A Comparative Study of Machine Learning Algorithms on Intrusion Detection System

[1] Anusha Manjunath Raykar and [2]Ashwini K B
[1]Student, Master of Computer Applications, RV College of Engineering, Bengaluru, India.
[2]Associate Professor, Master of Computer Applications, R V College of Engineering, Bengaluru, India.
[1]anushamr29@gmail.com, [2]ashwinikb@rvce.edu.in

ArticleInfo
Journal of Machine and Computing (http://anapub.co.ke/journals/jmc/jmc.html)
Doi : https://doi.org/10.53759/7669/jmc202202009
Received 30 December 2021; Revised form 25 February 2022; Accepted 05 March 2022.
Available online 05 April 2022.

Abstract – To detect malicious activity, an intrusion detection system (IDS) automates the procedure of observing and reasoning events that take place in the computer network. The existing intrusion detection system is confined to particular sorts of malicious activity, and it may not be able to identify new types of malicious activity, thus ML techniques were employed to implement the detection system at a faster rate. The intrusion detection system employs ML technologies such as random forest and support vector machines. This system has three main modules: data preparation, feature mapping, modelling and accuracy analyser. In this paper accuracy and sensitivity of both the support vector and random forest algorithms will be compared, with the results verified at a faster rate. The results show that machine learning approaches can aid intrusion detection using a dataset (KDD '99) that also highlights the findings of the prediction model which can differentiate between intrusions and normal connections.

Keywords – Detection, Intrusion, KDD’99, Dataset, Prediction, Attacks

I. INTRODUCTION

IDS’s have become a required element in most organizations' safety infrastructure as the quantity and severity of network attacks has escalated in recent years. Since it is theoretically not possible to put up a system without weaknesses, deploying effective detection systems is exceedingly difficult and has evolved as a prominent subject of research. Machine learning technologies can be applied to develop the detecting system at a faster rate. SVM and Random Forest are supervised ML techniques that are used to solve classification and regression problems. To train and test the model- KDD'99 dataset has been used. The KDD training set contains around 4,900,00 single connections, each of which has 41 columns and is labeled either as normal or intrusion(attack), with roughly only one attack types. After which the accuracy and efficiency of categorization modelling are evaluated empirically.

II. LITERATURE SURVEY

The purpose of IDS's is to detect harmful computer activity and network traffic that a regular firewall would miss. Paper[3] focuses on the issue faced by the users and the different types of systems that are present to solve this problem. Machine learning is a method for creating automated decision-making systems using data sets. In machine learning, the more data is utilized for system training, the higher the system's accuracy, assures good data is available. Anomaly-based systems operate in a similar manner, with three distinct phases: parameterization, training, and detection which is discussed in the paper.

Pattern identical algorithms are used by signature IDS to locate a known assault. Any major difference between observed and predicted behaviour is considered as something different, which might be construed as an incursion. This set of approaches is predicated on the idea that harmful behaviour is distinct from normal user behaviour. Intrusions are anomalous user activities that aren't typical, these are the two types of system which is mainly focused on paper[1]. In paper[2] functional modules of Id’s have been explained. In general, it is built around four functional modules: (i)Event boxes, (ii) Database boxes, (iii) Analysis boxes, and (iv) Response boxes, which is represented in the Fig1. The E block contains sensor elements for system monitoring and gathers event data for later analysis. This data must be saved so that it can be processed later. The information arriving from the E block is stored in D block elements for this purpose. A processing module is used to detect harmful behaviour by examining the events. Paper [4] focuses on how IoT devices are used to prevent the systems from intrusions. Tracing, debugging, and profiling are some of the methods that can be used to obtain data. IoT devices can be employed to accomplish the best intrusion detection which are made up of multiple sensors and actuators as well as an analysis system. Paper[5] presents an up-to-current taxonomy, as well as a critical analysis of the major explorer works on IoT IDSs, as well as a division of the suggested systems based on typology. It

presents an organized, complete analysis of existing IoT detection systems, allowing a finder to quickly result in acquaintance with the important components of IoT detection systems. In paper[7] an explanation about industrial implementation of Id's is presented. Modern Critical _Infrastructures includes the following: water treatment to plants, oil refineries, electricity grids, nuclear and thermal power plants, and industrial control systems (ICS). The term "ICS" refers to a system that controls a physical process by merging computational and communication components which is depicted in Fig 2.

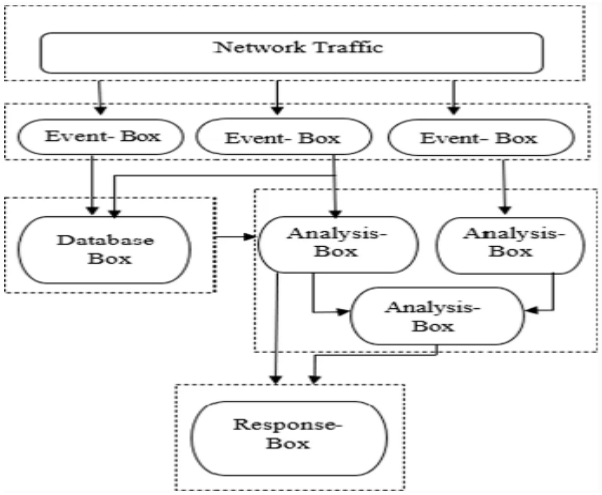

Fig 1. Common intrusion detection architecture for IDS [2]

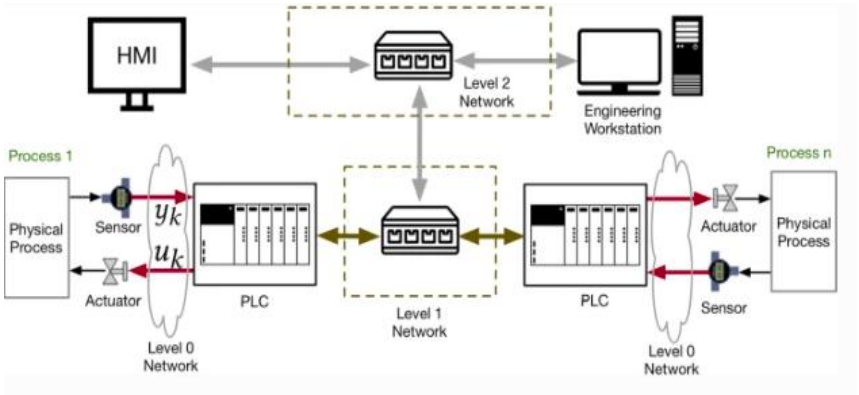

Fig 2. Abstract view of an Ics [7]

The Paper[6] provides a unique ML method for implementing a fast IDS utilizing the most recent CIC-IDS dataset. The following are key contributions: A realistic IDS which is capable of detecting the vast majority of the modern day threats, and the two features combining strategy with minimal prediction latency in mind. The classification of different division of attacks in the dataset is shown in the below figure.(Fig 3)

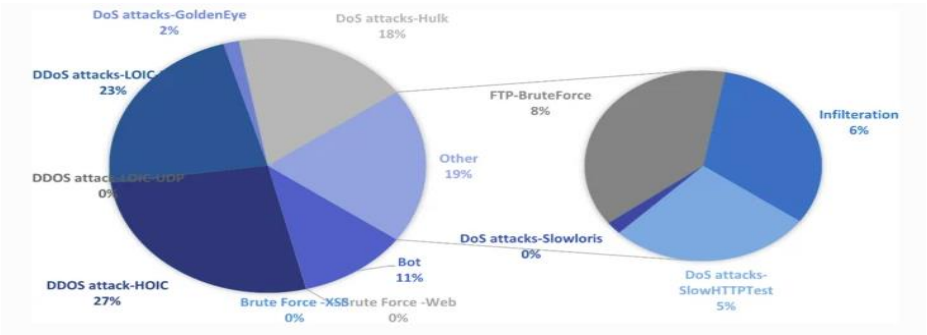

Fig 3. The classification of different categories of attacks [6]

The machine learning algorithms utilized in the paper [8], intrusion detection model included random forest, which is widely used in similar research. Classifier's production was improved by integrating data imbalance processing techniques with a random forest algorithm which was optimized for similarity.

The goal of this research (paper [9]) is to free position using a deep learning intrusion detection system with finer-grained-channel-state-information . To get blurry components of the CSI stage on various pathways SDS(sensitive-detection-signal), Deep learning's CNN (convolutional-neural-network) is used. The feature selection procedure is the subject of the paper [10]. This study presented a ranking of the feature importance depending on the decision tree approach for picking properties and methods for calculating the subset score using recursive feature removal.

III. PROBLEM STATEMENT

In order to detect malicious activity, an intrusion detection system (IDS) automates the methodology of observing and analyzing the events that happen in a CN(Computer-network). Software-based security systems cannot guarantee protection against new forms of attacks, and the behavior of the monitored environment may change over time, necessitating the system's maintenance. If there are attacks in the training set, the system will treat malevolent activity as normal. The existing intrusion detection system is confined to particular sorts of harmful activity; if new types of malicious activity emerge, it may not be able to identify them, thus ML technologies were employed for the development of the system at a faster rate.

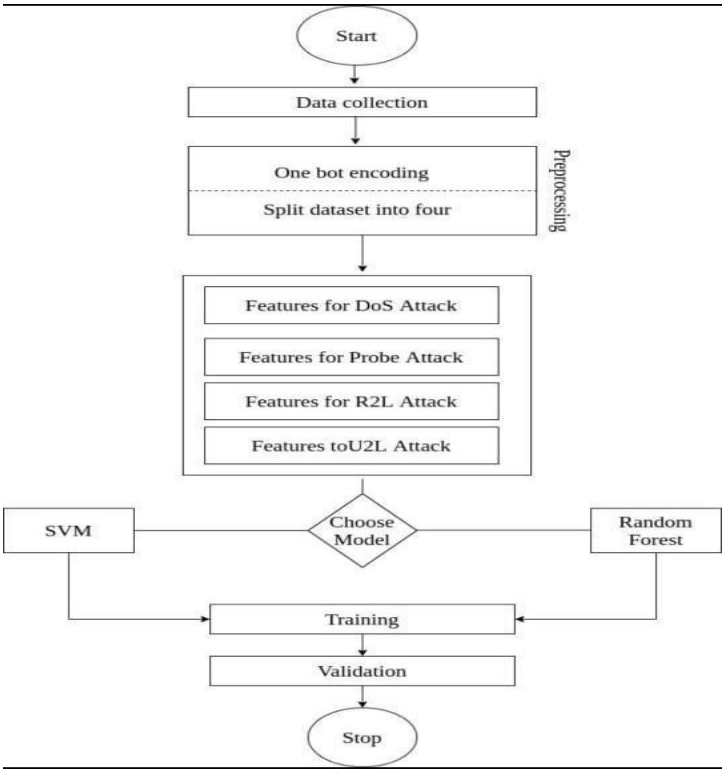

Fig 4. Workflow diagram

Fig 4 shows the many phases involved in training the model, including collection of data, data preparation, feature mapping, modeling, and accuracy analyzer.

Data acquisition includes the collection of NSL-KDD dataset , which has 41 characteristics in each training and testing dataset, and is utilized for intrusion detection. Preprocessing is required for the NSL-KDD dataset since it contains redundant data and categorical features. 24 attack types present in the training DS and 38 types of attacks present in the testing dataset, respectively. Raw information is transformed into a more consistent format through data pre-processing. To normalise and standardise the data, we use preprocessing. There are several techniques they are Identifying categorical Features which includes training, testing datasets with numerous groups, here a list of categorical characteristics will be established. Second one is One Hot Encoding, Categorical characteristics are not supported by the machine learning techniques currently in use. For training and testing, categorical features in training data sets are transformed to binary vectors. After mapping a categorical value to integer , each value can be recorded in the binary with all points of 0 excluding the integer index, that is 1.Third technique is Adding of missing categories in data testing. And the fourth one is splitting the dataset. Various other attack types observed in the training data and testing datasets which are mapped to R2L, Probe, DoS, and U2R after performing one-hot encoding. Giving training to the model for every sorts of strikes and

forecast inferences, it is divided into sets of four data depending on the type of attack i.e., DoS, Probe, R2L, and U2R. Feature Selection is the method of choosing the subset of an informative characteristic from the huge set of attributes. Recursive Feature Elimination Method Used. Mainly selecting the feature for four types of attacks. The next step is to choose a model for training and testing the system. After the comparison of accuracy and sensitivity of both support vector and random forest algorithms will be done and verifying results at the faster rate.

V. IMPLEMENTATION AND RESULT

Python used as a programming language, applications include,

- Software Development
- Data Analytics
- AI & ML
- Web Development
- Game Development

Dataset considered is KDD'99, No of rows it contains is 494021, No of columns it contains is 42. The KDD '99 data set has been the most often used for testing anomaly detection algorithms. The KDD dataset contains around 4,900,000 single connections, which has 41 columns and is labelled as normal or an intrusion(attack), with one different attack type.

Four types of attacks -

1) Denial-of-Service-Attack:users access a computer by making a computing or memory resource too busy or full to accept genuine requests.

2) User-to-Root-Attack: An attacker gains access to a system's normal user account before exploiting a vulnerability to gain root access.

3) Remote-to-Local-Attack: attacker with the ability to transmit packets to a system over n/w but no account on that machine takes advantage of a vulnerability. As a user of that machine, he or she gains local access.

4) Probing-Attack: An attempt to obtain metadata about a computer network in order to circumvent the network's security mechanisms.

Algorithms considered are:SVM (Support Vector Machine) is a type of supervised ML algorithm that can solve classification and regression problems. RF (Random forest) is a user-friendly ML method that produces results in most cases even without hyper-parameter adjustment.

The NSL-KDD dataset, which has 41 characteristics in each training and testing dataset, is utilized for intrusion detection. Preprocessing is required for the NSL-KDD dataset since it contains redundant data and categorical features.

The dataset is taken from kaggle[11] and this dataset is used for performing all the operations that are with respect to the proposed system and for training and testing purposes, divide the dataset into two partitions.The below figures (Fig. 5 and Fig. 6) represents the first three rows of train and testing dataset respectively. and Fig 7 represents the attack class distribution of both testing and training dataset for each type of the attacks.

Train dataset:

	duration	protocol_type	service	flag	src_bytes	dst_bytes	land	wrong_fragment	urgent	hot	num_failed_logins	logged_in	num_compromised
0	0	tcp	ftp_data	SF	491	0	0	0	0	0	0	0	0
1	0	udp	other	SF	146	0	0	0	0	0	0	0	0
2	0	tcp	private	S0	0	0	0	0	0	0	0	0	0

Train set dimension: 125973 rows, 42 columns

Fig 5. Train dataset output [11]

Test dataset:

	duration	protocol_type	service	flag	src_bytes	dst_bytes	land	wrong_fragment	urgent	hot	num_failed_logins	logged_in	num_compromised	root_sl
0	0	tcp	private	REJ	0	0	0	0	0	0	0	0	0	
1	0	tcp	private	REJ	0	0	0	0	0	0	0	0	0	
2	2	tcp	ftp_data	SF	12983	0	0	0	0	0	0	0	0	

Test set dimension: 22544 rows, 42 columns

Fig 6. Test dataset output [11]

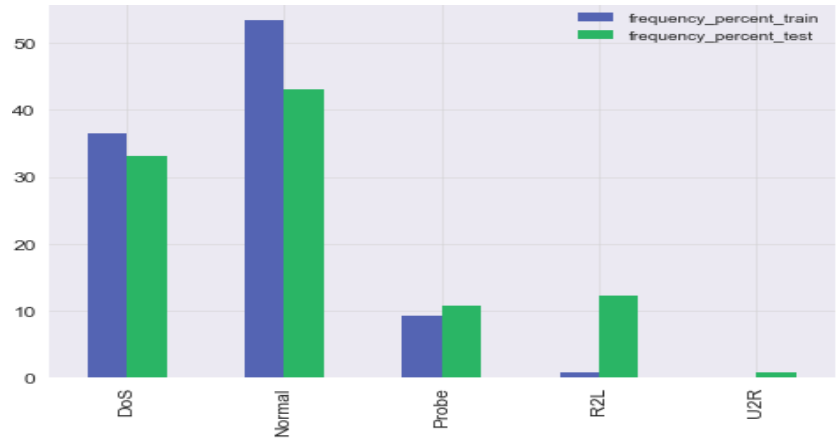

Fig 7. Attack class distribution

Feature-selection process:
RFE (Recursive Feature Elimination) Used in order to fetch the informative feature from a large set of features. The below figure (Fig 8) represents the features of the dataset that are considered with respect to the importance.

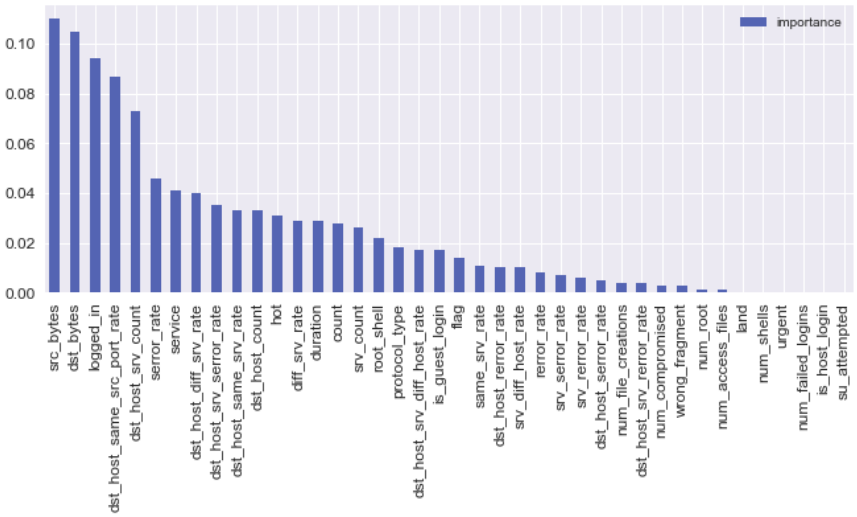

Fig 8. feature selection [11]

Classification:
A classification model tries to deduce something from seen data. Using RF and SVM to train particular data.

Prediction and Evaluation:
The model is trained using SVM or RF based on the set of data for various attacks displayed by the confusion matrix, and then predictions are made. A confusion matrix is a technique of briefing the classification algorithm performance. The metadata presented in a confusion matrix is used to calculate accuracy, precision, and recall. the Accuracy of both testing

and training dataset calculated for each of the algorithms and represented in Table 1 along with the train time

Table 1. Training and testing accuracy

Classifier	Training Accuracy	Testing Accuracy	Train Time
Support vector machine	99.79	92.86	44.15
Random Forest	100	91.40	10.71

Experimental Results and Discussions:

For the various attack kinds, attack count log values are framed. Normal packets are the most numerous in the data collection. All of the framed attacks are divided into 4 categories: DoS, User to Root (U2R), Probe, Remote to Location (R2L). The false negative rate (miss rate) and precision for each attack type is calculated and depicted in the below tables(Table 2 and Table 3) respectively.

Table 2. False negative rate for the each of the attack

Classifier	Do's	Probing	R2L	U2R	Average FNR
Support vector machine	3.49	17.7	72.74	32.86	31.69
Random Forest	6.82	8.7	86.63	95.7	49.45

Table 3. Precision for the each of the attack

Classifier	Do's	Probing	R2L	U2R	Average Precision
Support Vector machine	0.96	0.64	0.86	0.34	0.7
Random Forest	0.99	0.55	0.67	1	0.8

VI. CONCLUSION AND FUTURE WORK

Proposed intrusion detection model in this paper depends on two ML algorithms - SVM and Random Forest. Tested these two ML algorithms extensively by employing all of the characteristics and found them to be time consuming and provides poor performance. As a result, characteristic selection is even critical in the suggested study. To lower the dataset's dimensionality, Recursive Feature Elimination is used. Before feature selection, RF performed better compared to SVM. In the majority of the attacks, however, SVM performed better compared to Random Forest. So, deep learning models, which are a subset of machine learning, can be used to boost performance. Multiple deep networks can be employed to increase the quality of IDSs. DL models are known to have superior fitting and generalization skills when contrasted to ML models. Furthermore, unlike ML models, deep learning approaches are not reliant on feature engineering or domain expertise.

References

[1] Khraisat, A., Gondal, I., Vamplew, P. et al. Survey of intrusion detection systems: techniques, datasets and challenges. Cybersecur 2, 20 (2019). https://doi.org/10.1186/s42400-019-0038-7

[2] Disha, R.A., Waheed, S. Performance analysis of machine learning models for intrusion detection system using Gini Impurity-based Weighted Random Forest (GIWRF) feature selection technique. Cybersecurity 5, 1 (2022). https://doi.org/10.1186/s42400-021-00103-8

[3] Jadhav, A.D., Pellakuri, V. Highly accurate and efficient two phase-intrusion detection system (TP-IDS) using distributed processing of HADOOP and machine learning techniques. J Big Data 8, 131 (2021). https://doi.org/10.1186/s40537-021-00521-y

[4] Gassais, R., Ezzati-Jivan, N., Fernandez, J.M. et al. Multi-level host-based intrusion detection system for Internet of things. J Cloud Comp 9, 62 (2020). https://doi.org/10.1186/s13677-020-00206-6

[5] Khraisat, A., Alazab, A. A critical review of intrusion detection systems in the internet of things: techniques, deployment strategy, validation strategy, attacks, public datasets and challenges. Cybersecur 4, 18 (2021). https://doi.org/10.1186/s42400-021-00077-7

[6] Seth, S., Singh, G. & Kaur Chahal, K. A novel time efficient learning-based approach for smart intrusion detection system. J Big Data 8, 111 (2021). https://doi.org/10.1186/s40537-021-00498

[7] M. R., G.R., Ahmed, C.M. & Mathur, A. Machine learning for intrusion detection in industrial control systems: challenges and lessons from experimental evaluation. Cybersecur 4, 27 (2021). https://doi.org/10.1186/s42400-021-00095-5

[8] Wu, T., Fan, H., Zhu, H. et al. Intrusion detection system combined enhanced random forest with SMOTE algorithm. EURASIP J. Adv. Signal Process. 2022, 39 (2022). https://doi.org/10.1186/s13634-022-00871-6

[9] Hu, Y., Bai, F., Yang, X. et al. IDSDL: a sensitive intrusion detection system based on deep learning. J Wireless Com Network 2021, 95 (2021). https://doi.org/10.1186/s13638-021-01900-y

[10] Megantara, A.A., Ahmad, T. A hybrid machine learning method for increasing the performance of network intrusion detection systems. J Big Data 8, 142 (2021). https://doi.org/10.1186/s40537-021-00531-w

[11] Steven huang, Kaggle,2019, Https://Www.Kaggle.Com/Datasets/Galaxyh/Kdd-Cup-1999-Data/Metadata, 'Kddcup1999 Data Computer Network Intrusion Detection',