Comparative analysis of Machine Learning algorithms for Intrusion Detection

Vasudeva Pai 1, Devidas 2, Adesh N. D. 3
1NMAM Institute of Technology, Nitte, Udupi, Karnataka, India
2Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Udupi, Karnataka, India
3paivsassudeva@nitte.edu.in, *devidasbhat@nitte.edu.in, *adesh.nd@manipal.edu

Abstract - In this modern era, the network related applications, programs and services are growing enormously but the network security issues also grow along with them. Keeping the network secure is a challenging and a crucial task. To maintain the secure network there must be some system which can detect and identify any malicious activity happening in network. This system is called as Intrusion Detection System. There are many traditional network security tools and techniques of preventing intrusion like firewalls, anti-virus, encryption-decryption, access control etc. But all are not effective in protecting network from increasing attacks. The network traffic can be categories into normal and intrusive traffic using Machine Learning (ML) algorithms. Here, the preliminary comparative study regarding which type of machine learning algorithm performs better in identifying the attacks namely Denial of Service, Probe, User to Root and Remote to Local. The NSL-KDD dataset is used to study features and behavior of malicious attacker using machine learning techniques. This study can be taken as reference for mechanical engineers for developing a safe automation in industrial atmosphere and automation in automobile.

Index Terms—Intrusion Detection System, Machine Learning algorithms, NSL-KDD dataset, feature selection.

I. INTRODUCTION

The advancement of internet in this era has brought a positive impact to a wide variety of purpose. This improvement also has impact on emerging large amounts of data and information. Along with the progress of internet there are number of challenges to be faced to make it more reliable, stable and secure system. To make system more secure there are lot of alternatives like firewalls, some dynamic mechanisms, software etc., can be used. Out of these, Intrusion Detection System (IDS) is one among the most powerful dynamic mechanism that determines and detects the specific attacks in the network.

Intrusion detection system mainly monitors the process of network and analyzes them to detect any deviation from the normal operation or any sort of abnormalities. The software that scans the activity in the network to detect the suspicious and policy breaching. The IDS is classified into five categories such as Network Intrusion Detection System, Host Intrusion Detection System, Protocol-based Intrusion Detection System, Application Protocol-based Intrusion Detection System, and Hybrid Intrusion Detection System. There are two main methods of detection: misuse detection (signature-based detection) and anomaly detection.

Now a day, few researchers are looking towards machine learning algorithms to detect anomalies in the network. They are using large amount of training dataset to train the machine learning model and test the model by applying the input data that detect attacks or not [3]. It is also observed that some study has already been done [1-4] and only few algorithms are considered for the study purpose. Many of the research works were related to performance and evaluation of algorithms such as Support Vector Machine (SVM) [4], Artificial Neural Network (ANN) etc. But some ML techniques are more suitable for analyzing huge data for intrusion detection of network and information system.

In this paper, the research work is more intend to study the best algorithm, which is more prominently detects any anomaly in the network. Selecting the best possible algorithm for detection of anomaly in network plays a very important role in coming up with good solution to network security issues. The ML algorithms which are selected for the study are SVM, Decision Tree, Logistic Regression, Naive Bayes, and Random Forest. Here the NSL-KDD dataset is used to train and test the ML model. The dataset is preprocessed and features are extracted from it. These features are then applied to ML algorithms and the performance of these algorithms over different intrusions like Probe, Denial of Service (DoS), Remote to Local (R2L), and User to Root (U2R) are compared. The study gives a brief idea about which machine learning algorithm should be used by the Intrusion Detection System that will best identify the deviation in network. It also helps to build a more accurate and efficient detection system for identifying each attack.

The rest of the section is as follows, Section II gives the research work of other researchers pertaining to the machine learning and its application in network security. The Section III describes the methodology of the proposed study and Section IV explains the different machine learning algorithms. The brief description of performance metric is explained in Section V and the outcomes of experiments are discussed in Section VI. Lastly the highlights of the research work are drawn in section VII.
II. Literature Survey

The author [1] has studied the network security issues and conducted the experiment using Naïve Bayes, Random Forest Support Vector Machine and K-means ML algorithms to detect four types of attacks like DOS, PROBE, U2R and R2L. They concluded that the Random Forest Classifier (RFC) surpasses the other methods and also stated that hierarchical clustering method can be used to improve the performance of the system. In paper [2] the author has done a comparison using supervised machine learning classifiers namely, Random Forest, Support Vector Machine, Gaussian Naïve Bayes, and Logistic Regression are compared for an intrusion detection in network. Effective classifying algorithm is identified based on performance matrix namely F1-Score, accuracy, precision, and recall. Based on the observed results they have concluded that the Random forest classifier outperforms other classifiers for the considered data-set and parameters.

A light weight IDS method is proposed here [3] mainly concerned on pre-processing of the data, so that they can use important features of online data. The main step is to remove the redundant data from dataset to standardize the data. This helps the machine learning algorithms to give the unbiased and accurate result. In paper [4] the author proposed intrusion detection system (IDS) using supervised machine learning techniques to detect the online network data as normal or anomaly. The proposed method only identifies the Denial of Service (DOS) and probe attacks, but the other attacks are not taken into consideration.

The author proposed Intrusion detection method using Support Vector Machine (SVM) [5]. They also used feature removal method to improve the efficiency of the algorithm. Using the proposed feature removal method, they selected best nineteen features from the KDD-CUP99 data-set. The authors have proposed [6] anomaly intrusion detection using improved Self Adaptive Bayesian Algorithm to process the large amount of data. In this paper [7] authors proposed a novel idea to reduce the dimensionality of the data by using triangle-based K-NN approach. An Intrusion Detection system using fuzzy logic is tested [8]. This technique uses a set of fuzzy rules which are obtained from the definite rules using frequent items. The classification accuracy of this approach is above 90% for all types of attacks.

G. Meera Gandhi et al., [9] examined the performance measure of four supervised machine learning algorithms in detecting the four types of attack such as DoS, R2L, Probe, and U2R. The result shows that the C4.5 decision tree classifier performs best in prediction accuracy compared to Naïve Bayes. The authors [10] have compared the performance of the three machine learning algorithm namely Neural Network, Support Vector Machine and Decision Tree. The algorithms were measured based on false alarm rate, accuracy and detection rate of four categories of attacks classes. From these experiments they found that the Decision tree (J48) algorithm outperformed the other two algorithms.

The author [11] suggested using a collective of, Support Vector Machines (SVMs), Multivariate Adaptive Regression Splines (MARS), and Artificial Neural Networks (ANNs). Whereas [12] proposed a hybrid approach in which they have used a Support Vector Machine (SVM) and Radial Basis Function (RBF).

The sequential search strategy for feature selection or feature extraction through determining the importance of a given attribute by simply removing it and recording the performance [13]. If performance of the algorithm is increased, then the feature is unimportant and thus shall be removed. The author [14] suggested that every attribute in the dataset is not much important and it will not give the accurate result as expected. It is very important to reduce the no of features using feature selection technique and also they concluded that simple Cart algorithms gives accurate result than other five algorithm J4.8 Naïve Bayes, NBTree, Multi-Layer Perceptron, and SVM.

III. Methodology

In this paper, the research work is more intend to study the best machine learning algorithm, which is more prominently detects any anomaly in the network. Selecting the best possible algorithm for detection of anomaly in network plays a very important role in coming up with good solution to network security issues. There are various steps to be carried out in order to build a ML model such as data preparation, data preprocessing, feature selection, data visualization, clustering, model training and model validation. Different ML algorithms are selected for the study purposes are SVM, Decision Tree, Logistic Regression, Naïve Bayes, J48, and Random Forest. The NSL-KDD dataset [15] is used for training the machine learning model to detect the intrusion in network. NSL-KDD dataset consist of several attributes which gives the features of network-based intrusion detection systems. The dataset consist of 42 attributes out of which 41 attributes are features and one attribute is a class label as normal traffic or anomaly traffic is as shown in Figure 1. This makes the detection more accurate and realistic as the ML model is also verified for the unknown attacks. NSL KDD dataset is divided into train and test data as shown in Table 1. The train data set is used to train the ML model and test dataset is given as inputs to trained model for validation purpose. The advantage of this dataset is that it does not include any redundant records in the training data, so the classifiers are accurate and not biased towards more frequent records.
Figure 1: Attributes of NSL–KDD dataset [15]

Table 1: Distribution of instance in NSL-KDD dataset for training and testing

| Types of Attack          | Number of Records |
|--------------------------|-------------------|
|                          | Training dataset  | Test dataset   |
| Normal                   | 67343             | 9711            |
| Denial of Service        | 45927             | 7456            |
| Probe                    | 11656             | 2421            |
| Remote to Local          | 995               | 2756            |
| User to Root             | 52                | 200             |
| Total                    | 125973            | 22544           |

A. Data Preparation

Data preparation is the process of cleaning and transforming the raw data prior to data preprocessing and analysis. It is a mandatory step prior to processing and often involves making corrections to data, reformatting data and the combining of data sets to enrich data. The first step of data preparation is dividing connections into normal and anomaly classes based on 'labels' column. Then attacks are further divided into four main categories: DoS, Probe, U2R and R2L as shown in Figure 2. After this step, all the categories are indexed and furthermore, column ID is included to simplify the work with clustered data.

B. Data Preprocessing

Data preprocessing is an integral step in Machine Learning as the quality of data and the useful information that can be derived from it directly affects the ability of our model to learn. Some of the concepts that handles the data preprocessing are Handling Null values, Standardization, Handling Categorical Variables, One-Hot Encoding, and Multicol linearity. The real-world data is incomplete, inconsistent, lacking in certain behavior, and likely to contain many errors. Data pre-processing is proven to resolve these errors.
The technique of data pre-processing will get rid of irrelevant data and missing values. Here, the One Hot Encoding technique is used to convert nominal values into binary values in dataset. So, that the machine learning algorithms can use these data values to training a model.

C. Feature Selection
Feature selection in machine learning is one of the core concepts which largely impact the performance of the model. The features that are used to train machine learning models have a huge influence on the performance because partially relevant or irrelevant features can negatively impact the performance. The process of selecting the features manually or automatically from dataset mainly contributes to the prediction or output of model which are interested in. The technique used here for feature selection is attribute ratio. After feature selection data standardization is done which is the process that converts the structure of disparate datasets into a Common Data Format. Standardization is required as many distance based algorithms are used.

D. Data Visualization
Data visualization is used to understand how actually the data looks like and what type of correlation is held by the attributes of data. It is the best and quickest way to see if the features correspond properly to the output. In this study, the Principal Component Analysis (PCA) algorithm is used on NSL-KDD dataset for data visualization is as shown in Figure 3. One of the advantages of using PCA algorithm is for speeding up machine learning algorithms and also used in reducing the dimension of data. It is also used later on for preprocessing in Gaussian Mixture clustering.

E. Clustering
Clustering is the process of dividing the data points into several groups such that the data points in the group are more similar to each other than data points in other groups. In simple words, the task is to segregate groups with similar properties and assign them into clusters. In this study, the K-Means and Gaussian Mixture are used for clustering purpose because the output of K Means is different from Gaussian Mixture is as shown in the Figure 4. So, that the results of both algorithm are combined to achieve high performance. Here, it is divide into two categories, Frist category include clusters that contain both ‘attack’ and ‘normal’ connections and also have more than 25 connections. The second category includes all other clusters and maps these cluster to ‘attack’ or ‘normal’ based on the majority. The clusters that include less than or equal to 25 connections are considered as outliers.

F. Training
After all the previous process for NSL-KDD dataset, now the ML model can be trained using various different classification algorithms namely Random Forest, Logistic Regression, Naive Bayes, Decision Tree, SVM and J48 with NSL-KDD dataset.
IV. CLASSIFICATION ALGORITHMS

A. Random Forest
Random Forests are a boosting learning strategy for classification and regression strategies. It builds various decision trees at preparing time and yields the class that is the method of the classes yield by individual trees. It likewise endeavors to limit the issues of high change and high predisposition by averaging to locate a characteristic harmony between the two boundaries.

B. Naive Bayes
Naive Bayes is an upgraded version of Bayes Theorem as it considers strong independence among the attributes. It is based on Bayes probability theory. It is a classification technique which is based on an assumption of independence between predictors. This implies that the probability of one attribute does not affect the probability of the other attribute.

C. Support Vector Machine
SVM is a supervised learning technique which classifies data into two classes over a hyper plane. In SVM, each data item is plotted as a point in n-dimensional space where n represents the number of features with the value of each feature being the value of a particular coordinate.

D. J48
In this algorithm every aspect of the information is split into minor subsets based on decision. J48 look at the standardized data gain and split the information by choosing an attribute. To summarize, the attribute extreme standardized data gained is utilized. The minor subsets are returned by the algorithm. The split strategies stop if a subset has a place with a similar class in all the instances. J48 develops a decision node utilizing the expected estimations of the class. J48 decision tree can deal with particular characteristics, lost or missing attribute estimations of the data and varying attribute costs.

E. Logistic Regression
Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. Unlike linear regression which outputs continuous number values, logistic regression transforms its output using the logistic sigmoid function to return a probability value which can then be mapped to two or more discrete classes.

F. Decision Tree
A decision tree classifier is one of the most known supervised machine learning technique. A decision tree is composed of three basic elements. A decision node represents test or condition on a data item, and an edge or a branch which corresponds to the one of the possible attribute values which means one of the test attribute outcomes. A leaf determines the class to which the object belongs.

V. PERFORMANCE METRICS

A. Accuracy
It is the number of correctly predicted data points out of all the data points and it is represented in terms of percentage. The accuracy is calculated as shown in Equation 1.

\[
Accuracy = \frac{TP + TN}{TP + TN + FP + FN}
\]  

Where False positive (FP): It defined as the number of detected attacks which are actually normal, and False negative (FN): means the wrong prediction i.e. it detects the instances as normal but in actual its an attack. True positive (TP): is an instances that are correctly predicted as normal, and True negative (TN): is an instances that are correctly classified or detected as attack.

B. Precision
It is a measure which estimates the probability that a positive prediction is correct, and the formula is as shown in Equation 2.

\[
Precision = \frac{TP}{TP + FP}
\]

C. Recall
It is the proportion of instances belonging to the positive class that are correctly predicted as positive, and the formula is as shown in Equation 3.
\[ \text{Recall} = \frac{TP}{TP + FN} \]  

(3)

D. Kappa

Its value ranges from 0 to 1. 0 means totally disagreement and full agreement. It checks the reliability of classifying algorithm on dataset.

E. Receiver Operating Characteristics (ROC)

It is used to design the curve between true positive rate and false positive rate, and the Area Under Curve (AUC) gives the value of ROC. More the area under curve and more will be the value of ROC.

F. F1 Score

The F1 score or F-measure is a measure of the test’s accuracy. It considers both the precision P and the recall R of the test to compute the score.

VI. Results

A. DoS attack

Table 2: Performance of the selected machine learning classifiers against the DoS attack

| Machine Learning algorithms | Accuracy (%) | Precision | Recall   | F1-Score | ROC |
|-----------------------------|--------------|-----------|----------|----------|-----|
| Random Forest               | 99.83        | 1         | 0.999    | 0.999    | 0.999|
| J48                          | 99.76        | 0.99      | 0.999    | 0.999    | 0.658|
| Logistic Regression         | 99.47        | 0.99      | 0.998    | 0.999    | 0.727|
| Decision Tree               | 99.66        | 0.99      | 0.999    | 1.0      | 0.618|
| SVM                         | 99.19        | 0.98      | 0.999    | 0.999    | 0.5  |
| Naïve Bayes                 | 94.19        | 0.97      | 0.903    | 0.948    | 0.844|

B. Probe attack

Table 3: Performance of the selected machine learning classifiers against the Probe attack

| Machine Learning algorithms | Accuracy (%) | Precision | Recall   | F1-Score | ROC |
|-----------------------------|--------------|-----------|----------|----------|-----|
| Random Forest               | 99.94        | 0.999     | 1.0      | 1.0      | 1.0 |
| J48                          | 99.94        | 0.994     | 0.999    | 0.999    | 0.999|
| Logistic Regression         | 99.47        | 0.995     | 0.998    | 0.998    | 0.996|
| Decision Tree               | 99.91        | 0.993     | 0.999    | 0.998    | 1.0  |
| SVM                         | 99.90        | 0.994     | 1.0      | 1.0      | 1.0  |
| Naïve Bayes                 | 90.31        | 0.986     | 0.966    | 0.956    | 0.974|

C. Remote to Local attack

Table 4: Performance of the selected machine learning classifiers against the Remote to Local attack

| Machine Learning algorithms | Accuracy (%) | Precision | Recall   | F1-Score | ROC |
|-----------------------------|--------------|-----------|----------|----------|-----|
| Random Forest               | 99.99        | 0.999     | 1.0      | 0.999    | 1.0 |
| J48                          | 99.89        | 0.998     | 0.999    | 0.998    | 0.994|
| Logistic Regression         | 99.81        | 0.997     | 0.992    | 0.992    | 0.997|
| Decision Tree               | 99.74        | 0.997     | 0.993    | 0.993    | 0.999|
| SVM                         | 98.92        | 0.995     | 0.990    | 0.990    | 0.979|
| Naïve Bayes                 | 98.92        | 0.999     | 0.882    | 0.931    | 0.953|

D. User to Root attack

Table 5: Performance of the selected machine learning classifiers against the User to Root attack
In this study, the machine learning technique that can efficiently identify the anomalies in the network such as DoS, Probe, R2L and U2R. The machine learning technique considered for the study are Random forest, Naive Bayes, J48, Logistic Regression, Decision Tree, and SVM. The NSL-KDD dataset is used for training and testing the machine learning model. The dataset needs to be normalized before feeding the model. The performance of classifiers on the test data set and compute a classification metrics for each model. To check the performance of ML model is evaluated in terms of accuracy, precision, recall, F1-score, and ROC. The best classifier was the random forest classifier with an accuracy of 99.997% for the DoS attack and similarly J48 is the best classifier with an accuracy of 99.947% for U2R attack. Likewise, the performance of R2L and U2R attack is computed and do a comparative study.

There is an increasing demand for a reliable and real-world attacks dataset among research community. As a future work, it is possible to work in several directions, at one direction enhancing the current work by implementing various balancing algorithm and investigate its effects on model learning. Study the trouble with the two attacks (Web Attack-Brute Force Web Attack-XSS). In addition, applying deep learning algorithms on this data set such as convolution and recurrent neural networks.

### VII. Conclusion and Future Work

| Machine Learning algorithms | Accuracy (%) | Precision | Recall | F1-Score | ROC |
|-----------------------------|--------------|-----------|--------|----------|-----|
| Random Forest               | 99.85        | 0.999     | 0.998  | 0.998    | 0.997|
| J48                         | 99.67        | 1         | 0.998  | 0.998    | 0.936|
| Logistic Regression         | 99.16        | 1         | 0.995  | 0.995    | 0.982|
| Decision Tree               | 99.29        | 0.999     | 0.993  | 0.993    | 0.999|
| SVM                         | 98.97        | 0.999     | 0.992  | 0.992    | 0.826|
| Naïve Bayes                 | 88.85        | 0.999     | 0.942  | 0.960    | 0.948|

### REFERENCES

[1] A. T. Youssef El Mourabit, Anouar Bouindien and N. E. Moussaid, “Intrusion Detection Techniques in Wireless Sensor Network Using Data Mining Algorithms: Comparative Evaluation Based on Attacks Detection”, International Journal of Advanced Computer Science and Applications, vol. 6, no. 9, pp. 164–172, (2015).

[2] J. Manjula C. Belavagi and Balachandra Munyal, “Performance Evaluation of Supervised Machine Learning Algorithms for Intrusion Detection” Twelfth International Multi-Conference on Information Processing- 2016.

[3] S. S. S. Sundhu, S. Goetha and A. Kannan, “Decision Tree Based Light Weight Intrusion Detection Using a Wrapper Approach, Expert Systems with Applications”, vol. 39, no. 1, pp. 129–141, (2012).

[4] P.Sangkapatsane, N. Wattanapongsakorn and C. Chamsripinyo., “Practical Real-Time Intrusion Detection Using Machine Learning Approaches, Computer Communications”, vol. 34, no. 18, pp. 2227–2235, (2011).

[5] Y. Li, J. Xia, S. Zhang, J. Yan, X. Ai and K. Dai, “An Efficient Intrusion Detection System Based on Support Vector Machines and Gradually Feature Removal Method, Expert Systems with Applications”, vol. 39, no. 1, pp. 424430, (2012).

[6] D. M. Farid and M. Z. Rahman, “Anomaly network intrusion detection based on improved self-adaptive Bayesian algorithm” Journal of computers, vol. 5, no. 1, pp. 23–31, 2010.

[7] J.B. Luo and J. Xia, “A novel intrusion detection system based on feature generation with visualization strategy”, Expert Systems with Applications, vol. 41, no. 9, pp. 4139–4147,2014.

[8] R. Shannugavadivu and Dr. N. Nagarajan, “Network intrusion detection system using fuzzy logic”, Indian Journal of Computer Science and Engineering, Vol. 2, pp. 101-111.

[9] G. Meera Gandhi, “Machine learning approach for attack prediction and classification using supervised learning algorithms”, International Journal of Computer Science & Communication, Vol. 1, No. 2, pp. 247-250. July-December 2010.

[10] K. Abdalfall, and S. Mara, “Comparison of machine learning algorithms performance in detecting network intrusion”, In Proceedings of Networking and Information Technology (ICNIT), pp. 221 – 226, Manila 2010.

[11] S. Mukkamala, A. H. Sung, and A. Abraham, “Intrusion detection using an ensemble of intelligent paradigms” J. Netw. Comput. Appl., vol. 28, no. 2, pp. 167–182, 2005.

[12] M. Govindarajan and R. Chandrasekar, “Intrusion Detection using an Ensemble of Classification Methods,” In Proceedings of World Congr. Eng. Computer. Sci., vol. 1, no. October, 2012.

[13] Karam Bajaj, Amit 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 Applications (0975 – 887X) Volume 76– No.1, August 2013.

[14] C.Ingre B. and Yadav A., “Performance analysis of NSL-KDD dataset using ANN,” 2015 International Conference on Signal Processing and Communication Engineering Systems, Guntur, 2015, pp. 92-96.

[15] NSL-KDD dataset[online] available: http://nsl.cs.unb.ca/nsl-kdd/. Last accessed on 15-7-2020.