Detection of Cyber Attack in Networks using Machine Learning Techniques

Meghana M[1], Poornima G B[2], Ramya R Shet[3], Rashika N[4], Abdul Razak M S[5]
[1, 2, 3, 4] [5]
B.E, Student of CSE, Asst. Professor. CSE, BIET, Davangere

Abstract: The use of recent innovations provides unimaginable blessings to individuals, organizations, and governments, be that because it might, messes some up against them. for example, the protection of serious information, security of place away data stages, accessibility of knowledge so forth. Digital concern, that created an excellent deal of problems individuals and institutions, has received A level that might undermine open and nation security by totally different gatherings, as an example, criminal association, good individuals and digital activists. the foremost common risk to a network's security is an intrusion like brute force, denial of service or maybe an infiltration from inside a network. this can be wherever machine learning comes into play.
Intrusion Detection Systems (IDS) has been created to take care of a strategic distance from digital assaults.
Keywords: Cyber-attack, Intrusion detection system, digital assaults.

I. INTRODUCTION

As the world becomes more and more digital, we have a tendency to square measure unlocking additional price and growth than ever before. However, a challenge that governments, enterprises and also as people leverage technology square measure perpetually facing is that the growing threat of cyberattacks that looms giant over United States. With AN increasing variety of cyberattacks targeting important networked resources that can't be detected by ancient network watching tools, it becomes important to explore and leverage subtle tools for detection and reportage of such attacks. Machine Learning (ML) is that the technology trends that have the potential to remodel the trendy security design landscape.

II. LITERATURE SURVEY

[1] Port Scanning is one amongst the foremost standard techniques attackers use to get services that they will exploit to interrupt into systems. All systems that area unit connected to a local area network or the web via an electronic equipment run services that hear well-known and not therefore well-known ports. By port scanning, the offender will realize the subsequent info concerning the targeted systems: what services area unit running, what users own those services, whether or not anonymous logins area unit supported, and whether or not bound network services need authentication. [2] Port scanning could be a common activity of extensive importance. it's typically utilized by laptop attackers to characterize hosts or networks that they're considering hostile activity against. therefore, it's helpful for system directors and alternative network defenders to observe portscans as attainable preliminaries to a lot of serious attack. it's additionally wide utilized by network defenders to know and realize vulnerabilities in their own networks. therefore, it's of extensive interest to attackers to see whether or not or not the defenders of a network area unit portscanning it frequently. However, defenders won't typically would like to cover their portscanning, whereas attackers can. For predictability, within the remainder of this paper, we'll speak of the attackers scanning the network, and therefore the defenders attempting to observe the scan.

III. CLASSIFIERS USED

A. Decision Tree

Is a tree-structured classifier, wherever internal nodes represent the options of a dataset, branches represent the choice rules and every leaf node represents the result. In a call tree, there are 2 nodes, that are the choice Node and Leaf Node. call nodes are accustomed create any call and have multiple branches, whereas Leaf nodes are the output of these choices and don't contain any longer branches. The decisions or the check are performed on the idea of options of the given dataset. It is a graphical illustration for obtaining all the attainable solutions to a problem/decision supported given conditions.

B. Random Forest

Random Forest may be a in style machine learning formula that belongs to the supervised learning technique. It may be used for each Classification and Regression issues in milliliter. it's supported the thought of ensemble learning, that may be a method of mixing multiple classifiers to unravel a posh drawback and to enhance the performance of the model. because the name suggests, "Random Forest may be a classifier that contains variety of call trees on numerous subsets of the given dataset and takes the typical to enhance the prognosticative accuracy of that dataset

C. Support Vector Machine

Support Vector Machine or SVM is one in every of the foremost commonplace supervised Learning algorithms, that's used for Classification furthermore as Regression problems. However, primarily, it's used for Classification problems in Machine Learning. The goal of the SVM formula is to make the best line or decision boundary which is able to segregate n-dimensional space into classes so as that we'll merely place the new info among the right category among the long run. This best decision boundary is termed a hyperplane. SVM chooses the acute points/vectors that facilitate in creating the hyperplane. These extreme cases are observed as support vectors, and thence formula is termed as Support Vector Machine.

IV. METHODOLOGY

A. Dataset

The datasets are download from unsw.adfa.edu that consists of 43 columns.
https://www.unsw.adfa.edu.au/unsw-canberra-cyber/cybersecurity/ADFA-NB15Datasets/NUSW-NB15_features.csv Dataset info

Number of variables 44

Number of observations 125972

Missing cells 0 (0.0%)

Duplicate rows 0 (0.0%)

Total size in memory 42.3 MB

Average record size in memory 352.0 B

B. Problem Statement

Anomaly-based intrusion detection approaches are suffering from consistent and accurate performance evolutions. Therefore, there is a need for a system to detect the type of cyber attack in network using Machine Learning techniques.

C. Proposed Solution

Intrusion Detection System (IDSs) and Intrusion bar Systems (IPSs) square measure the foremost vital defence tools against the subtle and ever-growing network attacks. Therefore, the model is made that predict the cyber attack with reduced execution time

D. Objectives

1) To apply pre-processing techniques on KDD dataset

2) To extract the most prominent features and apply classification techniques.

3) To build a model that can detect the type of cyber attack.

4) To compare the accuracy rates of different classifiers.

V. SYSTEM IMPLEMENTATION

A. Model Design

As shown in the below fig: 1.1 the designed model follows the data pre-processing steps and later is fed to the application building. First all the required libraries are imported such as pandas, numpy, seaborn, mathplotlib, pandas profiling and sklearn. The train and test dataset are loaded and read using pandas library. Binomial and multinimial classsification is performed on train and test dataset. Binomial classification is done to identify 2 targets in our case it is normal or attack. Binomial Classification In attack_class normal means 0 and attack means 1 and identify the count in both train and test dataset. Multinomial classification done to identify more than one targets that is Normal, DOS, PROBE, REL & E2R. Mutinomial Classification in attack_class normal means 0, DOS means 1, PROBE means 2, R2L means 3 and U2R means 4

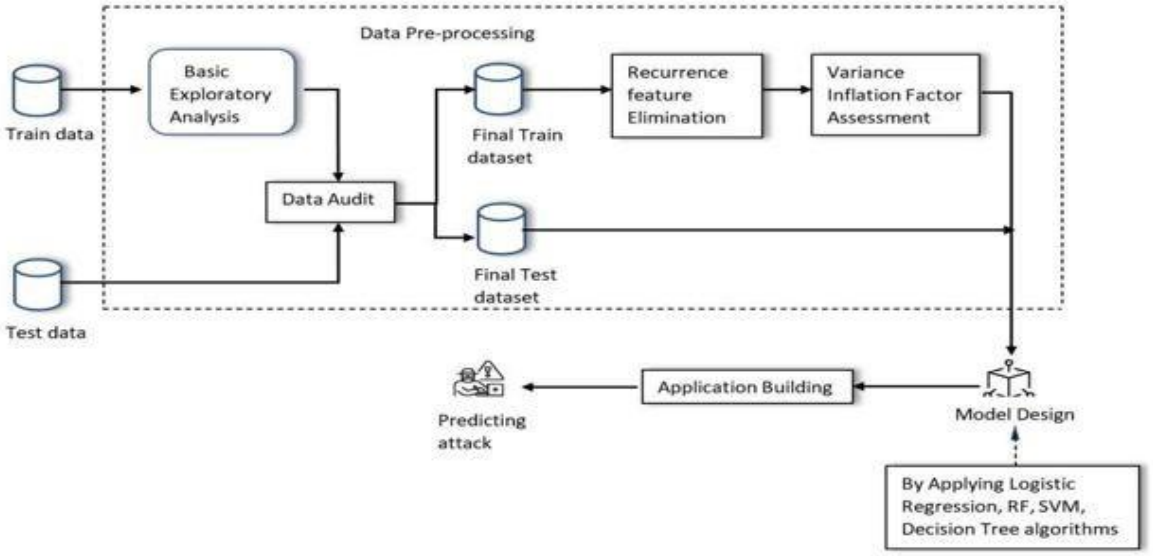


Fig 1.1: Model Design

B. Model Implementation

Model analysis is Associate in Nursing integral a part of the model development method. It helps to search out the most effective model that represents the information and the way well the chosen model can add the longer term. Here we've got taken four totally different classifier ways to make a model and eventually the one with high accuracy and best score is choosen.

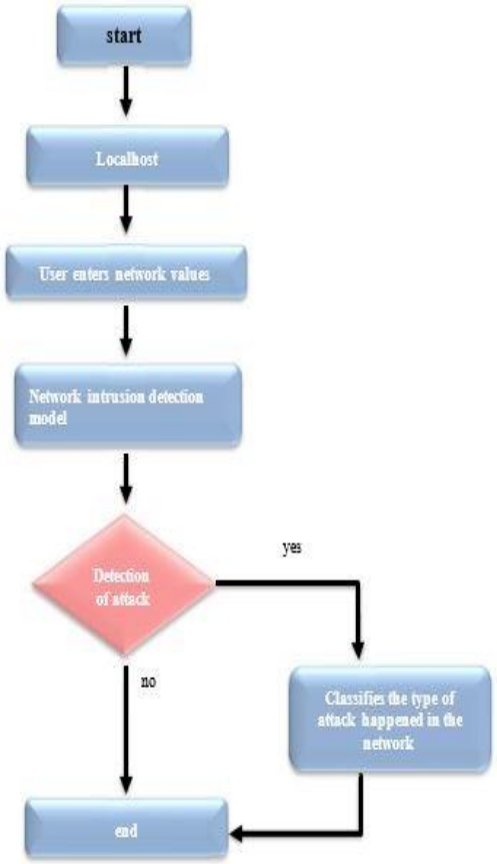
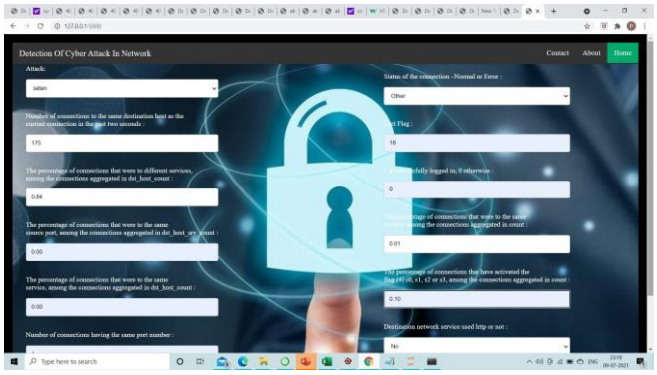


Fig 1.2: Flow diagram

VI. RESULT



OUTPUT: Attack class should be **PROBE**

VII. FUTURE SCOPE

A. Improving accuracy by using some variations or modifications for the existing model.
B. Support to detect other types of attacks in the dataset.
C. Commercialize the application using APIs which can be used in various industries like cyber-crime detection and FBI according to their needs.
D. Provide security measures to avoid misuse of technology

VIII. CONCLUSION

Right now, estimations of support vector machine, Random Forest and profound learning calculations hooked in to the dataset were introduced comparatively. Results show that the profound learning calculation performed basically preferred outcomes over SVM, RF and call tree. we tend to square measure aiming to utilize port sweep endeavours further as alternative assault sorts with AI and profound learning calculations, apache Hadoop and sparkle innovations along hooked in to this dataset presently. of these calculation helps U.S.A. to sight the cyber attack in network. It happens within the means that once we contemplate long back years there is also such a big number of attacks happened therefore once these attacks square measure recognized then the options at those values these attacks square measure happening are going to be hold on in some datasets. therefore, by exploitation these datasets we tend to square measure aiming to predict whether or not cyber attack is finished or not. These predictions are done by algorithms like SVM, RF and call tree. This paper helps to spot that formula predicts the simplest accuracy rates that helps to predict best results to spot the cyber attacks happened or not.

REFERENCES

[1] R. Saint Christopher, “Port scanning techniques and therefore the defense against them,” SANS Institute, 2001.
[2] S. Staniford, J. A. Hoagland, and J. M. McAlerney, “Practical machine-controlled detection of sneaky portscans,” Journal of pc Security, vol. 10, no. 1-2, pp. 105–136, 2002
[3] K. Graves, Ceh: Official certified moral hacker review guide: communicating 312-50. John Wiley & Sons, 2007.
[4] M. Baykara, R. Das¸, and I. Karado ˘gan, “Bilgi g ¨uvenli ˘gi sistemlerinde kullanilan arac¸larin incelenmesi,” in first international conference on Digital Forensics and