# Developing a framework for network security behavior integrated with the organization data management system to predict the threats

# تطوير إطار مفاهيمي لسلوك امن الشبكات متوافق مع نظام إدارة البيانات الخاص بالمؤسسة للتنبؤ بالتهديدات

**by**

**HANI ABDELHADI ABDULLAH ISMAIL**

**Dissertation submitted in fulfilment**

**of the requirements for the degree of**

**MSc INFORMATION TECHNOLOGY MANAGEMENT**

**at**

**The British University in Dubai**

**March 2022**

# DECLARATION

I warrant that the content of this research is the direct result of my own work and that any use made in it of published or unpublished copyright material falls within the limits permitted by international copyright conventions.

I understand that a copy of my research will be deposited in the University Library for permanent retention.

I hereby agree that the material mentioned above for which I am the author and copyright holder may be copied and distributed by The British University in Dubai for the purposes of research, private study, or education and that The British University in Dubai may recover from purchasers the costs incurred in such copying and distribution, where appropriate.

I understand that The British University in Dubai may make a digital copy available in the institutional repository.

I understand that I may apply to the University to retain the right to withhold or to restrict access to my thesis for a period which shall not normally exceed four calendar years from the congregation at which the degree is conferred, the length of the period to be specified in the application, to ether with the precise reasons for making that application.

Signature of the student

# COPYRIGHT AND INFORMATION TO USERS

## Abstract

Network security management becomes an essential task in all organizations to protect their information and communication. It became more critical, especially after the COVID 19 Pandemic, as most businesses and industries have moved to use more online technologies. This study aims to develop a framework for analyzing network security behavior integrated with the current data management system and to predict the threats for administrator remedial actions by using Machine-Learning techniques The primary objective of the study is to automatically provide an optimum set of rules that are summarized and generalized across various security devices for professionals to configure the best security solution with minimum configuration efforts. This is experimental analysis research method depends on collecting information from network security data flow based on selected events that matched with the actual organization's security rules and policies with a dataset of 123029 records collected from log files of the standard security system. Moreover, a framework is designed based on the network security events, including the threats prediction, which can be used to take proper actions by using the artificial intelligence method. The result of the studied framework showed that KNN and random forest models performed better with the precision of 91.84% and 91.48%, respectively, compared to the other models of SVM, decision tree, and Naïve Bayes. The future work of the study is to enhance the prediction of unknown threats and apply the model in the real world to establish a security baseline for similar organizations.

باتت إدارة أمن الشبكات مهمة أساسية في جميع المؤسسات على مستوى العالم لضمان حماية المعلومات والاتصالات الخاصة بها. وقد أصبح الأمر أكثر أهمية من ذي قبل، خاصة بعد جائحة COVID 19 حيث تحولت معظم المؤسسات والشركات

تحولا ذكياً وتم استخدام المزيد من التقنيات عبر الانترنت. تقوم هذه الدراسة على تطوير إطار عمل يقوم على تحليل سلوك تأمين الشبكة المتكامل مع نظام إدارة البيانات المستخدم للتنبؤ بالتهديدات المحتملة لاتخاذ الإجراءات التصحيحية ويتم ذلك عن طريق استخدام تقنيات التعلم الآلي. وتهدف الدراسة بشكل أساسي إلى توفير مجموعة مثالية من القواعد تلقائياً يتم تلخيصها وتعميمها على أنظمة الأمان التي يستخدمها مهندسي الشبكات المتخصصين للوصول إلى أفضل حلول تأمين الشبكات بأقل مجهود ممكن. تعتمد هذه الطريقة البحثية للتحليل التجريبي على جمع المعلومات من تدفق بيانات تأمين الشبكة استنادًا على الأحداث المحددة سابقاً والتي تتوافق مع قواعد وسياسات تأمين الشبكات الخاصة بالمؤسسة التي تجرى فيها الدراسة والمستخدمة فعلياً مع مجموعة بيانات مستخلصة من 123029 سجلًا تم جمعها من ملفات السجل الخاصة بنظام الأمان الأساسي. علاوة على ذلك، تم تصميم إطار عمل بناءً على أحداث تأمين الشبكة، بما في ذلك التنبؤ بالتهديدات، والتي يمكن استخدامها لاتخاذ الإجراءات المناسبة باستخدام تقنية الذكاء الاصطناعي. وأظهرت نتيجة الإطار المدروس أن نماذج KNN والنماذج العشوائية للغابات كان أداؤها أفضل و أكثر دقة بنسبة 91.84٪ و 91.48٪، مقارنة بالنماذج الأخرى من SVM وشجرة القرار و .Naïve Bayes ويتمثل العمل المستقبلي للدراسة في تعزيز التنبؤ بالتهديدات غير المعروفة وتطبيق نفس النموذج في مجال العمل على أرض الواقع لإنشاء خط أساس أمني للشبكات يمكن أن يستخدم من قبل المؤسسات ذات النشاط المماثل.

# Acknowledgment

I would like to firstly thank Allah the Almighty for providing me an opportunity to further my studies and work on the research project. By the Grace and Blessings of Allah, I am able to satisfactorily complete my research thesis.

I would like to express my sincere thankfulness and gratitude for Prof. Abdallah Sherief who has provided all the support and guidance that was required for the research project. I am also grateful for all my teachers especially Prof. Khaled Shalan that were part of my learning process.

I would like to extend my gratitude towards my friend Muhammad Atif who have supported and motivated me throughout my journey.

Lastly I would also like to especially thank my dearest family for showing all the support I needed for this valuable achievement.

# Contents

# List of Figures

# List of Tables

# Chapter I: Introduction.

This chapter introduces the thesis subject, the practical experience, the problem statement, the study aims and objectives, the intangible benefits, the nature of challenges, and the thesis structure.

## 1.1 Introduction

Technology is implemented everywhere in our daily lives, increasing the demand for security at different levels. Data over the network is one of the trends nowadays in various sectors like healthcare, education, enforcement, and many others. According to many studies, more than 70% of people around the globe will have access to the network by 2023 (Cisco Annual Internet Report (2018–2023). (2020)). These indicators will raise the crucial need to research network security and provide a safe environment for these users and the shared data. Most of the current applied works related to security are still behind the available advanced technologies and rely on rule-based algorithms designed by the expertise of the field. These rigid rules make the systems less flexible for future requests and cross-implemented into different scenarios (Chen, Bo et al.,2020).

On the other hand, some systems used signature-based techniques to solve network security problems. These systems contain a database of attack signatures. Human experts create the attack signatures by manually analyzing the attacked data. Generating signatures based on the problems is inefficient and requests a lot of manual work to be completed (Sultana, Nasrin, et al.,2019). Network security system with the ability to learn from the current attacks, adapt to a different scenario, and propose new methods to avoid the forthcoming threats is one of the major requests(Elsayed, M.S., Le-Khac, N.A., Dev, S. and Jurcut, A.D., 2020). More specifically,

intrusion detection over the network is one of the problems that can fall into machine learning tasks. It requires traffic data analysis and making decisions based on the data and the user behavior over the network. To reach that goal, this work will investigate the opportunities to use different machine learning algorithms to predict the threats for administrator remedial actions to keep the network highly responsive, efficient, and secure.

## 1.2 Practical experience

The organizations have multiple security devices and systems to prevent breaches at different levels. Each system requires specialized configuration rules and specialized engineers to perform the set up. There are generalized and specialized rules for each organization. Certain security rules should be classified as the baseline or standard rules that are must to be configured. The default rules for most of security systems is (Any to Any = Denay), or (Any to Any Allow) and the security engineers will start the configuration and build the rules.


*Figure 1: Example of Denial Rule*

An example of generalized rules is for NAT (Network Address translation) that allows the local network users to reach to the public network as shown in figure 2 . An example of an specialized rule will be to prevent the guest users to have access to the local network. There is a need for learning the required rules that can form an efficient standard and baseline set of rules for securing the network.

*Figure 2: Example of NAT policy*

The various types of security policy rules configuration that are implemented manually are shown as below

- **Gateway-Antivirus Policy**

All known viruses will be blocked by configuring the gateway-antivirus signature policy rule that will inspect all the traffic related to HTTP, FTP, IMAP, SMTP and POP3 protocols.

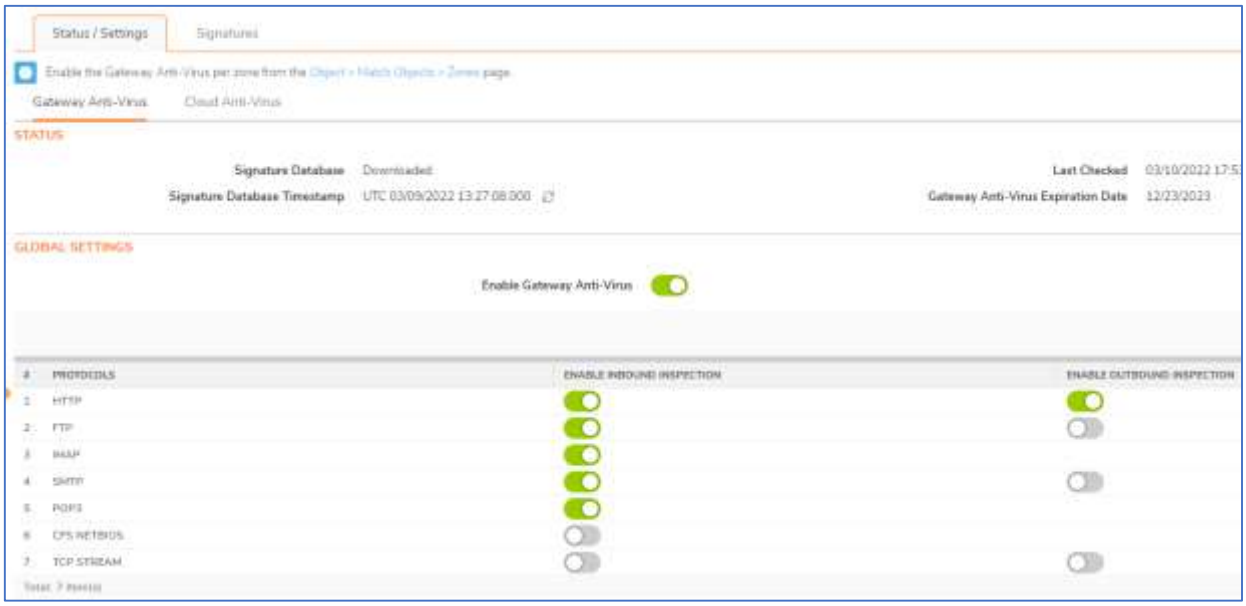



*Figure 3: Example of Gateway-Antivirus Policy rule*

- **Anti-Spyware Policy**

All known Spyware will be blocked Appling the rule of Anti-spyware signature.

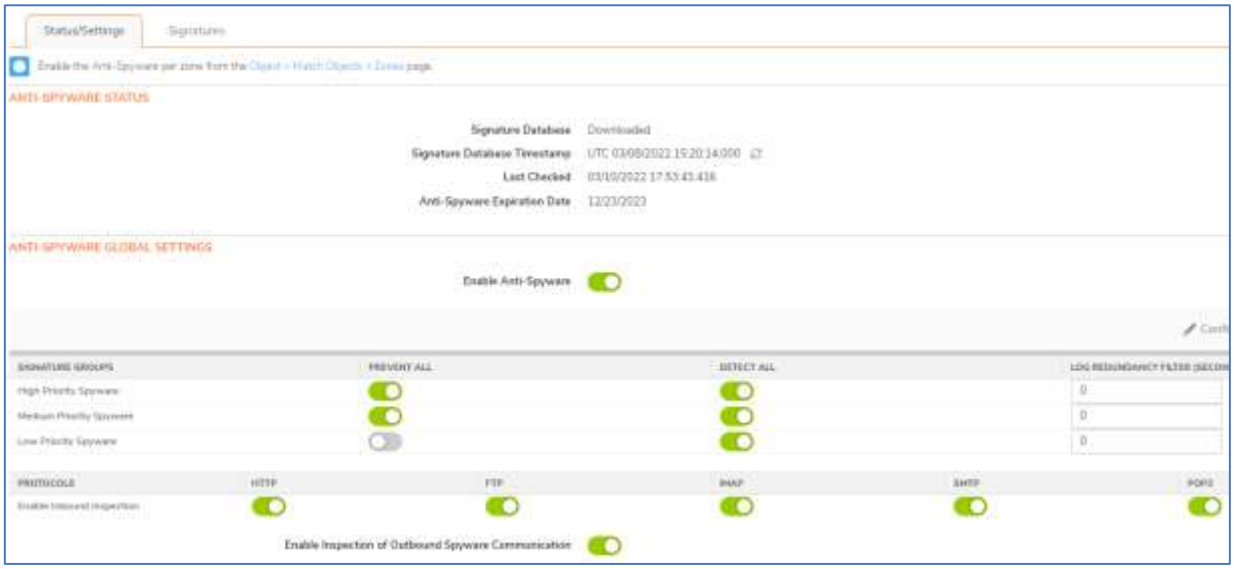



*Figure 4: Example of Anti-Spyware Policy*

- **Intrusion prevention policy**

The security system device will periodically update the IPS signature if any new type of attack vector found In the organization it will be blocked based on matching the signature with the updated IPS signature .

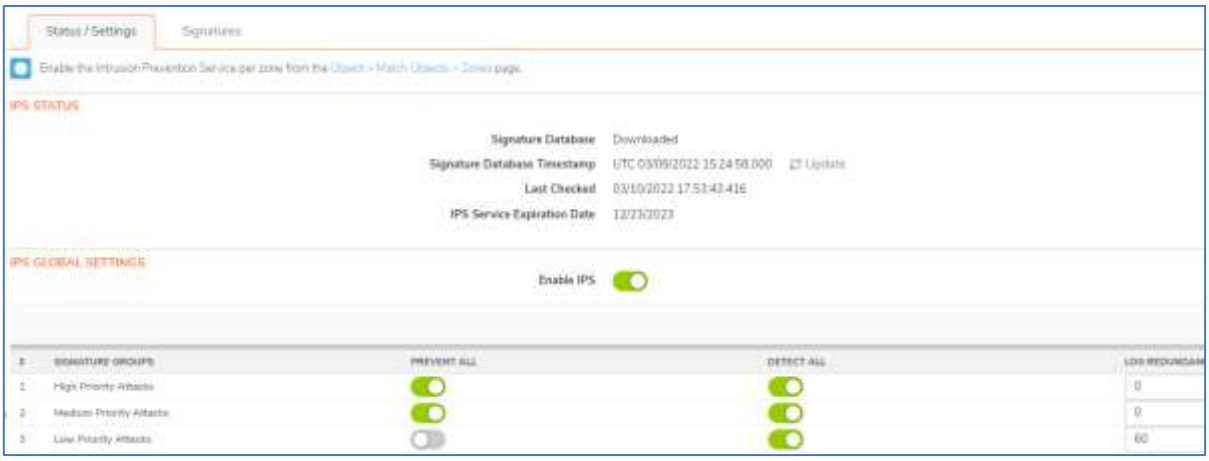



*Figure 5:Example of Intrusion prevention policy*

- **Botnet policy**

Botnet policy will detect the anomaly of the traffic and detect the traffic is belongs to original user traffic or Botnet triggered traffic.

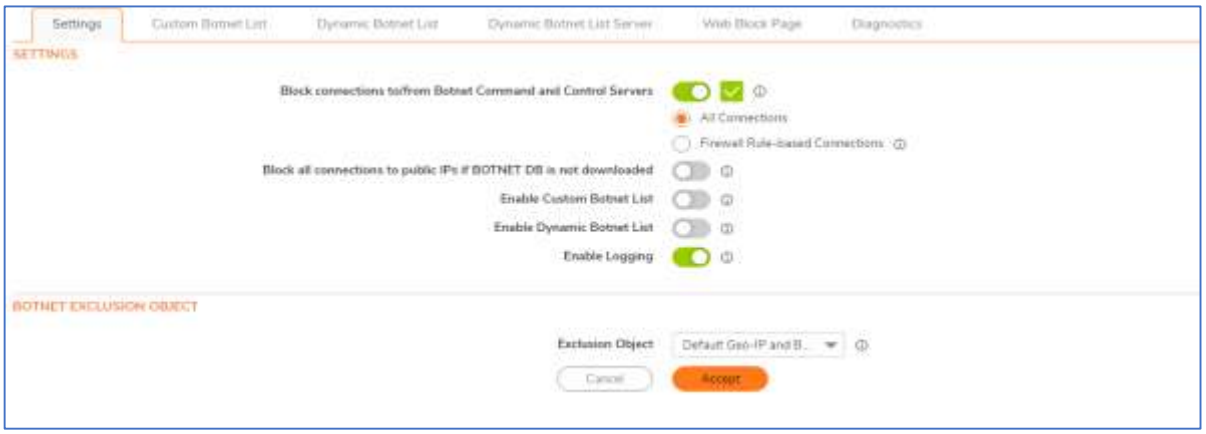



*Figure 6: Example of Botnet policy*

- **Geo-IP Filter**

In the organization the security rules can block the traffic from certain countries such as China and North Korea.

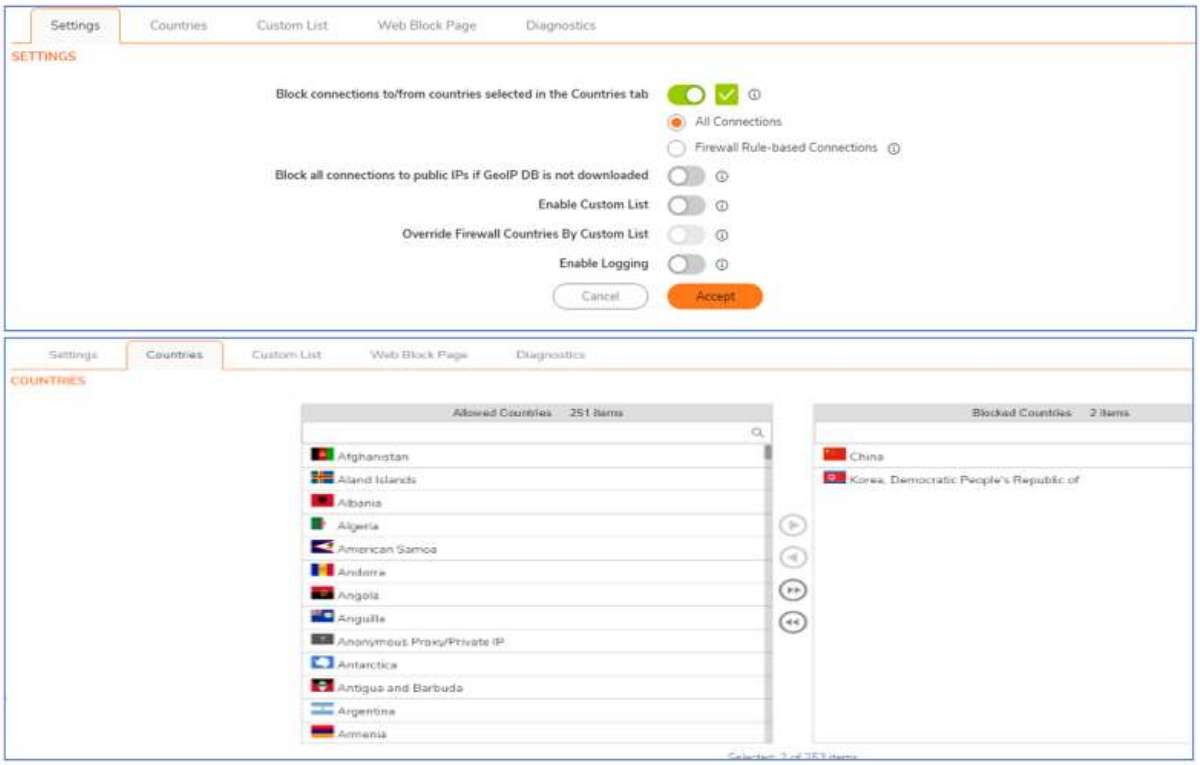



*Figure 7: Example of Geo-IP Filter*

- **SonicWall CFS Policy**

Content-Filter-Policy to block unwanted and malicious traffic content to organization's network.

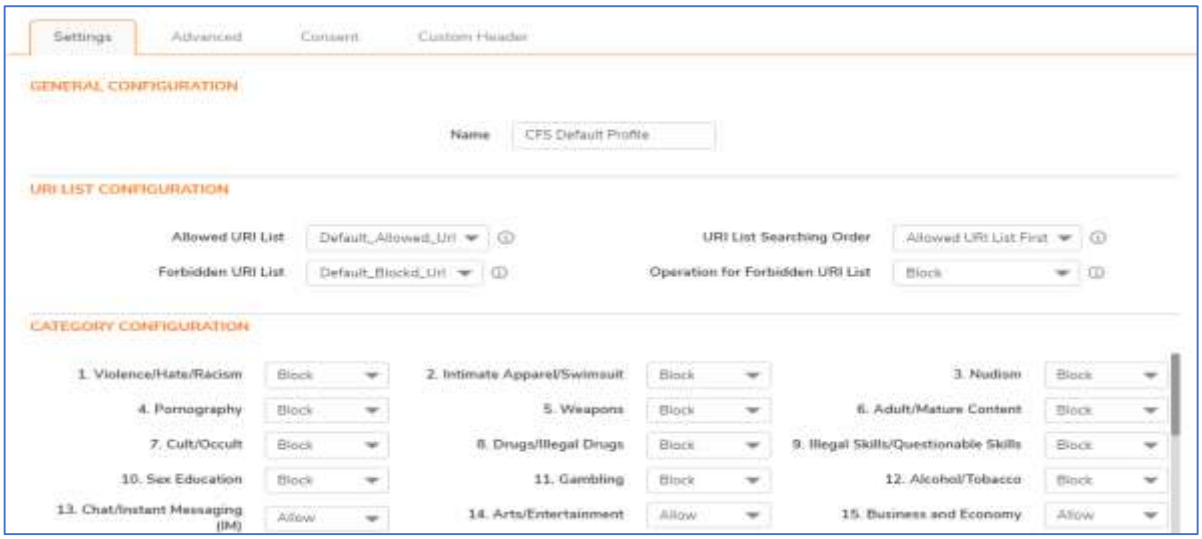



*Figure 8: Example of CFS Policy*

- **APP Control Policy**

All P2P and Proxy application was blocked by configuring the firewall app control policy, also all High/Severe alert application categorized by SonicWall was also blocked.

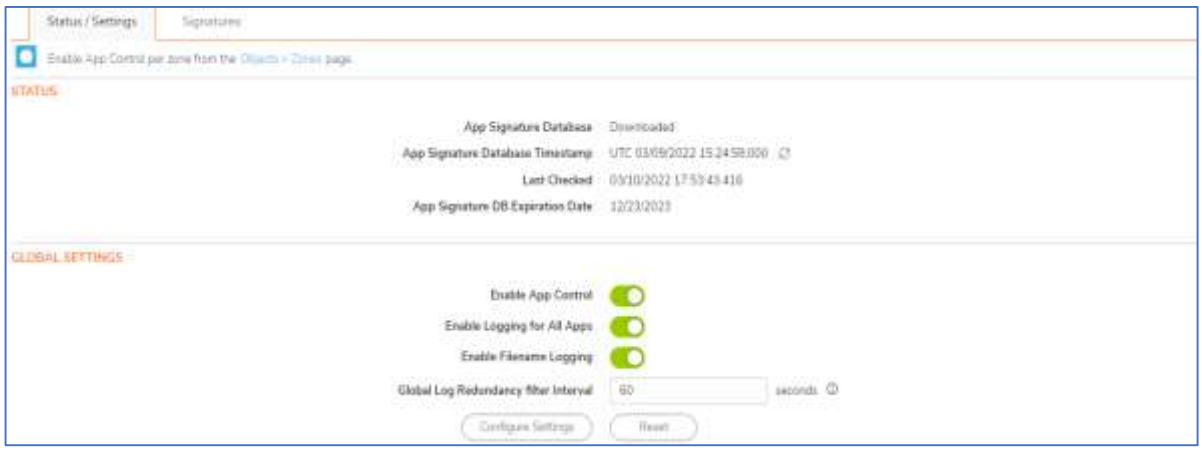



*Figure 9: Example of APP Control Policy*

The above examples are a subset of a larger security configuration set up of any organization that requires to prevent them from cyber-attacks. It requires huge efforts for security engineers to implement an efficient set of rules without compromising the security breaches on several kind of security devices within an organization. The study aims to help the security engineers to reduce the efforts and increase the efficiency and security of the network.

All the network TCP traffic needs to follow the rules of a "three-way handshake" to form a connection. The 3 Way Handshake that always require three steps:

- Step No.1: Establish A connection between the client and the server: Firstly, a connection between the server and the client is established, so a port must be opened in the target server to allow and introduce the new connections. Then the client starts to send an SYN (Synchronize Sequence Number) packet to the server over IP network, even in the same network or to the external network.

- Step No.2: the SYN packet to be received by the server: When the client sends the SYN packet and is received by the server, a confirmation receipt will be created by the server, which is called ACK (Acknowledgement Sequence Number) or can be called SYN/ACK.
- Step No.3: the server will respond and send the SYN/ACK packet to the client: In this step, it is the role of the client to acknowledge the packet sent from the server. Then after this process to complete the Connection will be created between the client and the server. So, verifying the serial numbers between both sides is necessary, and missing any segment's order, the Connection will not be established.

In a three-way handshake, if the first packet has any changes or invalid sync of its header, it is considered a Network attack by the firewall.

The devices that can connect over the internet should have an (IP) Internet Protocol address that gives the device a unique identity. The data is traveling over the internet in IP Packets containing IP header. This header carries and shares routing information such as the destination and the source IP addresses. Replacing the packet's header with a fake one is called IP spoofing. There are different types of IP spoofing as follows:

- Masking botnet devices. The IP spoofing Process can be used to access computers via masking honeypots, a collection of attached computers that execute mundane tasks to keep the websites running and up. These botnets are unseen by IP spoofing attacks that use their interconnectedness for malicious reasons. This includes providing spam and different forms of malware and flooding the targeted websites, networks, and servers with data, and causing crashing.

- DDoS attacks. IP spoofing is often utilized to release a distributed denial-of-service (DDoS) attack. DDoS attacks are brute-force attempts to bring a server down. Attackers might flood their victims with data packets by using faked IP addresses. This allows attackers to hide their identity while slowing or crashing a website or computer network with a deluge of internet traffic.
- Man In The Middle attacks. The attacker acts as a "man in the middle," capturing sensitive communications and utilizing them to commit crimes like identity theft and other forms of fraud Attacks by a man-in-the-middle, which works by stopping sessions between two computers, frequently use IP spoofing. In this case, IP spoofing changes packets before delivering them to the target computer, and neither the sender nor the recipient is aware of the changes.

Various machine learning algorithms can be practically applied for network monitoring and data analysis. Some of these techniques are based on statistical analysis like nearest neighbors, Decision Tree, which is like an if-else set of rules, or Random Forest as a set of proposed decision trees. Algorithms based on the human brain simulation like neural network can be proposed. In this method, the features should be manufactured and extracted as a pre-step of NN. NN will be trained over these features and predict the following unseen data coming from the network. Usually, the NN consists of three layers which are the “input layer,” “hidden layer,” and the “output layer.” More advanced neural networks with multiple layers can be proved for this work; this type of network is called a deep neural network.

## 1.3 Problem statement

Based on the previous shared practical experience related to networking, many problems related to the network architecture, network connections, and data transition over the network should

be taken into consideration. To build a reliable security network massive effort is required by IT security engineers in configuring security rules for internal and external communications. These IT engineers are expensive resources in terms of cost and time.

All the network TCP traffic needs to follow the rules of a "three-way handshake" to form a connection. In a three-way handshake, the first packet in the network connection needs to have a sync header to start the Connection. If the first packet didn't have any sync header, it is considered a Network attack by the firewall. Another way to attack the network is by using IP spoofing. It is a process that enables the attackers to replace the header's packet of the source IP address with a fake header or a spoofed address. This method will lead to intercepting and modifying the IP packet before sending, which will let the IP look like from a trusted source.

A perfection in decision-making cannot be achieved by using machine learning techniques; different problems should be evaluated while implementing the proposed machine learning technique (Tulabandhula, Theja, and Cynthia Rudin.,2014). One of them is the reliability and the efficiency in detecting the threats that the network is facing. Also, the proposed algorithms can generate several false-positive alarms (triggering an intrusion in the network while it is normal behavior). An additional factor that impacts the efficiency of a detection process is the number of false-negative threats that failed to be detected. The greater the number of false alarms, the slower detection activity is and is more expensive (Belavagi, Manjula C., and Balachandra Muniyal.,2016). Response time required to detect the abnormal events occurring through the network is another real challenge in real-time environments. The different possibilities of the violation over the network should also be a challenge in this work. The policies can be violated internally by the employees themselves or externally by the attackers.

## 1.4 The Thesis Aim and Objectives

The main aim of the proposed work is to provide a framework that analyses the network behaviour based on studying the currently used security systems in the organization. The main objective of this study is to provide an ideal set of rules that are minimized and generalized that can be used in similar organizations effectively with less effort needed from security engineers.

As a field of AI, Machine learning will be used to learn from the provided network traffic data to achieve a secure network with less human intervention. The objectives of the proposed work can be listed as follows:

1. To investigate the current implementations of network intrusion detection/attacks.
2. to investigate the various intrusion detection techniques
3. To propose the rules that define attacks and intrusion over the network
4. To analyze the existing method used in terms of performance and accuracy
5. To see if the proposed procedure is appropriate.

With the implementation of this proposed research, some benefits can be gained as follows:

| Function | Justification |
|---|---|
| Decrease in the number of network intrusions | Networks are monitored, and admins will be informed if there might be possible attacks over the network. |
| Amount of manpower needed to monitor the network behaviour | Network administrators do not need to monitor the performance continuously in person. |
| Actions can be taken in before threats happen | Because of the system's real-time nature, authorities can intercept the network before any attack or threat arises. |

| | |
|---|---|
| The system will record all incidences. | With the available logs of all the network traffics, trace back, and better-trained algorithms can be done |

*Table 1: Tangible benefits*

## 1.5 Intangible benefits

- Better and safer networks
- Increased compliance
- Trusted interaction

## 1.6 Nature of the Challenges

As my background is related to networking with no strong experience in machine learning algorithms and programming, the main challenge that I face is the lack of experience in understanding and developing ML to serve my research objectives. Analyzing the data to find the best features to be proposed for analyzing the different traffics and detecting as many types of attacks as possible is considered another highly challenging duty. Determining the best ML algorithms to be implemented for this purpose and the coding language/tool that led to reaching this aim is not an easily achievable task.

## 1.7 Thesis Structure

The following are the five chapters that make up this dissertation:

- Chapter1: present an overview of the research subject, the practical experience, the problem statement, the study aims and objectives, the intangible benefits, the nature of challenges, and the thesis structure.
- Chapter 2: present an overview of the experimental literature review that starts with analysis for prior research articles that were connected to this study in terms of the

algorithms employed, the paper's goal, the datasets utilized, the evaluation measures, the attributes, the publication year, and the conferences or journals. Also, discuss some of the study's issues, including artificial intelligence (AI), machine learning (ML), and the relationship between AI and network security.

- Chapter 3: Present an overview of the research methodology model that has been followed through this study, starting with the Business Understanding Phase that contains the domain requirement, the network structure, the access situation, and determining the data mining goal as a sub-phase and followed by the data understanding phase with a data collection as a sub-phase. The data preparation phase includes data integration, feature selection, future transformation, scaling, and data cleaning process. Finally, this chapter also introduces the modeling, evaluation, and deployment phases.
- Chapter 4: present an overview of the data and result, starting with data analysis, attribute relationships, machine learning performance by using Naïve Bayes, SVM, KNN, Logistic Regression, Decision Tree, and random forest. Finally, evaluate the models by using Accuracy, Precision, Recall, F1-Score, ROC AUC score, and AUC.
- Chapter 5: represent an overview of the conclusion and future work that contains the conclusion, contribution, practical implication, limitation, and future work.

# Chapter II: Literature Review

This Chapter provides an introduction about the importance of network security, analysis for previous research papers that related to this study in terms of the used algorithms, the objective of each paper, the used datasets, the evaluation metrics, the attributes, the published year, and the conferences or journals. Then the thesis explained some topics that will be used in the study, such as Artificial Intelligence (AI), Machine Learning (ML), and the relation between AI and network security.

## 2.1 Introduction

Securing the networks and keep free from attacks are a challenging and complicated task that becomes harder over time. Different methods, technologies, and threats are proposed innovatively and applied daily in this regard. Attacks worldwide are increasing, and the methods used to infiltrate the target machines are very wide. The responsible groups for these attacks are amateur persons and professional attackers, organizations, and even countries by all their advanced resources. All the previously mentioned points stress the need to have advanced research and find up-to-date solutions in this field to provide a reliable solution and secure environment. Most of the available solutions are rule-based systems that are able to handle only limited scenarios. With the ability of the machine nowadays to learn, analyze, and predict actions based on the available data, one of the advantages that might take into consideration is the wide availability of the data describing a different kind of problems related to networking. These data might formulate a solid ground and justification to explore the possibility of applying artificial intelligence and machine learning techniques to sort out these critical issues.

## 2.2 Research papers analyses

the literature review stage plays an important role in any research. It gives a wider understanding to any researcher, which helps him/her better understand the subject of the research. As part of the literature review, I went through more than 40 research papers on using Artificial Intelligence and Machine Learning to detect network attacks. I preferred to focus on newly published papers that were recently published, and I made sure that they were published in well-known journals of high citation as per the following details:

27% of the papers published by IEEE, 23 % of the papers by Springer, 14% by Elsevier, 9% by "International Conference on Advances in Computing, Communications, and Informatics" (IJCSIS), 5% for each of the following publishers arXiv , The Computer Journal, "Asia Conference on Computer and Communications Security," EURASIP Journal, "Information Security, Journal of Computer Networks" and "Communications and Information Security Journal."

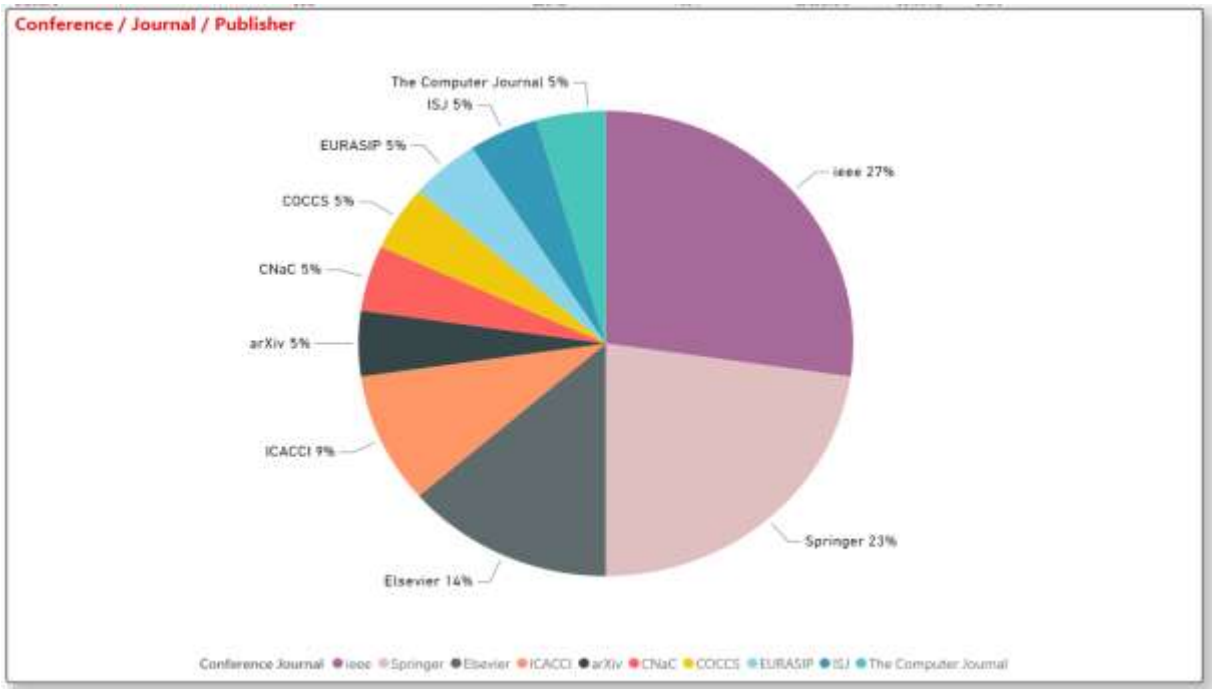



*Figure 10:Conference, Journal and Publishers comparison*

These papers which are used in the literature review can be categorized according to the year of their publishing. The majority of the papers (32%) were published in 2019, 27% were published in 2018, 23% were published in 2016 and the rest 19% were published between 2020 and 2021.

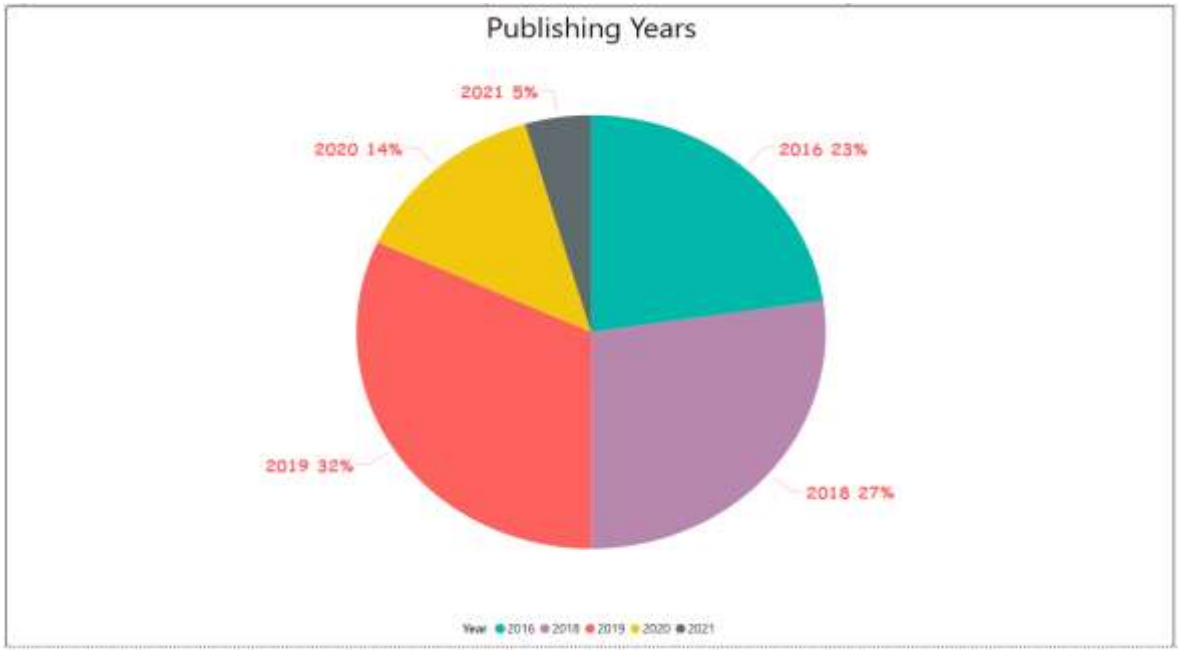



*Figure 11: Publish years comparison*

There are various types of evaluation techniques that have been used by the researchers of the papers mentioned above that I used for the literature reviews. After analysis, it was noticed that the Accuracy classification model measurement that was used to identify the relation between the variables in the dataset was used in 58% of the papers. The Precision that measures the quality and Recall that used to measure the quantity was used in 34% of the papers.

The rest of the papers used different types of evaluation techniques such as F-measure, AUC (Area Under The Curve), ROC (Receiver Operating Characteristics) curve, True Negative Rate (TNR), and False Positive Rate (FPR).

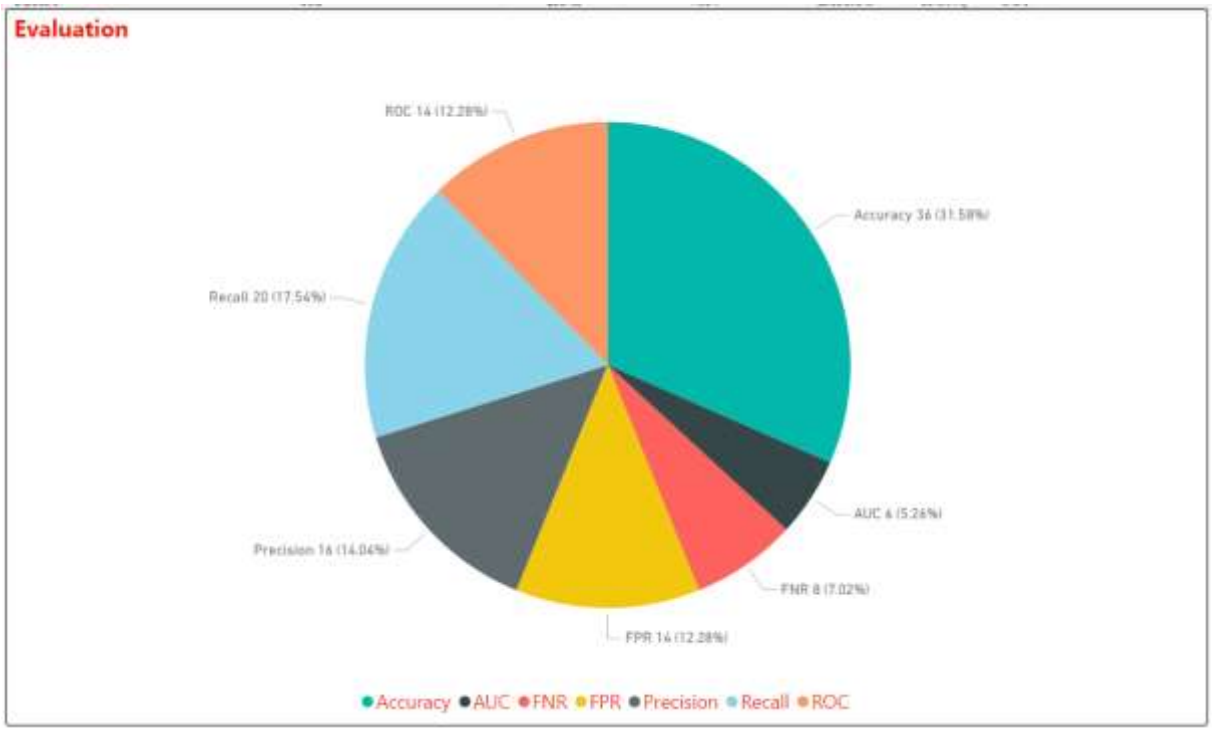



*Figure 12: Evaluation metrics comparison*

With thorough review and comparison conducted among the mentioned study papers, it was found that different Machine Learning algorithms have been used to get the most accurate evaluation results. It was noticed that most of them used SVM (Support Vector Machine), which has been used by 25% of the papers. Then the Artificial Neural Network (ANN)and Random Forest were used equally by 14% for each in the papers. But the other algorithms such as Naïve Bayes, Decision Tree, Convolutional Neural Network (CNN), K Nearest Neighbors (KNN) are used by less percentage.

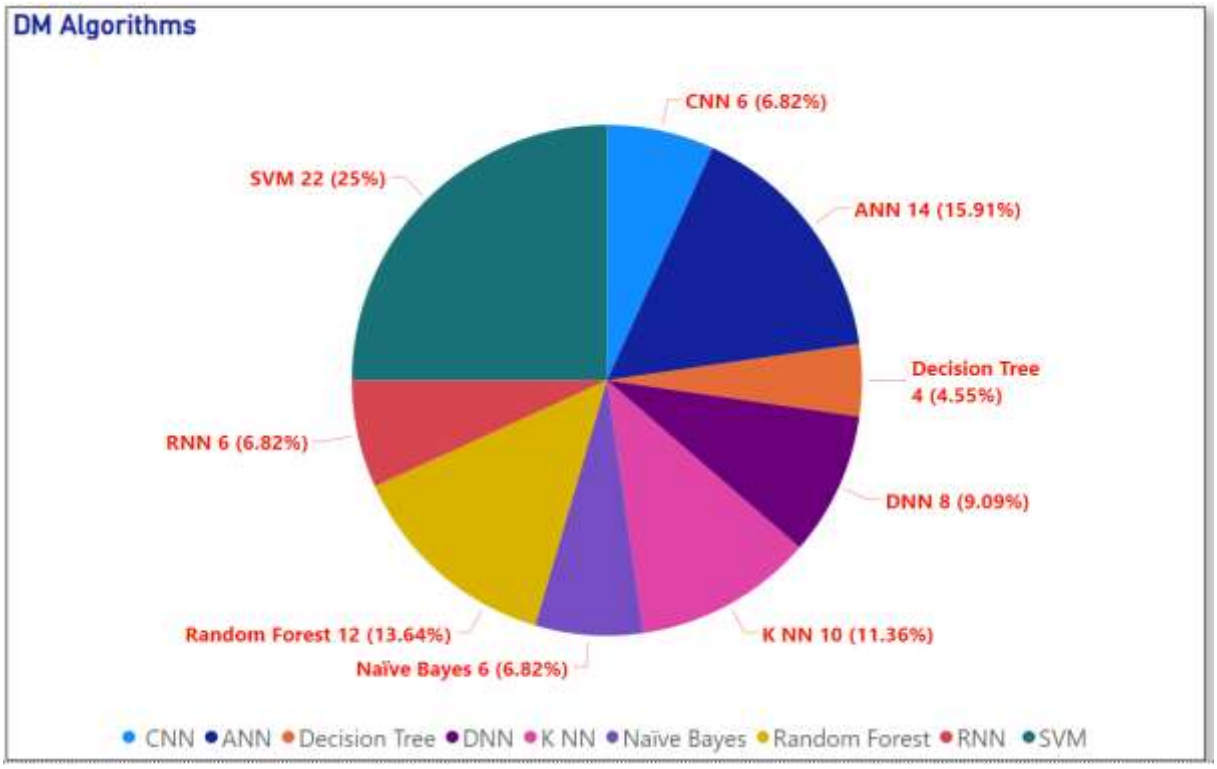



*Figure 13: Machine learning algorithm comparison*

Most of the papers used the same Dataset (NSL-KDD), so most of the papers used the same attributes in the research; these attributes are U2R (User to Root), R2L (Root to Local,) and Probe. But the DDoS attack attribute was hugely used in the papers. The rest, such as Web Attacks, Spam, Worms, Botnet, Brute Force, Services and Network (TCP, ICMP), are used but with less percentage.

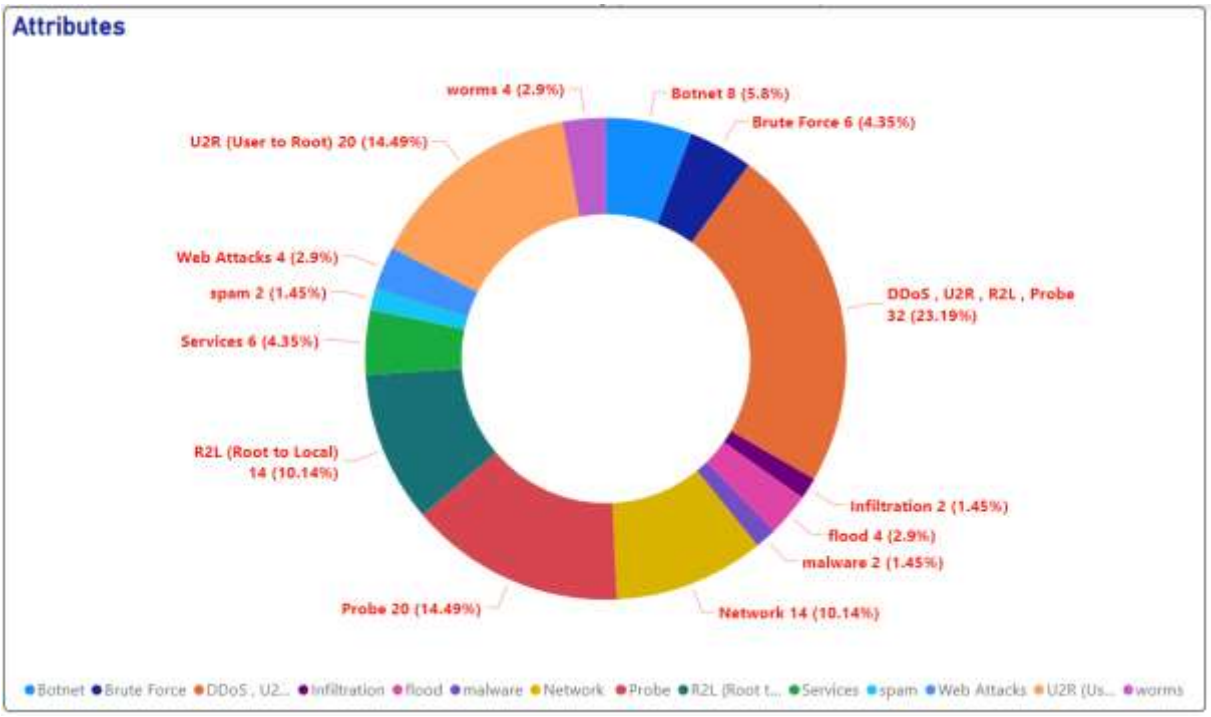



*Figure 14: Attributes comparison*

| **Name** | **Objective** | **Dataset** | **Evaluation** | ***Algorithms*** | **Attributes** | **Year** | **Conference Journal** |
|---|---|---|---|---|---|---|---|
| (Ieracitano, C., Adeel, A., Gogate, M., Dashtipour, K., A. and Hussain, A., 2018) | To propose IDS that extract optimized and correlated features using bigdata visualization and deep autoencoder to detect the potential threats. | NSL-KDD | Accuracy, Precision, Recall, F-Score | RNN, DNN | Network, Probe, U2R, R2L | 2018 | Springer |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| (Kunang, Y.N., Stiawan, D. and Suprapto, B.Y., 2021) | Propose to provide an alternative to deep learning building models by using automated hyperparameter process optimization that helps to verify the value of hyper-parameters and the optimal categorical hyperparameter configuration. | NSL-KDD, CSE-CIC-ID2018 | Precision, Recall, F-Score | DNN | Network, Brute Force, Botnet, Web Attacks | 2021 | Elsevier |
| (Singla, A., Bertino, E. and Verma, D., 2020) | to propose and analyse the use of aggressive domain adaptation to overcome the problem of a dataset lacking identified training data by knowledge transfer from an available network intrusion detection (NID) dataset. | NSL-KDD and UNSW-NB15 | Accuracy, F-Score | CNN, DNN | Network, Probe, U2R, flood, worms, Services | 2020 | COCCS |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| (Apruzzese, G., Guido, A. Colajanni, M., Ferretti, L., and Marchetti, M., 2018) | To analyse the existing maturity of these systems and identify the major obstacles preventing machine learning cyber detection schemes from being adopted immediately. | Drebin dataset | Accuracy, Precision, Recall, F-Score | Naïve Bayless, Random Forest, DNN | Network, malware, Brute Force, Botnet | 2018 | IEEE |
| (Alsughayyir, B., Qamar, A.M. and Khan, R., 2019) | to develop a framework to utilize deep learning in a more effective and better way for detecting the threats by focusing on distinguishing typical network behaviour from abnormal behaviour. | NSL-KDD Dataset | Accuracy, AUC, F-Score, FNR, FPR | SVM, Random Forest | Probe, U2R, Services | 2019 | Springer |
| (Sultana, Nasrin, et al.,2019) | to cover and evaluate the tools used in developing NIDS models in the SDN environment. | NSL-KDD | Accuracy, | SVM | Probe, U2R | 2019 | Springer |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| (Tulabandhula, Theja, and Cynthia Rudin,2014) | Propose a new framework called (MLOC) that considers how a classifier will be used for a successive job, connecting machine learning with a decision-making process. | Rudin et al | Recall, AUC, ROC | RNN | | 2014 | Springer |
| (Belavagi, Manjula C., and Balachandra Muniya ,2016) | To prove that the Random Forest Classifier operates better than the other classifier while detecting the traffic is an attack or regular. | NSL-KDD | Accuracy, Precision, Recall, F-Score, FNR, FPR | Naïve Bayes, SVM, Random Forest | Probe, U2R, R2L | 2016 | Elsevier |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| (Wang, Ping, et al.,2016) | This research presents a behaviour-based SVM (support vector machine) with machine learning for usage in security monitoring systems (SMS) to classify network threats for network intrusion detection systems. | NSL-KDD 1999 | Accuracy | SVM, Decision Tree | flood | 2016 | IEEE |
| (Lin, W.H., Lin, Wu, B.H. H.C., Wang, P., and Tsai, J.Y., 2018) | This research focused on detecting network intrusions using the LeNet-5-based convolutional neural networks (CNNs) Algorithm to categorize network risks. | KDD'Cup99 | Accuracy, Recall | CNN | spam, Services | 2018 | IEEE |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| (Khan, F.A., Derhab, A. Gumaei, A., and Hussain, A., 2019) | For effective network intrusion detection, the researcher proposes an innovative two-stage deep learning (TSDL) approach that relies on a stacked auto-encoder with a smooth classifier. | KDD99 + UNSW-NB15 | Accuracy, Precision, Recall, F-Score, AUC, FPR | RNN | Network, Probe, U2R, R2L | 2019 | IEEE |
| (Ayo, F.E., A.A., Adekunle, Folorunso, S.O.Abayomi-Alli, , A.O. and Awotunde, J.B., 2020) | to propose NIDS established on a deep learning pattern that is improved with rule-based mixed feature selection | UNSW-NB15 | Accuracy, Precision, Recall, F-Score, FNR, FPR | KNN, SVM, CNN, ANN | worms, Brute Force, Botnet | 2020 | ISJ |
| (Lin, P., Ye, K. and Xu, C.Z., 2019) | design and construct a deep learning-based dynamic network incident detection system | CSE-CICIDS2018 | Accuracy, Recall, ROC, F-Score | SVM, Decision Tree, Random Forest, ANN | A botnet, Web Attacks | 2019 | Springer |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| (Asad, M., Asim, Beg, M.O.M., Mujtaba, H. Javed, T., , and Abbas, S., 2020) | Offer a unique deep neural network-based detection technique for reliably detecting numerous application-layer DDoS attacks using feed-forward back-propagation. | CIC IDS 2017 | F-Score | SVM, ANN | spam, Services | 2020 | The Computer Journal |
| (Meng, F., Lou, F., Fu, Y. and Tian, Z., 2018) | introduce a new threat intelligence detection approach based on recurrent neural networks with long short-term memory for attribute categorization (LSTM-RNNs). | CERT insider threat dataset v6.2 | Accuracy, Precision, Recall, FPR | KNN, SVM, ANN | Network | 2018 | IEEE |
| (Fang, X., Xu, M., and Zhao, P., 2019) | Produce a deep learning model (framework) by using a bi-directional pre-trained model with short-long term memory, called BRNN-LSTM | ARIMA, ARMA+GARCH , Hybrid | Accuracy | ANN | Web Attacks, Services | 2019 | EURASIP |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| (Iqbal, M.F., Habib, D. Zahid, M.and John, L.K., 2019) | Investigates several predictors in search of one that has high accuracy, low processing complexity, and low power consumption. | CAIDA Traces | Accuracy | ANN | Web Attacks, Spam | 2019 | CNaC |
| (Reddy, R.R., and Sunitha, K.N., 2016) | provide an automatic analysis with performance improvement to enhance the identification of data mining approaches. | KDDCUP99, UNB ISCX , HTTP-CSIC | Accuracy, Recall, ROC, F-Score, FNR, FPR | SVM | Network | 2016 | ICACCI |
| (Alkasassbeh, Mouhammad, and Mohammad Almseidin, 2018) | This paper demonstrates how the KDD dataset (or Knowledge Discovery in Databases) may be used to test and evaluate various Machine Learning techniques. | KDD | Accuracy, ROC, FNR, FPR | Naïve Bayes ANN | Probe, U2R, R2L | 2018 | arXiv |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| (Rai, K., Devi, and Guleria, A., 2016) | The researcher developed the Decision tree algorithm based on C5.5 to avoid the issue related to split values and the feature selection process. | NSL-KDD | Accuracy | SVM, Decision Tree, ANN | Probe, U2R, R2L | 2016 | ICACCI |
| (Farnaaz, N. and Jabbar, M.A., 2016) | Use a random forest classifier to create a model for intrusion detection and compare the result of the effective classifiers and the traditional classifiers. | NSL-KDD | Accuracy | Random Forest | Probe, U2R, R2L | 2016 | Elsevier |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| (Negandhi, P., Trivedi, Y. and Mangrulkar, R., 2019) | used the well-known data mining approach of feature selection to improve the classification accuracy of the study model even more. Intelligent feature selection based on Gini relevance was used to limit the number of features. | NSL-KDD | Accuracy | Random Forest | Probe, U2R, R2L | 2019 | Springer |

*Table 2: Research Papers table*

## 2.3 Artificial intelligence (AI)

As a simple definition, AI is the ability to solve a problem based on what input it gets, to which it will produce a certain output. In another way, a type of software program(s) based on a statistical model that a large amount of data has trained, both input and output. Intelligent systems are caused by years of natural evolution. In the beginning, the system of predefined rules that is not self-learning was considered as AI, but it is not. The AI system should be able to decide by itself based on the learning cycle and the different inputs injected into the system. This field of study is currently applied in most fields and the current daily technologies in our hands. It is used for car navigation, voice commands, searching, reasoning, speech recognition, path finding, prediction, evaluation, and many others (M. Haenlein, A. Kaplan. (2019). AI is not that modern field and has been used since 250 BC by Ctesibius from Alexandria in Clepsydra (water clock) without knowing that it is AI. Clepsydra (water clock) regulated the water flow to a continuous value by pressure. This system is considered one of the first feedback and complex gearing systems that used the feedback to enhance accuracy. Also, the Mathematical theory of self-regulating systems that was established in the 19th is another example of AI usage. A clearer example of using AI can be found in the 20th century. Warren McCulloch and Walter Pitts proposed logical calculus of the concepts immanent in nervous activity in1943. The two scientists combined three concepts in their conducted research which are:

- The function of neurons in the brain
- Formal analysis of logic
- Programming theory of Turing

This work inspired many other works to investigate more in the field of AI and to come up with very innovative ideas. An example of these works is the ELIZA program. ELIZA was proposed in 1965 and can be considered as the first natural processing work that provided the machine with the ability to understand human language (Shah, Huma, et al.,2016).

Weak and strong AI are the most common types of artificial intelligence. Strong AI (commonly assumed definition of AI) imitates and replicates human intelligence with the ability to learn by itself and is self-conscious. On the other hand, weak AI can only imitate intelligence and is mainly driven by designed rules. The major fields of AI are:

- Robotics
    - Realize agents which can interact with real environments
    - It also involves sensors,
- Reasoning and information retrieval
    - Inferring answers based on given rules and facts
    - Finding search items or related "suggested" items
- Machine Learning
    - When presented with data, systems improve their performance
    - Deep learning: more complex (deeper) models
- Data science
    - Involves machine learning and intelligent algorithms on data
    - Statistics and visualization
    - But also, storage of data and representation (databases)
- Computer Science

- More general than just AI involves data science, but also programming of other algorithms and user interface design

## 2.4 Machine Learning

The major part of AI that will impact the conducted research is called machine learning ML. ML is the process of creating a machine model that is able to learn based on provided data. These data have been collected from similar scenarios to the problem that aims to be solved. The main advantage of ML is that it doesn't need that much human intervention. Other researchers also define ML to extract a pattern from the given data; this pattern will be executed later to find the unseen data and to decide and predict (El Naqa, Issam, and Martin J. Murphy,2015). ML field relies on different fields as a theory of statistics, linear algebra, probabilities, and others. ML can be applied to labeled data for classification and prediction purposes; this type of ML is called supervised learning. Also, ML might work on data without any label for clustering purposes; this kind of ML is called unsupervised.

**Supervised Learning:**

- Data: (X, Y)

X is the data

Y is the label (Ground Truth – made by a human)

- Goal: Learn function Y = f(X)
- Examples

Classification, Regression Object detection Image captioning Translation

**Unsupervised Learning:**

- Data: X

Just data, no labels!

- Goal: Learn some underlying hidden **structure** of the data

- Examples

  Clustering, Dimension Reduction, Feature Learning, Anomaly Detection, Learning generative models

In general, the different areas of ML can be summarized in the figure below.

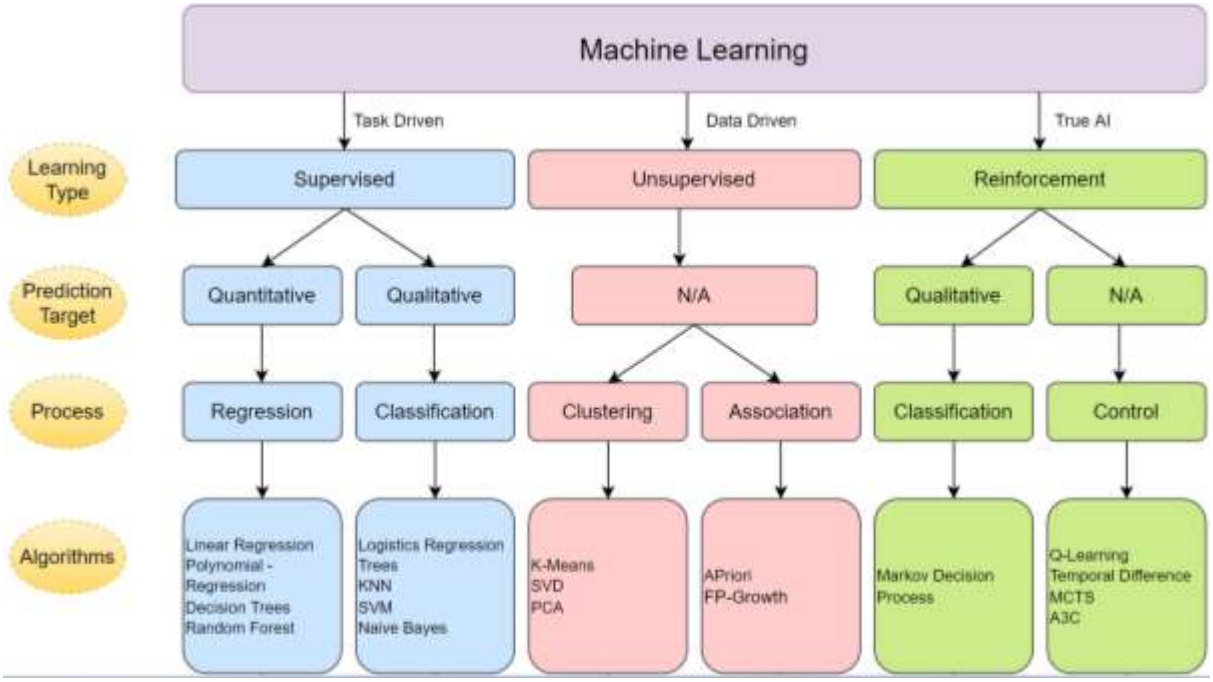



*Figure 15: Areas of Machine Learning*

To understand the ML better, all the concepts and the key terms related to ML and will be existed in the different papers have been listed in one table.

| Key term | Definition |
|---|---|
| Machine Learning | Computers learn without being specifically programmed |
| Pattern Recognition | The computer recognizes structures (text, objects) |
| Learning/Training | Give the computer samples to automatically adjust itself |
| Ground Truth | Labels generated by a human expert – what we expect to get |

| | |
|---|---|
| Training Data | Representative example patterns (images) with ground truth during training, the machine adjusts its internal parameters |
| Validation Data | A separate set to find the best method and parameters |
| Test Data | Another separate set for final evaluation test data should be independent; never look at the test data during developing |
| Simple classifiers | KNN (will be explained shortly in the next paragraph)<br>decision trees and Random Forest (will be explained shortly in the next paragraph) |
| Advanced classifiers | Support Vector Machine, Neural networks |
| Advanced Neural Networks | Convolutional Neural Networks (CNN) – for images (fixed size)<br>Long Short-Term Memory (LSTM) – for n-dimensional sequences |

*Table 3: Machine Learning Concepts*

## 2.5 Network Security and AI

From the previous sections, it is obvious that applying AI will achieve a big revolution in the network security field. In this part, an intensive literature review on the different proposed implementations of AI in the field of network security will be conducted. Different areas related to network security can benefit from AI, such as human error detection (van den Bosch, K., and Bronkhorst, A., 2018), decision making (Mokhtarzadeh, Nima Garousi, et al.,2021), threat detecting (Wang, Ping, et al.,2016), fraud detection (Viswam, Anju, and Gopu Darsan,2017), and many others. Threat detection is a kind of wide and highly demanded topic in network security, and the main challenge is related to the difficulties in predicting and capturing the different types of threats. There are many threats defined, but only fifteen of them have been identified by the European

agency as networking threats (Kettani and Polly Wainwright, 2019). Malicious behavior is one of the threats that we plan to handle using the proposed AI techniques. Many of the conducted research applied encryption to protect the network connection from any suspected vulnerabilities. This step is considered a pre-step to protect the network communication, but the work aims to analyze the network traffic and to predict if the traffic is normal or there is a possibility for any abnormal event that might affect the performance network. This kind of problem can be considered as a binary classification problem, and the proposed ML algorithm is trying to classify whether the traffic is normal or suspicious. Some of the binary classifiers are based on probability theorem like Bayes (Ncubukezi, Tabisa, L. Mwansa, and F. Rosaries., 2020 and Alsughayyir, B., Qamar, A.M. and Khan, R., 2019), packet analyzer knows that these types of data cause abnormal behavior in the network 50% of the time $P(S|M)$, by knowing the prior probabilities for both classes, Bayes might assist in finding how likely the data will cause a problem in the network by applying (Alkasassbeh, and Mohammad Almseidin.,2018):

$$P(M|S) = \frac{P(S|M)P(M)}{P(S)}$$

Bayes treats the attribute and class value as random variables. Suppose a given a record with the attributes (A1, A2,..., An), and the goal will be to predict the class C that will maximize P(C| A1, A2,…, An ). To do so, computation of the following probability P(C | A1, A2, …, An) for all the values of C by using the Bayes theorem is done as follows:

$$P(C|A_1A_2 \dots A_n) = \frac{P(A_1A_2 \dots A_n|C)P(C)}{P(A_1A_2 \dots A_n)}$$

As the network traffic data is continuous data which requests us to discretize the range into bins where one ordinal attribute will be used per bin and violates the independence assumption. To make the probability density estimation, work assumes that attribute will follow a normal distribution, then uses data to estimate the parameters of the distribution, for example (mean

and standard deviation). When probability distribution is known, these distributions will be used for estimating the conditional probability P(Ai|c). The main advantages of using Bayes for this kind of binary classification and network monitoring are as follows:

- Strong to isolate the noise points
- Have the ability to handle any missing values by ignoring that instance during probability estimate calculations
- Strong to irrelevant attributes

This method isn't an optimal case and consists of some drawbacks that might be considered too:

- For some attributes, the assumption of independence may not be valid.
- Make use of other tools, such as Bayesian Belief Networks (BBN)

Besides the statistical-based methods, supervised learning methods can be proposed as good candidates to solve this kind of binary problem. The main idea of the learning method is to learn the pattern from the training data and project the same pattern over the unseen data as a kind of predictor. The survey will start by finding out the networking relater work that used simple methods like a decision tree and random forest to solve the network intrusion issue. A decision tree leads to performing the prediction based on a group of questions on whether you belong to certain groups (Rai, Kajal, M. Syamala Devi, and Ajay Guleria.,2016). they proposed a decision tree for network intrusion detection. The researchers defined a set of rules to achieve the study's main objective. The rules start by evaluating if all the training samples fall into the same class, then the left node will be created for that relevant class. The gain ratio of each will be calculated to estimate the value of the node. The last step is based on calculating the information gain of each attribute as follows

$IG_a = Ent(S) - \sum \left(\frac{s-a}{a} * Ent(S_a)\right)$

The accuracy of the proposed method compared with other methods that implementing random forest, GA, and other techniques shows that there is no huge difference in performance.

On the other hand, other researchers proposed to use the advanced version of the decision tree, which is called random forest (Jin, Yunhu, et al.,2016). Random forest is a collaborative learning method made up of several decision trees that output the category that is the median of the class output by the individual trees.

The name was coined in 1995 by Bell Labs' Tin Kam Ho, who proposed the concept of random decision forests. It's not the same as decision trees, which are made up of individual learners. They're one of the most widely used learning approaches for data exploration. (Farnaaz, Nabila, and M. A. Jabbar, 2016) Developed the following strategy for dealing with binary classification issues using random forest.

The rules for constructing each tree are as follows:

• Assume that N is the amount of training instances and M is the amount of variables in the classifier.

• We are given the amount (m) of input variables that will be used to make a decision at a tree node, (m) should be substantially smaller than M.

• Select the training set for this tree by selecting n times with a substitute from all N training examples available. By predicting the classes of the remaining examples, you may assess the tree's inaccuracy.

• Choose (m) variables randomly for each node of the tree to base the choice on that node. Then calculate the optimum split according to the training set's (m) variables.

• Each tree is fully mature and hasn't been pruned (as may be done in building a normal tree classifier).

A new sample is dragged down the tree for prediction. In the terminal node where it ends up, it is given the label of the training sample. This technique is repeated for each tree in the group, with the average vote of all trees being presented as random forest prediction.

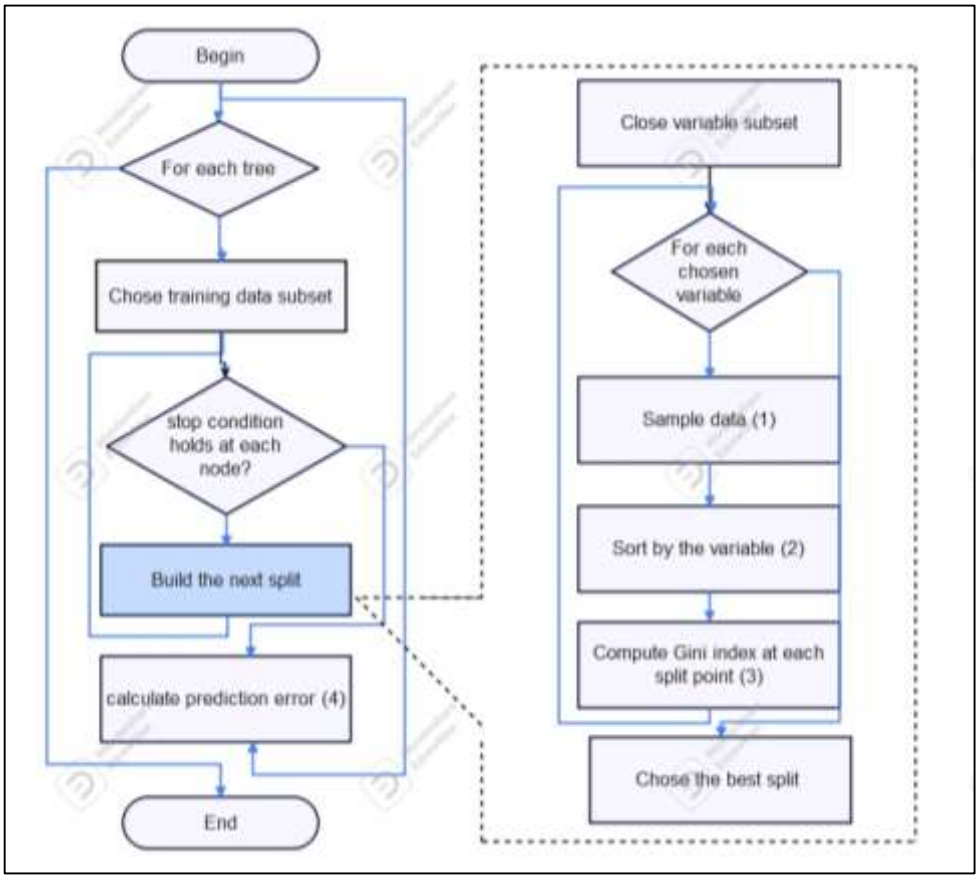



*Figure 16: Random Forests for ID in networking*

(Negandhi, Yash Trivedi, and Ramchandra Mangrulkar., 2019) covered in the research the practical factors that should be taken into consideration while designing the random forest for the ID problem. These factors can be listed as follows:

- Splits are selected according to the purity measure:

  - For Example, squared error, deviance (classification), or using Gini index

- How to choose N?
  - Build the trees till the error is no longer decreases
- How to choose M?
  - Try to suggest defaults, half of them and double of them, and pick the best.

The main advantages of the random forest can be listed as follows:

- It's one of the most precise learning algorithms in the industry. It generates an accurate classifier for different data sets.
- It works well when the dataset is huge.
- It can cope with tens of thousands of inputs without losing any.
- It calculates the relevance of numerous categorization parameters.
- The forest produces an approximate internal solution of the prediction error as it grows.
- It provides a method for estimating incomplete information that works well and is accurate even when a significant quantity of data is missing.
- Techniques for balancing error in uneven data sets with a class population are included.
- The forests that are built can be saved and later utilized with other data.
- The relationship between the independent dependent and the classification is revealed through prototypes.
- It estimates distances between pairs of instances, which might be useful for clustering, spotting outliers, or creating interesting data visualizations (by scale).
- Using the capabilities stated above, unlabeled data can be utilized to build unsupervised clustering, data visualizations, and outlier identification.
- It gives a method for experimenting with variable interactions in order to find them.

- While there are some limitations to dealing with random forests, it has been reported to overfit certain datasets with noisy classification/regression tasks.
- When data contains categorical variables with varying numbers of levels, then random forests are rigged in favor of qualities with more levels. Therefore, random forest variable significance ratings for this kind of data are unreliable.

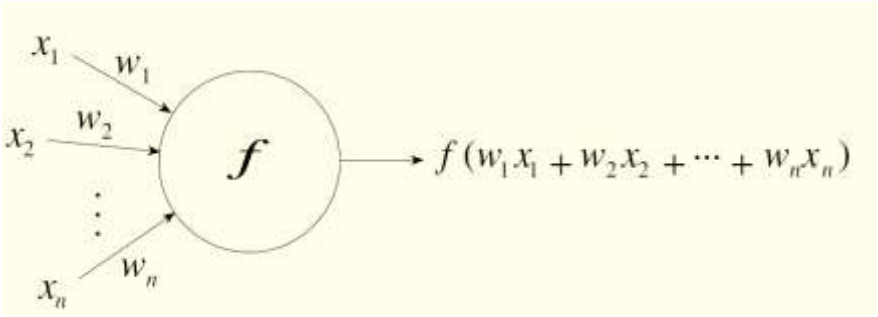
Some of the researchers proposed to apply clustering to handle the malware detection (Liu et al. 2017); the proposed method processed the data by using feature extraction, decision making and to detect the new malware by using the clustering. after evaluating 20000 or more malware instances, the proposed system classifies the unknown malware effectively with 98.9% accuracy and 86.7 % new malware has been detected successfully. In this experimental research, the researchers demonstrated a combined feature extraction technique based on a grayscale picture, an Op-code n-gram and the import function. The technique successfully describes the characteristics of the various programs by using the texture of the malware in the first stage; then, in the following stage, they offered a system for decision making to detect the malware based on the malware's high dimensionality, Shared Nearest Neighbor for clustering the samples.

Some other LR work tried to list and evaluate different ML techniques that can be applied for network analysis. various techniques of network traffic classification such as Machine learning technique, port-based technique, and pay load-based technique was discussed thoroughly, and researchers developed a real-time internet data set by using a tool called Wireshark to capture the network traffic such as FTP, WWW, P3P, TELNET and DNS applications then they used a NetMate tool to extract around 23 features from the traffic. Finally, they applied four machine learning classifiers which are Naïve Bayes classifier, support vector machine classifier, bays

Net, and decision tree classifier, which shows high accuracy in comparison with the other classifiers (Shafiq, Muhammad, et al. 2016).

Working on proposing a different kind of features out from the available networking data can be considered as another direction in this type of work.

More advanced techniques based on the neural network are also proposed to handle networking-related matters. The main idea behind NN is to make computer intelligence by simulating the neural behavior on PC. The simulation had been done by proposing concepts like perceptron and activation function. $y = f(\sum wi * xi)$ as demonstrated below

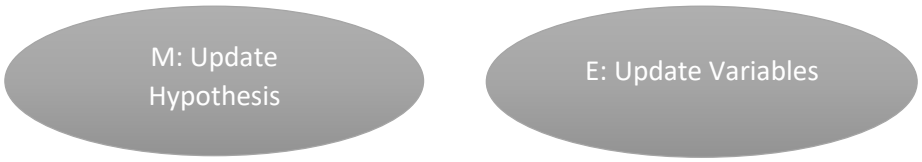



A lot of activation functions can be proposed to fire the neuron for binary classification; Relu was proposed by (Tang, Ying, and Mohamed Elhoseny. , 2019) to be used for the network intrusion problem. The idea behind the function is to define the relationship between input amplitude and output amplitude. In principle, the resulting input and output amplitudes correspond to spike rates in a quasi-stationary state. Thus, we consider a spike-rate code and neglect fast transients and the complex dynamic properties of biological neurons

On the other hand, (Lin, W.H., Lin, H.C., Wang, P., Wu, B.H., and Tsai, J.Y., 2018) implemented the sigmoid function for a similar task. The main idea behind learning is to maximize the expectation function. This goal can be accomplished by an iterative technique to find the

maximum likelihood of factors in statistical models. The model depends on unobserved latent variables that alternate between E and M

- Expectation (E) – using current estimate value for the parameters
- Maximization (M) – re-estimate parameters maximizing the expected log-probability found in the E step




This is a kind of shallow learning that has a limited number of layers has the advantage of fast processing, but it might fail in case there are huge data available for training. The progress of the neural simulation started with a work called "Receptive fields of single neurons in the cat's striate cortex" by Hubel & Wiesel 1959. The work after that tends to introduce multi-hidden layers to make the NN deeper and able to extract more features through these different layers. This kind of work was called a deep neural network. Deep learning provides the chance for the computational models to be consist of many layers to have a better understanding of the provided data (Singla, A., Bertino, E., and Verma, D., 2020). There are mainly two types of DNN which are feed forwarded network which doesn't have any feedback from the output layer to the input layer. At the same time, the back propagation network relies on getting a response from the output to the input. Another proposed DNN is called convolutional neural network (CNN), which was done by K. Fukushima, who proposed a work named Noncognition in 1988. Based on the idea proposed by Fukushima, Yann LeCun conducted work on a backdrop-style learning algorithm to CNN in 1989. A big revolution was achieved in this field by proposing ImageNet for large-scale visual recognition competition. The traditional topology of different

deep learning models such as Lenet, AlexNet, Inception, VG16, and many others started to appear from that period till nowadays.

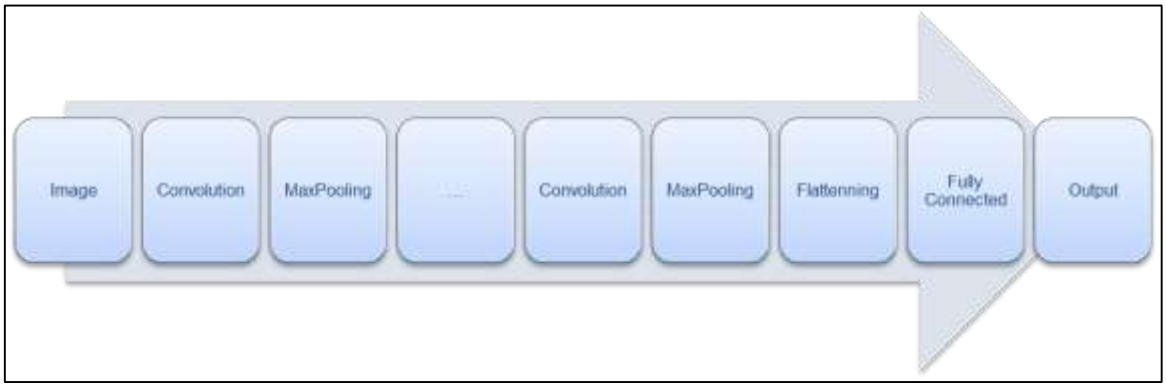



*Figure 17: Typical CNN Structure*

As an explanation of the layers proposed on the typical CNN structure, the convolution layer will try to implement a mask over the network to get a new set of values representing new extracted features. The masks can be designed for the purpose of extracting edged like a gaussian mask blur the image to decrease the noise like an average mask. The next proposed layer might be a pooling layer. The purpose of this layer is to decrease the dimensionality of the extracted features to simplify the task. The last layer in the convolution process is called a fully connected layer, and this layer will flatten the two-dimensional features to one dimension. The reason behind this flattening is to prepare data to be ready to inject to the (Amrollahi, Mahdi, et al., 2020) conducted a literature review that covered the different Machin learning models that applied to enhance the performance of networking. One of the works introduced in this field is related to the Recurrent Neural Network (RNN) and the long short-term memory neural network (LSTM). This research applied the following to reach the stated target:

- Recurrent connections are added to keep information of previous time stamps in the network

- Context information is used; however impossible to store precise information over long durations
- Vanishing Gradient problem: Usual RNN forget information after a short period of time
- LSTM NN adds a memory cell
  - LSTM unit is a memory cell that contains three gates: Input, Output, Forget.

(Yin, C., Zhu.,& He, X., 2017) implemented the recursion neural network to detect the malicious behavior over the network. The work conducted a comparison analysis between the traditional machine learning techniques and RNN-based methods. The performance of the RNN proves the value behind implementing these methods instead of using conventional machine learning techniques.

Using the DNN to solve the issues related to network security started in 2014. The work tried to combine both DNN with genetic algorithms using a dataset called KDD. The purpose behind this combination is to simplify the difficult structure of the network and to detect possible threats over the network. The work prompted a self-adaptation structure for the network to be able to handle different scenarios (Gao, N., Gao, Q., & Wang, H. 2014). Using CNN as a kind of classifier for different types of malwares that might attack the network is another promising field in network security (Yin, C., Zhu, & He, X. 2017). The proposed architecture that was used for this purpose had a masking layer to extract features followed by a feed-forward neural network. The features will be extracted from the provided logs about the network traffics. Then, these features will be trained to detect the intrusion and other abnormal behavior in the network. The overall performance of the proposed architecture was compared with other traditional ML methods to prove the efficiency of this technique (B. Kolosnjaji, G. Eraisha, Webster, A. Zarras, and C. Eckert,2017). Batching was recommended to be added as a pre-processing step for an

intrusion detection network solution using a convolutional neural network. The batching will assist in filtering the redundant data that might lead to negatively affect the performance of the feed forwarded NN.

A technique that is able to adapt to detect and classify different events that occurred in the network was explored by (Vinayakumar et al. 2019). Evaluation of the performance was conducted over different algorithms for this multi-task challenge. The researchers found out that the DNN outperformed in this complicated scenario compared with other proposed methods.

Other works suggest having a fusion between different architectures for better performance. Fusing the convolutional neural network with autoencoders was proposed by (Yu et al., 2017) to enhance the performance of network security. The number of the features extracted from the provided data was increased as it was provided from both different techniques. The data used for training is a combination of normal data and abnormal ones. The conducting results showed an efficient improvement in recall, precision, and accuracy.

## 2.6 Chapter II Summary

There are several studies about the network security threats detection, enhancing the intrusion detection systems (IDS) and Intrusion Prevention Systems (IPS)to detect and prevent the attacks in the networks. But the main challenge is most of the studies use the same dataset NSL-KDD (Knowledge Discovery in Database) for the experimental studies that have some problems and does not consider as a perfect representative of the current real networks; it also has a lack of public dataset for the IDS based network. Forty-four papers have been analyzed with the following outcomes, SVM is the most used algorithm in the papers with 25%, then Artificial Neural Network (ANN)and Random Forest were used equally by 14%. The Accuracy

evaluation technique is used in 58% of the papers. The Precision and the Recall are equally used in 34% of papers, then the other measure such as F1-Score, Roc, and AUC are also used. The majority of the papers used the KDD dataset with the attributes U2R (User to Root), R2L (Root to Local,) and Probe. DDoS attack attribute also was hugely used in the papers. The rest, such as Web Attacks, Spam, Worms, Botnet, Brute Force, Services and Network (TCP, ICMP), are used but with less percentage. The papers were recently published. The majority of the papers (32% of the papers) were published in 2019, 27% were published in 2018, 23% were published in 2016, and the rest, 19%, were published between 2020 and 2021.

Some useful definitions and terminology are presented in this chapter like Machine Learning, Machine Learning, Pattern Recognition, Learning/Training, Ground Truth, Training Data, Validation Data, Test Data, Simple classifiers, Advanced classifiers, Advanced Neural Networks.

# Chapter III Research Methodology

This chapter introduces the methodology that has been followed in this study. The CRISP-DM Model with six main phases and sub-phases are used to describe the data life cycle. The main phases are business understanding, data understanding, data preparation, prediction modeling, evaluation, and deployment phase. Each phase will be explained in detail.

## 3.1 Introduction

This research followed the CRoss-Industry Process for Data Mining (CRISP-DM) methodology, including six phases describing the data life cycle. These phases help the researchers organize, plan, and implement machine learning for the project, as shown in the below Figure.

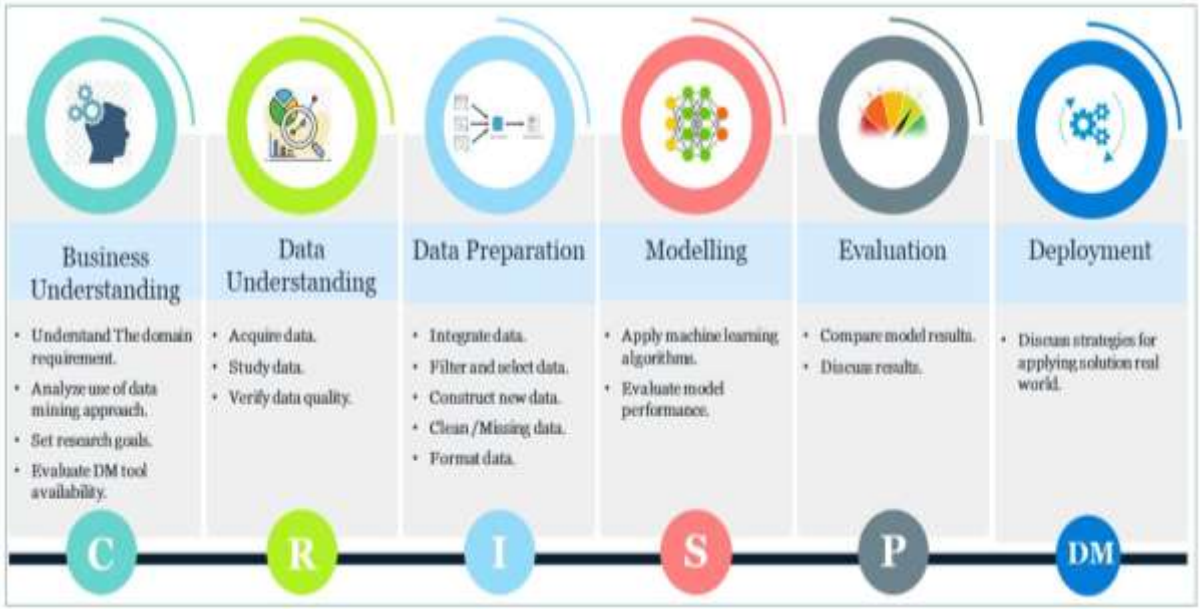



*Figure 18: CRISP-DM Model Phases*

This section will start to discuss the performed tasks in each phase.

## 3.2 The Business Understanding Phase.

The developed framework can be integrated with the existing data management system to predict a threat based on the network security behavior. Furthermore, the admin will take action depending on the machine learning recommendation, or remedial action can be taken by the system automatically.

### 3.2.1 The domain requirement

The organization model for this study is designed for one branch with two different buildings for a university. The primary data center or the Main Distribution Frame (MDF) room that located in the main building. Another building contains an Intermediate Distribution Frame (IDF) room connected by fiber optic cable with the MDF. The organization applications are hosted internally in the local data center. The services are published in the public domain in order to extend the service's availability internally and externally through the internet.

The organization stakeholders can access all services through the internet, such as registering new courses, adding a subject, dropping the course, checking the schedule, applying for letters, applying for leave, logging in to the library services, and more extra services can do. Also, the stakeholders can do the financial services through an online payment. Therefore, network security is essential for the organization and plays a vital role in securing connections and data.

### 3.2.2 Network structure:

The organization network has been designed with various layers of security, starting from the endpoint security with enterprise ani virus, then the network security by segmenting the network to three various networks (faculty, students, and staff). Even within the same network, there is a logical segmentation between departments such as the finance department with a VLAN (Virtual Local Network) different from the registration VLAN. The published application

services are also secured by separating the servers in a secure zoon called DMZ (Demilitarized Zoon) between two firewalls.

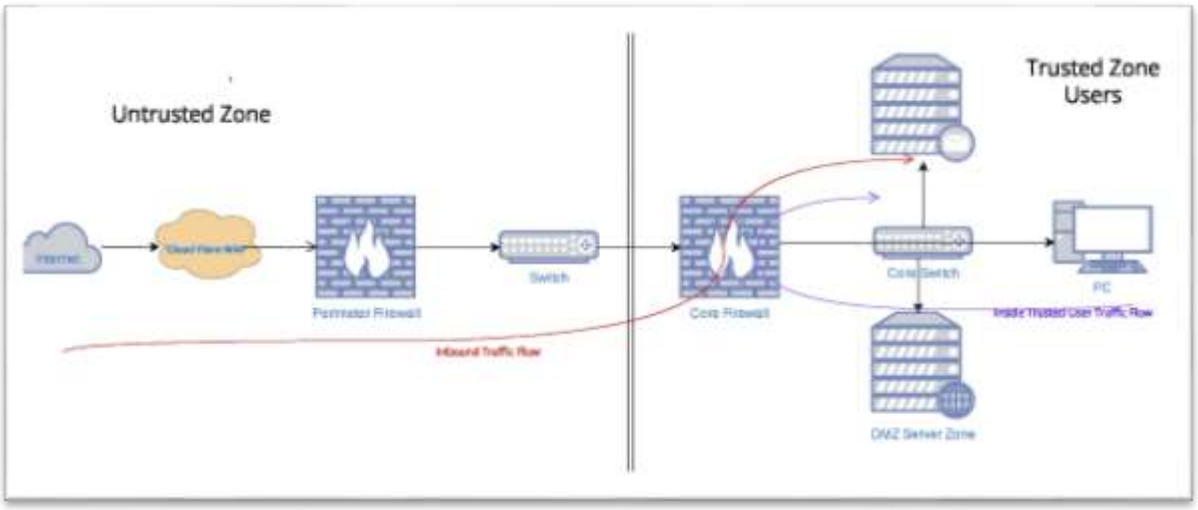



*Figure 19: Network Traffic Flow*

### 3.2.3 Access Situation:

The organization network has been designed to segregate the traffic based on the request and divide it into internal requests and external requests. For example, if there is a request from inside the campus to log in to the student portal, faculty portal, or staff portal, this request will be forwarded to the core firewall. Then action will be taken by the network server to accept or drop the request based on a list of configured policies/ rules set by the IT administrator. If the request is accepted, it will pass to the target server in the DMZ.

In case of an external request, two additional layers of security have been added. The cloud controls the first layer using Cloudflare WAF (Web Application Firewall). The second layer is accomplished by adding a firewall between the local network and the internet called a parameter firewall, which will work as a security gateway. The traffic procedure is described as follows,

1- The requester login to the internet seeking a particular service.

2- The WAF (Web Application Firewall) will check the request based on the security team's preconfigured rules and policies.

   a. If the request is invalid, then it will be dropped.

   b. If the request is valid, then it will pass the Cloudflare WAF.

3- The request will pass to the parameter firewall (Security Gateway)

   a. If the request is invalid, the request will be dropped

   b. If the request is valid, it will pass to the core firewall.

4- The core firewall will check the request then act as follows,

   a. If the request is invalid, the packet will be dropped.

   b. If the request is valid, the packet will pass to the DMZ.

5- The request will take the required service from the DMZ server services and return it to the requester.

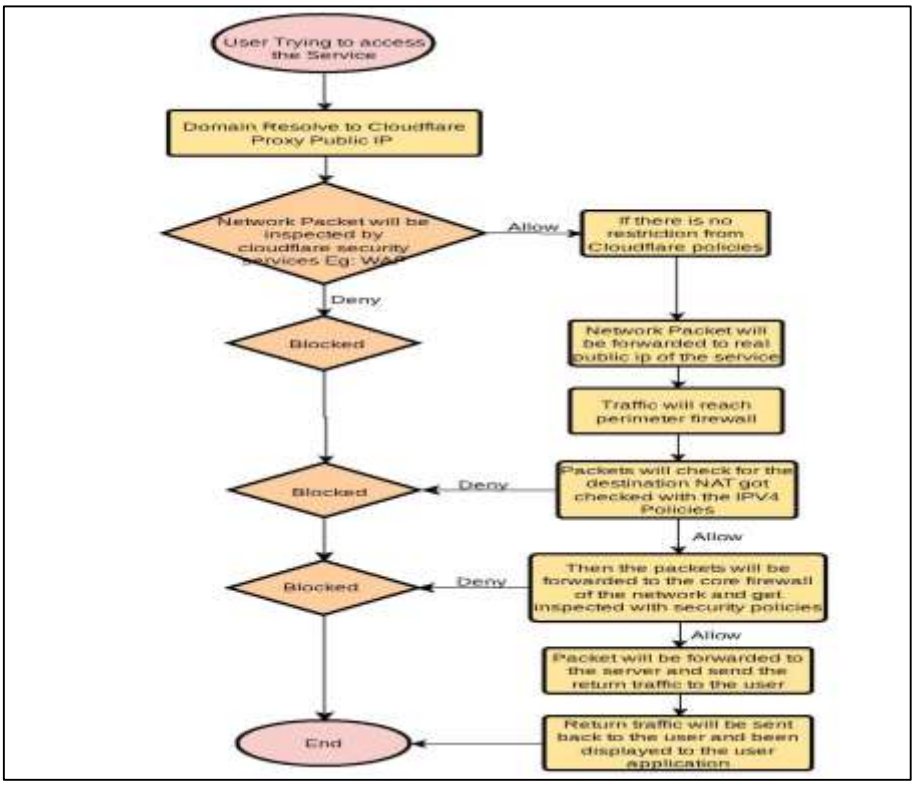



*Figure 20: Connection flowchart*

### 3.2.4 Determine the Data Mining Goal:

As mentioned before, this research aims to establish a framework for the network security behavior with the current data management system and then use the machine learning algorithm to predict if there are any threats or not then remedial action will be taken even to allow the clean data to pass or to drop the infected data. Using machine learning will help reduce the effort required by the technical security engineer to build the security rules for the organization.

The machine learning algorithm will be used for the prediction process, so there are some questions we need to be considered during the study, such as: -

***Question1***: does the machine learning algorithm will be able to detect the threats effectively.

***Question2***: How far ahead of time can machine learning anticipate possible threat detection?

To answer this question, a dataset that includes 17 parameters (features) has been considered during the feature selection process for the machine learning model.

***Question3:*** Which characteristics (attributes) are the most important predictors of the threats?

To answer the previous questions, we need to move to the next phase related to understanding the data.

## 3.3 Data Understanding Phase

The Data Understanding phase is obtained to study correlations between variables and detect quality issues such as outliers and missing data. During this phase, a strategy to address the issues that have been discovered. At this level, data selection strategies and pre-processing criteria are studied in greater depth before being executed in the Data Preparation phase. The datasets are described in the next section.

### 3.3.1 Data Collection:

Collecting data is the most important part of the research. The data has been collected from the firewall logs file, which recorded all the events related to data access or data transfer. The data has been collected in a different time range. The data size is related to the number of logs stored per each file. The maximum number of records in each log file is 10000 records, and then it will overwrite data. To generate a related data set, a custom report has been configured and customized by a security engineer to collect data with dependent attributes such as ID and category features. It has been noticed that all the ID 27 is related to the security service category and considered under attacks groups but under two priorities (Information, Alert). This type of data will be passed through the firewall as per the action. On the other hand, ID 23 is under the same category of security services and the same group of attacks, but with a different action, and the firewall will drop the data because it was detected as an IP Spoof Detection. The data can be extracted as a text file or a CSV file. There are different attributes that have been explained in detail in the below table.

| Item | Definition / Use |
|---|---|
| **Time** | Time of the traffic entered the firewall; it includes data /time. |
| **ID** | The signature pattern matched the firewall IPS, and the application control signature log will show the ID of IPS or Application control.<br>Each ID is related to a certain group, event, and priority ex. ID 23 is related to the security services category, attacks group with Alert priority. |
| **Category** | shows the log belongs to which category, e.g., the logs related to (Firewall setting, Security Services, Anti-SPAM, Users, VPN Connection, Network, High Availability setting ). |

| | |
|---|---|
| **Group** | From respective to the category, it will show a group type, e.g., In the security service category, the type of log will be generated with IPS or application control or botnet protection. Block types can be identified by the group.<br>Theirs is different types of groups such as (Application Control, Flood Protection, AppFlow, Radius Authentication, IPS (Intrusion Prevention Systems), Multicast, VPN IPsec, Content Filter, Attacks, Botnet Filter, Content Filter). |
| **Event** | What is the reason why the file has been blocked can be illustrated by event logs. There are different types of events such as (IP Spoof Detected, Website Blocked, Application Control Detection Alert, Using LDAP Without TLS, IPS Detection Alert, Link-Local/Multicast IPv6 Packet, LAN ICMPv6 Deny, Port Scan Possible, Wrong Admin Password). |
| **Msg. Type** | The massage that related to the event; (Standard Note Ethernet Network, Standard Application Control Message String, Standard IDP Message String, Standard Note Blocked). |
| **Priority** | There are seven types of priority in logs, e.g., emergency, critical, alert, error, warning, debug, informational, and notice. |
| **Ether Type** | It will show the traffic ether type IPV4 or IPV6 traffic. |
| **Src. MAC** | Source Mac address of the source device. |
| **Src. Vendor** | the Mac address will be searched in the public mac database and identify the device type (VMware, Cisco Systems, Cisco Meraki, TP-link technologies, Samsung electro-mechanics, Fortinet, Hewlett Packard, AZURE technology). |
| **Src. Int.** | Source interface from where the traffic is passing. |
| **Src. Zone** | Source Zone from where the traffic is passing. |
| **Dst. MAC** | Destination Mac address of the source device |

| | |
|---|---|
| **Dst. Vendor** | the Mac address will be searched in the public mac database and identify the device type (VMware, Cisco Systems, Cisco Meraki, TP-link technologies, Samsung electro-mechanics, Fortinet, Hewlett Packard, AZURE technology). |
| **Dst. Int.** | destination interface from where the traffic is passing |
| **Dst. Zone** | destination Zone from where the traffic is passing |
| **Src. IP** | Source IP address of the device that initiates the request |
| **Src. Port** | Source Port Number that used to pass the request |
| **Src. Name** | If the source device has any name registered, it will be displayed |
| **Src.NAT IP** | If there is any Source address that has been NATTED, it will be displayed here |
| **Src.NAT Port** | If there is any Source Port that has been NATTED, it will be displayed here |
| **In SPI** | It will check the IN stateful packet inspection table if the traffic is not stateful, it will be dropped |
| **Dst. IP** | The destination IP address for the target device |
| **Dst. Port** | destination Port Number |
| **Dst. Name** | If the destination device has any name registered, it will be displayed |
| **Dst.NAT IP** | If there is any destination address that has been NATTED, it will be displayed here |
| **Dst.NAT Port** | If there is any destination Port has been NATTED, it will be displayed here |
| **Out SPI** | It will check the OUT stateful packet inspection table if the traffic is not stateful, it will be dropped |
| **IP Protocol** | Which protocol will be used in the traffic, TCP (Transmission Control Protocol), UDP (User Datagram Protocol), or ICMP (Internet Control Message Protocol), ESP (Encapsulating Security Payload) |
| **ICMP Type** | If the traffic is ICMP, it will show the type here |
| **ICMP Code** | It will show the ICMP code Type |
| **RX Bytes** | Indicate the amount (volume) of data that Received in Bytes. |
| **TX Bytes** | Indicate the amount (volume) of data that Transferred in Bytes. |

| | |
|---|---|
| **Access Rule** | To indicate which access rule will be applied to process the traffic . |
| **NAT Policy** | If Nat is applied, it will show which NAT rule is used. |
| **Username** | The traffic is related to which user. |
| **Session Time** | Traffic Session Duration Timing for the event. |
| **Session Type** | Traffic Session type which event triggered. |
| **IDP Rule** | If any instruction detection policy is configured and triggered the event, it will appear here. |
| **IDP Priority** | IDP policy priority value if any. |
| **HTTP OP** | display the value of Sonic Wall OS assigned for content filter policy while processing<br>e.g: 0 = NO OPERATION<br>1 = HTTP GET<br>2 = HTTP POST<br>3 = HTTP HEAD |
| **URL** | the event triggered URL will be displayed here if the NPCS is enabled |
| **VPN Policy** | Any VPN policy applied for the event traffic |
| **HTTP Result** | Here it will display HTTP response code if any e.g: 200, 404, 500 HTTP codes |
| **Block Cat** | If any event is blocked by any category, it will be displayed here |
| **Application** | the application which triggered the event |
| **FW Action** | Firewall action for the event triggered, e.g., Allow/ Block |
| **DPI** | If any deep packet inspection is applied, it will show here |
| **Notes** | general Noted for the event |
| **Message** | The messages generated by the firewall for the event-triggered |

*Table 4: Attribute description*

## 3.4 Data Preparation Phase

The Data Preparation step is the most important and time-consuming of all the stages. At this stage, the sub-stages have been applied, such as the data integration process, data filter, selection, data construct, and data cleaning. The Rapid Miner application has been used in the data preparation process. It is used to remove the missing values in the dataset and create a feature selection process. The dataset was collected from different resources: the Cloudflare firewall, the parameter firewall, and the core firewall.

### 3.4.1 Data Integration and Pre-processing

As an initial pre-processing stage, the logs from the three resources were collected, filtered, and selected the related records. These records are related to firewall setting, which is 62823 records out of 123029 in total; the other types of records are related to Network logs (Network, VPN, High Availability) that are 10617 records out of the whole. The third type of logs is related to the security logs, which contain anti-Spam, Users, System, and Security services. These records are 49589 records out of 123029 in total. These data have been integrated and pre-processed by using Microsoft Access and Microsoft Excel as per the following figure

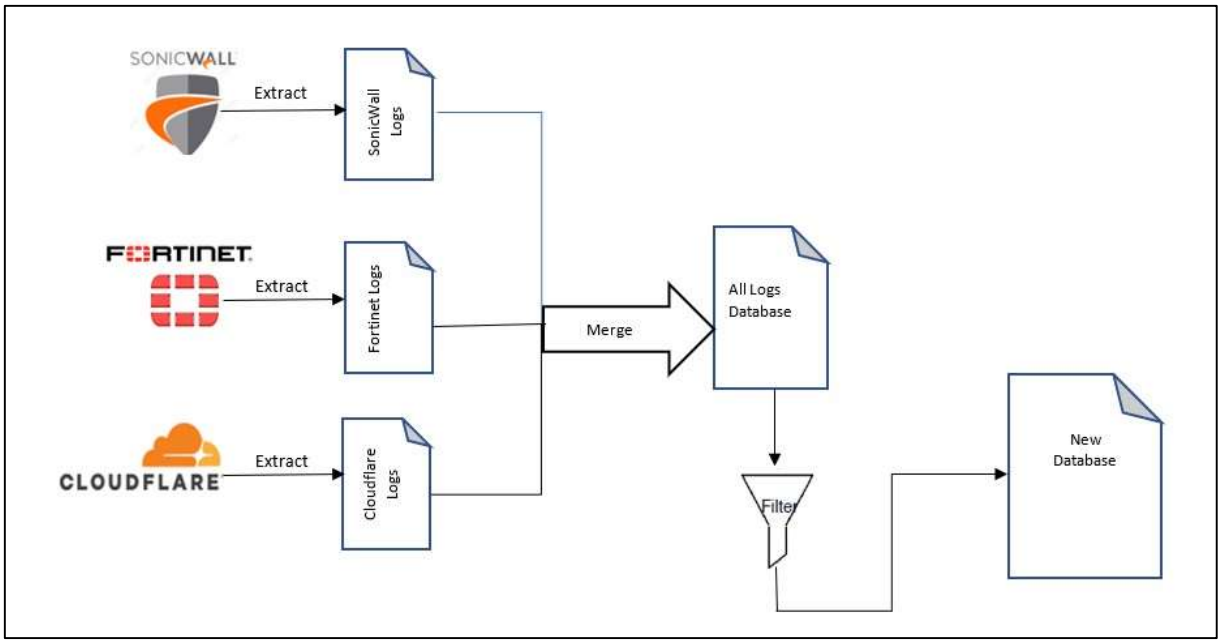



*Figure 21: Data Integration and Pre-processing*

### 3.4.2 Feature Selection

The complexity of data causes a slew of problems, including model overfitting, which means the model can't be generalized (Kunang, Y.N., Nurmaini and Suprapto, B.Y., 2021). in this research, the number of features has been reduced by selecting only that related to ID, Category, Group, Ether Type, Source vendor, Source Zone, Destination vendor, Destination Zone, IP Protocol, PX Byte, Application, Event, Priority, and the action. By using Python, the Correlation Matrix, and the heatmap, we found that all the numerical attributes have an impact on the class label power Action. Herewith the research will include these attributes in the study.

### 3.4.3 Feature Transformation (One hot Encoding)

In this research, there are some categorical variables that contain label values, not a numeric value in the data set such as priority, group, event, massage type, source vendor, source zone, destination zone, and IP protocol .so the researcher transfer all the categorical attributes to numerical attributes.

### 3.4.4 Feature Scaling (Data Normalization)

In this stage, the researcher will use this method to transfer the data of RX bytes attribute, which have the value of byte that varies widely into a scaling .by using Python code, the researcher tried to use the Min-Max normalization and Standardization normalization (Z-score Normalization).

### 3.4.5 Data Integration

Each firewall provider standardized an ID for the traffic logs. It is a unique value based on the provider. The logs tables were joined together and generated a combined dataset using MS Access.

### 3.4.6 clean-up / missing values

The dataset contains 53 attributes in total. The set has been cleaned up by removing the irrelevant features that will not help in the study, such as the log messages, which will give a text message for each log; this log message can be even from the vendor or customized during the configuration, also the source IP address and the destination IP address. These features have been removed from the study because the devices IP address has been assigned by the DHCP server, and this IP is temporary for a specific time based on the device owner; if it is related to a guest, it means the IP age will be 4 hours, but if the owner is a student then the IP age will be 12 hours, but if the owner is staff then the IP age will be four days. So, the source IP and the destination IP will not help in the study since the same device can log in to the network with different IP. Another feature has been removed, which is called RX byte, that is related to the data size that has been transferred during the Connection; we care about the data type, not the data size. The sender's name and the destination name also have been removed for the confidentiality of the data. Then the set has been left with only 16 attributes, 15features as an input feature, and one feature as an output or target for the network security behavior prediction.

## 3.5 Prediction Model Phase.

### 3.5.1 KNN

The (KNN) k-nearest neighbors' technique is a kind of data classification method that is used to estimate the probability. That data point will be part of one of two groups depending on the data points closest to it.IT is an algorithm for supervised Machine learning which is used for solving regression problems and classification problems; however, mainly it used for classification. It is a lazy learning algorithm, also known as a non-parametric algorithm. The Euclidean Distance is used to measure the distance between any two vectors X and Y as per the below equation

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2}$$

### 3.5.2 SVM

SVMs stand Support Vector Machines and are used to solve classification issues and regression issues [38]. But it is mostly utilized in Machine Learning for Classification situations. SVMs can professionally perform non-linear categorization as well as linear categorization utilizing a method or factor called Kernel, which completely translates their inputs to high-dimensional feature spaces.

SVM's purpose is mainly to partition datasets into classes in order to identify a maximum marginal hyperplane (MMH), which may be achieved in two steps:

- First, SVM will iteratively construct hyperplanes that best separate the classes.
- The hyperplane that accurately separates the classes will then be chosen.

Nominal properties are not supported by the SVM algorithm. As a result, for this classification algorithm, we convert all nominal Attributes to numerical values.

### 3.5.3 Decision Tree

Decision Tree is a kind of supervised learning algorithm that is used to contribute to solving Regression issues and Classification problems, but it is mostly to solve classification issues. In this tree-structured classifier, the interior points represent dataset features, the branches provide decision rules, and each leaf delivers the output. The tests or decisions are made based on the physical characteristics of the given dataset.

## 3.6 Evaluation Phase.

This study aims to develop a framework for network security behavior integrated with the current data management system to predict the threats for administrator remedial actions by using Machin-Learning (ML) technique based on analyzing the current network activity logs; then, predict the actions to the traffic using dual-class labels; that classes are Allow (Positive Class) and Drop (Negative Class). The main interest is to predict the dropped packet accurately while avoiding misclassification of the allowed packets. Inaccurate classification results in inefficient use of resources in dropping intervention of the packets that are likely to pass through the firewalls.

In this research, more than evaluation metrics will be used to measure the performance of the prediction model, such as Accuracy, Recall, Precision, F-Score, Receiver Characteristic Operator (ROC), and Area Under Curve (AUE).

Each instance of the testing data is classified into one of two classes by the classification algorithms: Y (Allowed Packets) or N (Dropped Packets). A confusion matrix captures the four alternative classifications of the instances. as per the following confusion matrix table

| Action | Actual (Y) | Actual (N) |
|---|---|---|
| Predicted (Y) | TP (True positive) | FP (False Positive) |
| Predicted (N) | FN (False Negative) | TN (True Negative) |

*Table 5: Confusion Matrix*

This matrix summarizes the predicted and the actual values as the following:

TP (True Positive): represent the number of the positive observations that are correctly classified.

FN (False Negative): represent the number of the positive observations that are incorrectly classified.

TN (True Negative): represent the number of the negative observations that are correctly classified.

FP (False Positive): represent the number of the negative observations that are incorrectly classified.

in the next part, a brief explanation about each performance evaluation matric will be introduced:

### 3.6.1 Accuracy:

The accuracy of any machine learning model is considered as a metric for defining which model is the best at identifying relationships and patterns across variables in a dataset depending on the inputs, or training, data (Iqbal, M.F., Zahid, and John, L.K., 2019). The deeper a model's generalization to 'unseen' data is, the deeper predictions and perceptions it can generate, and thus the more business profit it can deliver.

The accuracy of the following definition

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predections}$$

But for the binary classifications, the accuracy can be calculated in terms of negatives and positives as per the following

$$Accuracy = \frac{\text{True Positives} + \text{True Negatives}}{\text{True Positives} + \text{True Negatives} + \text{False Positives} + \text{False Negatives}}$$

### 3.6.2 Recall

The Recall is a measure of how well the model detects True Positives

$$Recall = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Negative\ (FN)}$$

This metric is also called sensitivity or True Positive Rate (TPR) metric, that measures the number of positive observations, which is correctly classified in this study; it represents the number of allowed packets as a percentage of all traffic.

### 3.6.3 Precision

Precision is one metric for a machine learning model's performance – it measures how accurate a model's positive prediction is (Lin, P., Ye and Xu, C.Z., 2019). Precision (i.e., the number of false positives + the number of true positives) is defined as the number of (TP) true positives divided by the total number of positive predictions.

$$Precision\ = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Negative\ (FN)}$$

The precision is known as (PPV) Positive Predictive Value that measures the number of positive observations among all the positive prediction

$$PPV = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Negative\ (FN)}$$

### 3.6.4 F-Score

The F-score, also called the F1-score, is a metric that shows how accurate a model is on a certain dataset. It's used to evaluate binary classification algorithms that categorize examples into " negative " and " positive " categories.

The F-score is a way for integrating the precision and recall of a model. It is defined as the harmonic mean of the model's precision and recall.

The F-score is a widely used metric for evaluating information retrieval systems such as search engines, as well as a wide range of machine learning techniques, particularly in natural language processing (NPL).

$$F1 = \frac{2}{\frac{1}{Recall} X \frac{1}{Precision}} = 2\frac{Precision\ x\ Recall}{Precision+Recall} = \frac{TP}{TP+\frac{(FP+FN)}{2}}$$

### 3.6.5 ROC

The Receiver Operator Characteristic (ROC) curve is a tool for evaluating binary classification issues. It's a probability curve that compares the TPR to the FPR at various thresholds.

$$\text{(True Positive Rate)}TPR = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Negative\ (FN)}$$

$$\text{(False Positive Rate)}TPR = \frac{False\ Positive\ (FP)}{False\ Positive\ (FP) + True\ Negative\ (TN)}$$

On a ROC curve, TPR vs. FPR for different categorization parameters are presented. More objects are categorized as positive as the classification threshold is dropped, resulting in a growth in both False Positives (FP) and True Positives (TP). The graphic below represents a typical ROC curve.

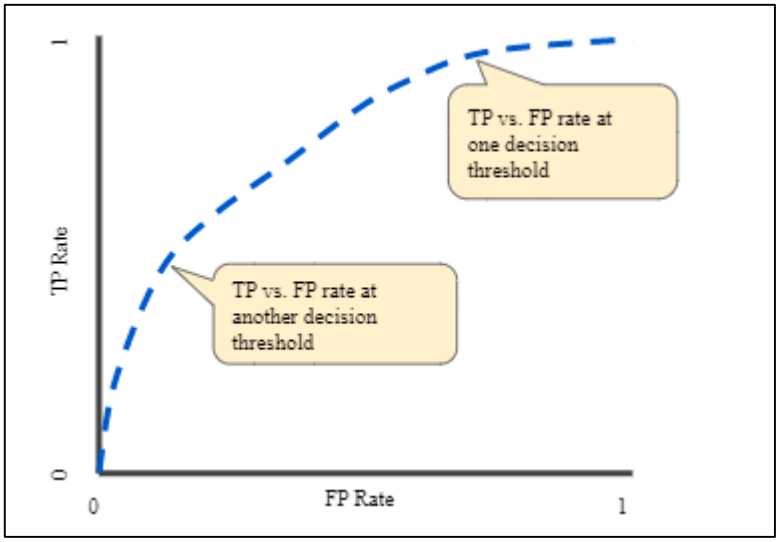



*Figure 22: FP vs TP rate at different classification values*

### 3.6.6 AUC

"Area under the ROC Curve" is abbreviated as AUC. In other words, the AUC evaluates the entire two-dimensional area beneath the complete ROC curve (think integral calculus) from (0,0) to (1,0). (1,0). (1,1). Scaling has no effect on the AUC. It measures how effectively predictions are arranged rather than measuring absolute values.

Categorization criteria have no effect on AUC. It evaluates the model's prediction accuracy independent of the categorization level chosen.

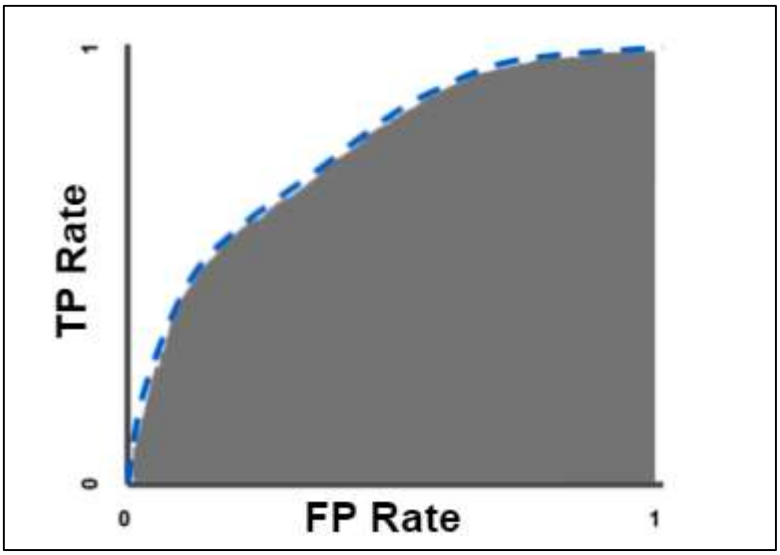



*Figure 23: Area Under the Curve*

## 3.7 Deployment Phase.

The data set has been divided into 70% as a training set and 30% as a testing set. Python code has been used for the modelling, evaluation, and deployment phase.

## 3.8 Chapter III: Summary

123029 raw data records as logs have been collected from three different resources in the organization; two of them are located in the local data center (SonicWall firewall and FortiGate firewall), and the third resource is located in the cloud, which is (Cloudflare). Initially, the data contained 53 attributes that have been reduced to 17 attributes with a class label (Action) that will be predicted in this study.

The CRISP-MD model was implemented in this study, and six phases have been followed. through the business understanding phase, the researcher explains the nature of the organization, the network structure, access structure, and the goal of using data mining in this scenario.

In the data understanding phase, a deep explanation about the data has been provided by the researcher, starting with the data collection process, data preparation, data integration and pre-processing, feature selection, feature transformation (one-hot encoding), feature scaling (data normalization) process, data integration and data clean -up process that has been applied to the dataset before starting the next phase. Then the prediction model phase started with providing a brief explanation about the six algorithms that have been used in the study (Naïve Bayes, SVM, KNN, Logistic Regression, Decision Tree, and Random Forest). also in the evaluation phase; different evaluation measures have been produced briefly such as (Accuracy, Precision, Recall, F1-Score, ROC AUC score, and AUC).

# Chapter IV: Data and Result

This chapter provides detailed information about the data, starting with the data analysis and insights into the attribute relationship using the Correlation matrix and the Heatmap. Then the result for each classifier model used in the study, along with the evaluation metrics comparison. Finally, a summary table includes all the results.

## 4.1 Data Analysis

This study section will focus on the data structure after collecting, integrating, pre-processing, and cleaning. The analysis starts with visualizing the data for a deep understanding of the attributes and finding the relationship between the attribute and the target class label for the study, the Action.

### 4.1.1 Data insights

After the data integration and pre-processing stage, the data analysis stage will start, which will help us to analyze and extract some valuable meaning from the data. We used the Power BI tool to visualize the data and find the relations between the attributes at this stage. For example, suppose we choose the data category. In that case, we will see 43 % from the data related to security service, 38 % from the data pertaining to Firewall services, and the rest for a system, users, firewall setting, and VPN.

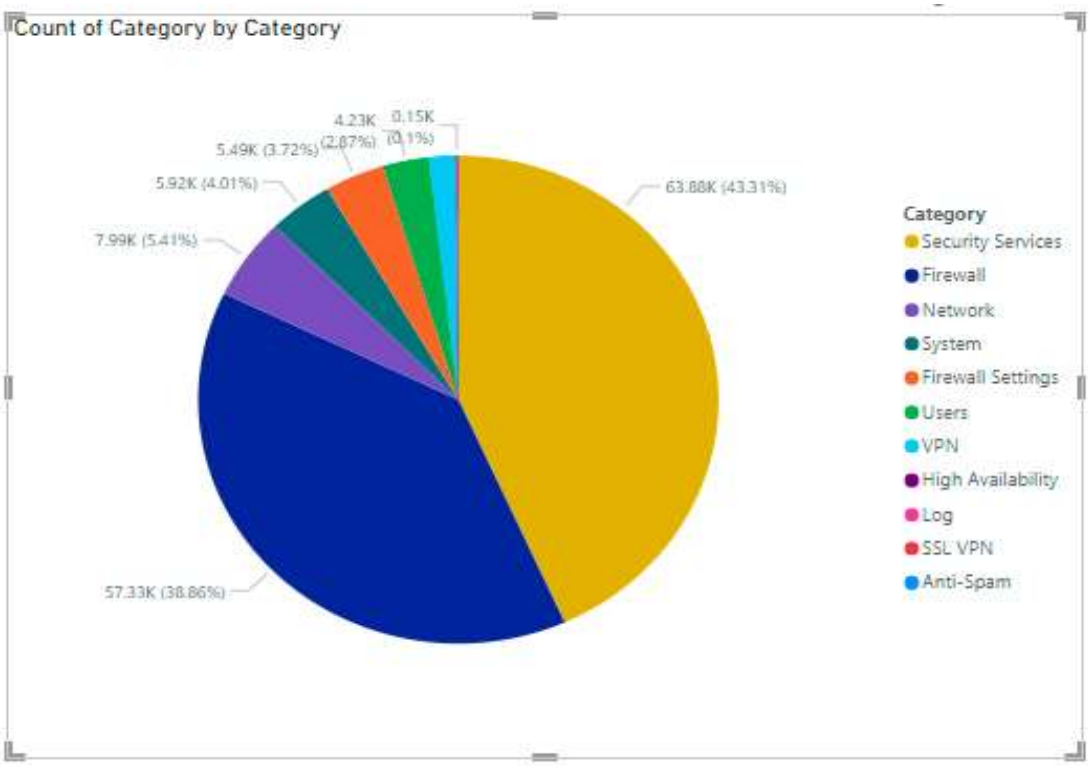



*Figure 24: Data Category*

Also, from the data, we can see that 66 % of the data traffic depends on TCP (Transmission Control Protocol ), 29 % of the traffic-related to UDP (User Datagram Protocol), and the rest for other protocols such as ICMP, IPv6-ICMP, and IPv6 Hop-to Hop protocol.

In the data set, we have six types of IP protocols that have standard numbers as per table 6. The protocol numbers are published and managed by the Internet Assigned Numbers Authority (IANA); these numbers are found in the IPv4 and IPv6 headers; it is essential because it is determined the layout of the data after the header and identifying the encapsulation protocol.

The Request for Comments or RFC is a formal document industrialized by the Internet Engineering Committee Task Force .this documents include technical information and organizational notes for the internet. These documents cover several networks, email, and computer aspects, including programs, protocols, concepts, and procedures.

| Hexadecimal | Protocol Number | Keyword | Protocol | References/RFC |
|---|---|---|---|---|
| **0x00** | 0 | HOPOPT | IPv6 Hop by Hop option | 8200 |
| **0x01** | 1 | ICMP | Internet Control Massaging Protocol | 792 |
| **0x06** | 6 | TCP | Transmission Control Protocol | 793 |
| **0x11** | 17 | UDP | User Datagram Protocol | 768 |
| **0x32** | 50 | ESP | Encapsulation Security Payload | 4303 |
| **0x3A** | 58 | IPv6-ICMP | ICMP for IPv6 | 4443 and 4884 |

*Table 6: List of IP protocol numbers*

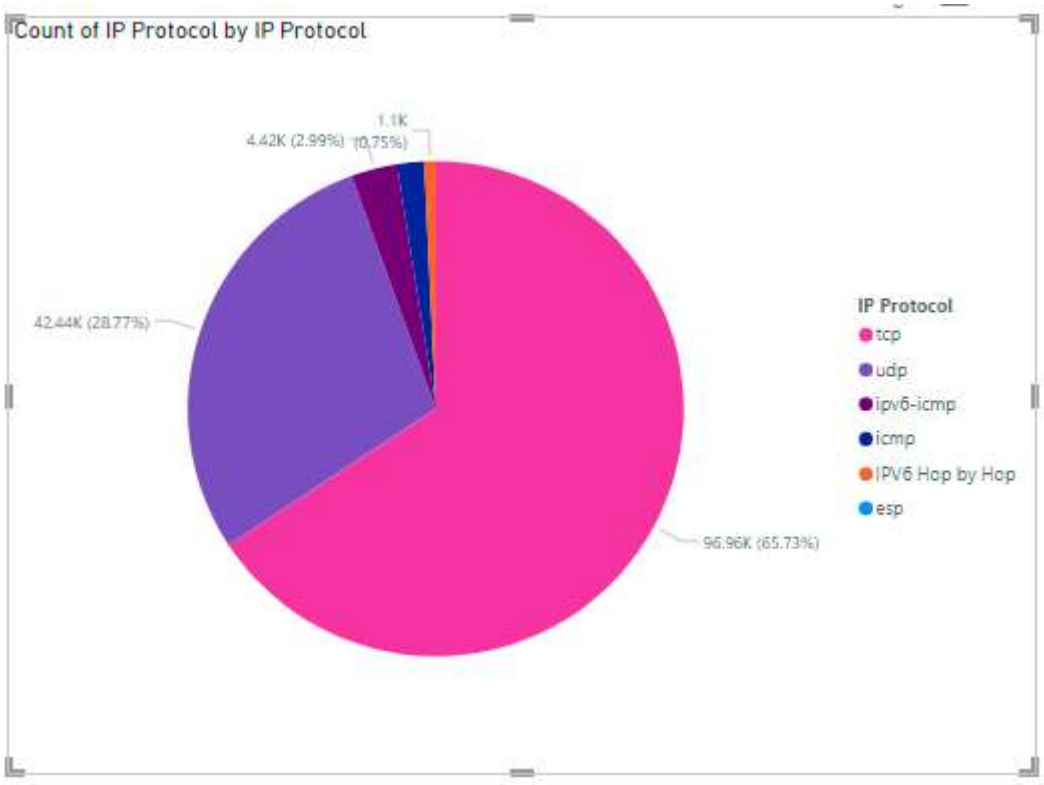



*Figure 25: IP Traffic Protocols*

The security system divides the data as per the priority into different categories based on the severity level from zero to seven. This priority has different names (Debug, Information, Notification, Warning, Error, Critical, Alert, and Emergency ); it helps the security

administrator determine logs messages nature if it is normal massage or critical and needs human interaction.

The logs level also allows the security engineers to filter the events as per the priority. The events with higher priority will pass, but those with low priority will drop. The loges massages can also be configured to be sent automatically by emails to the security concern to take action. These actions are based on the organization's business nature, so we can find some organizations with a very high level of security, such as banks, so the data drop level will be very high rather than other organizations such as the education sector. The security logs priority is demonstrated in the below table.

| Level | Name | Description |
|---|---|---|
| **0** | Emergency | The system is not responding, or it is unstable |
| **1** | Alert | It is used in the security logs, and Immediate action is required |
| **2** | Critical | One of the functions is affected |
| **3** | Error | There is an error, and one of the functions maybe will be affected |
| **4** | Warning | There could be affected on the function |
| **5** | Notification | There is some information about the normal events |
| **6** | Information | This event is used in the event log for recording the changes in the configuration files .also gives information about the system operation process |
| **7** | Debug | gave detailed information that will help in the debug purpose. |

*Table 7: priority level description*

Through the data analysis process, it was noticed that there is a relation between the attributes for example, in the data logs, there is an attribute named category, which indicates the data will belong to which category, it is related to system, Network, security setting, firewall setting or

Haigh Availability category. These categories are varied from one organization to another organization. There is a high relation between the category and the priority level. As per the below graph, there is a category Firewall setting connected with different types of priority (warning, notice, information, and alert), but in another category such as network, there is another priority which is (notice, alert, and information only ).as per the below figure.

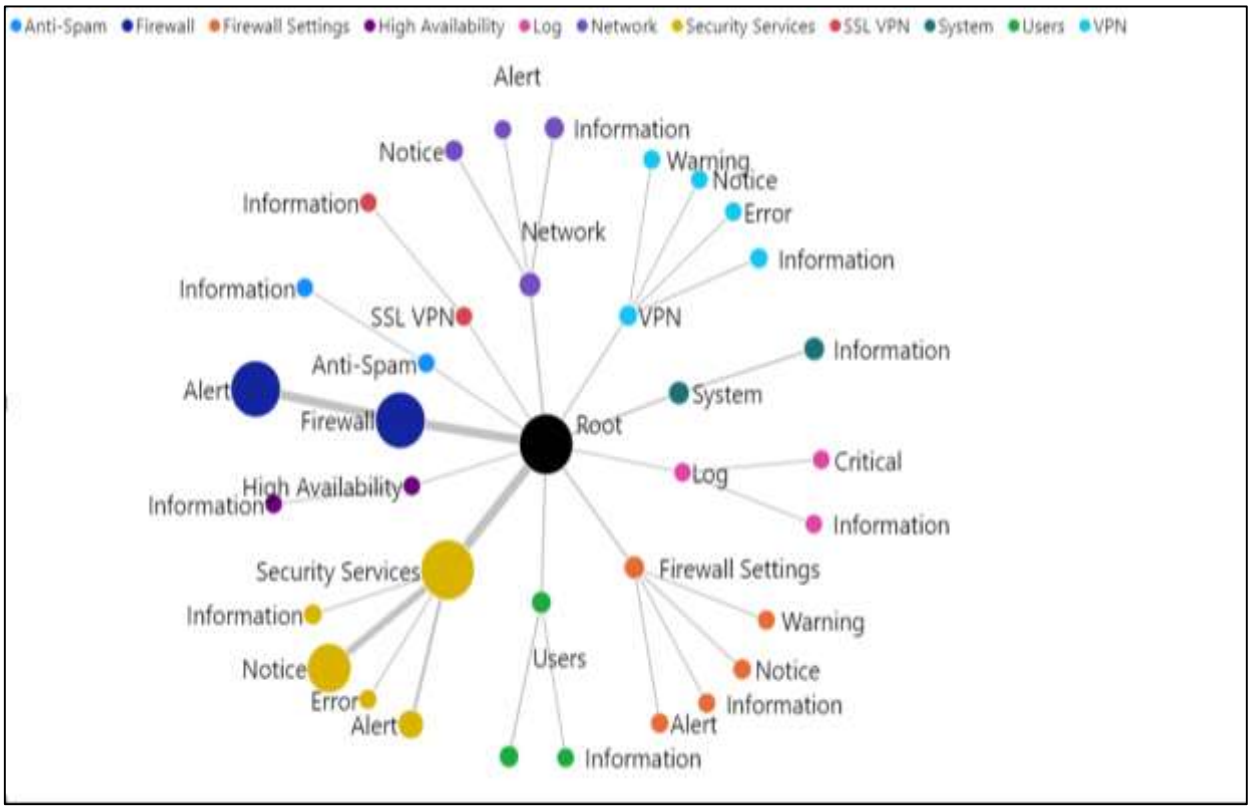



*Figure 26: Category, Priority Relationship*

### 4.1.2 Attributes Relationship

this section represents the relationship between the data attributes. Using the Correlation matrix and Heatmap, most of the variable shows negative of low correlation, thus there is no linear relationship between the variables. Therefore, the machine learning algorithms is carried out to solve the non-linear relation for specific target. As shown in the below heatmap figure and the correlation matrix table.

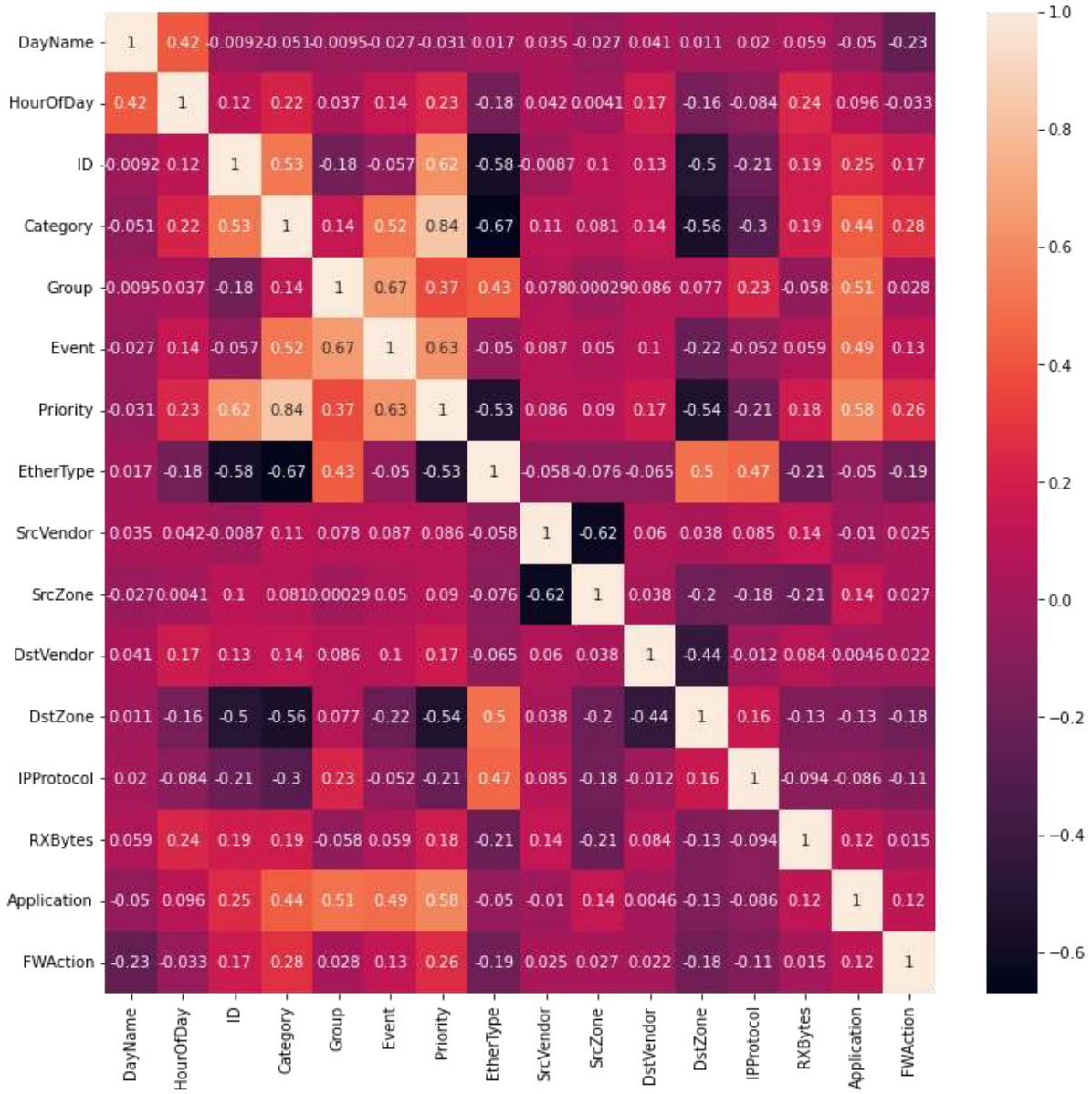



*Figure 27:Correlation Heatmap*

| | Day Name | HourOfDay | ID | Category | Group | Event | Priority | Ether Type | SrcVendor | SrcZone | DstVendor | DstZone | IP Protocol | RXBytes | Application | FWAction |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Day Name** | 1 | 0.421373 | -0.009245 | -0.051269 | -0.009537 | -0.026816 | -0.030951 | 0.016792 | 0.035427 | -0.027441 | 0.041377 | 0.010607 | 0.020156 | 0.059252 | -0.049836 | -0.234266 |
| **HourOfDay** | 0.421373 | 1 | 0.116517 | 0.21777 | 0.037407 | 0.143892 | 0.234764 | -0.177548 | 0.042166 | 0.00409 | 0.174653 | -0.161845 | -0.08412 | 0.235354 | 0.096309 | -0.032727 |
| **ID** | -0.009245 | 0.116517 | 1 | 0.527313 | -0.176666 | -0.056945 | 0.621008 | -0.580582 | -0.008735 | 0.103468 | 0.125722 | -0.498486 | -0.20892 | 0.186267 | 0.254371 | 0.170546 |
| **Category** | -0.051269 | 0.21777 | 0.527313 | 1 | 0.139168 | 0.524122 | 0.837418 | -0.669695 | 0.11088 | 0.081325 | 0.137055 | -0.556256 | -0.296251 | 0.188227 | 0.435324 | 0.277662 |
| **Group** | -0.009537 | 0.037407 | -0.176666 | 0.139168 | 1 | 0.666294 | 0.373003 | 0.428737 | 0.078359 | 0.000292 | 0.086159 | 0.076732 | 0.230125 | -0.057524 | 0.505531 | 0.028156 |
| **Event** | -0.026816 | 0.143892 | -0.056945 | 0.524122 | 0.666294 | 1 | 0.629051 | -0.04972 | 0.087383 | 0.049532 | 0.103431 | -0.223534 | -0.052383 | 0.05874 | 0.4897 | 0.127365 |
| **Priority** | -0.030951 | 0.234764 | 0.621008 | 0.837418 | 0.373003 | 0.629051 | 1 | -0.525771 | 0.086166 | 0.089949 | 0.170714 | -0.5359 | -0.214465 | 0.184303 | 0.579161 | 0.258399 |
| **Ether Type** | 0.016792 | -0.177548 | -0.580582 | -0.669695 | 0.428737 | -0.04972 | -0.525771 | 1 | -0.057873 | -0.076297 | -0.064945 | 0.498917 | 0.467916 | -0.212734 | -0.050393 | -0.191476 |
| **SrcVendor** | 0.035427 | 0.042166 | -0.008735 | 0.11088 | 0.078359 | 0.087383 | 0.086166 | -0.057873 | 1 | -0.618065 | 0.059888 | 0.03849 | 0.084958 | 0.139036 | -0.010377 | 0.024638 |
| **SrcZone** | -0.027441 | 0.00409 | 0.103468 | 0.081325 | 0.000292 | 0.049532 | 0.089949 | -0.076297 | -0.618065 | 1 | 0.037566 | -0.19812 | -0.17521 | -0.206888 | 0.139681 | 0.026668 |
| **DstVendor** | 0.041377 | 0.174653 | 0.125722 | 0.137055 | 0.086159 | 0.103431 | 0.170714 | -0.064945 | 0.059888 | 0.037566 | 1 | -0.440285 | -0.012171 | 0.084292 | 0.004633 | 0.021584 |
| **DstZone** | 0.010607 | -0.161845 | -0.498486 | -0.556256 | 0.076732 | -0.223534 | -0.5359 | 0.498917 | 0.03849 | -0.19812 | -0.440285 | 1 | 0.157834 | -0.125756 | -0.131364 | -0.183309 |
| **IP Protocol** | 0.020156 | -0.08412 | -0.20892 | -0.296251 | 0.230125 | -0.052383 | -0.214465 | 0.467916 | 0.084958 | -0.17521 | -0.012171 | 0.157834 | 1 | -0.093544 | -0.085801 | -0.112251 |
| **RXBytes** | 0.059252 | 0.235354 | 0.186267 | 0.188227 | -0.057524 | 0.05874 | 0.184303 | -0.212734 | 0.139036 | -0.206888 | 0.084292 | -0.125756 | -0.093544 | 1 | 0.117456 | 0.015313 |
| **Application** | -0.049836 | 0.096309 | 0.254371 | 0.435324 | 0.505531 | 0.4897 | 0.579161 | -0.050393 | -0.010377 | 0.139681 | 0.004633 | -0.131364 | -0.085801 | 0.117456 | 1 | 0.115451 |
| **FWAction** | -0.234266 | -0.032727 | 0.170546 | 0.277662 | 0.028156 | 0.127365 | 0.258399 | -0.191476 | 0.024638 | 0.026668 | 0.021584 | -0.183309 | -0.112251 | 0.015313 | 0.115451 | 1 |

*Table 8: Correlation matrix*

## 4.2 Machine Learning Performance

### 4.2.1 Naïve Bayes:

| Class Label | precision | recall | f1-score |
|---|---|---|---|
| **0** | 0.56 | 0.69 | 0.62 |
| **1** | 0.72 | 0.58 | 0.64 |

*Table 9: Naïve Bayes classification Report*

For the Naïve Bayes classification model report, we can notice that the data which does not have any type of malicious attack with the allowed action passing through the different layers of firewalls (1). It has a high precision than the data that will be dropped and consider having any type of attacks, whereas the recall shows the opposite result and F1- Score gives a similar approximate result for both actions.

| Accuracy | Precision | Recall | f1 Score | ROC AUC SCORE | AUC |
|---|---|---|---|---|---|
| **63.16%** | 71.69% | 58.47% | 64.41% | 63.92% | 77.4 % |

*Table 10: Naïve Bayes Metrics*

From the above evaluation metrics table, we notice that all the evaluations are similar, but the precision is a little high because it predicts that the uninfected data resulted in the precision score of 71.69%. The model did not predict the infected dropped data well, which resulted in reducing the accuracy to 63.16 % .further, the model predicted more uninfected data to be dropped, which reduced the recall and F1-score to 58.47% and 64.41%, respectively. The AUC measure shows better results, close to 100% that, indicates the model is able to differentiate between the infected and uninfected data. The ROC score indicates that the infected data allowed to pass is predicted higher, so the result shows 63.92%.

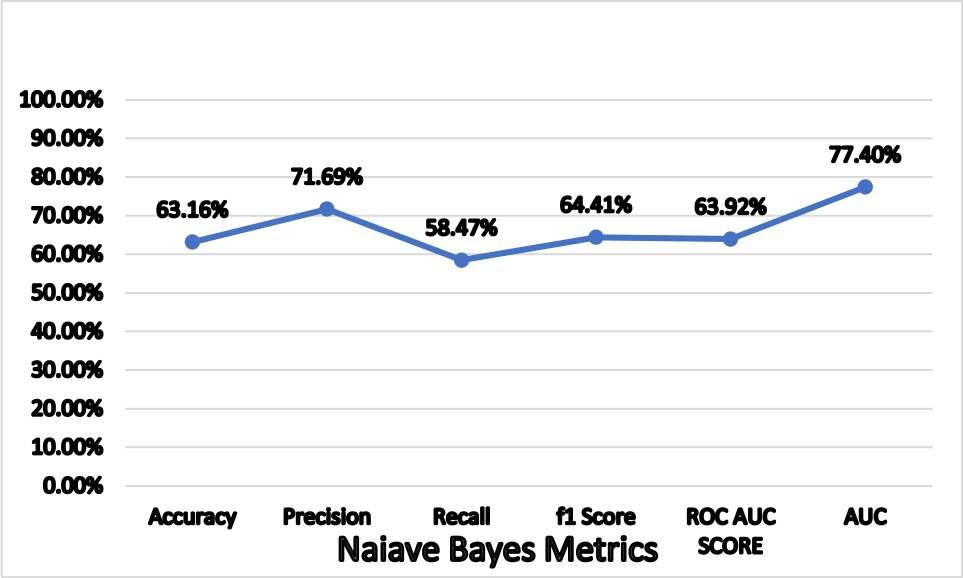



*Figure 28: Naïve Bayes Metrics*

The above is a visual representation of the evaluation matrix. The model showed a good AUC measure score and decent precision.

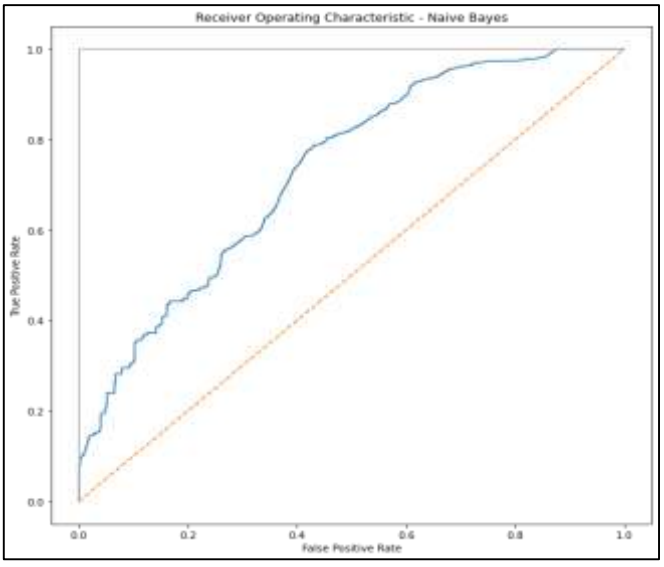



*Figure 29: TPR/FPR for Naïve Bayes*

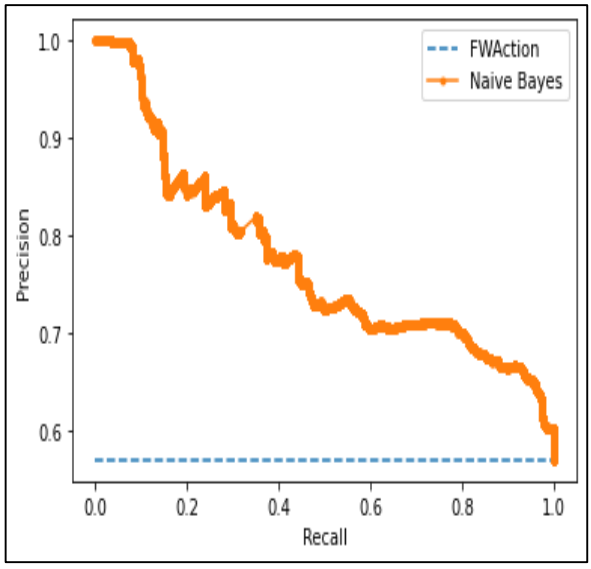



*Figure 30:Recall/ Precision for Naïve Bayes*

The above graphs show the relationship between the True Positive rate and the false positive rate, Precision, and recall with curve lines. The curve indicates the model is performing slightly well because the under curve area is averagely higher than the above curve area.

### 4.2.2 Support-Victor Machine

| Class Label | precision | recall | f1-score |
|---|---|---|---|
| **0** | 0.78 | 0.71 | 0.75 |
| **1** | 0.80 | 0.85 | 0.82 |

*Table 11: SVM Classification Report*

For the Support Vector Machine classifier model report, it is noticed that the recall value is high, which means the clean data will pass. Also, from the report, it is noticed that the precision value is high, which means that the infected data will be dropped. This is also seen by the high value of the F1-Score that the classifier model prediction is high.

| Accuracy | Precision | Recall | f1 Score | ROC AUC SCORE | AUC |
|---|---|---|---|---|---|
| **79.14%** | 79.65% | 85.16% | 82.31% | 78.15% | 90 % |

*Table 12: SVM Metrics*

From the above evaluation metric table, it is noticed that the recall measure 85.16% is high comparing with precision 79.65%, which means that the classifier prediction is the clean data will pass more than the infected data will be stopped. The accuracy of 79.14% means the clean data will pass compared with the whole data. The AUC measure shows a better result that indicates the model is able to differentiate between the clean and the data with suspected threats with 90%. The ROC score also indicates that the infected data allowed to pass is less.

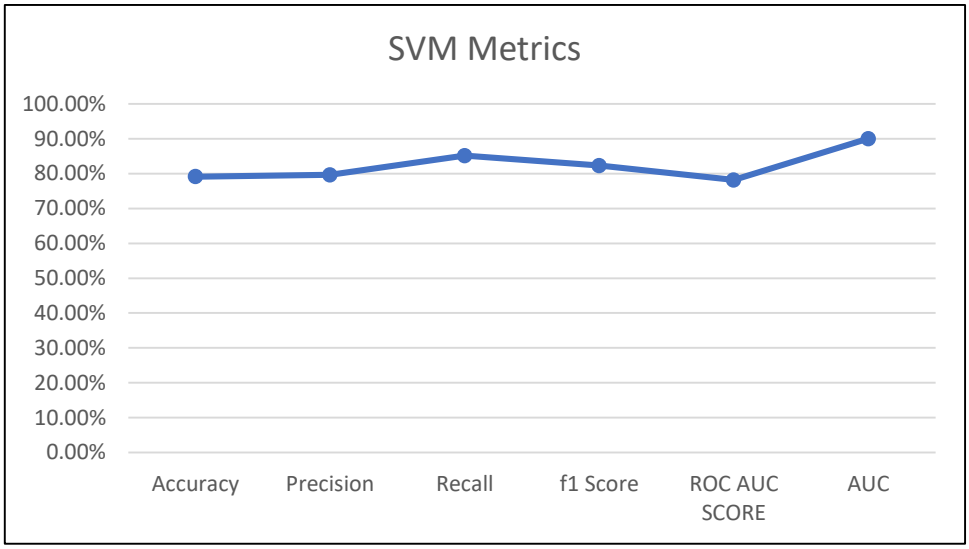



*Figure 31: Support-Victor Machine Metrics*

The above is a visual representation of the evaluation matrix. The model showed a good AUC measure score and adequate precision.

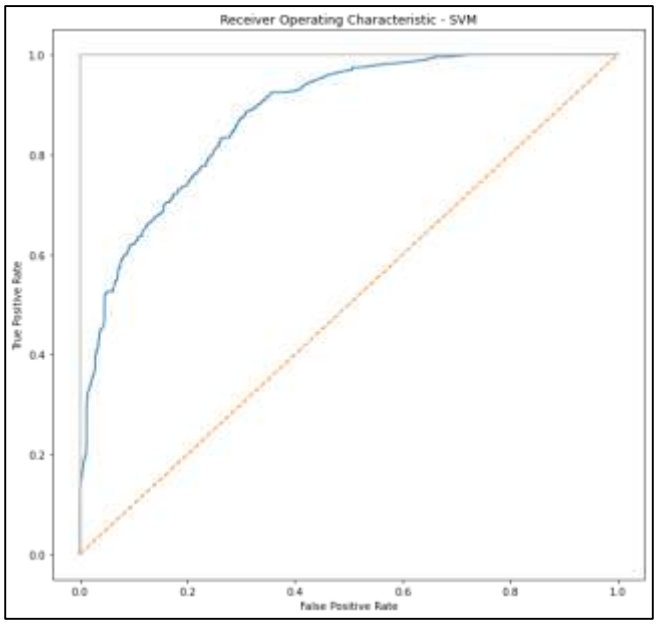



*Figure 32: TPR/FPR for SVM*

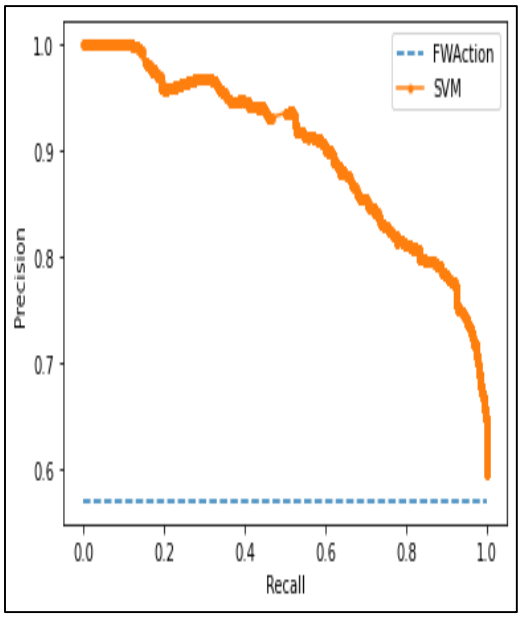



*Figure 33: Recall/ Precision for SVM*

The above graphs show the relationship between the True Positive rate and the false positive rate, Precision, and recall with curve lines. The curve indicates the model is performing well because the under curve area is higher than the above curve area and is closer to the top edges.

### 4.2.3 K-Nearest Neighbours

Initially, the best value of K needs to be determined; the minim error was found to be 8.4% at the value of K =16.

| Class Label | precision | recall | f1-score |
|---|---|---|---|
| **0** | 0.91 | 0.89 | 0.90 |
| **1** | 0.92 | 0.93 | 0.93 |

*Table 13: KNN Classification Report*

The KNN classifier model report shows that the clean data which does not have any kind of threats can pass directly without dropping actions from the security systems; this appears from the high value of precision and recall, which are near in the value. That means that the classifier model predicts the infected data to be dropped very near to allow the clean data to pass without action. Since the precision and recall are high in value, then the model gives a high F1 score also.

| Accuracy | Precision | Recall | f1 Score | ROC AUC SCORE | AUC |
|---|---|---|---|---|---|
| **91.47%** | 91.84% | 93.33% | 92.58% | 91.17% | 96.7% |

*Table 14: KNN Metrics*

From the above evaluation metrics, it is noticed that all the evaluation measures are similar in result with high value. The recall value of 93.33% shows that the model predicts all the clean data will pass and not stop. Also, as per the precision, the model predicts the infected data that have threats will be dropped and the clean data that should be passed with 91.47% Accuracy. The AUC measure shows better results, close to 100%, which indicates the model is able to differentiate between the infected and the clean data with 96.7%. the ROC score shows that the infected data not allowed to pass with a very high prediction of 91.17%

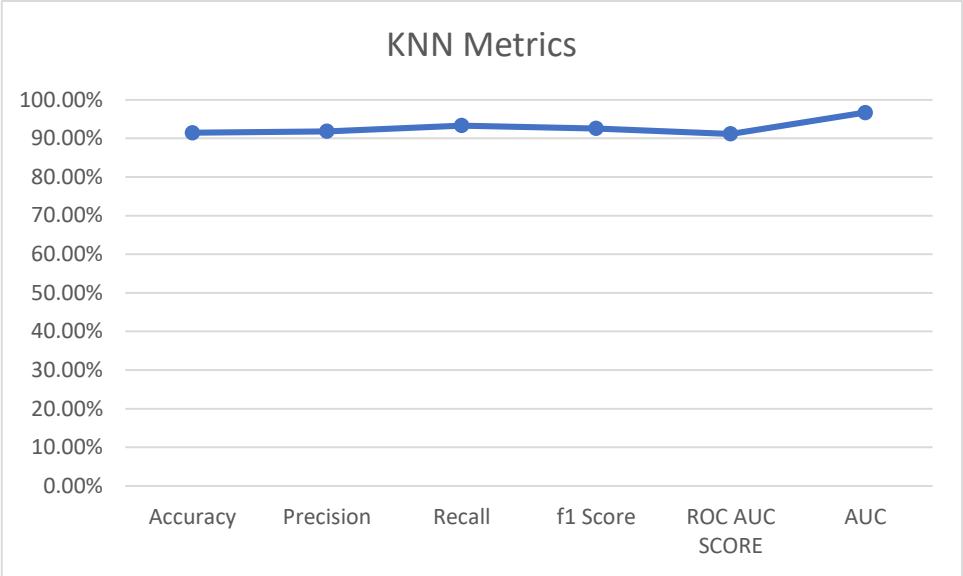



*Figure 34: KNN Metrics*

The above is a visual representation of the evaluation matrix. The model showed a very good AUC measure score and decent precision.

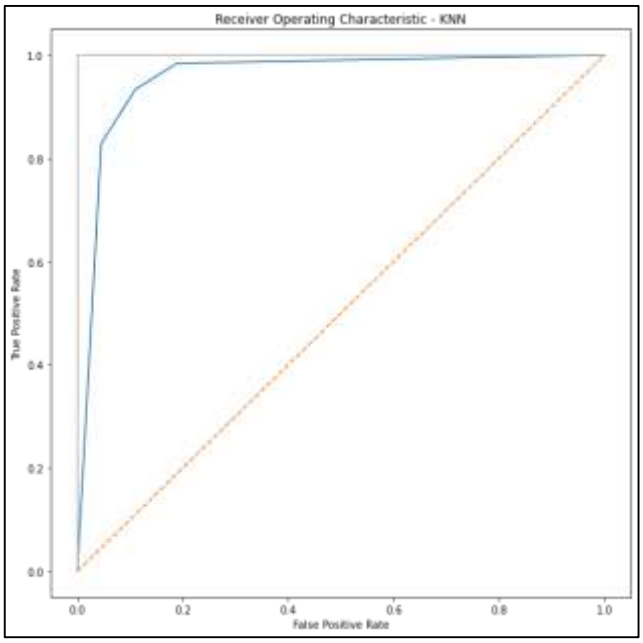



*Figure 35 :TPR/FPR for KNN*

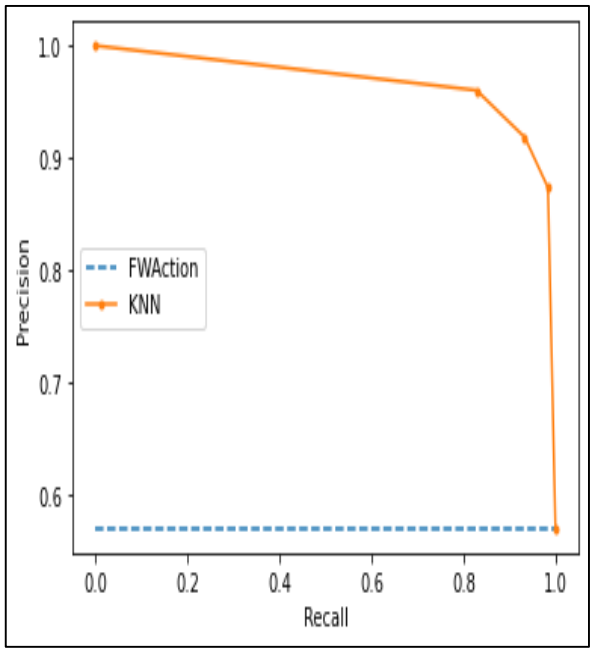



*Figure 36: Recall/ Precision for KNN*

The above graphs show the relationship between the True Positive rate is very high compared with the false positive rate. And the relation between precision and recall indicates that the

model is performing very well because the under curve area is higher than the above curve area and is next to the top edges of the graph.

### 4.2.4 Logistic Regression

| Class Label | precision | recall | f1-score |
|---|---|---|---|
| **0** | 0.78 | 0.71 | 0.74 |
| **1** | 0.79 | 0.85 | 0.82 |

*Table 15: Logistic Regression Classification Report*

For the Logistic Regression Machine learning classifier model report, it is noticed that the recall value of class label 1 is high, which means the clean data will pass compared with the value of precision that is slightly less, which means the infected data will be dropped but with less chance than the uninfected data to pass. This is also noticed from the F1-Score that indicates the classifier prediction is high.

| Accuracy | Precision | Recall | f1 Score | ROC AUC SCORE | AUC |
|---|---|---|---|---|---|
| **78.87%** | 79.40% | 84.99% | 82.10% | 77.88% | 90.1% |

*Table 16: Logistic Regression Metrics*

From the above evaluation metrics, it is noticed that recall measure is high compared with the precision and accuracy. The recall value is 84.99% shows that the model predicts the uninfected data to pass without stopping. But as per the precision, the model predicts the infected data that may have a threat will be dropped with the value of 79.40%. The AUC measure shows better results which indicates the model is able to differentiate between the infected and the clean data with 96.7%. the ROC score shows that the infected data is not allowed to pass with a little prediction of 77.88%

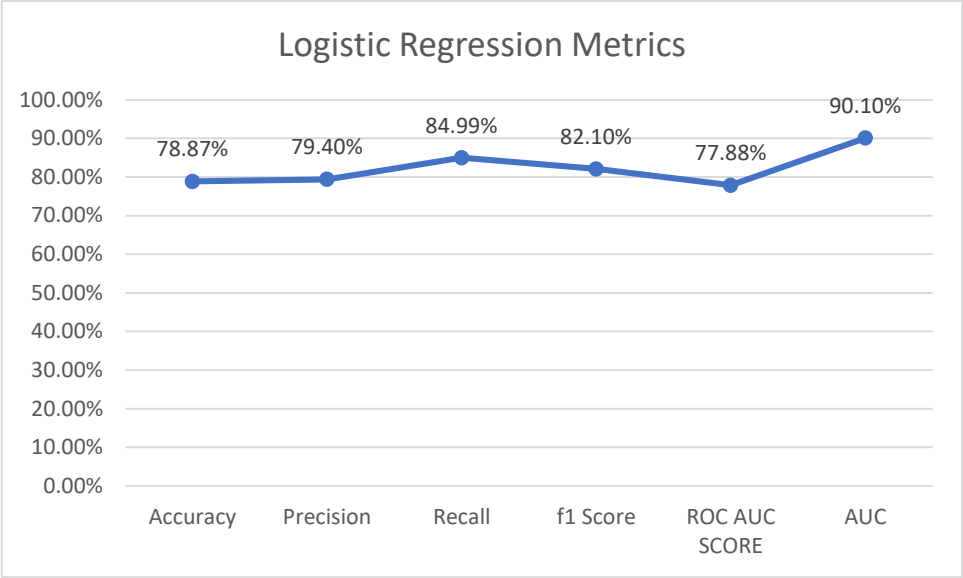



*Figure 37: Logistic Regression Metrics*

The above is a visual representation of the evaluation matrix. The model showed a very good AUC measure score and decent precision.

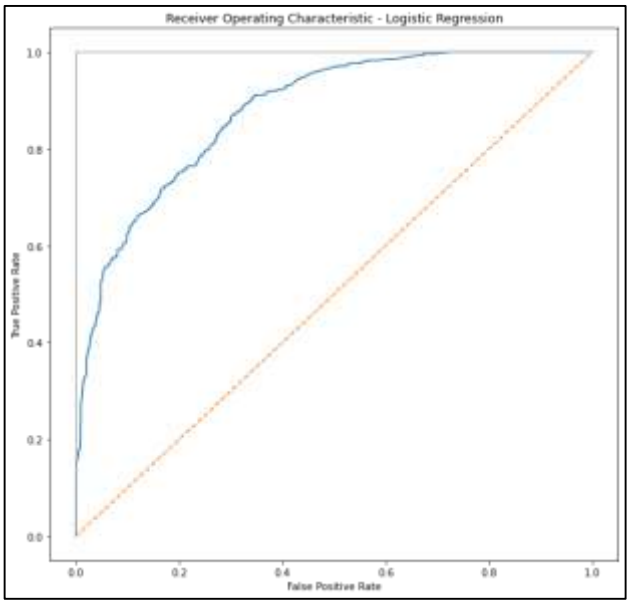



*Figure 38: TPR/FPR for Logistic Regression*

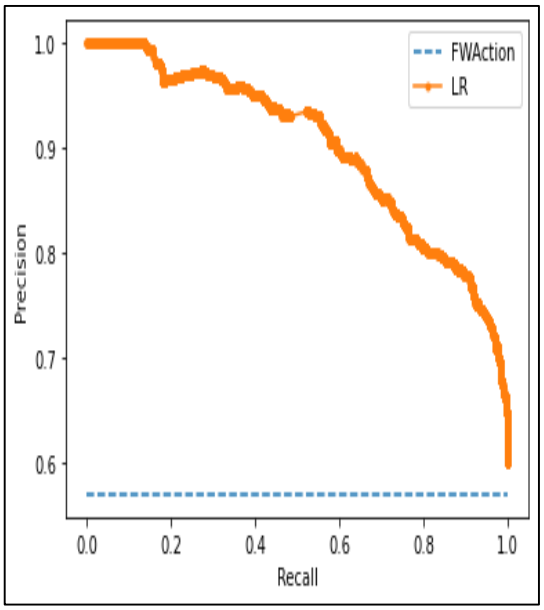



*Figure 39: Recall/ Precision for Logistic Regression*

The above graphs show the relationship between the True Positive rate is very high compared with the false positive rate. And the relation between Precision and recall indicates that the model is performing very well because the under-curve area is higher than the above curve area.

### 4.2.5 Decision Tree

As mentioned before, the data has been collected from three different resources (Cloudflare as a cloud firewall, Sonic Wall, and Fortinet firewall). Each firewall has its own rules and security configuration, which cannot be copied from one system to another. The number of rules are more than 200. But by using the decision tree algorithm, the number of rules has been optimized to 14 rules only, which means using machine learning will reduce the effort and time required for configuring and implementing the security rules. The below figure shows the visual decision tree .

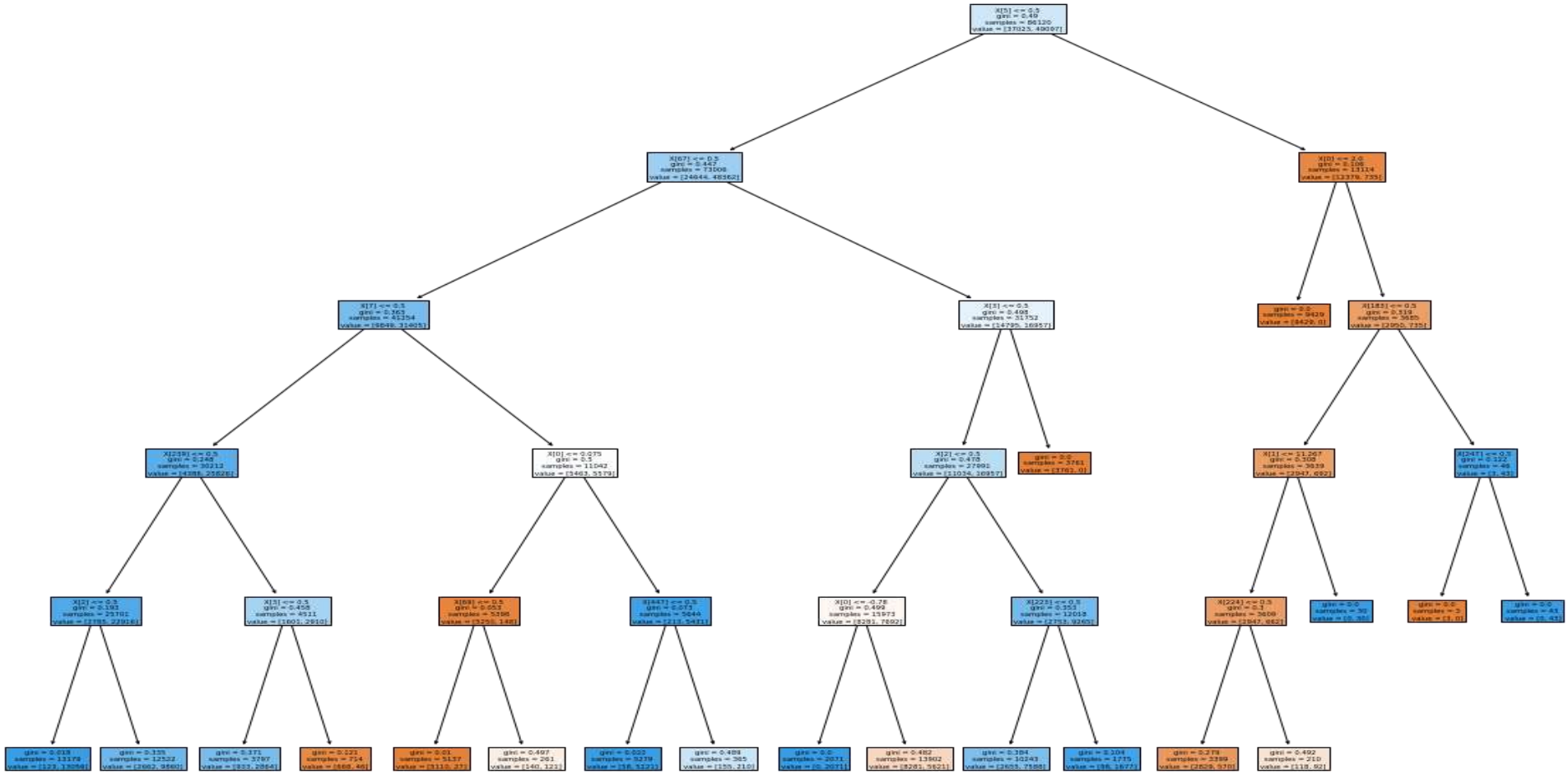


Figure 40: Decision Tree

| Class Label | precision | recall | f1-score |
|---|---|---|---|
| **0** | 0.83 | 0.82 | 0.83 |
| **1** | 0.87 | 0.87 | 0.87 |

*Table 17: Decision Tree Classification Report*

For the Decision Tree Machine learning classifier model report, it is noticed that the recall, precision, and F1-Score values are almost the same, which means the clean data will pass without any interruption from the security system, and the infected data will be dropped.

| Accuracy | Precision | Recall | f1 Score | ROC AUC SCORE | AUC |
|---|---|---|---|---|---|
| **85.03%** | 86.80% | 86.98% | 86.89% | 84.72% | 94.00% |

*Table 18: Decision Tree Metrics*

From the above evaluation metrics, it is noticed that recall, precision, F1-score, and accuracy are almost the same in value 86.80% with a little difference which means that the classifier model can predict the clean data will pass and the infected data will drop. The AUC measure shows better results which indicates the model can differentiate between the infected and the clean data with 94.00%. The ROC score shows that the infected data is not allowed to pass with a high prediction of 84.72%.

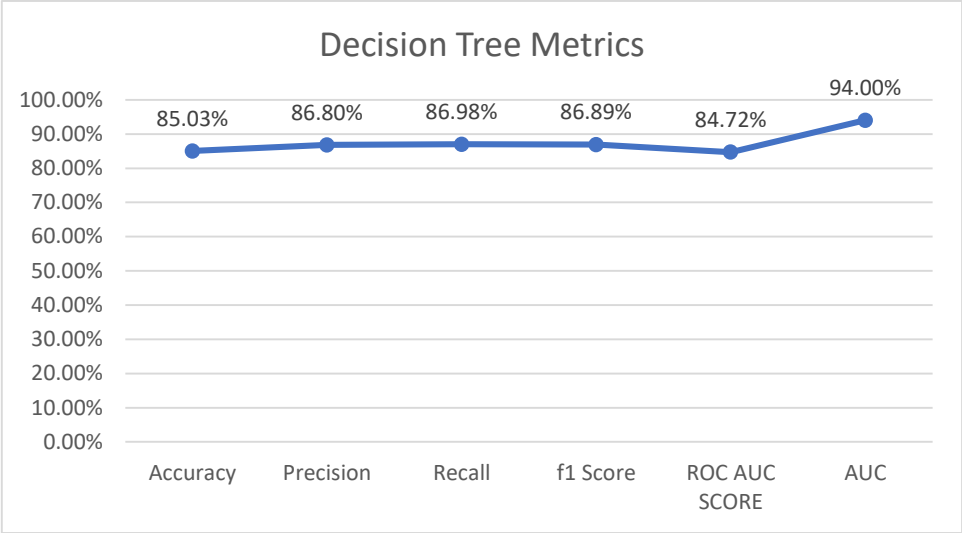



*Figure 41: Decision Tree Metrics*

The above is a visual representation of the evaluation matrix. The model showed a very good AUC measure score and decent precision.

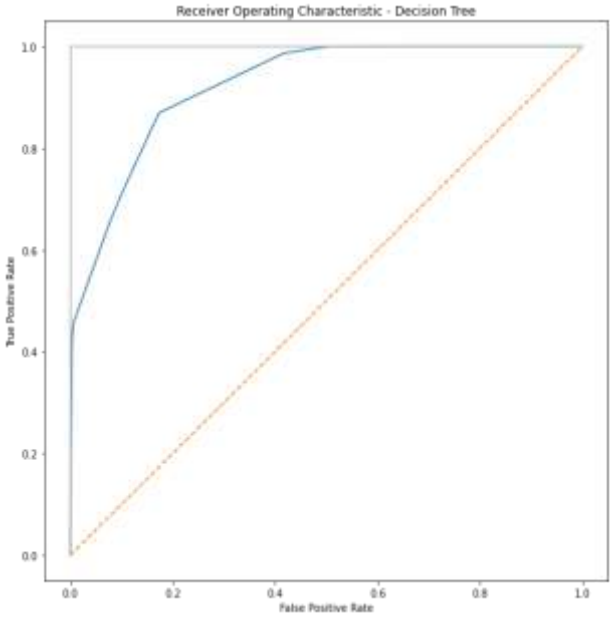



*Figure 42: TPR/FPR for Decision Tree*

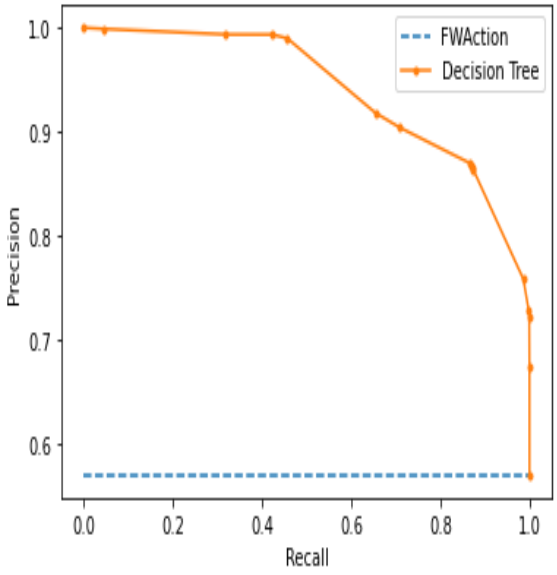



*Figure 43: Recall/ Precision for Decision Tree*

The above graphs show the relationship between the True Positive rate is very high compared with the false positive rate. And the relation between Precision and recall indicates that the model is performing very well because the under-curve area is higher than the above curve area.

### 4.2.6 Random Forest Model

| Class Label | precision | recall | f1-score |
|---|---|---|---|
| **0** | 0.93 | 0.88 | 0.91 |
| **1** | 0.91 | 0.95 | 0.93 |

*Table 19: Random Forest Classification Report*

For the Random Forest Machine learning classifier model report, it is noticed that the recall value is high, which means the clean data will pass compared with the value of precision that slightly less, which means the infected data will be dropped but with less chance than the uninfected data to pass. This is also noticed from the F1-Score that indicates the classifier prediction is high.

| Accuracy | Precision | Recall | f1 Score | ROC AUC SCORE | AUC |
|---|---|---|---|---|---|
| **92.24%** | 91.48% | 95.26% | 93.33% | 91.75% | 98.20% |

*Table 20: Random Forest Metrics*

From the above evaluation metrics, it is noticed that recall measure is high compared with the precision and accuracy. The recall value is 95.26% shows that the model predicts the uninfected data to pass without stopping. But as per the precision, the model predicts the infected data that may have a threat will be dropped with a value of 91.48%. The AUC measure shows better results which indicates the model is able to differentiate between the infected and the clean data with 98.20%. the ROC score shows that the infected data is not allowed to pass with a high prediction of 91.75%

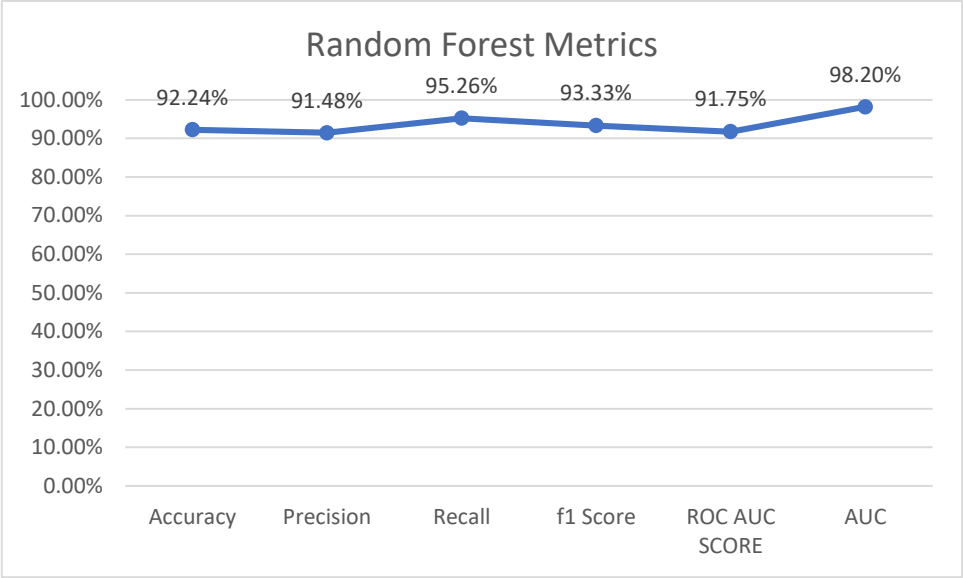



*Figure 44: Random Forest Metrics*

The above is a visual representation of the evaluation matrix. The model showed a very good AUC measure score and decent precision.

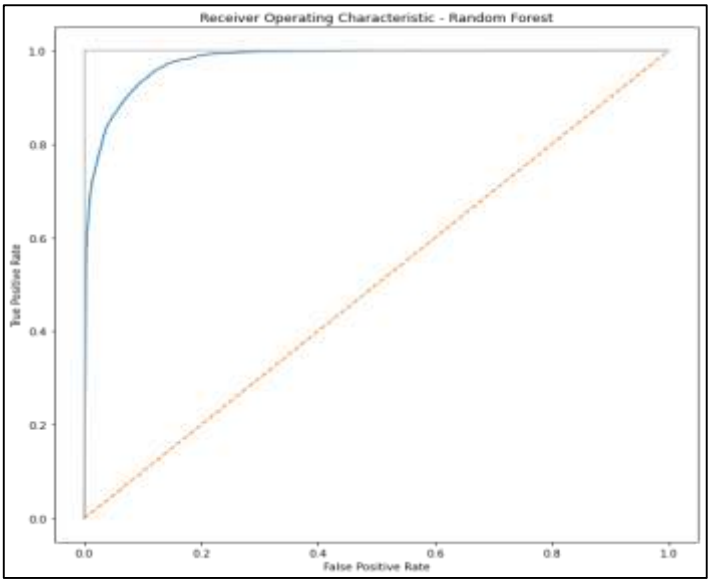



*Figure 45: TPR/FPR for Random Forest*

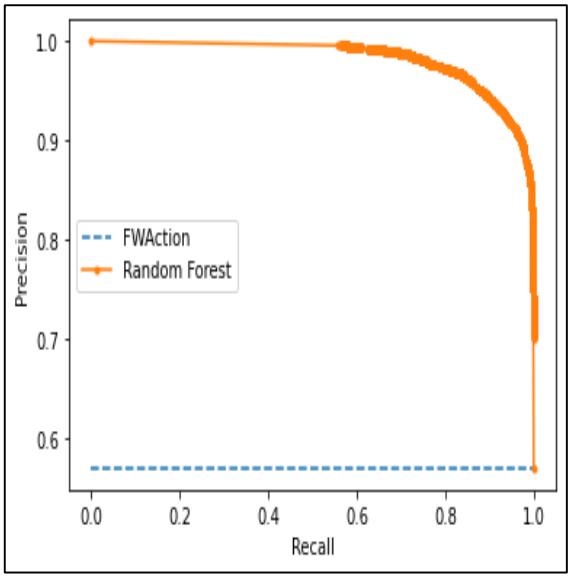



*Figure 46: Recall/ Precision for Random Forest*

The above graphs show the relationship between the True Positive rate is very high compared with the false positive rate. And the relation between Precision and recall indicates that the model is performing very well because the under-curve area is higher than the above curve area.

## 4.3 Machine Learning Deployment Summary

| Evaluation Metrics | Naïve Bayes | SVM | KNN | Logistic Regression | Decision Tree | Random Forest |
|---|---|---|---|---|---|---|
| **Accuracy** | 63.16% | 79.14% | 91.47% | 78.87% | 85.03% | 92.24% |
| **Precision** | 71.69% | 79.65% | 91.84% | 79.40% | 86.80% | 91.48% |
| **Recall** | 58.47% | 85.16% | 93.33% | 84.99% | 86.98% | 95.26% |
| **F1- Score** | 64.41% | 82.31% | 92.58% | 82.10% | 86.89% | 93.33% |
| **ROC AUC SCORE** | 63.92% | 78.15% | 91.17% | 77.88% | 84.72% | 91.75% |
| **AUC** | 77.40% | 90.00% | 96.70% | 90.10% | 94.00% | 98.20% |

*Table 21: Machine Learning Classification Summary*

The above table provides a summary of the machine learning classification models and their evaluation matrices.

## 4.4 Chapter IV: Summary

Power BI tool was used to visualize the data and find the relation between the attributes that will give a clear picture about the dataset, such as what is the levels of data priority (Debug, Information, Notification, Warning, Error, Critical, Alert, and Emergency), the connection protocol types (TCP, UDP or ICMP). Then the correlation matrix and the heatmap have been used to illustrate the relationship between the variables.

The machine learning performance classifiers models have been run by using Anaconda Navigator, JupyterLab to run the Python code. This code is a custom code that has been customized to run and evaluate six models (Naïve Bayes, SVM, KNN, Logistic Regression, Decision Tree, and Random Forest) and evaluate the result by using different evaluation metrics (Accuracy, Precision, Recall, F1-Score, ROC AUC Score and ACU).

# Chapter V: Conclusion and Future Work

The conclusion chapter provides the research summary, Concluding statements, Contributions of the research study, practical implications, the limitations, and finally, the future work recommendations.

## 5.1 Summary

The aim of the research study primarily focused on providing a framework that aids the educational institutions' information technology security teams to automate the enhancement of the security threats in the organization. The data mining technique enables the prediction of the threats by using models that are trained from the massive log file data generated by standard security systems. The research used data records of 123029 for training the model. The tools used during data collection, preparation, and analysis phases included FortiAnalyzer, MS EXCEL, MS Access, Power BI Desktop, and Python. The models were developed on six machine learning algorithms: KNN, Decision Tree, Naïve Bayes, Logistic Regression, Random Forest, and SVM. The models were evaluated based on six metrics are Accuracy, recall, F1 Score, Precision, AUC, and ROC AUC score.

## 5.2 Conclusion

Network security is one of the major concerns for any organization, and a lot of effort is required by the information technology personnel to protect the network connections by creating security rules and configuring the security systems. The risk of threats like WannaCry, ransomware, and other malicious activities can sabotage the organization. The research aims to provide a reliable framework that could assist the information technology personnel in protecting the network intelligently with less human interaction.

The class label characteristics that allow or drop the data packets of the network traffic can be seriously harmful if the model does not detect the malicious packets. So, the precision of the models should be high for evaluation. The KNN and Random Forest models provided high precision of 91.84% and 91.48%, respectively. The KNN performed accuracy 91.47%, recall of 93.33%, F1 Score of 92.58%, AUC of 96.70%, and ROC AUC Score of 91.17%. The random forest model performed in terms of accuracy 92.24%, recall of 95.26%, F1 Score of 93.33%, AUC of 98.20%, and ROC AUC Score of 91.75%.

The future enhancement for the research study would be to predict the unknown new threats. This can include the basis of more attributes relation between the attributes and data collected from different organizations to provide a more precise, reliable, and universal framework for implementation.

## 5.3 Contributions

Since most of the research papers and studies depended on the universal NSL-KDD that basically used certain attributes, this study used a real dataset log related to higher education organizations.

The proposed framework can help for the following

- Reduce the human effort for configuring and implementing the access data rules
- Establish a framework for security baseline for the similar organization.
- Use machine learning to discover the potential threats and take a quick action

## 5.4 Practical implication

Since the dataset which has been used in the study was a real dataset, there were driftnet attributes used in the investigation process such as malware, DDoS attack, App follows, Content filter, Application control, and spoof attack. The workflow can be enhanced and generalized to be used in different organizations. This workflow will also help to set baseline security for the organizations then additional rules and policies can be applied. Finally, by applying this workflow then, the human interaction with the security system will be reduced, and immediate actions will be taken automatically, which will reduce the security breach chances.

## 5.5 Limitations

The framework has been built based on the logs file that has been extracted from the security system that is related to one of the higher education organizations. The security action has been taken based on the set of rules and policies that has been set by the organization security engineers, which may be considered as a custom rule. The education sector has a special requirement that allows the organization to be secured, but at the same time, students can try and practice different technologies. These rules may be will not fit for other types of business such as the health care sector or many banks sector which need more roles and regulations.

## 5.6 Future work

- To enhance the framework so it can predict the new threats that are unknown.
- To train the model to predict the threats based on data collected from several organizations having different configured rules.
- To apply more machine learning algorithms for choosing the best performance.
- To apply the framework in an organization providing notifications and alarms for the suspected malicious data.

## References: -

[1] Cisco Annual Internet Report (2018–2023). (2020). Cisco Systems. https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html.

[2] Chen, Bo, et al. "Cybersecurity of Wide Area Monitoring, Protection, and Control Systems for HVDC Applications." IEEE Transactions on Power Systems 36.1 (2020): 592-602.

[3] Sultana, Nasrin, et al. "Survey on SDN based network intrusion detection system using machine learning approaches." Peer-to-Peer Networking and Applications 12.2 (2019): 493-501.

[4] Elsayed, M.S., Le-Khac, N.A., Dev, S. and Jurcut, A.D., 2020, August. Ddosnet: A deep-learning model for detecting network attacks. In *2020 IEEE 21st International Symposium on" A World of Wireless, Mobile and Multimedia Networks"(WoWMoM)* (pp. 391-396). IEEE.

[5] Tulabandhula, Theja, and Cynthia Rudin. "On combining machine learning with decision making." Machine learning 97.1-2 (2014): 33-64.

[6] Belavagi, Manjula C., and Balachandra Muniyal. "Performance evaluation of supervised machine learning algorithms for intrusion detection." Procedia Computer Science 89 (2016): 117-123.

[7] M. Haenlein, A. Kaplan. (2019). A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence.

[8] Shah, Huma, et al. "Can machines talk? Comparison of Eliza with modern dialogue systems." Computers in Human Behavior 58 (2016): 278-295.

[9] El Naqa, Issam, and Martin J. Murphy. "What is machine learning?." machine learning in radiation oncology. Springer, Cham, 2015. 3-11.

[10] van den Bosch, K. and Bronkhorst, A., 2018. Human-AI cooperation to benefit military decision-making. NATO.

[11] Mokhtarzadeh, Nima Garousi, et al. "Classification of inter-organizational knowledge mechanisms and their effects on networking capability: a multi-layer decision-making approach." Journal of Knowledge Management (2021).

[12] Wang, Ping, et al. "An efficient flow control approach for SDN-based network threat detection and migration using support vector machine." 2016 IEEE 13th international conference on e-business engineering (ICEBE). IEEE, 2016.

[13] Viswam, Anju, and Gopu Darsan. "An efficient bitcoin fraud detection in social media networks." 2017 International Conference on Circuit, Power and Computing Technologies (ICCPCT). IEEE, 2017.

[14] Kettani, Houssain, and Polly Wainwright. "On the top threats to cyber systems." 2019 IEEE 2nd International Conference on Information and Computer Technologies (ICICT). IEEE, 2019.

[15] Ncubukezi, Tabisa, L. Mwansa, and F. Rosaries. "A Proposed: Integration of the Monte Carlo model and the Bayes network to Propose Cyber Security Risk Assessment Tool for Small and Medium Enterprises in South Africa." IJISRT 3.18 (2020): 152-155.

[16] Devarakonda, Nagaraju, et al. "Integrated Bayes network and hidden Markov model for host-based IDS." International Journal of Computer Applications 41.20 (2012).

[17] Alkasassbeh, Mouhammad, and Mohammad Almseidin. "Machine learning methods for network intrusion detection." arXiv preprint arXiv:1809.02610 (2018).

[18] Alsughayyir, B., Qamar, A.M. and Khan, R., 2019, April. Developing a network attack detection system using deep learning. In 2019 International Conference on Computer and Information Sciences (ICCIS) (pp. 1-5). IEEE.

[19] Rai, Kajal, M. Syamala Devi, and Ajay Guleria. "Decision tree-based algorithm for intrusion detection." International Journal of Advanced Networking and Applications 7.4 (2016): 2828.

[20] Jin, Yunhu, et al. "The model of network security situation assessment based on random forest." 2016 7th IEEE international conference on software engineering and service science (ICSESS). IEEE, 2016.

[21] Negandhi, Prashil, Yash Trivedi, and Ramchandra Mangrulkar. "Intrusion detection system using random forest on the NSL-KDD dataset." Emerging Research in Computing, Information, Communication, and Applications. Springer, Singapore, 2019. 519-531.

[22] Farnaaz, Nabila, and M. A. Jabbar. "Random forest modeling for network intrusion detection system." Procedia Computer Science 89 (2016): 213-217.

[23] Lin, W.H., Lin, H.C., Wang, P., Wu, B.H. and Tsai, J.Y., 2018, April. Using convolutional neural networks to network intrusion detection for cyber threats. In 2018 IEEE International Conference on Applied System Invention (ICASI) (pp. 1107-1110). IEEE.

[24] Tang, Ying, and Mohamed Elhoseny. "Computer network security evaluation simulation model based on neural network." Journal of Intelligent & Fuzzy Systems 37.3 (2019): 3197-3204.

[25] Vinayakumar, R., K. P. Soman, and Prabaharan Poornachandran. "Applying convolutional neural network for network intrusion detection." 2017 International Conference on Advances in Computing, Communications, and Informatics (ICACCI). IEEE, 2017.

[26] Apruzzese, G., Colajanni, M., Ferretti, L., Guido, A. and Marchetti, M., 2018, May. On the effectiveness of machine and deep learning for cyber security. In 2018 10th international conference on cyber-Conflict (CyCon) (pp. 371-390). IEEE.

[27] Singla, A., Bertino, E., and Verma, D., 2020, October. Preparing network intrusion detection deep learning models with minimal data using adversarial domain adaptation. In Proceedings of the 15th ACM Asia Conference on Computer and Communications Security (pp. 127-140).

[28] Amrollahi, Mahdi, et al. "Enhancing network security via machine learning: opportunities and challenges." Handbook of big data privacy (2020): 165-189.

[29] Liu, Liu, et al. "Automatic malware classification and new malware detection using machine learning." Frontiers of Information Technology & Electronic Engineering 18.9 (2017): 1336-1347.

[30] Shafiq, Muhammad, et al. "Network traffic classification techniques and comparative analysis using machine learning algorithms." 2016 2nd IEEE International Conference on Computer and Communications (ICCC). IEEE, 2016.

[31] Kunang, Y.N., Nurmaini, S., Stiawan, D. and Suprapto, B.Y., 2021. Attack classification of an intrusion detection system using deep learning and hyperparameter optimization. Journal of Information Security and Applications, 58, p.102804.

[32] Yin, C., Zhu, Y., Fei, J., & He, X. (2017). A Deep Learning Approach for Intrusion Detection Using Recurrent Neural Networks. IEEE Access, 5, 21954–21961. doi:10.1109/access.2017.2762418

[33] B. Kolosnjaji, G. Eraisha, G. Webster, A. Zarras, and C. Eckert, "Empowering convolutional networks for malware classification and analysis," in Proc. Int. Joint Conf. Neural Netw. (IJCNN), 2017, pp. 3838–3845.

[34] Y. Yu, J. Long, and Z. Cai, "Network intrusion detection through stacking dilated convolutional autoencoders," Secure. Commun. Netw., vol. 2, no. 3, pp. 1–10, 2017.

[35] Xiao, Y., Xing, C., Zhang, T., & Zhao, Z. (2019). An intrusion detection model based on feature reduction and convolutional neural networks. IEEE Access, 7, 42210-42219

[36] Vinayakumar, R., Alazab, M., Soman, K. P., Poornachandran, P., Al-Nemrat, A., & Venkatraman, S. (2019). Deep learning approach for the intelligent intrusion detection system. IEEE Access, 7, 41525-41550.

[37] Kunang, Y.N., Nurmaini, S., Stiawan, D. and Suprapto, B.Y., 2021. Attack classification of an intrusion detection system using deep learning and hyperparameter optimization. Journal of Information Security and Applications, 58, p.102804

[38] Iqbal, M.F., Zahid, M., Habib, D., and John, L.K., 2019. Efficient prediction of network traffic for real-time applications. Journal of Computer Networks and Communications, 2019.

[39] Lin, P., Ye, K. and Xu, C.Z., 2019, June. Dynamic network anomaly detection system by using deep learning techniques. In International conference on cloud computing (pp. 161-17

[40] Reddy, R.R., Ramadevi, Y. and Sunitha, K.N., 2016, September. Effective discriminant function for intrusion detection using SVM. In 2016 International conference on advances in

[41] Ieracitano, C., Adeel, A., Gogate, M., Dashtipour, K., Morabito, F.C., Larijani, H., Raza, A., and Hussain, A., 2018, July. Statistical analysis-driven optimized deep learning system for intrusion detection. In International conference on brain inspired cognitive systems (pp. 759-769). Springer, Cham.

[42] Khan, F.A., Gumaei, A., Derhab, A., and Hussain, A., 2019. A novel two-stage deep learning model for efficient network intrusion detection. IEEE Access, 7, pp.30373-30385.

[43] Ayo, F.E., Folorunso, S.O., Abayomi-Alli, A.A., Adekunle, A.O. and Awotunde, J.B., 2020. Network intrusion detection based on a deep learning model optimized with rule-based hybrid feature selection. Information Security Journal: A Global Perspective, 29(6), pp.267-283.

[44] Asad, M., Asim, M., Javed, T., Beg, M.O., Mujtaba, H. and Abbas, S., 2020. Deepdetect: detection of distributed denial of service attacks using deep learning. The Computer Journal, 63(7), pp.983-994.

[45] Meng, F., Lou, F., Fu, Y. and Tian, Z., 2018, June. Deep learning-based attribute classification insider threat detection for data security. In 2018 IEEE Third International Conference on Data Science in Cyberspace (DSC) (pp. 576-581). IEEE.