TREE BASED INTRUSION DETECTION SYSTEM

ADITHYA M
SRN:PES2UG19CS015
Computer science and Engineering
PES University
Bengaluru, India
pes2ug19cs015@pesu.pes.edu

MAHANTESH M
SRN:PES2UG19CS210
Computer science and Engineering
PES University
Bengaluru, India
pes2ug19cs210@pesu.pes.edu

***Abstract*—**

No firewall is foolproof, and no network is impenetrable. Attackers continuously develop new exploits and attack techniques designed to circumvent your defenses. Many attacks leverage other malware or social engineering to obtain user credentials that grant them access to your network and data. An Intrusion Detection System is used to detect all types of malicious network traffic and computer usage that cannot be detected by a conventional firewall. This includes network attacks against vulnerable services, data driven attacks on applications, host-based attacks such as privilege escalation, unauthorized logins and access to sensitive files, and malware (viruses, Trojan horses, and worms). It will consist of an agent that will collect the necessary information from the stream of monitored events and then the analysis engine will detect the signs of intrusion on the system, then a response will be generated that will signify whether the system has been breached or not.

Our research Model involved creation of multiple machine learning models such as XG Boost, Decision Tree, Random Forest, Extra Tress, and these models which undergo some training, testing and optimization were able to accurately guess the type of intrusion with the lowest and highest accuracy after optimization being 0.949 and 0.971 respectively.

We then to get the best of all worlds combine these four models in-order to obtain an Ensemble model which can predict the type of system intrusion with an accuracy of about 0.968.

I. INTRODUCTION

In current time as more and more data and information is getting stored online in one or other form of services it becomes necessary that the organization data which is stored on these services or even personal data of the user though there are many services for this specific purpose, they can also be exploited in one way or another. Intrusion detection is one of the dynamic network threat detectors which analyzes and monitors network and alerts if there is an abnormality in the behavior or any violation of network security policy.

There are mainly one types of intrusion detection which will bediscussed here:

Network based IDS: Network intrusion detection systems (NIDS) are set up at a planned point within the network to examine traffic from all devices on the network. It performs an observation of passing traffic on the entire subnet and matches the traffic that is passed on the subnets to the collection of known attacks. Once an attack is identified or abnormal behavior is observed, the alert can be sent to the administrator. An example of a NIDS is installing it on the subnet where firewalls are to see if someone is trying to crack the firewall.

The other types of systems can be:

Host Intrusion Detection System (HIDS): Host intrusion detection systems (HIDS) run on independent hosts or devices on the network. A HIDS monitors the incoming and outgoing packets from the device only and will alert the administrator if suspicious or malicious activity is detected. It takes a snapshot of existing system files and compares it with the previous snapshot. If the analytical system files were edited or deleted, an

alert is sent to the administrator to investigate. An example of HIDS usage can be seen on mission-critical machines, which are not expected to change their layout.

Hybrid Intrusion Detection System: Hybrid intrusion detection system is made by the combination of two or more approaches of the intrusion detection system. In the hybrid intrusion detection system, host agent or system data is combined with network information to develop a complete view of the network system. Hybrid intrusion detection system is more effective in comparison to the other intrusion detection system. Prelude is an example of Hybrid IDS. There are two types of detection methods which can be performed by either one of those: Signature based, and Anomaly based.

Application Protocol-based Intrusion Detection System (APIDS): Application Protocol-based Intrusion Detection System (APIDS) is a system or agent that generally resides within a group of servers. It identifies the intrusions by monitoring and interpreting the communication on application-specific protocols. For example, this would monitor the SQL protocol explicit to the middleware as it transacts with the database in the web server.

- **K-Nearest Neighbors:** It is one of the simplest machine learning algorithms, based on the supervised learning method, is K-Nearest Neighbor's non-parametric nature means that is makes no assumptions about the underlying data. It is also known as a lazy learner algorithm since it saves the training dataset rather than learning from it immediately. Instead, it uses the dataset to execute an action when classifying data.
- **Extra Trees:** A decision tree-based ensemble learning approach called Extra Trees Class is used. Like Random Forest, Extra Trees Classifier randomizes some decisions and subsets of data to reduce over-learning from the data and overfitting.
- **Decision Trees:** A decision support tool known as a decision tree employs a tree-like model to represent options and their potential outcomes, including utility, resource costs and chance event outcomes. One technique to show an algorithm that solely uses conditional control statements is to use this method.
- **Random Forest:** The supervised learning method includes the well-known machine learning algorithm Random Forest. It may be applied to ML Classification and Regression issues. Its foundation is the idea of ensemble learning, which is the act of mixing many classifiers to solve a challenging issue and enhance the performance of the model.
- **XGBoost:** Extreme Gradient Boosting (XGBoost) is a distributed, scalable gradient-boosted decision tree ma- chine learning framework. The top machine learning library for regression, classification, and ranking issues, it offers parallel tree boosting.

II. NEED

- Now a day's internet has become part of our daily life infect, the business world is getting connected to Internet. Number of peoples are getting connected to the Internet every day to take advantage of the new business model which is known as e-Business.
- Connectivity enhancement has therefore become very critical aspect of today's e-business. There are two phases of business on the Internet. First phase is the Internet brings in outstanding potential to business in terms of reaching the users and at the same time it also brings a lot of risk to the business. There are both harmless and harmful users on the Internet. Whereas an organization makes its information system accessible to harmless Internet users. Malicious users or hackers can also get an access to organization's internal systems in various reasons. These are,
  - Software bugs called vulnerabilities in a system.
  - Failure in administration security
  - Leaving systems to default configuration.
- The intruders are use different types of techniques like Password cracking, peer-to-peer attack, sniffing attack, Dos attacks, Eavesdropping attack, Application layer attack etc. to exploit the system vulnerabilities mentioned above and compromise critical systems. Therefore, there required to be some kind of security to the private resources of the organization from the Internet as well as from users inside the organization.

III. METHODOLOGY

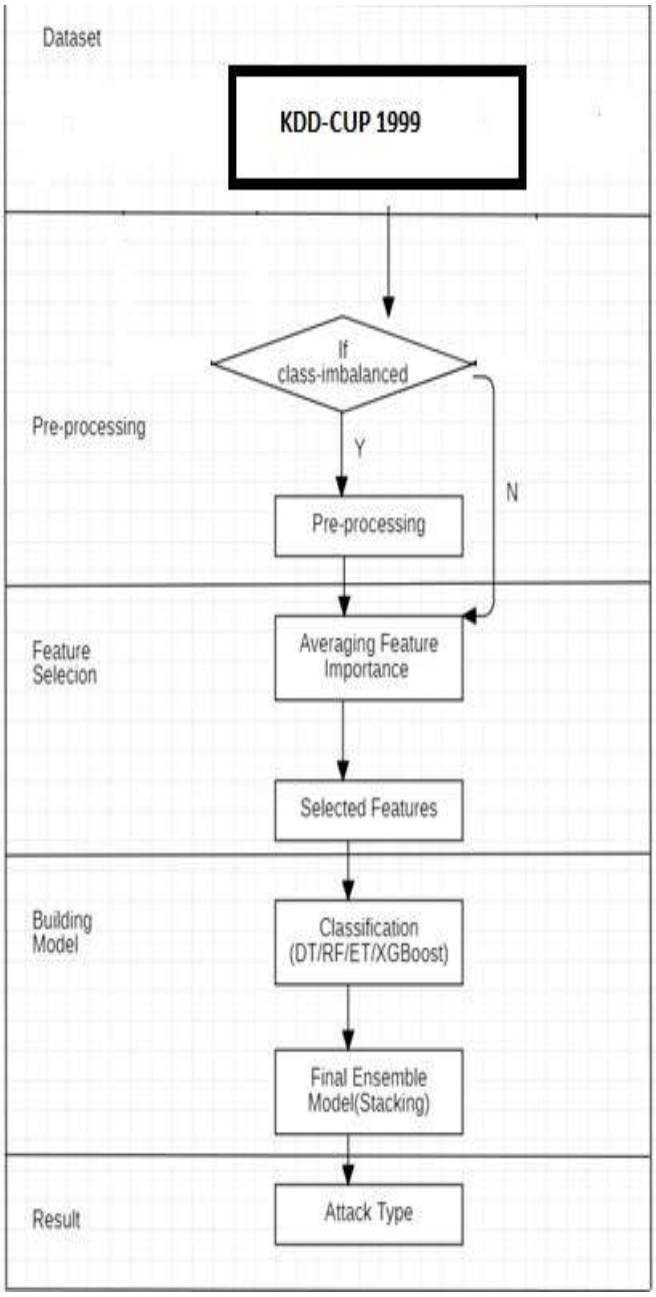

Figure 1: Model Architecture

The fundamental strategy entails:

- **Data selection:** we have chosen KDD-CUP 1999 dataset
  The no of data points are: 494021
  The no of features are: 42
  Some of the features are: ['duration', 'protocol type', 'service', 'flag', 'src_bytes', 'dst_bytes', 'land', 'wrong fragment', 'urgent', 'hot']
- Then, using k-means, data sampling is conducted for eachof these datasets
- **Preprocessing:** Procedures such as eliminating null values and z-score normalization are performed next.
- **Feature Engineering:** This phase entails choosing crucialcharacteristics needed for the model's training.
- **Machine learning:** Building the ML model is the next phase. We have selected five machine learning (ML) methods for this task:
  1) **XGBoost,**
  2) **Decision Trees,**
  3) **Random Forest**,
  4) **Extra Trees,**
  5) **Ensemble Model**
- The next step involves building the ML model. For this we have chosen 4 ML algorithms namely XG Boost, Decision Trees, Random Forest and Extra Trees.
- For each algorithm we have also done the optimization using the Bayesian optimization technique.
- Next, we have combined all 4 algorithms into one ensemble model. and we have applied the Bayesian optimization.
- Then Finally the results are displayed.

IV. RESULTS

The dataset consists of twenty-three different types of attacks and their frequency of these attacks varies as such

normal.	87832
neptune.	51820
back.	968
teardrop.	918
satan.	906
warezclient.	893
ipsweep.	651
smurf.	641
portsweep.	416
pod.	206
nmap.	158
guess_passwd.	53
buffer_overflow.	30
warezmaster.	20
land.	19

imap.	12
rootkit.	10
loadmodule.	9
ftp_write.	8
multihop.	7
phf.	4
perl.	3
spy.	2

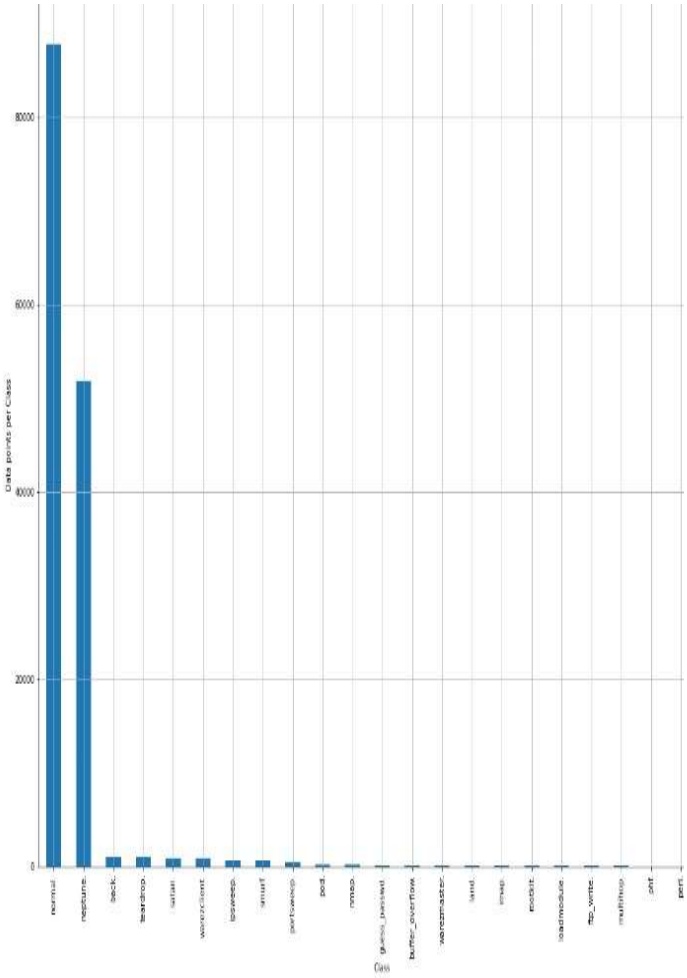

Figure 2: Description of attacks

Table 1: Performance Grades of XG Boost

	Accuracy	Precision	Recall	F1Score
XG Boost pre-optimization	0.966	0.955	0.966	0.959
XG boost post optimization	0.976	0.976	0.970	0.972

Table 2: Performance Grades of Random Forest

	Accuracy	Precision	Recall	F1Score
Random Forest pre-optimization	0.968	0.962	0.968	0.964
Random Forest post optimization	0.966	0.953	0.966	0.958

Table 3: Performance Grades of Decision Tree

	Accuracy	Precision	Recall	F1Score
Decision Tree pre-optimization	0.971	0.965	0.971	0.967
Decision Tree post optimization	0.949	0.937	0.949	0.941

Table 4: Performance Grades of E Tree

	Accuracy	Precision	Recall	F1Score
E Tree pre-optimization	0.973	0.969	0.973	0.970
E Tree post optimization	0.971	0.963	0.971	0.965

Table 5: Performance Grades of Ensemble Model

	Accuracy	Precision	Recall	F1Score
Ensemble learning pre-optimization	0.966	0.953	0.966	0.958
Ensemble learning post optimization	0.968	0.955	0.968	0.961

From the above 5 tables we can visualize that the highest accuracy was given by ET classifier before optimization and XG Boost gave the highest accuracy.

XG boost gave an accuracy of 0.966 and precision of 0.955 before optimization and after optimization it gave an accuracy of 0.976 and precision of 0. 976.So from this we can interpret that this model performed better after optimization.

Random Forest gave an accuracy of 0.968 and precision of 0.962 before optimization and after optimization it gave an accuracy of 0.966 and precision of 0. 953.So from this we can interpret that this model performed better before optimization this might be due to overfitting of data.

Decision Tree gave an accuracy of 0.971 and precision of 0.965 before optimization and after optimization it gave an accuracy of 0.949 and precision of 0. 937.So from this we can interpret that this model performed better before optimization this might be due to overfitting of data.

Extra Trees classifier gave an accuracy of 0.973 and precision of 0.969 before optimization and after optimization it gave an accuracy of 0.971 and precision of 0. 963.So from this we can interpret that this model performed better before optimization this might be due to overfitting of data.

Ensemble Model which is combination of the above 4 models gave an accuracy of 0.966 and precision of 0.953 before optimization and after optimization it gave an accuracy of 0.968 and precision of 0. 955.So from this we can interpret that this model performed better after optimization.

V. DISCUSSION

From the above results section we can conclude that the models XG Boost, and Extra Trees Classifier provided with better results than that of Random Forest classifier and Decision Tree.

The model which performed the best pre-optimization was XG Boost.

The model which performed the best post-optimization was also XG Boost itself.

Random Forest classifier and Decision Tree models accuracy and precision have decreased after the optimization this might be due to the case of overfitting of data.

Coming to the ensemble stacking model which is the combination of all the four models i.e., XG Boost, Extra Trees classifier, Random Forest classifier and Decision Tree provided great accuracy. The accuracy also increased after the optimization

The flaws or drawbacks from this experiment are as follows:

1)Since the dataset which was chosen for this experiment was prepared in the year 1999 so the models might not be able to detect newer kinds of attacks

2)Real-time testing was also not done.

Further research work can be done as follows

1)picking up a newer dataset which contains newer kinds of attacks

2)picking up a model which can provide better results and accuracy

3)Exploring the models which can be used to determine attacks in the case of connected vehicles, IoT etc.

VI. ACKNOWLEDGEMENT

I would like to express my deep gratitude to Dr Arti Arya, my research supervisors, for their patient guidance, enthusiastic encouragement, and useful critiques of this research work.

I take this opportunity to thank Dr. Sandesh B J, Chairperson, Department of Computer Science and Engineering, PES University – Electronic City Campus, for all the knowledge and support I have received from the department.

I am deeply grateful to Dr. M. R. Doreswamy, Chancellor, PES University, Prof. Jawahar Doreswamy, Pro Chancellor – PES University, Dr. Suryaprasad J, Vice-Chancellor, PES University for providing to me numerous opportunities and enlightenment every step of the way.

Nobody has been more important to me in the pursuit of this project than the members of my family. I would like to thank my parents, whose love and guidance are with me in whatever I pursue. They are the ultimate role models.

VII. References

[1]. A Unifying Review of Deep and Shallow Anomaly Detection Lukas Ruff, Jacob R. Kauffmann, Robert A. Vandermeulen, Gregoire Montavon, Wojciech Samek, ´ Member, IEEE, Marius Kloft* , Senior Member, IEEE, Thomas G. Dietterich* , Member, IEEE, Klaus-Robert Muller ¨ * , Member, IEEE. 8 February , 2021

[2] Anormal event detection in videos based on deep learning: School of Artificial Intelligence , Shenzhen Polytechnic,Shenzhen:Qinmin Ma: 5 August 2021

[3].Automatic Detection of Anomalies in video surveillance using artificial intelligence Sreedevi R Krishnan et al 2021 .

[4] B Ravi Kiran, Dilip Mathew Thomas, Ranjith Parakkal, “An overview of deep learning based methods for unsupervised and semisupervised anomaly detection in videos”, MDPI Journal of Imaging,arXiv:1801.03149v1, 2018