Machine-Learning Based TCP Security Action Prediction

Shivani Sharma[#1], Ankita Singh[*2], Muskan Singh[#3], Bhagyashree Solanki[*4],

Prof. Ganga Yadawad[*5]

[1,2,3,4]Students & [5]Assit. Prof. of Department of Computer Engineering,
JSPM's Jayawantrao Sawant College of Engineering
Hadapsar, Pune-28
Savitribai Phule Pune University, Pune

***Abstract*— As the internet continues to grow and evolve, the number of security incidents and cyber-attacks is also increasing. To combat these issues, computer network security is becoming more important than ever. One way to protect networks is through the use of a TCP firewall. These devices control the transmission of data according to specific rules, but traditionally this process is left up to network administrators. However, with the vast amount of data constantly flowing through the internet, this can be a daunting task. To address this problem, researchers are exploring the use of machine learning to automate this process and to improve network security. Machine learning algorithms can be trained to analysed network data and make decisions about which packets to allow or block, potentially reducing the workload for network administrators. Additionally, by using machine learning, the firewall can be more adaptive and responsive to new and emerging threats, improving overall network security.**

To achieve this goal, researchers will be using a dataset provided by UCI machine learning repository that contains TCP transmission characteristics. They will be implementing various machine learning models such as neural network, support vector machine (SVM), AdaBoost, and Logistic regression. The models will be evaluated and analysed for interpretability. By using the idea of ensemble-learning, the final result is expected to have an accuracy score of over 98%. This will greatly improve the efficiency of the firewall and will be able to predict the security action of TCP packets more accurately.

Keywords-TCP Security Action; Firewalls, Machine Learning; Ensemble learning; Prediction; Cyber Security

I. INTRODUCTION

The concept of a network firewall first appeared in the 1980s, and as firewall technology has continued to advance, new features like packet filtering, encryption, antivirus software, and proxy servers have been added. Software specifically designed for packet filtering can be used to see the header of packets passing through the firewall. It will only allow packets that are deemed secure to pass through; otherwise, it will trash them. Usually, the process requires a lot of calculation. However, using AI technology, a smart firewall can more effectively recognize network behavior with characteristic values. One of the developments in firewall technology is the use of intelligent firewalls. In addition to enhancing the firewall's functionality, artificial intelligence can provide it the ability to learn on its own and thwart the most recent network attacks.

Nowadays, there are an increasing number of cyberattacks, such as the Distributed Denial of Service (DDoS) attack, which is one of the most challenging security issues to resolve. DDoS assaults will become a significant concern with the development of the internet of things. The development of firewall technology will therefore be even more crucial in the 5G future. With the advent of 5G, the number of devices connected to the internet will increase exponentially, leading to a surge in the amount of data being transmitted. This will make it even more challenging for the firewall to keep up with the rate of data transmission. By using machine learning, firewall can process the data more quickly and make decisions on which packets to allow or block. Moreover, it will be able to detect and block new and emerging threats in real-time. This will greatly improve the efficiency of the firewall and will be able to predict the security action of TCP packets more accurately.

II. LITERATURE SURVEY

- The diversification of network services has generated massive amounts of Internet data, and traditional network security technologies have been difficult to meet the current needs of network security in terms of performance and self-adapt ability, is Discussed in paper [1] by the author Yong He. It's based on the key technology of Network Security.
- In this paper[2] author Sushant Sharma, Pavol Zavarsky, Sergey Butakov discussed about three tested algorithms, the J48 decision tree algorithm provided the highest True Positive rate, Precision, and Reca. The system introduces in this uses Classification Model for classifying authorized and unauthorized user.
- The author Agus Kurniawan, Marcel Kyas discusses the experimental results show that our proposed system can address
  security issues on machine learning computation with low time consumption is analysed in paper Securing Machine Learning Engines in IoT Applications with Attribute Based Encryption [3]
- The author Sean Oesch, Robert Bridges, Jared Smith, Justin Beaver, John Goodall,
  Kelly Huffer, Craig Miles, Dan Scofield of the paper [4] An Assessment of the Usability of Machine Learning Based Tools for the Security Operations Centre introduces that analysts lacked a clear mental model of how these tools generate scores, resulting in mistrust and/or misuse of the tools themselves.
- The author AKM Iqtidar Newaz, Amit Kumar Sikder, Mohammad Ashiqur Rahman discusses We trained Health Guard with data collected for eight different smart medical devices for twelve benign events including seven normal user activities and five
  disease affected events. [5].

III. METHODOLOGY

- **Methodologies: -**
  - In order to predict TCP security actions, it is recognized that this is a multi-class classification problem, which opens up the opportunity to use a variety of machine learning algorithms.
  - Since each algorithm has its own strengths and weaknesses, it makes sense to use multiple models. By using ensemble methods, which combine the predictions of multiple models, it is possible to create an integrated classifier that can achieve the highest level of accuracy.

- **Algorithms and Techniques: -**

1. **Support Vector Machine:**
   - In machine learning, support-vector machines (SVMs, also support-vector networks) are supervised learning models with associated learning algorithms that analyse data used for classification and regression analysis.
   - Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting).
   - An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on the side of the gap on which they fall.

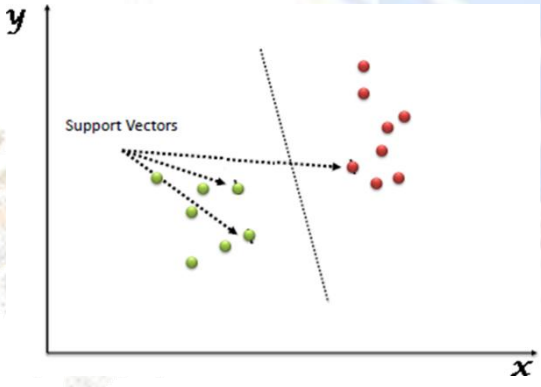

Fig 3.1 SVM Algorithm

2. **Naïve Bayes:**
   - It is a classification technique based on Bayes' Theorem with an independence assumption among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter.
   - Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.
   - An NB model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

$$P(A|B) = P(B|A) * P(A)/P(B)$$

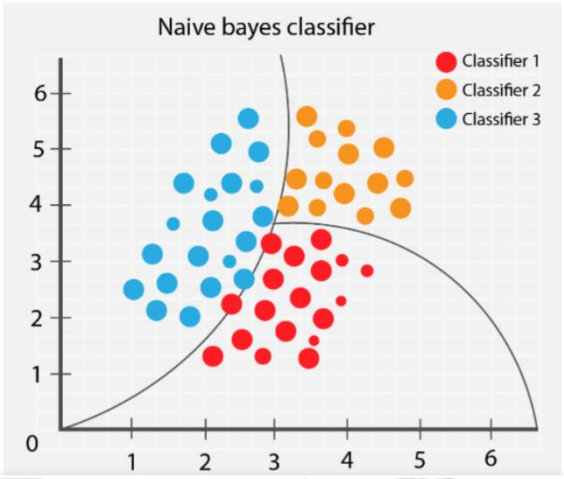

Fig. 3.2 Naïve Bayes Algorithm

3. **Decision Tree:**
   - Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.
   - In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.
   - The decisions or the test are performed on the basis of features of the given dataset. It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions. It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.

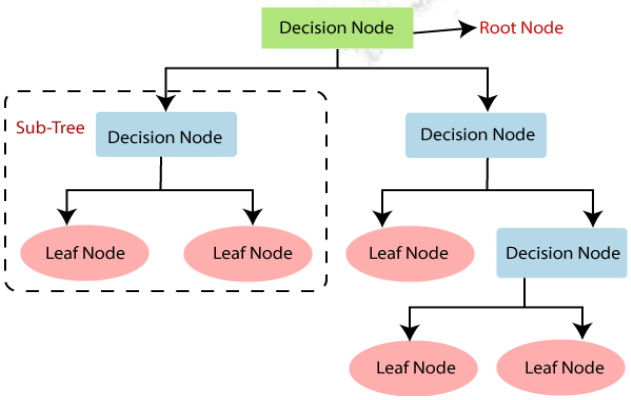

Fig. 3.3 Decision Tree Algorithm

IV. MATHEMATICAL MODEL

Let

S be Closed system defined as, S = Ip, Op, Ss, Su, Fi, A

To select the input from the system and perform various actions from the set of actions A so that Su state can be attained.

S=Ip, Op, Ss, Su, Fi, A

Where,

IP1=Username, Password, Parameters Set of actions=A=F1, F2,F3,F4 Where

F1= Image Capture F2= Preprocessing F3= Segmentation F4= Disease Detection S=Set of users

Ss=rest state, capturing image, processing image, detection of disease Su- success state is successful analysis

Fi- failure state Objects:

1) Input1: Ip1 = Username, Password
2) Input2: Ip2= Image
1) Output1: Op1 = Data Processing
2) Output2: Op2 = Feature Extraction
3) Output3: Op3 = Classification.

V. PROPOSED SYSTEM

This research focuses on the accuracy of the ensemble model for TCP security action prediction is determined by comparing the predicted labels from the model with the actual labels of the testing dataset. The overall architecture has been described in Figure 1. To differentiate between the classes "allow", "deny" and "drop", multiple models were tested. As a result, Support Vector Machine, Neural Network, and Logistic Regression performed the best individually. These models were chosen for the ensemble method as they gave the best results individually. However, it is possible to add more models if desired. The key is to find the right combination of models that will produce the best results. The grid search method was used to optimize the performance of each model and find the best set of parameters. Additionally, various ensemble techniques were applied such as bagging and boosting.

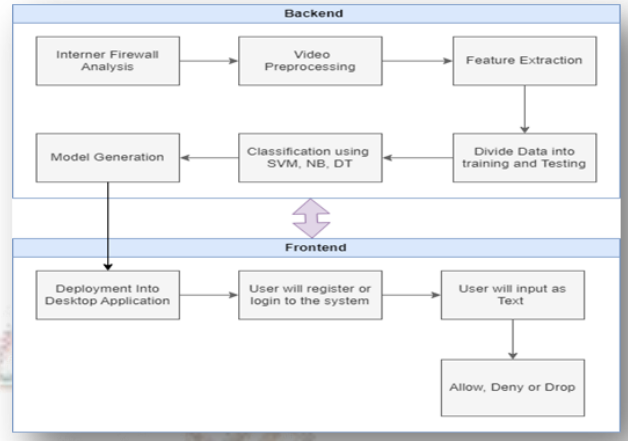

Fig.1: System Architecture

The final ensemble model is a combination of these well-performed models which were trained and evaluated on a dataset provided by UCI machine learning repository. The final result is expected to have an accuracy score of over 98%. This will greatly improve the efficiency of the firewall and will be able to predict the security action of TCP packets more accurately.

VI. RESULTS

- This work aims to address the challenge of analyzing firewall logs using machine learning (ML) and deep learning (DL) techniques. Firewall logs contain a vast amount of information about network traffic and security events, making manual analysis time-consuming and error-prone. By leveraging ML and DL, the objective is to develop multiclass models that can effectively analyze firewall logs and classify the appropriate actions to be taken in response to received sessions.

- The first step in this work involves data pre-processing and feature extraction from the firewall logs. The logs may include details such as source and destination IP addresses, port numbers, timestamp, packet size, and protocol type. These features are extracted and transformed into a suitable format for ML and DL algorithms to process.

- This work aims to tackle the difficulty of analyzing firewall logs using ML and DL by building multiclass ML and DL models that can analyze firewall logs and classify the actions to be taken in response to received sessions as “Allow”, “Drop”, “Deny”.

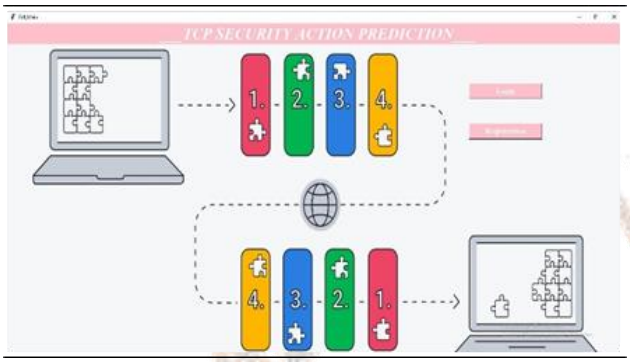

Figure 6.1: GUI Main

Figure 6.2. Registration

Figure 6.3. Login

Figure 6.4. SVM Prediction

Figure 6.5 DT Prediction

Figure 6.6 NB Prediction

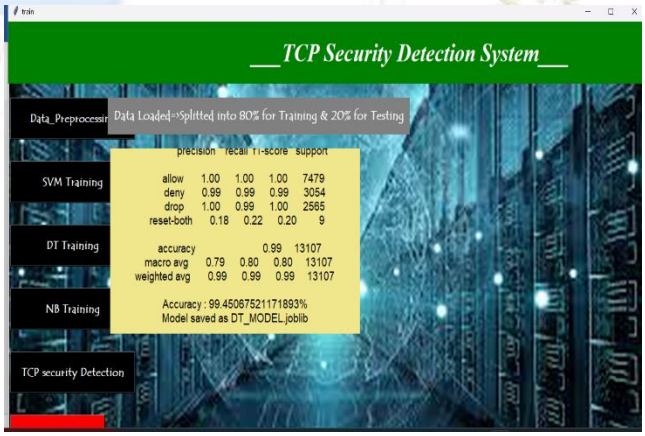

Figure 6.7 DT Report Analysis

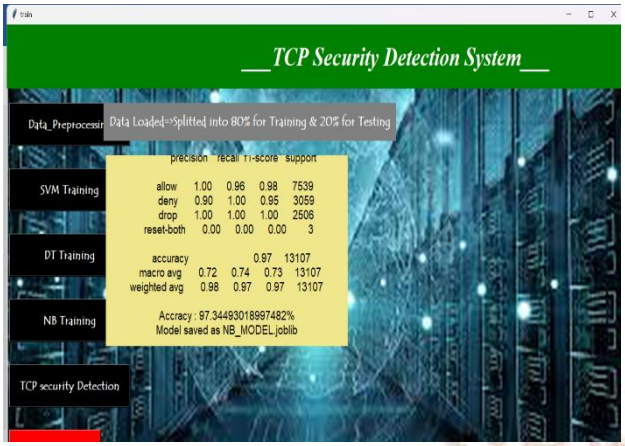

Figure 6.8 NB Report Analysis

- Our study also provides valuable insights into the field of intelligent firewalls and internet security technology.
- Our project aims to enable users to have more control over the permissions and access granted to apps and websites on their devices. By identifying and isolating only the necessary resources required by a specific app or website, users will be able to block any unwanted or unnecessary access. For instance, if someone downloads the Truecaller app, they will be able to deny access to non-essential resources such as images, resulting in improved security and privacy.
- By combining multiple machine learning models together, the accuracy of identifying security actions in TCP networks can be greatly improved and can exceed 98%.

VII. CONCLUSION

Our research looks into using machine learning models for predicting TCP security actions prediction. By refining and optimizing each model through pre-processing, the prediction accuracy of each model can be improved to over 98%. The result from the AdaBoost model provides a solution to the problem of class imbalance in the classification task. By using an ensemble learning technique, the prediction accuracy of TCP security action is able to reach over 98%. Furthermore, the ensemble model is able to classify all classes, regardless of how unevenly the classes are distributed in the dataset. Our research also makes a positive contribution to the study of intelligent firewalls and internet security technology.

VIII. ACKNOWLEDGMENT

We would like to express our gratitude to the researchers and publishers for making their resources available. We also appreciate the guidance and feedback provided by the reviewers and thank the college authorities for providing the necessary infrastructure and support throughout the research process.

IX. REFERENCES

[1] Yong He “Research on the Key Technology of Network Security Based on Machine Learning”,2021
[2] Sushant Sharma “Machine Learning based Intrusion Detection System for Web-Based Attacks”,2020.
[3] Agus Kurniawan “Securing Machine Learning Engines in IoT Applications with Attribute-Based Encryption”,2019.
[4] Jared Smith “An Assessment of the Usability of Machine Learning Based Tools for the Security Operations Center”,2020.
[5] Amit Kumar Sikder “Health Guard: A Machine Learning-Based Security Framework for Smart Healthcare Systems”,2019.
[6] Uday Hiwarale (2020) A brief overview of the TCP/IP model, SSL/TLS/HTTPS protocols and SSL certificates. https://medium.com/jspoint/a-brief-overview-of-the-tcp-ip-model-ssl-tls- https-protocols-and-ssl-certificates-d5a6269fe29e