International Journal of Database Theory and Application
Vol.9, No.12 (2016), pp.231-240
http://dx.doi.org/10.14257/ijdta.2016.9.12.21

Hybrid Intrusion Detection Method to Increase Anomaly Detection by Using Data Mining Techniques

Bilal Ahmad[1], Wang Jian[1], Bilal Hassan[1] and Sara Rehmatullah[2]

[1]Department of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China
Department of Computer Science, Sir Syed University of Engineering & Technology, Karachi, Pakistan
ahmad@nuaa.edu.cn, bilalhassan@nuaa.edu.cn

Abstract

An Intrusion Detection System is an application which observes movements or action happen on the network and determine it for any kind of harmful activity that can disturb computer security policy. With progress of increase the usage rate of the internet, there is a widely increase in the number of internet attacks as well, so contests arise towards the network security due to the arrival of new approaches of attacks. To classify these attacks, a new hybrid method with the help of data mining based on decision tree C4.5 and Meta algorithm is planned. This method gives a classifier which expands the whole accuracy of detection. Many data mining techniques have been settled for detecting intrusion. For recognition of anomalies a hybrid technique based on decision tree C4.5 with Meta algorithm is offered that provides better accuracy and reduces the problem of high false alarm ratio. The assessment of the given approach is made with other data mining techniques. With this given approach detection rate is improved significantly. KDD Cup 1999 dataset use for experimental work.

***Keywords**: Hybrid Intrusion Detection System, Data Mining, Decision Tree, Meta Algorithm*

1. Introduction

From the last decade, people become very much depends upon technology. In recent days we use internet to receive emails/banking/stock price/news and e-shopping. Highly use of the networks leads to violence. Due to that it educates the need of safe and nonviolent system. Due to dependence on information technology, it needs to meaningfully advance network security, so that data reliability, privacy and availability never obstruct. Every computer systems are vulnerable to negotiation and every network is at risk to unlawful contact and outflow of remote and sensitive information. A firewall is typically and keenly used in security appliances. It organizes security rule, but that has also been in unsuccessful, because it cannot stop from all kinds of malicious intent of the intruder or attacker. In case of a firewall, only its header content is examined whereas in "Intrusion Detection System" (IDS) both content and header of packets are examined. So, IDS is much more dynamic as compared to firewall in order to secure our private and complex data. IDS has confirmed an important tool for security, but firewall cannot interchange totally with intrusion detection system as there work is to make balance with each other by examining all kinds of misuse patterns on networks. An intrusion is known as any kind of act that compromises reliability, privacy or Availability. Although it plays a very important role to define and protect in security design, but IDS is still not mature and not considered as complete protection. IDS categorizes any kind of intrusion and warn right in the form of alert, so that assets can be safe. An IDS is also used in lawful proceedings as scientific evidence against the intruder because it delivers logs of any kind

ISSN: 2005-4270 IJDTA

of intrusion involved in cybercrime. An IDS is prepared to protect illegal access to resources or data. It can be hardware or software. Intrusion detection system is to protect single and all computer network. IDS provides user friendly interface to non-expert staff for managing the systems simply.

2. Literature Survey

In this section a literature survey of many models and techniques used to detect the intrusion. How IDS developed and various kinds of changes take place in existing and new models. The IDS notion has been presented in 1980 by James “Anderson's seminal” study, “Computer Security Threat Monitoring and Surveillance”. This concept has been approximately for 20 years, but due to the rise of the security structure leads to a dramatic popularity and progress in IDS. The idea of IDS was first came from the technical report from Anderson (1980). He was proposed the computer inspection system which should me transformed and able to provide risk and threats for computer security techniques. This idea should provide statistical methods which can apply on user behavior and detect intruders who can access the system illegally. In 1987, Dorothy suggested a prototype for intrusion detection Dorothy E. Denning and Peter Neumann (1987) were early pioneers in the Intrusion Detection arena. They had provided the structure for an intrusion detection expert system, which was called IDES (Intrusion Detection Expert System) [1] based off of the 1985 paper “Requirements and model for IDES”, the real time IDS [2]. Hoge and Austin (2004) provide “survey of anomaly detection using machine learning and statistical methods”. They introduce a survey of contemporary techniques for outlier detection. Markou and Singh [3] also presented extensive reviews for intrusion detection using ANN and statistical methods. Many books and article also written based on Outliers and intrusion detection (Douglas M. Hawkins 1980, V. Barnett, and T. Lewis 1994, Z. Bakar, R. Mohemad, A. Ahmad, M. Deris [4,5,6]. Various anomaly detection system such, as NIDES (Next generation Intrusion Detection Expert System) [7] ALAD (Application Layer Anomaly Detector) [8], PHAD (Packet Header Anomaly Detector) generate statistical model for normal network traffic and alarm generates if some deviation found in normal model. Most of then use feature extraction from network packet header. For example NIDES and ALAD use foundation, destination IP, port address and TCP connection state. Zhang, Yang, and Geng (2009) [9] presented review of network anomaly detection methods and techniques. Wu and Banzhaf (2010) [10] the area of this review include “artificial neural networks, fuzzy systems, evolutionary computation, artificial immune systems, swarm intelligence, and soft computing”. Dong, Hsu, Rajput (2010) [11] presented the method which is according to them is more authentic as compared to Markov and K. means. “Graph based Sequence Learning Algorithm” (GSLA) contains “data pre-processing, normal profile construction” and session marking”. In this approach, the standard profile is made over a “session learning” method, which can be used to know an anomaly session. Warusia and Udzir (2014) [12] purpose novel “Signature Based Anomaly Detection Scheme” (SADS) which applied to learn packet headers performance patterns more accurately is proposed. Mixing data mining algorithms like Naive Bayes and Random Forest can be used to make low false alarms and reducing processing time. Some researchers also use feature selection techniques for intrusion detection. Harbola and Jyoti (2014) [13] also use feature selection techniques to improve accuracy. The key objective of this study is to bring the wide analysis “Feature Selection” approach for NSL-KDD intrusion detection dataset. [14] P.G. Majeed and S. Kumar made implementation of genetic algorithms using pseudo code. This study provides an overview of the advantages and disadvantages of genetic algorithms in general, and as applied to intrusion detection in particular Anna L. Buczak [19] describe a survey with some good knowledge dataset used in ML/DM are designated. They describe complexity of Machine Learning/Data Mining algorithms with the addressing and discussion the challenges for ML/DM with the usage of internet security in recent days.

3. Intrusion Detection System

An IDS is a software or hardware that observers event happened at network and examining it to detect any kind of action that violate computer security rules. IDS, [15] [16] analyze the information collected from both user and system actions, examining formations of system and estimating the file and system reliability, identify irregular pattern and alert to system administrator if any kind of suspicious activity performed

3.1 Type of Alerts

IDS informs the system admin by alarm alerts if any unlawful access to the systems or network. According to IDS these alerts can be categories into IV groups: True Positive (TP) known as real intrusion which IDS alert the admin. False Positive (FP) means no intrusion but IDS made an alert. False Negative (FN) means real intrusion, but IDS not make any alert for admin. True Negative (TN) means no intrusion but IDS made alert for admin.

3.2. Type of IDS

I. Network Intrusion Detection System (NIDS)

This kind of IDS install with in the network and monitor network traffic from some specific network area, device and search protocol action to classify various types of suspicious activities

II. Host Intrusion Detection System (HIDS)

HIDS are used for single host and monitor the events going on within this host and classify malicious activities. HIDS monitor log files, running processes & applications, files access or modification, system application formation changes on a single host

3.3. Types of Detection

In Intrusion detection system presently there are many selections of detection methods, but two main approaches are:

I. Misuse/Signature based Detection

This method known as information based detection approach. It contains of some information based methods which database have attack signatures. This technique is especially for detecting known attacks. It makes a less number of “false positive alerts” but the basic restriction of this method is that it can only detect already known attacks which describe in the database and never spots the unknown attacks.

II. Anomaly/Statistical based Detection

The detection engine of this method detects usual and irregular performance of consumer so this method also known as performance based detection approach. Anomaly based detection method capable to spot unknown attacks by the statistical study approach, the disadvantage of this approach is that, It cannot spot the famous attacks. This is because of the fact that it generates a large number of false positive alerts.

4. Anomaly/Statistical based Detection

In signature based IDS there is a high accuracy of detecting known attack, but in current day's security is the main target in every field and every day a new type of attack is introduced. Signature based IDS are unable to notice that kind intrusion. So anomaly detection is the best technique to detect the new type of attack on the basis of their behavior.

Anomaly detection divides into two phases 1- training, 2- testing. On the basis of that behavior is differentiated. In training phase, we train our dataset about the normal and abnormal traffic profile. After that this training profile is tested on the dataset like KDD Cup 1999, or many more to check the accuracy of the detection approach [17].

Data mining has the facility to detect deviation from normal behavior by creating a boundary value of network activity between normal and abnormal behavior. Data mining technique is categorized into two types:

4.1. Supervised Learning

This organize audit data as a normal or abnormal flow using different classification algorithms. This classification algorithm defines a set of rules and patterns while detecting the intrusion. In this classification technique a particular process is followed: 1- Identify class and classes, attribute from training data 2- Identify attributes for classification 3- Learn about a model/algorithm used to train the data. There are so many classification algorithms are like Decision Tree, Naive Bayes, Random Forest, SVM, ANN, Naive Bayes, etc.

4.2. Unsupervised Learning

This is a clustering technique for finding patterns in many orders from unlabeled data using different clustering algorithms. Clustering has expertise to detect intrusion in the audit data without an explicit description around various attack classes.

5. Objectives

- To propose a new technique to improve the accuracy and reduce the false alarm rate.
- To improve the detection rate of anomalies.
- To validate the new proposed technique on the dataset.

6. KDD Cup 1999 Dataset

KDD Cup 1999 dataset is the famous dataset that is used for assessing the anomaly type intrusion. In 1998 DARPA conducted an evaluation program for intrusion detection in the MIT Lincoln Labs. The main objective of this program was to evaluate the intrusion [18]. In this standard dataset was provided which simulated a wide variety of intrusions according to the perspective of a military network environment. In this dataset, around 5 million connections and each one connection having 100 bytes were records. KDD Cup 1999 having 4, 90,000 single vector connection. This dataset contains 41 features which were labeled as normal or as abnormal. In this dataset attack lies in four categories they are:

I. **Denial of Service Attack (DOS):** It is a malicious attempt in which the attacker makes server or network resource unavailable or too busy to handle requests.
User to Root Attack (U2R): It is a type of exploit in which attacker has a local normal user account access the system. But an attacker takes the advantage of present vulnerabilities in the system like sniffing passwords, and gets the super user privilege access.

II. **Remote to Local Attack (R2L):** is a type of attack in which an attacker machine is able to send a packet remotely but does not have an account on the victim machine. So by taking advantage of any vulnerabilities on the victim machine, the attacker gets access to the victim machine.

III. **Probing Attack:** It is a very basic and initial step of misusing any system. The attacker scans a machine to find out the weakness or vulnerabilities in the network to exploit the victim machine.

Table 1. Attack Types with their Corresponding Type

Type	**Attacks**
DOS	apache, back, land, mailbomb, neptune, pod, processtable, smurf, teardrop, udpstorm
PROBE	ipsweep, mscan, nmap, portsweep, saint, satan
U2R	buffer_overflow, loadmodule, perl, rootkit, ps, sqlattack, xterm
R2L	ftp_write, guess_password, imap, multihop

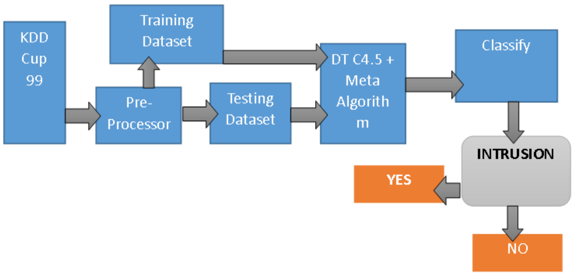

Figure 1. Architecture of Proposed Model

7. Pre-processor

To improve the detection capabilities of classifiers, there is a need to preprocess the data. Preprocessing provides a feature to clean and normalize the data so that any kind of irrelevant data never affects the accuracy of classifiers. To remove the redundant data there are various kinds of preprocessing filters which are available.

8. Proposed Algorithm

The proposed method (figure 1) is built on the Decision Tree (C4.5) algorithm and Meta algorithm. C4.5 algorithm is much accurate making any type of decision, but difficult while any kind of change occur in the dataset effects on the decision making and needs to train the dataset .It leads to variance in the classification of data. So to remove this variance Meta algorithm present with the C4.5 algorithm. In this proposed algorithm it provides a predictive feature using the randomization which recovers the variance while decision tree formation in C4.5 algorithm.

9. Experimental Results

The proposed method is based upon DT and Meta algorithms with a selection of 41 features in a KDD dataset. In this hybrid method, application of the Meta algorithm at the building time of decision tree helps to reduce the variance while decision tree improves the accuracy of anomaly detection. To evaluate the given model it compares with different supervised machine learning models and measure their accuracy, “false alarm rate” and “true positive rate”. The proposed approach compares popular machine learning model, for example, “decision tree, native Bayes, support vector machine (SVM), decision table and random tree". To measure the accuracy of these predictive models, 10-fold cross validation is used.

9. 1. Accuracy:

This mentions the ability of the classifiers to correctly measure the intrusions from the training dataset. This is defined as the ratio of correctly classified data to the total classified data.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

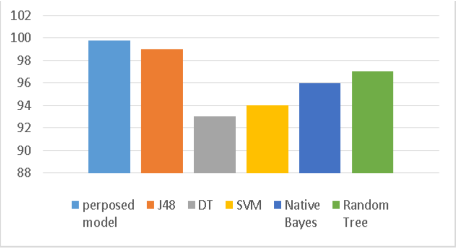

Figure 2. Accuracy

Experiment result shows (figure 2) that the proposed model is more accurate as compare to other data mining techniques. The accuracy of a proposed model is near about 100% as shown in the graph. This proposed method performs well than individual performance of the J48 (C4.5).

9.2. Detection Ratio

$$Detection\ Rate = \frac{TP}{TP+FP}$$

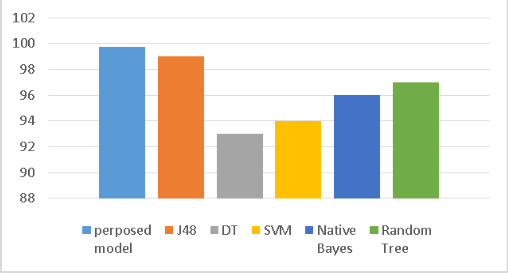

Figure 3. Comparison of Detection Ratio

In above experimentation, the result shows (figure 3) the average performance of 10-Cross validation. To measure the robustness and effectiveness of any model, comparison of different parameters like False Positive Rate, True Positive Rate, F-measure, Precision and Recall is computed and the performance of different models at the above parameters is evaluated.

11.3. True Positive Ratio

This is one in which correct classification of data has been performed. Means correctness in a system to detect normal or abnormal data. It is defined as:

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

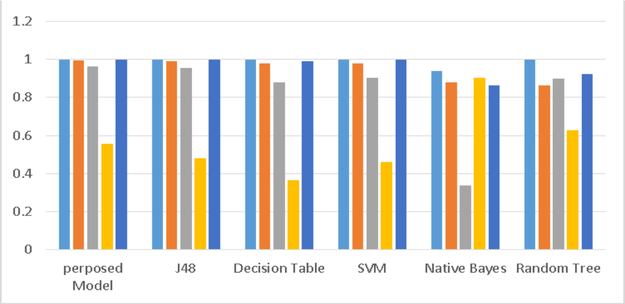

Figure 4. Comparison of True Positive Ratio

9.4. False Positive Ratio

This is one of the main parameters to find out the effectiveness of various models and also the major concern while network setup. A normal data is considered as abnormal or attack type data. It is defined as:

$$FPR = \frac{TP}{TN + FP}$$

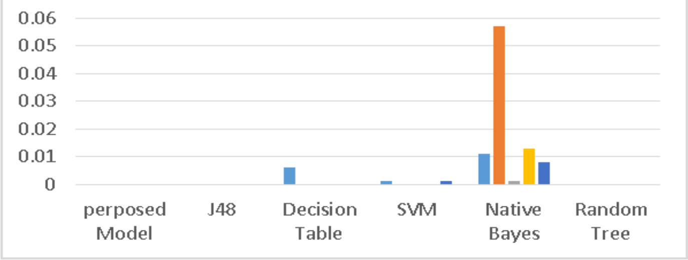

Figure 5. Comparison of False Positive Ratio

As revealed in comparison, both “true positive rate” and “false positive rate” proposed model executes better as compared to other models. These two parameters are very important measure to evaluate the performance of a model. Hence result shows that the proposed model performs better than Decision table, SVM and Naïve Bayes models.

9.5. Precision Ratio

It is also known as Positive Predictive Value (PPV). It measures the relevant instance that is retrieved after classification. A classifier that has high precision means that classifiers or algorithm returns more relevant results.

$$FPR = \frac{TP}{TP + FP}$$

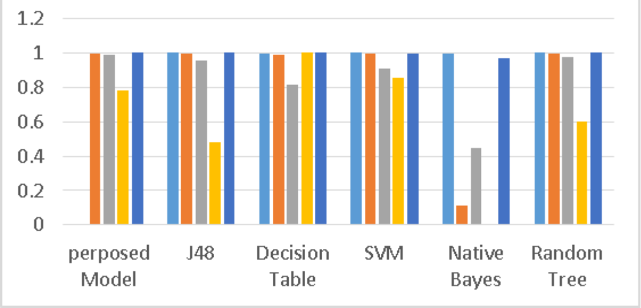

Figure 6. Comparison of Precision Ratio

As shown in the figure (Figure 6), "precision ratio" of the proposed model is high as compared to other models. Proved that proposed model provides the less relevant results.

9.6. Recall

It is also known as sensitivity. This is also used to measure the relevant instance that is selected. The higher value of recall more the relevant data is selected for classification. It is defined as:

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

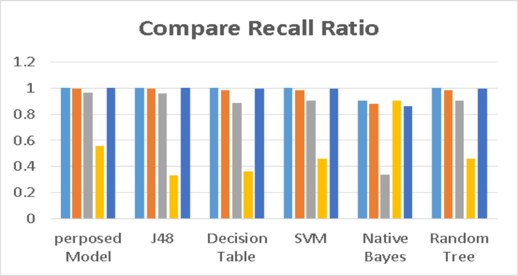

Figure 7. Comparison of Recall Ratio

The above result (figure 7) shows that the proposed model is having a high recall or sensitivity. Hence, the most relevant data is selected as compared to other classifiers.

9.7. F-Measure

It is basically used to measure the effectiveness of the classifiers. This is harmonic mean of precision and recall. It is also known as traditional F-measure or balanced F-score. It is defined as:

$$F - measure = 2\frac{Precision * Recall}{Precision + Recall}$$

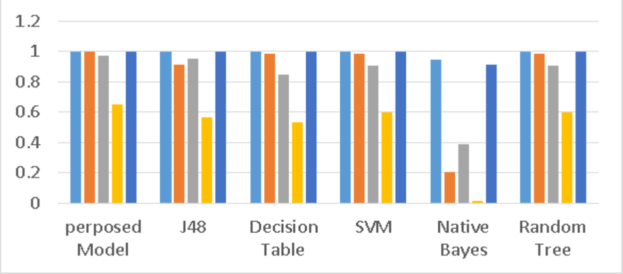

Figure 8. Comparison of F-Measure Ratio

The above result (figure 8) shows that the proposed model shows better results in every fold or round of evaluation. A proposed technique performs much better in all aspects of the evaluation parameter of anomaly detection.

9.8. Area under the ROC Curve

It defined that the accuracy of the classifier, how a normal or abnormal dataset is divided by using training dataset. More the area under the ROC bow the more accurate the classifier is.

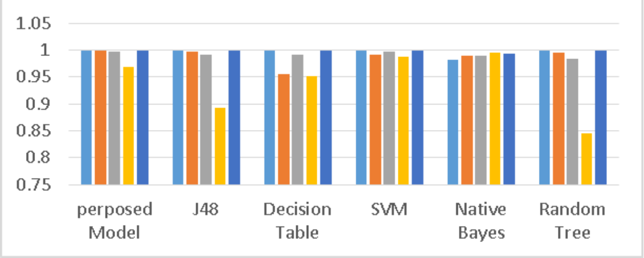

Figure 9. Comparison of Area under ROC

As above result (figure 9) shows that the proposed algorithm covers the maximum area means the maximum accuracy in the result while classification.

9.9 Area Under the PRC

This is a graph between the recall and precision value. In this, it defined the correctness. Area under the PRC means that classifiers are providing more correctness while classification

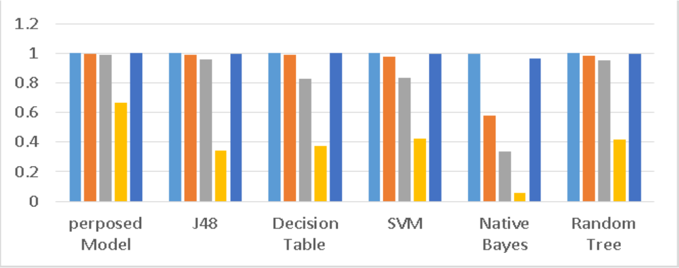

Figure 10. Comparison of Area under PRC

The above result (figure 10) shows that the proposed algorithm covers the maximum area means the maximum correctness in the result while classification.

From all the above experimentation results, it is shown that after applying all the evaluation parameters, proposed model found to be the best model in all scenarios. By applying the hybrid approach of data mining model on the dataset, the detection rate is improved for anomaly detection. So the main objective to improve the detection rate in anomaly detection has been met.

10. Conclusion

To improve the accuracy rate of intrusion detection in an anomaly based detection data mining technique is used. In this research a hybrid approach using data mining is applied, which is combination of two different methods, Decision Trees and Meta algorithm. The results of the proposed approach are compared with the results of other data mining techniques, and it outperforms them. The new approach is effective during detection of attacks. The detection ratio of the proposed algorithm is better than other techniques.

11. Future Scope

In the proposed approach data are classified into normal and abnormal data. The approach can be improved by classifying data further into the subclasses such as DOS, Probe, U2R and R2L. Better results were obtained while performing it on KDD Cup 1999 Dataset. It can also be used for real time traffic analysis for obtaining better results. While achieving the detection accuracy of data traffic in real time anomaly detection, better results can be obtained by combining it with supervised and unsupervised techniques. While analyzing the real time traffic for supervised learning method is a bit complicated due to the data size. To overcome this problem unsupervised learning is used to define the boundaries between normal and abnormal data. So after that when supervised learning model is applied to real time traffic, then it gets easily classified without any considerable delay.

References

[1] E. D. Dorothy, “An Intrusion-Detection Model”, IEEE Transactions on Software Engineering, vol. 13, no. 2, **(1987)**, pp. 222-232.

[2] D. E. Denning, “An Intrusion-Detection Model”, IEEE Transactions on Software Engineering, **(1987)**, pp. 222-232

[3] V. J. Hodge and J. Austin, “A Survey of Outlier Detection Methodologies”, Dept. of Computer Science, University of York, York, UK, vol. 22, no. 2, **(2004)**, pp. 85-126

[4] M. Markou and S. Singh, “Novelty detection: A review-part 1: Signal Processing archive, vol. 83, no. 12, **(2003)**, pp. 2481-2497.

[5] Z. Bakar, R. Mohemad, A. Ahmad and M. Deris, “A Comparative Study for Outlier Detection Techniques in Data Mining”, IEEE Conference on Cybernetics and Intelligent Systems, **(2006)**, pp. 1-6.

[6] D. Hawkins, “Identification of outliers” Monographs on Applied Probability and Statistics, **(1980)**

[7] H. Javits and A. Valdes, “The NIDES statistical component”, Description and justification, Technical report, SRI International, Computer Science Laboratory, **(1993)**.

[8] M. Mahoney, “Network Traffic Anomaly Detection Based on Packet Bytes”, Proceedings of ACM-SAC, Melbourne, **(2003)**, pp. 346-350.

[9] M. Mahoney and P. K. Chan, “Learning Non stationary Models of Normal Network Traffic for Detecting Novel Attacks”, ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmoton, **(2002)**, pp. 376-385.

[10] W. Zhang, Q. Yang, and Y. Geng, “A Survey of Anomaly Detection Methods in Networks”, in Proc. International Symposium on Computer Network and Multimedia Technology, Wuhan, vol. 18-20. **(2009)**, pp. 1-3.

[11] X. Wu and W. Banzhaf, “The use of computational intelligence in intrusion detection systems: A review”, Applied Soft Computing, vol. 10, no. 1, **(2010)**, pp. 1-35.

[12] Y. Dong, S. Hsu, S. Rajput, and B. Wu, “Experimental Analysis of Application Level Intrusion Detection Algorithms”, International J. Security and Networks, vol. 5, no. 2/3, **(2010)**.

[13] A. Harbola and J. Harbola, “Improved Intrusion Detection in DDoS Applying feature selection Using Rank & Score of Attributes in KDD-99 data set”, International Conference on Computational Intelligence and Communication network, **(2014)**, pp. 840-845.

[14] P.G. Majeed and S. Kumar, “Genetic algorithms in intrusion detection systems: A survey”, International Journal of Innovation and Applied Studies, vol. 5, no. 3, **(2014)**, pp. 233-236.

[15] F. Alserhani, M. Akhlaq, I. U. Awan, A. J. Cullen, J. Mellor, and P. Mirchandani, “Snort Performance Evaluation”, In Proceedings of Twenty Fifth UK Performance Engineering Workshop, Leeds, UK, **(2009)**.

[16] H. Njogu, L. Jiawei, J. Kiere and D. Hanyurwimfura, “A comprehensive vulnerability based alert management approach for large networks, Future Generation Computer Systems”, vol. 29, no. 1, **(2013)**, pp. 27-45.

[17] S. J. Stolfo, A. L. Prodromidis, S. Tselepis, W. Lee, D. W. Fan and P. K. Chan, “In KDD JAM: Java Agents for Meta-Learning over Distributed Databases”, vol. 97, **(1997)**, pp. 74-81.

[18] I. Levin, “KDD-99 classifier learning contest: LLSoft's results overview”, SIGKDD explorations, vol. 1, no. 2, **(2002)**, pp. 67-75.

[19] A. L. Buczak and E. Guven, “A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection”, IEEE Communications Surveys & Tutorials, vol. 18, no. 2, **(2016)**, pp. 1153-1176.