Intrusion Detection Using Supervised Learning Methods

Khushi[1], S. Vaishnavi[2], Yashvi Gupta[3] , Shelly Gupta[4]

Student[1, 2, 3] , Assistant Professor[4]
Department of Computer Science and Engineering,
Inderprastha Engineering College,
Ghaziabad, Uttar Pradesh, India

Abstract - If internet is helping people run their daily errands, it is also opening up the door to large number of intrusions, frauds, attacks such as phishing, viruses, trojans, etc. A lot of services are availed worldwide by various organizations, firms, schools, colleges, which in turn generates a lot of network traffic. This gives a chance to intruders to breach into the system. To safeguard individual's information and to also comply with the CIA triad, Intrusion Detection System (IDS) plays an important role. An IDS is a security system that monitors the network and alerts whenever an unauthorized access is made. To build a more effective system, machine learning techniques are employed. This paper provides a comprehensive study of NSL-KDD dataset using certain machine learning algorithms namely, Support Vector Machine (SVM), Random Forest (RF), and Extreme Learning Machine (ELM). In the experiment, a model is built making use of these algorithms, and is trained and tested on aforementioned dataset. A comparison is drawn between the algorithms on the basis of evaluation metrics. The results showed SVM having highest accuracy as well as precision score, and elm having highest recall score**.**

Key Words: Intrusion Detection, NSL-KDD dataset, Support Vector Machine**,** Random Forest**,** Extreme Learning Machine, Evaluation metrics.

1. Introduction

In a world where internet has become a crucial asset in day-to-day life, not much has to be done to entice people into giving up their information or jeopardizing somebody's information. The number of internet users is rising year by year with internet being the medium of communication between them. As per Internet World Stats, China, India, and USA are three countries having highest number of users, with India having approximately 560 million online users, and yet a large part of the country is still offline. Up until January 2021, there were 4.66 billion active internet users worldwide which accounts for 59.5 percent of the global population. Of this total, 92.6 percent (4.32 billion) accessed the internet via mobile devices. This itself speaks for the fact that a lot of private, confidential, or sensitive information is at stake with intruders having potential to gain illegal access to the system. A recent case of data breach showed up in April 2020 where credentials of around 500,000 Zoom accounts were up for sale on dark web.

To resolve these kind of issues, Intrusion Detection Systems come to an aid. It is a system or device that looks for any malicious activity happening into the system, and if there is, it generates an alert. Based on these alerts, appropriate steps are taken to remove the threat. The era of IDS started when a pioneer of Information Security department in U.S. Air Force produced a report on "Computer Security Threat Monitoring and Surveillance". After that, the first model was built that constantly monitored and compared network traffic against a list of known threats. The world then began using IDS as its saviour.
Intrusion Detection System follows different methods to detect malicious activities. First is Signature based IDS. A network consists of several data packets. These data packets consist of pattern or signature. An attacker would want to insert some malicious code into one of these packets. The newly generated packet is called as attack pattern or attack signature. IDS database already contains known intrusion attack signatures. Whenever a suspicious activity happens, the new attack pattern is matched with the existing data. If IDS finds a match, it recognizes it as an intrusion.

Second is Anomaly based IDS. It detects when system deviates from its normal behaviour. System is first trained with a normalized baseline and then the activity is compared against that baseline. If system behaves abnormally, an alert message is generated.

IDS can be classified into two major types, namely Network Intrusion Detection System (NIDS) and Host Intrusion Detection System (HIDS). NIDS is network-based Intrusion Detection System. It monitors, captures, analyses the network traffic, and helps find malicious content into the traffic. Analysis is done by matching traffic to the library of known attacks. NIDS is generally deployed at points in the network where traffic is more vulnerable to attack. Since it manages a vast network, it sometimes become difficult to detect a suspicious activity. HIDS, however, is host based. It is installed on a host or device from where it monitors the traffic. It takes a snapshot of existing system files and compares it with the previous snapshot. If the analytical system files are edited or deleted, an alert is sent to the administrator to investigate.

However, traditional IDSs have some limitations. Number of false alarms raised are greater than the number of real attacks. This makes real attacks often go unnoticed. These systems have low accuracy, high false rate, are not easily modifiable, and cannot identify some new malicious attack. To build a more efficient system, machine learning (ML) techniques can be applied which are well known for their classification methodology. These techniques will help in enhancing the accuracy by reducing false alarm rates and increasing detection rate. The accuracy of the model will depend upon its detection rate. ML learns from the past data and then predicts the outcome using new data. This data is provided by the chosen dataset. The dataset used for this study is NSL-KDD dataset which is a variant of KDD99 dataset. Several ML algorithms are used to train the system. To find out the most efficient algorithm out of all, a comparison is drawn between these algorithms on the basis of their evaluation metrics.

2. Background and Related work

Intrusion Detection System has become a proficient device to deal with malwares and attacks. With the help of machine learning algorithms, there has been an improvement in the accuracy of the system. Several researches in the field of algorithms & datasets have been carried out in order to increase the system's performance **.**

Corinna Cortes and Vladimir Vapnik[1] presented support-vector network as a new learning machine for two-group classification problems. The input vectors were mapped non-linearly to a very high dimension feature space and a study was conducted on non-separable training data. The performance of the support-vector network was also compared to various classical learning algorithms.

One of the most important asset in this research is the availability of dataset. Machine learning algorithms make use of the dataset in order to train themselves. KDD'99 has been the most widely used dataset for anomaly detection methods. Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani[2] conducted a statistical analysis on this data set, and found two important issues which resulted in a poor performance of the system. To solve these issues, a new data set namely, NSL-KDD was proposed which consisted of selected records of the complete KDD data set that overcame the deficiencies of existing dataset. Another study on NSL-KDD dataset is done by Gerry Saporito[14] where an in depth description of the origination of dataset is given and a knowledge of what comprises in the dataset is provided.

Preeti Singh, Amrish Tiwari[5] discussed about various approaches developed by different researchers for the intrusion detection. The authors presented methods such as neural network, decision tree, hidden markov model etc. with their advantages and disadvantages. The main concern was to build a system that can efficiently reduce the false alarm rate and decrease the training time.

The NSL-KDD data set is a refined version of KDD"99 data set. L. Dhanabal, Dr. S.P. Shantharajah[6] conducted an analysis on NSL-KDD dataset and found it the best to test the performance of intrusion detection systems. The study also brought to light the fact that most of the attacks are caused as a result of inherent drawbacks of the TCP protocol.

Saroj Kr. Biswas[9] used feature selection technique to select a set of significant features from the original set and it is then used to train different types of classifiers to model the IDS. The classifiers used in this research were k-NN, DT, NN, SVM and NB, after which k-NN classifier turned out to be the best performer among all.

Chibuzor John Ugochukwu, & E. O Bennett[10] carried out the research to find out the efficacy of several machine learning algorithms namely, Bayes Net, J48, Random Forest, and Random Tree. The dataset and the experimental tool that were worked upon are KDDCup99 and WEKA, respectively. The results showed that the Random Forest and Random Tree algorithms are the most efficient in performing the classification technique on the Test dataset. The evaluation metrics used to evaluate the model are Precision, Recall and F-measure.

Quang-Vinh Dang[11] used CICIDS'2012 dataset to evaluate IDS's performance. The author, with his study, showed that the problem in this dataset can be solved by by introducing gradient boosting technique. The author also suggested that algorithm like Naive Bayes can still achieve a high predictive performance when given a proper training dataset. Also, the problem of intrusion detection with the given dataset can be solved with ensemble machine learning techniques, even with a small training dataset.

Similar study has been performed by Iftikhar Ahmad, Mohammad Basheri, Muhammad Javed Iqbal,and Aneel Rahim[13]. The authors looked for various machine learning techniques like Support vector machine, Random Forest, Extreme Learning Machine that are well known for their classification capability. The dataset that have been worked upon are NSL–knowledge discovery and data mining data set. The aim of the research is to build a system which can improve the system's accuracy by reducing false alarms and increasing detection rate.

3. Description of Dataset

NSL-KDD dataset is a revised, cleaned-up version of the KDD Cup'99 data set. KDD Cup'99 is widely used to build intrusion detection systems. It consists of 4,898,430 records which was pre-processed into 41 features per record, labelled as either normal or an attack. The features are categorized into four groups i.e., Basic Features, Content Features, Time based traffic features, and Host based traffic features, and contains four categories of attacks namely, Denial of Service Attack (DoS), User to Root Attack (U2R), Remote to Local Attack (R2L), and Probing Attack.

However, there are some problems with the dataset which can have an impact on the performance of the system. One of the major problem the large number of redundant record. Due to this, the learning algorithms becomes biased towards the frequent records, and prevent them from learning un-frequent records which are usually more harmful to networks such as U2R and R2L attacks. To solve this problem, NSL-KDD dataset was proposed. NSL-KDD dataset have some advantages over the original KDD data set. Firstly, it doesn't include irrelevant records in the train set, so the classifiers will not be partial towards more repeated records. Secondly, there are no duplicate records in the proposed test sets, hence, the classifiers are not biased towards more frequent records. The dataset consisted of 43 features per record, with 41 of the features referring to the traffic input and the last two called labels (whether it is a normal or attack) and Score (the severity of the traffic input itself).

Table -1: Classification of attacks in KDD data set

Attack Class	Description	Attack Type
DOS	DoS is an attack that tries to shut down traffic flow to and from the target system.	Neptune, land, pod, smurf, teardrop, back, worm, udpstorm, processtable, apache2, mailbomb.
Probe	Probe or surveillance is an attack that tries to get information from a network. The goal here is to act like a thief and steal important information, whether it be personal information about clients or banking information.	ipsweep, satan, nmap, portsweep, mscan, saint.
U2R	U2R is an attack that starts off with a normal user account and tries to gain access to the system or network, as a super-user (root). The attacker attempts to exploit the vulnerabilities in a system to gain root privileges/access.	buffer_overflow, loadmodule, perl, rootkit, ps, xterm, sqlattack.
R2L	R2L is an attack that tries to gain local access to remote machine .An attacker does not have local access to the network and tries to "hack" their way into the network.	ftp write, guess password, imap, multihop, phf, spy, warezclient, warezmaster, snmpguess, named, xlock, xsnoop, snmpgetattack, httptunnel, sendmail.

4. Support Vector Machine (SVM)

SVM is a supervised machine learning algorithm which can be used for classification or regression challenges, but is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well. In the SVM classifier, it is easy to have a linear hyper-plane between these two classes. But the real problem occurs when classes are not linearly separable.

For this, the SVM algorithm uses a technique called the kernel function. The SVM kernel is a function that takes low dimensional input space and transforms it to a higher dimensional space i.e. it converts non separable problem to separable problem. It is mostly useful in non-linear separation problem. Simply put, it does some extremely complex data transformations, then finds out the process to separate the data based on the labels or outputs you've defined. Some important terminologies in SVM includes:

Support vectors, these are data points that are closest to the hyper plane. Separating line will be defined with the help of these data points. Hyper plane, it is a decision plane or space which is divided between a set of objects having different classes. Margin, it is the gap between two lines on the closet data points of different classes. It can be calculated as the perpendicular distance from the line to the support vectors. The main goal of SVM is to divide the datasets into classes to find a maximum marginal hyper plane.

5. Random Forest (RF)

RF is a flexible easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because of its simplicity and the fact that it can be used for both classification and regression tasks. RF is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. A forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result. RF works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

6. Extreme Learning Machine (ELM)

ELM is feed forward neural networks for classification, regression, clustering, sparse approximation, compression and feature learning with a single layer or multiple layers of hidden nodes, where the parameters of hidden nodes need not be tuned. These hidden nodes can be randomly assigned and never updated or can be inherited from their ancestors without being changed. In most cases, the output weights of hidden nodes are usually learned in a single step, which essentially amounts to learning a linear model. These models are able to produce good generalization and learn thousands of times faster than networks trained using back-propagation. It can even outperform SVM in both classification and regression challenges.

7. Material and Methods

The various stages are pre-processing of the captured dataset, classification process, training and testing using required data, and finally, implementation evaluation. The training dataset is fed to the pre-processing system; the output from the pre-processing stage is sent to the classifier model. Thereafter, the test dataset is fed into the IDS system and the performance evaluation is conducted where accuracy scores for each of them are obtained. Then the data is visualized using charts and graphs for better understanding of data.

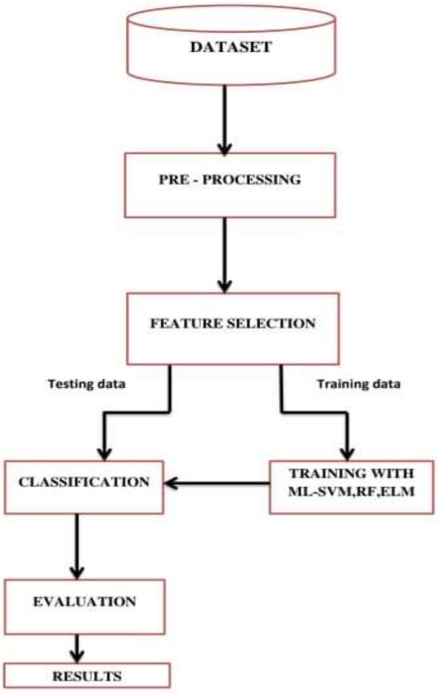

Fig -1: Methodolgy

8. Result and Analysis

The results obtained showed that SVM and ELM have highest accuracies as compared to RF whereas in case of precision, SVM had the highest score followed by RF. The recall score obtained was highest for ELM. A visual depiction for the comparison of algorithms is shown in the figures below in the form of graphs.

Table 2.Accuracy, Precision, Recall of the algorithms

Algorithms	Accuracy (in %)	Precision (in %)	Recall (in %)
SVM	83.94	97.43	69.73
RF	81.45	92.57	68.39
ELM	83.06	85.54	79.58

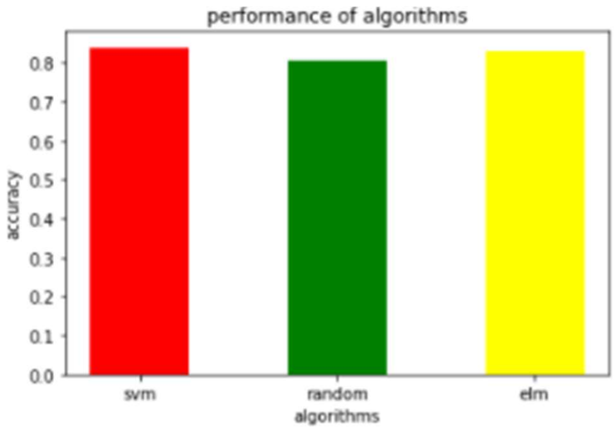

Figure 2.Accuracy comparison of algorithms

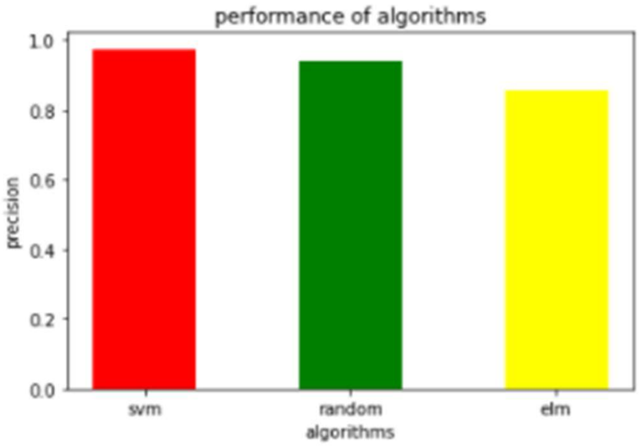

Figure 3.Precision comparison of algorithms

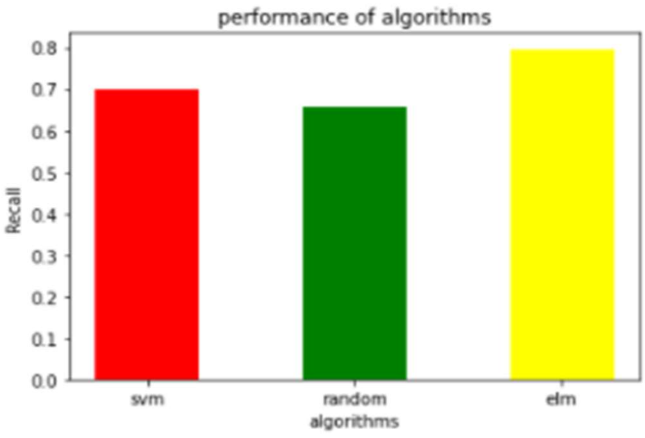

Figure 4.Recall comparison of algorithms

The **Receiver Operator Characteristic (ROC)** curve is an evaluation metric known for binary classification problems. It is a probability curve that plots the True Positive Rate (TPR) against False Positive Rate (FPR) at various threshold values and separates the 'signal' from the 'noise'. The **Area Under the Curve (AUC)** is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. ROC curve for SVM and RF with their AUC values is shown in the figures below.

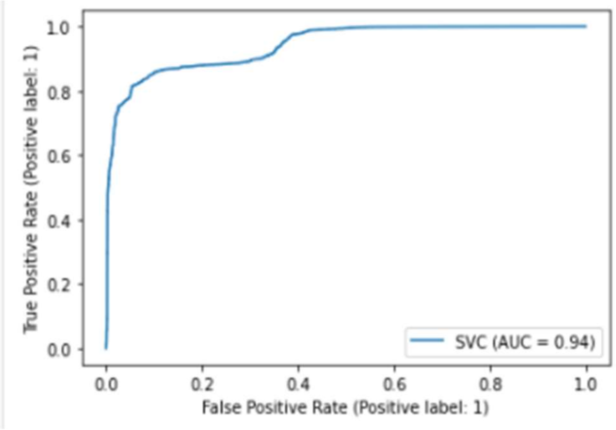

Figure 5.ROC curve for SVM

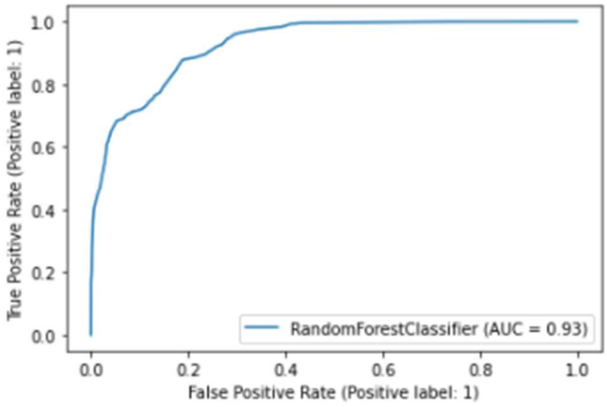

Figure 6.ROC curve for RF

9. Conclusion and Future Work

Computers are being targeted by using sophisticated techniques and strategies. Therefore, it becomes increasingly important for computer systems to be protected using advanced intrusion detection systems which are capable of detecting modern malware. Intrusion detection and prevention becomes essential to current and future networks and information systems, as our daily activities are heavily dependent on them.

Several techniques have been used in intrusion detection systems, but machine learning has turned out to be the most effective one. Additionally, different machine learning techniques can be used for analyzing huge data. To address this problem, different machine learning techniques, namely, SVM, RF, and ELM are investigated and compared in this work

As a future work, the proposed IDS can be implemented on network for protecting it from unlawful activity. In addition, it can be implemented for http services, ftp services for the detection of unauthorized work.

REFERENCES

[1] C. Cortes and V. Vapnik, Support-Vector Networks, Machine Learning, 20(3):273-297, September 1995.

[2] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani, "A Detailed analysis of the KDD CUP 99 Data Set", in the *Proc. of IEEE Symposium on Computational Intelligence in Security and Defense Applications (CISDA 2009)*, pp. 1-6, 2009.

[3] A.R. Vasudevan, E. Harshini, S. Selvakumar, "A Network Intrusion Detection System dataset and its comparison with KDD CUP 99 dataset", in *2011 Second Asian Himalayas International Conference on Internet* ***(AH-ICI*****)**, pp. 1-5, 2011.

[4] Y.H. Peng, "Research of Network Intrusion Detection system based on snort and NTOP", in the *Proc. of 2012 9th International Conference on Fuzzy Systems and Knowledge Discovery,* 2012.

[5] P. Singh, A. Tiwari, "Intrusion Detection System using KDD'99 Dataset", *International Journal of Engineering Research and Technology*, Vol. 03, No. 11, Nov. 2014.

[6] L Dhanabal, SP Shantharajah, "A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms", *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 04, No. 6, pp. 446-452, 2015.

[7] Loubna Dali, Ahmed Bentajer, Elmoutaoukkil Abdelmajid, Karim Abouelmehdi, Hoda Elsayed, Eladnani Fatiha, Benihssane Abderahim, "A survey of intrusion detection system", in the *Proc. of 2015 2nd World Symposium on Web Applications and Networking (WSWAN),* pp. 1-6**,** 2015.

[8] A. Dey, "Machine Learning Algorithms: A Review", *International Journal of Computer Science and Information Technologies (IJCSIT*), Vol. 7, No. 3, 2016.

[9] S.K. Biswas, "Intrusion Detection Using Machine Learning: A Comparison Study", *International Journal of Pure and Applied Mathematics*, Vol. 118, No. 19, 2018.

[10] C.J. Ugochukwn & E. O Bennett, "An Intrusion Detection System using Machine Learning algorithm", *International Journal of Computer Science and Mathematical Theory*, Vol. 4, No.1, 2018.

[11] Q.V. Dang, "Studying ML techniques for intrusion detection systems", *International Conference on Future Data and Security Engineering,* pp 411-426, Nov. 2019.

[12] J. Li, Y. Qu, F. Chao, H.P.H. Shum, E.S.L. Ho, L. Yang, "Machine Learning Algorithms for Network Intrusion Detection", *AI in Cybersecurity*, part of *Intelligent Systems Reference Library*, pp 151-179, 2018.

[13] Iftikhar Ahmad, Mohammad Basheri, Muhammad Javed Iqbal, and Aneel Rahim, "Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection", *IEEE Access*, Vol. 6, pp. 33789-33795, 2018.

[14] G. Saporito, "A deeper dive into the NSL-KDD Data Set", Sep 17, 2019.

[15] S.H. Kok, Azween Abdullah, N.Z. Jhanjhi, M. Supramaniam, "A Review of Intrusion Detection System using Machine Learning Approach", *International Journal of Engineering Research and Technology*, Vol.12, No.1, pp. 8-15, 2019.