
- Improving the performance

When the intruders introduce new type of viruses over the network, the computer systems are affected. If the systems are affected by the viruses then the process is denied, increasing the false alarm rate, reducing the performance and effectiveness of the system [22].

For these reasons the attacks are presented by following certain techniques:

3.1.1 Batch-Sequential Methods:

The batch and sequential methods combine in one unit to develop a multistage detection algorithm called batch-sequential. The main advantage of Batch-Sequential method is that it retains enough relevant information to detect network intrusions quickly, while maintaining the FAR (False Alarm Rate) below a selected level. The batch sequential method is also used to detect the network attacks very quickly and improve processing sequentially [6].

The method is designed to detect increase or decrease in the expected number of packets that are observed in all possible sets of size bins.

$$\chi^2_{pt,k} = \sum_{i=1}^{M_{pt}} \frac{\left(N_{k,i}^{pt} - \mu_i^{pt}\right)^2}{\mu_i^{pt}} \quad (1)$$

Which in this particular form measures the departure of the network traffic from the distribution P_0 under which the expectation $E_0 N_{k,i}^{pt} - \mu_i^{pt}$ simultaneously for all $i = 1,\ldots M_{pt}$.

3.1.2 Adaboost Algorithm:

The AdaBoost algorithm is a machine learning algorithm. It can do many pattern recognition problems like face recognition. This algorithm is used to correct the misclassifications done by weak classifiers. The intrusion detection system uses the Dataset to identify the weak classifier and this feature can be converted into strong classifier. Because this algorithm is very fast in identifying the weak classifier compared with other algorithms. This algorithm can be applied into four modules such as feature extraction, data labeling, design of the weak classifiers, and construction of the strong classifier. These features are used for detecting the intrusions; the set of data is used for training the labeled data; and the strong classifier is trained using the sample data and also obtained by combining the weak classifiers [8].

3.1.3 Conditional Random Fields and Layered Approach:

Conditional models are used to model the conditional distribution over a set of random variables and this gives better framework like Maxent classifiers, maximum entropy Markov models, and CRFs. The training data constrains this conditional distribution while ensuring maximum entropy and uniformity. The objective of using a layered model is to reduce computation complexity and the overall time needed to detect anomalous activity among the different layers. For example, four layers are grouped into four attacks in the data set. The dataset used for four types of layers are Probe layer, DoS layer, R2L layer, and U2R layer. Each layer is trained independently with a set of relevant features [9].

Let X be the random variable over data sequence to be labeled and Y the corresponding label sequence. In addition, let $G = (V; E)$ be a graph such that $Y = (Y_v)_{v\varepsilon(V)}$, so that Y is indexed by the vertices of G. Then, (X, Y) is a CRF, when conditioned on X, the random variables Y_v obey the Markov property with respect to the graph. $P(Y_v|X, Y_w, w \neq v) = P(Y_v|X, Y_w, w \sim v)$, where $w \sim v$ means that w and v are neighbors in G.

3.1.4 Hierarchical Gaussian Mixture Model:

It is the process of identifying the abnormal packets in the network. There are two phases in the process of Hierarchical Gaussian Mixture Model (HGMM). The first is the training phase that reference templates and second is the detection phase. The training phase trains the sample data provided by the traffic using statistical model. The detection phase is used to detect the abnormal packets that deviate from the stored reference [16], [21].

A Gaussian mixture density is a weighted sum of M component densities, as given by equation,

$$P(x/\lambda) = \sum_{i=1}^{M} w_i b_i(x) \quad (2)$$

where, x is a D-dimensional random vector, $b_i(x)$, $i = 1,\ldots M$, are the component densities and w_i, $i = 1,\ldots M$, are the mixture weights.

3.1.5 Voting Ensemble System:

The ensemble algorithms can be divided into two categories that are constructed as base classifiers and voting. The constructing base classifier is used to prepare and build the input training data for building base classifiers by perturbing the original training data. Voting system is used to combine the base models for better performance [12], [27].

There are different ensembles of classifiers using different features extracted from the KDDCup'99 intrusion detection dataset, and then these results are put into the voting system. Each classifier has a weight to denote the contributions of the classifier to the voting system. For each class to be identified, a weighted sum of base learners can be calculated as,

$$v_i = \sum_{d=1}^{N} \alpha * w_d, \qquad \alpha = \begin{cases} 1 & C_d = i \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where, N is the number of classifiers, i = 1, 2... C is the class label, C_d is the predicted class label by the d classifier, and w_d is the weight of the d classifier. For a given unknown pattern, the final class to be classified is determined by maximizing $\arg\max_{j=1}^{C} V_j$.

3.1.6 Neural Network:

i) *Back-Propagation Neural Network*: The Back-Propagation Algorithm is a supervised method, which uses steepest-method to reach global minima. This method involves two ways, Forward propagation and Reverse propagation to implement the Intrusion Detection System (IDS) [13], [15], [24], [11].

Forward Propagation: The output of each node in the successive layers is calculated as,

$$o(\text{output of a node}) = \frac{1}{1+e^{-\sum W_{ij} X_i}} \quad (4)$$

where,

W_{ij} = weight matrix connecting nodes of the previous layer i with nodes of next layer j.

X_i = variables of a pattern

o = output of a node in the successive layer

Reverse Propagation: The error δ for the nodes in the output layer is calculated as,

$$\delta(\text{output of a node}) = o(1-0)(d-o) \quad (5)$$

The new weights between output layer and hidden layer are updated

$$w(\eta + 1) = w(\eta) + (\text{output layer}) \text{ o } (\text{hidden layer}) \quad (6)$$

where,

η: is the learning factor $0 < \eta < 1$

The training of the network is stopped once the desired mean squared error (MSE) is reached as,

$$E(MSE) = \sum E(p) \quad (7)$$

The final updated weights are saved for detection the intrusion.

ii) *Genetic Algorithm*: The genetic algorithm can be applied to solve a variety of optimization problem. At every level, the genetic algorithm selects random individuals from the current population to be parents and uses them produce the children for the next generation shown in Fig.2 [2], [15].

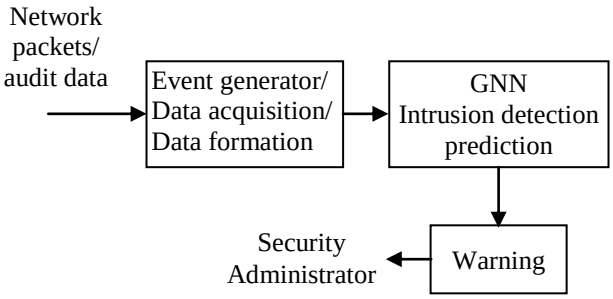

Fig.2. Intrusion detection model on GNN

Genetic algorithm can be defined as an eight-tuple:

$$SGA = (C, E, P_0, M, \Phi, \Gamma, \Psi, T) \quad (8)$$

where, C represents the chromosome representation; E represents the fitness function; P_o, the initial population; M, the population size; Φ, the selection operator; Γ, the crossover operator; Ψ, the mutation operator and T, the terminal conditions.

iii) *SVM and GA*: The Support Vector Machine and GA are used in the optimum selection of principal components which are used for classification. These methods are capable of achieving minimum amount of features and maximum amount of detection rates [30].

iv) *Fuzzy Clustering Neural Network*: The Fuzzy clustering neural network uses a hybrid framework experiment over the NSL dataset to test the stability and reliability of the technique. The hybrid approach performs better detection especially for lower frequency of over NSL datataset compared to original KDD dataset, due to the removal of redundancy and incomplete elements in the original dataset [38].

v) *Fuzzy rule-based systems*: Three fuzzy rulebased classifiers detect intrusions in a network. Results are then compared with other machine learning techniques like decision trees, support vector machines and linear genetic programming. Empirical results clearly show that soft computing approach could play a major role for intrusion detection and improve the efficiency [48].

3.2 MISUSE/SIGNATURE BASED APPROACH

The misuse detection systems rely on the definitions of misuse patterns i.e., the descriptions of attacks or unauthorized actions. The signature attacks are known attacks, which affect the network if the attacks match the database.

3.2.1 Signature IDS Methodology:

This IDS system follows the signature based methodology for ascertaining attacks. The signature based system will monitor

packets on the network and match with a database of malicious threats and signatures [17].

Pre intrusion activities prepare the network for intrusion. These include port scanning to find a way to get into the network and IP spoofing to disguise the identity of the attacker or intruder. Signature-based IDSs operate analogously to virus scanners, i.e. by searching a database of signatures for a known identity or signature for each specific intrusion event.

3.2.2 Genetic Algorithm:

The GA detects the network based misuse attacks. There are three basic genetic operators applied to each individual with certain probabilities like selection, cross over, and mutation and find the effectiveness of the system [18], [23], [33].

Analyzing the dataset, rules will be generated in the rule set. These rules will be in the form of an 'if then' format as follows,

$$\text{if } \{\text{condition}\} \text{ then } \{\text{act}\}. \tag{9}$$

Since the GA has to use such rules to detect intrusions, such rules in the rule set will be codified to the GA format. Each rule will be represented in a GA format.

The GA is used in the fitness function. The fitness function F determines whether a rule is good or bad. F is calculated for each rule using the support confidence framework.

$$\text{Support} = |A \text{ and } B| / N$$

$$\text{Confidence} = |A \text{ and } B| / |A|$$

$$\text{Fitness} = t1 * \text{support} + t2 * \text{confidence} \tag{10}$$

where, N is the total number of records, $|A|$ stands for the number of network connections matching the condition A, $|A$ and $B|$ is the number of records that matches the rule and $t1$ and $t2$ are the thresholds to balance the two terms.

3.2.3 Feature Reduction Techniques:

To enhance the learning capabilities and reduce the computational intensity of competitive learning neural network classifiers, different dimension reduction techniques have been used. These include: Principal Component Analysis, Linear Discriminant Analysis and Independent Component Analysis.

i) *Principal Component Analysis*: Principal Component Analysis uses dimensionality reduction techniques for data analysis and compression. This technique identifies the similarities and differences between the patterns [19], [26].

Given the data, if each datum has N features represented for instance by $X_{11}\ X_{12}\ \ldots\ X_{1N}$, $X_{21}\ X_{22}\ldots X_{2N}$, the data set can be represented by a matrix $X_{n\times m}$.

The average observation is defined as,

$$\mu = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{11}$$

The deviation from the average is defined as,

$$\phi_i = X_i - \mu \tag{12}$$

ii) *Linear Discriminant Analysis*: LDA is an optimal transformation matrix. LDA can be used to discriminate between the different classes. The analysis requires the data to have appropriate class labels and mathematically formulate the optimization procedure [19].

The analysis requires the data to have appropriate class labels. In order to mathematically formulate the optimization procedure,

$$\overline{x}_j = \frac{1}{N_j}\sum_{i=1}^{N_j} x_i. \tag{13}$$

To compute the mean vector and the covariance matrix for each class and for the complete data set,

$$\sum_{j=1}^{J} N_j = N \tag{14}$$

where, N denotes the total number of training tokens and N_j stands for the number of training tokens in class j. Naturally, the number of classes is j.

iii) *Independent Component Analysis*: ICA is a redundant feature, which is used to determine the performance or accuracy of the classifier. The ICA finds the irrelevant information. The ICA is used to reduce the dimensionality of the data so that a classifier can handle large volume of data [19].

The independent component analysis is expressed as the technique for deriving one particular W, $y = Wx$.

The general learning technique to find a suitable W is,

$$W = \eta(I - \Phi(y)y\mathrm{T})W \tag{15}$$

where, $\Phi(y)$ is a nonlinear function of the output vector y.

3.2.4 Recurrent Neural Network:

Recurrent Neural Network model used with four groups of input features has been proposed as misuse-based IDS and the experimental results have shown that the reduced-size neural classifier has improved classification rates, especially for R2L attack [31].

3.2.5 GNP and Fuzzy:

A novel fuzzy class-association rule mining method based on genetic network programming (GNP) method is used for detecting network intrusions. The Experimental results show that the proposed method provides competitively high detection rates compared with other machine-learning techniques and GNP with crisp data mining [32].

3.2.6 Fuzzy Decision Tree:

The Fuzzy decision tree uses Mutual Correlation for feature selection and Fuzzy Decision Tree classifier is used for detection and diagnosis of attacks. The Experimental results of the 10% KDD Cup 99 benchmark network intrusion detection dataset demonstrate that the proposed learning algorithms achieve good accuracy, high true positive rate (TPR) and reduce false positive rate (FP) significantly [41], [42], [46].

3.2.7 Fuzzy Systems and Ant Colony Optimization:

The fuzzy system with an Ant Colony Optimization procedure is used to generate high-quality fuzzy-classification rules. Hybrid learning approach is applied to network security and validated using the DARPA KDD-Cup99 benchmark data set. The results indicate that the proposed hybrid approach achieves better classification accuracies when comparison to several traditional and new techniques, [47].

3.2.8 C5 Decision Tree:

The multi-layer intrusion detection model is used to achieve high efficiency and improve the detection rate known and unknown attacks and classification rate accuracy by training the hybrid model on the known intrusion data. The experimental results show that the proposed multi-layer model using C5 decision tree achieves higher classification rate accuracy, and less false alarm rate [39], [40].

3.2.9 Genetic Algorithm and Fuzzy Logic:

The GA and Fuzzy logic focus on current development efforts and the solution of the problem of Intrusion Detection System to offer a realworld view of intrusion detection. The fuzzy membership value and fuzzy membership function are two different techniques used because the surface value is not always counted from the ground level. So, fuzzy sets can classify efficient rule sets and reduce the false alarm rate [43], [44], [45].

4. HOST BASED APPROACHES

4.1 ANOMALY BASED

The host based anomaly is a process of monitoring the events occurring in a host and analyzing them for intrusions. These attacks are attempts to bypass the security mechanisms. The anomaly based systems can detect known and unknown (i.e., new) attacks as long as the attack behavior deviates sufficiently from the normal behavior.

The following are the challenges for the host based anomaly:

- Speed
- Performance
- Accuracy
- Adaptability

4.1.1 Hidden Markov Model:

The Hidden Markov Model (HMM) is used for system-call-based anomaly intrusion detection. Experiments based on a public database demonstrate that this data preprocessing approach can reduce training time. A simple and efficient HMM anomaly intrusion algorithm is proposed as follows.

Assume that an HMM model parameter is,

$$\lambda = \{A, B, \Pi\} \tag{16}$$

where,

$A = \{a_{ij}\}$ represents the probability of being in state j at time $t + 1$, given that the state i at time t

$B = \{b_j^{(k)}\}$ represents the probability of observing symbol v_k at state j

$\pi = \{\pi_i\}$ is the probability of being at state i at time $t = 1$.

Three popular public databases have been used to test the HMM algorithm in detecting anomaly intrusions. The experiment demonstrate that up to 50 percent of the training cost saving can be achieved for a large data set without noticeable degradation of intrusion detection performance.

4.1.2 Multivariate Statistical Analysis:

The multivariate statistical analyses of audit trails use for detection of host-based intrusion. There are two types of statistical analysis used; T^2 test and X^2 test. Both tests are used to evaluate the performance [4].

***T*² test**: It is used to analyze audit trails of activities in an information system and detect host based intrusions into the information system that leave trails in the audit data. Let $X_i = (X_{i1}, X_{i2}, \ldots, X_{ip})$ denote an observation of p measures on a process or system at time i. Using a data sample of size n, the sample mean vector X and the sample covariance matrix S are usually used to estimate μ and $\sum$, where,

$$T^2 = \left(X_i - \overline{X}\right)' S^{-1} \left(X_i - \overline{X}\right) \tag{17}$$

A large value of T^2 indicates a large deviation of the observation X_i from the in-control population.

***X*² test**: The X^2 test performs well in intrusion detection, when tested on a small set of computer audit data containing sessions of both normal and intrusive activities. The X^2 test signals of all the intrusion sessions and produces no false alarms on the normal sessions. The P variable to measure and X_j denotes the observation of the j^{th} $(1 \le j \le p)$ variable at a particular time, the X^2 test statistic is given by the equation,

$$X^2 = \sum_{j=1}^{p} \frac{\left(X_j - \overline{X_j}\right)^2}{\overline{X_j}} \tag{18}$$

4.1.3 Rough Set Neural Network Algorithm:

The Rough Set theory algorithm used to filter out superfluous, redundant information and a trained artificial neural network identifies any kind of new attacks [16], [28].

Knowledge is represented by means of a table called an Information System given by $S = <U, A, V, f>$; where, $U = \{x_1, x_2, \ldots, x_n\}$ is a finite set of objects of the universe (n is the number of objects); A is a non empty finite set of features, $A = \{a_1, a_2, \ldots, a_m\}$; $V = \cup_{a \in A} V_a$ and V_a is a domain of feature a; $f: U \times A \rightarrow A$ is a total function such that $f(x, a) \in V_a$ for each $a \in A$, $x \in U$. If the features in A can be divided into condition set C and decision feature set D; i.e. $A = C \cup D$ and $C \cap D = \Phi$. The information system A is called decision system or decision table.

4.2 MISUSE OR SIGNATURE BASED APPROACH

The host-based system is a program that operates on a system and receives application or operating system audit logs. These programs are highly useful for detecting inside attack. If the user attempts unauthorized activity, host-based systems usually detect and collect the information quickly.

4.2.1 Genetic Algorithm based Clustering Algorithms:

The clustering algorithm detects the signature based intrusion detection. The fitness calculation process consists of two phases. In the first phase, the clusters are formed according to the centres encoded in the chromosome under consideration. This is done by assigning each point X_i. $i = 1, 2, \ldots, n$, to one of the clusters C_j with centre Z_j such as,

$$\left\|x_i - z_j\right\| < \left\|x_i - z_p\right\|, \; p = 1,2,..k \quad \text{and} \left(p \neq j\right) \tag{19}$$

Then the new centroids are calculated according to,

$$z_i = \frac{1}{n_i} \sum_{x_j \in C_i} x_j, \qquad i = 1,2,\ldots,K \tag{20}$$

where, z_i is the new centroid and nj is the number of points in the cluster *i*. After calculating new cluster centroids, cluster metrics must be computed for each cluster. It is the sum of the Euclidean distances of the points from their proper cluster centres [3], [34], [35].

4.2.2 Artificial Immune Network:

The Artificial Immune Network is a dynamic unsupervised learning method which consists of a set of cells called antibodies interconnected by links with certain strengths. These networked antibodies represent the network internal images of input patterns contained in the environment in which it is exposed. The author claims that, Artificial Immune Network is robust in detecting novel attacks [36].

4.2.3 Fuzzy C-means and Support Vector Machine algorithm:

Fuzzy C-means algorithm (FCM) is an efficient cluster algorithm which requires the number of clusters to be known beforehand for automatic clustering number determination. FCM aims to decide to what degree the sample data are affiliated to the cluster and to classify *n* sample data, $X = \{X_i \mid X_i \in R^D\ (i = 1, 2,\ldots,n)\}$ into *k* categories so as to compute the clustering central of each group $C = \{C_j \mid C_j \in R^D\ (j = 1,2,\ldots k)\}$.

Fuzzy support vector machine algorithm (SVM) has been used in intrusion detection for automatic clustering number determination. Here are marked samples (X_1, y_1), $(X_2, y_2),\ldots,$ (X_n, y_n). $X_i \in R^D$ belongs to one of two classes, $y_i \in \{-1,1\}$ is category mark. The main purpose of SVM is to construct a separating hyper-plane to separate the different samples so as to maximize the margin class. Then the optimizing question is shown below [25].

4.2.4 Fuzzy Rules:

This paper proposes a refined differential evolution search algorithm to generate fuzzy rules detects intrusive behaviors. In this algorithm the global population is divided into subpopulations, each is assigned a distinct processor and each subpopulation consists of the same class fuzzy rules. These rules evolve independently and also demonstrate with well-known KDD Cup 1999 Dataset [37].

5. COMPARATIVE ANALYSIS

The comparative analysis describes the Network Based Anomaly, Misuse/Signature with Network and Host Based Anomaly.

The Table.2 describes the various techniques used to find the detection rate which is used to measure the performance of the host or network.

Table.2 shows four groups of attacks such as Probe, DoS, U2R and R2L. The Probe attack detection rate achieved the minimum of 0.83% and maximum of 99.95%. The DoS attack detection rate achieved the minimum of 0.88% and maximum of 99.99%. The U2R attack detection rate achieved the minimum of 0.01% and maximum of 99.96%. The R2L attack detection rate achieved the minimum of 0.22% and maximum of 99.97%. The KDD Cup Dataset represents the training set is 494,021 and testing set is 311,028. The main aim of the proposed system is to improve the detection rate and reduce the false alarm rate.

The number of samples selected for training set and testing set for detecting the attacks and the detection rates compared is shown in Table.2.

Table.2. Comparison of Detection Rates in various classifiers

Ref. No.	Techniques used	Features	Normal %	Probe %	DoS %	U2R %	R2L %	Overall Results %	No. of samples for Training Set	No. of samples for Test Set
[8]	AdaBoost-Algorithm	41	-	-	-	-	-	90.04 - 90.88 %	494,021	311,029
[9]	Layered Conditional Random Fields	21	-	98.6%	97.4%	86.3%	29.6%	-	494,020	311,029
[16]	Gmix	41	98.97%	93.03%	88.24%	22.8%	9.6%	-	494,020	311,029
[16]	RBF	41	99.07%	91.31%	75.10%	7.01%	5.6%	-	-	
[16]	SOM	41	93.98 %	64.30%	96.10%	21.49%	11.7%	-		
[16]	Binary Tree	41	96.43 %	77.94%	96.45%	13.59%	0.44%	-		
[16]	ART	41	97.19 %	98.48%	97.09%	17.98%	11.3%	-		
[16]	LAMSTAR	41	99.69 %	98.48%	99.21%	28.94%	41.2%	-		
[16]	HGMM	41	88.14 %	99.33%	99.78%	96.01%	82.66%	-		
[12]	voting+J48+Rule	41	-	-	-	-	-	97.47%	Full Dataset	10 fold cross validation
[12]	voting+AdaBoost+J48	41	-	-	-	-	-	97.38%	-	

[16]	Rough Set Neural Network Algorithm	41	-	-	-	-	-	90%	6128 sets	257 sets
[16]	Rough Set Neural Network Algorithm	7	-	-	-	-	-	83%	-	
[18]	Genetic Algorithm	41	93.8%	-	Smurf - 67.3%	-	WarezMaster - 76.6%	-	-	-
[19]	Gmix	13	99.0 %	93.0%	88.3%	18.4%	10.5%	-	Full tree for testing	-
[19]	RBF	13	98.9 %	88.9%	75.1%	4.38%	5.40%	-	-	
[19]	Binary Tree	13	96.8 %	74.4%	96.4%	12.7%	0.44%	-		
[19]	LAMSTAR	13	99.7 %	98.9%	99.2%	30.3%	41.2%	-		
[19]	SOM	13	93.8 %	61.2%	96.1%	21.5%	10.9%	-		
[19]	ART	13	97.0 %	95.3%	97.0%	18.0%	11.0%	-		
[25]	SVM	41	19.48%	1.34%	73.90%	0.07%	5.20%	100.00%	24,701	15,551
[26]	Gmix	41	98.97 %	93.03%	88.24%	22.8 %	9.6%	-	494,020	311,029
[26]	RBF	41	99.07 %	91.31%	75.10%	7.01%	5.6%	-	-	
[26]	SOM	41	93.98 %	64.30%	96.10%	21.49%	11.70%	-		
[26]	Binary Tree	41	96.43 %	77.94%	96.45%	13.59%	0.44%	-		
[26]	ART	41	97.19 %	98.48%	97.09%	17.98%	11.29%	-		
[26]	LAMSTAR	41	99.69 %	98.48%	99.21%	28.94%	41.20%	-		
[27]	Ensemble Model	41	99.27%	99.88%	98.26%	99.96%	99.79%	-	5,092	6,890
[29]	Self Adaptive Bayesian Algorithm	12	99.97%	99.91%	99.99%	99,36%	99.53%	-	494,020	311,028
[29]	Self Adaptive Bayesian Algorithm	17	99.96%	99.95%	99.98%	99.46%	99.69%	-	494,020	311,028
[35]	Genetic Algorithm	41	69.5%	71.1%	99.4%	18.9%	5.4%	-	494,021	311,029
[38]	Fuzzy Clustering Neural Network	41	99.5%	88%	97.9%	87.9	46.8	-	18,285	311,089
[38]	Fuzzy Clustering Neural Network using NSL Dataset	41	98.2%	94.1%	99.1%	89	78	-	18,285	311,089
[42]	PSO based Fuzzy System	41	-	76.66	98.49	16.22	12.17	93.7	494,020	311,029
[46]	linguistic hedged fuzzy-XCS classifier	41	99.45	83.32	97.12	13.16	8.4	91.81	494,020	311,029
[47]	Evolutionary Fuzzy Systems and Ant Colony Optimization	41	96	86.25	98.83	72.8	33.45	-	752	311,029
[48]	Fuzzy rule-based systems	41	100	99.93	99.96	94.11	99.98	-	5,092	6,890

6. SUMMARY

The Multivariate Statistical Analysis methods are used to determine the anomaly detection and compared with the performance of the system [4]. The Hidden Markov Model is used to implement and determine the system on call based anomaly intrusion detection [5].

The Adaptive Sequential and Batch-Sequential Change-Point Detection Methods are used for detecting the attacks in the network traffic. This method uses the network simulator and real-life testing for detecting the attacks [6]. The model-free based approach using Markov modulated process mainly involves detecting the anomaly based attacks over the network [7].

The Adaboost Based Algorithm with decision rules provides both categorical and continuous features. This algorithm mainly focuses on four modules: feature extraction, data labeling, design of the weak classifiers, and construction of the strong classifier. [8]. Conditional Random Fields and Layered Approach are addressed by the two issues of Accuracy and Efficiency. This approach uses KDD cup '99 intrusion detection data set for detecting the attacks [9].

The Hierarchical Gaussian Mixture Model detects network based attacks as anomalies using statistical classification techniques. This model is evaluated by well known KDD99 dataset. Six classification techniques are used to verity the feasibility and effectiveness by reducing the missing alarm and accuracy of the attack in Intrusion Detection System [16]. The clustering algorithms such as K-Means and Fuzzy c-Means in data mining concepts are used for network intrusion detection and KDD Cup 99 data set is used for demonstration which performs both accuracy and computation time [14].

The system analyzes the performance of some data classifiers in a heterogeneous environment using voting ensemble system. The system is used to detect anomaly based network intrusions and demonstrated using KDD Cup 1999 benchmark dataset, which gives better result in detecting anomaly intrusion detection compared with other techniques [12].

The neural network is used to detect anomaly intrusions. Every day the system administrator checks the user's sessions. In case if there is no match in their normal pattern, the investigation can be launched. The NNID model implemented in a UNIX environment keep the log files when the commands executed to detect intrusions in a network computer system [13]. The genetic neural network combines the good global searching ability of genetic algorithm with the accurate local searching feature of Back Propagation networks to optimize the initial weights of neural networks. The result shows fast learning speed and high-accuracy categories [15], [18].

A Rough Set Neural Network Algorithm reduces a number of computer resources required to detect an attack from host based. The KDDCup'99 dataset is used to test the data and given the better and robust result [16]. The signature based intrusion detection system is used to monitor the packets from the network and this packet has been compared with signature database. VC++ software is used for implementation [17].

The feature reduction techniques such as Independent Component Analysis, Linear Discriminant Analysis and Principal Component Analysis reduce the computational intensity. KDD Cup 99 dataset is used to reduce computation time and improve the accuracy of the systems [19].

In this paper, various methods for Intrusion Detection System are reviewed. The host based or network based attacks comprises the information to protect the data or information from unauthorized users.

The main objective of the system is to detect the new intruder. The intrusion detection system has been developed using various methods and techniques, these are used to find the new threats over the hosts or networks. The dataset has been used for training and testing the different types of attacks using various techniques. The overall performance and accuracy of the system can be improved a lot.

REFERENCES

[1] KDD Cup 1999 Intrusion Detection Data available in the following link,
http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

[2] DARPA Intrusion Detection Data Sets available in the following link,
http://www.ll.mit.edu/mission/communications/ist/corpora/ideval/data/index.html.

[3] Hadi Bahrbegi, Ahmad Habibizad Navin, Amir Azimi Alasti Ahrabi, Mir Kamal Mimia and Amir Mollanejad, "A New System to Evaluate GA-based Clustering Algorithms in Intrusion Detection Alert Management System", *Second World Congress on Nature and Biologically Inspired Computing*, pp. 115-120, 2010.

[4] Nong Ye, Syed Masum Emran, Qiang Chen and Sean Vilbert, "Multivariate Statistical Analysis of Audit Trails for Host-Based Intrusion Detection", *IEEE Transactions on Computers*, Vol. 51, No. 7, pp. 810-820, 2002.

[5] Jiankun Hu, Xinghuo Yu, D. Qiu and Hsiao-Hwa Chen, "A Simple and Efficient Hidden Markov Model Scheme for Host-Based Anomaly Intrusion Detection", *IEEE Network*, Vol. 23, No. 1, pp. 42 – 47, 2009.

[6] Tartakovsky A. G, Rozovskii B. L, Blazek R.B and Hongjoong Kim, "A Novel Approach to Detection of Intrusions in Computer Networks via Adaptive Sequential and Batch-Sequential Change-Point Detection Methods", *IEEE Transactions on Signal Processing*, Vol. 54, No. 9, pp. 3372 – 3382, 2006.

[7] Ioannis Ch. Paschalidis and Georgios Smaragdakis, "Spatio-Temporal Network Anomaly Detection by Assessing Deviations of Empirical measures", *IEEE/ACM Transactions on Networking*, Vol. 17, No. 3, pp. 685 – 697, 2009.

[8] Weiming Hu, Wei Hu and Maybank S, "AdaBoost-Based Algorithm for Network Intrusion Detection", *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, Vol. 38, No. 2, pp. 577 – 583, 2008.

[9] Gupta K. K, Nath B and Kotagiri R, "Layered Approach Using Conditional Random Fields for Intrusion Detection", *IEEE Transactions on Dependable and Secure Computing*, Vol. 7, No. 1, pp. 35 – 49, 2010.

[10] M. Mehdi, S. Zair, A. Anou and M. Bensebti, "A Bayesian Networks in Intrusion Detection Systems", *Journal of Computer Science*, Vol. 3, No. 5, pp. 259-265, 2007.

[11] E. Anbalagan, C. Puttamadappa, E. Mohan, B. Jayaraman and Srinivasarao Madane, "Datamining and Intrusion Detection Using Back-Propagation Algorithm for Intrusion Detection", *International Journal of Soft Computing*, Vol. 3, No. 4, pp. 264-270, 2008.

[12] Mrutyunjaya Panda and Manas Ranjan Patra, "Ensemble Voting System for Anomaly Based Network Intrusion Detection", *International Journal of Recent Trends in Engineering*, Vol. 2, No. 5, pp. 8 – 13, 2009.

[13] Jake Ryan, Meng-Jang Lin and Risto Miikkulainen, "Intrusion Detection with Neural Networks", *Advances in Neural Information Processing Systems*, pp. 943-949, 1998.

[14] Mrutyunjaya Panda and Manas Ranjan Patra, "Some Clustering Algorithms to Enhance the Performance of the Network Intrusion Detection System", *Journal of Theoretical and Applied Information Technology*, Vol. 4, No. 8, pp. 795 – 801, 2008.

[15] Hua Jiang and Junhu Ruan, "The Application of Genetic Neural Network in Network Intrusion Detection", *Journal of Computers*, Vol. 4, No. 12, pp. 1223 – 1230, 2009.

[16] Neveen I. Ghali, "Feature Selection for Effective Anomaly-Based Intrusion Detection", *International Journal of Computer Science and Network Security*, Vol. 9, No. 3, pp. 285 – 289, 2009.

[17] Meera Gandhi and S.K. Srivatsa, "Detecting and preventing attacks using network intrusion detection systems", *International Journal of Computer Science and Security*, Vol. 2, No. 1, pp. 49 – 60, 2008.

[18] S. Selvakani and R.S. Rajesh, "Genetic Algorithm for framing rules for Intrusion Detection", *International Journal of Computer Science and Network Security*, Vol. 7, No. 11, pp. 285 – 290, 2007.

[19] V. Venkatachalam and S. Selvan, "Performance Comparison of Intrusion Detection System Classifiers Using Various Feature Reduction Techniques", *International Journal of Simulation*, Vol. 9, No. 1, pp. 30 – 39, 2008.

[20] Sang-Jun Han and Sung-Bae Cho, "Evolutionary Neural Networks for Anomaly Detection Based on the Behavior of a Program", *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, Vol. 36, No. 3, pp. 559-570, 2006.

[21] Suseela T. Sarasamma, Qiuming A. Zhu, and Julie Huff, "Hierarchical Kohonenen Net for Anomaly Detection in Network Security", *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, Vol. 35, No. 2, pp. 302-312, 2005.

[22] Di He and Henry Leung, "Network Intrusion Detection Using CFAR Abrupt-Change Detectors", *IEEE Transactions on Instrumentation and Measurement*, Vol. 57, No. 3, pp. 490-497, 2008.

[23] Dong Song, Malcolm I. Heywood and A. Nur Zincir-Heywood, "Training Genetic Programming on Half a Million Patterns: An Example from Anomaly Detection", *IEEE Transactions on Evolutionary Computation*, Vol. 9, No. 3, pp. 225-239, 2005.

[24] Naeem Seliya and Taghi M. Khoshgoftaar, "Active Learning with Neural Networks for Intrusion Detection", *IEEE International conference on Information Reuse and Integration*, pp. 49-54, 2010.

[25] Rung-Ching Chen, Kai-Fan Cheng, Ying-Hao Chen and Chia-Fen Hsieh, "Using Rough Set and Support Vector Machine for Network Intrusion Detection", *International Journal of Network Security & Its Applications*, Vol. 1, No. 1, pp. 1-13, 2009.

[26] V. Venkatachalam and S. Selvan, "Intrusion Detection using an Improved Competitive Learning Lamstar Neural Network", *International Journal of Computer Science and Network Security*, Vol. 7, No. 2, pp. 255-263, 2007.

[27] Anazida Zainal, Mohd Aizaini Maarof and Siti Mariyam Shamsuddin, "Ensemble Classifiers for Network Intrusion Detection System", *Journal of Information Assurance and Security, Special Issue on Intrusion and Malware Detection*, Vol. 4, No. 3, pp. 217-225, 2009.

[28] Mansour Sheikhan and Amir Ali Sha'bani, "Fast Neural Intrusion Detection System Based on Hidden Weight Optimization Algorithm and Feature Selection", *World Applied Sciences Journal 7 (Special Issue of Computer & IT)*, pp. 45-53, 2009.

[29] Dewan Md. Farid and Mohammad Zahidur Rahman, "Anomaly Network Intrusion Detection Based on Improved Self Adaptive Bayesian Algorithm", *Journal of Computers*, Vol. 5, No. 1, pp. 23-31, 2010.

[30] Iftikhar Ahmad, Azween Abdullah, Abdullah Alghamdi and Muhammad Hussain, "Optimized intrusion detection mechanism using soft computing techniques", *Telecommunication Systems*, Vol. 52, No. 4, pp. 2187-2195, 2013.

[31] Mansour Sheikhan, Zahra Jadidi and Ali Farrokhi, "Intrusion detection using reduced-size RNN based on feature grouping", Neural Computing & Applications, Vol. 21, No. 6, pp. 1185-1190, 2012.

[32] Shingo Mabu, Ci Chen, Nannan Lu, Kaoru Shimada and Kotaro Hirasawa, "An Intrusion-Detection Model Based on Fuzzy Class-Association-Rule Mining Using Genetic Network Programming", *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, Vol. 41, No. 1, pp. 130-139, 2011.

[33] S. Selvakani Kandeeban and R.S. Rajesh, "A Genetic Algorithm Based elucidation for improving Intrusion Detection through condensed feature set by KDD 99 data set", *Information and Knowledge Management*, Vol. 1, No. 1, pp. 1-9, 2011.

[34] A. M. Chandrashekhar and K. Raghuveer, "Performance evaluation of data clustering techniques using KDD Cup-99 Intrusion detection data set", *International Journal of Information & Network Security*, Vol. 1, No. 4, pp. 294-305, 2012.

[35] Mohammad Sazzadul Hoque, Md. Abdul Mukit and Md. Abu Naser Bikas, "An Implementation of Intrusion

Detection System Using Genetic Algorithm", *International Journal of Network Security & Its Applications*, Vol. 4, No. 2, pp. 109-120, 2012.

[36] Murad Abdo Rassam and Mohd. Aizaini Maarofi, "Artificial Immune Network Clustering Approach for Anomaly Intrusion Detection", *Journal of Advances in Information Technology*, Vol. 3, No. 3, pp. 147-154, 2012.

[37] T. Amalraj Victoire and M. Sakthivel, "A Refined Differential Evolution Algorithm Based Fuzzy Classifier for Intrusion Detection", *European Journal of Scientific Research*, Vol. 65, No. 2, pp. 246-259, 2011.

[38] Dahlia Asyiqin Ahmad Zainaddin and Zurina Mohd Hanapi, "Hybrid of Fuzzy Clustering Neural NetworkoOver NSL Dataset for Intrusion Detection System", *Journal of Computer Science*, Vol. 9, No. 3, pp. 391-403, 2013.

[39] Sherif M. Badr, "Implementation of Intelligent Multi-Layer Intrusion Detection Systems (IMLIDS)", *International Journal of Computer Applications*, Vol. 61, No. 4, pp. 41-49, 2013.

[40] Sherif M. Badr, "Adaptive Layered Approach using C5.0 Decision Tree for Intrusion Detection Systems (ALIDS)", *International Journal of Computer Applications*, Vol. 66, No. 22, pp. 18-22, 2013.

[41] Thuzar Hlaing, "Feature Selection and Fuzzy Decision Tree for Network Intrusion Detection", *International Journal of Informatics and Communication Technology*, Vol. 1, No. 2, pp. 109-118, 2012.

[42] Amin Einipour, "Intelligent Intrusion Detection in Computer Networks Using Fuzzy Systems", *Global Journal of Computer Science and Technology, Neural & Artificial Intelligence*, Vol. 12, No. 11, pp. 19-29, 2012.

[43] Mostaque Md. Morshedur Hassan, "Current Studies on Intrusion Detection System, Genetic Algorithm and Fuzzy Logic", *International Journal of Distributed and Parallel Systems*, Vol. 4, No. 2, pp. 35-47, 2013.

[44] Thakare S. P and Ali M. S, "Network Intrusion Detection System & Fuzzy Logic", *BIOINFO Security Informatics*, Vol. 2, No. 1, pp. 23-27, 2012.

[45] Swati Dhopte and N. Z. Tarapore, "Design of Intrusion Detection System using Fuzzy Class-Association Rule Mining based on Genetic Algorithm", *International Journal of Computer Applications*, Vol. 53, No. 14, pp. 20-27, 2012.

[46] Javier G. Marín-Blázquez and Gregorio Martínez Pérez, "Intrusion detection using a linguistic hedged fuzzy-XCS classifier system", *Soft Computing – A Fusion of Foundations, Methodologies and Applications – Special Issue on Evolutionary and Metaheuristics based Data Mining*, Vol. 13, No. 3, pp. 273–290.

[47] Mohammad Saniee Abadeh and Jafar Habibi, "A Hybridization of Evolutionary Fuzzy Systems and Ant Colony Optimization for Intrusion Detection", *The ISC International Journal of Information Security*, Vol. 2, No. 1, pp. 33-46, 2010.

[48] Ajith Abraham, Ravi Jain, Johnson Thomas and Sang Yong Han, "D-SCIDS: Distributed soft computing intrusion detection system", *Journal of Network and Computer Applications*, Vol. 30, No. 1, pp. 81-98, 2007.