Anomaly based IDS using Backpropagation Neural Network

Vrushali D. Mane
ME (Electronics)
JNEC, Aurangabad

S.N. Pawar
Associate Professor
JNEC, Aurangabad

ABSTRACT
Intrusion means illegal entry or unwelcome addition of the system. So, Intrusion detection system is used to find out the signatures of an intrusion. The goal of the system is to protect system for various network attacks like Dos, U2R, R2L, Probing etc. Intrusion detection system (IDS) collects information from various parts of network and system. This paper introduces the Anomaly Intrusion Detection System that can detect various network attacks.The aim of this work is to identify those attacks with the support of supervised neural network, i.e. back propagation artificial neural network algorithm and make complete data safe. In this paper, system comprises experimenting neural networks that use only the (17 of 41) most significant features of the KDD 99 dataset**.** The proposed IDS use a supervised neural network to study system's performance.

General Terms
Artificial Neural Network, Back Propagation algorithm, Intrusion Detection, KDD 99 dataset.

Keywords
KDD 99 dataset, Network Attack.

1. INTRODUCTION
The Evolution in the computer technology is outstanding in 19th century where human being uses computers for the different purposes. The Computers helps human to reduce his overhead in terms of data collection, education up gradation, home shopping, and one click data transmission and so on. These services help as well as threaten human if one consider security and privacy of the data. The human uses internet but become victim of various types of attacks, misuse, viruses, anomaly etc. The organizations, enterprises, institutes has database and they share it through the internet. Such large network requires lot of security from attacker, viruses, anomalies, misuse. The security and privacy still facing perfection issue. The network administrator must address elements of security like availability, utility, integrity, authenticity, confidentiality, possession [5]. The network not only open door to many positive information but also to the things which are harmful to the network. The unauthenticated and stalkers effectively step in to the networks and pull down the prominent companies. Much research work going on in. this direction much technology came and became successful. The methods and the performance make effect on the quality of services for security. The techniques like firewall observe network incoming and outgoing traffic at the edges. The attacks, anomalies, misuse are monitored and detected effectively by some techniques; such system is called as Intrusion Detection System (IDS). The IDS system examines and collects the data from all over the network to detect the susceptible components. The difference between traditional firewall and IDS is the passiveness of the firewall where IDS works online to sniff the packet. The examination and observation done by IDS after catching packets. The report generated after such process is stored for the administrator. The report contains both affected and not affected packets. IDS performance is remarkable comparatively but it is hard to attain. The IDS can notify the admin of the network if the unlawful activities are found in the network. IDSs are categorized, based on their action, as misuse detectors and anomaly detectors. Misuse detection system practices well defined forms of event which are matched against operator performance to detect intrusions. The misuse finding is easy than anomaly finding from the network. The misuse found by just comparisons of rules, protocols, and digital signatures. The anomaly detection must know the normal behavior of the network and then by inspecting the data, anomaly is analyzed [1].

Misuse discovery is founded on the arrays of famous bouts. The unidentified attacks are out of vicinity of misuse discovery methods where anomaly discovery systems are actual approach to sense unidentified attacks. If the current behavior of the system is uncertain it can have anomaly [2]. There are several approaches for intrusion detection process but the detection capability is not perfect of any one approach. The IDS system can collect and examine the data for the normal and abnormal behavior of the network. This event can be warned to the network administrator. The high rate of the warning decreases the quality of the system. The effective system of the network will warn only some time in a day and make the system usable. [3]. The effective Intrusion system can observe and examine user and system action, check system configurations and susceptibilities, evaluating the veracity of critical system and data files, gives statistics of normal and abnormal patterns of the data based on the attack, Operating system inspection[4]. The intrusion detection system can detect various types of threats from the system like DOS, Disclosure, Manipulation, Masqueraders, replay, repudiation, physical Impossibilities, Device Malfunctions [6]. An IDS is the old technique and evolved from 1960 in which computer access is restricted by the time sharing system. IDS upgraded to network intrusion detector, Network Anomaly Detection and Intrusion Reporter, Distributed Intrusion detection System [7]. The Intrusion Detection System will not work independently to observe normal and abnormal data. The classifier is the one which classify the data into categories. The classifier can be of various types and in which different methods can be used to identify the class of the data. The abnormal data of the system can be classified among DOS, PROB, R2L, U2L and detailed classification mode. The rest of the section organized as follows: Section 2 represents a related work in IDSs. The design stages of the proposed IDS are introduced in Section 3. Next, Section 4 discusses some implementation issues and experimental results. Finally, the paper is concluded in Section 5

2. REVIEW OF LITERATURE
The Intrusion Detection System contains classifier for categorizing anomalies affected data as abnormal and rest as a normal data. The applications for IDS use number of classifier

like Support Vector Machine (SVM), Artificial Neural Network (ANN), Fuzzy Logic etc. The classifier can use verity of methods to fulfill the task, some of the classifier and their methods are as Self-Organizing Maps (SOMs). The anomalies are identified by this method uses time window method which allows simplifying inputs further. This method is clustering technique to group inputs if they are dynamic. The ANN is trained for such inputs and gives decisions quickly for real time identifications [8]. The one class support vector machine categorizes negative class and positive class of data by identifying the start of anomalies and output organization possibility to monitor status of the system [9]. The KDD Cup 1999 dataset is analyzed using linear SVM for discovering the abnormal behavior of the data in network. This method classifies the intrusions into categories like DoS, probe, R2L and U2R. The KDD dataset is disturbed with all such categories [10]. The Decision Tree and SVM is another combination for multiclass problems. This method increases efficiency of the system by reducing the time of training and testing inputs to machine. The binary trees can be formed by categorizing datasets into two sets from root node to leaf node. The division is continued till every set contains only one class. The tree formation has great influence on the performance of the data categorization [1]. A hierarchical off-line anomaly network intrusion detection system based on Distributed Time-Delay Artificial Neural Network is used for IDS [11].Hierarchical clustering method of SVM improves training time for large datasets [12].The intrusion Detection can occur using Rough Set Theory (RST) method and Support Vector Machine. The feature selection can be done by RST to make an input samples. The SVM is trained and tested respectively for these samples [13]. The classifier Support Vector Machines (SVM) uses a weighted voting schema to detect intrusions. The features of the sample inputs are selected by the entropy and TF-IDF (term frequency and inverse documents frequency) and sent to the SVM model for learning and testing. To detect the intrusion in such samples voting schema named Weighted Voting SVM (WV-SVM) is used [21].

The fuzzy logic is a algorithm to classify the data and can be used with different methods. The fuzzy logic and neural network collectively used for the detection of intrusion which detects new attacks with high rate of detection and low false rate. The system effectively identifies attack in the network by using logic of ANN and Fuzzy Logic [14]. Mobile Ad-Hoc Network is used for the data transmission and communication. This network can also be affected by the intrusions and anomalies. IDS application used to identify attacks and intrusions in the MANETs. The possible attack in the network is black-hole. The fuzzy logic works to identify such attack in the MANETs and also report the behavior of the network to different nodes present in the network. The fuzzy logic contains set of rules applied on the set of features selected by the method of Principal Component Analysis (PCA). The reduced number of samples provided to fuzzy logic based system to monitor the network for the anomalies and intrusions. The fuzzy rules are generated by automated strategy using definite rules and frequent data sets. The fuzzy logic used for data mining technique to identify anomaly based intrusion which shows different behavior than the stored normal data. The rule based expert system is used to identify the misuse in the network. The changes in the normal data are captured to find out the intrusion in the normal data called as misuse components [17].

The general regression neural network (GRNN) is an algorithm which uses variables to meet the fundamental regression surface. It is one pass algorithm having parallel structure. The algorithm provides smooth outcome for multidimensional sparse data. The algorithm is best suited for the problems having no linearity [18]. The anomaly detection using reduced number of samples can be possible with hyper graph representation. The Expectation-Maximization algorithm is used with hyper graph method without feature reduction [19].

In distributed systems security is provided by authentication, digital signatures, confidentiality in the data transmission. But such security system is not sufficient to prevent malicious attacks, intrusions anomaly entry. The distributed system can be secured by the Grid and Cloud Computing Intrusion Detection System which integrates knowledge and behavior analysis to detect intrusions. Intrusion-detection systems (IDSs) can provide extra security actions for these networks by investigating configurations, logs, network traffic, and user actions to identify typical attack behavior to stop attackers [20].

3. SYSTEM IMPLEMENTATION

The proposed system for Intrusion Detection System uses Artificial Neural Network (ANN) as a classifier. The performance of the system is evaluated using KDD Cup 1999 dataset samples. Figure 3.1 shows the step by step flow of the system which includes stages like Data collection, Data Preprocessing, Representation & Normalization, Dimensionality Reduction, and Selection of Network Structure, Training & Testing & Attacks Classes (Normal, DOS, Probe, U2R, and R2L).

3.1 Data collection

The standard KDD Cup 1999 database encloses connection records of attacks and intrusions in a network. The system uses KDD dataset for testing and training samples. Each record from KDD dataset, describes one connection in the form of 41 features. The features are categorized as basic features which are took from packet headers only and without examining payload. Features 1 to 6 are in this category content features contain information of original tcp packets examined with support of domain knowledge. An example of this category is number of "hot" indicators. To capture Time-based Traffic types of features a window of 2 second intermission is defined. In this intermission, some properties of packets are measured. For example number of connections to the same service as the current connection in the past two seconds. In Host-based Traffic Features category instead of a time based window, a number of connections are used for building the window. This category is designed so that attacks longer than 2 second can be detected. To do so, KDD 99 dataset should be divided into two parts of Testing and Training. Before using this database, it is necessary to mention that this database has a high capacity and proposed system uses only 10% of the records for testing and training the designed network. Of course, this 10% is chosen in such a way that contains different types of attacks includes different modes of the network. So KDD Cup 99 data is preprocessed and added as input to neural network. Table 1 gives details of the features used to learn and test the system.

Table 1: A Selected input features from KDD 99 dataset

Sr. No.	Feature name	Type
1	protocol_type	Symbolic
2	Service	Symbolic
3	src_bytes	Continuous
4	wrong_fragment	Continuous
5	Flag	Symbolic
6	num_failed_login	Continuous
7	logged_in	Symbolic
8	root_shell	Continuous
9	Count	Continuous
10	serror_rate	Continuous
11	serror_rate	Continuous
12	rerror_rate	Continuous
13	srv_serror_rate	Continuous
14	srv_rerror_rate	Continuous
15	same_srv_rate	Continuous
16	diff_srv_rate	Continuous
17	srv_count	Continuous

Table 2: Attacks Type

1	Denial of Service Attacks	Back, land, neptune, pod, smurf, teardrop
2	User to Root Attacks	Buffer_overflow, loadmodule, perl, rootkit,
3	Remote to Local Attacks	Ftp_write, guess_passwd, imap, multihop, phf, spy, warezclient, warezmaster
4	Probes	Satan, ipsweep, nmap, portsweep

3.2 Data Preprocessing and Normalization

The preprocessing is performed for raw data. In the other word raw packets are converted to structured data which depicts the meaning of the connection. These structures have set of features describing connection. The feature selection regarding connection record requires vast understanding. So that there are only two datasets present for Intrusion Detection: KDD 99[12] and IDEval, while the former is only a refinement version of the later one. The preprocessing converts the features of the samples like protocol type, service and flag into numeric forms. After this transformation the data is normalized in the range of [0, 1] to avoid feature influence. To improve the performance of the system normalization is needed which is applied on the each feature selected. To normalize data in the range of MinX and MaxX. For this all the minimum and maximum values for feature X are converted to [New MinX, New MaxX]. As per equation each value of v is converted to a new value as follows:

$$new_v = \frac{v - min\ X}{max\ X - min\ X} \qquad (3.2.1)$$

3.3 Feature Reduction

If the all measured variables are used as input to neural network, it results in large size of network & hence larger training time. So, to make the NN approach applicable to large scale Intrusion detection problem, some dimensionality reduction is required.

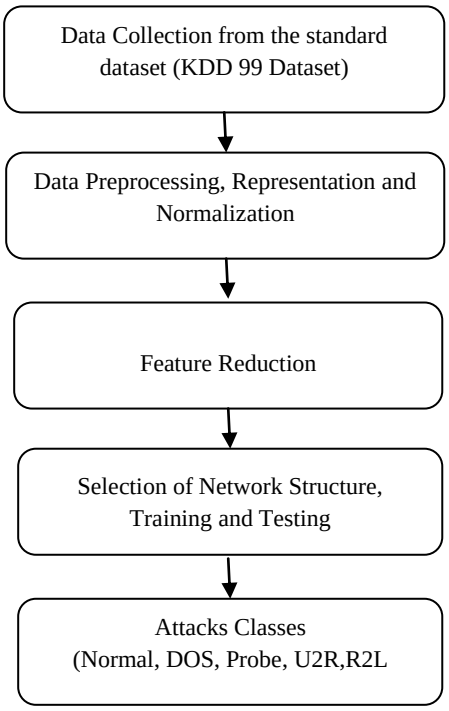

Fig 1: Flow of proposed IDS system.

3.4 Selection of Neural Network

A two layer feedforward network with sigmoid hidden neurons & linear output neurons fit multidimensional database. The number of input nodes of the ANN (17 or 41) represents the number of the selected features of the network. To implement the neural network algorithm, proposed system applied the MATLAB software (version R2012a 7.14.0.739, 32 bit (Win 32 bit)). In order to implement this algorithm, the

system should firstly train the designed neural network using training data, and then analyze the efficiency of the network using the experiment data. The algorithm used in the proposed anomaly intrusion detection system is a neural network of Multilayer type. The computer used for implementation was a Intel (R) core(TM) i3 @ 2.40 GHz CPU with 4.00 GB RAM. After training process the ANN generate the basic Simulink model of the neural network

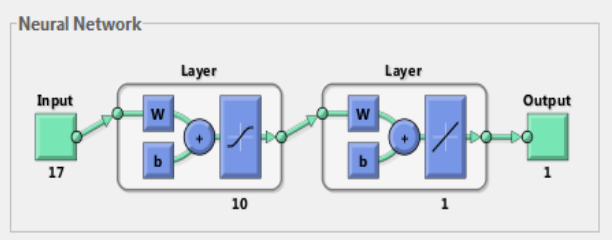

Fig 2: Neural network training model.

3.5 Classification Algorithm

The following steps are

1) Select the appropriate 17 features from standard data set i.e. KDD 99 dataset.

2) Data pre-processing & Normalization

   a) Select nominal features and convert it into numeric form.

   b) Determine the probability using probability density function & change the nominal value with the form of numbers.

3) Create a feedforward network (backpropagation neural network) using the MATLAB software (version R2012a 7) 14.0.739, 32 bit (Win 32 bit)).

4) Add input data to the backpropagation neural network.

5) The MATLAB function neural net used for training a backpropagation neural network & calculate the performance of system.

6) Using cross validation techniques test the input samples and compute accuracy of the system.

7) Compute the normal & abnormal samples in percentages form applied input data

3.6 Back propagation Algorithm

1) Get input output samples patterns from experimental simulation or KDD 99 dataset.

2) Select ANN Characteristic with no of layers, input nodes, hidden nodes, output nodes & activation function, threshold value etc. Activation function is used to calculate the output response of neural net.

3) Set some random value i.e weights.

4) Select an input data & output (target) data.

5) Calculate output & compute error, if error occurs then changes weights by training backpropagation algorithm

   Error=Target value – Output value.

   Error is used to adjustment the weights in such a way that the error will get reduced or it is smaller. The process is frequently use again and again until the error is negligible.

6) Train network with input data patterns which is selected form a standard KDD 99 dataset.

7) Again error occurs, change numbers of neurons in hidden layers. Observe the neural network performance. If it is better than previous results then test the network performance.

4. SYSTEM PERFORMANCE

To implement the neural network algorithm, proposed system applied the MATLAB software (version R2012a 7.14.0.739, 32 bit (Win 32 bit)). In order to implement this algorithm, the system should firstly train the designed neural network using training data, and then analyze the efficiency of the network using the experiment data. Note that the algorithm used in the proposed intrusion detection system is a neural network of MLP type. The computer used for implementation was a Intel (R) core(TM) i3 @ 2.40 GHz CPU with 4.00 GB RAM

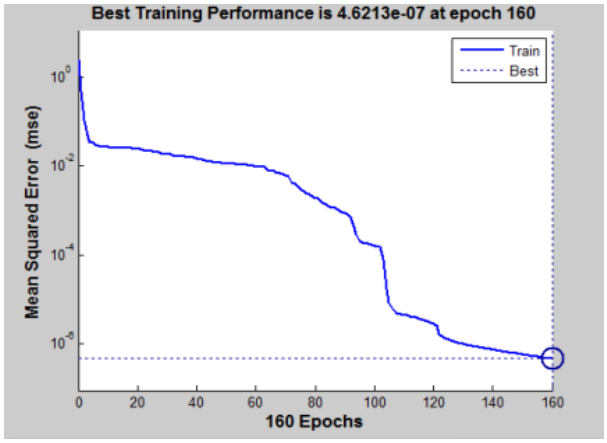

Fig 3: Training performance of ANN

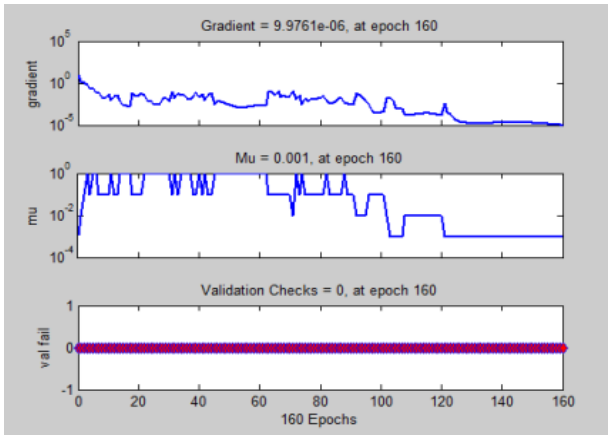

Fig 4: Training state of neural network

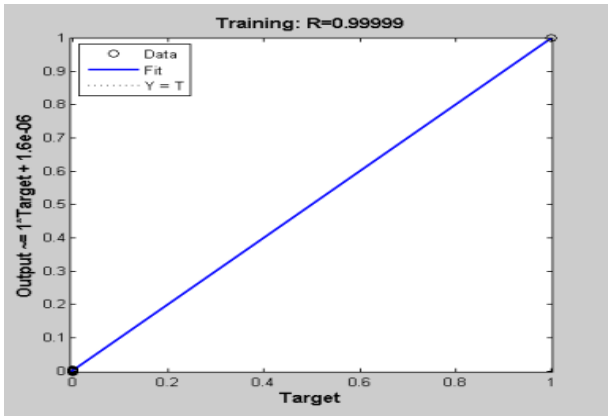

Fig 5: Training Regression of Neural network

5. EXPERIMENTAL RESULTS

Following tables and figures shows the result analysis of the system tested for the 20,000 and 10000 samples. The samples are divided into sets for the better performance of the system. Table shows the result in percentage for normal as well as abnormal data found by the system after training and testing

Table 3: The Detection rates for 20,000 Samples

Test samples	Normal Data (%)	Abnormal Data (%)
2500	**92.80**	**07.20**
3500	**93.50**	**06.40**
4000	**93.20**	**06.70**
10,000	**93.50**	**06.40**

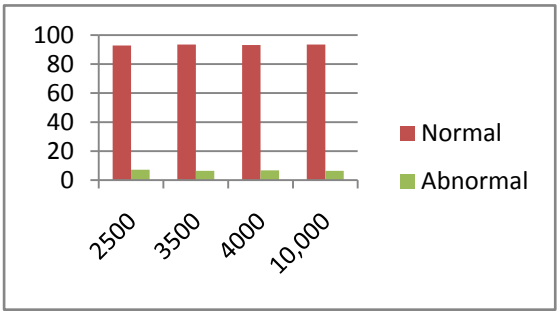

Fig 6: The graph the Detection rates for 20,000 Samples

Table 4: The Detection rates for 10,000 Samples

Test samples	Normal Data (%)	Abnormal Data (%)
1500	**96.60**	**03.33**
3000	**95.00**	**04.90**
4000	**96.60**	**03.33**
2500	**96.20**	**03.70**

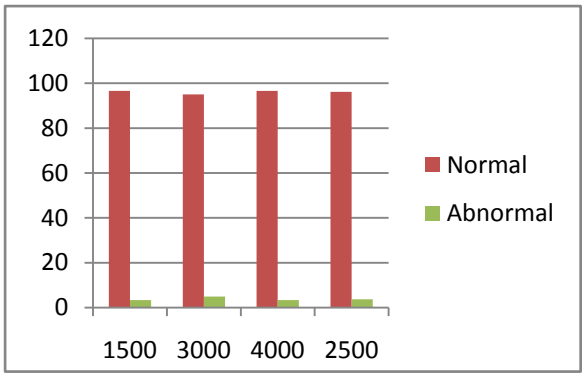

Fig 7: The graph the Detection rates for 10,000 Samples

6. CONCLUSION

The proposed IDS, present a practical solution to hierarchical anomaly intrusion detection system using supervised learning method. The performance of the system is analyzed using 10% of the data from the KDD 99 data set. All classifications were performed on the binary (attack / normal) basis. The KDD training dataset consists of approximately 4,900,000 single connection vectors each of which contains 41 features and is labeled as either normal or an attack, with exactly one specific attack type. Neural network takes a huge amount of time in days for training and testing if totals KDD data set given input to it. So, the performance of the system goes down. So to improve the performance and accuracy, IDS uses only 10 % KDD 99 data. The results obtained shows the capability of the proposed network IDS by reducing the number of features from 41 to 17, which leads to high detection accuracy (98.0%).Also, results of testing with normalization have been generally better than the results of testing without normalizing. The results show the efficiency with 17 features compared to the 41 features, with reduced training and testing times. Hence, feature reduction techniques to improve the efficiency and reduce the false alarm rate. There are different algorithms for implementation and networks intrusion detection in computer networks. In the proposed IDS uses supervised neural network to detect intrusions in KDD 99 database and also improved the speed of implementation.

7. ACKNOWLEDGEMENT

I would like to thanks to my guide Prof. Dr. S.N. Pawar in Electronics and Telecommunication department, Jawaharlal Nehru Engineering College, Aurangabad for guiding me to understand the work conceptually and their constant support during project work. A Very special thanks to Mr. Mahesh Korade, Assistant professor in Computer Engineering Department, Sandip Foundation's Sandip Institute of Technology & Research Center, Nashik for their great source of motivation.

8. REFERENCES

[1] Snehal A. Mulay, P.R. Devale & G.V. Garje "Intrusion Detection System Using Support Vector Machine and Decision Tree", Volume 3 – No.3, June 2010.

[2] E. Eskin, M. Miller, Z. Zhong, et al. Adaptive Model Generation for Intrusion Detection Systems. Proc. Of Workshop on Intrusion Detection and Prevention, 7th ACM Conference on Computer Security, Athens, and GR: 2000.

[3] K. Anup, S. Ghosh. "A Study in Using Neural Networks for Anomaly and Misuse Detection" Proc. of the 8th USENIX Security Symposium, USENIX press, Washington, D.C., 1999, 141-151.

[4] Intrusion Detection: Challenges and myths by Marcus J. Ranum.

[5] Denning D (Feb 1987) "An Intrusion-Detection Model." IEEE Transactions on Software Engineering, Vol. SE-13, No 2.

[6] Sufyan T. Faraj Al-Janabi and Hadeel Amjed Saeed "A Neural Network Based Anomaly Intrusion Detection System", 2011 Developments in E-systems engineering

[7] Joseph S. Sherif & Tommy G. Dearmond, "Intrusion Detection: Systems and Models".

[8] Yao Yu,Yang Wei & Gao Fu-xiang "Anomaly Intrusion Detection Approach Using Hybrid MLP/CNN neural Network" 2006

[9] Vasilis A. Sotiris, Peter W. Tse, and Michael G. Pecht, Fellow "Anomaly Detection through a Bayesian Support Vector Machine" IEEE Transactions On Reliability, Vol. 59, No. 2, June 2010.

[10] Carolina Fortuna, Blaž Fortuna & Mihael "Anomaly Detection In Computer Networks Using Linear Svms" 2007.

[11] Laheeb Mohammad Ibrahim, "Anomaly Network Intrusion Detection System based on Distributed Time-Delay Neural Network (DTDNN)", Journal of Engineering Science and Technology Vol. 5, No. 4 (2010) 457 – 471.

[12] J. Arokia Renjit and K.L. Shunmuganathan, "Network based anomaly intrusion detection system using SVM", Indian Journal of Science and Technology, Vol. 4 No. 9 (Sep 2011) ISSN: 0974- 6846.

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

[14] Muna Mhammad T. Jawhar & Monica Mehrotra, "Design Network Intrusion Detection System using Hybrid Fuzzy-Neural Network".

[15] Kulbhushan & Jagpreet Singh "Fuzzy Logic based Intrusion Detection System against Blackhole Attack on AODV in MANET", IJCA Special Issue on "Network Security and Cryptography" NSC, 2011.

[16] Tanveer Fatema Khan, Zuber Farooqui, Vineet Richhariya, "Identification of Intrusions in Network for Large Data Base using Soft Computing Approach", IJCST Vol. 3, Iss ue 1, Jan. - March 2012.

[17] Susan M. Bridges & Rayford B. Vaughn, "Intrusion Detection via Fuzzy Data Mining", Accepted for Presentation at The Twelfth Annual Canadian Information Technology Security Symposium June 19-23, 2000, The Ottawa Congress Centre.

[18] Donald F. Specht, "**A** General Regression Neural Network", IEEE Transactions on Neural Networks. Vol. ***2.*** No. 6. November 1991**.**

[19] Jorge Silva & Rebecca Willett, "Hypergraph-Based Anomaly Detection of High-Dimensional Co-Occurrences", EEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 31,no. 3, March 2009.

[20] Kleber Vieira, Alexandre Schulter, Carlos Becker Westphall, and Carla Merkle Westphall, Federal University of Santa Catarina, Brazil, "Intrusion Detection for grid & cloud computing.

[21] Rung-Ching Chen and Su-Ping Chen, "Intrusion Detection using a Hybrid Support Vector Machine based on Entropy and TF-IDF", International Journal of Innovative Computing, Information and Control ICIC International 2008, ISSN 1349-4198 Volume 4, Number 2, February 2008.