Network Intrusion Detection Using Supervised Machine Learning Technique

Kolukuluri Sri Lalitha
PG scholar, Department of MCA, CDNR collage, Bhimavaram, Andhra Pradesh.
K.Venkatesh
(Assistant Professor), Master of Computer Applications, DNR collage, Bhimavaram, Andhra Pradesh.

Abstract

In the modern computer world, use of the internet is increasing day by day. However, the increasing use of the internet creates some security issues. These days, such new type of security attacks occurs every day and it is not easy to detect and prevent those attacks effectively. One common method of attack involves sending large amount of request to site or server and server will be unable to handle such huge requests and site will be offline for many days. This type of attack is called distributed denial of service (DDOS) attack, which act as a major security threat to internet services and most critical attack for cyber security world. Detection and prevention of Distributed Denial of Service Attack (DDoS) becomes a crucial process for the commercial organizations that uses the internet. Different approaches have been adapted to process traffic information collected by monitoring stations (Routers and Servers) to distinguish malicious traffic such as DDoS attack from normal traffic in Intrusion Detection Systems (IDS). In general, Machine learning techniques can be designed and implemented with the intrusion systems to protect the organizations from malicious traffic. Specifically, supervised clustering techniques allow to effectively distinguishing the normal traffic from malicious traffic with good accuracy. In this paper, machine learning algorithms are used to detect DDoS attacks collected from “KDD cup 99 Dataset”, pre-processing and feature selection technique is used on the dataset to enhance the performance of the classifiers and reduce the detection time. The classification algorithms such as C4.5 decision tree and Navie Bayes is applied on the training dataset and the implementation of the algorithm is done using spyder tool. The performance comparison of algorithms is shown using confusion matrix and it is found that C4.5 decision is more efficient in detection of DDOS attack .The proposed method can be used as DDoS defense system.

KeyWords: C 4.5 Decision Tree, DoS attack detection, IDS, KDD Dataset, Naive Bayesian classifier, Machine learning.

INTRODUCTION

With the rapid development of information technology, the Computer networks are widely used by industry, business and various fields of the human life. As a result, it is very important for IT administrators to create a trusted network. In addition, the rapid development of information technology has created some problems in building a reliable network. This is a very difficult task.

Many types of attacks are threatening the availability, integrity, and privacy of computer networks. A Denial of Service (DOS) attack is considered one of the most common attacks. The purpose of a DOS attack is to temporarily deny multiple end-user services. Typically, network resources are typically consumed and unwanted request systems are overloaded.

For this reason DOS acts as a large umbrella for all types of attacks which aim to consume computer and network resources. Hence, it is very difficult to detect all types of attacks hence the intrusion detection system (IDS) has become an essential part of network security. It is implemented to monitor network traffic in order to generate alerts when any attacks appear.

IDS can be implemented to monitor network traffic of a specific device (host intrusion detection system) or to monitor all network traffics (network intrusion detection system) which is the common type used. In general, there are two types of IDS (anomaly base or misuse base). Anomaly based intrusion detection system is implemented to detect attacks based on recorded normal behavior.

LITERATURE SURVEY

[1]. M. Almseidin, M. Alzubi, S. Kovacs and M. Alkasassbeh, "Evaluation of machine learning algorithms for intrusion detection system," 2017 IEEE 15th International Symposium on Intelligent Systems and Informatics (SISY), Subotica, 2017, pp. 000277- 000282.

Intrusion detection system (IDS) is one of the implemented solutions against harmful attacks. Furthermore, attackers always keep changing their tools and techniques. However, implementing an accepted IDS system is also a challenging task. In this paper, several experiments have been performed and evaluated to assess various machine learning classifiers based on KDD intrusion dataset. It succeeded to compute several performance metrics in order to evaluate the selected classifiers.

The focus was on false negative and false positive performance metrics in order to enhance the detection rate of the intrusion detection system. The implemented experiments demonstrated that the decision table classifier achieved the lowest value of false negative while the random forest classifier has achieved the highest average accuracy rate.

[2]. Arul, Amudha & Subburathinam, Karthik & Sivakumari, S. (2013). Classification Techniques for Intrusion Detection an Overview. International Journal of Computer Applications. 76. 33-40. 10.5120/13334-0928.

Security is becoming a critical part of organizational information systems and security of a computer system or network is compromised when an intrusion takes place. In the field of computer networks security, the detection of threats or attacks is nowadays a critical problem to solve. Intrusion Detection Systems (IDS) have become a standard component in network security infrastructures and is an essential mechanism to protect computer systems from many attacks.

In recent years, intrusion detection using data mining have attracted researchers more and more interests. Different researchers propose a different algorithm in different categories. Classifier construction is another research challenge to build an efficient intrusion detection system.

[3]. Performance Evaluation of Supervised Machine Learning Algorithms for Intrusion Detection

Intrusion detection system plays an important role in network security. Intrusion detection model is a predictive model used to predict the network data traffic as normal or intrusion. Machine Learning algorithms are used to build accurate models for clustering, classification and prediction.

In this paper classification and predictive models for intrusion detection are built by using machine learning classification algorithms namely Logistic Regression, Gaussian Naive Bayes, Support Vector Machine and Random Forest. These algorithms are tested with NSL-KDD data set. Experimental results shows that Random Forest Classifier out performs the other methods in identifying whether the data traffic is normal or an attack.

[4]. Kanagalakshmi.R, V. Naveenantony Raj,“ Network Intrusion Detection Using Hidden Naïve Bayes Multiclass Classifier Model,“ International Journal of Science, Technology & Management ,Volume No.03, Issue No. 12, December 2014.

With increasing Internet connectivity and traffic volume, recent intrusion incidents have reemphasized the importance of network intrusion detection systems for combating increasingly sophisticated network attacks. Techniques such as pattern recognition and the data mining of network events are often used by intrusion detection systems to classify the network events as either normal events or attack events.

Our research study claims that the Hidden Naïve Bayes (HNB) model can be applied to intrusion detection problems that suffer from dimensionality, highly correlated features and high network data stream volumes. HNB is a data mining model that relaxes the Naïve Bayes method's conditional independence assumption.

PROPOSED METHOD

The workflow of the proposed method is setup as shown in the Figure 1, starting with data collection (KDD-99 Dataset), Pre-Processing: Training and testing dataset, building model and result analysis

3.1 KDD cup 1999 dataset Collection

In 1998, the DARPA Intrusion Detection Assessment Program was prepared and managed by MIT Lincoln Labs. Its purpose was to study and evaluate intrusion detection research.

Standard data sets include various simulation intrusions in military network environments. The connection to the dataset includes a sequence of TCP packets beginning and ending at a well-defined time between the source IP address and the destination IP address using a well-defined protocol.

Each connection is categorized as a normal or specific type of attack. Data sets are categorized into five sub-sets: denial-of-service attacks, local or remote network attacks, user / root attacks, sample attacks, and generic data. Each record is classified as normal or attack with exactly one type of attack.

They are categorized as follows:

□ **Denial of service (DoS)** Denial of Service (DOS) allows a legitimate user to gain access to the machine by creating too much or too much computer resources or memory for an attacker to handle legitimate requests.

□ **R2L (Local Remote Attack (User))** Local Remote Attack (R2L) is a type of attack in which an attacker can send packets to a computer over the network and then exploit a vulnerability in the computer to illegally attack local access. On the machine.

• **Root User Attack (U2R)** Root User Attack (U2R) is the attack class that an attacker first accesses a regular user account on a system. The vulnerability could be exploited to gain root access to the system.

• Monitoring (monitoring and other discovery) detection is an attack type in which an attacker scans the network for known information or vulnerabilities. An attacker with a map of systems and services available on the network will use the information to detect attacks.

RESULT

In this paper author is evaluating performance of two supervised machine learning algorithms such as SVM (Support Vector Machine) and ANN (Artificial Neural Networks). Machine learning algorithms will be used to detect whether request data contains normal or attack (anomaly) signatures.

Now-a-days all services are available on internet and malicious users can attack client or server machines through this internet and to avoid such attack request IDS (Network Intrusion Detection System) will be used, IDS will monitor request data and then check if its contains normal or attack signatures, if contains attack signatures then request will be dropped.

IDS will be trained with all possible attacks signatures with machine learning algorithms and then generate train model, whenever new request signatures arrived then this model applied on new request to determine whether it contains normal or attack signatures. In this paper we are evaluating performance of two machine learning algorithms such as SVM and ANN and through experiment we conclude that ANN outperform existing SVM in terms of accuracy.

To avoid all attacks IDS systems has developed which process each incoming request to detect such attacks and if request is coming from genuine users then only it will forward to server for processing, if request contains attack signatures then IDS will drop that request and log such request data into dataset for future detection purpose.

To detect such attacks IDS will be prior train with all possible attacks signatures coming from malicious user's request and then generate a training model. Upon receiving new request IDS will apply that request on that train model to predict it class whether request belongs to normal class or attack class. To train such models and prediction various data mining classification or prediction algorithms will be used.

In this paper author is evaluating performance of SVM and ANN.

In this algorithms author has applied Correlation Based and Chi-Square Based feature selection algorithms to reduce dataset size, this feature selection algorithms removed irrelevant data from dataset and then used model with important features, due to this features selection algorithms dataset size will reduce and accuracy of prediction will increase.

To conduct experiment author has used NSL KDD Dataset and below is some example records of that dataset which contains request signatures. I have also used same dataset and this dataset is available inside 'dataset' folder.

Dataset example

duration,protocol_type,service,flag,src_bytes,dst_bytes,land,wrong_fragment,urgent,hot,num_failed_logins,logged_in,num_compromised,root_shell,su_attempted,num_root,num_file_creations,num_shells,num_access_files,num_outbound_cmds,is_host_login,is_guest_login,count,srv_count,serror_rate,srv_serror_rate,rerror_rate,srv_rerror_rate,same_srv_rate,diff_srv_rate,srv_diff_host_rate,dst_host_count,dst_host_srv_count,dst_host_same_srv_rate,dst_host_diff_srv_rate,dst_host_same_src_port_rate,dst_host_srv_diff_host_rate,dst_host_serror_rate,dst_host_srv_serror_rate,dst_host_rerror_rate,dst_host_srv_rerror_rate,label

All above comma separated names in bold format are the names of request signature

0,tcp,ftp_data,SF,491,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,2,0,0,0,0,1,0,0,150,25,0.17,0.03,0.17,0,0,0,0.05,0,normal

0,tcp,private,S0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,166,9,1,1,0,0,0.05,0.06,0,255,9,0.04,0.05,0,0,1,1,0,0,anamoly

Above two records are the signature values and last value contains class label such as normal request signature or attack signature. In second record 'Neptune' is a name of attack. Similarly in dataset you can find nearly 30 different names of attacks.

In above dataset records we can see some values are in string format such as tcp, ftp_data and these values are not important for prediction and these values will be remove out by applying PREPROCESSING Concept.

All attack names will not be identified by algorithm if it's given in string format so we need to assign numeric value for each attack. All this will be done in PREPROCESS steps and then new file will be generated called 'clean.txt' which will use to generate training model.

In below line i am assigning numeric id to each attack

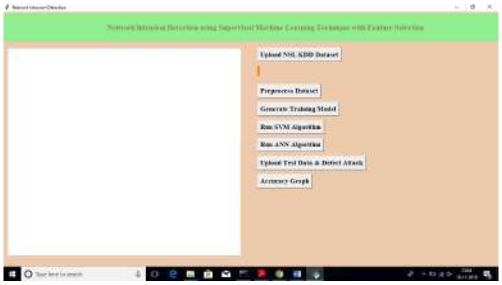
"normal":0,"anamoly":1

In above lines we can see normal is having id 0 and Anomaly has id 1 and goes on for all attacks.

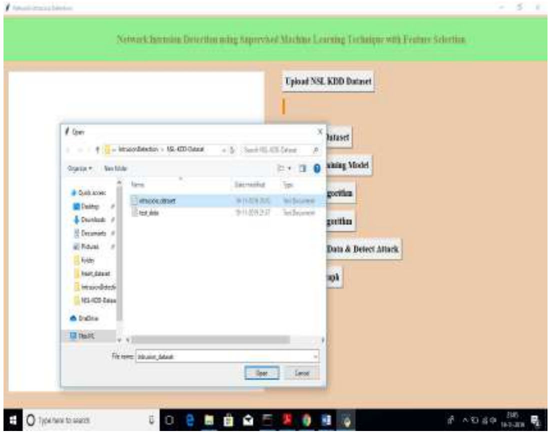
Before running code execute below two commands

Screen shots

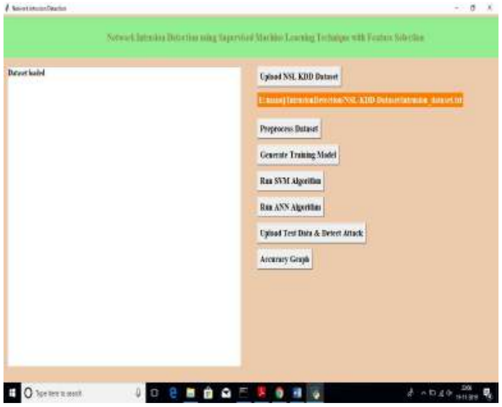
Double click on ‘run.bat’ file to get below screen


Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection
Upload NSL KDD Dataset
Preprocess Dataset
Generate Training Model
Run SVM Algorithm
Run ANN Algorithm
Upload Test Data & Detect Attack
Accuracy Graph

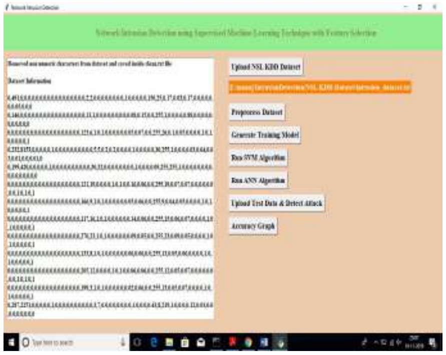
In above screen click on ‘Upload NSL KDD Dataset’ button and upload dataset


Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection
Upload NSL KDD Dataset
Open
intrusion_dataset
test_data
File name: intrusion_dataset
Open
Cancel

In above screen I am uploading ‘intrusion_dataset.txt’ file, after uploading dataset will get below screen

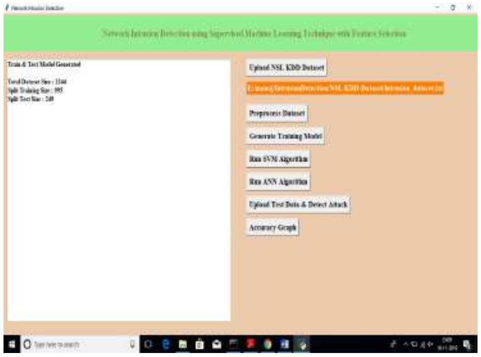

Dataset loaded
Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection
Upload NSL KDD Dataset
Preprocess Dataset
Generate Training Model
Run SVM Algorithm
Run ANN Algorithm
Upload Test Data & Detect Attack
Accuracy Graph

Now click on ‘Pre-process Dataset’ button to clean dataset to remove string values from dataset and to convert attack names to numeric values


Dataset Information
Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection
Upload NSL KDD Dataset
Preprocess Dataset
Generate Training Model
Run SVM Algorithm
Run ANN Algorithm
Upload Test Data & Detect Attack
Accuracy Graph

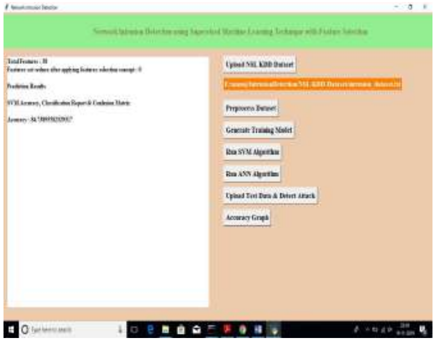
After pre-processing all string values removed and convert string attack names to numeric values such as normal signature contains id 0 and anomaly attack contains signature id 1.

Now click on ‘Generate Training Model’ to split train and test data to generate model for prediction using SVM and ANN

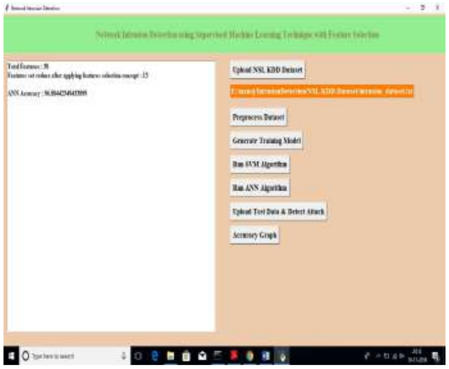

Train & Test Model Generated

Total Dataset Size : 1244
Split Training Size : 995
Split Test Size : 249
Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection
Upload NSL KDD Dataset
Preprocess Dataset
Generate Training Model
Run SVM Algorithm
Run ANN Algorithm
Upload Test Data & Detect Attack
Accuracy Graph

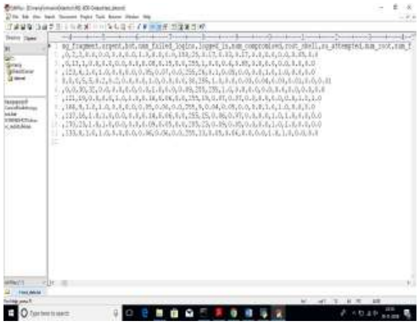
In above screen we can see dataset contains total 1244 records and 995 used for training and 249 used for testing. Now click on ‘Run SVM Algorithm’ to generate SVM model and calculate its model accuracy


Network Intrusion Detection
Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection
Upload NSL KDD Dataset
Preprocess Dataset
Generate Training Model
Run SVM Algorithm
Run ANN Algorithm
Upload Test Data & Detect Attack
Accuracy Graph

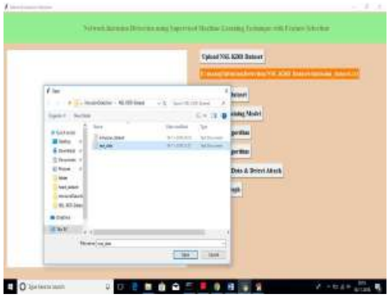
In above screen we can see with SVM we got 84.73% accuracy, now click on 'Run ANN Algorithm' to calculate ANN accuracy


Network Intrusion Detection
Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection
Upload NSL KDD Dataset
Preprocess Dataset
Generate Training Model
Run SVM Algorithm
Run ANN Algorithm
Upload Test Data & Detect Attack
Accuracy Graph

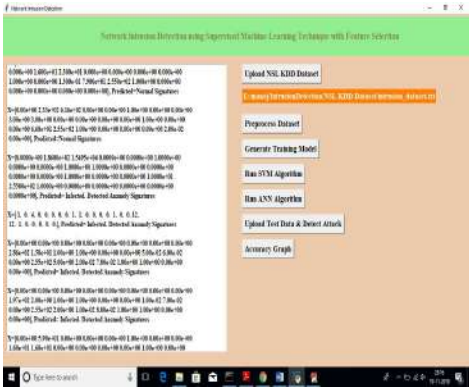
In above screen we got 96.88% accuracy, now we will click on 'Upload Test Data & Detect Attack' button to upload test data and to predict whether test data is normal or contains attack. All test data has no class either 0 or 1 and application will predict and give us result. See below some records from test data



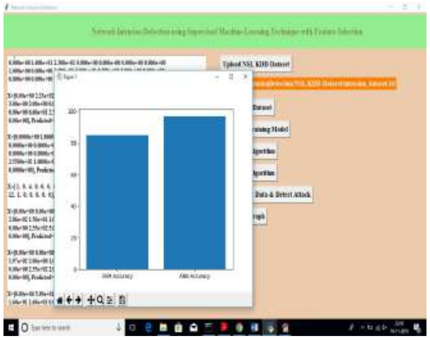
In above test data we don't have either '0' or '1' and application will detect and give us result


Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection
Upload NSL KDD Dataset
Open
Open
Cancel

In above screen I am uploading 'test_data' file which contains test record, after prediction will get below results


Network Intrusion Detection
Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection
Upload NSL KDD Dataset
Preprocess Dataset
Generate Training Model
Run SVM Algorithm
Run ANN Algorithm
Upload Test Data & Detect Attack
Accuracy Graph

In above screen for each test data we got predicted results as 'Normal Signatures' or 'infected' record for each test record. Now click on 'Accuracy Graph' button to see SVM and ANN accuracy comparison in graph format



From above graph we can see ANN got better accuracy compare to SVM, in above graph x-axis contains algorithm name and y-axis represents accuracy of that algorithms

CONCLUSION

In this paper Intrusion detection is considered as a classification problem where each record can be classified into normal or a particular kind of intrusion. Intrusion detection using machine learning has attracted more and more interests in recent years. As an important application of machine learning, an accurate intrusion detection model is built by choosing an effective classification approach. This paper shows the comparison of the most well-known classification algorithms like C4.5 decision trees and Naive Bayes has been carried out . These algorithms are tested with the KDD data-set. Effective classifier is identified by comparing the performances based on the accuracy and confusion matrix. Performance calculation is done by considering only the important attributes for the intrusion detection. From the observed results it can be concluded that the C4.5 decision trees classifier outperforms other classifiers for the considered data-set and parameters. It has the accuracy of 99%.

REFERENCES

[1]. M. Almseidin, M. Alzubi, S. Kovacs and M. Alkasassbeh, "Evaluation of machine learning algorithms for intrusion detection system," 2017 IEEE 15th International Symposium on Intelligent Systems and Informatics (SISY), Subotica, 2017, pp. 000277- 000282.

[2]. Arul, Amudha & Subburathinam, Karthik & Sivakumari, S. (2013). Classification Techniques for Intrusion Detection an Overview. International Journal of Computer Applications. 76. 33-40. 10.5120/13334-0928.

[3]. Performance Evaluation of Supervised Machine Learning Algorithms for Intrusion Detection

[4]. Kanagalakshmi.R, V. Naveenantony Raj,‟ Network Intrusion Detection Using Hidden Naïve Bayes Multiclass Classifier Model,‟ International Journal of Science, Technology & Management ,Volume No.03, Issue No. 12, December 2014.

[5]. M. Alkasassbeh, G. Al-Naymat et.al,‛ Detecting Distributed Denial of Service Attacks Using Data Mining Technique,‛ (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, pp. 436-445, 2016. Science and Information Technologies, Vol. 6 (2), pp. 1096-1099, 2015.

[6]. Jasreena Kaur Bains ,Kiran Kumar Kaki ,Kapil Sharma,‟ Intrusion Detection System with MultiLayer using Bayesian Networks‟ , International Journal of Computer Applications (0975 – 8887) Volume 67– No.5, April 2013.

[7]. Dewan Md. Farid, Nouria Harbi, Mohammad Zahidur Rahman , Combining Naive Bayes and Decision Tree for Adaptive Intrusion Detection, Proc. of Intl. Journal of Network Security & Its Applications (IJNSA), Volume 2, Number 2, 2010, pp.12-25.

[8]. Domingos P. and Pazzani M., Beyond Independence: Conditions for the optimality of the simple Bayesian Classifier, In proceedings of the 13th Intnl. Conference on Machine Learning, 1996, pp.105-110.

[9]. V. Hema and C. Emilin Shyni, ‟ DoS Attack Detection Based on Naive Bayes Classifier, ‟ Middle-East Journal of Scientific Research 23 (Sensing, Signal Processing and Security): 398-405, 2015.

[10]. Yi-Chi Wu, Huei-Ru Tseng, Wu Yang* and RongHong Jan,‛ DDoS detection and trackback with decision tree and grey relational analysis‛, Int. J. Ad Hoc and Ubiquitous Computing, Vol. 7, No. 2, 2011.

[11]. Dewan Md. Farid, Nouria Harbi, Emna Bahri, Mohammad Zahid ur Rahman, Chowdhury Mofizur Rahman,‟ Attacks Classification in Adaptive Intrusion Detection using Decision Tree ,International Journal of Computer, Electrical, Automation, Control and Information Engineering, Vol:4, No:3, 2010.

[12]. Quinlan, C4.5: Programs for Machine Learning, 1993, Morgan Kaufmann Publishers, San Mateo, CA.

[13]. Sabhnani M, Serpen G(2003), Application of Machine Learning Algorithms to KDD Intrusion Detection Dataset within Misuse Detection Context, In Proc. of the Intl. Conference on Machine Learning, Models, Technologies and Applications, vol. 1, pp. 209–215.

[14]. Gharibian F, Ghorbani A.A , Comparative Study of Supervised Machine Learning Techniques for Intrusion Detection, Proc. of the Fifth Annual Conference on Communication Networks and Services Research, 2007, pp. 350–358.

[15]. Ohta S, R. Kurebayashi and K. Kobayashi. , Minimizing false positives of a decision tree classifier for intrusion detection on the internet, Journal of Networks System Management, vol.16, 2008, pp.399– 419. ISSN 1064-7570.

[16] M. Kemiche and R. Beghdad, Intelligent Systems in Science and Information 2014: Extended and Selected Results from the Science and Information Conference 2014, Cham: Springer International Publishing, ch. Towards Using Games Theory to Detect New U2R Attacks, pp. 351–367, (2015). [Online]. Available: http://dx.doi.org/10.1007/978-3-319-14654-6-22

[17] S. Patil, D. V. K. B. P, S. Singha and R. Jamil, A Survey on Authentication Techniques for Wireless Sensor Networks, International Journal of Applied Engineering Research, vol. 7, (2012).

[18] T. M. Mitchell, Machine Learning, 1st ed, New York, NY, USA: McGraw-Hill, Inc., (1997).

[19] D. S. Kim and J. S. Park, Network-Based Intrusion Detection with Support Vector Machines, Information Networking: International Conference, ICOIN 2003, Cheju Island, Korea, February 12–14, (2003). Revised Selected Papers, Berlin, Heidelberg: Springer Berlin Heidelberg, ch. pp. 747–756, (2003). [Online]. Available: http://dx.doi.org/10.1007/978-3-540-45235-5-73

[20] H. Altwaijry and S. Algarny, Bayesian Based Intrusion Detection System, Journal of King Saud University – Computer and Information Sciences, vol. 24, no. 1, pp. 1–6, (2012). [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1319157811000292