International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 09 | Sep 2020 www.irjet.net p-ISSN: 2395-0072

Performance Analysis of Different Machine Learning Techniques for Anomaly-based Intrusion Detection

Ambreen Sabha[1], Lalit Sen Sharma[2]

[1]*M.Tech. Student, Department of Computer Science and IT, University of Jammu, J&K, India*
[2]*Professor, Department of Computer Science and IT, University of Jammu, J&K, India*

Abstract - *An Intrusion is an activity that compromises the confidentiality or the availability of the resource. An Intrusion Detection System is a device or the software that monitors the state of the network for any unauthorized access or any policy violations. The objective of the current research work is to compare the performance between different machine learning techniques using an anomaly-based intrusion dataset. For the proposed study, three supervised machine learning techniques namely Naïve Bayes, Decision Tree, and Random Forest have been applied to the dataset. To assess the performance of each machine learning technique; four parameters namely accuracy, recall, precision, and f-score have been evaluated. Experimentation is performed on the NSL-KDD dataset, which is based on the different sets of features. The detection accuracy and the execution time taken by the machine learning algorithms are analyzed. Random Forest obtained the highest accuracy of 97.8% and execution time of 0.998 milliseconds compared to that of the Decision Tree and Naïve Bayes. The detection accuracy of all the four attacks which were present in the dataset is DoS 99%, Probe 99%, R2L 98%, and U2R 99%, using the proposed research machine learning algorithm as Random Forest.*

Key Words: **Decision Tree, IDS, Machine Learning, NSL-KDD, Naïve Bayes, Network Security, Random Forest.**

1. INTRODUCTION

The Intrusion detection system (IDS) is a device or software which monitors the network for any malicious activity. An IDS is a tool that works with the network to keep it secure and alert when somebody is trying to break into your system [6]. Intrusion detection is the problem of identifying unauthorized use and abuse of computer systems by system insiders and external intruders and it is the process of detecting malicious patterns in the large data sets. Intrusion detection systems is classified into two different categories as **Host-based** intrusion detection system i.e. HIDS and **Network-based** intrusion detection system i.e. NIDS. Host-based IDS runs in any individual host or device. HIDS monitor only the inbound and outbound packets in the network traffic and when suspicious or harmful activities are identified it sends the alert to the administrator [6]. Whereas Network-based IDS monitors, capture and analyze the data packets in the network traffic. A typical network-based IDS makes use of Signature detection and Anomaly detection.

Signature-based IDS are designed to detect only known attacks and it uses a database of known attack signatures which is developed by the experts or intrusion analysts. The Signature detection monitors the packets in the network and compared them to the known signature or entries in this database. If there is a match, the IDS generates an alert message. **Anomaly-based IDS** looks for the kinds of unknown attacks that signature-based IDS, find hard to detect, and they function on the assumption that attacks are different from "normal" activity and can, therefore, be detected by the systems.

This research paper is organized as: Section 2 gives a brief Literature Review, Section 3 explains the Research Methodology, Section 4 shows the Experimental Results using the NSL-KDD dataset and Section 5 includes the Conclusion and Scope for future work.

2. LITERATURE REVIEW

A review of different Machine Learning techniques in the field of intrusion detection systems from the past few years is presented as under.

S. Revathi et.al. [10] published a paper on detailed analysis on the various intrusion dataset i.e. DARPA98, KDD-cup99, and NSL-KDD. They focused on the NSL-KDD dataset which contains only selected records, and those selected records provide a good analysis of various machine learning techniques for intrusion detection. NSL-KDD improves the accuracy of the system and reduces the false positive rate compared to that of DARPA98 and KDD99.

S. Taruna R et.al. [8] proposed a new method of Naïve Bayes Algorithm i.e. Enhanced Naïve Bayes. The results showed that the proposed algorithm more efficiently detects the intrusions, compared to the neural network and it also improved the detection rate and reduces the false positive rate. The experimentation was performed using the KDD-cup99 dataset.

Balogun et.al. [13] proposed a hybrid classification algorithm based on decision tree and K-Nearest neighbour. Firstly, the node information is generated according to the rules generated by the Decision Tree and then this node information is passed through KNN to obtain the final output. The results showed that the hybrid classifier (DT-KNN) gives the best result in terms of accuracy and efficiency when compared with the individual base classifiers i.e. decision tree and KNN.

Gaikwad and Thool [15] proposed a Decision Tree Algorithm using the NSL-KDD dataset and the Genetic Algorithm (GA) is used for the feature selection process. The GA selects 15 features out of all the 41 from the dataset and gave accuracy of 79% on the test data using the decision tree classifier. The execution time taken by the classifier to build the model is 176 seconds.

Kajal Rai et.al. [9] proposed a decision tree split (DTS) algorithm based on the C4.5 decision tree approach. Feature selection and split value are important issues for the construction of DTS. The proposed algorithm performed better when compared with the existing tree algorithms such as Classification and Regression Tree (CART), C4.5, and AD Tree. The DTS algorithm was implemented using tools WEKA and MATLAB.

M. Gupta et.al. [4] proposed the J48 decision tree algorithm. The proposed J48 algorithm gave higher accuracy of 99.73% over other machine learning algorithms i.e. Naïve Bayes and SVM.

Revathi and Malathi et.al. [10] published a paper, which performed a comparative analysis of machine learning algorithms such as Random Forest, C4.5 decision tree, SVM, CART, and Naïve Bayes. It selects 15 features using the Correlation-based Feature Selection (CFS) technique. The experimental results of the above-mentioned algorithms have been compared and the outcome shows that Random Forest gave the highest accuracy of 98% in detecting the attacks.

A. Nur Cahyo et.al. [3] performed a comparative analysis of different machine learning algorithms i.e. Artificial neural network (ANN) and Support Vector Machine (SVM). ANN obtained high accuracy in all categories compared to that of SVM and it uses all the features of the dataset. The detection rate of DoS is 92.20%, probe is 90.60%, R2L is 89%, and that of U2R is 90.80%, and it showed that the performance of ANN is better than SVM.

A. Abd Ali Hadi et.al. [12] proposed the Random forest algorithm to classify the network data. The Information gain method is used as a feature selection process, the 13 most significant features were generated from the original set of 41 features. The accuracy of the proposed model is 99.33%, and it performs better than the existing machine learning classifiers. The implementation of the proposed Random Forest algorithm was done using WEKA and MATLAB.

Shilpashree. S. et.al. [2] published a paper that measured the performance of the intrusion detection system by applying the machine learning techniques based on decision trees. The Bayesian three modes were analyzed for different sizes of datasets. The Multinomial naïve Bayes gets the least computation time than Bernoulli naïve Bayes, and Gaussian naive Bayes is the last one among all the test cases. Information gathering is obtained through, some network capturing devices, such as Libdump, TCPdump, and Wireshark. The accuracy and execution time taken by the classifier to build the model is analyzed.

3. RESEARCH METHODOLOGY

The main objective of the current research work is to detect the normal behaviour or attacks in the test data using Machine learning techniques and it also finds the detection accuracy of all the four types of attack which were present in the NSL-KDD dataset. The methodology steps which are used to build the model are shown in Fig.1 and steps are described below:

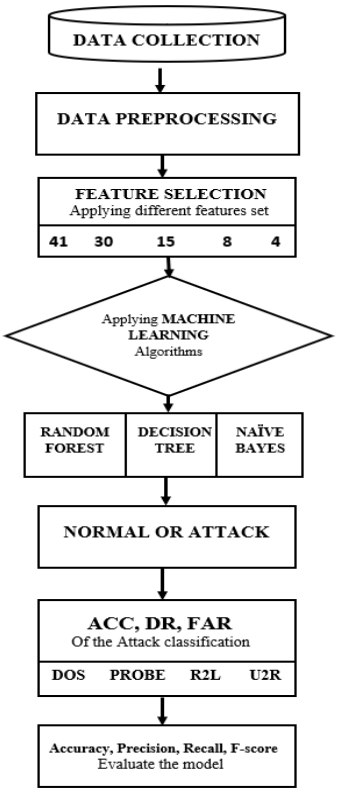

Fig –1: Flowchart of the current research work

To attain the objective of the proposed research work experimentation has been carried out using the NSL-KDD (Network Security Laboratory Knowledge Discovery and Data Mining) dataset, which is the revised version of KDD-Cup99.

3.1 Data Collection

In the research work, the NSL-KDD dataset is used and is downloaded from the **"Canadian Institute for Cybersecurity (//www.unb.ca/cic/)"**. The reason for using the NSL-KDD dataset for the research purpose is that the KDDcup99 data set has a large number of redundant and duplicate records in the training and testing dataset [10], which causes the machine learning algorithms to be biased towards the frequent records. NSL-KDD dataset consists of 42 features out of which 4 are symbolic(categorical) features and 38 are digital(numeric) features.

Table –1: Types of features

Type of Features	Features present in the NSL-KDD dataset
Symbolic features	Protocol_type, Service, Flag, Class
Numeric features	Duration, src_bytes, dst_bytes, wrong_fragment, urgent, hot, num_failed_logins, num_compromised, root_shell, su_attempted num_root, num_file_creations, num_shells, num_access_files, num_outbound_cmds, Land, logged_in, 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

3.2 Data Pre-processing

After Data collection next is the data pre-processing phase. This phase deals with the missing, noisy and inconsistent values in the dataset, but after checking the dataset, it shows that NSL-KDD doesn't contain any missing and noisy values. Min-max normalization is used for normalizing the data and normalization has been done to transform all the input attributes in the range [0,1]. Label Encoder has been performed to transform the categorical features to numeric by removing categories (alphabetical) from all the attributes to make the dataset whole numeric.

Next, the experiments have been carried out using the **Training** and **Testing** set. The NSL-KDD dataset classifies the network traffic into two classes, namely, Normal and Anomaly. The training set consist of 41 attributes and one label or class and the testing set has only 41 attributes and no label or class. The experiments were performed on training data set having 25000 records (or rows) out of which 13348 are normal and rest 11652 are Anomaly and test data having 12000 records out of which 5182 are normal and rest 6818 are Anomaly. There are records in the testing set that are not in the training set to ensure the prediction quality of the proposed machine learning model. The attacks are further classified into four types which were present in the NSL-KDD dataset i.e. Dos, Probe, R2L, and U2R. The training and testing instance are shown in Table 2 below:

Table –2: Number of training and testing records

NSL-KDD		Training instances	Testing instances
Normal		13348	5182
ANOMALY	DOS	9160	3986
	Probe	2274	1256
	R2L	207	1538
	U2R	11	38
Total		**25000**	**12000**

3.3 Feature Selection

This phase is based on the selection of attributes from the dataset and these selected attributes were used in the proposed model to check the performance of the machine learning model. It shows that the accuracy of mostly all the algorithms, namely, Random Forest, Decision Tree, and Naïve Bayes increases using feature selection technique compared to that of without feature selection. The comparison between with and without feature selection is shown in Table 3 below.

Table –3: Comparison between with and without feature selection techniques

Algorithm	With feature selection	Without feature selection
Random Forest	0.978	0.915
Decision Tree	0.941	0.916
Naïve Bayes	0.824	0.837

For the different feature sets, different techniques were used for the selection of features, to check the performance that, which set predicted the highest accuracy compared to other

sets. The different feature selection techniques were described below.

3.3.1 Variance threshold:

The first feature selection technique used is the Variance Threshold. This technique is used for dropping features with variance below threshold variance. By default, it removes all zero-variance features, i.e. features that have the same value in all the records.

$$Var[X] = p(1-p)$$

Where, p is threshold variance.
And for this study, threshold variance is 0.09 i.e. 0.9*(1-0.9) =0.09. It drops those features which are the same 90% of the time, and it selects **30** features out of all the 41 features present in the dataset and gives an accuracy of 97.8%.

3.3.2 Univariate feature selection:

The second technique used is univariate feature selection and this technique works by selecting the best features based on univariate statistical tests. In this technique each feature is compared to the target variable, to see whether there is any statistically significant relationship. The Univariate feature selection technique uses two tests, namely, Chi-square and Anova.

chi-square statistical test examines each feature individually to determine the strength of relationship of the feature with the target variable. In this test the SelectKBest method is used, which removes all features but only the specified number of the highest-scoring feature are selected. The value of k is 15 and it selects the 15 relevant features and gives an accuracy of 92.8%.

Anova f-test method is a statistical test and it is used when one variable is numeric and the other is categorical i.e. numerical input variable and categorical target variable. It captures the linear dependency between the variables and those features that are independent of the target variable can be removed and those that are dependent on target variable are selected and the value of k for this test is 4. It selects 4 most relevant features that are dependent on the target variable and gives an accuracy of 87.4%.

3.3.3 Mutual information gain:

This technique calculates the mutual dependency between the variables i.e. amount of information obtained about one variable through the other with numeric input and categorical output. It is equal to zero if the variables are independent, and a higher value means higher dependency. It selects those features having the highest score and are highly dependent on each other, 8 relevant features having the highest value are selected from all the 41 features and gives an accuracy of 93.6%.

3.4 Model Selection

This is the process of selecting or choosing a model between different machine learning approaches for the proposed research work. The experimental results have been carried out using different machine learning techniques i.e. Random Forest, Decision Tree, and Naïve Bayes.

3.4.1 Random Forest:

Random Forest is the Supervised Ensemble Machine Learning algorithm, which is used for both classifications as well as regression. Random forest is made up of a bundle of decision trees and more trees mean a more robust forest. This Random forest algorithm gets a prediction from each of the individual Decision tree classifiers and finally selects the best solution by means of voting.

For the experimental set-up, different feature sets were used, namely, variance threshold, univariate statistical test, and mutual information gain to check the accuracy of the Random Forest algorithm. And from the result, it is shown that random forest gives better accuracy using the variance threshold feature selection technique, and 30 best features were selected from this technique and it gives the highest accuracy of 97.8% compared to that of other sets.

3.4.2 Decision Tree:

Decision Tree is the supervised machine learning algorithm and is mostly used in classification problems. On the basis of the split the decision tree uses multiple algorithms to split the node into two or more sub-nodes and then selects the split which results in a more homogenous sub-node.

For the current experimental research study Decision tree algorithm has been implemented using the variance threshold technique, which performs better than other techniques i.e. univariate test and mutual information gain and it selects 30 features and gives the highest accuracy of 94.1% compared to that of other features.

3.4.3 Naïve Bayes Classifier:

Naïve Bayes is the probabilistic supervised machine learning algorithm that is mainly used in classification problems. It is based on the Bayes theorem and there is class conditional independence between every pair of features i.e. effect of the particular feature on the class is independent on the other features.

For the current research work, the Naïve Bayes classifier has been implemented using all the 41 features and it gives an accuracy of 82.4% compared to that of other sets.

4. EXPERIMENTAL RESULTS

The experiments have been carried out on the Windows platform and the editor used for the implementation is Jupyter Notebook.
Microsoft Excel 2019 is used for dataset preparation. And from the experimental result, it can be found that the Random Forest algorithm predicted the highest accuracy of 97.8% and execution time of 0.998 millisecond compared to that of Decision tree and Naïve Bayes having an accuracy of 94.1% and 82.4% and execution time of 0.999 milliseconds and 1.97 milliseconds. There are 4 main classes of attacks, which were presented in the NSL-KDD dataset i.e. DoS, Probe, R2L, and U2R. The accuracy (ACC), Detection rate (DR), and False alarm rate (FAR) of these attacks are shown in Table-IV below, using the proposed research algorithm as Random Forest. Figure 2 represents the comparison of the different types of attacks based on the detection accuracy and the false alarm rate.

Table -4: Experimental Results of types of Attack

Attack Types	Accuracy (ACC)	Detection rate (DR)	False alarm rate (FAR)
Dos	0.99858	0.99749	0.00142
Probe	0.99612	0.99247	0.00388
R2L	0.98155	0.97111	0.01845
U2R	0.99693	0.82888	0.00307

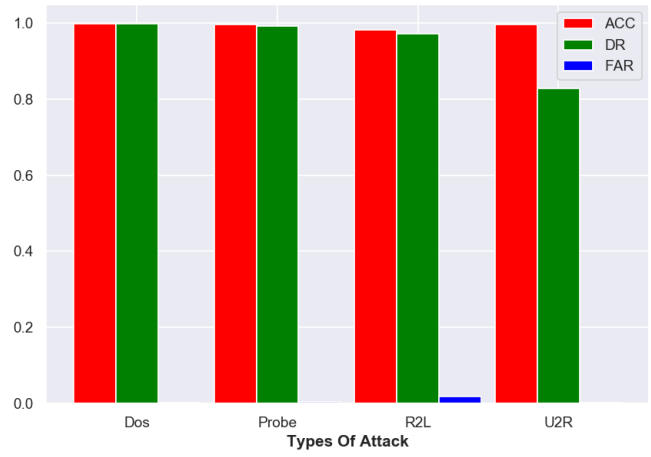

Fig –2: Comparison of different types of Attack

The comparative analysis of the machine learning techniques i.e. Random Forest, Decision Tree, and Naïve Bayes are shown in Table-V and the algorithms have been evaluated, based on parameters like Accuracy, Precision, Recall, and F-score and figure 3 shows the comparative analysis between different ML techniques.

Table -5: Experimental Results of ML Techniques

ML Algorithms	Accuracy	Precision	Recall	F-score
Decision Tree	0.941	0.94	0.94	0.94
Random Forest	0.978	0.97	0.98	0.97
Naïve Bayes	0.824	0.94	0.82	0.87

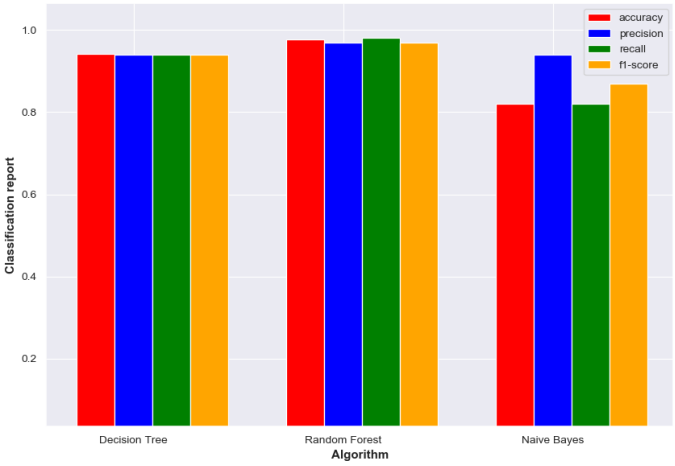

Fig –3: Comparison between the different Machine Learning Techniques

5. CONCLUSION AND FUTURE SCOPE

From the experimental results and the evaluation based on the proposed research model, we conclude that NSL-KDD is one of the best datasets used in the intrusion detection system. From the experimental set up of the proposed research work, we conclude that, among the other machine learning techniques Random Forest found to be the best algorithm in detecting the attacks for anomaly intrusion detection, which improves the detection rate and reduces the false alarm rate. For future work, the researchers recommend the use of different feature selection and extraction techniques. And to use the real-time dataset for the detection of intrusion using machine learning techniques and integrate the characteristics of modern deep learning algorithms to form the comparative analysis between machine learning and deep learning techniques.

REFERENCES

[1] M. Tabash, M. A. Allah, B. Tawfik **"Intrusion Detection Model Using Naive Bayes and Deep Learning Technique"** International Arab Journal of Information Technology, 2018.

[2] Shilpashree. S, S. C. Lingareddy, N. G. Bhat, S. Kumar G **"Decision Tree: A Machine Learning for Intrusion Detection"** International Journal of Innovative Technology and Exploring Engineering, Vol. 8, Issue. 6, pp. 1126-1130, 2019.

[3] A. N. Cahyo, R. Hidayat, D. Adhipta **"Performance Comparison of Intrusion Detection System based on Anomaly Detection using Artificial Neural Network and Support Vector Machine"** Advances of Science and Technology for Society AIP Conf. Proc. 1755, pp. 1-7 ,2017.

[4] M. Gupta, J. Shriwas, S. Farzana **"Intrusion Detection Using Decision Tree Based Data Mining Technique"** International Journal for Research in Applied Science & Engineering Technology**,** Vol. 4, Issue. 7, pp. 24-28, 2016.

[5] R. Wankhede, V. Chole, S. Kolte **"A Review on Intrusion Detection System Using Classification Technique"** International Journal of Advanced Computational Engineering and Networking, Vol.3, Issue 12, pp. 62-65, 2015.

[6] "DNS stuff. **Intrusion detection system". Available at:** https://www.dnsstuff.com/intrusion-detection system Accessed on October 2019.

[7] A. Juneja, "**Dzone. Machine learning algorithm for Intrusion detection system"**. Available at : https://dzone.com/articles/evaluation-of-machine-learning-algorithms-for-intrusion Accessed on May 2019.

[8] S. Taruna R., S. Hiranwal **"Enhanced Naïve Bayes Algorithm for Intrusion Detection in Data Mining"** International Journal of Computer Science and Information Technologies Vol. 4, Issue. 6, pp. 960-962, 2013.

[9] K. Rai, M. S. Devi, A. Guleria **"Decision Tree Based Algorithm for Intrusion Detection"** International Journal of Advanced Networking and Applications Vol. 07, Issue. 4, pp. 2828-2834, 2016.

[10] S. Revathi, A. Malathi **"A Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection"** International Journal of Engineering Research & Technology, Vol.2, Issue. 12, pp. 1848- 1853, 2013.

[11] R. R. Devi, M. Abualkibash **"Intrusion Detection System Classification using different Machine Learning Algorithms on KDD-99 and NSL-KDD datasets - A Review Paper"** International Journal of Computer Science & Information Technology, Vol.11, Issue. 3, pp. 65-80, 2019.

[12] A. A. Ali Hadi **"Performance Analysis of Big Data Intrusion Detection System Over Random Forest Algorithm"**, International Journal of Applied Engineering Research, Vol.13, Issue. 2, pp. 1520-1527, 2018.

[13] Balogun, A. O., Jimoh, R. G. **"Anomaly Intrusion Detection Using an Hybrid of Decision Tree And K-Nearest Neighbor"**, A Multidisciplinary Journal Publication of the Faculty of Science, Vol. 2, pp. 67-74, 2015.

[14] S. Omar, A. Ngadi, H. H. Jebur **"Machine Learning Techniques for Anomaly Detection"** International Journal of Computer Applications, Vol. 79, Issue. 2, pp. 33-37, 2013.

[15] D.P. Gaikwad, S. Jagtap, K. Thakare, V. Budhawant **"Anomaly Based Intrusion Detection System Using Artificial Neural Network and Fuzzy Clustering"** International Journal of Engineering Research & Technology, Vol. 1, Issue. 9, pp. 1-6, 2012.

[16] O. P. Akomolafe, A. I. Adegboyega **"An Improved KNN Classifier for Anomaly Intrusion Detection System Using Cluster Optimization"** International Journal of Computer Science and Telecommunications Vol. 8, Issue. 2, 2017.

[17] U. Kumari, U. Soni "**A Review of Intrusion Detection using Anomaly based Detection"** in Proceedings of IEEE 2nd International Conference on Communication and Electronics Systems, pp.824-826, 2017.

[18] V. Jyothsna, V. V. R. Prasad "**A Review of Anomaly based Intrusion Detection Systems"** International Journal of Computer Applications, Vol. 28, No.7, pp. 26-35, 2011.

[19] M. Kumar, M. Hanumanthappa, T.V. Suresh Kumar **"Intrusion Detection System Using Decision Tree Algorithm"** IEEE, pp.629-634, 2012.

[20] V. D. Mane, A. Sayar, S. Pawar "**Anomaly Intrusion Detection System Using Neural Network"** International Journal of Computer Science and Mobile Computing Vol.2, Issue. 8, pp. 76-81, 2013.

[21] K. Bajaj, and A. Arora, **"Improving the Intrusion Detection using Discriminative Machine Learning Approach and Improve the Time Complexity by Data Mining Feature Selection Methods"**, International Journal of Computer Science, vol. 76, 2013.

[22] R.M. Elbasiony, E.A. Sallan, T.E. Eltobely, and M.M. Fahmy, **"A hybrid network intrusion detection framework based on random forests and weighted kmeans",** Ain Shams Engineering Journal, vol.4, Issue 4, Dec, 2013, pp. 753-762.

[23] D.P. Gaikwad, and R.C. Thool, **"Intrusion Detection System using Ripple Down Rule learner and Genetic Algorithm",** International Journal of Computer Science and Information Technologies, vol. 5, 2014, pp. 6976-6980.

[24] L.M. Ibrahim, D.T. Basheer, and M.S. Mahmod, **"A comparison study for intrusion database (KDD99, NSLKDD) based on Self Organization Map (SOM) Artificial Neural Network"**, Journal of Engineering Science and Technology, vol. 8, No. 1, 2013, pp. 107-119.

[25] Bhavsar Y. B. and Waghmare K. C., **"Intrusion Detection System Using Data Mining Technique: Support Vector Machine,"** International Journal of Emerging Technology and Advanced Engineering, Vol. 3, Issue. 2, pp. 581-586, 2013.

AUTHORS BIOGRAPHY

Ambreen Sabha is a Master of Technology (M. Tech) student in the Department of Computer Science & IT, University of Jammu, J&K India. She has received her Bachelor in Engineering degree (B.E) in Computer Science from Model Institute of Engineering and Technology (MIET) Kot-Bhalwal Jammu. Her areas of research include Networking, Network Security and Machine Learning.

Lalit Sen Sharma has received his Master's degree in Mathematics and Computer Applications from Guru Nanak Dev University, Amritsar, Punjab, India. He has also received his Doctorate of Philosophy (PhD) in Computer Science and Engineering from Guru Nanak Dev University. He has been teaching to post graduate students in computer applications of University of Jammu for more than 20 years. Currently, he is working as a Professor and Head of the department of Computer Science and Information Technology in the University of Jammu, India. He is specialized in Data Communication and Network, Internet and WWW and Data Structures. He is a member of Indian Science Congress Association, Computer Society of India, Institute of Electronics and Communication Engineers, and National HRD Network, India.