An Ensemble Model for Classification of Attacks with Feature Selection based on KDD99 and NSL-KDD Data Set

Akhilesh Kumar Shrivas
Bilaspur (C.G.), India

Amit Kumar Dewangan
M. E. (CSE), SRIT Jabalpur
(M.P.), India

ABSTRACT

Information security is extremely critical issues for every organization to protect information from unauthorized access. Intrusion detection system has one of the important roles to prevent data or information from malicious behaviours. Basically Intrusion detection system is a classifier that can classify the data as normal or attacks. In this research paper, we have proposed ANN-Bayesian Net-GR technique that means ensemble of Artificial Neural Network (ANN) and Bayesian Net with Gain Ratio (GR) feature selection technique. We have applied various individual classification techniques and its ensemble model on KDD99 and NSL-KDD data set to check the robustness of model. Due to irrelevant features in data set, also applied Gain Ratio feature selection technique on best model. Finally our proposed model produces highest accuracy compare to others.

Keywords

Intrusion Detection System, Artificial Neural Network (ANN), Ensemble Model, Feature Selection (FS), Gain Ratio (GR).

1. INTRODUCTION

Now a days the rapid development and popularity of Internet and Intranet, the security is very important for network. IDS is an emerging area of research in computer security and network with growing usages of Internet and Intranet in everyday life. IDS can identify the user's activity as either normal or anomaly (Intrusion) and protect system for unauthorized users or attackers .There are various techniques applied by different authors to develop an Intrusion Detection System (IDS) in which data mining technique is one of the most widely used for classification of data. Li, Y. et al. [6] have applied various feature reduction method on KDD99 data set. In case of Gradually Feature Reduced (GFR) with 19 features, Support Vector Machine (SVM) classifier achieved high accuracy with 98.62% for intrusion detection. Koc, L. et al. [5] have introduced Hidden Naive Bayes (HNB) model with promotional k-interval discretization and INTERACT feature selection method to develop IDS. They have compared proposed model with traditional Naive Bayes methods. Our proposed model gives satisfactory result with 93.72% of accuracy in multiclass classification problem for intrusion detection in case of KDDCUP99 data set. Altwaijry, H. et al. [7] have suggested Bayesian network to improve the accuracy of R2L type of attack.

Different feature subset of KDD99 data set are applied on proposed model which gives better results for R2L attack with detection rate 85.35% using 3 features. Chung, Y.Y. et al. [8] have proposed a new hybrid approach known as network intrusion detection system using intelligent dynamic swarm based rough set (IDS-RS) for feature selection and simplified swarm optimization with weighted local search (SSO-WLS) strategy for intrusion data classification. Proposed hybrid model SSO-WLS improve the overall performance of the network intrusion detection system with 99.3% of accuracy .Amira Sayed A. Aziz, et al. [9] have proposed Minkowski distance technique based on genetic algorithm to develop a classifier (IDS) to detect anomalies. The proposed Minkowski distance techniques applied on NSL-KDD data which gives satisfactory detection rate. They have also compared proposed technique with Euclidean distance as well as other techniques but achieved high accuracy with 82.13% in case of Minkowski distance.

2. METHODS AND MATERIALS

There are various data mining, statistical techniques and benchmark data set used in this research work for intrusion detection system are explained below:

2.1 Classification and Regression Tree (CART)

CART (Classification and Regression Tree) [2] is one of the popular methods of building decision tree in the machine learning community. CART builds a binary decision tree by splitting the record at each node, according to a function of a single attribute. CART uses the gini index for determining the best split. The initial split produces the nodes, each of which we now attempt to split in the same manner as the root node. Once again, we examine the entire input field to find the candidate splitters. If no split can be found then significantly decreases the diversity of a given node, we label it as a leaf node. Eventually, only leaf nodes remain and we have grown the full decision tree. The full tree may generally not be the tree that does the best job of classifying a new set of records, because of overfitting.

2.2 Artificial Neural Network (ANN)

An Artificial Neural Network [3] is composed of a set of elementary computational units, called neurons, connected together through weighted connections. These units are organized in layers so that every neuron in a layer is exclusively connected to the neurons of the preceding layer and the subsequent layer. A multilayer feed-forward neural network consists of an input layer, one or more hidden layers, and an output layer.ANN is known as best classifier and is able to mine huge amount of data for classification.

2.3 Bayesian Net

Bayesian Net [3] is statistical classifiers which can predict class membership probabilities, such as the probability that a given tuple belong to a particular class. Let X is a data sample whose class label is unknown. Let H be some hypothesis, such as that the data sample X belongs to a specified class C. For classification problems, we want to determine P(H|X), the probability that the hypothesis H holds given the observed data sample X. P(H|X) is the posterior probability, or a posteriori probability of H conditioned on X. In contrast, P(H) is the prior probability, or a priori probability of H. The posterior probability, P(H|X) is based on more information (such as background knowledge) than the prior probability, P(H), which is independent of X. Bayesian theorem is useful in that it provides a way of calculating the posterior probability (H|X) from P(H), P(X), and P(X|H). Bayesian theorem is:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

2.4 Ensemble Approach

When Two or more trained models are ensemble together to form a new model known as ensemble model. An ensemble model is defined as a set of individually trained classifier whose predictions are combined when classifying a new data. An ensemble model [1] combines the output of several classifier produced by weak learner into a single composite classification. It can be used to reduce the error of any weak learning algorithm. The purpose of combining all these classifier together is to build a hybrid model which will improve classification accuracy as compared to each individual classifier.

2.5 Feature Selection

Feature selection is one of the important roles that can reduce irrelevant feature and improve the performance of model. The main goal of feature selection is to find a feature subset that produces higher classification accuracy. In this paper we have used Gain Ratio (GR) raking based feature selection technique. The extension to information gain known as gain ratio [3] based on ranking, which attempts to overcome bias. It applies a kind of normalization to information gain using a "split information" value defined analogously with *Info*(*D*) as

$$\text{SplitInfo}_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right)$$

This value represents the potential information generated by splitting the training data set, *D*, into *v* partitions, corresponding to the *v* outcomes of a test on attribute *A*. For each outcome, it considers the number of tuples having that outcome with respect to the total number of tuples in *D*. It differs from information gain, which measures the information with respect to classification that is acquired based on the same partitioning. The gain ratio is defined as

$$GainRatio(A) = Gain(A)/SplitInfo(A)$$

The attribute with the maximum gain ratio is selected as the splitting attribute. However, that as the split information approaches 0, the ratio becomes unstable. A constraint is added to avoid this, whereby the information gain of the test selected must be large at least as great as the average gain over all tests examined.

2.6 Proposed ANN- Bayesian Net-GR Technique

The proposed technique ANN- Bayesian Net-GR is based on ensemble and feature selection technique. In this proposed model, we have ensemble two techniques as Artificial Neural Network (ANN) and Bayesian Net. This ensemble model gives higher accuracy compared two each individual model like ANN and Bayesian Net. Feature selection is also one of the most important roles to reduce the irrelevant features and improve classification accuracy. Gain Ratio (GR) feature selection applied on ensemble of ANN and Bayesian Net techniques which gives higher accuracy with less number of features. Figure 1 depicted proposed model of this research work.

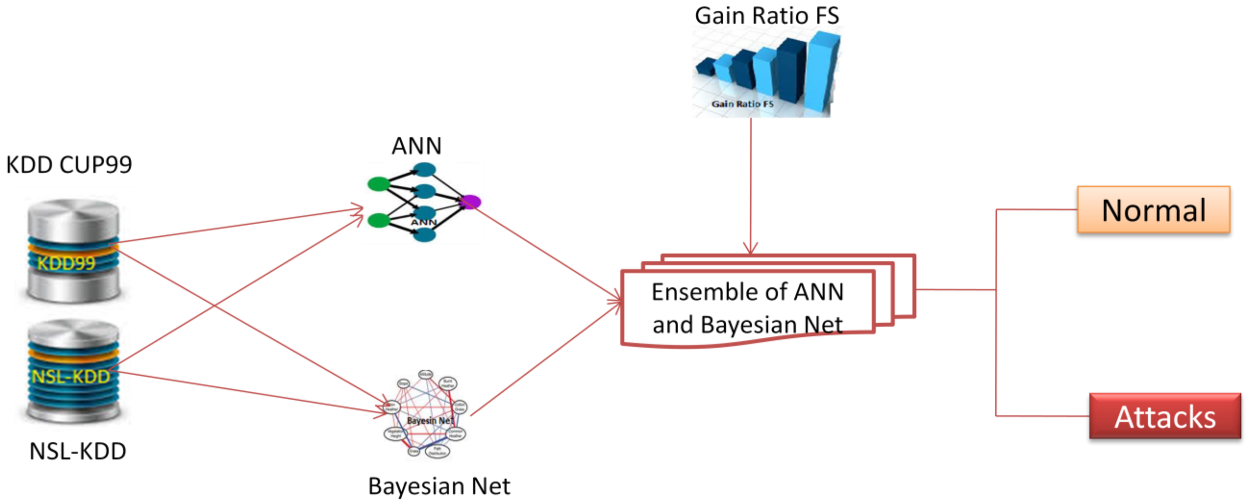
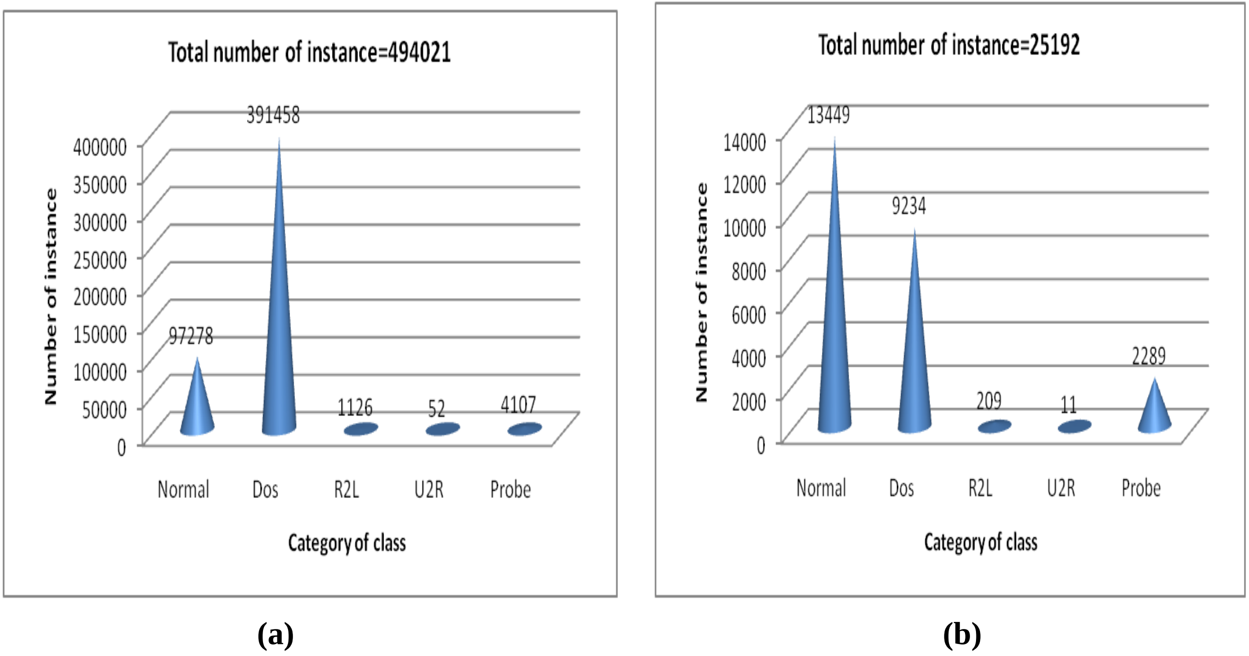


Figure 1: Proposed model for classification of attacks

2.7 Data Set

Publicly data set available for the evaluation of intrusion detection system are KDD99 and NSL-KDD data set [4].The NSL-KDD data set solving some of the inherent problems of the KDD99 data set. One of the most important efficiencies in KDD99 data set is the huge number of redundant records, which causes the learning algorithms to be biased towards the frequent records, and thus prevent them from learning infrequent records which are usually more harmful to network such as U2R and R2L attacks.

In this experiment we have used 494021 records of KDD99 data set and 25192 records of NSL-KDD data set .This data set contains one type of normal and four type of attacks data Like DoS,R2L,U2R and Probe. The experiment done with both data set to check the robustness of model.



(a)

(b)

Figure 2: Different attacks and normal category along with sample size of (a) KDDCUP99 (b) NSL-KDD data set

From the above figure , it is clear that data set highly unbalanced or there is no uniform distribution of samples for each type of attacks. Number of samples of DoS type attack is high while on the other hand U2R type of attack has less samples. This unbalanced distribution of samples may create problem during learning of any data mining based classification model. The features of NSL-KDD data set same as KDD99 data set as shown in figure 3 and different types of attacks and it category shown in table 1.

Table 1: Different types of attack and its category

Category of attack	Attack Name
DoS	Back,land,Neptune,pod,teardrop,smurf
R2L	ftp_write,guess_passwd,imap,multihop,phf,spy,warezclient,warezmaster
U2R	Buffer_overflow,load_module,perl,rootkit
Probe	Ipsweep,nmap,portsweep,satan
Normal	Normal

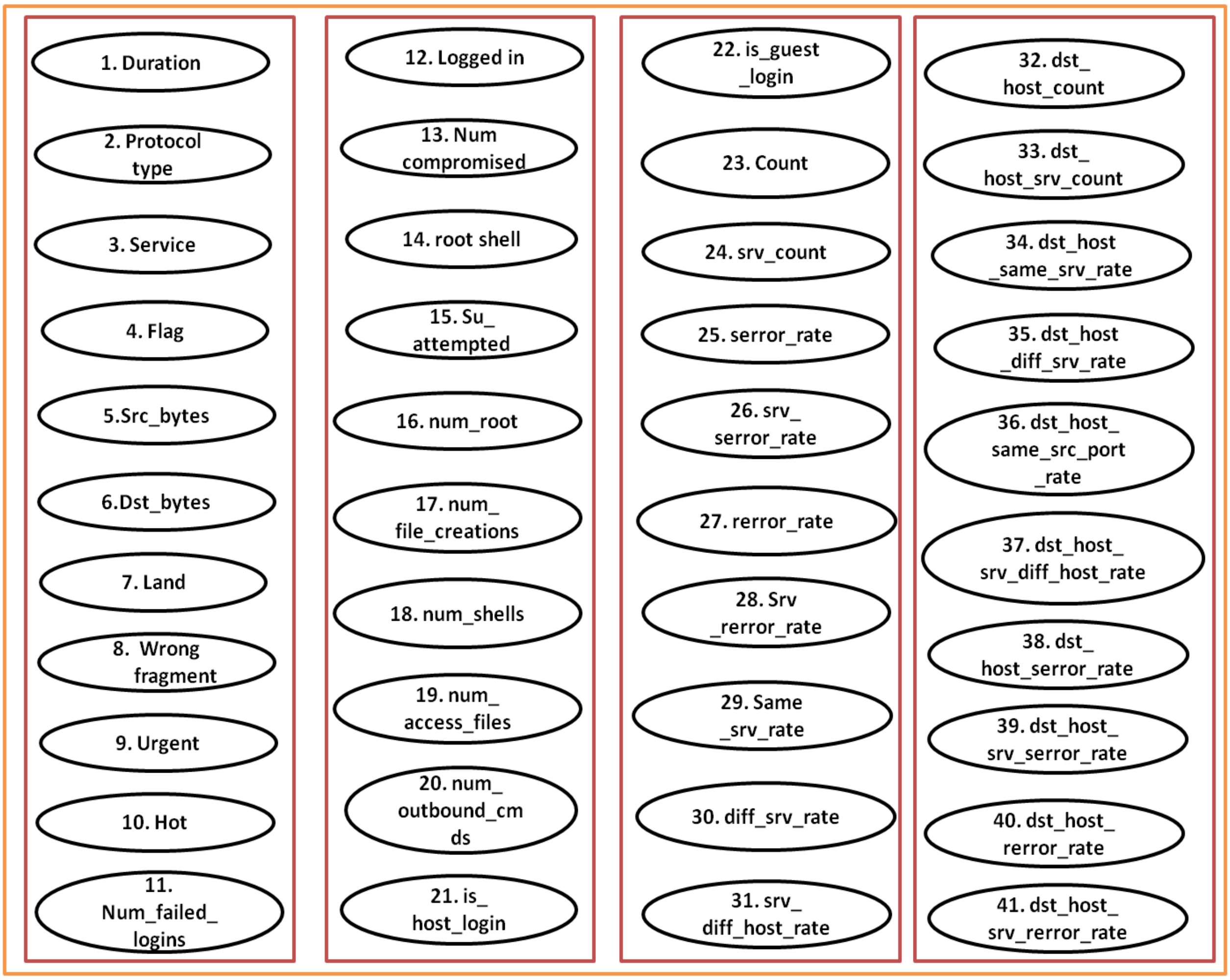


Figure 3: Features of KDD 99 and NSL-KDD Data set

3. EXPERIMENTAL WORK

The experiment done into two parts: Firstly various classification techniques applied on different partitions of NSL-KDD and KDD99 data set and secondly feature selection technique applied on best model in case of both data set.

In this experiment, we have applied various partitions of NSL-KDD and KDD99 data into different classification techniques like CART, ANN, Bayesian Net and its ensemble techniques as shown in table 2. From table 2, it is clear that accuracy of models is varying from one portion to another. Simulated result shows that accuracy for proposed ensemble of ANN and Bayesian Net is the best as compare to its individuals and other ensemble models. Accuracy of proposed model is consistent (99.41%) in case of KDD99 data set with all partitions of data set like 70-30%, 80-20% and 90-10% as training-testing , but accuracy of proposed model is highest 97.76% in case of NSL-KDD data set with 80-20% training-testing partitions. We can say that proposed model achieved highest accuracy and developed a robust model for Intrusion detection.

Table 2: Accuracy of model with different partitions of data set

Models	**NSL-KDD**			**KDD**		
	70-30% Partition	***80-20% Partition***	***90-10% Partition***	***70-30% Partition***	***80-20% Partition***	***90-10% Partition***
CART	95.90	96.58	96.56	97.57	97.55	97.51
ANN	96.97	97.06	95.98	99.36	99.09	99.17
Bayes net	97.13	97.37	97.02	99.27	99.28	99.27
ANN+Bayes Net	**97.53**	**97.76**	**97.53**	**99.41**	**99.41**	**99.41**
CART+BayesNet	97.58	97.61	97.49	99.35	99.34	99.33

Irrelevant features are also increase computational time and decrease performance, due to this reason we have applied feature selection technique on best model as ensemble of ANN and Bayesian Net. Table 3 shows that accuracy of best model with reduced number of features in case of NSL-KDD and KDD99 data set. We have applied Gain Ratio (GR) feature selection techniques on this model. In case of NSL-KDD data set, our model gives 98.07% and 97.78% accuracy with 35 and 29 features respectively, Which is higher than with all other feature subsets.In KDD99 data set, our proposed model gives 99.42% with 31 features which is higher than 0.1% compare to with all features and 99.38% with 29 features which is less than 0.3% compare to with all features but 10 numbers of features is reduced which computationally improve the performance of model. Figure 4 shows that accuracy of proposed model where x-axis represents number of features and y-axis represent accuracy of proposed model.

Table 3: Accuracy of best model with Gain Ratio (GR) feature selection technique

Data Set	***Number of features***	***Features _ID***	***Accuracy***
NSL – KDD	*35*	*9,26,25,4,12,39,30,38,6,29,5,37,11,3,22,35,34,14,33,23,8,10,31,27,28,32,1,36,2 ,41,40,17,13,16,19*	***98.07***
	29	*9,26,25,4,12,39,30,38,6,29,5,37,11,3,22,35,34,14,33,23,8,10,31,27,28,32,1,36,2*	***97.78***
KDD 99	*31*	*9,26,25,4,12,39,30,38,6,29,5,37,11,3,22,35,34,14,33,23,8,10,31,27,28,32,1,36,2 ,41,40*	***99.42***
	29	*9,26,25,4,12,39,30,38,6,29,5,37,11,3,22,35,34,14,33,23,8,10,31,27,28,32,1,36,2*	***99.38***

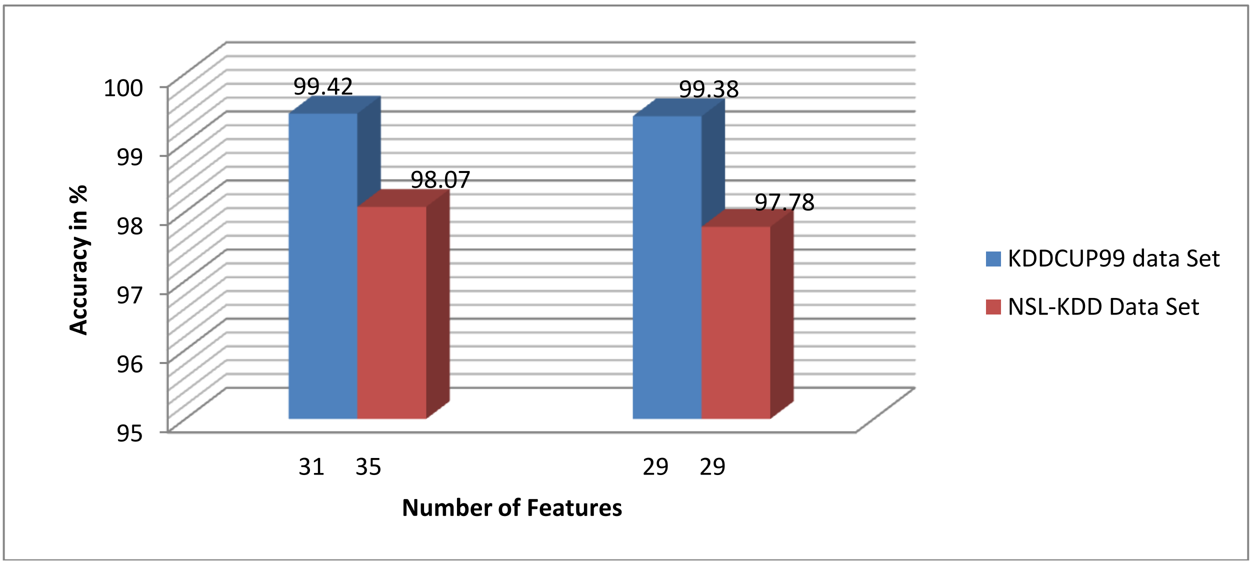


Figure 4: Accuracy of proposed model

4. CONCLUSION

To rapid development of information technology, protecting of the information is crucial issues from malicious behaviours or attackers. Intrusion detection system is analyser that can analyse or identity information (packets) and classifies this information as attacks or normal. In this paper, we have proposed ensemble of ANN and Bayesian Net classifiers with Gain Ratio (GR) feature selection technique for intrusion detection system. This proposed model gives accuracy of 99.42% with KDD99 data set and 98.07% with NSL-KDD data set in case of 35 and 31 features respectively. Finally proposed model is a robust classifier as intrusion detection system which achieved the highest accuracy.

5. REFERENCES

[1] Pal, M. 2007. Ensemble learning with decision tree for remote sensing classification. World Academy of Science, Engineering and Technology , Vol. 36, pp. 258-260.

[2] Pujari, A. K. 2001. Data mining techniques. Universities Press (India) Private Limited, Fourth Edition.

[3] Han, J. and Kamber, M. 2006. Data Mining Concepts and Techniques. Morgan Kaufmann, Second Edition.

[4] UCI Machine Learning Repository of machine learning databases 2010. University of California, school of Information and Computer Science, Irvine. C.A. web site: http://www.ics.uci.edu/~mlram,?ML.Repositary.html. Last accessed (Oct 2013).

[5] Koc, L., Thomas A. M. and Sarkani S. 2012. A network intrusion detection system based on hidden Naive bayes multiclass classifier .Journal of Expert system with applications, Vol. 39, pp. 13492-13500.

[6] Li, Y. ,Xia J., Zhang S. , Yan J., Ai X. and Dai K. 2012. An efficient intrusion detection system based on support vector machines and gradually feature removal method. Expert systems with Applications, Vol. 39, pp. 424-430.

[7] Altwaijry, H., and Algarny S. 2012. Bayesian based intrusion detection system . Journal of king saud University-computer and information sciences. Vol. 24, pp. 1-6.

[8] Chung, Y.Y. and Wahid N. 2012. A hybrid network intrusion detection system using simplified swarm optimization (SSO). Applied soft computing, Vol. 12 , pp. 3014-3022.

[9] Amira Sayed A. Aziz, Mostafa A. Salama, Hassanien A. E. and Sanaa El-Ola Hanafi 2012. Artificial Immune System Inspired Intrusion Detection System Using Genetic Algorithm. Informatica, Vol. 36, pp. 347–357.