International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 11, November-2014

A Review Intrusion Detection System using KDD'99 Dataset

Preeti Singh
M TECH (Student)
Vits satna (RGPV Bhopal)
M. P india

Amrish Tiwari
Asst. Prof.(Vits satna RGPV Bhopal)
M. P india

Abstract: **Development of internet or mobile technology number of users increases day by day which also makes ease of communication. But the use of lots of internet network compromise with rigorous type of security threats or intrusion which persuade the performance of the system. Due to the lack of satisfaction in intrusion detection is makes possible to develop such system which can detect the intrusion efficiently. Many author or researchers has developed the system for intrusion detection using KDD'99 dataset. In this paper we discuss about the various approached/method developed by different researchers for the detection of intrusion. We also present the shortcomings of these methodologies.**

Keywords: - Internet, Intrusion detection, KDD'99, Threats.

I. INTRODUCTION

Now a days the use of internet and multimedia technology is growing rapidly, it facilitate us for the purpose of communication or sharing the resources but use of these technology our network suffered from the network security or information security. The security can be break by injecting the intrusion or threats in the network of the system. The security of these now becomes a very challenging task and for detection of the intrusion or threats various organization has started their work. With the use of the effective intrusion detection system we can protect or network or information. Intrusion detection is the ability of detection unsuitable, erroneous, or anomalous activity. It is the method of monitoring and analyzing the incident occurring in a computer system in order to perceive signs of security problems. Intrusion detection is an imperative component of infrastructure protection mechanism. An intrusion detection system is the most requisite part of the security infrastructure for the network linked to the internet because of the numerous ways to conciliating the stability and security of the network. IDS can be used to scrutinize computer or network for unauthorized activities. Predominantly, network based IDS scrutinize the network traffic coming into the network to detect, identify and track the intruders [1]. An intrusion detection system is categorized into two kinds: network based or host based sytem.

The network based attacks may be either misuse or anomaly based attacks. The network based attacks [2] are detected from the interconnection of computer systems. Since the system communicates with each other, the attack is sent from one computer system to another computer system by the way of routers and switches. The misuse or signature based intrusion detection system detects the intrusion by comparing with its existing signatures in the database. If the detecting attacks and signatures match, it is an intrusion. The signature based intrusions are called known attacks whenever the users are detecting the intrusion by matching with the signatures log files. The log file contains the list of known attacks detected from the computer system or networks. The anomaly based intrusion detection is called as unknown attacks and this attack is observed from network as it deviates from the normal attacks. The host based attacks [2] are detected only from a single computer system and is easy to prevent the attacks. These attacks mainly occur from some external devices which are connected. The web based attacks are possible when systems are connected over the internet and the attacks can be spread into different systems through the email, chatting, downloading the material etc. Nowadays many computer systems are affected from web based dangerous attacks. To ensure performance for intrusion detection system, we can evaluate it basically using KDD'99 intrusion detection datasets [2]. In KDD99 dataset attacks are separated into four classes (DoS, U2R, R2L, and probe) are divided into 22 different attack classes that are tabulated in Table 1.

Table 1 Diverse types of attack in KDD'99 Dataset

4 Main Attack Classes	22 Attack Classes
Denial of Service (DoS)	back, land, neptune, pod, smurt, teardrop
Remote to User (R2L)	ftp_write, guess_passwd, imap, multihop, phf,spy, warezclient, warezmaster
User to Root (U2R)	buffer_overflow, perl, loadmodule, rootkit
Probing(Information Gathering)	ipsweep, nmap, portsweep, satan

The KDD 1999 datasets are alienated into two parts: the training dataset and the testing dataset. The testing dataset includes not only known attacks from the training data but also unknown attacks. Since 1999, KDD'99 has been the most passionately used data set for the assessment of anomaly detection methods. This data set is prepared by [3] and is built based on the data captured in DARPA'98 IDS assessment program [4]. The DARPA 1998 is about 4 GB of squashed raw (binary) TCP abandon data of 7 weeks of network traffic which can be developed into about 5 million association records each with about 100 bytes. For each TCP/IP connection, 41 diverse quantitative (continuous data type) and qualitative (discrete data type) features were extracted between the 41 features, 34 features (numeric) and

7 features (symbolic). The description about the KDD'99 dataset is described in appendix [5]. The process for detection of intrusions or threats is shown through figure 1.

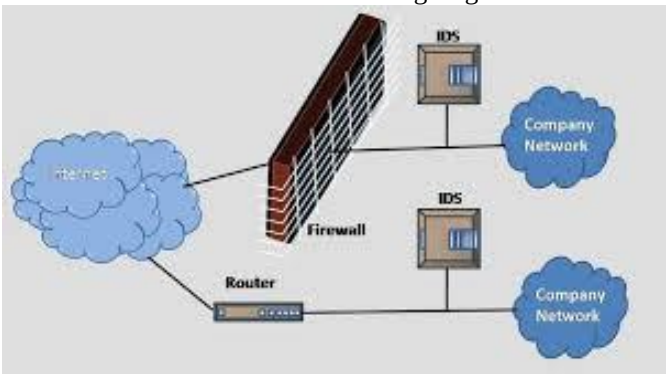

Fig.1 process for Intrusion detection system

The organization of the rest section of the research paper is in this manner: In section II literature about the previous work done in the field of intrusion detection. Section III describes about some techniques to prevent the network or information from the intruders and last section gives conclusion of the paper.

II. RELATED WORK

The intrusion or intimidation cracks the security or hack our private information so to thwart from such issues a multiplicity of techniques and methodologies have been proposed by different researchers. In this paper literature of the work done is discussed below:

Heba F. Eid, Ashraf Darwish, Aboul Ella Hassanien and Ajith Abraham [6] efficiently introduced intrusion detection system by using Principal Component Analysis (PCA) with Support Vector Machines (SVMs) as method to choose the optimum feature subset. They substantiate the efficiencies and the practicability of the proposed IDS system by abundant experiments on NSL-KDD dataset. The reduction method has been used to trim down the number of features in order to diminish the complication of the system. The experimental results show that the proposed system is proficient to speed up the process of intrusion detection and to minimize the memory space and CPU time rate.

Pratibha Soni, Prabhakar Sharma [7] proposed a method which uses two methods C5.0 and artificial neural network (ANN) are utilized with feature selection. Feature selection methods will dispose of some inappropriate features while C5.0 and ANN acts as a classifier to categorize the data in either normal type or one of the five types of attack.KDD99 data set is used to train and test the models ,C5.0 model with numbers of features is producing improved results with all most 100% accurateness.

Jiankun Hu [8] introduced a new host-based anomaly intrusion detection methodology using discontiguous system call patterns, in an endeavor to increase detection rates whilst plummeting false alarm rates. The main idea is to apply a semantic structure to kernel level system calls in order to replicate inherent activities hidden in high-level programming languages which can help comprehend program anomaly behavior. Outstanding results were demonstrated using a multiplicity of decision engines evaluating the KDD98 and UNM data sets and a new, modern data set. The ADFA Linux data set was created as part of this research using a recent operating system and contemporary hacking methods and is now openly available. Additionally, the new semantic method possesses an inherent flexibility to mimicry attacks and demonstrated a high level of portability between dissimilar operating system versions.

Joong-Hee Lee, Jong-Hyouk Lee, Seon-Gyoung Sohn Jong-Ho Ryu and Tai-Myoung Chung [9] instigated decision tree method for detection of intrusion. In intrusion detection systems (IDSs) the data mining methods are useful to notice the attack particularly in anomaly detection. Intended for the decision tree, we employ the DARPA 98 Lincoln Laboratory assessment Data Set (DARPA Set) as the training dataset and the testing data set. The KDD' 99 Intrusion Detection data set is also based on the DARPA set. These three units are comprehensively used in IDSs. Consequently, they demonstrated the total process to engender the decision tree learned from the DARPA Sets. In this paper also guesstimate the efficient value of the decision tree as the data mining method for the IDSs and the DARPA set as the learning dataset for the decision trees.

A. M. Chandrasekhar, K. Raghuveer [10] proposed a method which is divided into four steps: initial step, k-means clustering is used to generate different training subset then based on the obtained subset, various neuro-fuzzy data model are trained. Consequently, a vector for SVM classification is obtained and in last, classification using radial SVM is applied to detect the intrusion occurred or not. To make obvious the applicability and ability of the new method, the result of KDD dataset is confirmed in which it shows that the proposed methods produce better result than the BP, multiclass SVM and other approach such as decision tree etc.

Preecha Somwang, Woraphon Lilakiatsakun [11] proposed the clustering method by using hybrid method based on Principal Component Analysis (PCA) and Fuzzy Adaptive Resonance Theory (FART) for identifying diverse attacks. The PCA is apprehensive to random selects the best provenance and reduction the feature space. The FART is implementing which is used to classifying dissimilarity in collection of data, regular and irregular. The proposed method can enhances the high performance of the detection rate and to reduce the false alarm rate and this is computed approach on the benchmark data from KDD Cup 99 data set.

Fan Li [12] proposed an Intrusion Detection system (IDS) based Hybrid Evolutionary Neural Network (HENN). In order to construct a precise model for normal behaviors and achieve better detection performance. The genetic algorithm is employed to evolve input features, network structure and connection weights. The experimental results show that the proposed method accomplishes feature selection and structure optimization effectively. Through the comparative analysis, it can be seen that the HENN achieves better detection performance in terms of detection rate and false positive rate.

Singh, Ritu Ranjani, Neetesh Gupta, and Shiv Kumar [13] proposed here a using Self Organizing Map to diminish alarm in IDS and demonstrated as an intrusion detection systems intend to distinguish attacks with a high detection rate and a low false alarm rate. The classification-based data mining models for intrusion detection are often unsuccessful

in dealing with self-motivated changes in intrusion patterns and features. Accordingly, unsupervised learning methods have been given a closer look for network intrusion detection. Conventional instance-based learning methods can only be used to identify known intrusions since these methods categorize instances based on what they have learned. They rarely detect new intrusions since these intrusion classes has not been able to detect new intrusions as well as known intrusions. Author proposed a soft Computing technique such as Self organizing map for detecting the intrusion in network intrusion detection. Problems with k-mean clustering are hard cluster to class assignment, class dominance, and null class problems.

III. INTRUSION DETECTION TECHNIQUES

In this section various techniques for the detection of intrusion is described such as network based intrusion system and host based intrusion system: Some of these techniques are described below

1. Neural Network approach

Increasing amount of research is going on Artificial Neural Network (ANN) [15], [16]. ANN consists of base units called neurons, which are grouped, in several levels. Neurons are connected to neighbor neurons and those connections are weighed. An ANN has input level, one or several hidden layers, and output level. Neural Networks architecture can be distinguished as follow:

- Supervised training algorithm [14], [15]: The neural network learns the preferred output for a given input or prototype in the learning phase. Ex. Multi-Level Perception (MLP): the MLP is employed for Pattern Recognition difficulties.
- Unsupervised training algorithm [14], [15]: The network learns without specifying desired output in the learning phase. Ex. Self-Organizing Maps (SOM)

It finds a topological mapping from the input space to clusters. Generally used for classification problems. For IDS using ANN approach has two phases: i) Training and ii) Testing

i. Training: To recognize various normal and abnormal traffic behavior one has to train the network. In the research it is done by using a dataset. The KDD99 dataset is publically available and it is mostly used for evaluating IDS.

ii. Testing: It is similar to the training. After training NN IDS tested using a test dataset. This dataset is smaller than the training dataset to ensure that the network can detect intrusions it was trained to detect.

Advantages:

It has more potential to identify and classify

Disadvantage:

It splits the dataset based on completely the attribute values

2. Honey pot deception

A honeypot is ambushes set to sense, deflects or in some manner thwart attempts at unauthorized use of information systems. It consists of a computer data or a network site that appears to be part of a network, but is truly isolated and monitored and which seems to restrain information or a resource of value to attackers. Honeypot can be classified based on their exploitation and based on their level of involvement. Based on deployment, honeypot may be categorized as: Production Honey Pot (PHP) and Research Honey Pot (RHP).

PHP are simple to use, incarcerate only limited information and are used principally by companies or corporations. These are placed within the production network with other production servers in an organization to advance their overall state of security. Generally, PHP are low-interaction honey-pot which are easier to organize. They give few information concerning the attacks or attackers than research honey-pot do.

The opinion of a PHP is to help diminish risk in an organization. The honey-pot inserts extra value to the security measures of an association. RHP is setup by a volunteer, non-profit research association or an educational institution to congregate information about the intentions and policies of the Black hat community targeting diverse networks. These honey-pot do not add on direct value to a definite organization instead they are used to research the threats of an organizations facade and to learn how to better defend against those threats. RHP is difficult to systematize and sustain, imprison widespread information and used principally by research, military, or government association.

Advantage:

- Fewer false positives since no legitimate traffic uses honeypot
- Collect smaller, higher-value, datasets since they only log illegitimate activity
- Work in encrypted environments
- Do not require known attack signatures

Disadvantage:

- Can be used by attacker to attack other systems
- Only monitor interactions made directly with the honeypot - the honeypot cannot detect attacks against other systems
- Can potentially be detected by the attacker

3. SVM Classifier

SVM is developed on the principle of structural risk minimization. It is one of the learning machines that map the training patterns into the high-dimensional feature space through some nonlinear mapping. SVM has been successively applied to many applications in the multiclass classification [18-19]. By computing the hyper plain of a given set of training samples, a support vector machine builds up a mechanism to predict which category a new sample falls into (Figure 1).

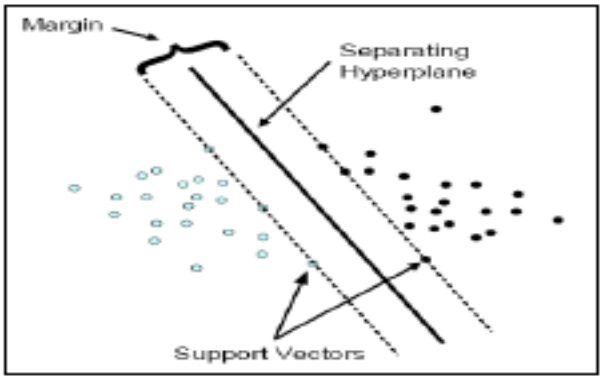

Fig.2 Separating Hyperplane with SVM

In an SVM, a data point is viewed as a vector in the d-dimensional feature space. Assume that all data points belong to either class A or class B. Each training data point can be labeled by based on (1):

$$Y_i = \begin{cases} -1 & x_i \in class\ A \\ 1 & x_i \in class\ B \end{cases} \quad (1)$$

Therefore as it is revealed in Fig. 2, the training data set can be designated as:

$$D = \{x_i y_i | 1,2,3 \dots \dots\} \quad (2)$$

Data points with label 1 and −1 are referred to as positive and negative points, respectively. In the linear separable case, there are many hyper-planes which might separate the positive from the negative points. The algorithm merely looks for the major margin separating hyper-plane where the ‘‘margin’’ of a separating hyper-plane is defined to be the sum of the distances from the hyper-plane to the closest positive and negative points. In order to calculate the margin of a separating hyper-plane H, consider the hyper-planes H1 and H2 that include the closest positive training points and the closest negative training points to H, respectively:

$$\begin{aligned} &H: w.x - b = 0, \quad x \ R^d \\ &H_1: w.x - b = 1 \quad x \ R^d \\ &H_2: w.x - b = -1 \quad x \ R^d \end{aligned} \quad (3)$$

Where w is the normal to H and b is the distance from H to the origin. Obviously, H2, H1, and are parallel. In addition,

$$\begin{aligned} &w.x_i - b \geq 1 \text{ for } y_i = 1 \\ &w.x_i - b \leq -1 \text{ for } y_i = 1 \end{aligned} \quad (4)$$

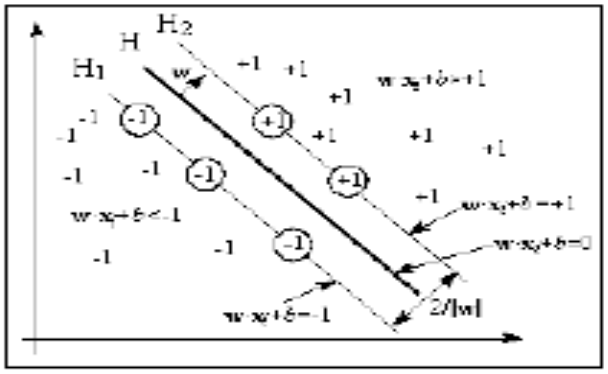

Fig.3 Data Points and Their Classes

In the case of the data points are linearly separable, above method can classify points by a linear hyperplane. However, data points are non-linearly discrete, other approaches ore used for data categorization such as the use of a kernel function to generate a non-linear decision boundary. A kernel function takes a data set and convert it into a higher dimension through the use of some function (general ones comprise (RBF) radial basis functions, Gaussian functions, and sigmoidal functions).

Advantages:

- diminish the wrong alarm rate
- produce an accurate detection model from a mass of

Disadvantages

- It takes extensive training time
- difficult to train system for dynamic changing environment

4. Hidden MARKOV Models

The Hidden Markov Model (HMM) is a twofold stochastic model [20]. The model is represented by λ (A, B, π), where A is the set of observables, B is the set of hidden situations, and π is the set of transition probabilities i.e., the probabilities from going to one hidden state to a different. This model is known as twofold stochastic since there is a hidden layer that contains some hidden states. This hidden layer follows the principles of Markov process. The other layer contains the states of the observables in a particular time t of the model construction. This is also a Markov process where the observable outputs can be seen, unlike the hidden layer. The HMM algorithm works in two steps.

The HMM is trained in the initial step using the training sequences. At the initial state (at time t0), the state transition probabilities and the discernible output probabilities are randomly assigned. However, assigning these probabilities according to prior knowledge of the system, instead of the random assignment, can improve the performance of HMM. At this point, the model is denoted with λ_0. Then, applying the Baum-Welch algorithm, the HMM λ_0 is adjusted according to the input training sequences and construct the new model λ_1 [20]. After every adjustment of λ, the probability difference of the previous model and the adjusted model is calculated. If the difference is below the preset probability difference threshold, the model is known to be the final HMM. Otherwise, further adjustment is required. In the

next step, the unknown sequences are applied to the model and the likelihood of the sequences (i.e., the probability of how much a sequence conforms the HMM) are determined. If the probability is above the predefined acceptable probability, the sequence is concluded as a non-anomalous sequence. Otherwise, it is concluded as an anomalous one. The HMM algorithm has very accurate prediction of anomaly and has been used for complex sequence analysis. However, the model training time is very high in HMM algorithm.

5. Machine Learning

The machine learning method is to distinguish outliers in datasets from a diversity of fields were designed by Gardener employ a One-Class Support Vector Machine to notice anomalies in EEG data from epilepsy patients) and Barbara (proposed an algorithm to recognize outliers in raucous datasets where no information is accessible regarding ground truth based on a transductive confidence machine [21]. Dissimilar induction that uses all data points to persuade a model, transduction, and an alternative uses small subset of them to approximate unknown attributes of test points. To implement online anomaly detection on time series data in [22] Ma and Perkins presented an algorithm using support vector regression. They offered an adaptive anomaly detection algorithm that is based on a Markov-modulated Poisson progression method and use Markov Chain Monte Carlo approaches in a Bayesian approach to learn the model parameters [23].

Advantages:

- Shortening the effective geographical distances and sharing information

Disadvantages:

- It cannot bitterly reduce false positives, active platform or event based classification

6. Boosted Decision Tree

Boosted Tree (BT), that uses ADA Boost algorithm [24] to generate many Decision Trees classifiers trained by different sample sets drawn from the original training set, is implemented in many IDS successfully. Every supposition fashioned from each of these classifiers is united to calculate total learning error, thus received at a concluding fused hypothesis.

Advantages:

- It gives better visibility of behavior of each

Disadvantages

- It is difficult to train system for dynamic

7. User Intention

Building the profile of normal behavior and attempting to identify certain pattern or activity deviations from normal profile. Anomaly detection is used to find mysterious attacks by using the perception of profiling normal behaviors. However, noteworthy false alarm may be caused because it is complicated to achieve complete normal behaviors. Intrusion detection can be designed upon multiple levels in a authentic computer network system. It will be a choosing the features that characterize the user or the system usage patterns in the best way such that distinguishing abnormal activities from normal activities is done clearly. Data sources like unix shell commands, audit events, keystroke, system calls and net work packages can be used. The first crucial step in building a profiling method for intrusion detection is selecting a data source. Throughout the early on reconsideration on anomaly detection, the key hub was on profiling system or user behaviours from monitored system log or accounting log data.

Advantages:

- Lowered false positive rates
- Increase the accuracy of network *IDS*

Disadvantages:

- Defeating much of the intent of the monitor

IV. CONCLUSION

To make our information or network protected, it is the key problem to detect the intruders while various authors designed the system for the detection of intrusion and also proposed the method such as neural network, decision tree, hidden markov model etc. which has some advantages and disadvantages. In this study paper we described few methods with its advantages and disadvantages. Most of the method is less efficient to reduce the false alarm rate and takes extensive training time so in future design such system which can efficiently reduce the false alarm rate and decreases the training time.

REFERENCE

[1]. David Wanger and Paolo, " Mimicry Attack on Host Based Intrusion detection System" Proceeding of the 9th ACM conference on Computer and communication security, pp.255-264,2002

[2]. S. Devaraju and S. Ramakrishnan " Performance Comparison For Intrusion Detection System Using Neural Network With KDD Dataset" ICTACT Journal On Soft Computing, April 2014, Volume: 04, Issue: 03 743 ISSN: 2229-6956(Online)

[3]. Chaturvedi, Anshul, and Vineet Richharia. "A Novel Method for Intrusion Detection Based on SARSA and Radial Bias Feed Forward Network (RBFFN)." International Journal of Computers & Technology 7, no. 3 (2013): 646-653.

[4]. Mohammad Behdad, Luigi Barone, Mohammed Bennamoun and Tim French "Nature-Inspired Techniques in the Context of Fraud Detection" in IEEE transactions on systems, man, and cybernetics—part c: applications and reviews, vol. 42, no. 6, November 2012.

[5]. Rashmi Singh and Diwakar Singh "A Review of Network Intrusion Detection System Based on KDD Dataset" Int J Engg Techsci Vol 5(1) 2014, 10 – 15

[6]. Heba F. Eid, Ashraf Darwish, Aboul Ella Hassanien and Ajith Abraham Principle Components Analysis and Support Vector Machine based Intrusion Detection System", in proceeding of IEEE (2010)

[7]. Pratibha Soni, Prabhakar Sharma "An Intrusion Detection System Based on KDD-99 Data using Data Mining Techniques and Feature Selection", International

Journal of Soft Computing and Engineering (IJSCE), ISSN: 2231-2307, Volume-4 Issue-3, July 2014

[8]. Jiankun Hu, "A Semantic Approach to Host-Based Intrusion Detection Systems Using Contiguous and Discontiguous System Call Patterns", IEEE Transactions on Computers, vol.63, no. 4, pp. 807-819, April 2014, doi:10.1109/TC.2013.13

[9]. Joong-Hee Lee, Jong-Hyouk Lee, Seon-Gyoung Sohn Jong-Ho Ryu and Tai-Myoung Chung "Effective Value of Decision Tree with KDD 99 Intrusion Detection Datasets for Intrusion Detection System ",. Advanced Communication Technology, 2008. ICACT 2008. 10th International Conference on (Volume:2)Feb. 2008 Page(s):1170 - 1175 ISSN :1738-9445

[10]. A. M. Chandrasekhar, K. Raghuveer " Intrusion Detection Techniques by using K-means, Fuzzy Neural network and SVM classifier", ICCCI-2013, Jan. 04-06, 2013, Coimbatore, INDIA

[11]. Preecha Somwang, Woraphon Lilakiatsakun: "intrusion detection technique by using fuzzy ART on computer network security" 2011, 978-1-4577-2119, in IEEE.

[12]. Fan Li ―Hybrid Neural Network Intrusion Detection System using Genetic Algorithm‖ in IEEE 2010.

[13]. Singh, Ritu Ranjani, Neetesh Gupta, and Shiv Kumar. "To reduce the false alarm in intrusion detection system using self-organizing map." International Journal of Soft Computing and Engineering (IJSCE) ISSN (2011): 2231-2307.

[14]. Jean-Philippe "Application of Neural Networks to Intrusion Detection", SANS Institute Reading Room site.

[15]. Deepika P Vinchurkar, Alpa Reshamwala " A Review of Intrusion Detection System Using Neural Network and Machine Learning Technique", International Journal of Engineering Science and Innovative Technology (IJESIT), Volume 1, Issue 2, November 2012

[16]. Shahbaz Pervez, Iftikhar Ahmad, Adeel Akram, Sami Ullah Swati, "A Comparative Analysis of Artificial Neural Network Technologies in Intrusion Detection Systems", Proceedings of the 6th WSEAS International Conference on Multimedia, Internet Video Technologies, Lisbon, Portugal, September 22-24, 2006.

[17]. Dr.K.V.Kulhalli, S.R.Khot "Network Based Intrusion Detection Using Honey pot Deception" International Journal on Recent and Innovation Trends in Computing and Communication Volume: 2 Issue: 4 ISSN: 2321-8169 pp-805 – 810

[18]. Vahid Golmah "An Efficient Hybrid Intrusion Detection System based on C5.0 and SVM" International Journal of Database Theory and Application Vol.7, No.2 (2014), pp.59-70

[19]. S.-W. Lin, K.-C. Ying, C.-Y. Lee and Z.-J. Lee, "An intelligent algorithm with feature selection and decision rules applied to anomaly intrusion detection", Applied Soft Computing, vol. 12, (2012), pp. 3285-3290.

[20]. Afroza Sultana and Abdelwahab Hamou-Lhadj, Mario Couture "An Improved Hidden Markov Model for Anomaly Patterns Detection Using Frequent Common" ICC-2012

[21]. D. Barbar´a, C. Domeniconi and J. Rogers, "Detecting outliers using transduction and statistical testing" ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), Philadelphia, PA, Aug. 2003.

[22]. J. Ma and S. Perkins, "Online novelty detection on temporal sequences" ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), Washington, DC, Aug. 2003.

[23]. A. Ihler, J. Hutchins, and P. Smyth, "Adaptive event detection with time-varying Poisson processes" ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD), Philadelphia, PA, Aug. 2006.

[24]. Y. Freund, Schapire.R. "Experiments with a new boosting algorithm" Thirteenth International Conference on Machine Learning, Italy, 1996.