Contents

[1. Introduction: 2](#_Toc105246194)

[1.1 Dataset: 3](#_Toc105246195)

[2. Related work: 3](#_Toc105246196)

[2.1 Anomaly Detection: 3](#_Toc105246197)

[2.2 Anomalies detection using machine learning: 4](#_Toc105246198)

[3. Problem and Solution statement: 4](#_Toc105246199)

[a. Problem Statement: 4](#_Toc105246200)

[b. The solution to the problem: 5](#_Toc105246201)

[4. Purpose of research: 5](#_Toc105246202)

[5. Who will get benefits from this research? 6](#_Toc105246203)

[6. Methodology: 6](#_Toc105246204)

[6.1 Random Forest: 6](#_Toc105246205)

[6.2 Math behind Random Forest: 7](#_Toc105246206)

[References 9](#_Toc105246207)

# Introduction:

A network is a set of computers and users connected, their connectivity with each other is because they share any kind of resources over the network easily and time efficiently. We are connected to the world via the internet, 5 billion people are using the internet and they are connected. There are many types of networks, LAN, WAN, Military networks, etc. This is not always possible that every user of the network uses the network ethically, many people use the internet for various unethical purposes, like as stealing someone’s identity, bank details, etc. There are various ways and technologies used today to secure a network, among these technologies, there is a system called the Network Behavioral Analysis, by utilizing this system, we check the behavior of different activities and requests over the network.

A cyber attack is defined and explained in different ways. A cyber-attack is an attack or an assault that is launched by hackers, or criminals by using different types of technologies and fast computer systems against other people, the attack can be against any individual computer or hijacking the whole network, etc. the purpose of cyberattacks can be anything, the criminals may steal important data, installation of malicious software in the computers over the network, etc. One of the most used techniques used by hackers is to hack one computer over the network and make breach other computers easily, this way they can paralyze the whole network very easily because every computer over the network is connected. To avoid such fatalities, every computer connected over the network should be highly secured.

If you are connected to any network then there is a high chance that your computer can be attacked by some outsiders. Organizations invested a huge amount of money to secure their networks, they used different techniques and different systems which protect the network to some extent. These systems are used to detect threats and attacks over the network. This is often known as Threat Detection; threat detection is the process of analyzing the activities over the network and identifying any malicious activity over the network. If there is any threat detected, the process will immediately take over the threat and inform the authorities. For this purpose, many different systems are used such as NBA, etc.

Network Behavior Analysis often called NBA is a program that is used to monitor networks. The purpose of NBA is to check the activities over the network and it helps us secure the network by looking at the traffic over networks and detecting unusual activities. NBA helps us to detect malicious activities, packet checking, and signature recognition [1].

The network behavior analysis is used to detect and respond to the anomalies in real-time, Anomaly detection means detecting abnormality, it’s the pattern that does not match the other patterns. It is often called outliers detection which means detecting events and activities over the network which are unexpected. Network anomaly detection is a device or a program that is used to monitor networks to detect any abnormal activity over the network [2].

Cyber-attacks can occur in many ways, hackers and intruders can easily take control over your network, and your data and credentials can be stolen easily if you are not secured over a network. To prevent any abnormal activity over a network, organizations use different types of security systems, these systems include NBA (Network Behavior Analysis). NBA is a program or a physical device that is installed over a network to monitor activities and users, this system detects any abnormal activity immediately and blocks the activity, and informs the authorities. By using this system, we can secure our networks and prevent cyber-attacks [3].

Cybersecurity helps us to protect the networks we are connected to; it helps us prevent any unethical use of the network and all the hardware and software connected over the network. Cybersecurity is also helpful to protect the data over the network from theft and any other damage. The data can be of any type just like individual personal details, intellectual property, etc. without cybersecurity, we can’t prevent and defend our networks and organizations from cyber-attacks [3].

The network behavior analysis helps us to strengthen our network security. This system monitors the networks constantly and immediately notify when it detects any abnormal activity over the network. using this system, we can take immediate actions against any attack that occurs over the network. These systems have strengthened cybersecurity [4].

There are many benefits of using the NBA over other tools. The best thing about NBA is that it helps us detect the root cause of the problem, issue, or any abnormal activity over the network, NBA helps us detect the problem very quickly and it is capable of immediate responses to the anomaly it detects, so it is very time efficient and responsive to the issues over the network [5].

## 1.1 Dataset:

The dataset used in this project is collected over a private military network, the dataset contains information on cyber-attacks based on various features given in the dataset. It established an environment to get crude TCP/IP dump information for an organization by mimicking a regular US Air Force LAN. The LAN was engaged like a genuine climate and impacted by different assaults. There is a variety of different TCP packets sequenced through a network, these TCP packets are the flow of the data between two different IP addresses during a specific time duration under some well-defined protocols. Each connection and activity over the network are labeled as normal or an attack over the network. the total size of each connection over the network is almost 100 bytes. The dataset contains 41 different features, these features are collected and obtained from normal and abnormal activities over the network. There is also a class feature that is categorized into two different categories, [Normal & Abnormal].

The link to the dataset: <https://www.kaggle.com/datasets/sampadab17/network-intrusion-detection>

# Related work:

## 2.1 Anomaly Detection:

The meaning of anomaly detection is to detect any abnormal activity over the network. the activities over the network make different types of patterns, when the pattern of a request is different from the rest of the pattern, the system detects that pattern as an anomaly because the behavior of that request is different from the rest of the request roaming over the network. in other words, the anomalies are often referred to as outlier detections, because the outlier is a value that does not meet the rest of the values. Anomaly detection is the process that is used to detect these types of different activities over the network, it monitors the network and detects abnormal requests [2].

Anomaly is a completely different type of observation in the data from the rest of the observation. Anomalies detection in various domains has been applied to different dynamic research, Researchers proposed a very basic structural overview of the anomaly’s detection techniques. They have proposed different methods which include statistical methods and machine learning methods. The statistical methods and machine learning methods have been compared [6].

Network security is nowadays a big issue and researchers are working very hard to propose the best way to secure the networks, for this purpose, many researchers have worked on proposing different methods. Researchers have proposed a robust method that helps detect anomalies over the networks. They are detecting the anomalies over the network based on the flow monitoring, they have proposed four different metrics which capture the flow of requests and activities over the network [7].

## 2.2 Anomalies detection using machine learning:

Machine learning is very helpful in detecting anomalies over the networks, the classical methods like intrusion detection and other systems have secured the networks to some extent but they are not highly secured. Researchers from all over the world are now focusing on using different machine learning techniques to overcome this problem. Researchers have tried to implement a Support Vector Machine (SVM) to detect anomalies over the network [8].

The networks we use are at high risk of different types of attacks like Dos/DDos attacks, etc. to avoid these attacks, companies are using different network security systems like IDS, BNA, etc. all these network security systems are based on using machine learning in the backend. Researchers have used the same technique as other researchers used that is the Support Vector Machine, but the difference is that researchers used two machine learning techniques, the first one used to be the Genetic Algorithm (GA) for features selection, and the Support vector machine (SVM) for the classification of the activities over the network [8] [9].

Deep learning on the other hand is also a very helpful way of detecting anomalies over the network, deep learning is always used for image classifications but it can also help us detect the anomalies over the network easily and accurately. Different deep learning neural networks like RNN, DNN, and CNN can be utilized to detect abnormal activities over the network. Researchers have used the Recurrent Neural Networks (RNN) to detect anomalies over the network [10].

# Problem and Solution statement:

## Problem Statement:

We use networks for different purposes because they are beneficial and time-efficient. From small organizations to big ones, we use the network to connect the employees and staff. Similarly, we use the internet on daily basis for different purposes. These networks can be sometimes hijacked and controlled by hackers and intruders, when these types of attacks happen, the network users may find it very devastating because their personal information can be stolen anytime. To avoid such fatalities, different systems and protection systems have been introduced which help us secure the network from any foreign attacks.

The main objective of this project is to detect abnormal activities over the network, I have a dataset that is collected over a military network. The dataset has many features like protocol type, flag, src bytes, DST bytes, etc. using these features, I will train different machine learning models to detect any abnormal activity over the network.

## The solution to the problem:

The proposed solution to the problem is machine learning. Cyber-attacks can occur in many ways, hackers and intruders can easily take control over your network, and your data and credentials can be stolen easily if you are not secured over a network. To prevent any abnormal activity over a network, organizations use different types of security systems, these systems include NBA (Network Behavior Analysis). NBA is a program or a physical device that is installed over a network to monitor activities and users, this system detects any abnormal activity immediately and blocks the activity, and informs the authorities. All these proposed systems are based on machine learning, these systems are highly trained that they can easily detect any abnormal activity over the network.

Our approach is also by using machine learning to detect anomalies over the network. The activity over any network is based on different factors like the protocol of the request, DST, Flags, etc. We will train the machine learning model on the data collected over a military network; the machine learning model will be intelligent enough to detect any anomaly over the network based on the factors of the activity.

Machine learning is a way that makes a machine capable of taking decisions on its own. With the help of machine learning, we can train models and machines for different purposes. Machine learning is very useful in the field of cybersecurity, with the help of ML, the cybersecurity systems like NBA can analyze the different patterns of activities over the network and learn from those activities which then prevent any similar attacks it detects over the network. Machine learning helps cybersecurity experts to be more proactive and prevent any unusual activity and respond to it in real-time.

For the solution to the problem, we have used the Random Forest model. We will train this model on the data we have and will test it using the testing data to see the performance level of the model. Why did we choose Random Forest? The answer to this question is that Random Forest is the most useful and accurate model which is used for different regression and classification problems. In our project, the problem is the classification, we have a dataset having 41 features and a class feature, this class feature has two sub-classes which are normal and abnormal which means that activity over the network will be normal or abnormal based on all the 41 features. For this classification, the Random Forest is the best choice.

Another reason is that our dataset has more than 25,000 observations, this is a large dataset and Random Forest can handle large datasets very easily, the predictions of random forest are very accurate and easy to understand. The random forest algorithm is based on decision trees so it produces a high level of accuracy in predicting the outcomes.

Random Forest is also known as Random Decision Trees because it is based on different decision trees. This model is an ensemble model which means that the output of this model is the class selected by the majority of the decision trees. This way, this model performs very well. We will use the training data to train Random Forest and evaluate the model using the testing data.

# Purpose of research:

The purpose of this research work is to propose a way to detect normal and abnormal activities over the network using machine learning. There are many different techniques used to detect abnormal request over any network, they all use machine learning in the backend. For our research work, we proposed a machine learning algorithm know as Random Forest Classifier. The reason why we chosen this model is because this model performs very well on Regression and Classification both. This model is very easy and fast to train on the training data, by the use of this algorithm, we can minimize the problem of data overfitting.

We have a data collected over a private network, the dataset has different features and a class categorized as normal or abnormal, for this classification problem, we used the random forest and our proposed model performed very well.

# Who will get benefits from this research?

This research work is based on anomalies detection using machine learning. With the help of machine learning, we can easily detect normal and abnormal requests over the network. From this research work, the users of any network can get huge benefits by detecting any abnormal activity over the network they are connected with. The abnormal activity over the network can be sometime very dangerous and if there is no security of the network, hackers and intruders can easily get into the network and can steal important information. We have proposed a machine learning way to detect any insecure and irreverent request over the network, we have trained random forest algorithm over the data of many different requests labeled as normal or abnormal. This research work can be highly beneficial for organizations, companies, etc to secure their networks and prevent any attacks from the intruders.

# Methodology:

The overall methodology is divided into stages, in the first stage, we are doing the advanced data preprocessing, in the data preprocessing, the first thing we are checking is to see whether our dataset has any missing values? The real-world dataset can have missing values but fortunately, we didn’t notice any missing values in the dataset. During the data preprocessing, we have performed the feature selection, we have selected some very useful features from the dataset which identify the activity over the network as normal or abnormal. After the data preprocessing, we split the data into testing and training sets, the training set will be used to train our classification model and the testing data will be used to evaluate the model that we have trained. The goal of this project is to train machine learning Random Forest classifier and see the accuracy of the model, the higher the accuracy

* Random Forest

## 6.1 Random Forest:

We have already explained the basics of the Random Forest classifier, this model consists of many different decision trees and each tree have nodes. This algorithm works as a voting classifier which means that it will look at each tree’s decision. The final prediction by this model will be the result given by the majority of decision trees. The working diagram of Random Forest is given below:

Training set

Test set

Voting

(Averaging)

Predictions

Figure Working Diagram of Random Forest

The Random Forest algorithm is an unsupervised machine learning algorithm which is used for both regression and classification, this model works on labeled data. There are many advantages of using the Random Forest, the main advantage of this model is that it works on both regression and classification, this model is very fast, means it can be easily trained and tested. This model prevents the problem of data overfitting.

## 6.2 Math behind Random Forest:

The Random Forest algorithm can be used for Regression and as well as for Classification. the goal of this project is to detect anomalies over the network by using this model, our problem is the classification means the model will decide whether the activity over the network is normal or abnormal.

To use this model for the purpose of classification, we use the ‘Gini Index’.

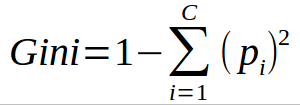


Figure Gini Index [11]

Class and probability are used by this formula which us used to determine the *gini* of each branch of the node. It means that which branch and node is the most likely to occur. The relevant frequency of the class is represented by the *pi* in the formula and the number of classes are represented by *c.*

The Random Forest can also use *entropy.* the formula of entropy is given below:

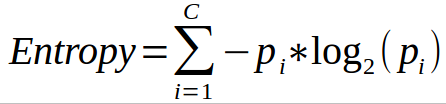


Figure Entropy of Random Forest [11]

Entropy is used to calculate the probability of the outcomes of the model to see how the nodes are branched. Dissimilar to the Gini file, it is more numerical escalated because of the logarithmic capacity utilized in computing it.

We have trained the above model on the training data, after the training process, the next stage was to evaluate the model so we perform the testing using the testing data. After training and testing, we got the following results. The code and the results are shown below:

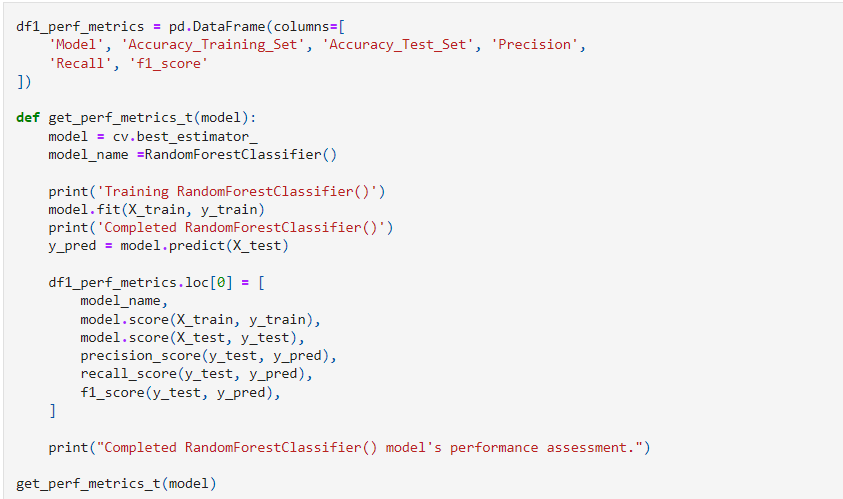


Figure Training Random Forest (Code)

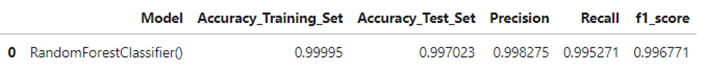


Figure 5 Random Forest accuracy

After training and testing the model performance, the next goal was to perform the predictions on the trained model, we got the following results.

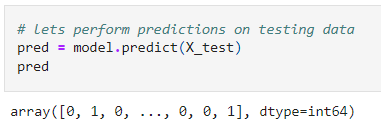


Figure Random Forest Predictions

The predictions are stored in an array, we can see that the model predicted the activities over the network in the testing data as 0 and 1, [0 means the activity is ‘NORMAL’, and 1 means the activity is ‘ABNORMAL’].

# References

|  |  |
| --- | --- |
| [1] | J. K. D. C. S. M. H. W. V. Z. René Veenstra, "Network–Behavior Dynamics," *Research on Adolescence,* 2013. |
| [2] | M. H. Bhuyan, D. K. Bhattacharyya and J. K. Kalita, "Network Anomaly Detection: Methods, Systems and Tools," *IEEE,* p. 303, 2014. |
| [3] | H. Sedjelmaci, S. M. Senouci and M.-A. Messous, "How to Detect Cyber-Attacks in Unmanned Aerial Vehicles Network?," *IEEE,* 2016. |
| [4] | R. M. L. P. D. C. KELLY BISSELL, "INNOVATE FOR CYBER RESILIENCE," in *INNOVATE FOR CYBER RESILIENCE*, 2022, p. 48. |
| [5] | L. M. L. Yang HongMei, "The Advantages and disadvantages of network Organization Behavior Teaching," *IETI Transactions on Social Sciences and Humanities,* p. 37, 2022. |
| [6] | G. Muruti, F. A. Rahim and Z.-A. b. Ibrahim, "A Survey on Anomalies Detection Techniques and Measurement Methods," *IEEE,* 2018. |
| [7] | H. A. Nguyen, T. V. Nguyen, D. I. Kim and D. Choi, "Network traffic anomalies detection and identification with flow monitoring," *IEEE,* 2008. |
| [8] | J. TaeshikShona, "A hybrid machine learning approach to network anomaly detection," *Information Sciences,* 2007. |
| [9] | T. Shon, Y. Kim, C. Lee and J. Moon, "A machine learning framework for network anomaly detection using SVM and GA," *IEEE,* 2020. |
| [10] | H. K. J. K. S. C. S. I. K. &. K. J. K. Donghwoon Kwon, "A survey of deep learning-based network anomaly detection," *Springer,* 2017. |
| [11] | M. Schott, "Random Forest Algorithm for Machine Learning," Capital One Tech, 25 April 2019. [Online]. Available: https://medium.com/capital-one-tech/random-forest-algorithm-for-machine-learning-c4b2c8cc9feb. |
| [12] | A. T. Tunggal, "Why is Cybersecurity Important?," upguard.com, 12 May 2022. [Online]. Available: https://www.upguard.com/blog/cybersecurity-important. |