Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 6 –Issue 2,August 2023

Paper 47

MACHINE LEARNING TO DETECT MALICIOUS NETWORK ACTIVITY

1st Sadhvi Selvaraj
Dept. of Electronics and Communication
SRM Institute of Science and Technology
Chennai, India
ss7007@srmist.edu.in

2nd Spandana Shree K
Dept. of Electronics and Communication
SRM Institute of Science and Technology
Chennai, India
sk7269@srmist.edu.in

3rd Anoushka Singh
Dept. of Electronics and Communication
SRM Institute of Science and Technology
Chennai, India
as8159@srmist.edu.in

***Abstract*—Computer networks have developed to pose serious threats since they are constantly targeted by various attacks. New attacks and trends are being passed along; these attacks target all open ports that are accessible on the network. For this, a number of technologies are available, including network mapping and vulnerability scanning. In recent years, Machine learning (ML) is already a widely useda way to feed the capacity of intrusion detection systems, also known as IDS, to detect harmful network activity. How efficiently ML models identify anomalies depends on the accuracy of the dataset used to train the model. In order to help Intrusion Detection Systems (IDS) discover network traffic anomalies, this study suggests an ML-based detection mechanism. This detection technique makes use of a dataset that is made up of both harmful and eligible traffic. The main dificulties in this studyconsist of the extracted attribute needed to develop the machine learning model about various assaults todiferentiate between anomalous and regular data. A network traffic dataset is used in the ML model's training phase.**

Keywords—IDS, feature extraction, malicious,model,feature,attack,network traffic

I. INTRODUCTION

Over the past three years, there have been numerous documented cases of severe cybersecurity attacks on networks across various industries. Unfortunately, experts predict that the frequency and intensity of network hacks and data security threats will only continue to rise.In a network attack, the attacker must be aware of the network's topology, current addresses and services that are offered. Network scanners can detect open ports associated with shared services that are implemented, if they belong to ports with UDP protocol or ports with TCP protocol. A threat might potentially Send data packets to all ports if desired. Additionally, TCP fingerprinting capabilities of diferent vendors' systems can detect responses to unauthorized packet forms. Many TCP/IP stack responses are sent to illegal packets. Therefore, attackers can take advantage of network weaknesses by transmitting a wide range of unauthorized packet combinations, initiating an association RST packet with, or mixing some strange andillegitimatecode bits for TCP.The perpetrator is unaware of whether a computer utilizes either Windows, Linux, or another operating system. This knowledge aids in the attack's refinement and aids in the hunt for access points to vulnerable services and systems. IDS was recently trained to catch malicious network traffic using ML approaches. The fundamental idea behind IDS that utilizes machine learning (ML) methods is identifying sequences and developing IDS depending inside the dataset.IDS is efficient in finding. With the goal to recognize malicious traffic, we therefore int end to provide a detection framework that uses a model and relies on a collectioncomposed of network traffic characteristics that feed IDS. Sklearn, sys, Numpy, matplotlib, and Pandas are used to prepare, create, fit, and assess the presented model in Python. An alluring model must create and accommodate in memory, listeningthe features gathered between network traffic and anticipate irregularities

The TrainDataset.TXT and TestDataset.TXT were used as databases, which include 1,25,973 and 22,544 Network logs, respectively, were utilized in this work. The training and test sets include 42 characteristics categorized into regular traffic or specialized attack types, which are They are further split into four distinct groups: basic features, time-based traffic features, content features and host-based traffic features.

- *Basic Features:*The fundamental components of a TCP/IP connection are encapsulated in the Basic Features, including the kind of protocol, Service, time duration and among othersattributes.
- *Time-based traffic features*: Attributes of traffic dependent on time capture behavior patterns that have emerged over a period of two seconds.
- *Content features*: Content attributes, such as hot, no-of-root, and is-guest-login, use specializedunderstanding to examine the primary TCP packets content.

- *Host-based traffic features:*Attacks that continue longer than a few seconds and possesses the equivalent target location for the present relation are detected using

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 6 –Issue 2,August 2023

Paper 41

required traffic characteristics. Instances with these attributes include dsti-host-count, dsti-host-serv count, and others.

In short, feature set is comprehensive, and it provides various methods to analyze and interpret network traffic.

Thedataset's classes or labels are separated into five groups, four of which designate one as routine traffic and the targeted category:

- *Denial of Service attack:*DoS (denial-of-service) attacks are a sort of cyberattack that try to prevent devices such as computers from operating normally. The attacker achieves this by bombarding the targeted machine with an overwhelmingamount of requests, rendering it unable to handle regular traffic. This attack is usually executed by a single computer.
- *Probe attack*: Attacks of the sort of probing in which the intruder searches a network for potential vulnerabilities, such as open ports that can reveal information about services running on a resource. They may then gain privileged access to an unsuspecting host by exploiting a known vulnerability.
- *Root to Local attack:*Remote to Local attacks involve sending packets to a network with the aim of exploiting vulnerabilities and gaining unauthorized local access to network resources..
- *User to Root escalation:*User to Root attacks begin with obtaining access to a regular user account, followed by attempts to discover passwords and gain access to a computer service as a root user. These attacks are carried out with the intention of escalating user privileges and gaining unauthorized access to sensitive resources.

  In summary, these attacks are malicious activities aimed at exploiting vulnerabilities in computer systems and networks, with the aim of gaining unauthorized access and disrupting normal operations.

II. RELATED WORK

There has been a surge of interest in anomaly detection among researchers, leading to several studies and research attempt for this area. Giving a quick summary of the most important related works, which have been categorized based on the type of proposed solution.

The performance of ML models in detecting anomalies significantly depends on the standard of the training dataset. In this work, a detection architecture that employs an machine learningstructure to IDS stream to identify irregularities in network operations. The detection technique utilizes a dataset containing both harmful and normal traffic. Primaryobstacles in this study include identifying the retrievedattributes required for trainingmachine learning model in detecting multiple assaults and distinguishing between abnormal and normal traffic.

To enhance the effectiveness of the detection method and improve its efficiency, a data processing strategy may be employed in this study. This strategy involves dividing the flow information into separate sections, allowing the model to process information more quickly. Additionally, a fraudulent traffic detection model that utilizes a modular attention mechanism has been developed to handle this type of data.

To improve the model's detection capabilities, feature information from multiple hierarchies is combined in three main phases: pre-processing, assessment, and training. The Based IDS is designed to provide a binary classification, determining whether a traffic sample is normal or malicious. The model's inputs consist of 13 selected indicators, while the IDS output is either 0 or 1.

Overall, this study focuses on enhancing the performance and accuracy of the detection model, utilizing advanced techniques and strategies to achieve optimal results.

A. Supervised Learning

Supervised learning is a two-stage process that includes applying a model to training data, training the model on fresh samples, and using the model that was developed to make predictions. To begin, a group of items are displayed, each with their own set of characteristics labelled as X. These items are then associated with the correct response or identified as being labelled. When the training is underway, we must select a family of models, such as neural networks or decision trees, and determine the parameters of each model within that family.

Once the model is trained, it is applied to new objects, without any changes made to its type or parameters. This is the protection phase, which is crucial in the case of malware detection. Typically, sellers provide users with a pre-trained model, whose predictions are then used by the product to make decisions on its own. However, errors in prediction can have catastrophic effects on the user, such as uninstalling an OS driver.

As a seller, it is crucial to make an informed decision when selecting a model family. For the model to have a rapid pace detection and a low frequency of false positives, an imperative to implement successful training approach.

B. Unsupervised Learning

Unsupervised learning is among the strategies used, where data set is provided without any solutions to the task, and the goal is to find the structure or rule governing data generation.

Unsupervised learning is particularly useful in threat detection, especially for cybersecurity companies that have access to large unlabeled datasets, and manual labelling is a costly affair.

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 6 –Issue 2,August 2023
Paper 41

C. *Deep Learning*

To get the best out of basic data, computers need to work flawlessly. That's why cybersecurity companies are using deep learning techniques to detect malware. By creating complex feature hierarchies and merging diferent malware detection pipeline phases, deep learning models can be trained in a comprehensive way. This allows all components to be learned simultaneously, resulting in a powerful model that can identify malware with accuracy.

Cybersecurity firms have found deep learning methods to be useful in identifying malware from basic data. By incorporating intricate feature hierarchies and combining various malware detection pipeline phases into a powerful model, deep learning models can be trained comprehensively, enabling the simultaneous learning of all its components.

III. METHODOLGY

A. *Data Pre-Processing*

The dataset has to be cleaned in order to eliminate any duplicate entries. Since the dataset comprises both numerical and non-numerical data, it requires a pre-processing step. Although the classifierwith arithmetic inputs, using one-of-K as well as one-hot encoding, sci-kit-learning succeeds wellto make the necessary changes.

B. *Feature Scaling*

A critical component of machine learning techniques is feature scaling. It helps to avoid the problem ofattributes with high levels that could heavily impact the final results. To address this, for every attribute,calculate the median by dividing the result by their standard deviation, thensubtracting the median value from the attribute value. Once scaling, each attribute will have a standard deviation of 1 and an average of 0.

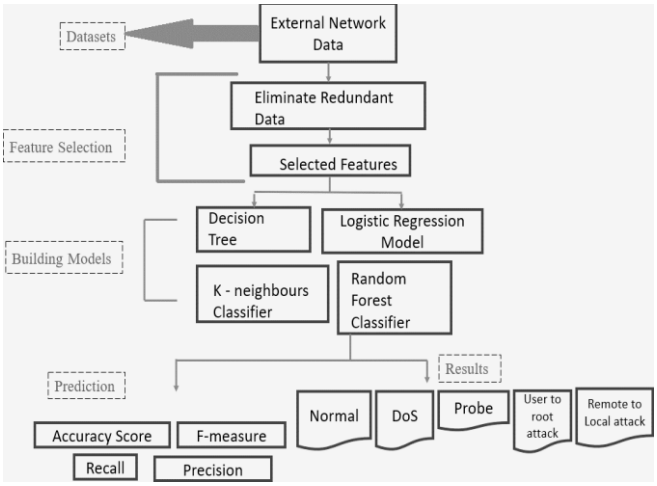


Fig. 1. Flowchart for predicting of various Malicious attacks

C. *Feature Extraction*

The process of feature selection is a smart way to eliminate any redundant or unnecessary data. It involves selecting only the most relevant features that represent the problem effectively while minimizing presentation loss.

The process commences with a univariate feature extraction that utilizes the analysis of variance (ANOVA) F-test to evaluate features. This methodanalyses each feature individually to evaluate its correlation with the labels.

When the most favorable subset of features was identified, Recursive Feature Elimination (RFE) was used as a strategy to incrementally create a model, feature was ignored, and the procedure was repeated with the other values until every attribute in the dataset had been used. As a result, it is an efficient enhancement for identifying the highest performing a collection of characteristics. The objective is to produce a feature ranking utilizing weights of a classifier.

D. *Split-validation evaluation*

This approachdivides the dataset into two categories: one to be tested and one to be trained on. Post memory tuning the machine learning model and training it using a diferentmethod, the model's correctness is calculated by constructing a confusion matrix, having 4 values. The True Positive (TP) represents amount of positive present and displayed positive observations, while the True Negative (TN) shows how many observations are positive and how many are expected negatives. The False Positive (FP) represents number of observations that are incorrectly anticipated to be positive but are in fact negative, and the False Negative (FN) reflects the amount of observations that are positive yet are projected negative.

IV. MODEL

A. *Decision Tree Classifier*

The classifier tree relies on the extracted features to function, with each internal node representing a specific feature. Using this model, four distinct attacks are analyzed, allowing for quick identification of the most significant features and prediction of their relation. The classification tree's response variable defines whether the input value falls under an attack category or not.

B. *Random Forest Classifier*

This supervised learning algorithm creates clusters of decision trees, and the end result is calculated based on the vast majority of trees chosen. Given the large dataset, this model will take a fair amount of time to compute the final answer. Compared to the decision tree classifier, random forest achieves higher accuracy.

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 6 –Issue 2,August 2023
Paper 41

V. EXPERIMENT

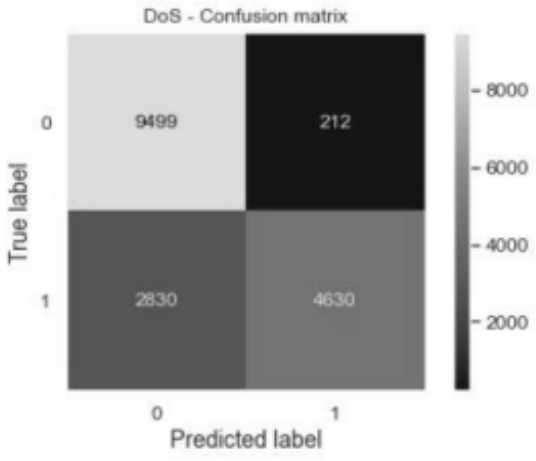
With the goal to get the most effective values to suit the algorithm, a thorough grid search parameters were adjusted, and information gain was achieved. was employed to choose features. Decision trees tend to overfit models at times. As a consequence, a tree was created using the training data, with the leaves representing class labels. A single character is utilized simultaneously to separate the node and the data when generating a decision tree.

Secondly, we employ random forest technique to create clusters of decision trees to determine the final result based on the majority of trees chosen. However, it's essential to note that this model may take a substantial amount of time to compute the final answer due to the vast dataset. Despite the time-consuming process, random forest surpasses the decision tree classifier by achieving a higher level of accuracy.

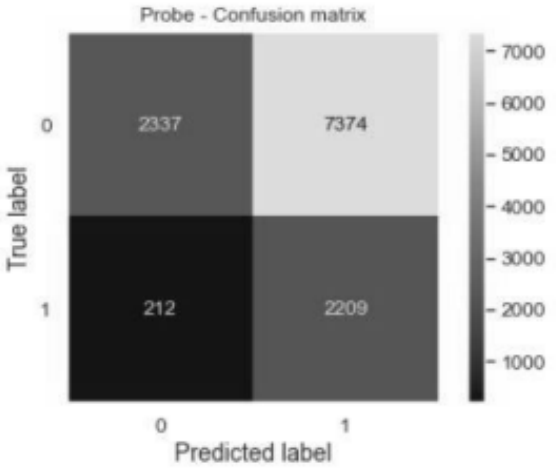
In order to achieve optimal results, the features are used individually. The univariate selection process is utilized to gather the necessary number of features. Once this is complete, the RecursiveFeature Elimination (RFE) is used to identify the selected characteristics, employing a set of attributes as input.Throughout the RFE procedure, the initial set of features is used to train the classifier and assign weights to every feature. The attributes with the least relative weights are then pruned based on the current feature set. It constitutes a pruning procedure thatis repeated recursively till the required feature count is obtained.

VI. DISCUSSION AND RESULTS

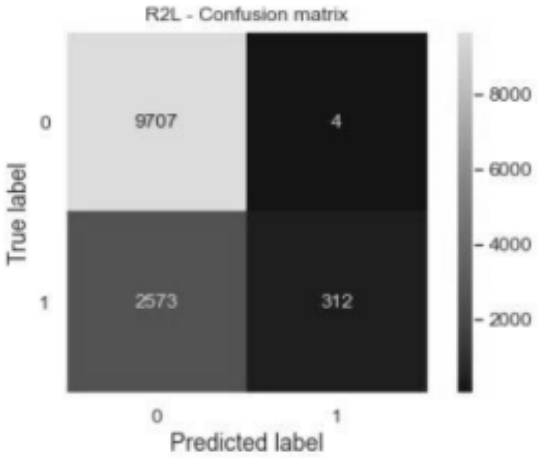
Feature selection represents methods for determininga part of important characteristics with the lowest graphical loss.Therefore, utilizing limited characteristics may generate better results. The confusion matrices for each attack were plotted.



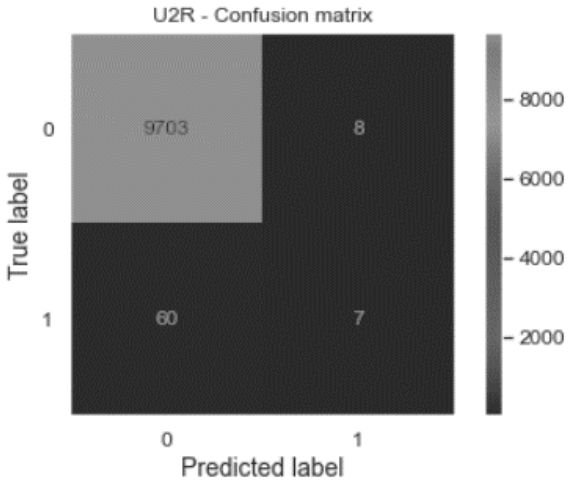
(a)



(b)



(c)



(d)

Fig. 2. Confusion Matrix for

(a)DoS (Denial of Service Attack) (b) Probe Attack (c) R2L (Root to Local Attack) (d) U2R(User to Root Attack)

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 6 –Issue 2,August 2023

Paper 41

In terms of measurement, accuracy and precision are two important concepts. Accuracy is the degree of proximity between the measured value and the actual measurement of the object. On the other hand, precision refers to the consistency of multiple measurements of the same object, irrespective of the actual measurement of the object.

$$Accuracy = \frac{(TN + TP)}{(T\,N + F\,P + F\,N + T\,P)}$$

$$Precision = \frac{TP}{(TP + F\,P)}$$

$$Recall = \frac{TP}{(TP + F\,N)}$$

$$F\text{-}measure = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

TABLE I. CYBER ATTACK PERFORMANCE MATRIX

Attack	*Accuracy*	*Precision*	*Recall*	*F-Measure*
DoS	99.7	99.6	99.7	99.6
Probe	99.6	99.4	99.3	99.3
R2L	99.8	97.0	99.3	99.3
U2R	97.9	97.1	97.0	97.0

Table 2 presents the important characteristics after a recursive feature removal was conducted on the dataset. Feature retrieval is based on rank.

TABLE II. REVELANT FEATURES

Target	Features Selected
Denial of Service attack	'log-in', 'count', 'ser-error-rate', 'serv-ser-error-rate', 'serv-same-rate', 'dsti-to-hst-counts', 'dsti-hst-serv-count', 'dsti-hst-same-serv-rate', 'dsti-hst-ser-error-rate', 'dsti-hst-serv-ser-error-rate', 'servi-http-count', 'S0flag', 'SFflag
Probe attack	'log-in', 'rej-error-rate', 'serv-rej-error-rate', 'dsti-hst-serv-count', 'dsti-hst-dif-serv-rate', 'dsti-hst-same-source-port-rate', 'dsti-hst-serv-dif-hst-rate', 'dsti-hst-rej-error-rate', 'dsti-hst-serv-rej-error-rate', 'Protocol-icmp', 'servi-eco', 'servi-priv', 'SFflag'
Root to Local attack	'source-bytes', 'dsti-bytes', 'hot', 'num-failed-logins', 'guest-login-count', 'dsti-hst-serv-count', 'dsti-same-hst-source-port-rate', 'dsti-hst-serv-dif-hst-rate', 'ftp-servi-count', 'ftp-servi-count', 'http-servi-count', 'imap4-servi', 'count-RSTOflag'
User to Local escalation	'urgent', 'hot', 'root-shell', 'num-file-creation', 'num-shells', 'serv-dif-hst-rate', 'dsti-hst-count', 'dsti-hst-serv-count', 'dsti-hst-same-source-port-rate', 'dsti-hst-servs-dif-hst-rate', ‘ftp-servi', http-servi-rate', telnet-servi-rate'

VII. CONCLUSION

For the purpose of modelling an IDS, this paper demonstrates the significance of utilizing a suitable classification learning algorithm in conjunction with a set of relevant features. Using a decision tree classifier to find important features, a method for selecting features that combine univariate feature selection with recursive feature elimination has been presented and proposed. This procedure repeatedly builds a model by putting the feature aside and then continuing with the other features until all of the features in the dataset are used up. Various classification metric measurements were used to assess the method's efficacy, and it was found that reducing the number of features increased the model's accuracy. The accuracy of the feature selection method proposed in this paper was high, and features were identified using a ranking and information gain technique

REFERENCES

[1] Andrey Ferriyan, Keiji Takeda, Jun Murai. “Generating Network Intrusion Detection Dataset Based on Real and Encrypted Synthetic Attack Traffic”, August 2021.

[2] Robin Sommer, Vern Paxson. “Outside the Closed World: On Using Machine Learning For Network Intrusion Detection.”

[3] Amirah Alshammari, AbdulazizAldribi.“Apply machine learning techniques to detect malicious network traffic in cloud computing”, July 2021.

[4] Xiaoyang Liu, Jiamiao Liu. “Malicious traffic detection combined deep neural network with hierarchical attention mechanism”. July 2021.