International Journal of Multidisciplinary Engineering in Current Research - IJMEC

Volume 10, Issue 5, May-2025, http://ijmec.com/, ISSN: 2456-4265

Network Intrusion Detection Using Supervised Machine Learning

Kotha Jyothiratnam
PG scholar, Department of MCA, CDNR collage, Bhimavaram, Andhra Pradesh.
K.Venkatesh
(Assistant Professor), Master of Computer Applications, DNR collage, Bhimavaram, Andhra Pradesh.

Abstract In the modern computer world, use of the internet is increasing day by day. However, the increasing use of the internet creates some security issues. These days, such new type of security attacks occurs every day and it is not easy to detect and prevent those attacks effectively. One common method of attack involves sending large amount of request to site or server and server will be unable to handle such huge requests and site will be offline for many days. This type of attack is called distributed denial of service (DDOS) attack, which act as a major security threat to internet services and most critical attack for cyber security world. Detection and prevention of Distributed Denial of Service Attack (DDoS) becomes a crucial process for the commercial organizations that uses the internet. Different approaches have been adapted to process traffic information collected by monitoring stations (Routers and Servers) to distinguish malicious traffic such as DDoS attack from normal traffic in Intrusion Detection Systems (IDS). In general, Machine learning techniques can be designed and implemented with the intrusion systems to protect the organizations from malicious traffic. Specifically, supervised clustering techniques allow to effectively distinguishing the normal traffic from malicious traffic with good accuracy. In this paper, machine learning algorithms are used to detect DDoS attacks collected from "KDD cup 99 Dataset", pre-processing and feature selection technique is used on the dataset to enhance the performance of the classifiers and reduce the detection time. The classification algorithms such as C4.5 decision tree and Navie Bayes is applied on the training dataset and the implementation of the algorithm is done using spyder tool. The performance comparison of algorithms is shown using confusion matrix and it is found that C4.5 decision is more efficient in detection of DDOS attack .The proposed method can be used as DDoS defense system.

Key Words: C 4.5 Decision Tree, DoS attack detection, IDS, KDD Dataset, Naive Bayesian classifier, Machine learning.

I. Introduction

With the rapid expansion of information technology, computer networks are now widely used across businesses, industries, and everyday life. This makes it crucial for IT administrators to ensure network security and trustworthiness. However, the rapid growth of technology has also led to challenges in establishing reliable networks, which is a complex task. Various types of cyberattacks threaten the privacy, integrity, and availability of computer networks, with denial-of-service (DOS) attacks being particularly common. These attacks aim to disrupt multiple end-user services by overwhelming network resources and overloading systems.

In recent years, there has been growing interest in using machine learning techniques, particularly classification algorithms, to build more accurate intrusion detection models. This paper proposes leveraging machine learning classifiers such as Naive Bayes and Decision Tree classifiers to enhance the accuracy of intrusion detection systems using a Knowledge Discovery in Databases (KDD) dataset.

This paper conducts experiments to evaluate different machine learning classifiers using the KDD intrusion dataset, assessing their performance based on various metrics. Specifically, it focuses on improving the detection rate of the intrusion detection system by analysing false positive and false negative metrics. [1]

The KDDCup 1999 intrusion detection dataset is crucial for refining intrusion detection systems and is extensively utilized by researchers in the field. In addition to analysing the KDDCup'99 dataset as well as providing an extensive review of the classification methods used in intrusion detection, this paper gives an overview of intrusion detection. [2]

By relaxing the conditional independence requirement, the Hidden Naive Bayes (HNB) model differs from the conventional naive Bayes method. This paper focuses primarily on the Hidden Naive Bayes model, and preliminary findings suggest that it outperforms traditional naive Bayes models in terms of error rate, misclassification cost, and accuracy, particularly when applied to the Knowledge Discovery and Data Mining (KDD) Cup 1999 dataset. [3]

To address the multitude of attacks targeting network resources, intrusion detection systems (IDS) have become essential. IDS monitors network traffic to detect and alert administrators to any suspicious activity or potential attacks. They can focus on monitoring traffic for specific devices (host-based IDS) or all network traffic (network-based IDS), with the latter being more common. There are two main types of IDS: anomaly-based and exploit-based. Anomaly-based IDS detect deviations from normal behaviour by comparing current traffic patterns with historical data, while exploit-based IDS rely on predefined attack signatures. Both types have their advantages and drawbacks, leading to challenges in accurately detecting all types of attacks.

This paper presents a newly collected dataset that includes modern DDoS attacks across various network layers, including SIDDoS and HTTP Flood, which were not commonly available in existing datasets. The study evaluates the performance of three popular classification techniques: Naïve Bayes, Random Forest, and Multilayer Perceptron (MLP). [4]

This paper proposes a hierarchical, layered approach to enhance the detection rate of both majority and minority attacks. For each attack class, the proposed approach uses a Naive Bayes classifier with a K2 learning procedure on a NSL KDD dataset. With this approach, every layer is shown separately to identify a certain kind of attack category. The layer receives the output from the previous layer, which helps to increase the detection rate and better classify majority and minority attacks. [5]

Given the complexity of distinguishing normal and attack packets, machine learning algorithms are increasingly being explored as an alternative approach for intrusion detection. These algorithms automate the process of identifying patterns of normal and intrusive behaviour, reducing the need for manual intervention.

This paper introduces a novel learning algorithm for adaptive network intrusion detection, employing a naive decision tree and Bayesian classifier. It aims to achieve balanced detections while maintaining acceptable false-positive rates across various network attack types. Additionally, the algorithm tackles challenges in data mining, including reducing noise in training data, managing missing attribute values, and handling continuous attributes. [6]

This paper introduces a traffic classification scheme aimed at enhancing classification accuracy in scenarios with limited training data. It proposes a method to combine the predictions of naive Bayes for different traffic flows. [7]

1. LITERATURE SURVEY

In this paper, we explore the effectiveness of various machine learning classifiers for intrusion detection using the KDD intrusion dataset. Our experiments aimed to evaluate performance metrics like false negatives and false positives to improve detection rates. We found that the decision table classifier had the lowest false negative rate, while the random forest classifier achieved the highest average accuracy. This research underscores the importance of robust intrusion detection systems against evolving cyber threats. [1]

As security becomes increasingly crucial for organisational information systems, detecting threats and attacks on computer networks has become a pressing issue. Intrusion detection systems (IDS) have emerged as essential tools to safeguard computer systems from various attacks. In recent years, data mining-based intrusion detection has gained significant attention from researchers, who propose different algorithms across various categories. Constructing effective classifiers presents another challenge in building efficient intrusion detection systems. The KDDCup 1999 intrusion detection dataset serves as a cornerstone for refining such systems and is widely utilized by researchers in this field. This paper provides an overview of intrusion detection, highlights the importance of the KDDCup'99 dataset, and offers a detailed analysis of classification techniques employed in intrusion detection. [2]

As Internet connectivity grows and network traffic increases, the need for effective network intrusion detection systems (NIDS) becomes more critical to combating sophisticated attacks. These systems often rely on techniques like pattern recognition and data mining to classify network events as normal or malicious. The Hidden Naive Bayes (HNB) model is particularly useful for intrusion detection tasks with highly correlated features and large data stream volumes. Compared to traditional naive Bayes models and other advanced methods like support vector machines, the HNB model demonstrates superior performance in terms of accuracy, error rate, and misclassification cost. Specifically, it shows enhanced accuracy in detecting denial-of-service (DoS) attacks, highlighting its effectiveness in addressing complex intrusion scenarios. [3]

Distributed denial of service (DDoS) attacks remains a persistent challenge for users and organizations. Security engineers strive to ensure service availability by using intrusion-detection systems (IDS) to detect and classify any unusual activity. These systems need to stay updated with the latest attack prevention methods to maintain service confidentiality, integrity, and availability. This paper addresses the need for updated datasets containing modern DDoS attacks across different network layers, such as SIDDoS and HTTP Flood. The study employs three well-known classification techniques: multilayer perceptron (MLP), Naïve Bayes, and random forest, and finds that MLP achieved the highest accuracy rate at 98.63%. [4]

The workflow of the proposed method is setup as shown in the Figure 1, starting with data collection (KDD-99 Dataset), Pre-Processing: Training and testing dataset, building model and result analysis

3.1 KDD cup 1999 dataset Collection

In 1998, the DARPA Intrusion Detection Assessment Program was prepared and managed by MIT Lincoln Labs. Its purpose was to study and evaluate intrusion detection research.

Standard data sets include various simulation intrusions in military network environments. The connection to the dataset includes a sequence of TCP packets beginning and ending at a well-defined time between the source IP address and the destination IP address using a well-defined protocol.

Each connection is categorized as a normal or specific type of attack. Data sets are categorized into five sub-sets: denial-of-service attacks, local or remote network attacks, user / root attacks, sample attacks, and generic data. Each record is classified as normal or attack with exactly one type of attack.

They are categorized as follows:

□ **Denial of service (DoS)** Denial of Service (DOS) allows a legitimate user to gain access to the machine by creating too much or too much computer resources or memory for an attacker to handle legitimate requests.

□ **R2L (Local Remote Attack (User))** Local Remote Attack (R2L) is a type of attack in which an attacker can send packets to a computer over the network and then exploit a vulnerability in the computer to illegally attack local access. On the machine.

• **Root User Attack (U2R)** Root User Attack (U2R) is the attack class that an attacker first accesses a regular user account on a system. The vulnerability could be exploited to gain root access to the system.

• Monitoring (monitoring and other discovery) detection is an attack type in which an attacker scans the network for known information or vulnerabilities. An attacker with a map of systems and services available on the network will use the information to detect attacks.

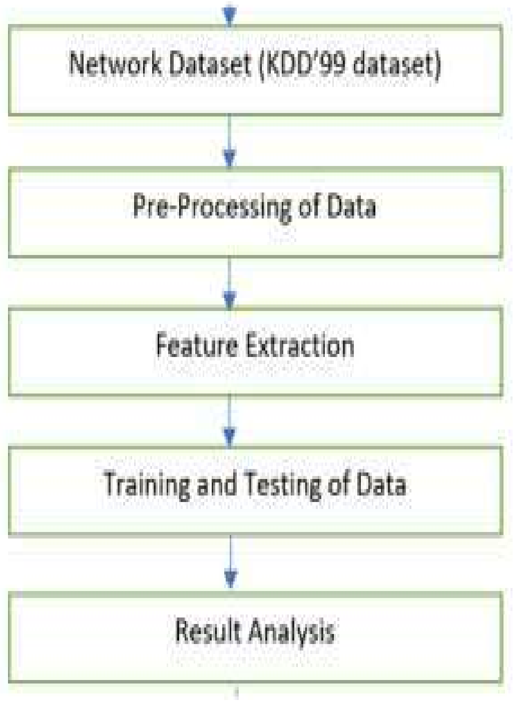

Network Dataset (KDD'99 dataset)
Pre-Processing of Data
Feature Extraction
Training and Testing of Data
Result Analysis

Figure 1: System design

3.2 Data Pre-processing:

Preprocessing involves Handling Null Values: - is null() method to check whether a null values is present in dataset . Label Encoder: le=label Encoder () method label_ encoder is used to transferring Categorical data into Numerical data

3.3 Feature Selection

Information Gain Ratio based feature selection: Features selected based on only information gain is biased towards attributes having many values. Information Gain Ratio (IGR) based Feature Selection removes this drawback by taking the splitting information of an attribute into account. Splitting information of an attribute is the entropy of pattern distribution into branches. Gain ratio of attribute decreases as value of split information increases.

Algorithm:

1. Start with the full set of attributes (set containing all attributes of the dataset) and null selected feature set.

2. Calculate information gain ratio of each attribute.

3. Choose an attribute from the total set with the highest information gain ratio.

4. Split the dataset into sub datasets depending on the attribute values.

5. Add the attribute to selected feature set and remove from set of attributes.

6. Repeat step 2 to 5 for each of the sub-datasets with the set of attributes, if instance in a sub-dataset belongs to more than one class.

7. Output the selected feature set

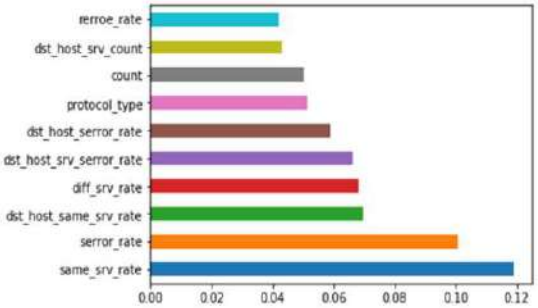


Figure 2: Feature Selection

3.4 Model building

The supervised machine learning algorithms [1] are those algorithms which needs external assistance. The input dataset is divided into train and test dataset. The train dataset has output variable which needs to be predicted or classified. All algorithms learn some kind of patterns from the training dataset and apply them to the test dataset for prediction or classification.

The proposed method distinguish the normal traffic from malicious traffic using supervised algorithms such as C4.5 Decision tree and Naïve Bayesian Classifier a. C4.5 Decision tree C4.5 Decision tree is one of the simple technique used in the machine learning and data mining. It is used as a predictive model in which observations about an item are mapped to conclusions about the item's target value.

In the process of decision analysis, a decision tree can be used to represent decisions and decision making visually and explicitly. In this algorithm, the data set is learnt and modelled. Therefore, whenever a new data item is given for classification, it will be classified accordingly learned from the previous dataset

The steps of the algorithm are as follows:

1. If all the given training examples belong to the same class, then a leaf node is created for the decision tree by choosing that class.

2. For every feature 'a', calculate the gain ratio by dividing the information gain of an attribute with splitting value of that attribute.

The formula for gain ratio is Gain Ratio a(a) = IG(a) / Split(a)

where, S is the set of all the examples in the given training set.

RESULT

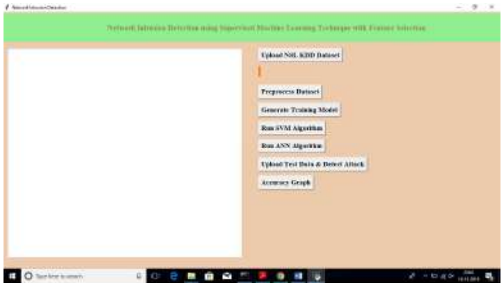
In this paper author is evaluating performance of two supervised machine learning algorithms such as SVM (Support Vector Machine) and ANN (Artificial Neural Networks). Machine learning algorithms will be used to detect whether request data contains normal or attack (anomaly) signatures.

In above lines we can see normal is having id 0 and Anomaly has id 1 and goes on for all attacks.

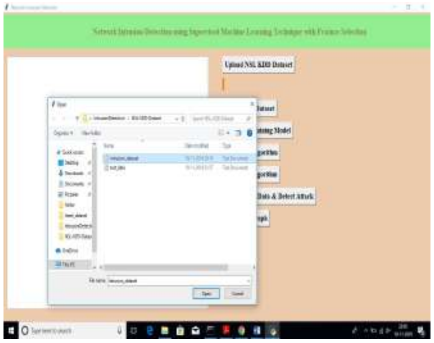
Before running code execute below two commands

Screen shots

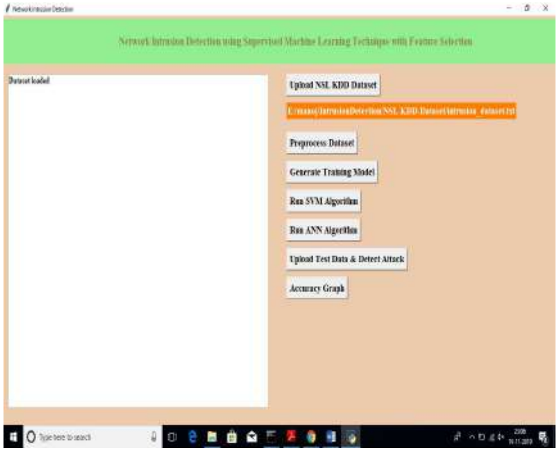
Double click on ‘run.bat’ file to get below screen


Network Intrusion Detection
Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection
Upload NSL KDD Dataset
Preprocess Dataset
Generate Training Model
Run SVM Algorithm
Run ANN Algorithm
Upload Test Data & Detect Attack
Accuracy Graph

In above screen click on ‘Upload NSL KDD Dataset’ button and upload dataset


Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection
Upload NSL KDD Dataset
Open
intrusion_dataset
test_data
Open
Cancel

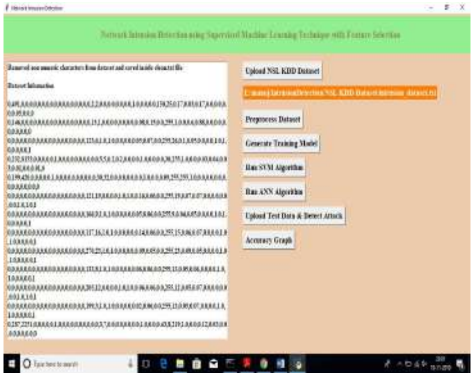
In above screen I am uploading ‘intrusion_dataset.txt’ file, after uploading dataset will get below screen


Network Intrusion Detection
Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection
Upload NSL KDD Dataset
Preprocess Dataset
Generate Training Model
Run SVM Algorithm
Run ANN Algorithm
Upload Test Data & Detect Attack
Accuracy Graph

Dataset loaded

C:/manoj/IntrusionDetection/NSL KDD Dataset/intrusion_dataset.txt

Now click on ‘Pre-process Dataset’ button to clean dataset to remove string values from dataset and to convert attack names to numeric values


Network Intrusion Detection
Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection
Upload NSL KDD Dataset
Preprocess Dataset
Generate Training Model
Run SVM Algorithm
Run ANN Algorithm
Upload Test Data & Detect Attack
Accuracy Graph

Removed non numeric characters from dataset and saved inside clean.txt file

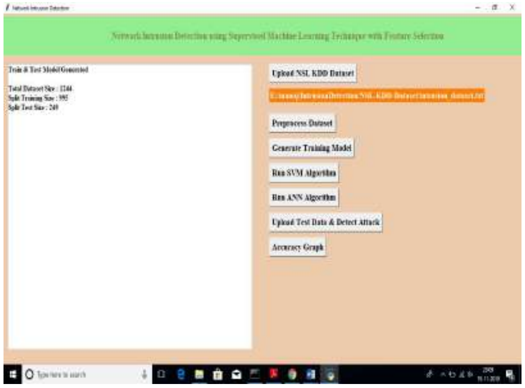
Dataset Information

[illegible]

C:/manoj/IntrusionDetection/NSL KDD Dataset/intrusion_dataset.txt

After pre-processing all string values removed and convert string attack names to numeric values such as normal signature contains id 0 and anomaly attack contains signature id 1.

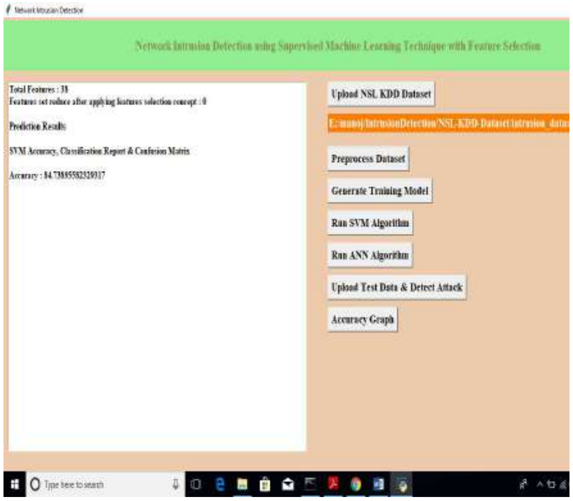
Now click on ‘Generate Training Model’ to split train and test data to generate model for prediction using SVM and ANN


Network Intrusion Detection
Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection
Upload NSL KDD Dataset
Preprocess Dataset
Generate Training Model
Run SVM Algorithm
Run ANN Algorithm
Upload Test Data & Detect Attack
Accuracy Graph
Type here to search

E:/manoj/IntrusionDetection/NSL-KDD-Dataset/intrusion_dataset.txt

```
Train & Test Model Generated

Total Dataset Size : 1244
Split Training Size : 995
Split Test Size : 249
```

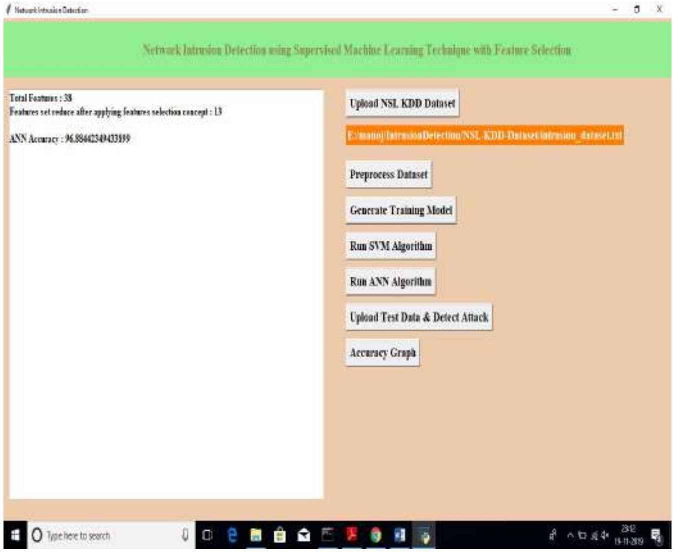
In above screen we can see dataset contains total 1244 records and 995 used for training and 249 used for testing. Now click on ‘Run SVM Algorithm’ to generate SVM model and calculate its model accuracy


Network Intrusion Detection
Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection
Upload NSL KDD Dataset
Preprocess Dataset
Generate Training Model
Run SVM Algorithm
Run ANN Algorithm
Upload Test Data & Detect Attack
Accuracy Graph
Type here to search

E:/manoj/IntrusionDetection/NSL-KDD-Dataset/intrusion_data

```
Total Features : 38
Features set reduce after applying features selection concept : 0

Prediction Results

SVM Accuracy, Classification Report & Confusion Matrix

Accuracy : 84.73895582329317
```

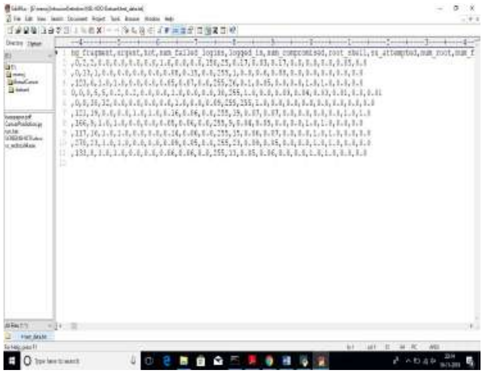
In above screen we can see with SVM we got 84.73% accuracy, now click on ‘Run ANN Algorithm’ to calculate ANN accuracy


Network Intrusion Detection
Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection
Upload NSL KDD Dataset
Preprocess Dataset
Generate Training Model
Run SVM Algorithm
Run ANN Algorithm
Upload Test Data & Detect Attack
Accuracy Graph
Type here to search

E:/manoj/IntrusionDetection/NSL-KDD-Dataset/intrusion_dataset.txt

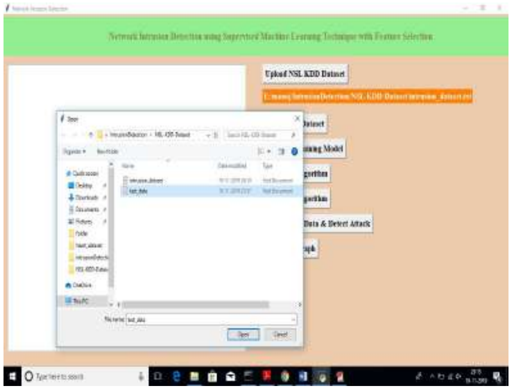
```
Total Features : 38
Features set reduce after applying features selection concept : 13

ANN Accuracy : 96.88442340433899
```

In above screen we got 96.88% accuracy, now we will click on ‘Upload Test Data & Detect Attack’ button to upload test data and to predict whether test data is normal or contains attack. All test data has no class either 0 or 1 and application will predict and give us result. See below some records from test data

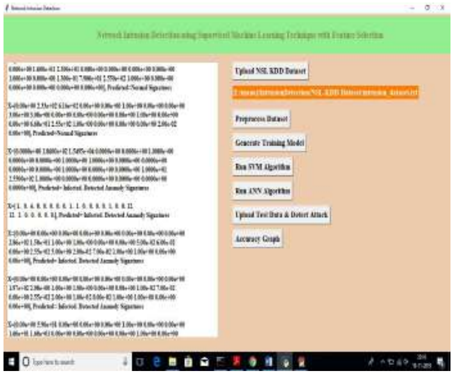


```
ng_fragment,urgent,hot,num_failed_logins,logged_in,num_compromised,root_shell,su_attempted,num_root,num_f
,0,2,2,0.0,0.0,0.0,0.0,1.0,0.0,0.0,150,25,0.17,0.03,0.17,0.0,0.0,0.0,0.05,0.0
,0,13,1,0.0,0.0,0.0,0.0,0.08,0.15,0.0,255,1,0.0,0.6,0.88,0.0,0.0,0.0,0.0,0.0
,123,6,1.0,1.0,0.0,0.0,0.05,0.07,0.0,255,26,0.1,0.05,0.0,0.0,1.0,1.0,0.0,0.0
0,0,0,5,5,0.2,0.2,0.0,0.0,1.0,0.0,0.0,30,255,1.0,0.0,0.03,0.04,0.03,0.01,0.0,0.01
,0,0,30,32,0.0,0.0,0.0,0.0,1.0,0.0,0.09,255,255,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
,121,19,0.0,0.0,1.0,1.0,0.16,0.06,0.0,255,19,0.07,0.07,0.0,0.0,0.0,0.0,1.0,1.0
,166,9,1.0,1.0,0.0,0.0,0.05,0.06,0.0,255,9,0.04,0.05,0.0,0.0,1.0,1.0,0.0,0.0
,117,16,1.0,1.0,0.0,0.0,0.14,0.06,0.0,255,15,0.06,0.07,0.0,0.0,1.0,1.0,0.0,0.0
,270,23,1.0,1.0,0.0,0.0,0.09,0.05,0.0,255,23,0.09,0.05,0.0,0.0,1.0,1.0,0.0,0.0
,133,8,1.0,1.0,0.0,0.0,0.06,0.06,0.0,255,33,0.05,0.06,0.0,0.0,1.0,1.0,0.0,0.0
```

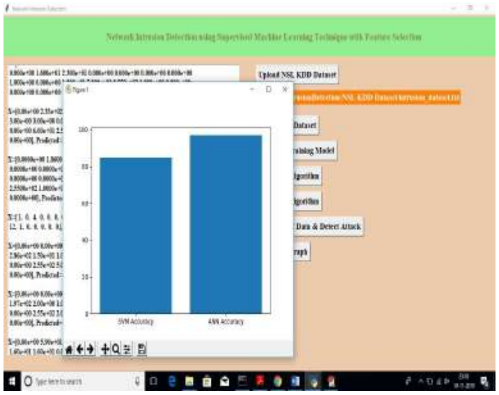
In above test data we don’t have either ‘0’ or ‘1’ and application will detect and give us result



In above screen I am uploading ‘test_data’ file which contains test record, after prediction will get below results



In above screen for each test data we got predicted results as ‘Normal Signatures’ or ‘infected’ record for each test record. Now click on ‘Accuracy Graph’ button to see SVM and ANN accuracy comparison in graph format



From above graph we can see ANN got better accuracy compare to SVM, in above graph x-axis contains algorithm name and y-axis represents accuracy of that algorithms

CONCLUSION

In this paper Intrusion detection is considered as a classification problem where each record can be classified into normal or a particular kind of intrusion. Intrusion detection using machine learning has attracted more and more interests in recent years. As an important application of machine learning, an accurate intrusion detection model is built by choosing an effective classification approach. This paper shows the comparison of the most well-known classification algorithms like C4.5 decision trees and Naive Bayes has been carried out . These algorithms are tested with the KDD data-set. Effective classifier is identified by comparing the performances based on the accuracy and confusion matrix. Performance calculation is done by considering only the important attributes for the intrusion detection. From the observed results it can be concluded that the C4.5 decision trees classifier outperforms other classifiers for the considered data-set and parameters. It has the accuracy of 99%.

REFERENCES

[1]. M. Almseidin, M. Alzubi, S. Kovacs and M. Alkasassbeh, "Evaluation of machine learning algorithms for intrusion detection system," 2017 IEEE 15th International Symposium on Intelligent Systems and Informatics (SISY), Subotica, 2017, pp. 000277- 000282.

[2]. Arul, Amudha & Subburathinam, Karthik & Sivakumari, S. (2013). Classification Techniques for Intrusion Detection an Overview. International Journal of Computer Applications. 76. 33-40. 10.5120/13334-0928.

[3]. Kanagalakshmi.R, V. Naveenantony Raj,‟ Network Intrusion Detection Using Hidden Naïve Bayes Multiclass Classifier Model,‟ International Journal of Science, Technology & Management ,Volume No.03, Issue No. 12, December 2014.

[4]. M. Alkasassbeh, G. Al-Naymat et.al,‛ Detecting Distributed Denial of Service Attacks Using Data Mining Technique,‛ (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, pp. 436-445,

2016. Science and Information Technologies, Vol. 6 (2), pp. 1096-1099, 2015.

[5]. Jasreena Kaur Bains ,Kiran Kumar Kaki ,Kapil Sharma,‟ Intrusion Detection System with MultiLayer using Bayesian Networks‟ , International Journal of Computer Applications (0975 – 8887) Volume 67– No.5, April 2013.

[6]. Dewan Md. Farid, Nouria Harbi, Mohammad Zahidur Rahman , Combining Naive Bayes and Decision Tree for Adaptive Intrusion Detection, Proc. of Intl. Journal of Network Security & Its Applications (IJNSA), Volume 2, Number 2, 2010, pp.12-25.

[7]. V. Hema and C. Emilin Shyni, ‟ DoS Attack Detection Based on Naive Bayes Classifier, ‟ Middle-East Journal of Scientific Research 23 (Sensing, Signal Processing and Security): 398-405, 2015.

[8]. Domingos P. and Pazzani M., Beyond Independence: Conditions for the optimality of the simple Bayesian Classifier, In proceedings of the 13th Intnl. Conference on Machine Learning, 1996, pp.105-110.

[9]. Yi-Chi Wu, Huei-Ru Tseng, Wu Yang* and RongHong Jan,ʻ DDoS detection and trackback with decision tree and grey relational analysisʻ, Int. J. Ad Hoc and Ubiquitous Computing, Vol. 7, No. 2, 2011.

[10]. Dewan Md. Farid, Nouria Harbi, Emna Bahri, Mohammad Zahid ur Rahman, Chowdhury Mofizur Rahman,‟ Attacks Classification in Adaptive Intrusion Detection using Decision Tree ,International Journal of Computer, Electrical, Automation, Control and Information Engineering, Vol:4, No:3, 2010.

[11]. Quinlan, C4.5: Programs for Machine Learning, 1993, Morgan Kaufmann Publishers, San Mateo, CA.

[12]. Sabhnani M, Serpen G(2003), Application of Machine Learning Algorithms to KDD Intrusion Detection Dataset within Misuse Detection Context, In Proc. of the Intl. Conference on Machine Learning, Models, Technologies and Applications, vol. 1, pp. 209–215.

[13]. Gharibian F, Ghorbani A.A , Comparative Study of Supervised Machine Learning Techniques for Intrusion Detection, Proc. of the Fifth Annual Conference on Communication Networks and Services Research, 2007, pp. 350–358.

[14]. Ohta S, R. Kurebayashi and K. Kobayashi. , Minimizing false positives of a decision tree classifier for intrusion detection on the internet, Journal of Networks System Management, vol.16, 2008, pp.399– 419. ISSN 1064-7570.

[15] M. Kemiche and R. Beghdad, Intelligent Systems in Science and Information 2014: Extended and Selected Results from the Science and Information Conference 2014, Cham: Springer International Publishing, ch. Towards Using Games Theory to Detect New U2R Attacks, pp. 351–367, (2015). [Online]. Available: http://dx.doi.org/10.1007/978-3-319-14654-6-22