REPUBLIC OF TÜRKİYE

ALTINBAŞ UNIVERSITY

Institute of Graduate Studies

Electrical and Computer Engineering

MACHINE LEARNING FOR NETWORK ANOMALY DETECTION

Rusul Tareq KHUDHAIR

Master's Thesis

Supervisor

Asst. Prof. Dr. Abdullahi Abdu IBRAHIM

Istanbul, 2023

MACHINE LEARNING FOR NETWORK ANOMALY DETECTION

Rusul Tareq KHUDHAIR

Electrical and Computer Engineering

Master's Thesis

ALTINBAŞ UNIVERSITY

2023

The thesis titled MACHINE LEARNING FOR NETWORK ANOMALY DETECTION prepared by RUSUL TAREQ KHUDHAIR and submitted on 00/04/2023 has been **accepted unanimously** for the degree of Master of Science in Electrical and Computer Engineering.

Asst. Prof. Dr. Abdullahi Abdu IBRAHIM

Supervisor

Thesis Defense Committee Members:

Asst. prof. Dr. Abdullahi Abdu IBRAHIM	Department of Computer Engineering, Altınbaş university	_____
Asst. prof. Dr. Ayça Kurnaz TÜRKBEN	Department of Software Engineering, Altınbaş university	_____
Asst. prof. Dr Tariq Mohammed	Department of Biomedical Engineering, İstanbul Arel University	_____

I hereby declare that this thesis meets all format and submission requirements of a Master's thesis.

Submission date of the thesis to Institute of Graduate Studies: ____/____/____

I hereby declare that all information/data presented in this graduation project has been obtained in full accordance with academic rules and ethical conduct. I also declare all unoriginal materials and conclusions have been cited in the text and all references mentioned in the Reference List have been cited in the text, and vice versa as required by the abovementioned rules and conduct.

Rusul Tareq KHUDHAIR

Signature

DEDICATION

I would like to start by thanking God Almighty for giving me the knowledge, health, and stamina to finish my research, and I thank my supervisor Dr Abdullahi Abdu IBRAHIM for all the support during my academic career. I dedicate this letter to my parents. There are not enough words to express, and I cannot repay what they have done for me. I will continue to work hard to meet their expectations, and finally, I dedicate this to my close family members and friends who have supported me through this study.

ABSTRACT

MACHINE LEARNING FOR NETWORK ANOMALY DETECTION

KHUDHAIR, Rusul Tareq

M.Sc., Electrical and Computer Engineering, Altınbaş University,

Supervisor: Asst. Prof. Dr. Abdullahi Abdu IBRAHIM

Date: April / 2023

Pages: 60

In previous years, the Internet of Things (IoT) developed rapidly from its wide application in many fields, and the use of Internet of Things devices recently, for example, in health, agriculture, industry, and so on. It devotes integrated attention to AI with the Internet of Things and defines Artificial Intelligence of Things (AIOT) technology that aims to collect, process, and make decisions autonomously and without human intervention in relation to IoT networks. Many problems occur in the Internet of Things networks, such as bank fraud, an increase in the number of attacks on Internet of Things networks, and other problems. The Internet of Things is the result of an increase in the number of attacks and distortions that hinder movement, and the failure of the IoT system. This study demonstrated the performance of five machine learning classification algorithms (support vector machine, decision trees, random forests, artificial neural network, and k-nearest neighbours). It was measured based on a set of data taken from the open-source Kugel website, built using the Intelligent Space Distribution System DS2OS, and compared with the prediction of many anomalies and attacks on the network, then the performance evaluation measures were used, which are accuracy, recall, Precision, and f1 score. In the thesis, high performance was obtained with several algorithms, but one algorithm gave the support vector machine little performance compared to other algorithms, and the best performance of the KNN algorithm was 94, 99% in measurement accuracy, and the Precision is 1, Recall 0.65, Recall 0.65 and F1- score 0.79. the confusion matrix of the KNN algorithm works correctly.

Keywords: Anomaly Detection, Machine Learning, Internet of Things (IoT), IoT Security, Classification.

TABLE OF CONTENT

LIST OF TABLES

Pages

LIST OF FIGURES

Pages

ABBREVIATIONS

ML : Machine Learning

KNN : K-Nearest Neighbours

RF : Random Forest

DT : Decision Trees

SVM : Support Vector Machine

ANN : Artificial Neural Network

IoT : Internet of Things

TP : True Positive

FP : False Positive

TN : True Negative

FN : False Negative

DOS : Denial of Service

D.P : Data Type Probing

M.C : Malicious Control

SC : Scan

M.O : Malicious Operation

W.S : Wrong Setup

SP : Spying

1. INTRODUCTION

The intensive development of Internet network technologies, over the past years, in the field of the Internet. Through these many developments, a new type of Internet device has emerged, and Internet of Things (IoT) devices communicate with each other without human intervention. IoT is the notion of linking a large number of devices together to enable automated, data flow between them. IoT technology has developed as a consequence of technological developments like Machine Learning (ML). Almost every industry is seeing the effects of IoT applications. This article will discuss a few of the many IoT uses. It intervenes in many fields, for example in the field of x-rays, intervention in the automated pathological system for cancer detection, and in the industrial and military fields and aircraft to detect defects in aviation, agriculture, mechanics, and many fields, thus increasing the use of the number of IoT devices and applications [1]. The use of ML techniques is a great option. When it comes to reliable IoT software. ML is a cutting-edge AI tool that is also a cutting-edge intelligence synthesis tool; so, it is not difficult, and powerful claims may go beyond it. All kinds of attacks from all kinds of networks. ML in artificial intelligence is used to develop systems and can learn by itself without explicit programming. Programming in artificial intelligence is minimal because it allows the machine to think. Integration of IoT technology with other advanced technologies such as artificial intelligence (artificial intelligence processing algorithms and data analysis), or big data (processing a huge volume of data from sensing IoT) [2]. Despite their popularity, IoT devices still present serious security risks due to issues such as insufficient protection for sensitive data, a lack of encryption, and inadequate network service protection. In parallel with this change. There are IoT of problems such as Internet traffic congestion, bank fraud, cyber-attacks on the network, etc [3]. And to solve these problems through artificial intelligence with machine learning and deep ML techniques Security between IoT devices becomes very important. The detection model used by ML algorithms requires a huge amount of data, which may be provided by IoT devices. The vast variety of assaults and their manifestations also makes it difficult for human operators to recognize and classify them. As the risks increase, you may be exposed to a host of Internet attacks and anomalies daily Security is now a concern, and despite the widespread use of IoT devices, they still pose serious security risks. Therefore, IoT devices are an easier target for attack because they spread data over wireless media.

While IoT system attacks spread widely and had catastrophic repercussions on IoT websites. IoT nodes create a backdoor that allows an attacker to obtain sensitive information from any important company. In recent years, experts have been investigating more sophisticated security technologies to address this problem. IoT systems can provide massive amounts of data that ML algorithms need to create their detection models [4].

1.1 GENERAL INFORMATION

1.1.1 Internet of Things (IoT)

Modern technology in information technology is a very large field. The rapid development of the IoT has become massive developments in smart IoT technology supported by systems and applications connected to devices such as connected cars, Smart cities, linked health, smart homes, wearable technology, smart grids, smart retail, and smart farming [1].

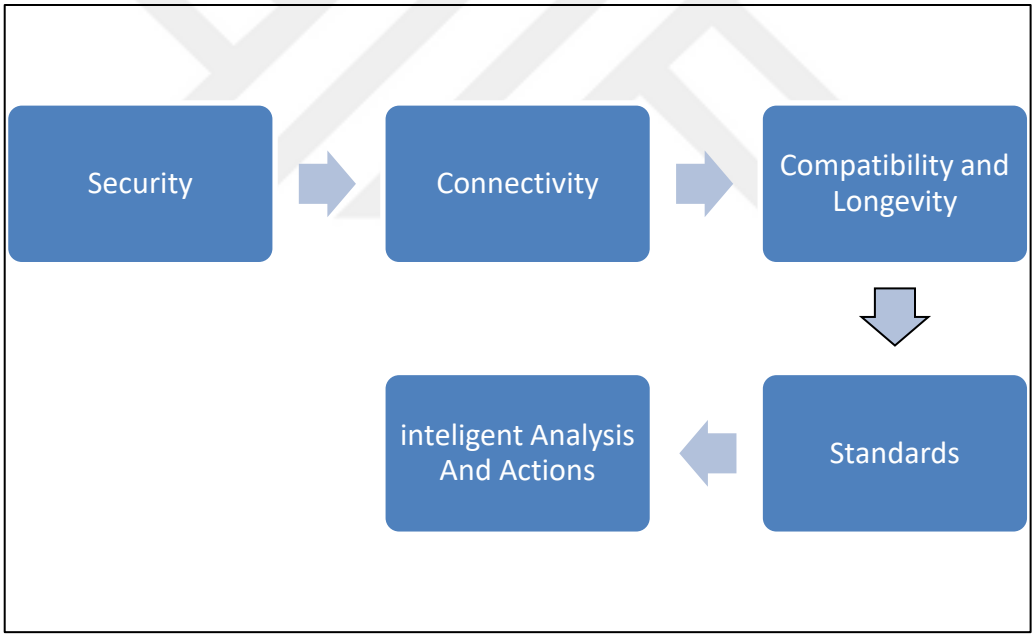


Figure 1.1: Internet of Things Challenges [1].

The IoT has a vast number of linked devices making the process of massive data collection and unidentified assaults a major security issue. Because user behaviour in the IoT environment is more unpredictable and hacking an IoT device is simpler than attacking a firewall-equipped PC system, hackers have more assaults as more portable devices enter the network with extremely poor security protections than ever before [2]. Without sensors, the Internet of Things would not exist because it cannot identify changes in the environment or

detect them to produce data. It is important for the Internet of Things, any physical object that can use the Internet of Things to sense data, store it, process it, and transmit it to the application [3]. There are many applications used in our lives. For example, smart homes, smart cities, healthcare, agriculture, wearable devices, and industrial automation.

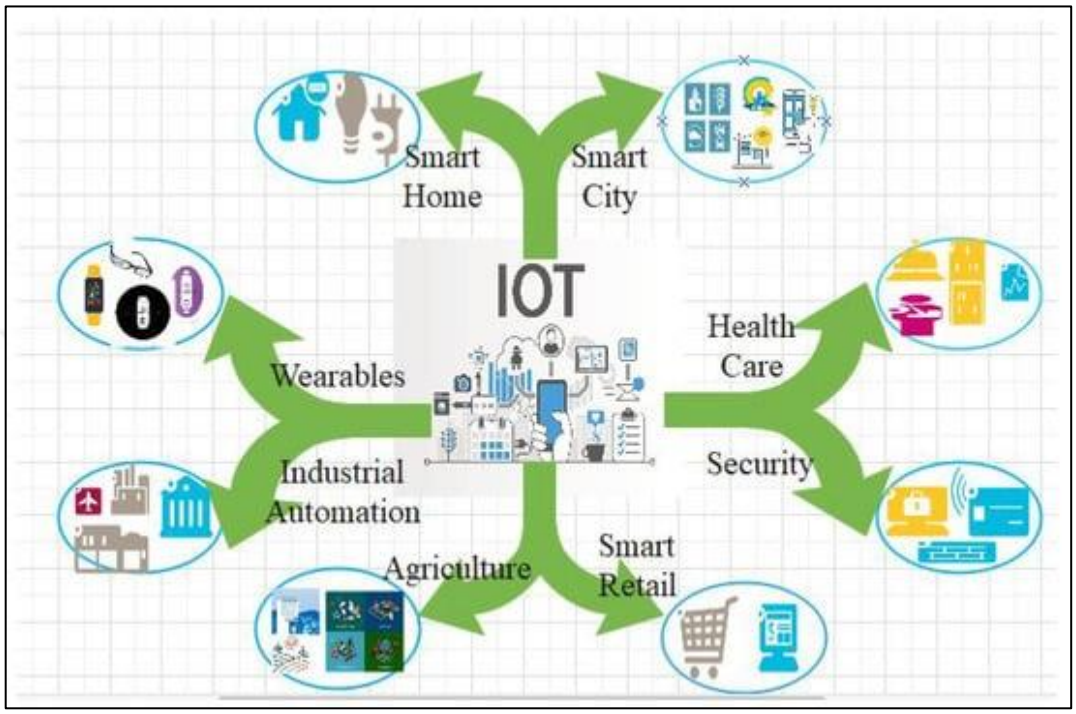


Figure 1.2: IoT Applications [3].

To provide smart applications and services, it is necessary to build an Internet of Things model. The structure of the IoT system consists of three layers. The first layer is the application layer, the second layer is the network layer, and the third layer is the presentation layer. The cyber system is the IoT system from small sensors to large global positioning units, and for radio.

frequency identification, it communicates with sensors in the near field, such as detectors and alarms. All IoT devices communicate, understand their environments, and classify and use them to store all kinds of information. Audio data, temperature reading and gs, and light intensity[2].

1.1.2 Anomaly Detection

Anomaly detection (freshness detection). When analysing a data set, abnormal cases that are outside the normal range are identified, that is, they are not consistent with the rest of the

data, or the observations may deviate significantly from the majority of the data and there is no agreement between them. There are broad categories for detecting anomalies.

a. Moderated anomaly detection techniques the dataset is categorized as normal and abnormal.

b. Anomaly detection techniques Unsupervised.

In light of the Answers In recent years, ML built on IoT devices has emerged as a promising new field of research. Academics who are interested in contributing to this field are starting to take notice. This section describes many ML algorithms that might be used to secure IoT devices [9]. The three basic architectural layers that make up the IoT architecture were used to examine these potential solutions. The Surface's typical web and app authentication techniques sit atop the physical/cognitive layer and below the network layer. Due to the precise threshold setting for identifying unwanted transmissions that create false alerts, Surface's standard web/app authentication techniques are unsuitable. In short, it lets authorized users interact with the system in a secure environment while protecting the data of other authorized users [10]. Under the proposed peer-to-peer encryption protocol architecture, clients will need to sign up with the cloud Registry before they can establish an IoT connection.

Domain. Allam et al. (2018) proposed using either the Neural or Beauty algorithm models to keep IoT devices safe from attacks and functioning normally. They were pressed by the public-private key encryption mechanism that was already in place. The information is first classified and then linked to its appropriate training sets. This is in addition to what Baracaldo et al. Create a situation in which you are being observed at random by suggesting a new security measure, such as a way to find and clear up collected hazardous data. Regardless of how regular attacks are, the challenge of network security is the connecting element between the actual world and the virtual one. This calls for a wide range of ml methods, including SVM, ANN, and KNN, to identify intruder attacks [2].

1.1.3 Challenges and Possibilities of Anomaly Detection in the IOT

The IoT for short is a verbal concept of current technology problems and security flaws that are prevalent in the field of traditional IT. However, the basic problem with IoT is the lack

of available resources, which makes information technology difficult to update ever-evolving security solutions [13]. However, the networks that make up the IoT host these solutions. To ensure the success of the IoT, it is imping its infrastructure and computational frameworks mubendned without compromising the security of users' private data. Optimization of networked systems may face other hurdles due to the proliferation of IoT devices. On the one hand, because of computers, there are additional problems with protection protocols in the IoT, including encryption and other methods of security and privacy. hardware. The most insurmountable obstacles are rarely isolated issues [3]. For example, security issues such as denial-of-service attacks and intrusions may be met with poor treatments, such as false positives. As the market no longer believes in these solutions, their effectiveness has also decreased. There will be many uses for a structured approach to securing the IoT and the privacy of its users, including in the development of new smart, reliable, evolutionary, and scalable IoT security. A trustworthy IoT framework can be built with the help-based technologies Help-based artificial intelligent techniques that can operate without being explicitly programmed and can be bypassed in complex network architectures [12]. Effective adoption of the IoT requires extensive research into the opportunities presented by security and privacy issues. Device and network security measures learned for IoT The scope and significance of the IoT have reached every corner of the world in recent years. Many researchers in the field of networks and devices have been interested in the issue of IoT security [11]. Future study avenues can be revealed by examining the deployment, use, and impact of IoT networks, which reveal a wide range of difficulties and exposures.

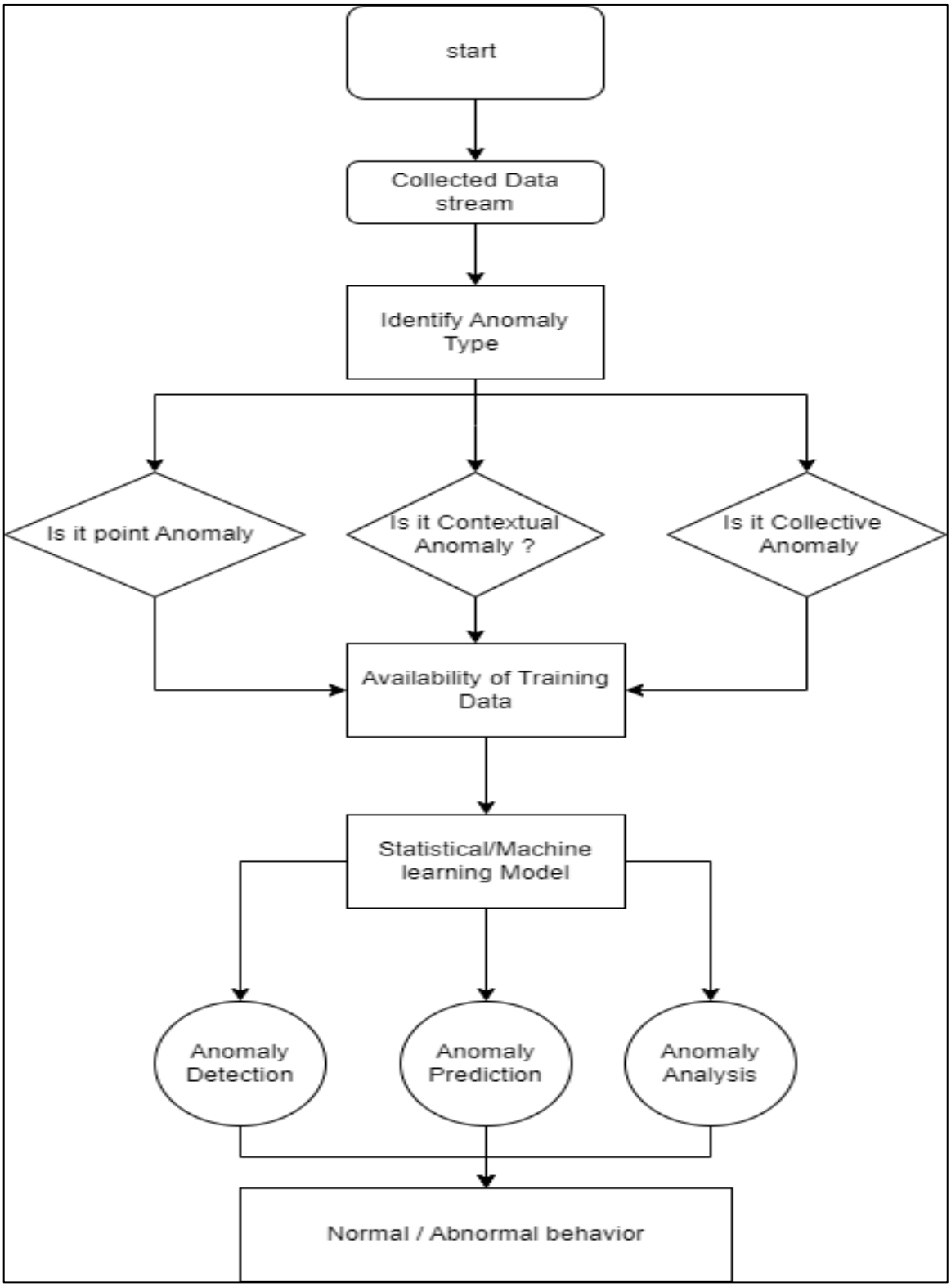


Figure 1.3: Structural Anomaly and Attacks[4].

1.1.4 Security for IOT Using Machine Learning

Several modern applications make use of the distinctive approach to tackling complicated issues offered by ML techniques. ML algorithms are designed to tackle difficult problems and boost efficiency via exposure to training and experience. ML algorithms from experience create and regulate machines on their own. In the IoT, gadgets serve several roles and are easily accessible. Because anyone can access some IoT devices from anywhere without the user's permission, the security of IoT devices is a concern because, on the one hand, they make human life more technologically advanced, easy-going, and conformable; on the other hand, they put the users' privacy in greater danger from various threats and attacks. Protecting devices connected to the internet requires many layers of protection. Complex security mechanisms would be ideal, but IoT devices' limited processing capabilities prevent their implementation. A breach occurs when an unauthorized person gains access to a system and discloses sensitive data without the knowledge or consent of the system's owner or operator. As a sort of artificial intelligence, ML technology employs the usage of trained algorithms to provide predicted performance in response to input data to identify threats. To identify illnesses, bank fraud, network assaults, and a wide variety of other network-based crimes and intrusions. Network threat analysis is only one example of how ML is being used in cybersecurity to foil potential cyberattacks [5].

1.1.5 Machine Learning Technologies

ML is by far the most common kind of schooling Since computers learn from experience, ML is utilized to provide robots with decision-making abilities comparable to those of humans. As the goal of artificial intelligence (al) is to impart "intelligence" onto computers, it is not surprising that the two fields are tied to one another in practice al. No specific instructions are supplied to the machine in ML instead, the ML is on its own. This is accomplished by identifying patterns in data for use in making inferences and taking action [12, 13].

Several ML algorithms, which serve as a means by which computers may be trained, are required to improve their performance. This approach eliminates the need for direct human involvement in the issue.

Several fields largely rely on ml since traditional approaches are ineffective. The applications may be specified. It can understand massive amounts of data and solve difficult issues that cannot be addressed by more conventional means. ML techniques may be used to discover answers to problems when conventional approaches either aren't applicable or need a high level of human involvement to maintain updates. Analysing the data allows ML techniques to be transferred to novel settings. Typically, to develop a high-accuracy classification model, it is extremely crucial to choose a strong ML algorithm as well as to tune its parameters. If their parameters are set up correctly, most ML algorithms may provide optimum results. Bad categorization results were obtained due to insufficient parameter values. When there are a lot of variables in the learning process, manual parameter tuning might take a long time. Hence, the grid search was first developed as an exhaustive search of a restricted region of the whole hyper-parameter field. To do this, the grid search method employs a cross-validation (CV) approach to improve the model's parameters. The objective is to find an optimal set of hyper-parameters for the classifier so that it can make reliable predictions about new data [6]. The issue of over-fitting is avoided by using the CV method. Many factors, including which features are picked from the dataset and how much weight is assigned to each feature, may influence how well an ML system performs. Overfitting the model may occur when more characteristics are included than are needed [7]. Based on whether or not the training data are labelled level of supervision the algorithms got during training, there are three distinct categories of ML algorithms. We may classify these approaches as either supervised, unsupervised, or semi-supervised [8]. Training under close supervision: It relied on properly labelledategorized training data. Every one of the flows in the dataset, for instance, has attributes (labels) that specify what kind of flow it is (such as normal or harmful). Supervised learning Decision trees, K-nearest neighbours, random forests ...etc. are all examples of such algorithms. It's important to keep in mind that not all ML techniques are created equal. for example, are less complex than other approaches yet provide just a limited variety of possible estimates due to the limited variety of shapes they can generate. One of the most popular types of education is instructional learning. There is a meaningful label attached to each data point. We employ an ML technique called a learning algorithm to sort the outputs in acbyinputs. This approach cannot be used without a previously trained data set. Supervised learning encompasses both classification and regression[5].

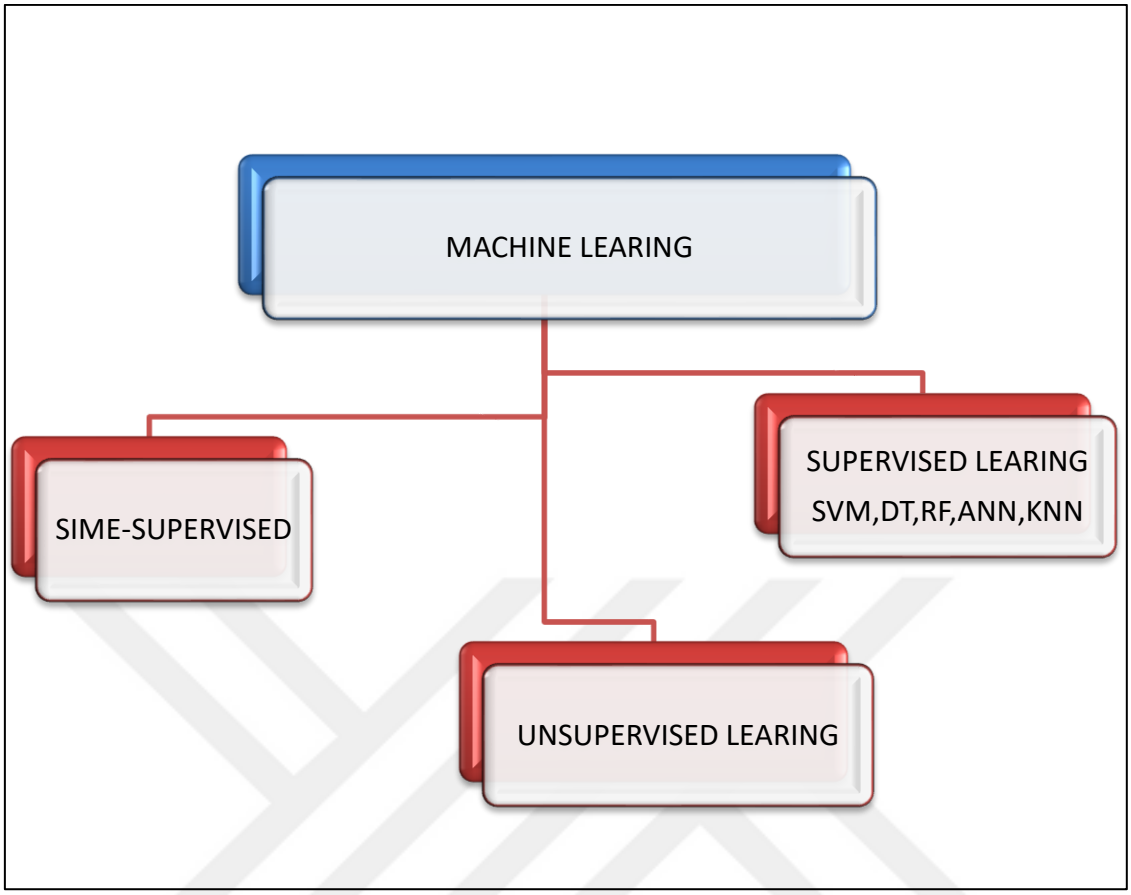


Figure 1.4: Type Machine Learning [5].

1.1.6 Supervised Machine Learning

Supervised Machine Learning Includes Two Types

a. Classification-based learning: - A classification learning algorithm produces a discrete value or class, such as [true, false], [yes, no], etc. as its output [6].

b. Regression-Based Learning: Regression-based ml is a subfield of supervised ML in which we manipulate numerical variables to represent probabilities.

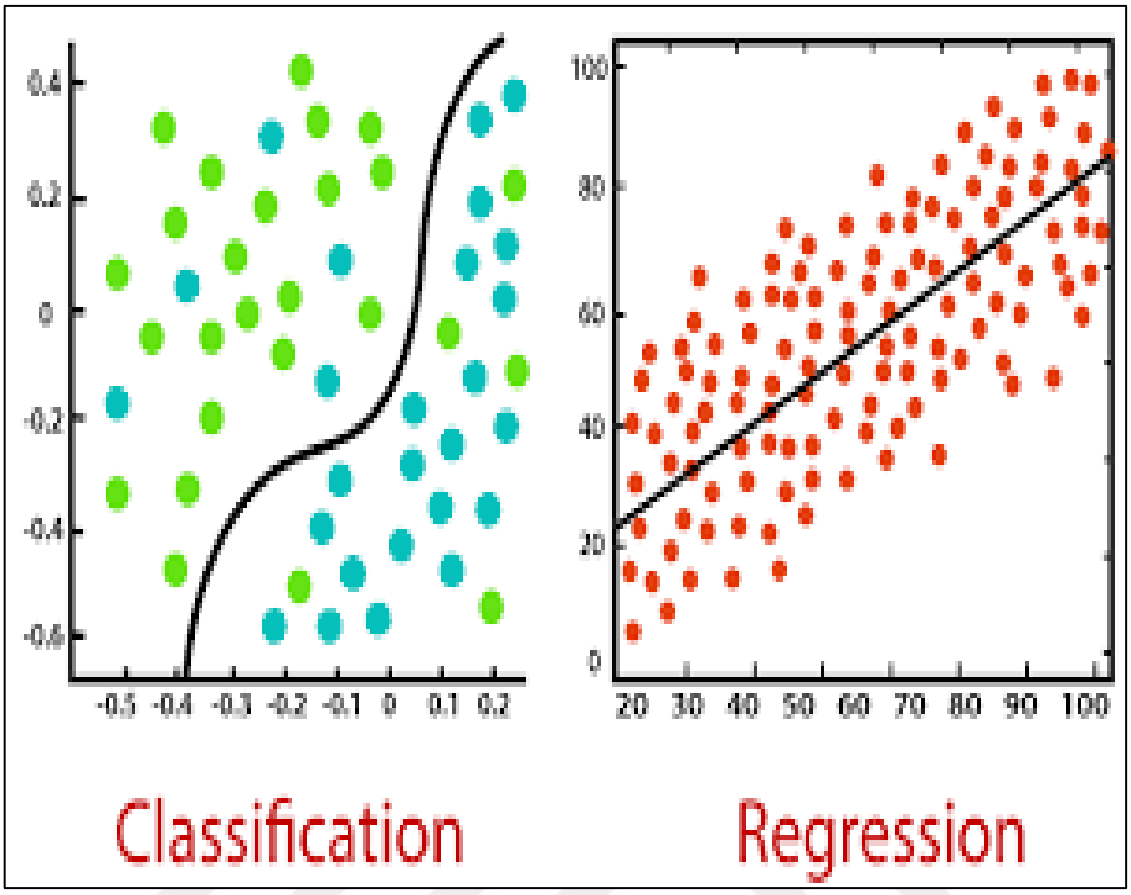


Figure 1.5: Show Structure Classification and Regression [6].

1.1.7 Unsupervised Learning

Unsupervised education In this technology, there is no labelling procedure. The algorithm separates the data into groups based on different characteristics and examines how they are related. It is widely utilized in fields including dimension reduction, connection learning, and anomaly detection [7]. When supervised learning and unsupervised learning are mixed, a hybrid approach known as semi-supervised learning results. Typically, only a tiny fraction of the data gets labeled, leaving the majority un labeled. This approach combines the very effective [8].

1.2 CONTRIBUTIONS

a. categorical datasets, we develop and put into practice an attack and anomaly detection strategy based on machine learning, employing multi-class classification rather than binary classification.

b. In contrast to conventional machine learning in IoT contexts, we show the usefulness of machine learning models in attack and anomaly detection systems.

c. For the categorical dataset, our suggested method selects a machine model that produces the best anomaly detection and the fewest false positives.

1.3 QUESTIONS ABOUT THE SEARCH

Questions addressed and answered in this letter.

What are the best-performing machine learning algorithms for detecting attacks and anomalies in IoT devices?

The results of the research to compare the algorithms are discussed as shown in the fourth section of the thesis.

How are algorithms tested on a given DS2OS dataset?

The following approach is used for this research

a. Explanation, classification, and processing of the data set.

b. The algorithms will be executed on a data set.

c. The performance results of the selected algorithms were discussed and the used algorithms were compared with each other in terms of performance and classification.

1.4 THESIS OBJECTIVE

The primary objective of this study is to develop an intelligent system design for a secure infrastructure and implement a set of machine learning algorithms to identify network-based attacks and anomalies based on IoT, and to evaluate these algorithms using the DS2OS dataset, a new dataset that includes both malicious networks captures and benign. from a variety of IoT devices. In this project, these data were tested on five supervised machine learning classification algorithms (SVM, DT, KNN, ANN, RF). For this task using the Python language, the results were presented and the performance of a set of classification algorithms was discussed and compared with each other with several procedures to identify attacks and abnormal data.

1.5 RESEARCH STRUCTURE

The structure of this dissertation is as follows:

In the second chapter, previous studies related to this research will be presented, In the third section, a hypothetical dataset will be explained and its types will be classified, its properties and types of classification classes will be explained, and its pre-processing will be explained. The details of the device that was used to implement the program are also explained, and then an explanation of the work of the algorithms that were used in the criteria for evaluating research work, confusion matrix, and performance measurement, In the fourth section, the results of applying the test algorithms are presented, compared, and summarized. In the fifth chapter, a summary of what is going on about the research, its details, conclusions and purpose will be explained.

2. RELATED WORK

2.1 Structure Work

Several new studies are being conducted to evaluate the detection of assaults and irregularities in the IoT networking environment, and we drew heavily on works published in the previous few years. You may read about this part of the research below.

a. Pajouh et al. presented a DL strategy based on RNN for IoT malware analysis. The authors took into account IoT applications built on Advanced RISC Machines (ARM). The authors use several datasets of current malware to train their models, and they test their framework with fresh malware. The authors came to the conclusion that LSTM classifiers outperform other classifiers through experimentation. More specifically, the RNN-based DL approach detected new viruses inside IoT apps with 98% accuracy. includes a detailed look at a smart home system that uses Dense Random Neural Network (DRNN) deep learning to check for security flaws. The bulk of their attention was given to detailing the DoS attack process and how to conduct a DoS attack on a basic IoT site t[9].

b. The author Al-Sumairi and Al-Subhi January 2019 used the Bot-IoT dataset. This is implemented using the older dataset, does not cover all types of attacks, and uses classification algorithms technology. For feature selection over MLA, all these results are obtained With 7 features. Accuracy rate KNN = 99%, ID3 = 97%, RF = 97%, AdaBoost = 97%, QDA = 87%, MLP = 84%, NB = 79%[10].

c. The author Wehbi et al. in April 2019 used the Cleveland dataset on heart disease dataset. This dataset is based on the diagnosis of heart diseases and uses classification algorithms technology, DT, ada boosting, and multilayer feed-forward neural network Accuracy rate of 59.73% (Ada Boost)[11].

d. The author, El-Sayed et al. In 2020, it uses the dataset CICDDOS2019 and through classification algorithm technology. A new model with which this accuracy is achieved has been proposed In the full feature set, a 99% accuracy rate results obtained[12].

e. Nicolau et al. use datasets CTU13, UNSW-NB15, AND NSL-KDD and ML models UL: AE. anomaly detection of various attacks, such as DOS attacks uses the methodology to push normal data towards the origin and aberrant data distant from the origin, update AE models. he was a disadvantage of these recall is under 90% in some cases[13].

f. Nomm et al. ML models UL: I Forest and OCSVM use UL models with less than 10 selected features to detect attacks. detection anomaly IoT by botnet attacks. were the detection rate low in some cases[14].

g. Rathore et al. detection anomaly IoT models DOS attacks, etc. use By double-checking the data and categorizing it when both outputs have high confidence, pseudo-label unknown data was Accuracy is lower than 90%[15].

h. Weinger et al, in 2022, use Dataset TON-IoT and DS2OS. performance Metrics(Accuracy , Precision , Recall, f1-secor) .Improving detection performance for IoT anomaly detection(AD) using Federated Learning (FL) [16].

i. Pajaud et al. suggested a model that incorporates a dual-tier classification module and a dual-tier dimension reduction module. Those who require a citation: The purpose of this approach was to identify malicious actions like User to Root (U2R) and Remote Local (R2L) attacks. These terms designate two distinct forms of cybercrime. It is possible to minimize the size of an issue in several ways; two examples are component analysis and linear discriminate analysis. The research relied heavily on data accessed from the NSL-KDD repository. Naive Bayes and Certainty Factor K-Nearest Neighbour algorithms were utilized in the two-tier classification module that was developed to spot suspicious behaviour [17].

j. Ukic et al.looked at the potential for anomaly detection using Internet of Things (IoT) healthcare analytics. Also, this study contributes a model for the diagnosis of heart issues using a smartphone. Big data mining, predictive analytics, medical image analysis, biomedical signal analysis, and Internet of Things-connected sensors are all used in healthcare anomaly identification (IoT) [18].

3. METHODOLOGY

3.1 INTRODUCTION

this section details the many machine-learning techniques we use throughout the categorization multi-class classification model. All out-of-the-ordinary occurrences are lumped together as anomalies in our data. Based on the DS2OS dataset that was generated using the Intelligent Space Distribution System. To detect attacks and anomalies using machine learning we created a training and test set using our dataset's features. as 80 per cent and 20per cent, respectively. Both the training and testing datasets were utilized to master the algorithm and evaluate the effectiveness of the machine-learning techniques. In this study, we want to categorize ML algorithms to identify the most effective, precise method. In order to accurately categorize data and provide insight into how to construct supervised machine learning models, it's important to first establish the performance of various methods on big and smaller data sets. Then the proposed model was implemented using five supervised classification algorithms: Decision Tree, Support Vector Machine (SVM), Artificial Neural Network (ANN), Random Forest (RF), and K-Nearest Neighbors (KNN). An illustration Figurers 3.1 below will outline the proposed anomaly detection approach.

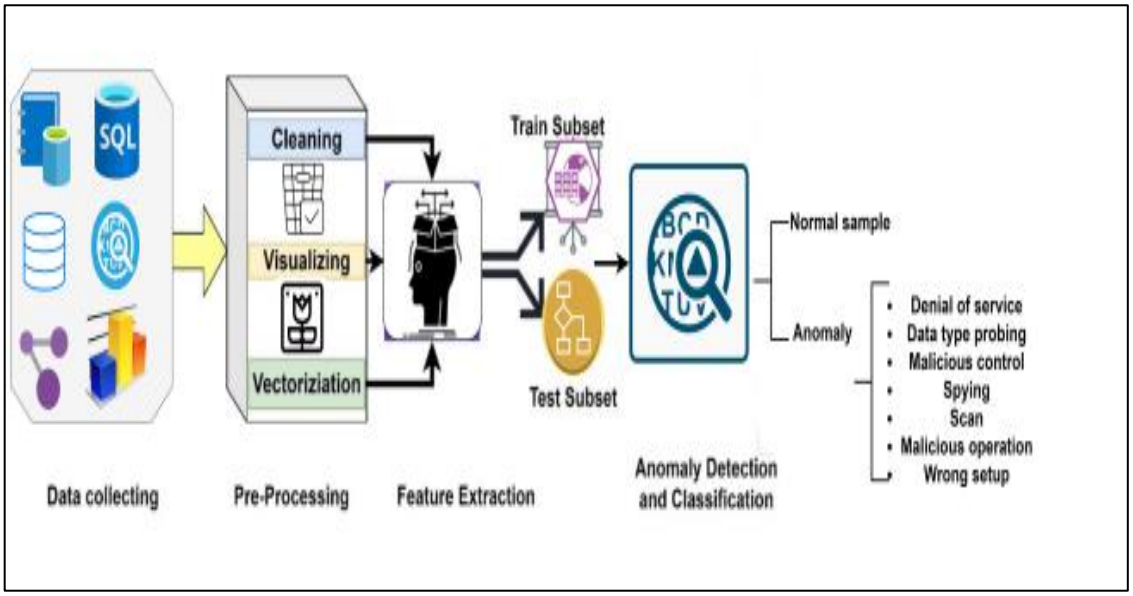


Figure 3.1: The Schema of the Proposed Anomaly Detection Approach [19].

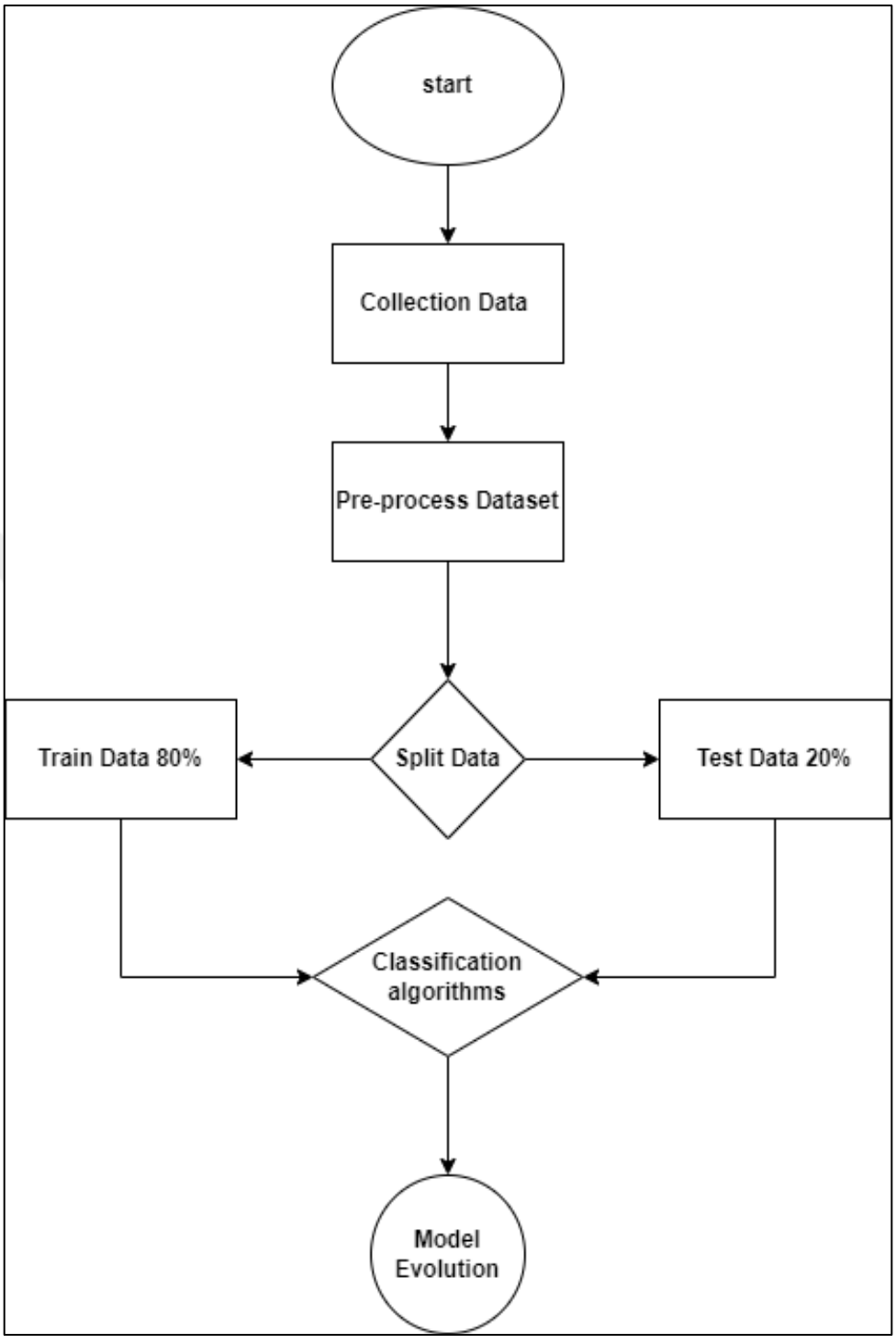


Figure 3.2: The Flowchart Approach Proposed That is Implemented to Detect Attacks and Anomalies [20].

3.2 DATASET

In this section, the data set is explained in detail. Data collection is one of the challenging tasks especially when it comes to detecting several attacks and anomalies. The data set is taken from the Kaggle website, [14]. which is open source. This data was generated by using a distributed intelligent system DS2OS for IoT the environment. This data is categorically containing nominal values and numerical values. Each set includes data from 357,952 samples. The data set consisted of 347,935 normal data and 10,017 abnormal (abnormal) data. Consists of 13 features in the dataset.

3.2.1 Type Features in the Dataset.

we classified the features as normal or discrete as shown in Table 3.1 below.

Table 3.1: Feature Type [14].

Frequency distribution Source ID (SID)	nominal
Frequency distribution Source Address (SA)	nominal
Frequency distribution Source Type (ST)	nominal
Frequency distribution Source Location	nominal
Frequency distribution Destination Service Address	nominal
Frequency distribution Destination location	nominal
Frequency distribution Accessed Nod Address	nominal
Frequency distribution Accessed Node Type	nominal
Frequency distribution Operation	nominal
Frequency distribution Value	nominal
Frequency distribution Normally	nominal
Timestamp	Discrete
Destination Service Type	nominal

Several frequency features are described and photographed on the site in the dataset as we note in below Figure 3.3 (Accessed nod type**),** figure 3.4 (operation), figure 3.5 (source location), figure 3.6 (Destination location), figure 3,7 (Source type), figure 3.8 (Destination service type).

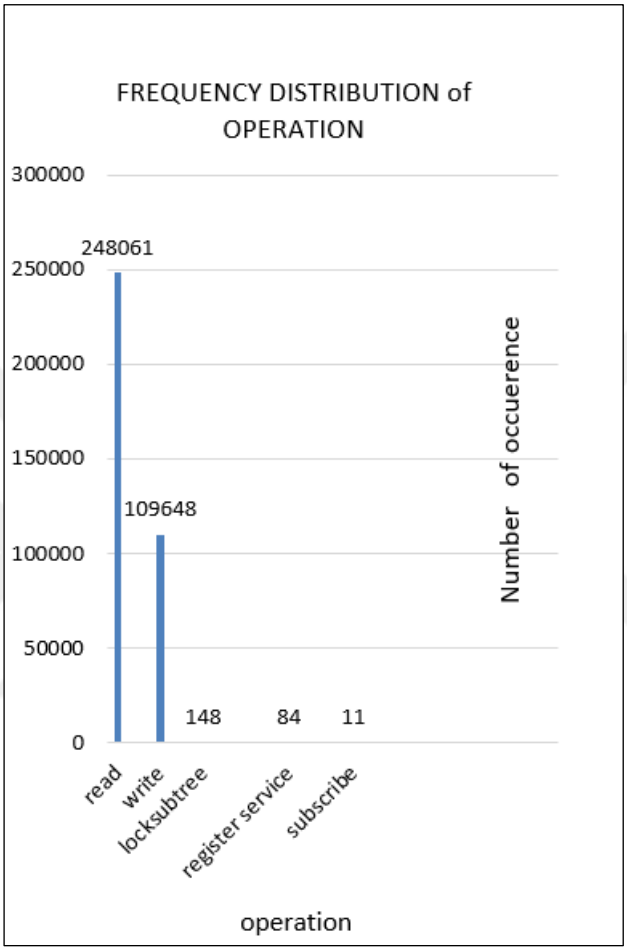


Figure 3.3: Accessed Nod Type [14].

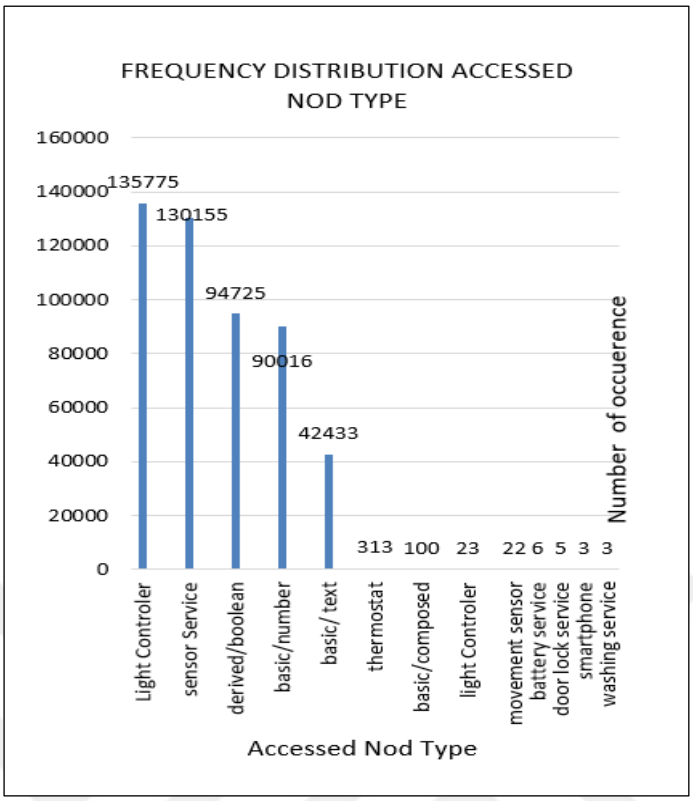


Figure 3.4: Operation [14].

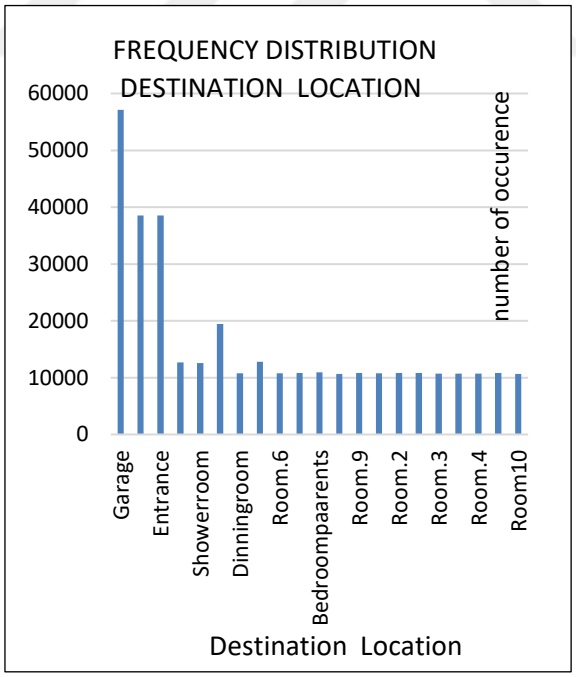


Figure 3.5: Source Location [14].

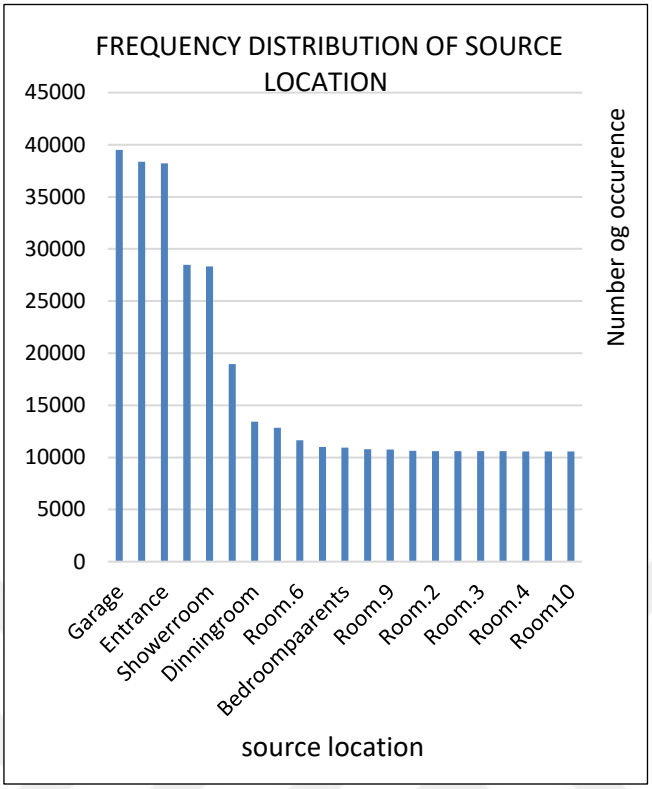


Figure 3.6: Destination Location [14].

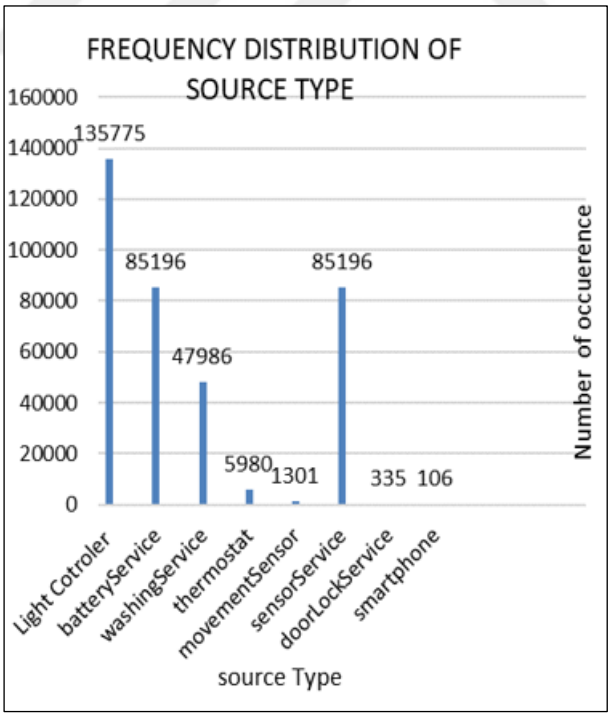


Figure 3.7: Source Type [14].

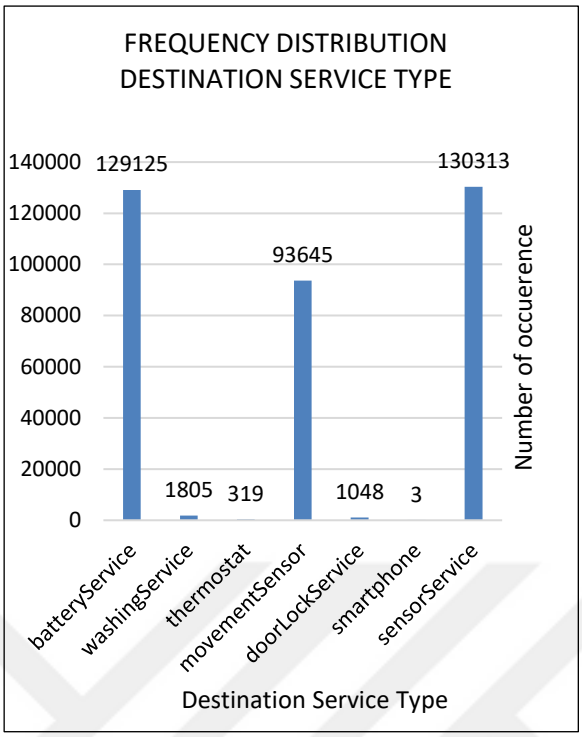


Figure 3.8: Destination Service [14].

3.2.2 Pre-Processing Dataset.

To process the data, we perform several operations on the raw data and convert it into a format that the machine learning can understand, ftoananalysehe the data, and engineer the features [16]. To extract the necessary and error-free data. Each inclua des dataset from 357,952 samples. The dataset consists of 347,935 normal data and 10,017 abnormal (abnormal) data containing eight categories and categorized respectively as the type of node accessed, data missing and value type [19]. The timestamp is omitted from the dataset because it contains the lowest value. The most important goal of this research is to convert the dataset type in several ways from nominal data to feature vector files using label coding techniques and one hot frequency count. The use of this technique is the most common for any categorical variable in which this technique is applied, it may require a lot of special attention before applying any of the machine learning algorithms[20]. Eight Categories of Classification Data.

3.2.3 Type Categories of Classification Data

i. Denial of Service (Dos): - The attacker sends junk requests on a device to disable the server and the user cannot access it[21].

ii. Data Type Probing (D.P) :These attacks involve a different type of node that generates new data instead of real data[22].

iii. Malicious control (MC): An attacker can possess the session key or manage network traffic through vulnerabilities (holes). In this way, a malicious attacker may gain control of the entire system[23].

iv. Malicious process (M.O): Malware creates malicious processes

v. Scan: - When the system is scanned it will get data from the hardware and often produce corrupted data[24]

vi. Spying (SP): Intrusion occurs through the system port and the attacker exploits system flaws to find important data.

vii. Wrong Setup (W.S): Data can be corrupted due to wrong system setup.

viii. Normal (NL): - pure (correct) and accurate data [25].

It will be explained in table 3.2 below.

Table 3.2: The Frequency Distribution of the Attacks and the Percentage for the Dataset Will Be Shown in the Table [19].

digital serial	Name attackers	frequencies Count in data	The percentage of the attacker of total data	The percentage of the attacker of Anomaly
1.	Denial of Service (DOS)	5780	01.61%	57.70%
2.	Data Type Probing (D.P)	342	00.09%	03.41%
3.	Malicious Control (M.C)	889	00.24%	08.87%

Table 3.2: The Frequency Distribution of the Attacks and the Percentage for the Dataset Will Be Shown in the Table. "Table Continued"

4.	Malicious Operation (M.O)	805	00.22%	08.03%
5.	Scan (SC)	1547	00.43%	15.44%
6.	Spying (SP)	532	00.14%	05.31%
7.	Wrong Setup (W.S)	122	00.03%	01.21%

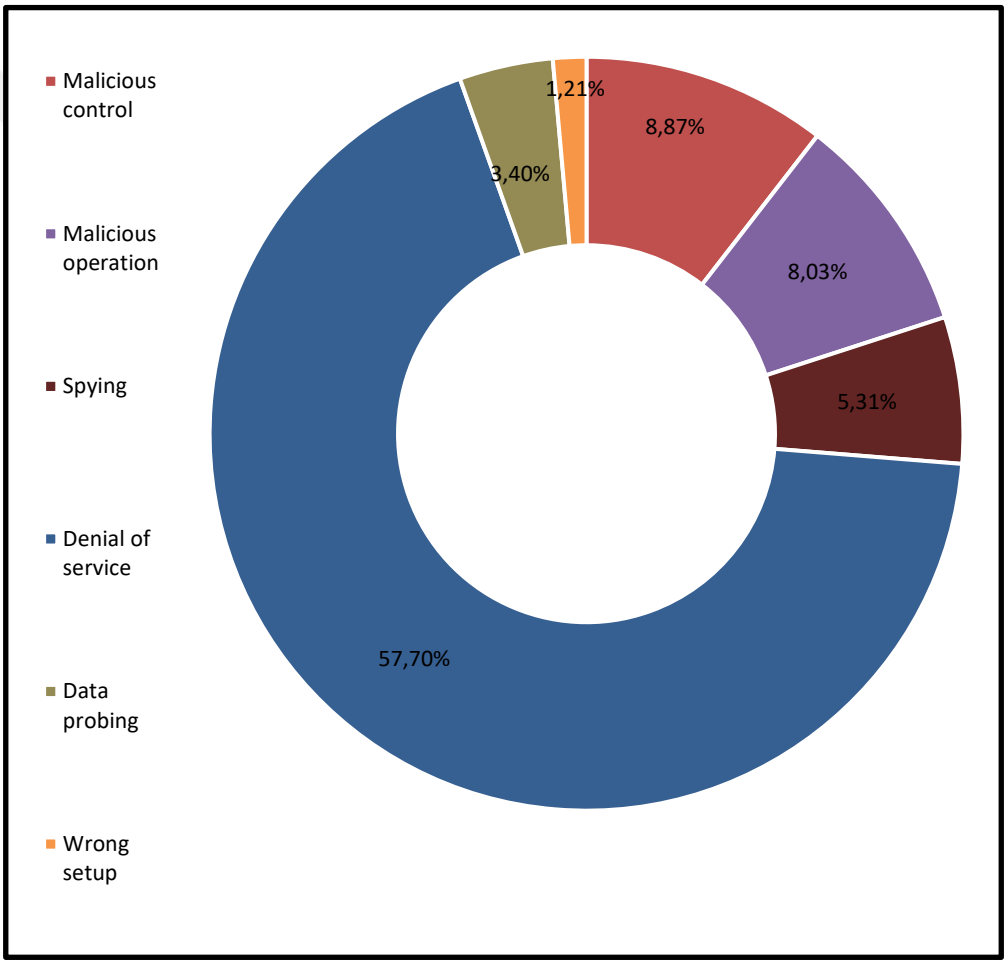


Figure 3.9: Describe the Percentage Distribution of Frequency [19].

Illustrated in Figure 3.9 above The is frequency distribution of the attacks (DOS, D.P, M.C, M.O,S C, SP ,W.S) and the percentage.

3.3 HARDWARE AND EQUIPMENT

The specifications of the device that was used to implement several Python algorithms on the default data set of the Internet of Things environment will be explained in the cola program on which the research was conducted, can view table 3.3 below.

Table 3.3: The Specifications of the Device [19].

User	Types
Device Name	HP (Hp Laptop - DESKTOP-GJQI0D3) desktop
System Type	64-bit operating system, x64-based processor
Operating system	Windows 10 pro-64-bit (10.0, Build 19045)
OS build	19045.2486
Memory	8192 MB RAN
Processor	Intel(R) Core (TM) i5-6200U CPU @ 2.30GHz 2.40 GHz
Program implementation	Use colab

3.4 EXPLANATION OF SUPERVISED CLASSIFICATION ALGORITHMS

In this section, details of the supervised machine-learning classification algorithms will be explained.

3.4.1 Decision Tree (DT)

DT is one of the learning algorithms Belonging to a supervised type that solves problems using classification and regression [8]. The working machine is based on the divisible root node and the non-divisible leaf node and has a hierarchical model consisting of nodes and then multiple nodes and each node gives the ability to compare expected workflows based on their associated benefits. costs and capabilities became the tree more complex and become more form-fitting as they evolve [19].

Root Node: To create any DT, a root node must be provided. Using a statistical approach, this feature is finalized, to reduce impurities in the data set for a particular feature we use the segmentation technique.

Inner node: Child nodes that are still executing or still could fork.

Leaf knots: No cleavage is known as terminal knots, i.e., the maximal inclusions are removed by the style. As shown in figure 3.10 below.

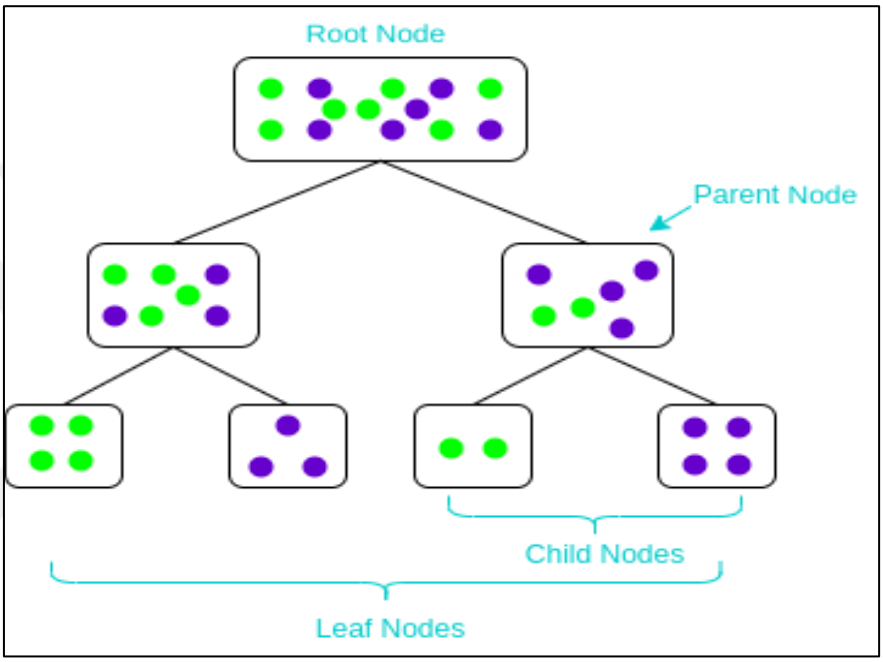


Figure 3.10: Decision Tree (DT) [8].

3.4.2 Support Vector Machine (SVM)

SVM is a form of supervised learning. Techniques for classification and regression are employed. To widen the distance between each class, SVM is based on the super level between the two classes. After examining the data, SVM executes the kernel function to linearize the hyper level when it is nonlinear. Due to its great precision, it is appropriate for Internet of Things security applications [9]. As shown in figure 3.11 below.

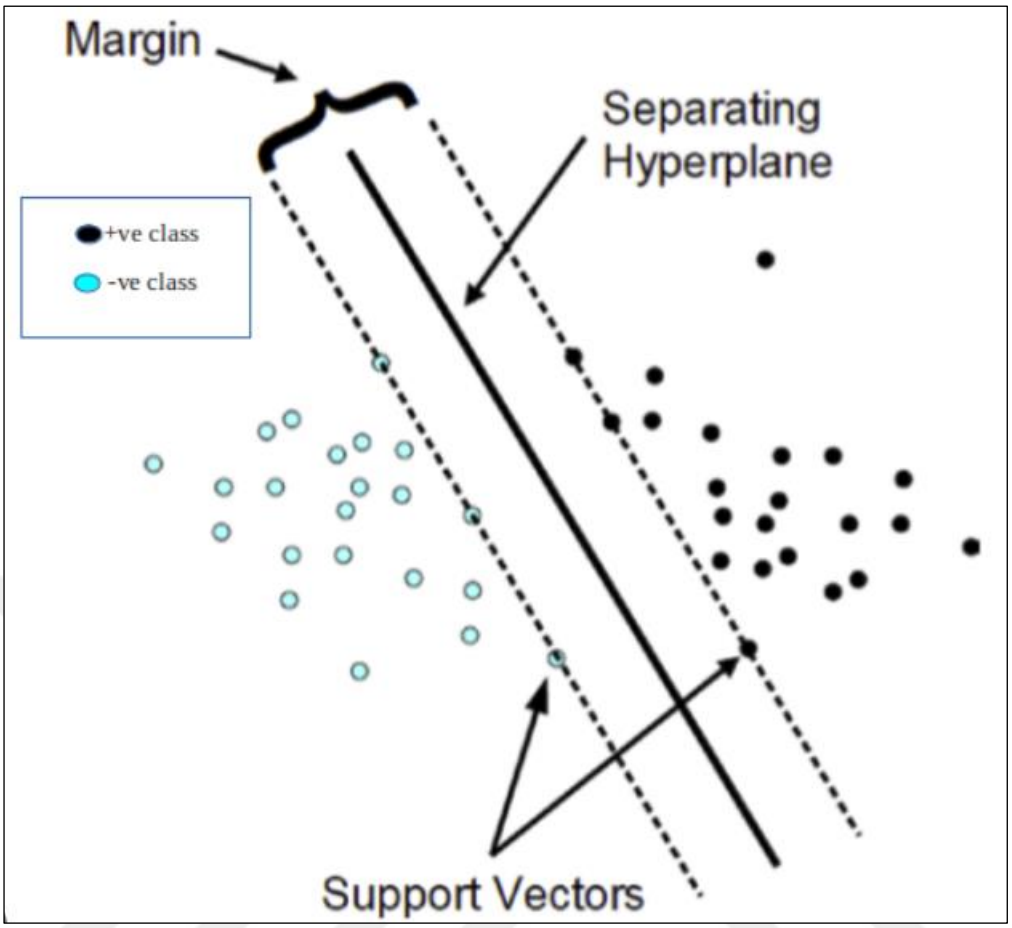


Figure 3.11: Support Vector Machine (SVM) [9].

3.4.3 Artificial Neural Network (ANN)

Deep learning encompasses a wide range of machine learning techniques, including the ANN method. The research origins of ANN are a single-layer cognitive algorithm. The model can solve problems by changing the coupling strength between artificial lattices in the networks. ANNs are often capable of relatively high performance on multidimensional and non-separable problems. The ANN algorithm is trained from scratch using just the raw data, unlike other classifiers, but it still takes a long time before the error may be minimized by using several features. Each individual neurzed to train the neural network. As there are so many customizable features, its architecture may be considered complex. Any X in the range [X1]–[Xn] is equivalent to X. (where X1…. Xn represents individual characteristics). Random weights W = W1, W2, W3...Wn is added to more features, and so are bias values b = b1, b2... one billion [28]. As shown in the figure 3.12 below.

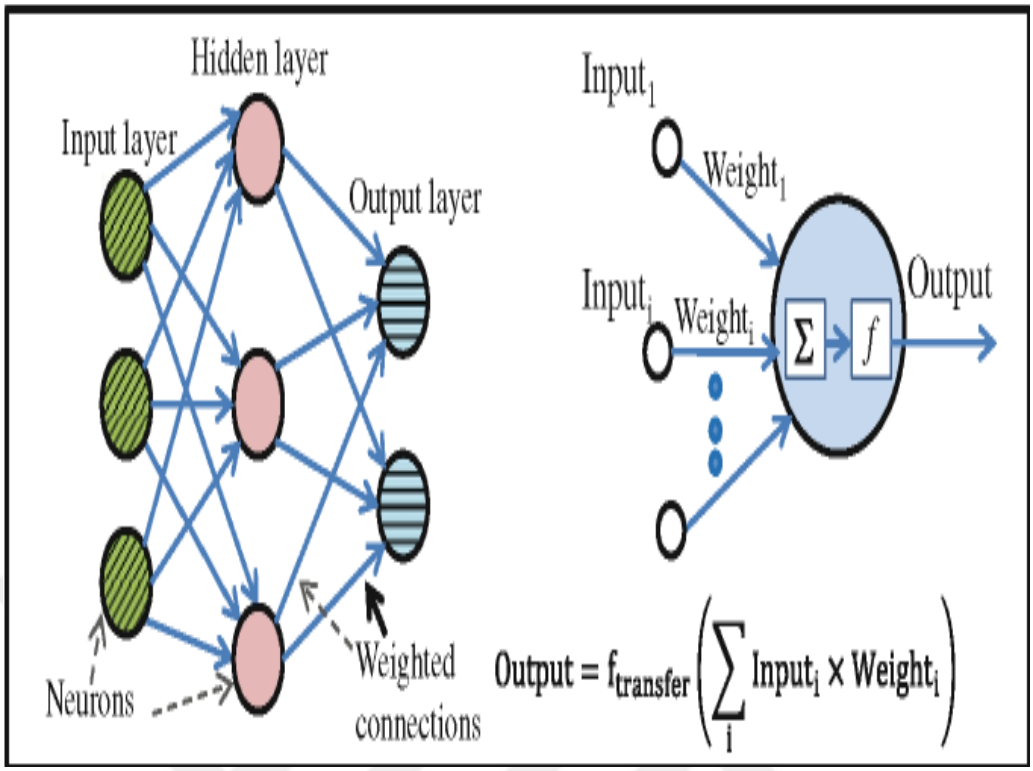


Figure 3.12: Artificial Neural Network (ANN) [28].

3.4.4 Random Forest (RF)

Random forest machine learning, within a supervised classification. The number of combinations of decision trees forms a random forest cluster, so the prediction accuracy of a single decision tree is lower than the prediction accuracy of raa ndom forest. Therefore, high-accuracy classification is performed on multiple trees with different data, multiple training, and a different number of features to extract predictions. The appearance of the forest also increases, when the number of trees in the forest increases [26]. As shown ie figure 3.13 below.

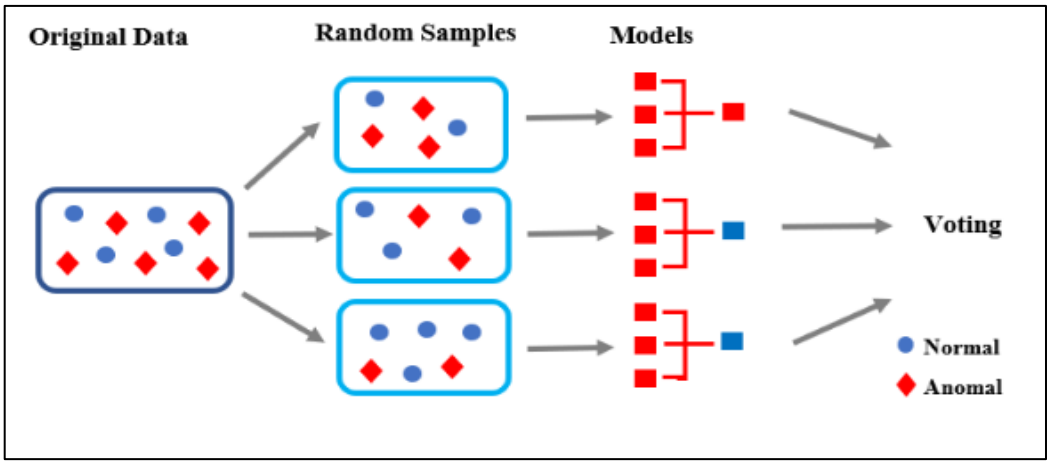


Figure 3.13: Random Forest Algorithm [26].

3.4.5 K-Nearest Neighbour's (KNN)

KNN deep learning it is collects similar training points using the value K. Find the nearest data points. It can be predicted based on the average value of the nearest neighbours. Although this number is not exact, it can be used to locate a possible missing node.Too detect intrusion and malware and to disco defects in the Internet of Things, the algorithm model is used. Our algorithm takes a long time to implement [27]. As shown in the figure below.

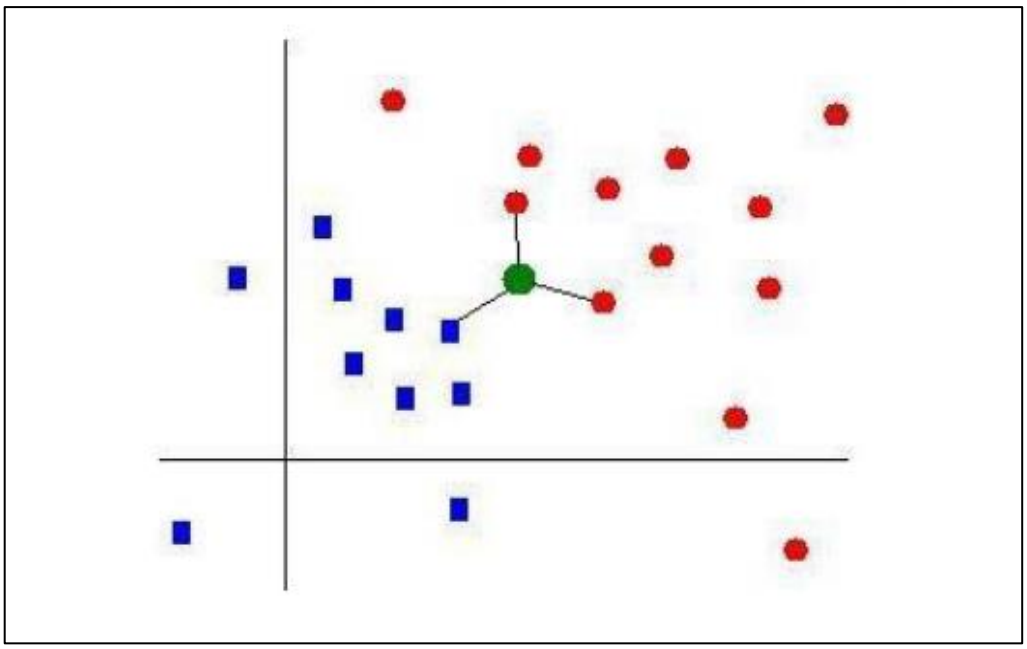


Figure 3.14: K- Nearest Neighbours [27].

3.5 CROSS-VALIDATION

It's an important topic about testing your machine learning models. Provides the ability to estimate model performance on unseen data that was not used during training One of the most popular data resampling techniques for assessing generalizability is cross-checking. the predictive model's capacity to avoid overfitting. A learning function (or learning algorithm) f is often applied to the complete learning set to construct the final model to predict actual future states. There is no way to authenticate this final form. During the model-building process, cross-validation is used to evaluate how well the final model will perform with fresh data. Hence, only the training set and not the whole learning set must be used to derive the predictive characteristics. The alternative is admiration. Assuming that the learning set is first partitioned into validation sets and the learning set is then used to calculate the predictive characteristics practice sets. This indicates that the decision to employ certain predictive characteristics was made using data from validation groups. This indicates that the genuine forecast error is understated. CV is widely used to modify model parameters; for instance, in the k-classifier, the ideal number of nearest neighbors are nneighboursghbors. Altneighboursss-validation is used here repeatedly for various values, the value of k should be selected for k-fold validation. Given that the prediction error estimate in the 10-fold cross-validation is nearly unbiased, k = 10 would be a suitable option. nevertheless, for small sample datasets, verified performance measures are not accurate [28].

3.6 EVALUATION CRITERIA

It is necessary to calculate the performance rating of the classification algorithm to ensure that its model may be good or not. To evaluate the performance of a multiple classifiers by using several metrics, including[29].

3.7 CONFUSION MATRIX

The table used to illustrate the performance of the multiple classification models on the test data se, and generally shows us the various errors that our classifier makes[30].

a. Predicted values (PV): -The value guessed by the machine.

b. Actual values (AV): - The real value in the data.

c. True Positive (TP): - True Positive denotes the existence of both positive actual and expected values[6].

d. False Positive (FP): - False positives occur when a positive result is expected but a negative value occurs.

e. True Negative (TN): - The terms "true negative" and "predicted values" both refer to negative numbers.

f. False Negative (FN): - False negative refers to a number that is projected to be negative but is positive.

3.8 MEASURE OF PERFORMANCE.

3.8.1 Accuracy

is a metric for evaluating model erformance that counts the proportion of accurate predictions over all possible outcomes. The equation for calculating the correctness of a single class is as follows. [31]

$$Accuracy = \frac{(\mathrm{TP}) + (\mathrm{TP})}{(\mathrm{TP}) + (\mathrm{TN}) + (\mathrm{FP}) + (\mathrm{FN})} \quad (3.1)$$

3.8.2 Precision

Calculates the precision value by comparing the number of predicted positives to the number of true positives that the model asserts in the equation for one class. [32].

$$Precision = \frac{(\mathrm{TP})}{(\mathrm{TP}) + (\mathrm{FP})} \quad (3.2)$$

3.8.3 F1-Score

It is a weighted average of the precision and recall of a model, mean that measures the performance of the model and produces the best performance for incorrectly classified cases of accuracy[33].

$$F1 - scoe = \frac{2 * (\mathrm{TP})}{2 * (\mathrm{TP}) + (\mathrm{FP}) + (\mathrm{FN})} \quad (3.3)$$

3.8.4 Recall

It is the percentage rate of true positive for comparing the number of positives in the model to the total number of positives in the data[34].

$$Recall = \frac{(\mathrm{TP})}{(\mathrm{TP}) + (\mathrm{FN})} \tag{3.4}$$

4. RESULTS ANALYSIS

It was mentioned the dataset in the Data Analysis part was subjected to a significant number of machine learning algorithms. Using each of these techniques, the dataset was cross-validated ten times. ten-fold cross-validation after, the accuracy results converged. the performance of the classification algorithms model for solving problems will be measured using machine learning techniques. This includes two groups, the first group is to train the model on the data set 80%, and the second group is to test the classification model. The model's test 20% performance.

4.1 PERFORMANCE OF SUM MODEL

As the table below shows us the values for the metrics for precision, recall, and f1 score, from the results below we can see that the category wrong setup got the highest rates for the metrics mentioned while the DoS attack, malicious operation, spying, and data probing, these categories could not be detected by the algorithm as we got zero values. The worst results were for the scan category as we got low values for recall and f1 score metrics. Normal category detection was somehow acceptable regarding the metrics values. As shown in the table 4.1 below.

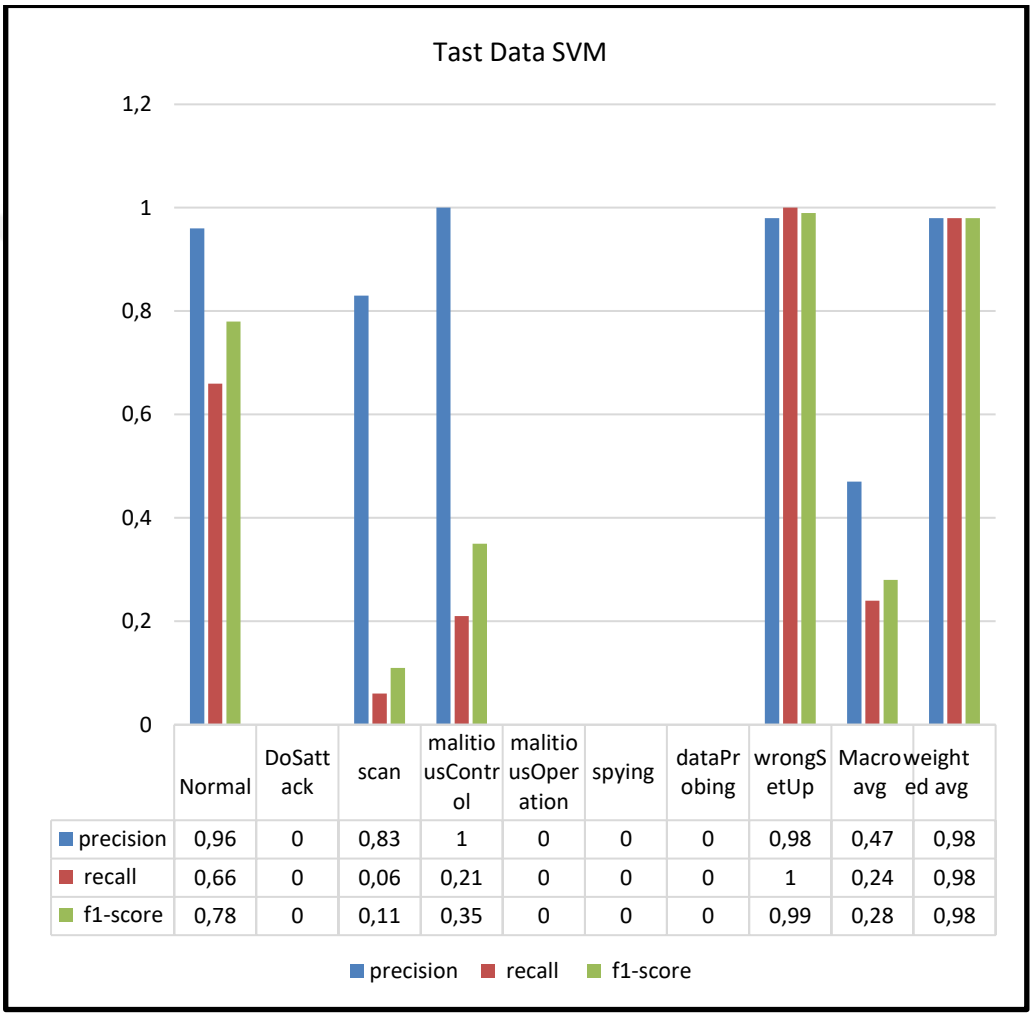
Table 4.1: The Model's Test Performance SVM For 10-Fold Cross Validation.

Type Attack	Precision	Recall	F1 score
Normal	0.96	0.66	0.78
DOS attack	0.00	0.00	0.00
scan	0.83	0.06	0.11
Malitious Control	1.00	0.21	0.35
Malitiouus operation	0.00	0.00	0.00
spying	0.00	0.00	0.00
Data Probing	0.00	0.00	0.00
Wrong setup	0.98	1.00	0.99

Table 4.1: The Model's Test Performance SVM For 10-Fold Cross Validation.

"Table Continued"

Macro avg	0.47	0.24	0.28
Weighted avg	0.98	0.98	0.98



	Normal	DoSattack	scan	malitiousControl	malitiousOperation	spying	dataProbing	wrongSetUp	Macro avg	weighted avg
precision	0,96	0	0,83	1	0	0	0	0,98	0,47	0,98
recall	0,66	0	0,06	0,21	0	0	0	1	0,24	0,98
f1-score	0,78	0	0,11	0,35	0	0	0	0,99	0,28	0,98

Figure 4.1: Evaluation Metrics Calculations for SVM Algorithm.

In figure 4.1 above diagram shows the relationship between the ratio of the evaluation performance measures and the classification categories in the SVM algorithm, and the scores of the evaluation performance measures are mentioned.

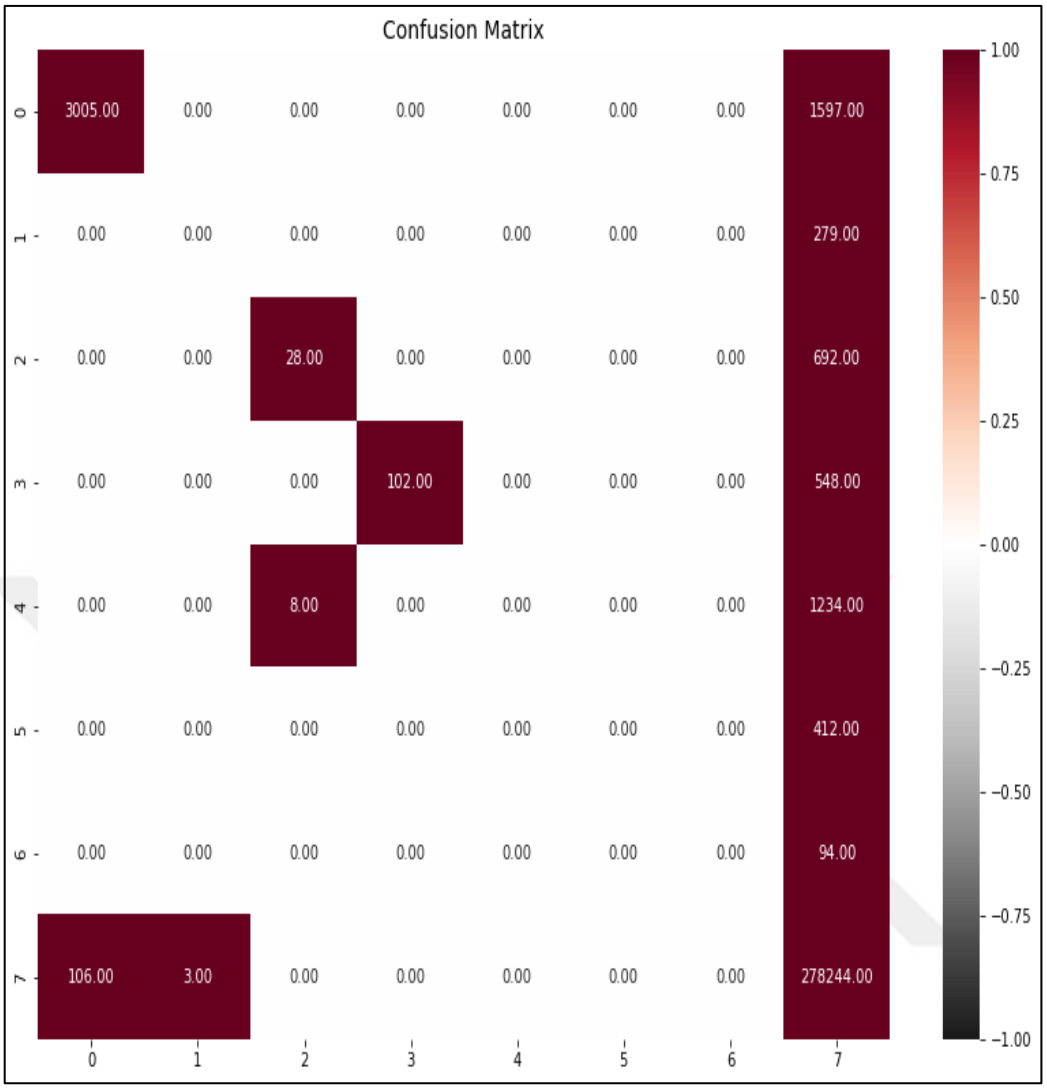


Figure 4.2: Confusion Matrix SVM Algorithm.

As shown in the figure 4.2 above, explain the confusion matrix. In the multiple confusion matrix, 3005 samples classified a category correctly, but in the classification, they made a mistake, with 1597 samples as a normal category. All the samples (checking data, spying, wrong numbers) were wrong on the supporting machine as normal, but according to him, he misclassified 692 as normal out of 720. For malicious processes, he misclassified 548 out of 650 samples. To illustrate, it classified eight as das out of 1234. Normality misclassified 106 samples as das out of 278,244. These classes with zero values could not be detected.

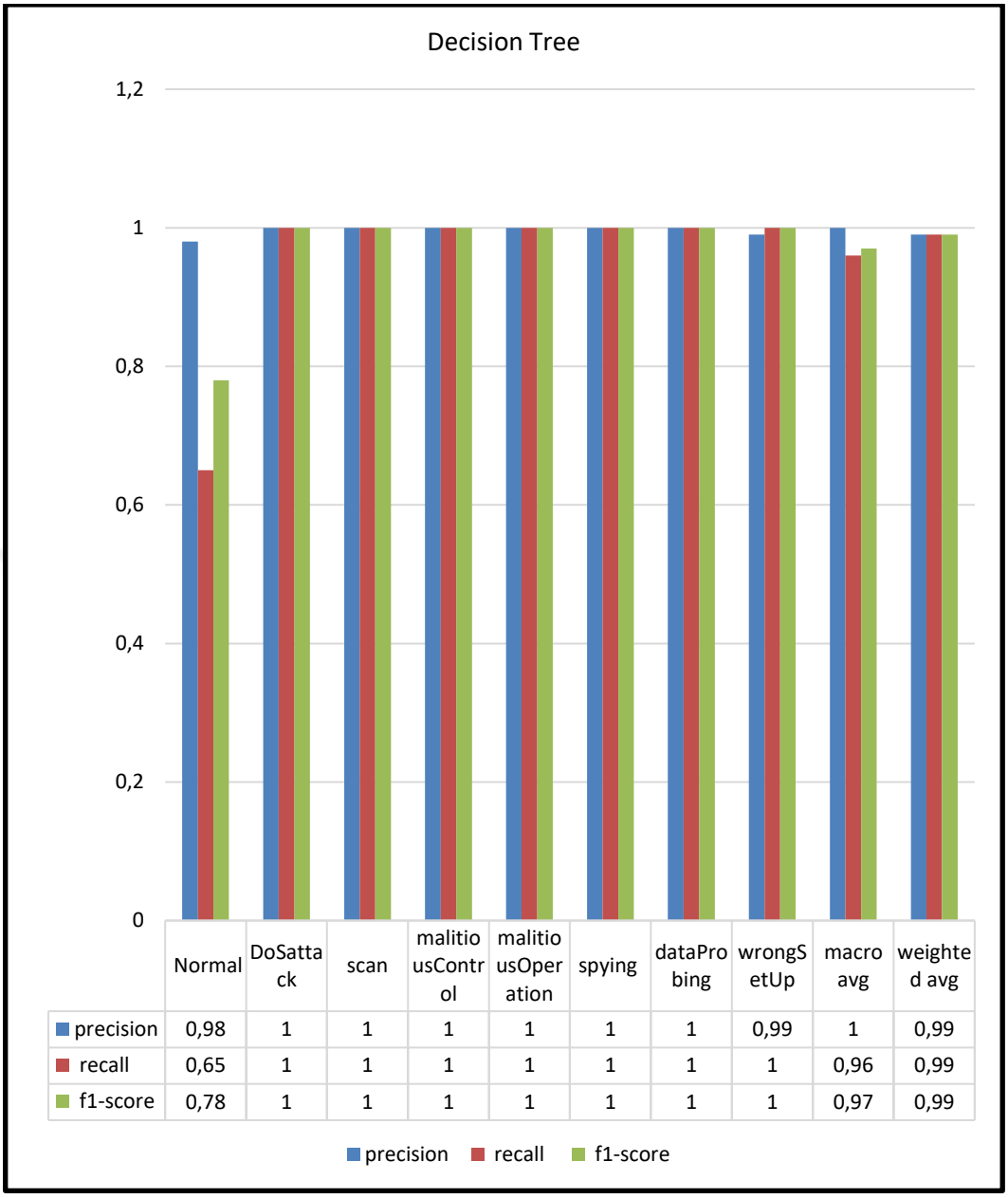
4.2 PERFORMANCE OF DECISION TREE MODEL

As the below shows us represents different evaluation metrics for techniques ML tested on the dataset. The DT had more deviations than other techniques in the values for the metrics

for precision, recall, and f1 score considering the decision tree algorithm, from the results below we can see that the category wrong setup, DoS attack, malicious operation, spying, scan, and data probing got the highest rates for the metrics to mention before and they are similar while the normal is the least among the others. Results of evaluation metrics calculations for the DT algorithm attack. As shown in the table 4.2 below.

Table 4.2: The Model's Test Performance DT for 10-Fold Cross Validation.

Type Attack	Precision	Recall	F1 score
Normal	0.98	0.66	0.79
DOS attack	1.00	1.00	1.00
Scan	1.00	1.00	1.00
Malitious control	1.00	1.00	1.00
Malitious operation	0.99	1.00	1.00
Spying	1.00	1.00	1.00
Data probing	1.00	1.00	1.00
Wrong setup	0.99	1.00	1.00
Macro avg	1.00	0.96	0.97
Weighted avg	0.99	0.99	0.99



	Normal	DoSattack	scan	maliciousControl	maliciousOperation	spying	dataProbing	wrongSetUp	macro avg	weighted avg
precision	0,98	1	1	1	1	1	1	0,99	1	0,99
recall	0,65	1	1	1	1	1	1	1	0,96	0,99
f1-score	0,78	1	1	1	1	1	1	1	0,97	0,99

Figure 4.3: Evaluation Metrics Calculations for DT.

As shown in the figure 4.3 above. the summary of the model, testing a decision tree algorithm for the model performance evaluation metrics. The above illustration shows the relationship between evaluation performance measures (precision, recall, f1-score) and rating categories (normal, DOS attack, scan, spying, scan, Malicious operation, Malicious control, Weighted avg, Macro avg, Wrong setup Data probing) in a decision tree algorithm.

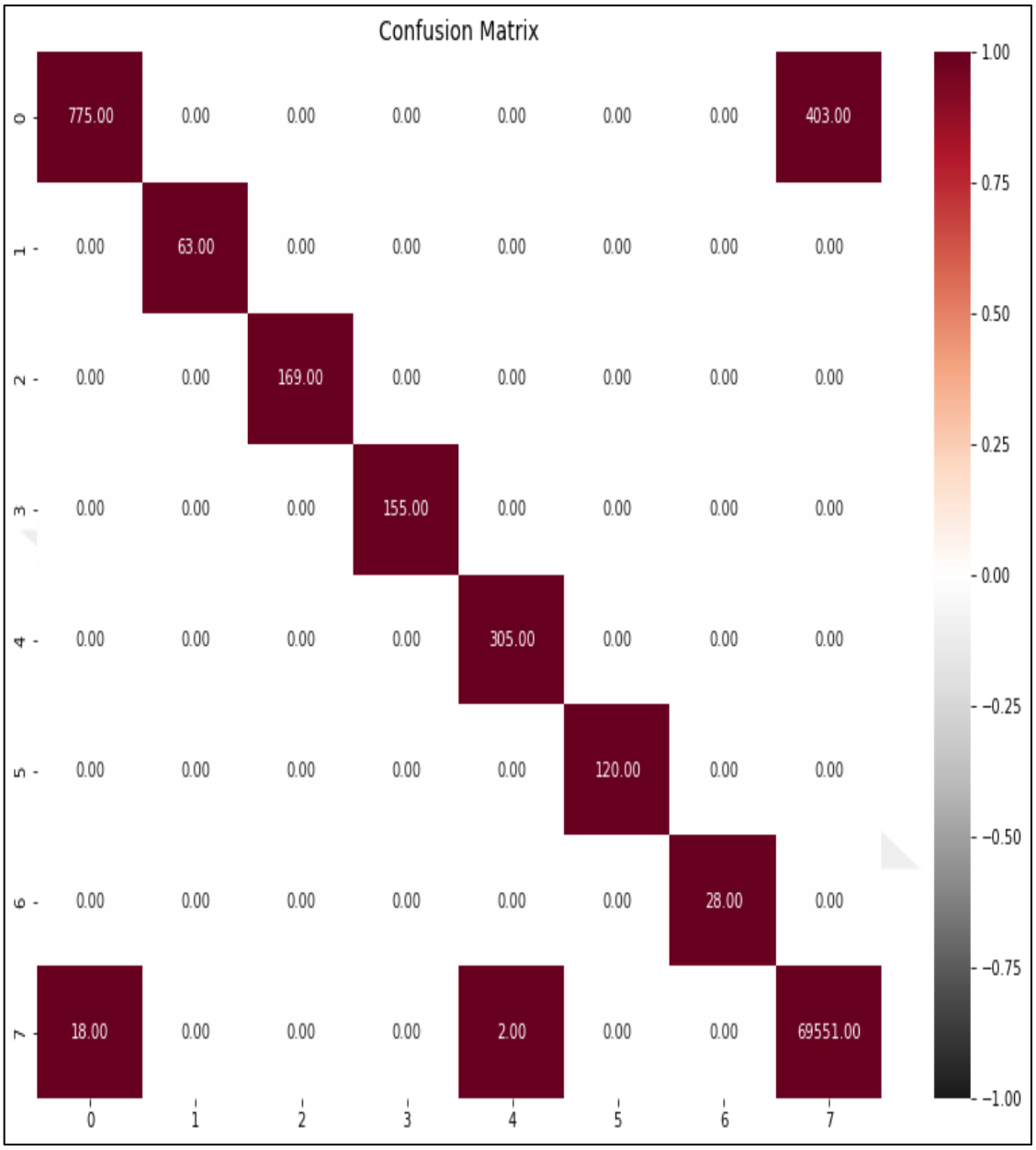


Figure 4.4: Confusion Matrix DT.

In the figure 4.4 above Confusion Matrix for A DT. In the Case Normal Category, out of 69,571 Samples, It misclassified 403 samples as normal while incorrectly classifying 18 samples as DOS and 2 samples as spying. The DT performed badly at initially during testing and had greater deviations than other approaches.

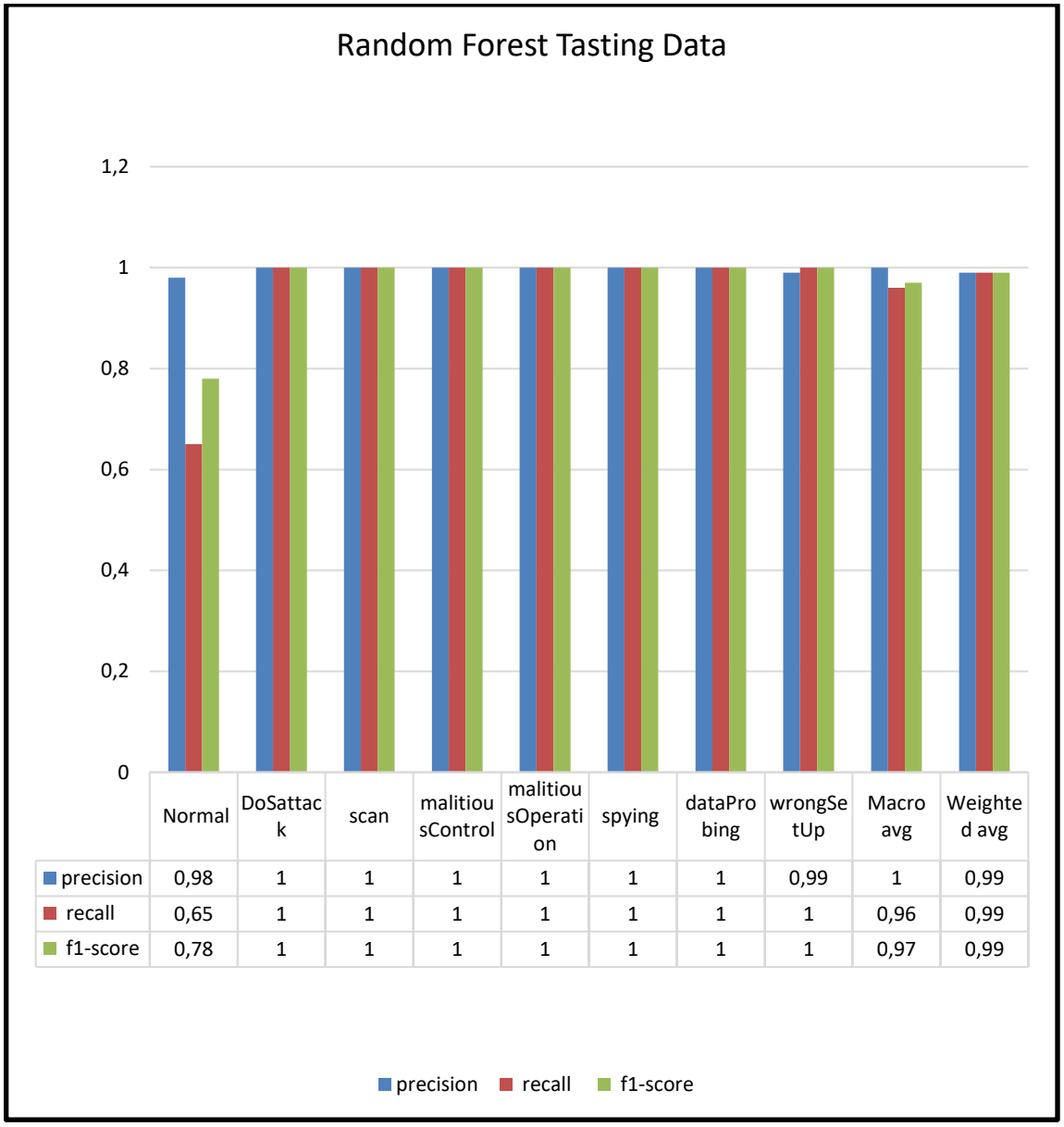
4.3 PERFORMANCE OF RANDOM FOREST MODEL

As the table below shows us the values for the metrics for precision, recall and f1 score considering the random forest algorithm, from the results below we can see that the category

wrong setup, DoS attack, malitious operation, spying, scan and data probing got the highest rates for the metrics mention before and they are similar and same as the results form decision tree in table which mean that these two algorithms have the same performance and detection pattern. As shown in the table 4.3 below.

Table 4.3: The Model's Test Performance RF for 10-Fold Cross Validation.

Type Attack	Precision	Recall	F1 score
Normal	0.98	0.66	0.79
DOS attack	1.00	1.00	1.00
Scan	1.00	1.00	1.00
Malitious control	1.00	1.00	1.00
Malitious operation	1.00	1.00	1.00
spying	1.00	1.00	1.00
Data probing	1.00	1.00	1.00
Wrong setup	0.99	1.00	1.00
Macro avg	1.00	0.96	0.97
Weighted avg	0.99	0.99	0.99



	Normal	DoSattack	scan	malitiousControl	malitiousOperation	spying	dataProbing	wrongSetUp	Macro avg	Weighted avg
precision	0,98	1	1	1	1	1	1	0,99	1	0,99
recall	0,65	1	1	1	1	1	1	1	0,96	0,99
f1-score	0,78	1	1	1	1	1	1	1	0,97	0,99

Figure4.5: Evaluation Metrics Calculations RF.

The above graph shows in figure 4.5. the representation of the evaluation metrics calculations for the RF algorithm and the relationship between the ratio of the rating performance metrics and the rating classes in the RF model. RF was performed with approximate similarity to DT have more accuracy, precision, recall, and F1 score values than other techniques.

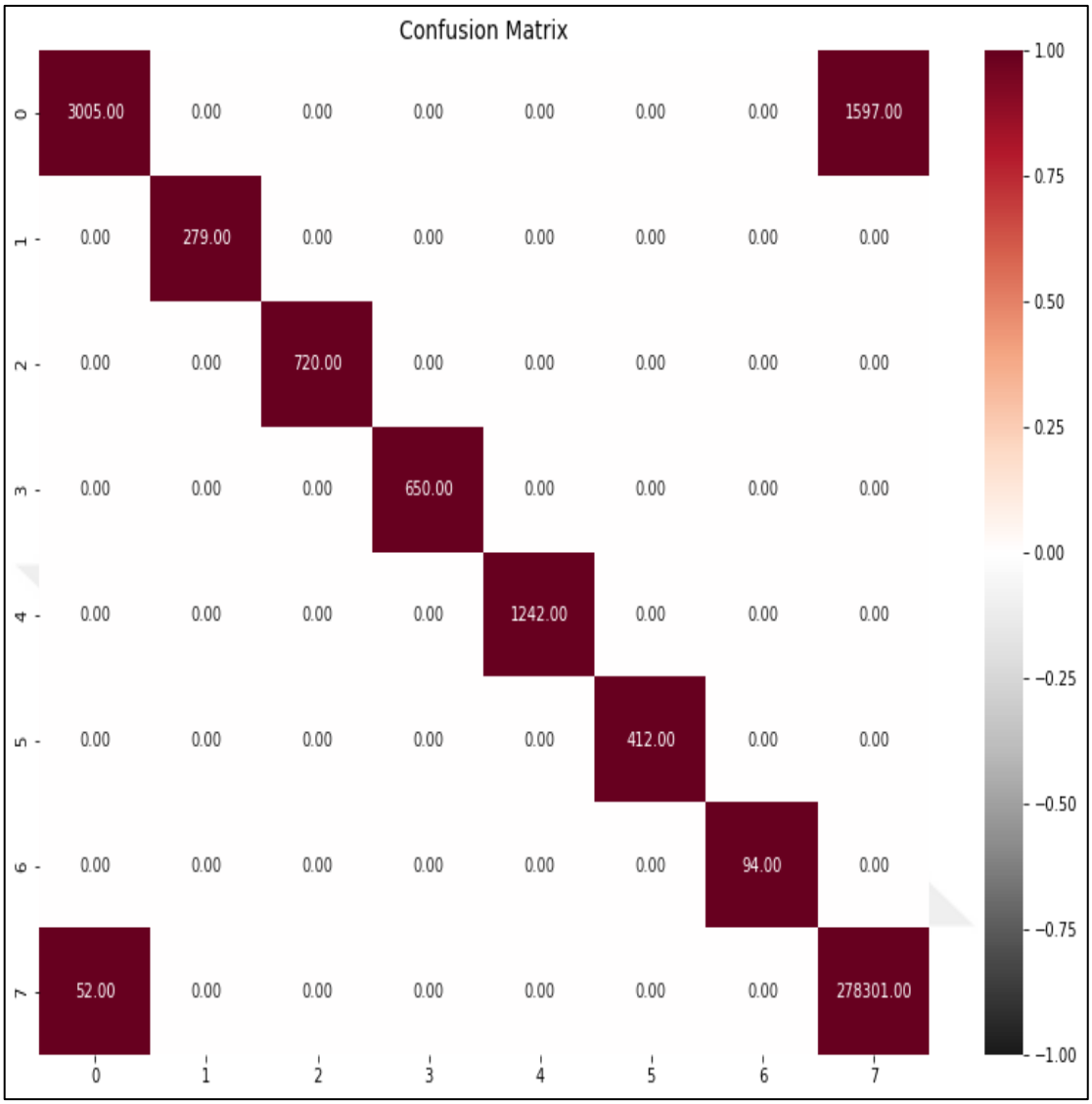


Figure 4.6: Confusion Matrix For RF.

As show above the figure 4.6 as show there are a number of similarities between the confusion matrix and the decision tree matrix. Random Forest has successfully recognized all classes except DoS and Normality. There was an imprecise classification of 1,597 of the 4,602 samples of individuals with DoS. Moreover, it wrongly classified 18 normal samples as DoS out of a total of 278,283.

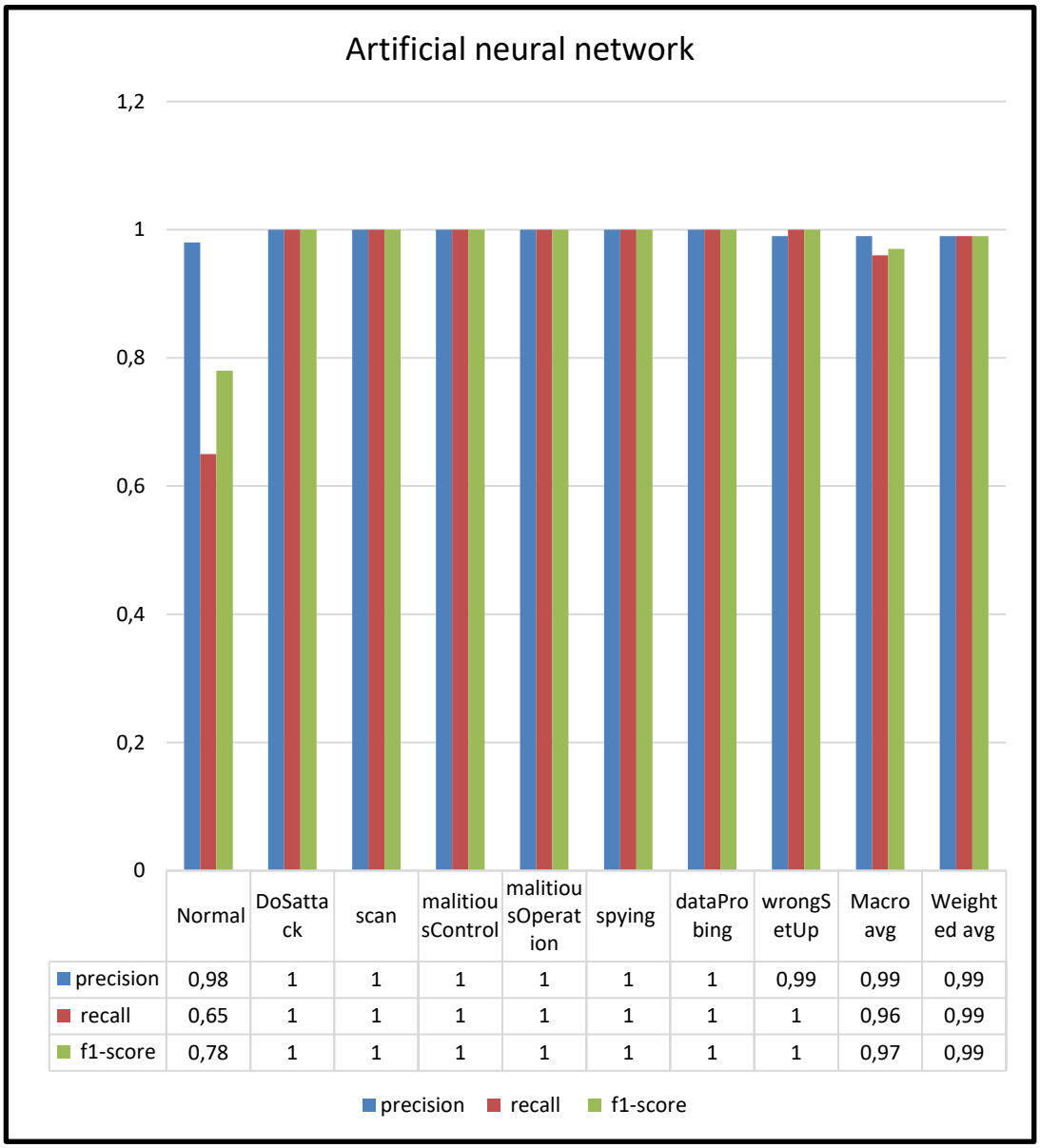
4.4 PERFORMANCE OF ARTIFICIAL NEURAL NETWORKS MODEL.

It was said that a number of machine learning algorithms were used on the dataset in the Data Analysis portion. The dataset was subjected to ten-fold cross-validation using each of these methods. After ten-fold cross-validation, the accuracy values converged. It may be

deduced from the cross-validation that Artificial Neural Networks have excelled in both training and testing accuracy. The pattern of abnormal activity detection can be very obvious in the dataset and highly detectable in ordinary machine learning and artificial neural networks. For normal activities, it can be seen that the algorithms have some difference in performance which means that there is a possibility that normal activities may be abnormal. As shown in the table 4.4 below.

Table 4.4: The Model's Test Performance ANN for 10-Fold Cross Validation.

Type Attack	Precision	Recall	F1 score
Normal	0.98	0.66	0.79
DOS attack	1.00	1.00	1.00
Scan	0.99	1.00	1.00
Malitious control	1.00	1.00	1.00
Malitious operation	1.00	1.00	1.00
spying	0.98	1.00	0.99
Data probing	1.00	1.00	1.00
Wrong setup	0.99	1.00	1.00
Macro avg	0.99	0.96	0.97
Weighted avg	0.99	0.99	0.99



	Normal	DoSattack	scan	malitiousControl	malitiousOperation	spying	dataProbing	wrongSetUp	Macro avg	Weighted avg
precision	0,98	1	1	1	1	1	1	0,99	0,99	0,99
recall	0,65	1	1	1	1	1	1	1	0,96	0,99
f1-score	0,78	1	1	1	1	1	1	1	0,97	0,99

Figure 4.7: Evaluation Metrics Calculations ANN.

In the figure 4.7 above, the scores for the performance measures are described by the classification categories in the ANN algorithm model. They are marked in red for precision, green for recall, and blue for f1-score.

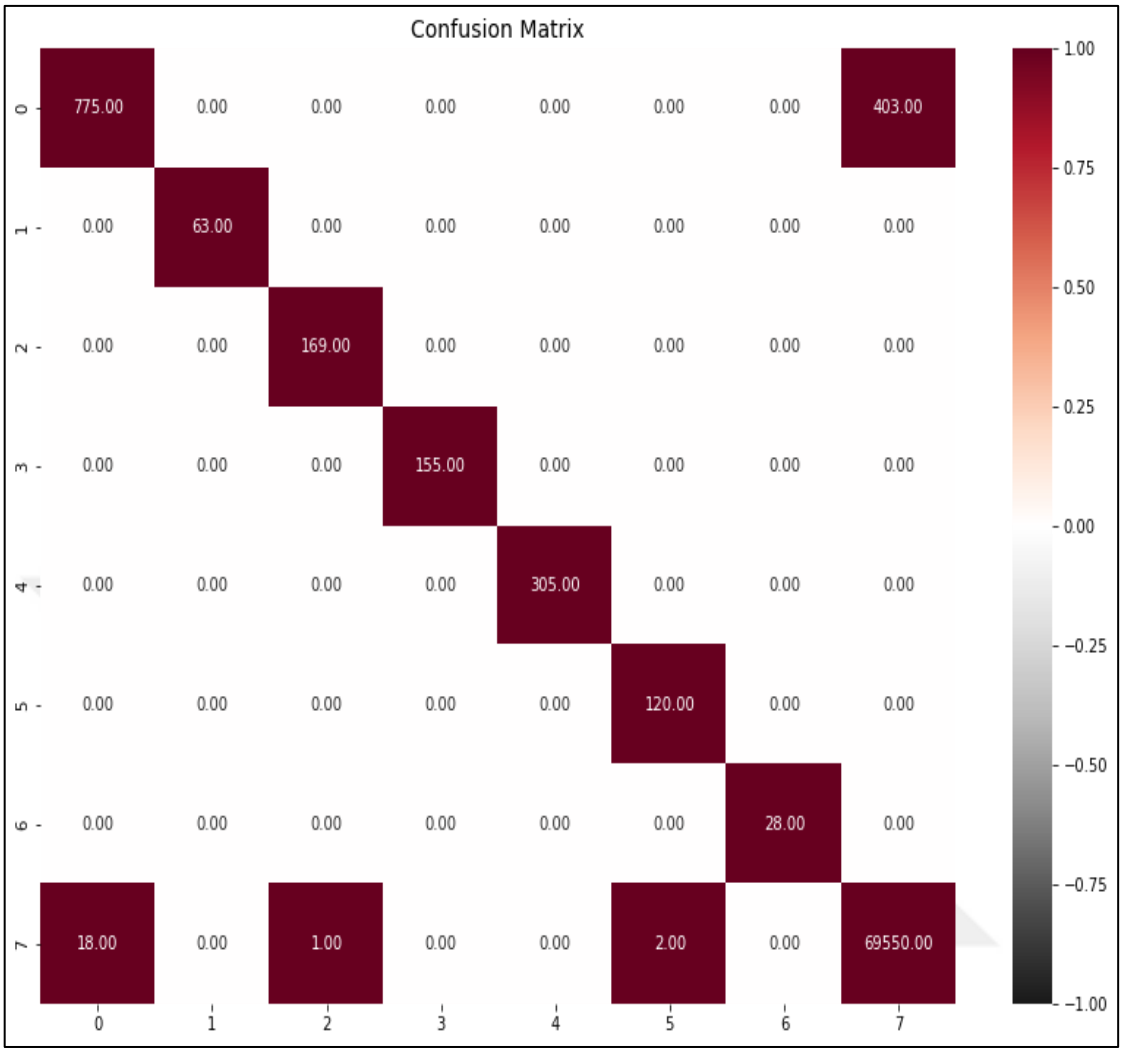


Figure 4.8: Confusion Matrix for ANN.

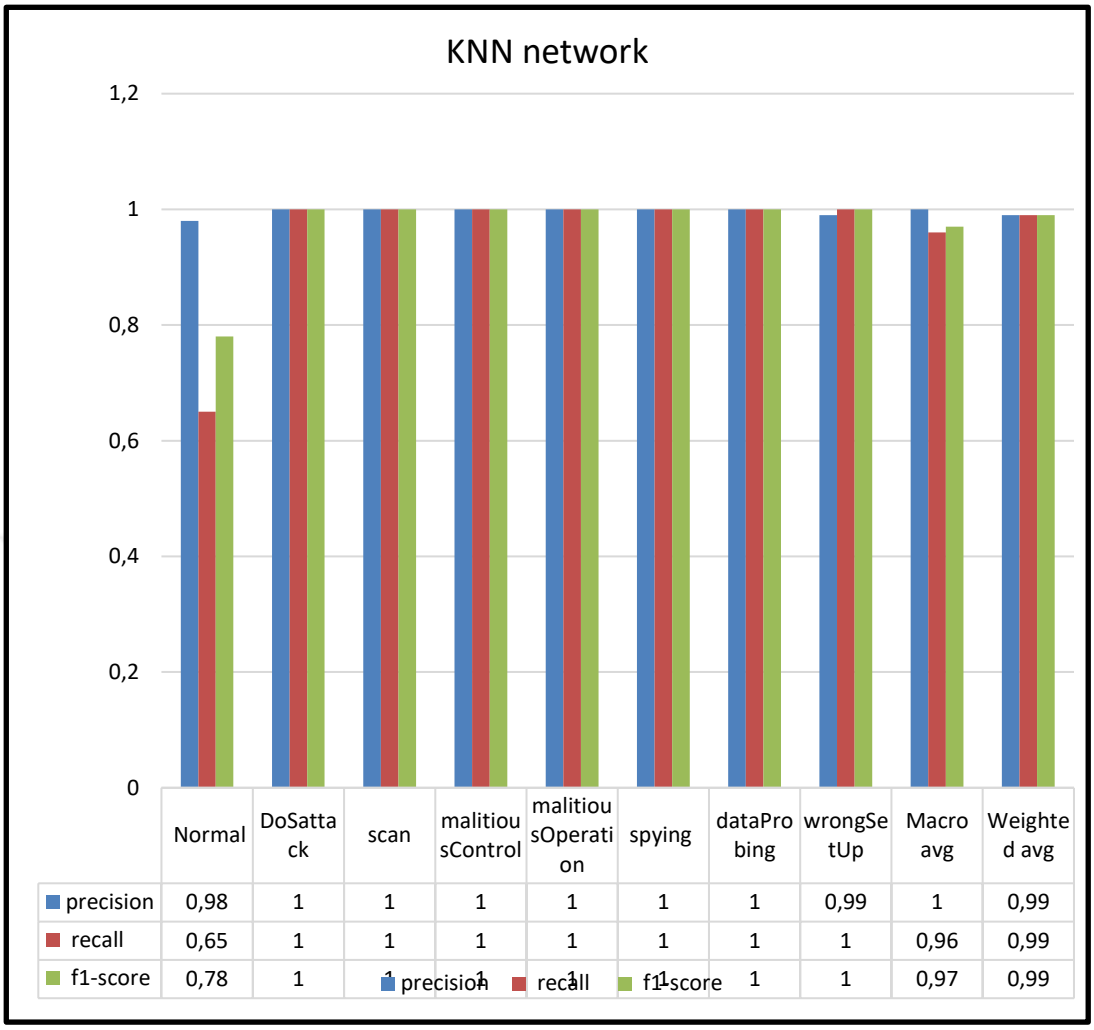
As show above the figure (4.8) There was little difference in the performance of the artificial neural network between it and the DT. Because he misclassified only one sample of the decision tree. The artificial neural network correctly predicted each sample of 5 labels out of 8 categories. In the case of DoS, the artificial neural network misclassified 403 samples as normal out of 1178 samples. While for normal labels, the artificial neural network misclassified 18 samples as DoS, 1 spy sample, and 1 sample as a malicious control sample out of 69,571 samples. Data discovery and misconfiguration attacks are close to zero, which means that this system cannot recognize these attacks. This is due to the lack of basic data used in testing the samples that preceded these attacks.

4.5 PERFORMANCE OF K-NEAREST NEIGHBORS MODEL.

It is found that the K- nearest neighbors' algorithm is the best one that is applied to these types of datasets because it is more accurately predicted due to having a k value of 5 through data investigation, espionage, malicious processes, malicious control, scanning, and misconfiguration attacks accurately compared to the other method. Five-fold cross-validation was performed on the dataset using each of these techniques. the accuracy results converged after ten-fold cross-validation. approximate similarity to Random Forest and Artificial Neural it has been described that several machine learning techniques were applied to the dataset. can follow the table 4.5 below.

Table 4.5: The Model's Test Performance KNN for 10-Fold Cross Validation.

Type Attack	**Precision**	**Recall**	**F1 score**
Normal	1.00	0.66	0.79
DOS attack	1.00	1.00	1.00
Scan	1.00	1.00	1.00
Malitious control	1.00	1.00	1.00
Malitious operation	1.00	1.00	1.00
Spying	1.00	1.00	1.00
Data probing	1.00	1.00	1.00
Wrong setup	0.99	1.00	1.00
Macro avg	1.00	0.96	0.97
Weighted avg	0.99	0.99	0.99



	Normal	DoSattack	scan	malitiousControl	malitiousOperation	spying	dataProbing	wrongSetUp	Macro avg	Weighted avg
precision	0,98	1	1	1	1	1	1	0,99	1	0,99
recall	0,65	1	1	1	1	1	1	1	0,96	0,99
f1-score	0,78	1	1	1	1	1	1	1	0,97	0,99

Figure 4.9: Evaluation Metrics Calculation for KNN.

In the figure 4.9 above, performance measures are described by many rating categories (precision, recall, f1-score) in the k-nearest neighbors algorithm model. It is a supervised classification based on the testing dataset since the model is from the distance of the data. The model is distributed from the distance of data, which means it is a supervised classification based on the testing dataset. By using this approach, properties can be separated into several classes to predict the probability of emergency repairs to a new sample. different evaluation metrics for different techniques trained on the dataset.

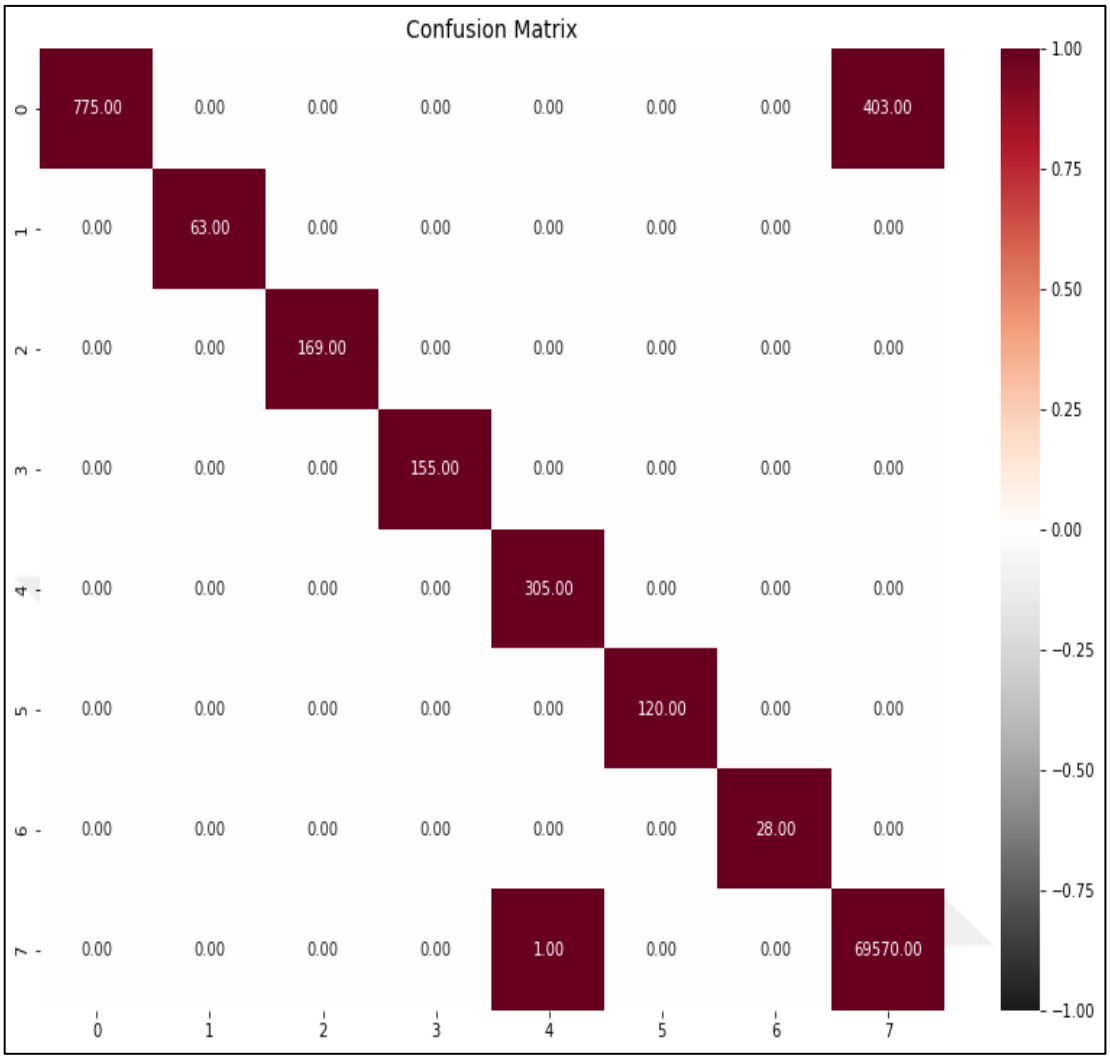


Figure 4.10: Confusion Matrix for KNN

As show above the figure (4.10) k nearest neighbour's has categorized each category correctly Except for the DoS and Normality categories. Out of 4602 samples from DoS, misclassified 403 samples were as normal. Moreover, the Normal category misclassified 52 samples as DoS out of 278,353 samples. Data discovery and misconfiguration attacks are close to zero, which means that this system cannot recognize these attacks. This is due to the lack of basic data used in testing the samples that preceded these attacks.

4.6 PERFORMANCE COMPARISON

In the table below, a comparison of five algorithms (SVM, DT, RF, ANN, KNN) and several metrics (Accuracy, Precision, recall, f1-score) is shown. We note that the KNN algorithm, RF and ANN are the best performance algorithms for this type of task according to the results

shown in the table, where it became 99.4% in this algorithm and the performance of the SVM algorithm was less. As shown in the figure (4.11) below.

	SVM	DT	RF	ANN	KNN
ACCURECY	0,9827	0,9941	0,9941	0,994	0,9943
precision	0,96	0,98	0,98	0,98	1
recall	0,66	0,66	0,66	0,66	0,65
f1-score	0,78	0,79	0,79	0,79	0,79

Figure 4.11: Performance Comparison for All Algorithms Used.

5. CONCLUSIONS

According to what was said in the first chapter, this thesis has four main objectives. Priority one is to show that detectable abnormalities may be found in network data using these technologies. The second objective is to put some of the suggested methods to use in a machine learning framework that can be put to use for comparing algorithms and for detecting anomalies in real time, even on a computer with low resources. The machine learning system will be built on top of this foundation. Ultimately, a real-world dataset will be used to put the answer to the test. Lastly, we'll look at how time has had a role in the provided data sets.

Many methods for spotting anomalies were presented, such as statistical and machine learning-based structures. A simple statistical technique based on a box plot was provided, and a machine learning algorithm using k-means and Random Forest was also suggested. It was expected that both of these algorithms would run quickly and provide respectable outcomes. These multi-instance anomaly detection techniques were selected because their performance was equivalent to that of other algorithms. More so, it was determined to put these algorithms into action.

In this thesis, discusses the comparison of classification algorithms on the analysis of a non-parallel data set using DS2OS described above with several metrics to evaluate the prediction of attacks and anomalies in the IoT environment. We note that the performance of the four classification algorithms was the same result and almost the same performance, but the best performance was the algorithm KNN in the Internet of Things environment. While the SVM classification algorithm had a lower result 98.2% and performance was lower in prediction compared to the rest of the other algorithms.

As for future work, there might be better solutions or approaches to detect anomalous attacks between the services of IoT to be able to develop and improve better and more accurately. we intend to apply a similar approach To a much larger categorical data set for the Internet of Things, another technique can be used with another data set. Finally, we intend to test the method in a real-world IoT environment to detect attacks and anomalies.

REFERENCES

[1] A. Nascita, F. Cerasuolo, D. Di Monda, J. T. A. Garcia, A. Montieri, and A. Pescape, "Machine and Deep Learning Approaches for IoT Attack Classification," *INFOCOM WKSHPS 2022 - IEEE Conf. Comput. Commun. Work.*, no. May, 2022, doi: 10.1109/INFOCOMWKSHPS54753.2022.9797971.

[2] K. Lakshmanna *et al.*, "A Review on Deep Learning Techniques for IoT Data," *Electron.*, vol. 11, no. 10, pp. 1–23, 2022, doi: 10.3390/electronics11101604.

[3] H. Agarwal, A. Bhat, and K. N. Venkat, "Attack and Anomaly Detection in IoT Sensors and".

[4] S. Manimurugan, S. Al-Mutairi, M. M. Aborokbah, N. Chilamkurti, S. Ganesan, and R. Patan, "Effective attack detection in internet of medical things smart environment using a deep belief neural network," *IEEE Access*, vol. 8, pp. 77396–77404, 2020, doi: 10.1109/ACCESS.2020.2986013.

[5] O. F.Y, A. J.E.T, A. O, H. J. O, O. O, and A. J, "Supervised Machine Learning Algorithms: Classification and Comparison," *Int. J. Comput. Trends Technol.*, vol. 48, no. 3, pp. 128–138, 2017, doi: 10.14445/22312803/ijctt-v48p126.

[6] M. A. Alsoufi *et al.*, "Anomaly-based intrusion detection systems in iot using deep learning: A systematic literature review," *Appl. Sci.*, vol. 11, no. 18, 2021, doi: 10.3390/app11188383.

[7] PETER HANNA ERIK SWARTLING, "Anomaly Detection in Time Series Data using Unsupervised Machine Learning Methods," no. November 2018, 2020, doi: 10.13140/RG.2.2.15967.07849.

[8] D. R. Thamaraiselvi and S. Anitha Selva Mary, "Attack and Anomaly Detection in IoT Networks using Machine Learning," *Int. J. Comput. Sci. Mob. Comput.*, vol. 9, no. 10, pp. 95–103, 2020, doi: 10.47760/ijcsmc.2020.v09i10.012.

[9] M. H. Aysa, A. A. Ibrahim, and A. H. Mohammed, "IoT Ddos Attack Detection Using Machine Learning," *4th Int. Symp. Multidiscip. Stud. Innov. Technol. ISMSIT 2020 - Proc.*, no. November, 2020, doi: 10.1109/ISMSIT50672.2020.9254703.

[10] M. Hasan, M. M. Islam, M. I. I. Zarif, and M. M. A. Hashem, "Attack and anomaly detection in IoT sensors in IoT sites using machine learning approaches," *Internet of Things (Netherlands)*, vol. 7, p. 100059, 2019, doi: 10.1016/j.iot.2019.100059.

[11] K. Wehbi, L. Hong, T. Al-Salah, and A. A. Bhutta, "A Survey on Machine Learning Based Detection on DDoS Attacks for IoT Systems," *Conf. Proc. - IEEE SOUTHEASTCON*, vol. 2019-April, no. 1, 2019,

[12] M. Alauthman, N. Aslam, M. Al-kasassbeh, S. Khan, A. Al-Qerem, and K. K. Raymond Choo, “An efficient reinforcement learning-based Botnet detection approach,” *J. Netw. Comput. Appl.*, vol. 150, no. October 2019, p. 102479, 2020, doi: 10.1016/j.jnca.2019.102479.

[13] V. Kotu and B. Deshpande, “Anomaly Detection,” *Data Sci.*, vol. 49, no. 8, pp. 447–465, 2019, doi: 10.1016/b978-0-12-814761-0.00013-7.

[14] S. Ahmed, Y. Lee, S. H. Hyun, and I. Koo, “Unsupervised Machine Learning-Based Detection of Covert Data Integrity Assault in Smart Grid Networks Utilizing Isolation Forest,” *IEEE Trans. Inf. Forensics Secur.*, vol. 14, no. 10, pp. 2765–2777, 2019, doi: 10.1109/TIFS.2019.2902822.

[15] S. Rathore and J. H. Park, “Semi-supervised learning based distributed attack detection framework for IoT,” *Appl. Soft Comput. J.*, vol. 72, pp. 79–89, 2018, doi: 10.1016/j.asoc.2018.05.049.

[16] B. Weinger, J. Kim, A. Sim, M. Nakashima, N. Moustafa, and K. J. Wu, “Enhancing IoT anomaly detection performance for federated learning,” *Digit. Commun. Networks*, vol. 8, no. 3, pp. 314–323, 2022, doi: 10.1016/j.dcan.2022.02.007.

[17] H. H. Pajouh, R. Javidan, R. Khayami, A. Dehghantanha, and K. K. R. Choo, “A Two-Layer Dimension Reduction and Two-Tier Classification Model for Anomaly-Based Intrusion Detection in IoT Backbone Networks,” *IEEE Trans. Emerg. Top. Comput.*, vol. 7, no. 2, pp. 314–323, 2019, doi: 10.1109/TETC.2016.2633228.

[18] A. Ukil, S. Bandyoapdhyay, C. Puri, and A. Pal, “IoT healthcare analytics: The importance of anomaly detection,” *Proc. - Int. Conf. Adv. Inf. Netw. Appl. AINA*, vol. 2016-May, pp. 994–997, 2016, doi: 10.1109/AINA.2016.158.

[19] O. B. J. Rabie, P. K. Balachandran, M. Khojah, and S. Selvarajan, “A Proficient ZESO-DRKFC Model for Smart Grid SCADA Security,” *Electron.*, vol. 11, no. 24, 2022, doi: 10.3390/electronics11244144.

[20] F.-X. Aubet, “Machine Learning-Based Adaptive Anomaly Detection in Smart Spaces,” no. January, 2018, doi: 10.13140/RG.2.2.35293.26088.

[21] D. S. A. M Vinitha, Dr D.V.V.Krishna Prasad, Dr.R.Krishnamoorthy, Dr.V.Balajivijayan , Sowmiya, Dr.R.Thiagarajan, “Cyber Attack Detection on IOT Using Network Traffic Mechanism by Neural Network Predictive Approach,” *Eur. J. Mol. & Clin. Med.*, vol. 07, no. 10, pp. 3690–3698, 2020.

[22] K. Kostas, “Anomaly Detection in Networks Using Machine Learning Kahraman Kostas A thesis submitted for the degree of Master of Science in Computer Networks and Security Supervisor : Dr . Martin Reed School of Computer Science and Electronic Engineering University of,” no. October, pp. 1–71, 2018.

[23] A. Yazdinejad, B. Zolfaghari, A. Dehghantanha, H. Karimipour, G. Srivastava, and R. M. Parizi, "Accurate threat hunting in industrial internet of things edge devices," *Digit. Commun. Networks*, 2022, doi: 10.1016/j.dcan.2022.09.010.

[24] M. Labonne, "Anomaly-based network intrusion detection using Maxime Labonne To cite this version : HAL Id : tel-02988296 Anomaly-Based Network Intrusion Detection Using Machine Learning," 2020.

[25] I. Idrissi, M. Boukabous, M. Azizi, O. Moussaoui, and H. El Fadili, "Toward a deep learning-based intrusion detection system for iot against botnet attacks," *IAES Int. J. Artif. Intell.*, vol. 10, no. 1, pp. 110–120, 2021, doi: 10.11591/ijai.v10.i1.pp110-120.

[26] V. Gaur and R. Kumar, "Analysis of Machine Learning Classifiers for Early Detection of DDoS Attacks on IoT Devices," *Arab. J. Sci. Eng.*, vol. 47, no. 2, pp. 1353–1374, 2022, doi: 10.1007/s13369-021-05947-3.

[27] T. Submitted, T. Degree, and R. Rafi, "MACHINE LEARNING BASED ATTACK DETECTION IN," 2021.

[28] D. Berrar, "Cross-validation," *Encycl. Bioinforma. Comput. Biol. ABC Bioinforma.*, vol. 1–3, no. January 2018, pp. 542–545, 2018, doi: 10.1016/B978-0-12-809633-8.20349-X.

[29] R. Harada *et al.*, "Quick Suppression of DDoS Attacks by Frame Priority Control in IoT Backhaul With Construction of Mirai-Based Attacks," *IEEE Access*, vol. 10, pp. 22392–22399, 2022, doi: 10.1109/ACCESS.2022.3153067.

[30] N. Mishra and S. Pandya, "Internet of Things Applications, Security Challenges, Attacks, Intrusion Detection, and Future Visions: A Systematic Review," *IEEE Access*, vol. 9, pp. 59353–59377, 2021, doi: 10.1109/ACCESS.2021.3073408.

[31] V. Timčenko and S. Gajin, "Machine Learning based Network Anomaly Detection for IoT environments," *ICIST-2018 Conf.*, vol. 8, no. 4, pp. 542–548, 2019.

[32] I. Ullah and Q. H. Mahmoud, "Design and Development of RNN Anomaly Detection Model for IoT Networks," *IEEE Access*, vol. 10, pp. 62722–62750, 2022, doi: 10.1109/ACCESS.2022.3176317.

[33] M. Savic *et al.*, "Deep Learning Anomaly Detection for Cellular IoT with Applications in Smart Logistics," *IEEE Access*, vol. 9, pp. 59406–59419, 2021, doi: 10.1109/ACCESS.2021.3072916.

[34] C. W. Tien, T. Y. Huang, P. C. Chen, and J. H. Wang, "Using autoencoders for anomaly detection and transfer learning in iot," *Computers*, vol. 10, no. 7, pp. 1–14, 2021, doi: 10.3390/computers10070088.