International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 4, Issue 12, ISSN No. 2455-2143, Pages 324-330
Published Online April 2020 in IJEAST (http://www.ijeast.com)

BEST ACCURACY PREDICTION TO NETWORK ATTACKS USING SUPERVISED MACHINE LEARNING ALGORITHM

Dr. G. Prabaharan
Department of computer science and engineering
Assistant Professor, Adhiparasakthi Engineering College

D. Sangeetha Devi, K. Sowmiya, D. Ramani
Department of computer science and engineering
UG Students, Adhiparasakthi Engineering College

ABSTRACT---**To create data for the Intrusion Detection System (IDS), it is necessary to set the real working environment to explore all the possibilities of attacks. The existing system is expensive because of its hardware implementation. Software to detect network intrusion protects a computer network from unauthorized users. The intrusion detector learning task is to build a predictive model (i.e. a classifier) capable of distinguishing between "intrusions" or "attacks", and "normal" connections. To prevent this problem in network, sectors have to predict whether the connection is attacked or not from KDDCup99 (Knowledge Discovery and Data mining) dataset using machine learning techniques. The aim is to investigate machine learning based techniques for better network connection by finding accuracy prediction results. Machine learning based method is proposed to accurately predict the DOS (Denial of Service) attack, R2L (Remote to User), U2R (User to Root), Probe and overall attacks when compared with the existing supervised classification machine learning algorithms. This accuracy predictor shows that the effectiveness of the proposed machine learning algorithm technique with the existing algorithms are compared for best accuracy by various parameters precision, Recall and F1 Score.**

*KEYWORDS---**Dataset, Supervised Machine learning Classification method, Prediction of Accuracy result.***

I. INTRODUCTION

An IDS (Intrusion Detection System) is a device that is placed inside a protected network to monitor what occurs within the network. Many organizations use Intrusion Detection Systems to help them to determine if their networks have been harmed or not. Major focus of machine learning research is to automatically learn complex patterns and make intelligent decisions based on data. Main difficulty lies in the fact that the set of all possible behaviors are hard to describe. IDS may complement other preventive controls as the next line of defense within the organization.

The Association for Computing Machinery (ACM) has a special interest group on Knowledge Discovery and Data mining (KDD) 38 which is the most popular professional organization of data miners. The KDD organized the annual Data Mining and Knowledge Discovery competition called KDD Cup in different areas. (http://www.sigkdd.org/kddcup)Various focused areas of KDD have been tabulated in Table 1.

Year	Focused Area
KDD-CUP 1997	Direct marketing for lift curve optimization
KDD-CUP 1998	Direct marketing for profit optimization
KDD-CUP 1999	Computer network intrusion detection
KDD-CUP 2000	Online retailer website click stream analysis
KDD-CUP 2001	Molecular bioactivity and Protein locale prediction
KDD-CUP 2002	Bio Medical document and Gene role classification
KDD-CUP 2003	Network mining and usage log analysis
KDD-CUP 2004	Particle physics; plus Protein homology prediction
KDD-CUP 2005	Internet user search query categorization
KDD-CUP 2006	Pulmonary embolisms detection from image data
KDD-CUP 2007	Consumer recommendations
KDD-CUP 2008	Breast cancer
KDD-CUP 2009	Fast scoring on a large database

Table.1: KDD-CUP center of attention

A. Aim

The aim is to investigate machine learning based techniques for better packet connection transfers forecasting by prediction results in best accuracy.

International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 4, Issue 12, ISSN No. 2455-2143, Pages 324-330
Published Online April 2020 in IJEAST (http://www.ijeast.com)

B. Scope

The scope of this project is to investigate a dataset of network connection attacks for KDD records for medical sector using machine learning technique. To identifying network connection is attacked or not.

C. Objective

The major objectives of intrusion detection system are:

- ✓ To accurately detect anomalous network behavior or misuse of resources.
- ✓ To Sort out true attacks from false alarms.
- ✓ To notify the Network administrators of the activity.

II. PROBLEM DESCRIPTION

Lately, an internet network company in Japan has been facing huge losses due to malicious server attacks. They've encountered breach in data security, reduced data transfer speed and intermittent breakdowns in user-user & user-network connections. When asked, a company official said, "there's a significant dip in the number of active users on our network ". The company is looking are some predictive analytics solution to help them understand, detect and counter the attacks and make their network connection secure. Think of a connection as a sequence of TCP packets starting and ending at some well-defined times, between which data flows to and from a source IP address to a target IP address under some well-defined protocol. In total, there are 3 major type of attacks to which their network is vulnerable to. But, 3 of them cause the maximum damage. In this challenge, you are given an anonymized sample dataset of server connections.

III. RELATED WORK

A connection is a sequence of TCP packets starting and ending at some time duration between which data flows to and from a source IP address to a target IP address under some well-defined protocol. Also, each connection is labeled as either normal or as an attack with exactly one specific attack type. Each connection record consists of about 100 bytes. The raw training data is about four gigabytes of compressed binary TCP dump data obtained from seven weeks of network traffic. Finally, the completed process generated around five million connection records. Likewise, the two weeks of test data gives around two million connection records. For each TCP/IP connection, 41 various quantitative and qualitative features are obtained with normal and attack data.

In 1999, the KDD acknowledged and approved DARPA data as the conventional benchmark data base for IDS called KDD Cup99 which is available in http://www.kdd.ics.uci.edu /databases/kddcuip99/task.html. There are 41 features used to represent each group of packets.

The data set in KDD Cup99 have normal and 22 attack type data with 41 features and Table 2 shows few data set. All generated traffic patterns end with a label either as 'normal' or any type of 'attack' for upcoming analysis.

Feature Name	**Packet-1 (normal)**	**Packet-2 (neptune)**
duration	0	0
protocol_type	tcp	tcp
service	http	private
flag	SF	REJ
src_bytes	327	0
dst_bytes	467	0
land	0	0
wrong_fragment	0	0
urgent	0	0
hot	0	0
num_failed_logins	0	0
logged_in	1	0
num_compromised	0	0
root_shell	0	0
su_attempted	0	0
num_root	0	0
num_file_creations	0	0
num_shells	0	0
num_access_files	0	0
num_outbound_cmds	0	0
is_hot_login	0	0
is_guest_login	0	0
count	33	136
srv_count	47	1
serror_rate	0.00	0.00
srv_serror_rate	0.00	0.00
rerror_rate	0.00	1.00
srv_rerror_rate	0.00	1.00

International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 4, Issue 12, ISSN No. 2455-2143, Pages 324-330
Published Online April 2020 in IJEAST (http://www.ijeast.com)

same_srv_rate	1.00	0.01

Table.2: Sample packet Data

There are varieties of attacks which are entering into the network over a period of time and the attacks are classified into the following four main classes.

- ✓ Denial of Service (DoS)
- ✓ User to Root (U2R)
- ✓ Remote to User (R2L)
- ✓ Probing

A. Denial of Service

Denial of Service is a class of attacks where an attacker makes some computing or memory resource too busy or too full to handle legitimate requests, denying legitimate users access to a machine. The different ways to launch a DoS attack are:

- ✓ By abusing the computer's legitimate features.
- ✓ By targeting the implementation bugs.
- ✓ By exploiting the misconfiguration of systems.

DoS attacks are classified based on the services that an attacker renders unavailable to legitimate users.

B. User to Root

In User to Root attack, an attacker starts with access to a normal user account on the system and gains root access. Regular programming mistakes and environment assumption give an attacker the opportunity to exploit the vulnerability of root access.

C. Remote to User

In Remote to User attack, an attacker sends packets to a machine over a network that exploits the machine's vulnerability to gain local access as a user illegally. There are different types of R2L attacks and the most common attack in this class is done by using social engineering.

D. Probing

Probing is a class of attacks where an attacker scans a network to gather information in order to find known vulnerabilities. An attacker with a map of machines and services that are available on a network can manipulate the information to look for exploits. There are different types of probes: some of them abuse the computer's legitimate features and some of them use social engineering techniques. This class of attacks is the most common because it requires very little technical expertise.

Category of attacks	**Types of attacks**
Denial of Service (DOS)	back, Neptune, ping of death, land, pod, smurf, teardrop,
Remote to Local (R2L)	ftp_write, multihop, phf, spy, warezclient, warezmaster, imap, guess_passwd
User to root (U2R)	buffer_overflow, loadmodule, perl, rootkit
Probe	ipsweep, nmap, satan, portsweep

Table.3: Attack Types Grouped to respective Class

IV. EXISTING SYSTEM

The system focuses on the conception of a monitoring network that can able to detect and classify jamming and protocol-based attacks. To achieve this goal, the system proposed to outsource the attack detection function to protect the network and used an antenna to monitor the spectrum over the time. The Wi-Fi network and the attacks were carried out in an anechoic chamber to avoid disturbing other Wi-Fi communication networks in the vicinity. The spectra highlights that the frequencies of interest belong to the communication channel between 2.402 and 2.422 GHz. Focusing the analysis on this 20-MHz frequency band permits to construct a classification model to overcome the problems induced by the utilization of the adjacent channels that can be or not occupied by other Wi-Fi communications. On these frequencies, the proposed estimation model shows good results in the prediction of attacks. In addition, the correction using the K spectra nearest in time permits to correct most of the miss classification.

A. Drawbacks

- ✓ The model doesn't discuss to know how our model can evolve in the case where unknown attack occurs with all types of attacks by popular machine learning algorithms.
- ✓ The model doesn't describe each categorized of DOS attacks like back, Neptune etc. based on the network connections.

V. PROPOSED SYSTEM

Exploratory Data Analysis (EDA) is not meant to be providing a final conclusion on the reasons leading to network sector as it doesn't involve using any inferential statistics techniques/machine learning algorithms. Data sets of various network connects are applied to Machine learning supervised classification algorithms, which help in extracting patterns and predicting the likely network affected or not, thereby helping to making better decisions to avoid network related attacks. Multiple datasets from different sources are combined to form a generalized dataset, and are applied to different machine learning algorithms for obtaining better results with maximum accuracy.

A. Data Wrangling

Data Wrangling is used to load the data, check for cleanliness, trim and clean given dataset for analysis. The

International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 4, Issue 12, ISSN No. 2455-2143, Pages 324-330
Published Online April 2020 in IJEAST (http://www.ijeast.com)

document is ensured to step carefully and justify for perfect decisions.

B. Data Collection

The data set collected for predicting the network attacks is split into Training set and Test set. Generally, 7:3 ratios are applied to split the Training set and Test set. The Data Model which was created using Random Forest, logistic regression, Decision tree algorithms, K-Nearest Neighbor (KNN) and Support vector classifier (SVC) are applied on the Training set and based on the test result accuracy, Test set prediction is done.

C. Preprocessing

The data which was collected might contain missing values that may lead to inconsistency. To gain better results, data need to be preprocessed so as to improve the efficiency of the algorithm. The outliers have to be removed and also variable conversion need to be done. The correlation among attributes can be identified using plot diagram in data visualization process. Data preprocessing is the most time consuming phase of a data mining process. Data cleaning of connections, data removed several attributes that has no significance about the behavior of a packet transfers. Data integration, data reduction and data transformation are also to be applicable for network connections dataset. For easy analysis, the data is reduced to some minimum amount of records. Initially the Attributes which are critical to make a loan credibility prediction is identified with information gain as the attribute-evaluator and Ranker as the search-method.

VI. ALGORITHMS

In machine learning and statistics, classification is a supervised learning approach in which the computer program learns from the data input given to it and then uses this learning to classify new observation. This data set may simply be bi-class (like identifying whether the person is male or female or that the mail is spam or non-spam) or it may be multi-class too. Some examples of classification problems are: speech recognition, handwriting recognition, bio metric identification, document classification etc. In Supervised Learning, algorithms learn from labeled data. After understanding the data, the algorithm determines which label should be given to new data based on pattern and associating the patterns to the unlabeled new data.

Used Python Packages:

- ✓ Sklearn
- ✓ NumPy
- ✓ Pandas
- ✓ Matplotlib

1. Logistic Regression

It is a statistical method for analyzing a data set in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables. Logistic regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.).

2. Decision Tree

It is one of the most powerful and popular algorithm. Decision-tree algorithm falls under the category of supervised learning algorithms. It works for both continuous as well as categorical output variables. Decision tree builds classification or regression models in the form of a tree structure. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. A decision node has two or more branches and a leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data. Decision tree builds classification or regression models in the form of a tree structure. It utilizes an if-then rule set which is mutually exclusive and exhaustive for classification. The rules are learned sequentially using the training data one at a time. Each time a rule is learned, the tuples covered by the rules are removed.

This process is continued on the training set until meeting a termination condition. It is constructed in a top-down recursive divide-and-conquer manner. All the attributes should be categorical. Otherwise, they should be discretized in advance. Attributes in the top of the tree have more impact towards in the classification and they are identified using the information gain concept. A decision tree can be easily over-fitted generating too many branches and may reflect anomalies due to noise or outliers.

3. Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set. Random forest is a type of supervised machine learning algorithm based on ensemble learning. Ensemble learning is a type of learning where you join different types of algorithms or same algorithm multiple times to form a more powerful prediction model. The random forest algorithm combines multiple algorithm of the same type i.e. multiple decision

International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 4, Issue 12, ISSN No. 2455-2143, Pages 324-330
Published Online April 2020 in IJEAST (http://www.ijeast.com)

trees, resulting in a forest of trees, hence the name "Random Forest". The random forest algorithm can be used for both regression and classification tasks.
The following are the basic steps involved in performing the random forest algorithm:

- ✓ Pick N random records from the dataset.
- ✓ Build a decision tree based on these N records.
- ✓ Choose the number of trees you want in your algorithm and repeat steps 1 and 2.
- ✓ In case of regression problem, for new record, each tree in forest predicts value for Y (output).

The final value can be calculated by taking the average of all the values predicted by all the trees in forest. Or, in case of a classification problem, each tree in the forest predicts the category to which the new record belongs. Finally, the new record is assigned to the category that wins the majority vote.

VII. SYSTEM ARCHITECTURE

System configuration is the hypothetical model that portrays the structure, lead, and more points of view on a structure. A designing delineation is a traditional depiction and depiction of a system, created to such an extent that supports contemplating the structures and practices of the structure. A structure configuration can contain system parts and the sub-systems made, that will coordinate to execute the general system.

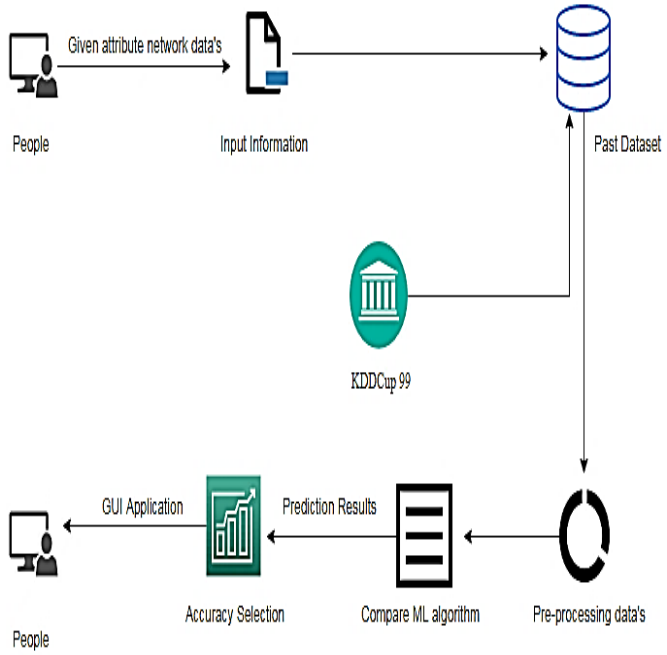
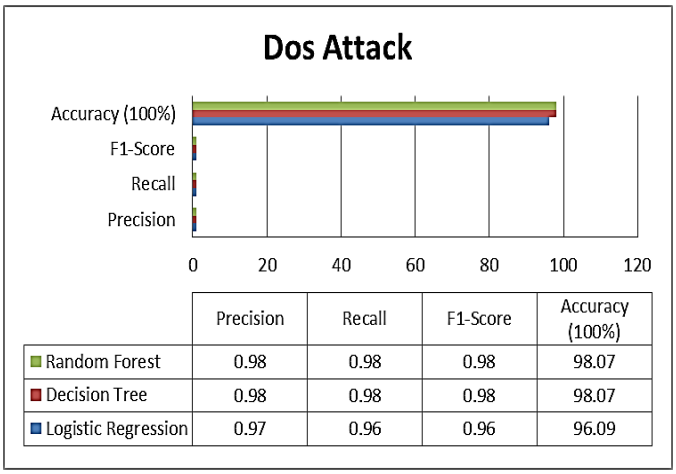


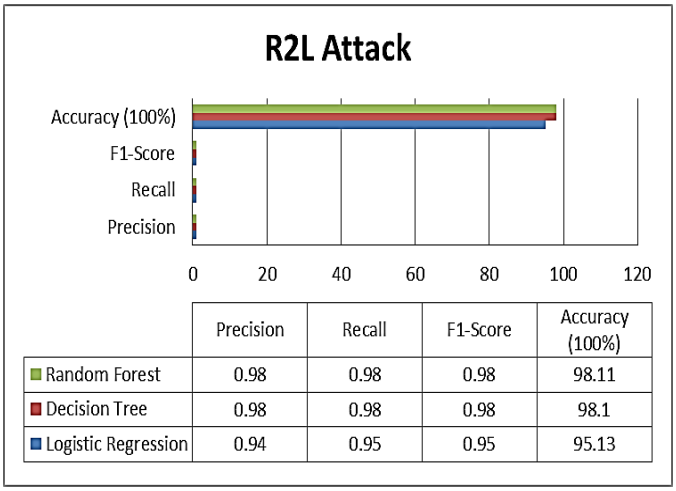
Fig.1: System Architecture

VIII. RESULT ANALYSIS



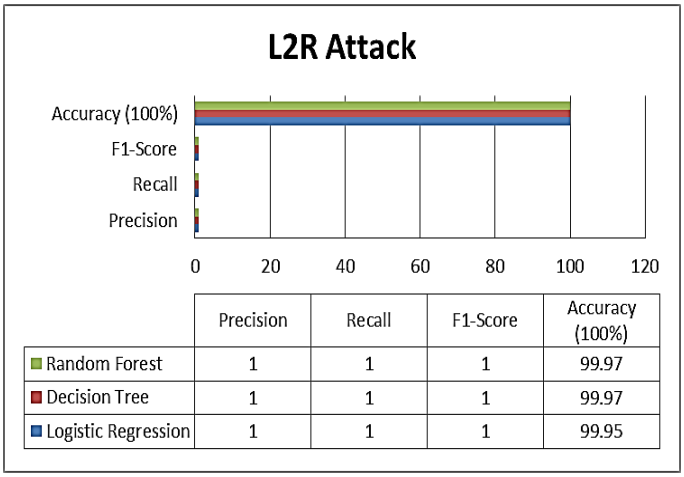
	Precision	Recall	F1-Score	Accuracy (100%)
Random Forest	0.98	0.98	0.98	98.07
Decision Tree	0.98	0.98	0.98	98.07
Logistic Regression	0.97	0.96	0.96	96.09

Fig.2: DOS Attack



	Precision	Recall	F1-Score	Accuracy (100%)
Random Forest	0.98	0.98	0.98	98.11
Decision Tree	0.98	0.98	0.98	98.1
Logistic Regression	0.94	0.95	0.95	95.13

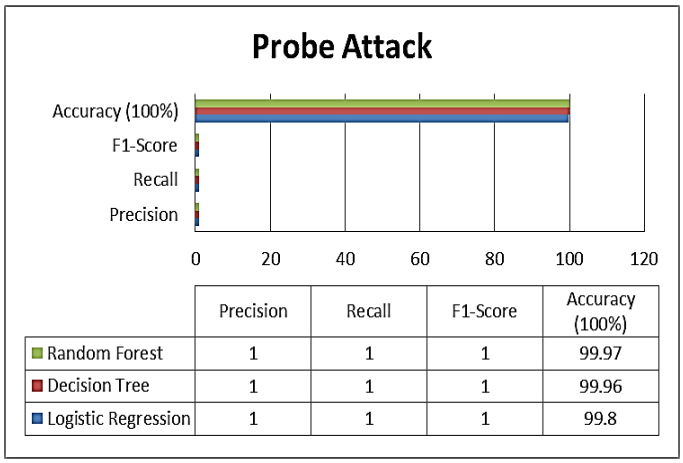
Fig.3: R2L Attack



	Precision	Recall	F1-Score	Accuracy (100%)
Random Forest	1	1	1	99.97
Decision Tree	1	1	1	99.97
Logistic Regression	1	1	1	99.95

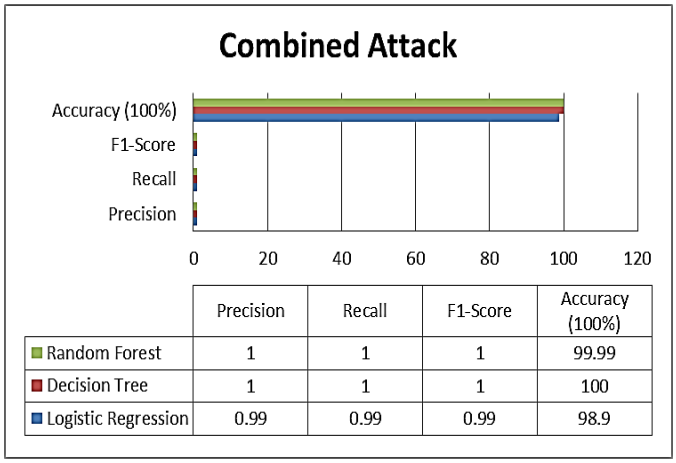
Fig.4: L2R Attack

International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 4, Issue 12, ISSN No. 2455-2143, Pages 324-330
Published Online April 2020 in IJEAST (http://www.ijeast.com)



	Precision	Recall	F1-Score	Accuracy (100%)
Random Forest	1	1	1	99.97
Decision Tree	1	1	1	99.96
Logistic Regression	1	1	1	99.8

Fig.5: Probe Attack



	Precision	Recall	F1-Score	Accuracy (100%)
Random Forest	1	1	1	99.99
Decision Tree	1	1	1	100
Logistic Regression	0.99	0.99	0.99	98.9

Fig.6: Combined Attack

IX. FUTURE ENHANCEMENT

Network sector want to automate the detecting the attacks of packet transfers from eligibility process (real time) based on connection detail. To automate this process by show the prediction result in web application or desktop application. To optimize the work to implement in artificial intelligence environment.

X. ACKNOWLEDGEMENT

The authors acknowledged the anonymous reviewers and editors for their efforts and valuable comments and suggestions.

XI. CONCLUSION

The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. The best accuracy on public test set is higher accuracy score of Decision tree algorithm which brings some insights about diagnosing network attack for new connection. A prediction model is presented with aid of AI to improve human accuracy and provide early detection scope. This model inferred that, area analysis and use of ML technique is useful in developing prediction models that can helps to network sectors reduce the long process of diagnosis and eradicate any human error.

XII. REFERENCES

[1] Bindra, Naveen & Sood, Manu. (2019), Detecting DDoS Attacks Using Machine Learning Techniques and Contemporary Intrusion Detection Dataset Automatic Control and Computer Sciences. 53. 419-428. 10.3103/S0146411619050043.

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

[3] Mellor, A., Haywood, A., Stone, C., and Jones, S., (2013) The performance of random forests in an operational setting for large area sclerophyll forest classification, Remote Sens., vol. 5, no. 6, pp. 2838–2856.

[4] 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.

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

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

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

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

[9] 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.

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

[11] Yi-Chi Wu, Huei-Ru Tseng, Wu Yang* and RongHong Jan, (2011)‘ DDoS detection and trackback with decision tree and grey relational analysis‘, Int. J. Ad Hoc and Ubiquitous Computing, Vol. 7, No. 2.
[12] C. H. Rowland, (2002) Intrusion detection system, U.S. Patent 6 405 318.