ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 9, September 2018

Ensemble Methodology Approach for Improving Anomaly Detection Accuracy

Anuja Sawant [1]

P.G. Student, Department of Computer Engineering, Pune Institute of Computer Technology, Dhankwadi, Maharashtra, India[1]

ABSTRACT: The use of network based system is increasing continuously due to reduced cost of devices and reduced cost of communication as 4G and 5G technologies have reached to the door steps of consumers. However the security aspects are still being researched. Many a time networks are attacked causing losses. Firewall attacks are one of the security concerns in communication. This paper presents a work to detect anomalous communication in the network using ensemble based methodology. Two ensemble methodologies are presented, each consisting of three well known classifiers. The ensemble presented here gives better results than the individual algorithms considered.

KEYWORDS: Anomaly Detection, Classification, Feature Selection, Network Security.

I. Introduction

With increase in number of Internet usage wide numbers of attacks are also increasing daily. IDS help to detect these attacks using ML techniques. IDS is designed to monitor the network traffic and identify the suspicious patterns representing network intrusion that may compromise the system. That is, it continuously inspects network traffic for potential vulnerabilities [1]. The main purpose of ensemble methodology is to minimize the false negative rate.

Last ten years, witnessed the tremendous advancement in technology, specifically the communication technology. As well, it has also marked the easy adaption of the technology by the people from all strata of life. People are using this technology not only for simple text or telephonic application but also for financial transaction as well as computational purposes. With tremendous growth in personal and organizational transactions and data, secure communication has become one of the most important research area. So in communication systems it is equally important to analyze the traffic by looking at messages from source to destination as it may give valuable information not just about imminent attacks but also about the unit movements and other routine matters [2]. Over the time many attacks on computer communication have been discussed and researched for the solutions [3]. Out of such attacks we focus on anomalous communications.

Anomalies are unusual events which differ from normal or something that deviates from anything that is standard or expected, like a sudden rise in temperature, country with left-hand driving rule, seen with a vehicle moving in opposite direction, such situations would be termed as anomalies. Anomalies can be classified as normal or abnormal. Anomaly in network communication could be different attacks, such as DoS (Service to authorized user gets denied due to busy network by malicious node), Probe (an action taken to learn something about the state of network, simply by sending a ping), U2R (gives root access to normal user), R2L (designed to give local access to target systems). Any sudden change in upload or download speed can also be considered as anomaly. It also includes malicious attacks such as worms, viruses, trojan horses and spyware, also abrupt node or link disconnection, multiple authentication attempts on same object. Anomalies could occur in network communication due to mis-configuration of the systems or system that run on undocumented services.

The proposed system uses Decision Tree (DT), Naive Bayes (NB), Multilayer Perceptron (MLP) and Gradient Boosting (GB) to find the accuracy of anomalies detected. To reduce the False Negative Rate Ensemble Methodology is applied.

DOI: 10.15680/IJIRCCE.2018. 0609005

As we know firewall system was able to detect and block the unauthorized access and programs to our local network. But Firewall system does not protect against malicious contents coming through the permitted ways into your local network.

Following are the detail description of algorithms used:

- Decision Tree - J48 : Decision tree is a predictive modelling tool. It is used to identify the ways to split a dataset into subsets based on different conditions. The simple aim of decision tree to create a model that predicts the value of a class labels by learning decision rules deduced from the data features. One of the most used model of DT constructs a tree from a set of available training data using the concept of information entropy. At each node of the tree, the algorithm selects the attribute of the data that most effectively splits its set of samples into subsets enrich in one class or other. The splitting criteria is the normalized information gain. The attribute with highest normalized information gain is chosen to make the decision[4].

- Naive Bayes : Naive Bayes is a probabilistic classifier that makes classification using posterior decision rules. To get the probability it analysis the relation between the dependent and the independent variables. It is faster and easy to interpret, but limits the requirement of prior probability [4][5].

- Multilayer Perceptron (MLP) : It is classification based technique, used to classify sets that are linearly separable, and if the instance cannot be separated linearly then the process will never be able to classify the instances properly. The solution is to use MLP which is also known as feed forward networks [5].

- Gradient Boosting (GB) : Gradient boosting is a Machine Learning technique used for regression and classification. It an ensemble method called as boosting which uses sequential predictors and learns from mistakes of previous predictors [6].

- Ensemble Method : It is a technique which use Machine Learning algorithms to obtain better predictive performance than it could be obtained from any of the constituent Learning algorithms alone. Ensemble methods usually produces more accurate solutions than a single model would. An ensemble is itself a supervised Learning algorithm, because it can be trained and then used to make predictions [7][8]. Ensemble helps to improve the performance of the system as it combines the results of several models. This approach gives better predictive performance as compared to a single model.

- Ensemble-1 : Group-1 includes combination of DT-J48, GNB and MLP algorithms to perform predictions on the same target labels.

- Ensemble-2 : Group-2 includes combination of DT-J48, GNB and GB algorithms to perform predictions on the same target labels.

II. Related work

Koc *et al.* [9] proposed an HNB (Hidden Naive Bayes) technique to identify intrusion which is then compared with the naive bayes technique. To test this technique KDD-CUP dataset is used. HNB performs best in terms of accuracy, error rate and miss-classification cost. It has shown improved accuracy in detecting DoS attack only.

Prerau *et al.* [10] proposed an optimized KNN algorithm for unsupervised anomaly detection. This has been tested on KDD-CUP dataset. The proposed optimization part in this paper means breaking down the search space into smaller subsets. Kulling, which eliminates data in linear time from cluster information is another form of optimization on KNN algorithm. Canopy clustering is used for breaking down the space into smaller subsets which will help new instances to be quickly tested against the smaller number of similarly located instances.

Krishnan *et al.* [11] did a research to address the issue of predictive and decision making of false positive attacks in network system. To detect and classify network-based attacks an IDS based on Multilayer perceptron feedforward artificial neural network is proposed. This technique is tested on KDD-CUP dataset which classifies the normal and abnormal data. It improves the false positive rate but is suitable for detecting only four types of attacks, viz. probe, DoS, U2R and R2L.

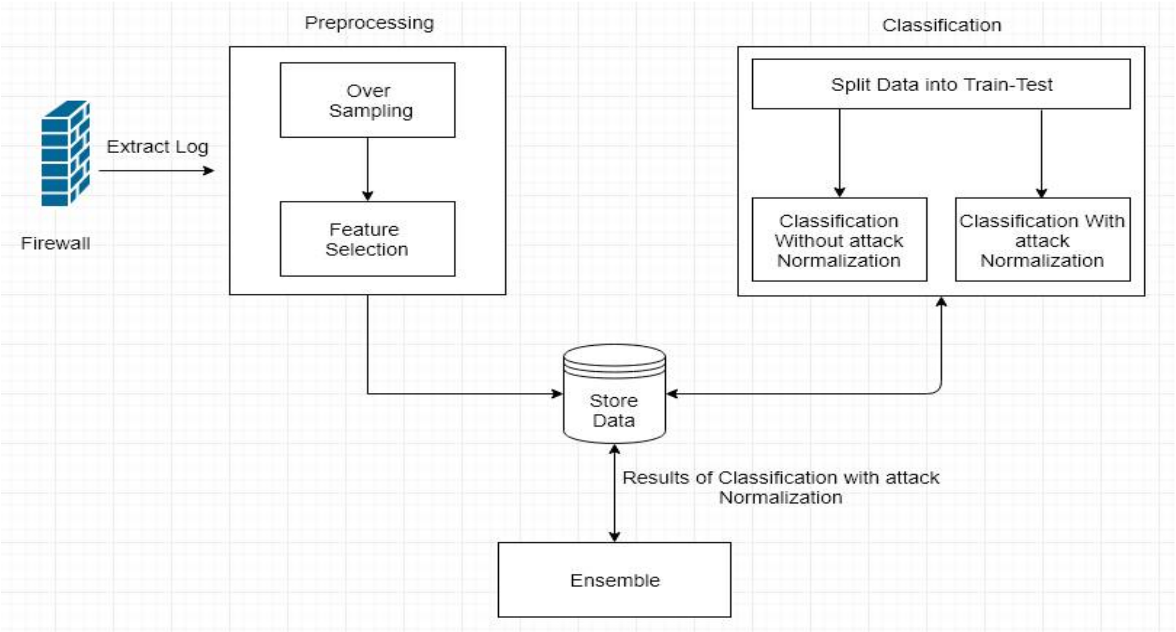
Kumari *et al.* [12] had tested a k-means clustering technique on KDD-CUP dataset. Spark technology is used to process the dataset which helps to obtain specific features from the data. Streaming K-means clustering technique is applied on the processed data which will update a cluster as new data arrives. This cluster is then normalized using euclidean distance which determines the closeness of the data points not only to one cluster but also to other clusters as well. This technique is not just useful for anomaly detection but can also be applied to study financial data, behavior of customers, market basket analysis.

Shakya *et al.* [13] proposed a hybrid approach for anomaly detection. This proposed hybrid approach uses Support Vector Machine and Naive Bayes technique for anomaly detection. The proposed algorithm was tested on 10\% KDD-CUP dataset. The results of SVM and Naive Bayes are compared with hybrid algorithm where hybrid shows better precision, recall, accuracy and F1-score than SVM and Naive Bayes. So hybrid algorithm is efficient in terms of reducing the false alarm ratio for anomaly detection.

In the above survey many researchers have found Machine Learning approaches to be quite useful in detecting the anomalies. Some researchers have also demonstrated hybrid approaches to detect anomaly. With such a study, we believe that there is an opportunity to try and test ensemble kind of approach for detecting anomalies. Such ensemble will not only detect but will also give a confidence of correctness on the detected anomalies.

III. Proposed System

A. *Architecture:*


Preprocessing
Classification
Over Sampling
Feature Selection
Split Data into Train-Test
Classification Without attack Normalization
Classification With attack Normalization
Extract Log
Firewall
Store Data
Results of Classification with attack Normalization
Ensemble

Firewall logs are extracted from network firewall and given for preprocessing of data. These preprocessed results are collected in a data store. In preprocessing task oversampling is done on classes of extracted data, then important features are selected from it. Classification is performed on this preprocessed data, where data is split into train-test and prediction results are calculated on both the data, with attack normalization and without attack normalization, accordingly results are stored separately in the data store. Ensemble is done on classification results and ensemble results are stored in the data store for further analysis.

B. *Description of Algorithm Used:*

- Decision Tree

$$E(S) = \sum_{i=1}^{c} - P_i \log_2 P_i$$

$$E(T, X) = \sum_{c€X} P(c)\, E(c)$$

$$Gain(T, X) = Entropy(T) - Entropy(T, X)$$

- Naive Bayes

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

- Multilayer Perceptron

$$y = \sum_{i=1}^{m} (w_i x_i) + \text{bias}$$

C. *Mathematical Model*

Model:
Let S be programmers perspective for proposed system for problem. where,

$S = \{s,e,X,Y,f_{main},f_f,DD,NDD,MEM_{shared}|Ø\}$

Where, it will have meaning respective with system. s, e will be respectively start and end of system. X denotes input for system, Y indicates output got from system. f_{main} will show functions contribution done by author. f_f will show used library functions. DD represents Deterministic data working and NDD represents non-deterministic data. denotes Ø constraints used into system.

Let X be the input (firewall log)

X = L

$L = \{x_1, x_2, .., x_n\}$

Where, x denotes input features and n=41

Let Y be the output (Anomaly detected data)

Y = {AD}

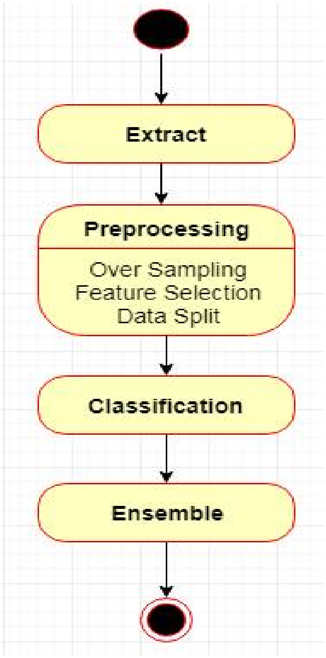
Where, AD is Anomaly Detected
Functional Division:

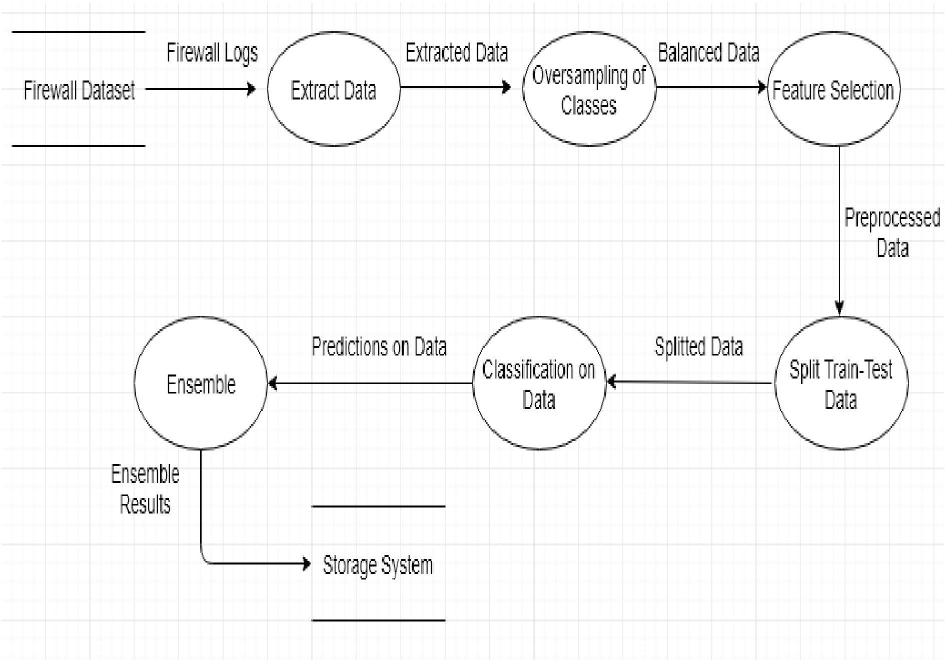
f_{main} = Anomaly detection
$f_{extract}$ = To extract data
$f_{classify}$ = F { Normal, DOS, Probe, U2R, R2L | labels }
$f_{malicious}$ = F { Attack/Normal | NB, DT, GB, MLP }
$f_{ensemble1}$ = { Malacious(x) | NB(x)=T & DT(x)=T || NB(x)=T & MLP(x)=T || DT(x)=T & MLP(x)=T }
$f_{ensemble2}$ = { Malacious(x) | NB(x)=T & DT(x)=T || NB(x)=T & GB(x)=T || DT(x)=T & GB(x)=T }

D. *Flow chart and Data Flow Diagram:*

System is initially in ideal state. Data is extracted from firewall logs after which it is given for preprocessing. In preprocessing task over sampling, feature selection and data split is performed. Classification is done on preprocessed data and Ensemble is performed on classified data.

In DFD, oversampling is performed on classes of extracted data. Important features are selected and data is split into train-test. Classification is performed and results of it are saved. Results of classification are then combined and used for ensemble, these ensemble results are stored in storage system.





IV. RESULT TABLE AND DISCUSSION

Dataset description:
KDD Cup '99 dataset is used which contains 41 features and 21 classes. The dataset is downloaded from the below link.
http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

Features:
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 outboundcmds, 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 ho-st 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.

Class Labels:
normal, buffer overflow, loadmodule, perl, neptune, smurf, guess passwd, pod, teardrop, portsweep, ipsweep, land, ftp write, back, imap, satan, phf, nmap, multihop, warezmaster, warezclient, spy, rootkit.

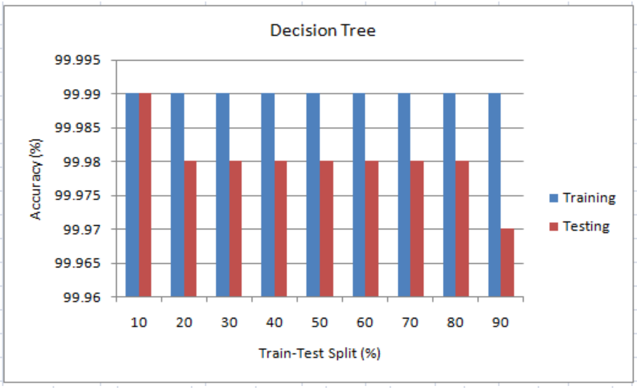
Classified Attacks:
Probe : ipsweep, nmap, portsweep, satan.
DOS : back, land, neptune, pod, smurf, teardrop.
U2R : buffer overflow, loadmodule, perl, rootkit.
R2L : ftp write, guess passwd, imap, multihop, phf, spy, warezclient, warezmaster.

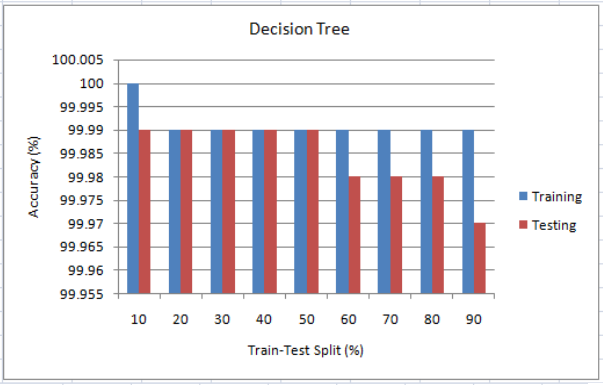
Preprocessing is carried out on dataset to reduce the number of features and use only the important one. This is done by calculating feature score and selecting top important features.
To evaluate performance of the system train dataset is split into train and test data. Testing is carried out on trained as well as test data. This is done on both preclassified data and classified.

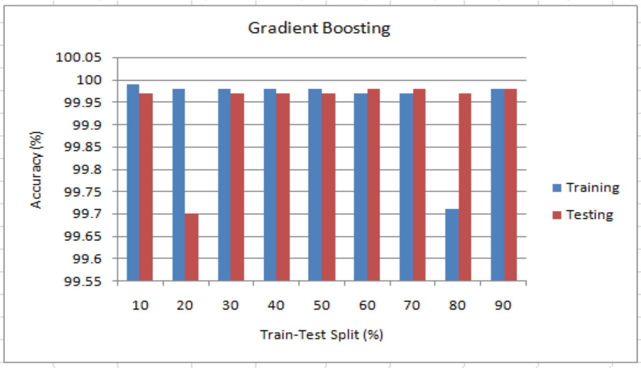
Ensemble is performed on classified data, where results of classified DT, NB, GB and MLP are used. Ensemble-1 gives results by combining DT, NB and MLP and Ensemble-2 gives results by combining DT, NB and GB.

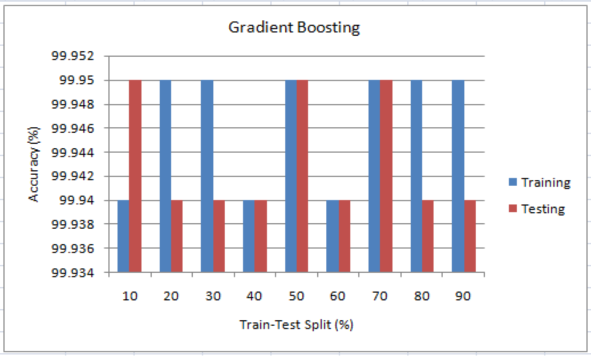
From the below mentioned result tables DT gives highest accuracy that is 99.99% on training data and 92.39% on test data as compared to NB, GB and MLP. Proposed Ensemble approach gives 92% accuracy which is similar to that of DT.

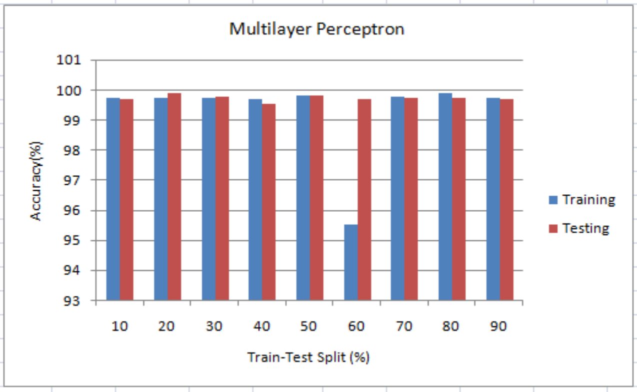
Results represented in confusion matrix, shows Ensemble-1 has 249 False Negative Rate which is lowest amongst all as compared to Ensemble-2 which is 264.

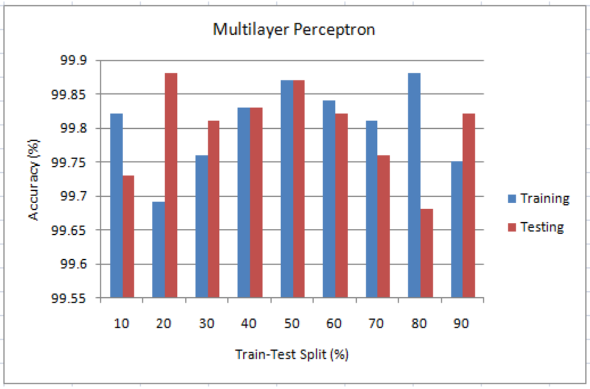

Decision Tree
99.995
99.99
99.985
99.98
99.975
99.97
99.965
99.96
Accuracy (%)
10
20
30
40
50
60
70
80
90
Train-Test Split (%)
Training
Testing

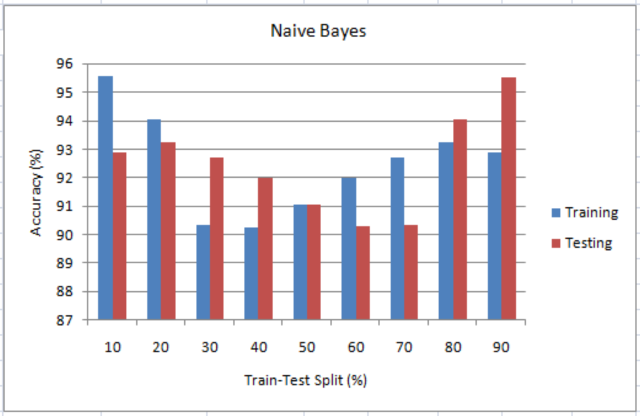

Decision Tree
100.005
100
99.995
99.99
99.985
99.98
99.975
99.97
99.965
99.96
99.955
Accuracy (%)
10
20
30
40
50
60
70
80
90
Train-Test Split (%)
Training
Testing

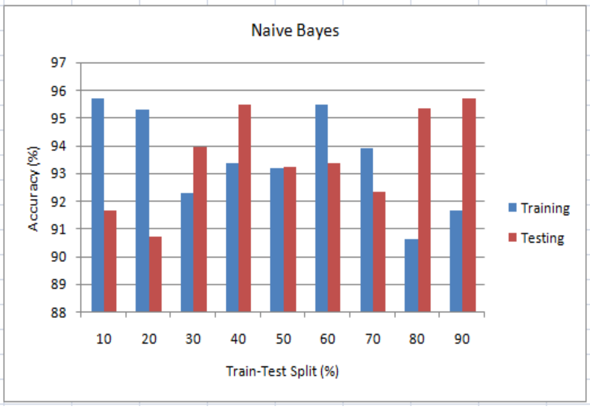

Gradient Boosting
100.05
100
99.95
99.9
99.85
99.8
99.75
99.7
99.65
99.6
99.55
Accuracy (%)
10
20
30
40
50
60
70
80
90
Train-Test Split (%)
Training
Testing

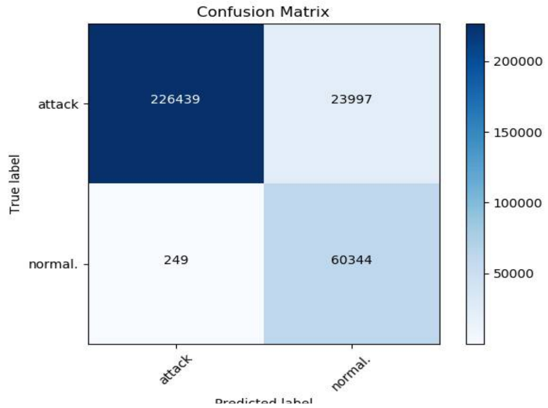

Gradient Boosting
99.952
99.95
99.948
99.946
99.944
99.942
99.94
99.938
99.936
99.934
Accuracy (%)
10
20
30
40
50
60
70
80
90
Train-Test Split (%)
Training
Testing

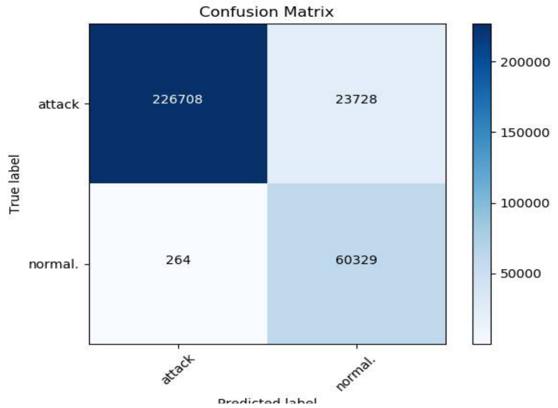

Multilayer Perceptron
101
100
99
98
97
96
95
94
93
Accuracy(%)
10
20
30
40
50
60
70
80
90
Train-Test Split (%)
Training
Testing


Multilayer Perceptron
99.9
99.85
99.8
99.75
99.7
99.65
99.6
99.55
Accuracy (%)
10
20
30
40
50
60
70
80
90
Train-Test Split (%)
Training
Testing


Naive Bayes
96
95
94
93
92
91
90
89
88
87
Accuracy (%)
10
20
30
40
50
60
70
80
90
Train-Test Split (%)
Training
Testing


Naive Bayes
97
96
95
94
93
92
91
90
89
88
Accuracy (%)
10
20
30
40
50
60
70
80
90
Train-Test Split (%)
Training
Testing





V. CONCLUSION

In this paper four different Machine Learning algorithms from supervised learning category are implemented. From the experiments done on firewall log dataset, it is observed that the proposed technique can identify both known as well as unknown attacks. Though individual techniques give good results, it suffers from False Negative Rate. To reduce this False Negative Rate, the proposed work has used two ensembles. These two ensembles are implemented and it is seen that Ensemble-1 consisting DT, NB and MLP results into minimum False Negative Rate.

REFERENCES

1. S. Uma, M. Suresh, "Security enrichment in intrusion detection system using classifier ensemble," Journal of Electrical and Computer Engineering, Vol. 2017, pp. 1-6, 2017.
2. R. Anderson, "Security Engineering, A guide to building dependable distributed systems," 1st edition, p. 325, 2003.
3. C. Pfleeger, S. Pfleeger, "Security in Computing," 4th edition, Prentice Hall, p. 484, 2008.
4. A. Sahasrabuddhe, S. Naikade, A. Ramaswamy, B. Sadliwala, P. Futane, "Survey on intrusion detection system using data mining techniques," International Research Journal of Engineering and Technology (IRJET), vol. 04, no. 05, pp. 1780-1784, 2017.
5. A. Abdel-Aziz, A. Hassanien, "Machine Learning techniques for anomalies detection and classification," Springer Conference of Advances in security of information and communication networks, pp. 219-229, 2013.
6. Gradient Boosting, [Online]. Available : https://medium.com/mlreview/gradie nt-boosting-from-scratch-1e317ae4587d, [Accessed : 10-Jun-2018].
7. Y. Zhang, Z. Han, J. Ren, "A network anomaly detection method based on relative entropy theory," IEEE Second International Symposium on Electronic Commerce and Security, pp. 231-235, 2009.
8. M. Jianliang, S. Haikum, B. Ling, "The application on intrusion detection based on k-means cluster algorithm," IEEE International Forum on Information Technology and Applications, pp. 150-152, 2009.
9. J. Dromard, G. Roudière, P. Owezarski, "Online and Scalable Unsupervised Network Anomaly Detection Method," IEEE Transactions on Network and Service Management, vol. 14, no. 01, pp. 34-47, 2017.
10. M. Prerau, E. Eskin, "Unsupervised anomaly detection using an optimized k-nearest neighbors algorithm," Article, pp. 1-14.
11. R. Krishnan, N. Raajan, "An enhanced multilayer perceptron based approach for efficient intrusion detection system," International Journal of Pharmacy and Technology, vol. 08, no. 04, pp. 23139-23156, 2016.
12. R. Kumari, Sheetanshu, M. Singh, R. Jha, N. Singh, "Anomaly detection in network traffic using k-means clustering," IEEE 3rd Intnternational Conference on Recent Advances in Information Technology, pp. 1-7, 2016.
13. S. Shakya, S. Sigdel, "An approach to develop a hybrid algorithm based on support vector machine and naive bayes for anomaly detection," IEEE International Conference on Computing, Communication and Automation (ICCCA), pp. 323-327, 2017.