International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 5, May 2020 ||

Comparative Analysis of Supervised Machine Learning Algorithms by Prediction of Network Attacks with Best Accuracy

Dr.Joe Prathap PM[1], Deepan.D[2], Giriram.S[2], Iyer Pradip.S[2], Jayandhan.S[2]

Associate Professor, Department of Information Technology, R.M.D Engineering College, Thiruvallur, Tamilnadu, India[1]

UG Student, Department of Information Technology, R.M.D Engineering College,Thiruvallur,Tamilnadu, India[2]

ABSTRACT: It is very difficult to create data-sets for the Intrusion Detection System and to know the possibilities of multiple attacks; a software to protect a network from intruders and perhaps insiders and to detect network intrusions. The aim of the intrusion detector is to be capable of distinguishing between good and bad connections by forming a predictive model. Generally, the network sectors have to predict whether the connection is attacked or not by using the KDDCup99 data set using machine learning techniques. To predict the better network packet connections with the help of machine learning techniques and obtain results based on best accuracy. To use the supervised classification machine learning algorithms in order to accurately predict DOS, R2L, U2R, Probe and overall attacks by proposing a machine learning based method. To understand the performance of multiple machine learning algorithms by comparing it withan evaluation classification report, a confusion matrix and a priority orderfor the given dataset. The efficiency of the proposed machine learning algorithm technique is understood by the results by means of best accuracy and precision through the F1 score calculated by Precision and Recall.

KEYWORDS: dataset, Machine learning Classification method, python, Best Accuracy result.

I. INTRODUCTION

Machine learning is the method to accurately predict the future data by the help of used past data.. The computers obtain the ability to understand data and learn without any programmed code by the help of machine learning. When a computer program is introduced to new data, multiple changes are observed which can be known by the help of various machine learning algorithms using python language. Different types of specialized algorithms are used to train the data. After training the data, the trained data is fed to an algorithm and the predictions over new data are done by the help of the algorithm. Machine learning is split into three categories i.e. supervised learning, unsupervised learning and reinforcement learning. Supervised learning method is the combination of both the given input data and the labeling of the the input data which has to be labeled by a human being beforehand. Unsupervised learning is a method without any labels. The data is provided to the learning algorithm whichperforms the clustering of the input data. Reinforcement learning is a method which can improve its own performance by the negative or positive feedback received from its environment dynamically.

A. Supervised Machine Learning

Supervised learning program occurs when a system is given input and output variables with the intentions of learning how they are mapped together, or related. The goal is to produce an enough mapping function that when new input is given, the algorithm can predict the output.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 5, May 2020 ||

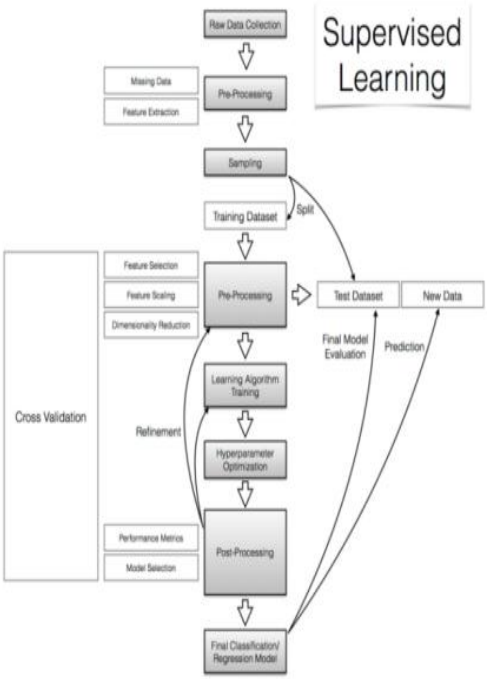

Figure: Supervised Machine Learning

B. Supervised Machine Learning

Supervised learning program occurs when a system is given input and output variables with the intentions of learning how they are mapped together, or related. The goal is to produce an enough mapping function that when new input is given, the algorithm can predict the output.

II. LITERATURE SURVEY

The vision of this paper is to understand and learn the input data and to use the various machine learning based techniques for better packet connection transfers and to predict the results obtained with best accuracy. To understand and evaluate the performance of various machine learning algorithms from the given data set. To show the effectiveness of the proposed machine learning algorithm technique by obtaining best accuracy and precision through F1 score.

This paper addresses the challenges to identify and predict all the different possibilities of attacks, as it is difficult to create data sets for the Intrusion Detection Systemto protect a network from intruders and perhaps insiders and to detect network intrusions.

A. DOMAIN OVERVIEW

Machine learning may be a sort of Artificial Intelligence (AI) that gives the ability to the computers in order to understand data and learn without any programmed code. It is implemented using python and various specialized algorithms are involved to train the data.It is classified as supervised and unsupervised learning techniques.Supervised learning method is the combination of both the given input data and the labeling of the the input data. Unsupervised learning is a method without any labels. The data is provided to the learning algorithm to cluster the data.

Figure: Domain Overview

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 5, May 2020 ||

B. CLASSIFICATION OF ATTACKS

The KDDCup99 is the data-set used for machine learning which has both normal and attack-type data. The KDDCup99 consists of 41 different features of data.The data generated by this data-set are labeled as either 'normal' or 'any type of attack'.

Denial of Service (DOS) : The intended users become unable to access the system as the machine or network is shutdown. In the case of a distributed denial-of-service attack, it becomes a very difficult task to prevent any attack by blocking a single source due to the flooding of the incoming network traffic.**Example : Syn flood**

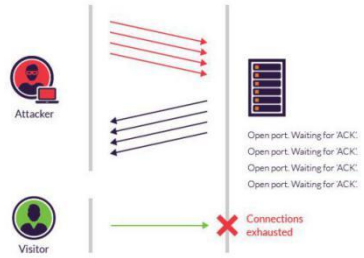

Figure: Denial Of Service attack

User to Root (U2R) : The main aim of attackers or hackers is to gain root access by obtaining the access rights of the system. These attacks are exploitation during which the hacker begins with taking the system with a traditional user account and then attempts to cause vulnerabilities in order to gain superuser privileges. **Example : Buffer overflow**

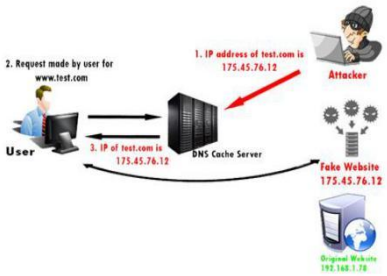

Figure: User to Root attack

Remote to User (R2L) : To make the system vulnerable by performing various actions in order to steal or view data illegally and manipulate the data by the use of any malicious software to cause damage to a targeted system. A type of attack that is performed to access a particular network address remotely and illegally. **Example : Guessing password**

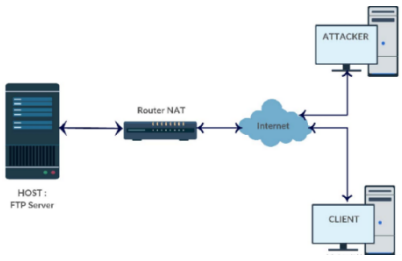

Figure: Remote to User attack

Probing : Probing is an exploitation of the targeted system or network in which the attacker scans a machine or a networking device in order to work out weaknesses or vulnerabilities which will can be used later to compromise the system.**Example : Port scanning**

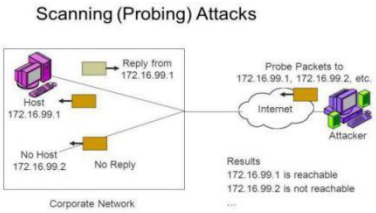

Figure: Probe attack

Overall network attack:

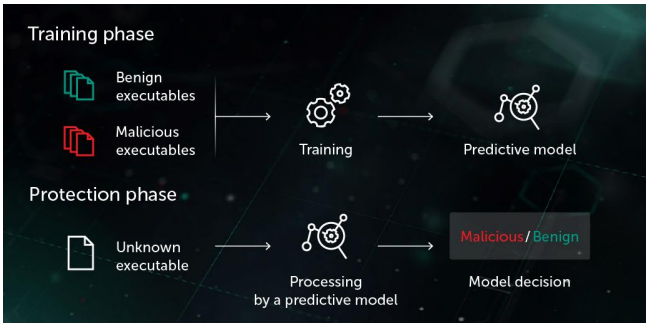

Figure: Overall network attacks

III. EXISTING SYSTEM

The proposal of the system is to outsource the attack detection function from the current network. It should be done in order to monitor the system to help us classify the protocol as well as jamming attacks The detection of the required attacks can be done in the frequency range of 2.402 to 2.422 GHz whereas the rest of the frequency range belongs to Wi-Fi communications. As a result of this analysis we can observe that the 20-MHz frequency band allows us to develop a classification model to overcome the problems caused by the utilization of the other channels that are occupied by other Wi-Fi communications.

As the Wi-Fi wireless networks are omnipresent, they are quite vulnerable and it can be easily attacked by hackers and attackers where the sensitive data as well assystem access is lost. In order to prevent these misuse of the networks some defense strategies have to be developed. The first defense mechanism of this study is to identify the occurrence of attacks and keep it independent of the communication networks by proposing a monitoring solution.The second defense mechanism is to propose a method which is able to classify different types of attacks. The received input data from the monitoring antenna is analyzed and monitored properly to perform these mechanisms. The input data is collected and classified based on a Support Vector Machine (SVM) classification model which helps usto predict the type of attack.

Disadvantages:

1. We cannot know how our model can evaluate in the case where an unknown attack occurs among all types of attacks by popular machine learning algorithms.
2. It cannot describe each category of DOS attacks like back, Neptune etc. based on the network connections.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 5, May 2020 ||

IV. PROPOSED SYSTEM

A.Exploratory Data Analysis :

Supervised machine learning classification algorithms will be used to obtain data sets and extract patterns, helping to avoid the attacks and make better decisions in the future and to obtain results with maximum accuracy.

B.Data Wrangling :

Load the data, check for cleanliness and then trim and clean the given dataset for analysis.

Figure: Data Analysis & Data Wrangling

C.Data Collection :

The collected input data is split intoTraining set and Testset. The training set and Test set are split in 7:3 ratio. Decision tree algorithms, K-Nearest Neighbor (KNN) and Support vector classifier (SVC) are applied on the Training set and results are obtained with best accuracy.

D. Pre-Processing :

The data is pre-pocessed in order to gain better results and make the algorithm efficient . Data integration, data reduction and data transformation are also performed for the given input dataset.

Figure: Data Collection and Pre-Processing

Advantages:

1. Improvement in the efficiency of the algorithm.
2. By comparison and analysis of the attacks with the logic, it is easy to obtain the best accuracy for each attack.

E. Applications:

1. Extract the structural features to learn a mapping function.
2. Predict the values from known targets.
3. Pre-labelled data trains a model to predict new outcomes.
4. Risk evaluation and prevent the access of third party.

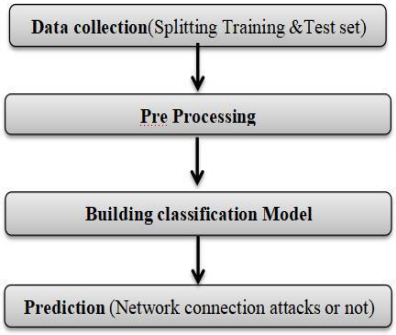

Figure: Proposed Model

V. RESULT AND DISCUSSION

A.Data Pre-processing:

After Pre-Processing:

```
In [32]: df.columns

Out[32]: Index(['duration', 'protocol_type', 'service', 'flag', 'src_bytes',
       'dst_bytes', 'land', 'Wrong_fragment', 'Urgent', 'hot',
       'num_failed_login', 'logged_in', 'num_compromised', 'root_shell',
       'su_attempted', 'num_root', 'num_file_creations', 'num_shells',
       'num_access_files', 'num_outbound_cmds', 'is_host_login',
       'is_guest_login', 'count', 'srv_count', 'serror_rate',
       'srv_serror_rate', 'rerror_rate', 'srv_rerror_rate', 'same_srv_rate',
       'diff_srv_rate', 'srv_diff_host_rate', 'dst_host_count',
       'dst_host_srv_count', 'dst_host_same_srv_rate',
       'dst_host_diff_ srv_rate', 'dst_host_same_src_port_rate',
       'dst_host_srv_diff_host _rate', 'dst_host_serror_rate',
       'dst_host_srv_serror_rate', 'dst_host_rerror_rate',
       'dst_host_srv_rerror_rate', 'class', 'DOSland', 'DOSlandclass', 'DOS',
       'R2L', 'U2R', 'Probe', 'attack'],
      dtype='object')
```

Figure: Data Pre-Processing

B.Prediction of DOS attack:

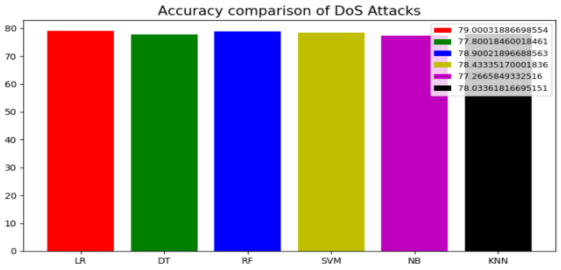

Figure: DOS attack

Result: The highest accuracy of DOS attack is **Logistic Regression algorithm.**

C.Prediction of R2L attack:

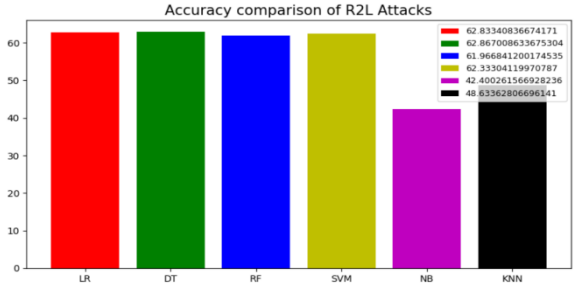

Figure: R2L attack

Result: The highest accuracy of R2L attack is **Logistic Regression algorithm.**

D.Prediction of U2R attack:

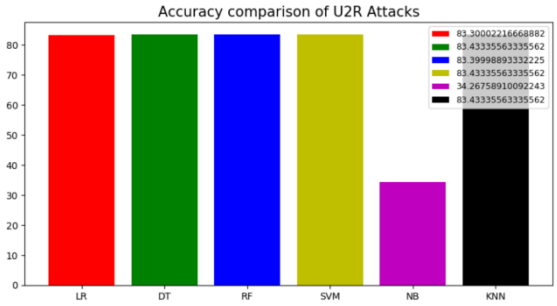

Figure: U2R attack

Result: The highest accuracy of U2R attack is **Decision Tree, Support Vector Classifier** and **K-Nearest Neighbor algorithm.**

E.Prediction of Probe attack:

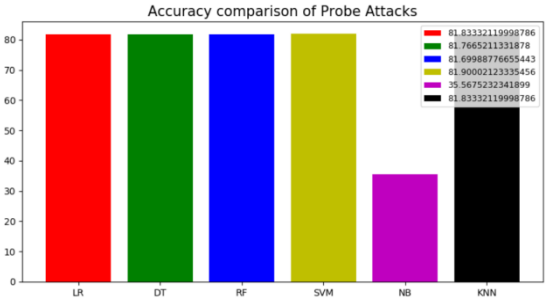

Figure: Probe attack

Result: The highest accuracy of Probe attack is **Support Vector Classifier algorithm.**

F.Prediction of Overall attacks:

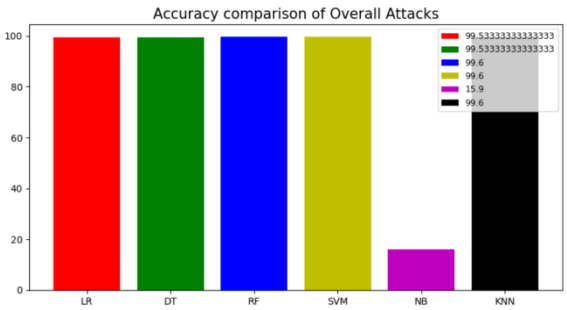

Figure: Overall attacks

Result: The highest accuracy from Overall attacks is **Support Vector Classifier algorithm.**

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 5, May 2020 ||

VI. INPUT AND OUTPUT

A. INPUT:

Figure: Input

B. OUTPUT:

Figure: Test 1

Figure: Test 2

VII. DESIGN ARCHITECTURE

The process of building something is said to be Design in engineering. The requirements are accurately collected and perfectly transformed into a finished product by the help of software design. Design helps us to understand the requirements by creating a model, providing detail about software data structure, architecture, interfaces and components that helps us to implement a system.

Ifdevelopment blends the attitude of change in settings, design, and integrating maps into one approach to development, then design is the act of taking the user profile settings values information and creating the planning of the merchandise to be developed.Systems design is therefore the method of defining and developing systems to satisfy specified requirements of the user.

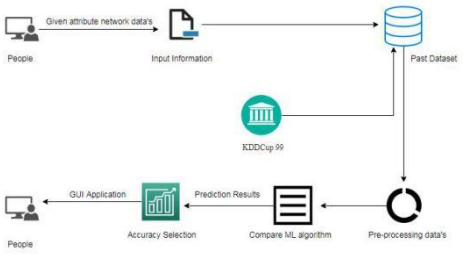

Figure: Design Architecture

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 5, May 2020 ||

VIII.FUTURE WORK

The best accuracy of public test set will be obtained by the highest accuracy score that can be evaluated by comparing all type of network attacks over each algorithm. Network sector wants to automate the process of dynamically detecting the attacks of packet transfers in real time based on the details of the network connection. To automate this process by predicting the results in a web application or desktop application. To optimize the work to implement in Artificial Intelligence environment.

IX.CONCLUSION

The evaluation of various processes such as data cleaning and data pre-processing, missing value, exploratory analysis is done. The best accuracy of public test set is obtained by the Random Forest algorithm with the highest accuracy score of 98.44% that is evaluated by comparing all type of network attacks over each algorithm. To provide an accurate prediction model which scopes artificial intelligence and avoids human error and shows early detection. This model helps us to understand the importance of machine learning in creating prediction and classification models which can be used in the network sectors to prevent human error.

REFERENCES

[1] R. Olfati-Saber, J. A. Fax, and R. M. Murray, "Consensus and cooperation in networked multi-agent systems," Proc. IEEE, vol. 95, no. 1,pp. 215–233, Jan. 2007.
[2] Y. Su and J. Huang, "Cooperative output regulation of linear multi-agent systems," IEEE Trans. Autom. Control, vol. 57, no. 4, pp. 1062–1066, Apr. 2012.
[3] Z. Feng, G. Hu, W. Ren, W. E. Dixon, and J. Mei, "Distributed coordination of multiple unknown Euler-Lagrange systems," IEEE Trans.Control Netw. Syst., vol. 5, no. 1, pp. 55–66, Mar. 2018.
[4] Z. Feng, C. Sun, and G. Hu, "Robust connectivity preserving rendezvous of multirobot systems under unknown dynamics and disturbances," IEEE Trans. Control Netw. Syst., vol. 4, no. 1, pp. 725–735, Dec. 2017.
[5] X. Dong and G. Hu, "Time-varying formation control for general linear multi-agent systems with switching directed topologies," Automatica, vol. 73, pp. 47–55, Nov. 2016.
[6] Z. Li, G. Wen, Z. Duan, and W. Ren, "Designing fully distributed consensus protocols for linear multi-agent systems with directed graphs," IEEE Trans. Autom. Control, vol. 60, no. 4, pp. 1152–1157, Apr. 2015.