Application of Machine Learning for Cybersecurity Threat Detection and Analysis

Report to be submitted in partial fulfillment of the requirements for the degree

Of

Master of Technology in Industrial Engineering & Management

By

Ankit Patel (21IM60R22)

Under the guidance of

Prof. Sayak Roychowdhury

DEPARTMENT OF INDUSTRIAL & SYSTEMS ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR
APRIL 2023

INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR
Department of Industrial and Systems Engineering
Indian Institute of Technology, Kharagpur
India – 721302

CERTIFICATE

This is to certify that we have examined the thesis entitled **Application of Machine Learning for Cybersecurity Threat Detection and Analysis**, submitted by **Ankit Patel** (Roll Number: **21IM60R22**) a postgraduate student of **Department of Industrial and Systems Engineering** in partial fulfilment for the award of degree of Master of Technology (MTech). We hereby accord our approval of it as a study carried out and presented in a manner required for its acceptance in partial fulfilment for the Post Graduate Degree for which it has been submitted. The thesis has fulfilled all the requirements as per the regulations of the Institute and has reached the standard needed for submission.

Prof. Sayak Roychowdhury
MTech Project Supervisor
Department of Industrial and Systems Engineering
Indian Institute of Technology, Kharagpur

Place: Kharagpur
Date: 26 / April / 2023

INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR
Department of Industrial and Systems Engineering
Indian Institute of Technology, Kharagpur
India – 721302

DECLARATION

I certify that

1. The work contained in this report has been done by me under the guidance of my supervisor.

2. The work is original and has not been submitted to any other Institute for any degree or diploma.

3. I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.

4. I have given due credit to other sources by citing them in the thesis´s text and providing their details in the references whenever I have used materials (data, figures, and text). Further, I have taken permission from the copyright owners of the sources whenever necessary.

Ankit Patel
21IM60R22
Department of Industrial and Systems Engineering
Indian Institute of Technology, Kharagpur

Place: Kharagpur
Date: 26 / April / 2023

ACKNOWLEDGEMENTS

I express my sincere gratitude to Prof. Sayak Roychowdhury, Assistant Professor, Industrial and Systems Engineering Department, Indian Institute of Technology, Kharagpur for his supervision and guidance during the project. His valuable advises, encouragement, suggestions and friendly attitude towards me, during this work helped a long way in bringing this project report to this stage. I am also thankful to him for extending all kind of help and for providing the necessary facilities required during this work.

I am very much thankful to Prof. Jhareswar Maiti, Head, Department of Industrial and Systems Engineering, IIT Kharagpur for providing necessary facilities during the project work.

I am thankful to Faculty Advisor Prof. Sri Krishna Kumar and all the faculty members of the department for their valuable suggestions, which helped me improve on this work.

Ankit Patel

IIT Kharagpur

Date: 26 / April / 2023

List of Figures

List of Tables

ABSTRACT

The importance of machine learning is on the rise within the field of cybersecurity. The main justification for employing machine learning in the realm of cybersecurity is to augment the effectiveness, scalability, and feasibility of malware identification, as opposed to traditional methods that depend on human involvement. Efficient and systematic handling of machine learning issues holds great significance in the field of cybersecurity. Statistical and machine learning methodologies, including logistic regression, k-nearest neighbours (KNN), and deep learning, have exhibited effectiveness in addressing cyber-attacks. The discernment of fundamental patterns and insights within network data and the establishment of a machine learning model predicated on this data is of paramount significance in the advancement of sophisticated security systems. The utilisation of machine learning methodologies for the purpose of addressing contemporary cybersecurity risks has been emphasised.

The utilisation of data has become an essential component for the efficient operation of modern society. The utilisation of data science solutions, such as decision support and consumer behaviour model, holds the capability to transform large volumes of data into actionable insights. Presently, modern methodologies for data cleansing predominantly centre on addressing one or two quality-related apprehensions, notwithstanding the presence of a plethora of additional challenges. The effectiveness of these methods is dependent on the availability of reference data, instructional data, or user feedback during the data refinement stage. The employed techniques for rectifying errors may rely on either impeccable master data or annotated training data. The aim of this investigation is to assess the efficacy of machine learning techniques in addressing the proliferation of malware, which poses an increasingly significant threat to the security of the internet.

Keywords: Cybersecurity, Cyber-attack, Data Quality, Machine Learning, Deep Learning, Feature Selection, Threat Detection

Contents

Chapter 1

Introduction

1.1 Introduction

Cybersecurity prevents malicious cyber activity and safeguards critical infrastructure. Cybersecurity practises safeguard computer networks and their applications from malicious intrusion. Inconsistent technology and a lack of in-house expertise can increase the cost of a security system. However, businesses with a comprehensive cybersecurity plan that is governed by best practises and automated using advanced analytics, AI, and machine learning can combat cyberthreats more effectively and reduce breach lifecycles and impacts. As internet use grows, so are cyberattacks. Global cybersecurity talent shortages are well known. Machine learning can detect advanced threats including organisation profiling, infrastructure vulnerabilities, and interdependent vulnerabilities and exploits.

Business data management requires high data quality. Good data quality can help a company improve its strategies and identify patterns. Data Quality ensures data suitability. It describes a dataset's usefulness and capacity to be handled and analysed. Data quality affects machine learning system performance. High operational costs, low customer satisfaction, and inaccurate decision-making have come from poor data [1]. Today, there are more data with unknown quality to examine and use for the organisation.

High data quality is data that is suitable for use and can meet user requirements [1–4]. This definition showed that data quality depends on its context, consumer needs, usability, and accessibility. Now we must ask: Organisational data quality: how good? Answering this question requires data quality metrics. This thesis describes how to identify data content using useful data quality criteria and applies ML/DL algorithms to cybersecurity challenges.

1.2 Motivation

The motivation for selecting "Application of Machine Learning for Cybersecurity Threat Detection and Analysis" as the topic for an MTech thesis is multifaceted. To begin, with an increase in the frequency of cyber-attacks on organisations, governments, and individuals in recent years, cybersecurity has become a critical concern. Traditional cybersecurity approaches, such as rule-based and signature-based systems, have difficulties in identifying new and sophisticated attacks. As a result, more complex approaches, such as machine learning, are required to detect and prevent cyber threats.

Second, machine learning has demonstrated significant potential in a variety of areas, including cybersecurity. Machine learning algorithms can find trends and abnormalities in massive datasets, which can help predict possible risks. They can also adapt to new and changing attack patterns, which is critical for cybersecurity because attackers' strategies are continuously evolving.

Finally, research in the field of machine learning for cybersecurity is still in its early stages, with plenty of space for experimentation and creativity. By selecting this topic for an MTech thesis, one can help to develop new approaches for detecting and preventing cyber-attacks, which will have a big impact on the field of cybersecurity.

1.3 Objectives and Scope of Research Work

After identifying problems, concerns, and questions related to using machine learning to cybersecurity threat detection and analysis, the following research objectives are established:

1. The present study aims to employ machine learning and deep learning techniques for the purpose of analysing and comprehending cybersecurity datasets. Comprehending the various types of cyber-attacks, their characteristic patterns, and the requisite algorithms and methodologies for their classification and detection is imperative.

2. To highlight crucial dataset elements for accurate cyber-attack classification: This research also identifies the most critical cyber-attack classification features. Features will be selected and engineered from the dataset. A machine learning model that effectively classifies cyber assaults based on these features would create a precise and reliable detection system.

3. Improve data quality to boost machine learning performance: Machine learning model performance depends on data quality. This research addresses missing values, outliers, and noise to improve data quality. To ensure the machine learning model uses high-quality data and improves performance, data cleaning, pre-processing, and augmentation will be used.

This study analyses and detects cyberattacks using machine learning and deep learning. Network traffic, log files, and malware will be analysed. The inquiry will use machine learning and deep learning methods. Data quality and feature selection are improved to improve machine learning system performance. This work improves practical cybersecurity threat detection and analysis.

1.4 Outline of the Thesis

The thesis is split up into its four individual segments. The following is a summary of each chapter in order:

Chapter-1, named as **‘Introduction’** in this chapter, we will discuss what cybersecurity is and why it is so crucial in the modern digital world. The study's rationale, as well as its goals and scope, will be discussed. Methodology and an outline of the study's framework will also be included in this section.

Chapter-2, named as **'Literature Review'** in this chapter we will provide a comprehensive analysis of the research that has previously been conducted on cybersecurity, potential risks associated with cybersecurity, and preventative strategies. The significance of data analysis in relation to both data integrity and cybersecurity will also be covered in this chapter.

Chapter-3, named as **'Methodology'** in this chapter we will provide a full discussion of the study process, including a description of the problem and dataset. The UNSW-NB 15, BoT-IoT, and DDoS Evaluation datasets, which were used in the study, will be described in detail. The chapter will also talk about judging the quality of the data, pre-processing the data, exploratory data analysis, selecting features, and the machine learning and deep learning models used in the study. Some of the models that will be looked at are Logistic Regression (LR), K-Nearest Neighbour (KNN), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and Artificial Neural Network (ANN).

Chapter-4, named as **'Results & Discussions'** in this chapter the research findings will be presented, and then a thorough analysis of those results will be given. In addition, a conclusion that is founded on the findings of the research will be presented, as well as some suggestions for further work in the subject of cybersecurity.

The concluding section of the thesis will comprise the Bibliography, encompassing a comprehensive enumeration of all the sources that have been cited throughout the thesis.

Chapter 2

Literature Review

2.1 Cybersecurity

Information and communication technology (ICT) has advanced significantly over the past 50 years and has become a necessary part of daily life. Security policymakers have recently placed a high priority on safeguarding ICT systems and applications against cyberattacks [18]. Protecting information and communication technology (ICT) infrastructure from potential online threats and attacks is a key component of cybersecurity [5]. A wide range of topics are covered by the field of cybersecurity, including the steps required to safeguard information and communication technologies, the processing and transmission of data and the information contained therein, as well as the virtual and physical system components involved. The measures implemented directly contributed to the level of security attained. Reference 19 supports this. According to Craigen et al. (20), "Cybersecurity" refers to a group of methods, guidelines, and regulations used to protect data, software applications, and computer networks against unauthorised access, damage, or exploitation. A set of practises and technologies are used to protect information systems, networks, programmes, and data from destruction, abuse, and unauthorised access, according to reference [6].

2.2 Cybersecurity Threats

Threats, or who is attacking, Vulnerabilities, or where the assault is weak, and Impacts, or what the attack does, make up the trifecta of security concerns that make up the risks involved with every attack [5]. A security incident is any event that compromises the safety of sensitive data, hardware, or software. There are several different kinds of cyberattacks that can compromise the safety of a company's data systems or an individual's personal information [7].

2.3 Cybersecurity Defensive Measures

Information, computer systems, and networks need to be protected from cyber-attacks, thus defensive measures are essential. Specifically, their job is to keep an eye out for intrusions, which may be defined as "any kind of unauthorised activity that causes damage to an information system," [21] and stop them before they cause any harm. An IDS is "a device or software application that monitors a computer network or systems for malicious activity or policy violations" [22]. Anti-virus, firewall, user authentication, access control, data encryption, and cryptography systems are all examples of well-known security solutions, however they may not be adequate for the modern needs of the cyber business.

2.4 Analysing Cybersecurity Information

The presence of data is crucial to the advancement of data science, according to scholarly sources [23]. The discipline of cybersecurity data science relies on datasets, which are exhaustive compilations of information records that frequently include multiple attributes or characteristics and associated pieces of data. Consequently, it is essential to comprehend the structure of cybersecurity data, which includes a variety of cyberattack categories and pertinent attributes. The justification behind this approach is that a security model that is based on data can be formulated through the examination of unprocessed security data sourced from pertinent cyber channels, with the aim of detecting trends in security incidents or malicious operations. There exists a plethora of cybersecurity-related datasets, encompassing analyses of intrusion, virus, anomaly, fraud, and spam, among others.

2.5 Data Integrity

As the globe progresses, data quality has emerged as a top concern for all businesses [8]. Because useful data is available in several formats, organisations must come up with coordinated and imaginative methods to handle issues with data quality [11]. Multiple places provide mediocre information. System failures, poor data migration procedures, data entry mistakes by staff and customers, and other outside factors all contribute [12]. Data can be defined as the "primary basis of information that represents real-world objects in a format that can be saved, retrieved, processed by a software procedure, and communicated across a network" [13].

Since the characteristics of flawed data have not been demonstrated, error-free information is accepted without question [14]. Even if the flawed information is corrected, it may still be wrong in some circumstances. Incomplete data, accurate data, ambiguous data, and unclear data are all examples of incorrect data. Incomplete data should not be used for decision making, regardless of whether they are transformed to real data.

The value of accurate data in businesses has risen as a result of users' easy access to it [15]. Poor data quality results in lost time, income, unhappy customers, and damaged reputations for brands [12]. Some major contributors to poor data quality [16] include the following:

- Numerous data.
- The convenience of making copies, storing, and sharing information online.
- The meteoric rise in available information delivery channels like radio, television, print media, online resources, electronic mail, mobile phones, and RSS feeds.
- The amount of data that has been recorded is increasing.
- There are not any straightforward methods for rapidly cleaning, inspecting, and assessing data sources.

The success of today's businesses depends on having high quality data. One of the main causes of poor data quality, among many others, is the existence of filthy data in data sources. An understanding of the types and scope of filthy data inside data sources is given by the review of four existing research studies from the literature that are related to identifying dirty data that affect data quality. Data cleaning processes that can monitor, evaluate, and maintain the quality of data are highly advised in order to assure high quality data in an organisation by properly cleaning the filthy data that already exists in data sources.

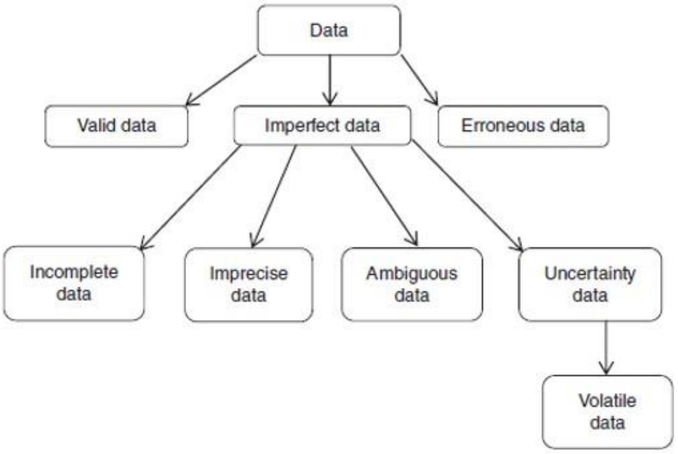

Figure 1: Types of data by quality [14]

Table 1 presents a summary of the applications of deep learning and other machine learning methodologies in addressing the most critical cybersecurity challenges of the present era. Consequently, it can be inferred that the use of machine learning and deep learning techniques, as well as their variants or ensembles, modified lightweight methodologies, or recently proposed algorithms, may have a significant impact on facilitating the achievement of our objective in the field of security analytics.

Table 1: An overview of Machine learning tasks for cybersecurity

Author and Year	Objective	Technique
Chandrasekhar, *et al.* (2014)	IDS Network Construction and Modelling	ANN and SVM
Almiani, *et al.* (2019), Yin, *et al.* (2017), Alrawashdeh, *et al.* (2016) and Kim, *et al.* (2016)	Attack and intrusion anomaly detection and classification	Deep Learning Recurrent, RNN, LSTM

Alrashdi, *et al.* (2019) and Chang, *et al.* (2017)	Cyber-anomaly detection	RF
Doshi, *et al.* (2018)	Monitoring for Denial-of-Service Attacks	
Mohamed, *et al.* (2018) and Resende, *et al.* (2018)	System for detecting intrusions	
Jo, *et al.* (2015)		Decision Tree
Shapoorifard, *et al.* (2017) and Vishwakarma, *et al.* (2017)		KNN
Bapat, *et al.* (2018) and Prokofiev, *et al.* (2018)	Threatening Botnet Detection	
Jaganathan, *et al.* (2015)	Cyberattack effect forecasting	LR
Meng, *et al.* (2015)	False alarm rate reduction	
Hoang, *et al.* (2018)	Managing complex security information	PCA
Hagos, *et al.* (2017)		Regression Regularization
Raman, *et al.* (2019), Pervez, *et al.* (2014), Li, *et al.* (2012) and Yan, *et al.* (2010)	Choosing protective measures, identifying threats, and labelling them	SVM
Kotpalliwar, *et al.* (2015)	Cyber-attacks classified as DoS, U2R, R2L, and Probing	

Machine learning techniques are extensively employed for the identification of cybersecurity vulnerabilities within the realm of cybersecurity. In recent years, machine learning techniques have been effectively utilised in various application domains within the realm of cybersecurity to address a multitude of issues. This technology can detect intrusions, malware, spam, anomalies, fraud, zero-day attacks, cyberbullying, Internet of Things attacks, and threat analysis.

Chapter 3

Methodology

3.1 Problem Description

The exponential growth of technology and the widespread adoption of the internet have given rise to cybersecurity risks. The escalation of cybersecurity threats has surfaced as a significant concern in the era of digitalization. The system's security measures are persistently targeted by malicious actors with the intention of compromising data integrity or disrupting system functionality. The conventional tools of cyber defence, namely antivirus software, firewalls, and intrusion detection systems, are inadequate in the face of contemporary security threats.

Recent advances in machine learning and deep learning offer promising options for detecting and preventing cyber-attacks. However, the efficiency of these methods relies on the accuracy of the training data. Unreliable forecasts and incorrect classification of cyberattacks due to incomplete or noisy data can compromise data security.

This project aims to create a reliable cybersecurity threat detection and analysis system that uses machine learning to identify and classify intrusions reliably and efficiently. Machine learning and deep learning will assess and understand cybersecurity datasets. Critical dataset features are needed to classify cyberattacks accurately and reliably. The study also improves machine learning framework training data. The objectives of this study are intended to facilitate the development of more resilient cybersecurity tools for the digital realm.

3.2 Dataset Description

The following is a description of the datasets that were utilised to write this thesis:

3.2.1 UNSW-NB 15 Dataset

The UNSW-NB15 dataset can be used to test the efficacy of network-based NIDS. The dataset is an annotated one, and it contains over 2.5 million samples of network activity. Real-world network data was collected, which included both benign and malicious intrusion attempts across a variety of operating systems, applications, and user actions.

Basic aspects of network connections, such as length and number of bytes, are among the 49 features in the dataset, which are divided into 5 feature classes. Informational characteristics like the volume of flows, Finally, the Attack features that categorise the traffic as normal or one of the 10 attack types, based on characteristics such as the standard deviation of packet sizes. DoS, Probing, Remote-to-Local, User-to-Root, and Other Types of Attacks.

The dataset is frequently used in academic research on intrusion detection systems and cybersecurity; it was created by academics at the University of New South Wales.

Data Source: https://research.unsw.edu.au/projects/unsw-nb15-dataset

3.2.2 BoT-IoT Dataset

The BoT-IoT dataset is a freely available resource for conducting security analysis on Internet of Things (IoT) gadgets. Data from both safe and risky Internet of Things (IoT) device network activity is included in the dataset. Information from 83 distinct Internet of Things gadgets in a smart home setting is included. In a lab setting, we were able to capture data on benign traffic, while simulating various assaults on the devices yielded data on malicious traffic. There is a total of 28.5 GB in the dataset, which is the result of 5 days' worth of network activity. The provided files are in the PCAP format and contain a mixture of plaintext and encrypted network traffic. This dataset can test machine learning and deep learning methods for identifying and classifying IoT-based cyberattacks.

Data Source: https://research.unsw.edu.au/projects/bot-iot-dataset

3.2.3 DDoS Evaluation Dataset

The DDoS Evaluation Dataset is publicly available to evaluate DDoS detection and mitigation solutions. Data was collected in a lab using real and simulated user traffic. The high-speed network dataset includes UDP flood, TCP SYN flood, and HTTP flood attacks. The dataset consists of a training and test set. There are a total of 4,227 network flows in the training set, with 3,840 in the test set. Researchers and practitioners can use the dataset to test and refine their DDoS detection and mitigation strategies.

Data Source: https://www.unb.ca/cic/datasets/ddos-2019.html

3.3 Methodology

This thesis examines how machine learning and deep learning can identify and assess cyber threats in cybersecurity. To improve machine learning, we need better data. Machine learning and deep learning will be used to analyse and comprehend varied cybersecurity datasets. The thesis helps design real-time cybersecurity systems that can detect and neutralise novel attacks. This paper shows how to apply these strategies and how machine learning can improve organisational cybersecurity.

The thesis aims to provide recommendations and suggestions for the implementation of machine learning-based cybersecurity systems, focusing on best practises for data collection, processing, and model selection. We have addressed various data quality challenges that arise due to multiple causes, without necessitating users' active involvement or presuming the existence of master or training data. Based on our computations, the density of data offers valuable insights that may inform the process of data cleansing. Our initial focus was on exploring various techniques for addressing a range of data quality concerns, such as those arising from the presence of duplicate entries, inconsistent data, inaccurate data, and outdated data. The primary objective of the thesis is to contribute towards enhancing digital system and network security, as well as to facilitate the advancement of cybersecurity research and application. The methodology employed in this thesis is depicted in Figure 2:

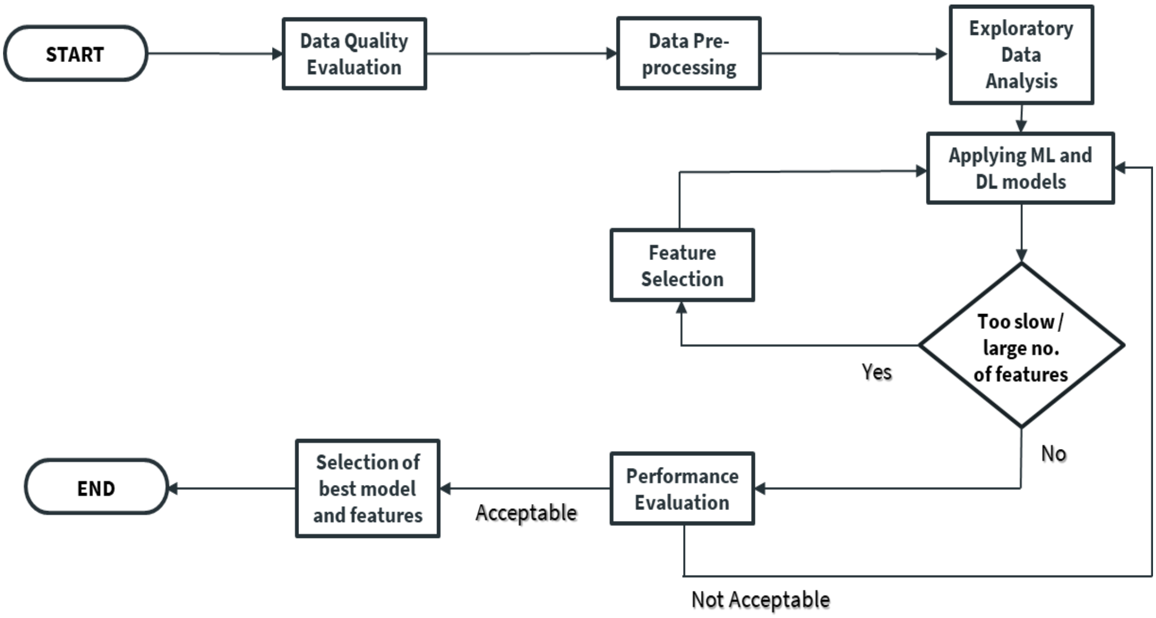
START
Data Quality Evaluation
Data Pre-processing
Exploratory Data Analysis
Applying ML and DL models
Too slow / large no. of features
Yes
Feature Selection
No
Performance Evaluation
Acceptable
Selection of best model and features
END
Not Acceptable

Figure 2: Methodology

3.3.1 Data Quality Evaluation

In practise, data is never complete, accurate, or error-free. Errors such as omissions, inconsistencies, and duplications are inherent to data and cannot be avoided. Understanding data quality would help many businesses organise their data. Table 2 shows data quality evaluation dimensions:

Accuracy	Is there nothing wrong with the data?
Completeness	How in-depth is this data?
Reliability	Does the data conflict with other trusted sources?
Relevance	Is this information necessary?

Timeliness	How current is the data? Is it feasible to use for instantaneous reporting?
Validity	Is the data available in a practical format, does it conform to recognised commercial standards, or is it non-existent?
Uniqueness	Is this the only occurrence of this data in the system?

Table 2: Data Quality Evaluation Dimensions [24]

3.3.2 Data Pre-processing

Machine learning and data analysis rely heavily on the preparation of data. Pre-processing entails modifying data so that it can be read and processed by machine learning programmes. The goal of data pre-processing is to enhance the model's functionality by eliminating noise, standardising the data, and decreasing the number of dimensions it occupies. Data pre-processing steps can include:

1. **Data Cleaning:** Eliminating or assuming missing values, fixing mistakes, weeding out outliers, and dealing with duplication.
2. **Data Transformation:** Methods for reducing dimensionality include feature selection and feature extraction, variable transformations including scaling and normalisation, and the conversion of categorical to numeric variables.
3. **Data Integration:** Bringing together disparate data sets while guaranteeing uniformity of data type and meaning across sources.
4. **Data Reduction:** Smaller datasets can be achieved through data cleaning, data aggregation, and sampling methods.
5. **Imbalanced Data:** One definition of "imbalanced data" is a dataset in which one or more classes has disproportionately fewer samples than the rest. Models trained with machine learning that are optimised for accuracy may suffer as a result. Low accuracy and excessive false negatives for the minority class may result in imbalanced datasets since the classifier may tend to predict the majority class.

Many real-world situations, like those involving fraud detection, medical diagnosis, and anomaly identification, include the presence of imbalanced data. As a result, accurate identification of instances from the minority class (the one with fewer samples) may be of utmost importance. Imbalanced data can be addressed using many methods, including:

i **Resampling:** To achieve this, either the majority or the minority group must be under-represented in the sample.

ii **Cost-sensitive learning:** This includes revising the misclassification costs for each group to account for the inequity between them.

iii **Ensemble methods:** The goal is to boost performance by merging multiple models or classifiers.

iv **Synthetic data generation:** To do this, artificial examples of the minority class must be generated.

v **Algorithm-specific techniques:** Weighting the instances or employing specialised loss functions are two ways in which some algorithms are pre-designed to deal with unbalanced data.

The specifics of the dataset and the issue at hand will determine which method is most suitable for resolving unbalanced data.

Overall, data pre-processing's major purpose is to get the data in shape so that the machine learning algorithm can learn and produce correct predictions.

3.3.3 Exploratory Data Analysis

EDA, or exploratory data analysis, is a technique for quickly understanding and summarising the key aspects of a dataset by the application of visual tools. Learning to recognise trends and patterns in the data and spotting outliers are all part of this process. Summary statistics, histograms, scatter plots, box plots, and correlation matrices are all examples of statistical and visual methods that can be used in EDA.

The primary objectives of EDA are:

1. Learn how to recognise and interpret the most important aspects of a dataset, such as its structure, variables, and relationships.
2. Find associations and correlations in the data that were previously hidden.
3. Find any data anomalies or outliers that might need checking or cleaning.
4. Identify the factors that contribute most to the data's variability.
5. Help researchers decide which statistical tools and methods to use.

EDA helps assist data analysis and modelling by revealing the data's structure and trends.

3.3.4 Feature Selection

Predictive modelling uses feature selection to reduce input variables. The goal is to minimise the number of inputs. The reduction of input variables can lead to a decrease in computational expenses associated with modelling, and under certain circumstances, it can enhance model performance. Machine learning employs a range of feature selection strategies, which may include but are not limited to:

- **Correlation Coefficient -** When comparing the linearity of relationships between multiple variables, correlation is the gold standard. One variable can be predicted from another using correlation. Good variables should have a high correlation with the target; hence this is the reasoning behind utilising correlation for feature selection. In addition, variables should have a positive correlation with the result but no relationship with other factors.

- **Information Gain -** Information gain measures how much entropy is lost when a dataset is transformed. It can be used for feature selection by comparing each candidate variable's Information gain against that of the target variable.

- **Wrapper Methods -** Wrappers must select the best feature subset and train and test a classifier on it. We base feature selection on a machine learning technique to improve a dataset. The utilisation of a greedy search technique involves testing all possible permutations of characteristics against the evaluation criterion. Predictive accuracy is typically higher when using wrapper techniques rather than filter methods.

  1. **Forward Feature Selection -** This is an iterative process where we first compare the successful features to our ideal set of characteristics. We then choose a second variable whose performance improves when combined with the first. This procedure is repeated until the desired result has been attained.

  2. **Backward Feature Elimination -** In contrast to Forward Feature Selection, this technique selects features backwards. In this case, we begin by constructing a model using all the data at our disposal. Then, we take the best-valued evaluation measure variable from the model. This procedure is repeated until the target is reached.

- **LASSO Regularization (L1) -** To prevent the machine learning model from becoming over-fit, regularisation can be applied to the model's parameters in the form of a penalty. The penalty is calculated by multiplying each predictor by its respective coefficient in a linear model. Lasso or L1 regularisation can reduce some coefficients to zero, which distinguishes it from other methods of regularisation. Thus, it is unnecessary to keep that aspect of the model.

3.3.5 ML and DL models

Both supervised and unsupervised models exist for machine learning. There are two types of supervised models: regression models and classification models. Here are some of the machine learning and deep learning models I used:

- Logistic Regression (LR)
- K-Nearest Neighbors (KNN)
- Linear Discriminant Analysis (LDA)
- Quadratic Discriminant Analysis (QDA)
- Artificial Neural Network (ANN)

3.3.5.1 Logistic Regression (LR)

A categorical (binary or multi-class) dependent variable can be analysed in connection to one or more independent variables using a statistical model known as Logistic Regression (LR). A logistic function models the dependent variable as a function of the independent factors and provides a score that indicates the likelihood of the dependent variable falling into a certain category. Unlike multinomial logistic regression, in which the dependent variable can have more than two categories, binary logistic regression only allows for two.

Logistic regression is a prominent machine learning method for classifying samples. A supervised learning algorithm, it uses previously tagged data to "learn" how to understand and predict future data. The model learns input feature weights to maximise label likelihood. Logistic regression's many benefits stem from its adaptability to either categorical or continuous input features, as well as its ease of use and interpretability. However, it makes the unrealistic assumption that there is a linear connection between the input features and the log-odds of the output variable. Also, it presumes that the observations are unrelated to one another, which is not always the case, especially with time series data.

3.3.5.2 K-Nearest Neighbor (KNN)

KNN is a non-parametric machine learning algorithm for classification and regression. The KNN algorithm predicts a new event by finding its k-nearest neighbours in the training set and using the majority class for classification or averaging their values for regression. Here is how the KNN algorithm functions:

1. Compare the new instance to the instances in the training set, using a distance measure such the Euclidean distance to establish how different it is.
2. Find the newly created instance's k nearest neighbours.
3. The outcome of a classification using k-nearest neighbours is the class with the largest percentage of neighbours. The result of using k-nearest neighbours for regression is a weighted average of their individual values.

Tuning the value of k, a hyperparameter, improves performance. Overfitting can occur with a small value of k, whereas underfitting can occur with a big number.

K-Nearest Neighbour (KNN) algorithm can be a suitable approach for datasets with limited observations and non-linear patterns, owing to its user-friendly and intuitive architecture. However, in the case of extensive datasets, the computational cost may be high, and the performance may be adversely affected if the dataset comprises redundant features.

3.3.5.3 Linear Discriminant Analysis (LDA)

LDA is used for supervised classification in machine learning and pattern recognition. This statistical method finds the best linear combination of features for classifying a dataset. In order to preserve as much class discriminating as feasible, LDA uses a linear transformation to map the original high-dimensional dataset onto a lower-dimensional space. The discriminant functions are used to categorise new samples into the lower-dimensional space, which is known as the LDA subspace.

LDA presumes that classes have equivalent covariance matrices and that the data follows a multivariate normal distribution. Its goal is to reduce the spread within each class while increasing it between them. How well the classes are separated from one another is quantified by the between-class scatter, whereas the amount by which samples within a class vary from one another is quantified by the within-class scatter.

After the LDA subspace is built, it may be used to project fresh samples onto it and use the discriminant functions to assign classes to those examples. Face recognition, text classification, and bioinformatics are just some of the many fields where LDA has found success.

3.3.5.4 Quadratic Discriminant Analysis (QDA)

Within the larger group of discriminant analysis techniques is the classification algorithm known as Quadratic Discriminant Analysis (QDA). QDA, like LDA, makes assumptions about the data's distribution, but it also allows for a quadratic decision boundary, which can be more adaptable.

For each training class, QDA determines its own mean and covariance matrix of predictors. These numbers are what we plug into a posterior probability model to estimate the likelihood of each class for a given data point, given the known predictors. Then, we use the highest posterior probability to assign the new data point to a category.

When the line of demarcation between classes is not straight, QDA can do well. Overfitting is possible with little training data since it requires estimating a covariance matrix for each class independently.

3.3.5.5 Artificial Neural Network (ANN)

ANNs, a machine learning model, are inspired by the human brain. Nodes—artificial neurons—communicate to analyse non-linear data. ANNs are effective in speech, image, NLP, and economic forecasting. ANNs have input layers, hidden processing layers, and output layers. Neurons (input nodes) in each layer are fixed. The network processes data at the input layer before sending it to the hidden layers. Hidden layers process signals by weighting and biassing input data. Output layer outputs processed input signals. Backpropagation trains ANN models. To reduce the difference between predicted and observed outputs, neurons' weights and biases are adjusted during training. Repeat this until the error is low enough to train the model.

ANN can learn alone and supervised. Supervised learning uses an existing dataset to predict an outcome for unlabelled input. Unsupervised learning finds data structures and patterns without labels. ANN can learn complex non-linear relationships, tolerate faults, and scale to new data sources. ANN requires a lot of training data and computer power.

Section 3.3 describes how conventional and deep learning models are used to analyse and categorise cybersecurity threat data. Pre-processing and exploratory analysis followed data quality assessment. Identifying critical features for accurate and reliable classification required feature selection approaches. Logistic regression, K-nearest neighbours, linear and quadratic discriminant analysis, and artificial neural networks were tested. This technology uses machine learning and deep learning to analyse cybersecurity datasets and classify risks.

Chapter 4

Results & Discussions

4.1 Results

I evaluated the three datasets. The data was found to have a large number of duplicate entries. So, it was required to eliminate duplicates in order to obtain objective results for the data analysis. There was also a variety of undesirable data. So, my initial objective was to exclude all the undesirable data before applying any model to the data. After deduplication, data must be pre-processed before analysis and prediction model construction. Pareto chart illustrating undesirable data in the datasets is shown in the Figure 3 below:

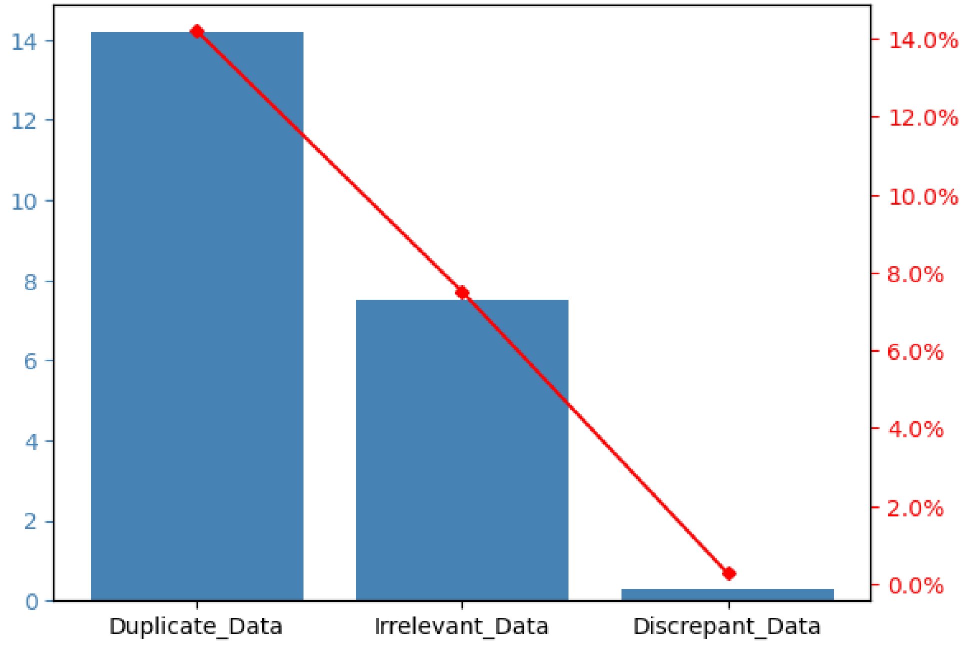

Figure 3: Pareto chart illustrating undesirable data in the datasets

Certain ML models can be interpreted by default. Decision trees and linear regression are two examples. The weights and coefficients for a linear regression model completely tell the narrative. We can determine which features have the greatest influence on the prediction and in what direction by looking at the weights. With decision trees, we are aware of the steps the model took to reach a conclusion. So, we can explain why a particular prediction was made by the model. Yet, these models might not produce excellent predictions for challenging issues. We thus employ models that perform more effectively on most tasks.

4.1.1 Outcomes of the UNSW-NB 15 Dataset

The output from UNSW-NB 15 Dataset is depicted in the following tables [3-5] and figures [4-7] below:

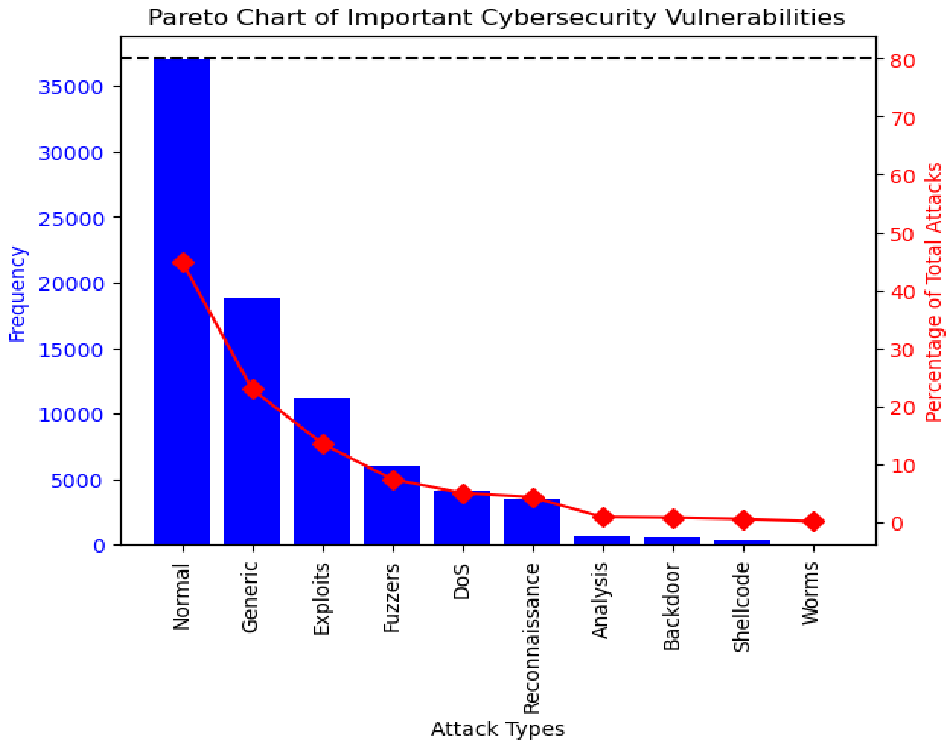

Figure 4: Vulnerabilities [UNSW-NB15 Cybersecurity Dataset]

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.86	0.84	0.87	0.86
KNN (Best k = 16)	0.86	0.90	0.89	0.89
LDA	0.87	0.85	0.85	0.85
QDA	0.89	0.88	0.87	0.87
ANN	**0.89**	**0.87**	**0.91**	**0.89**

Table 3: Performance of different models on UNSW-NB15 Cybersecurity Test Dataset

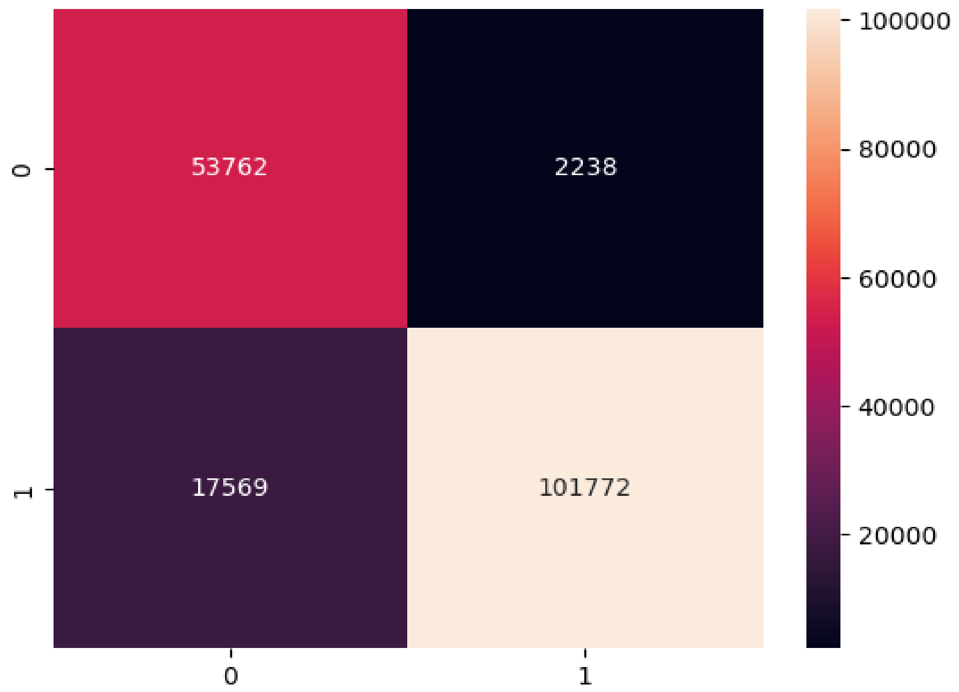

Figure 5: Confusion Matrix (ANN) [UNSW-NB15 Cybersecurity Test Dataset]

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.77	0.87	0.81	0.84
KNN (Best k = 16)	0.71	0.75	0.79	0.77
LDA	0.87	0.85	0.85	0.85
QDA	0.81	0.80	0.84	0.82
ANN	**0.89**	**0.87**	**0.90**	**0.88**

Table 4: Performance of different models on UNSW-NB15 Cybersecurity Test Dataset (After Feature Selection)

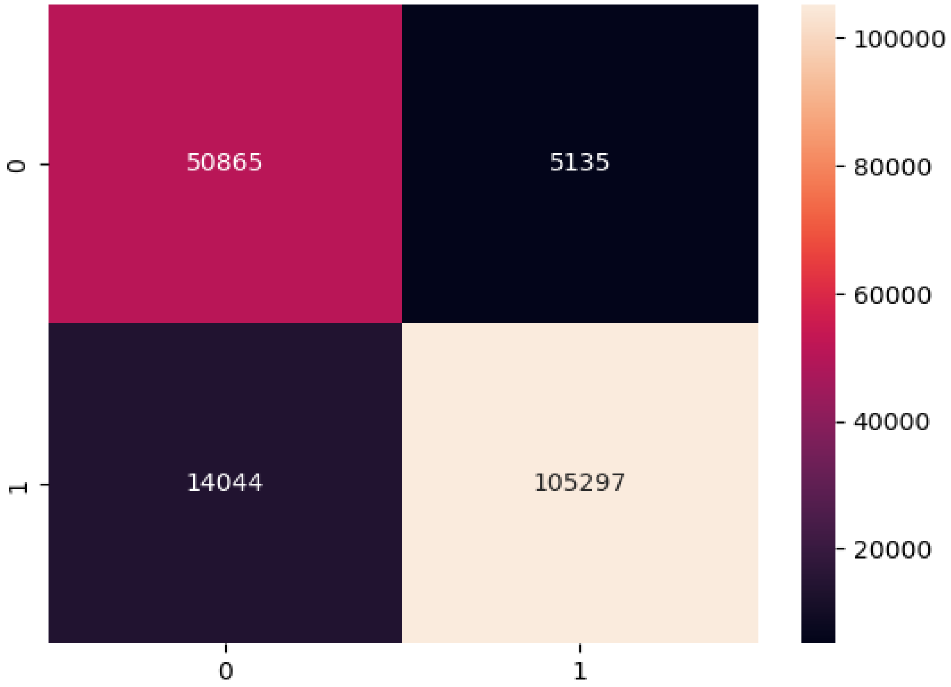

Figure 6: Confusion Matrix (ANN) [UNSW-NB15 Cybersecurity Test Dataset (After Feature Selection)]

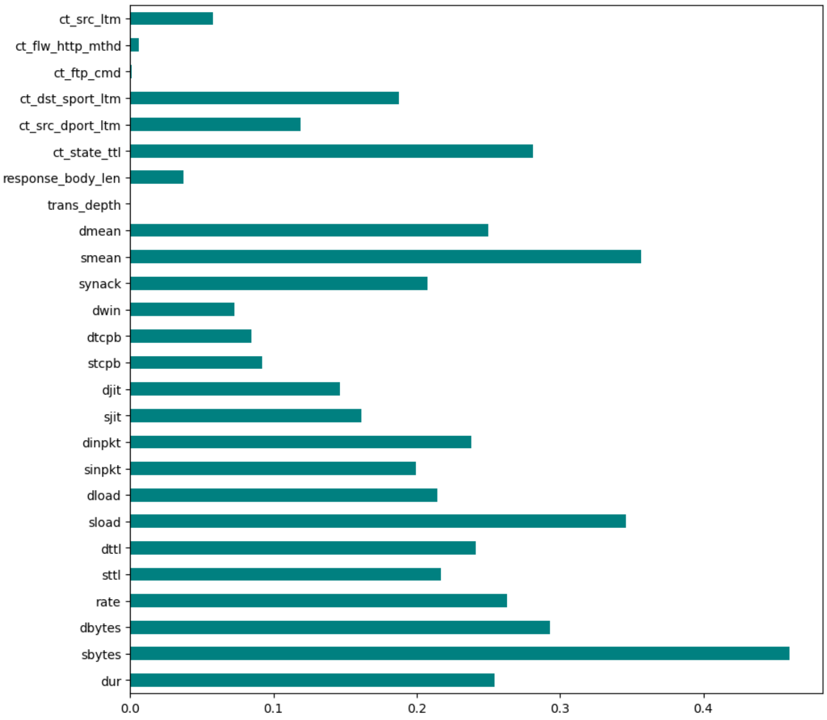

Figure 7: Feature Selection (Information Gain) [UNSW-NB15 Cybersecurity Test Dataset]

Features	Description
sttl	From origin to destination, an evaluation of life's time value
dttl	Value of remaining time to return to the point of origin
dwin	Advertisement of the TCP target window's value
synack	During a TCP connection establishment, the time between the SYN and SYN_ACK packets is measured
dmean	The typical size of the dst's row packets
trans_depth	Finds out how far along in the pipeline an HTTP request-and-response exchange has gotten
ct_dst_sport_ltm	Shared destination IP and source port percentage of the last 100 connections

Table 5: Selected Features of UNSW-NB15 Cybersecurity Data

4.1.2 Outcomes of the BoT-IoT Dataset

The output from BoT-IoT Dataset is depicted in the following tables [6-8] and figures [8-11] below:

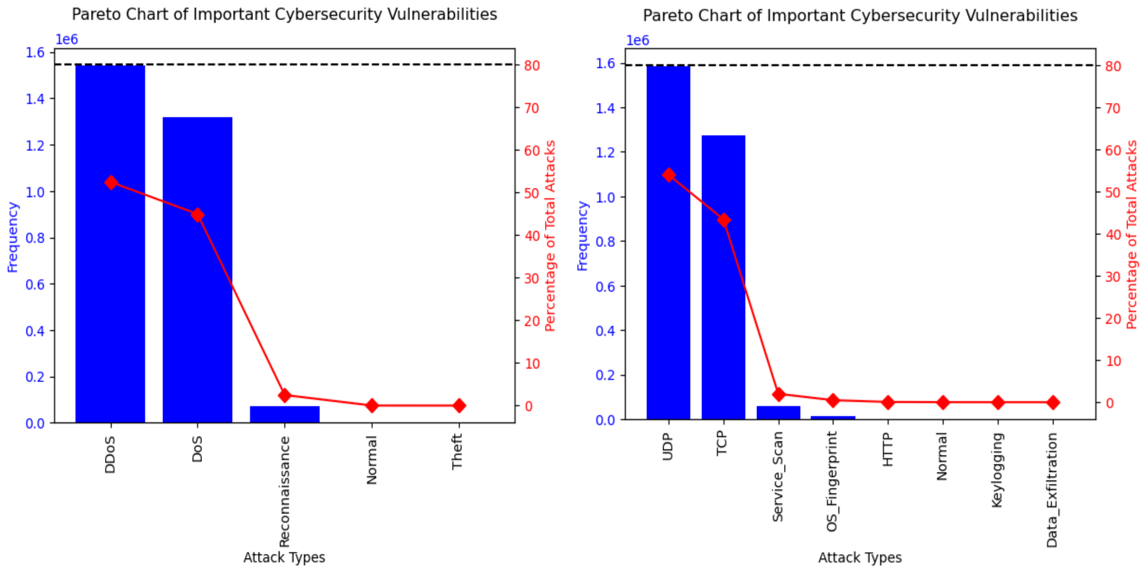

Figure 8: Vulnerabilities [IoT-Botnet Cybersecurity Dataset]

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.88	0.89	0.88	0.88
KNN (Best k = 1)	0.84	0.84	0.85	0.84
LDA	0.87	0.89	0.91	0.90
QDA	0.82	0.88	0.81	0.84
ANN	**0.89**	**0.89**	**0.89**	**0.89**

Table 6: Performance of different models on IoT-Botnet Cybersecurity Test Dataset

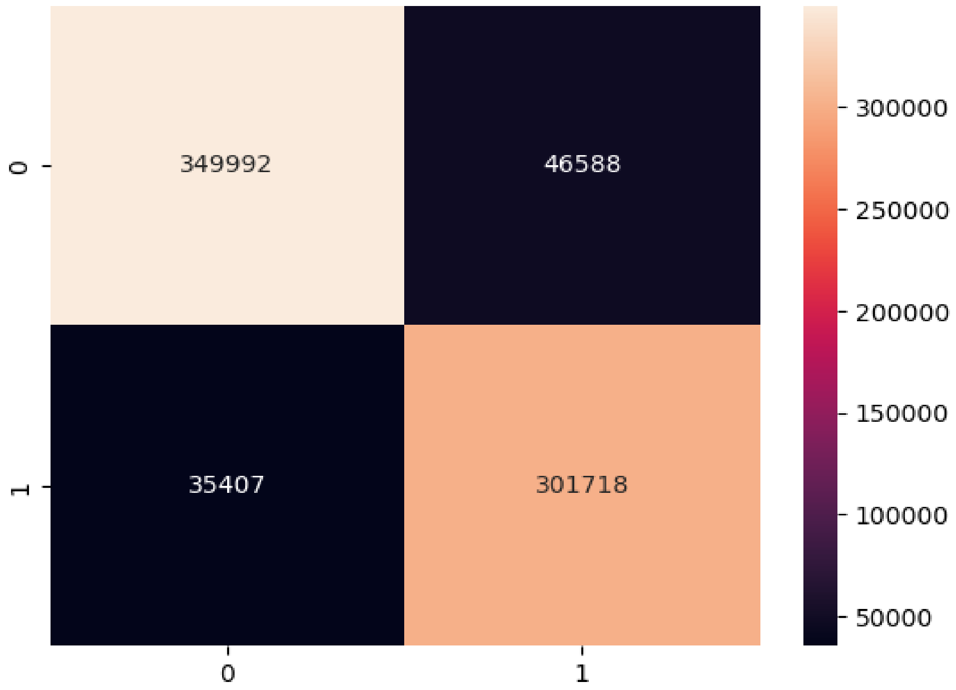

Figure 9: Confusion Matrix (ANN) [IoT-Botnet Cybersecurity Test Dataset]

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Logistic Regression	0.89	0.90	0.89	0.89
KNN (Best k = 1)	0.65	0.78	0.68	0.72
LDA	0.87	0.90	0.86	0.88
QDA	0.83	0.88	0.82	0.85
ANN	**0.90**	**0.91**	**0.89**	**0.90**

Table 7: Performance of different models on IoT-Botnet Cybersecurity Test Dataset (After Feature Selection)

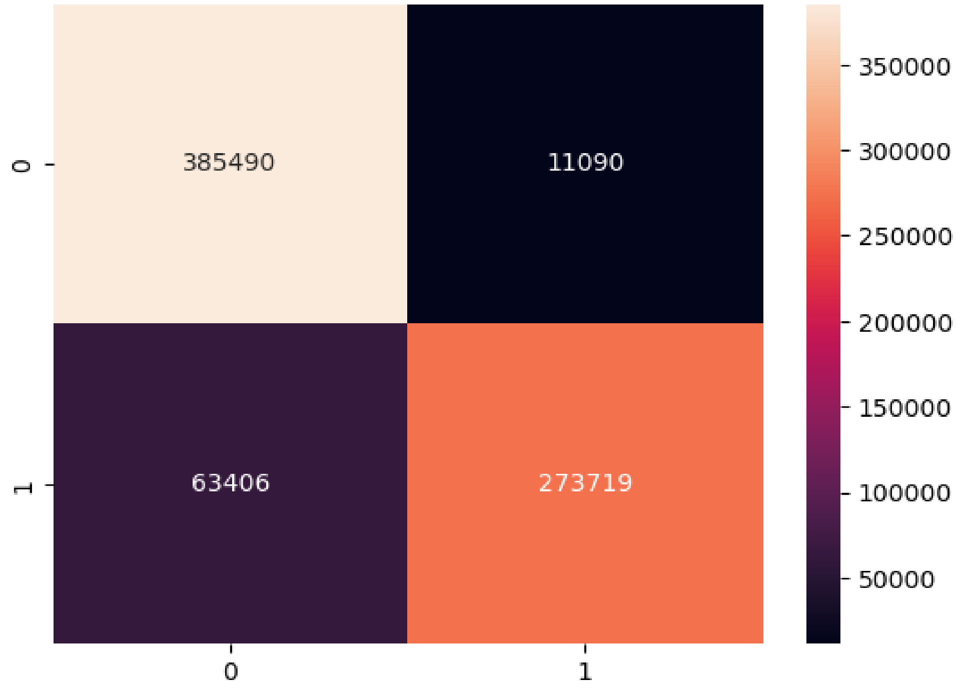

Figure 10: Confusion Matrix (ANN) [IoT-Botnet Cybersecurity Test Dataset (After Feature Selection)]

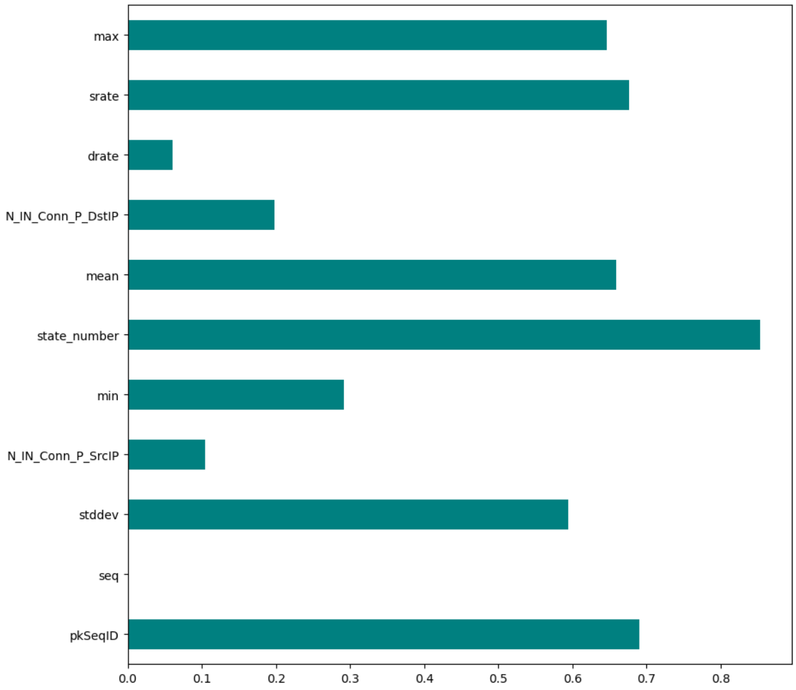

Figure 11: Feature Selection (Information Gain) [IoT-Botnet Cybersecurity Test Dataset]

Features	Description
min	Aggregate record retention requirements
state_number	Feature states are represented numerically
mean	The average age of all records
N_IN_Conn_P_DstIP	Total incoming connections broken down by destination IP
drate	Packets sent from a destination to a source per second
max	Aggregated data retention limits

Table 8: Selected Features of IoT-Botnet Cybersecurity Data

4.1.3 Outcomes of the DDoS Evaluation Dataset

The output from DDoS Evaluation Dataset is depicted in the following tables [9-12] and figures [12-15] below:

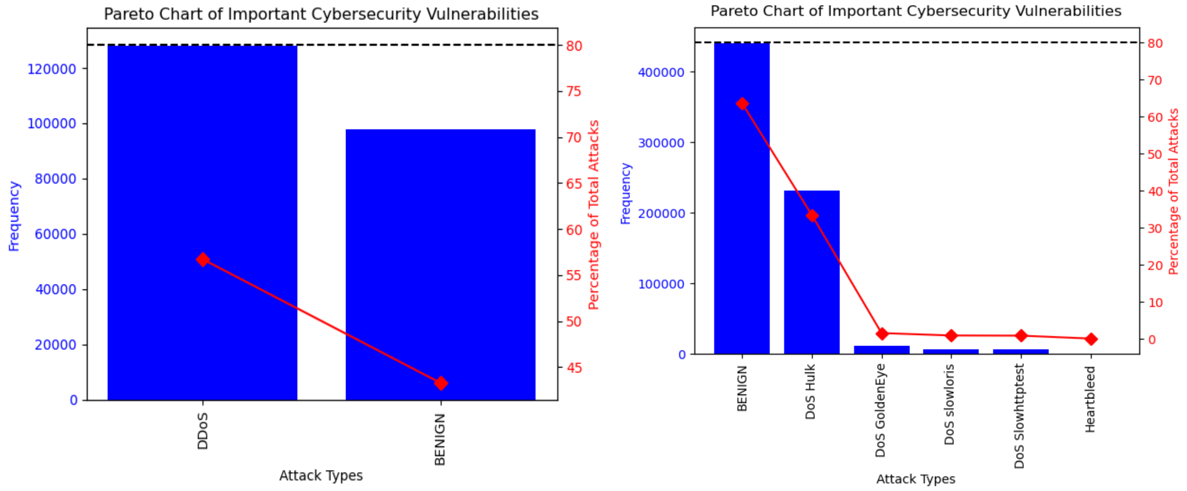

Figure 12: Vulnerabilities [DDoS Cybersecurity Dataset]

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.78	0.81	0.76	0.78
KNN (Best k = 2)	0.89	0.91	0.87	0.89
LDA	0.90	0.89	0.90	0.89
QDA	0.71	0.83	0.67	0.74
ANN	**0.96**	**0.96**	**0.96**	**0.96**

Table 9: Performance of different models on DDoS Cybersecurity Test Dataset

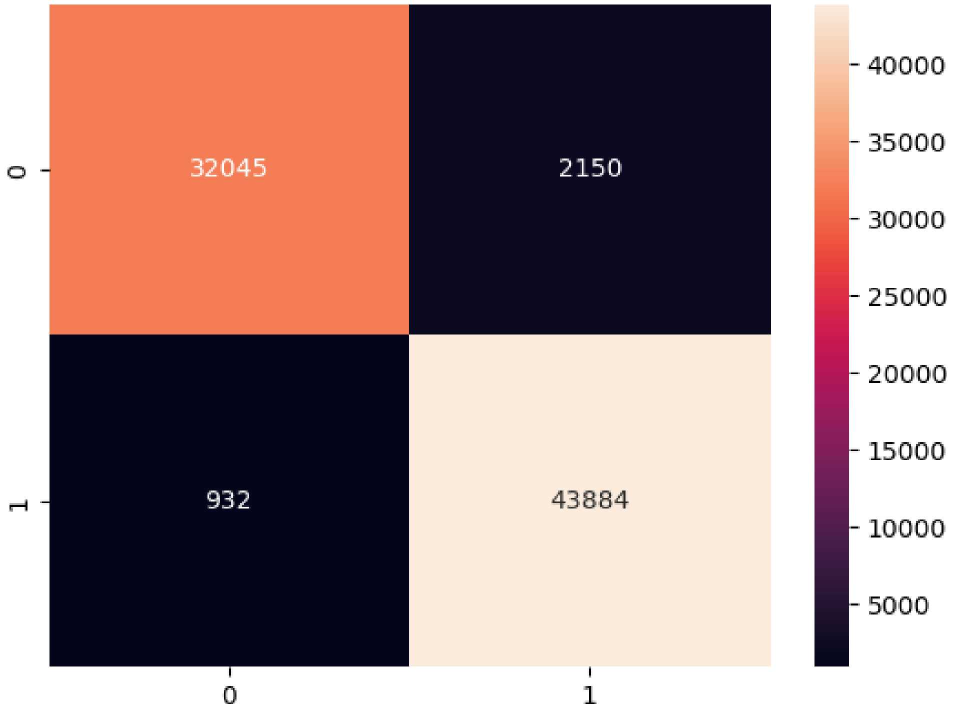

Figure 13: Confusion Matrix (ANN) [DDoS Cybersecurity Test Dataset]

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Logistic Regression	0.89	0.91	0.87	0.89
KNN (Best k = 2)	0.90	0.89	0.90	0.89
LDA	0.79	0.83	0.81	0.82
QDA	0.80	0.87	0.77	0.82
ANN	**0.97**	**0.97**	**0.96**	**0.96**

Table 10: Performance of different models on DDoS Cybersecurity Test Dataset (After Feature Selection)

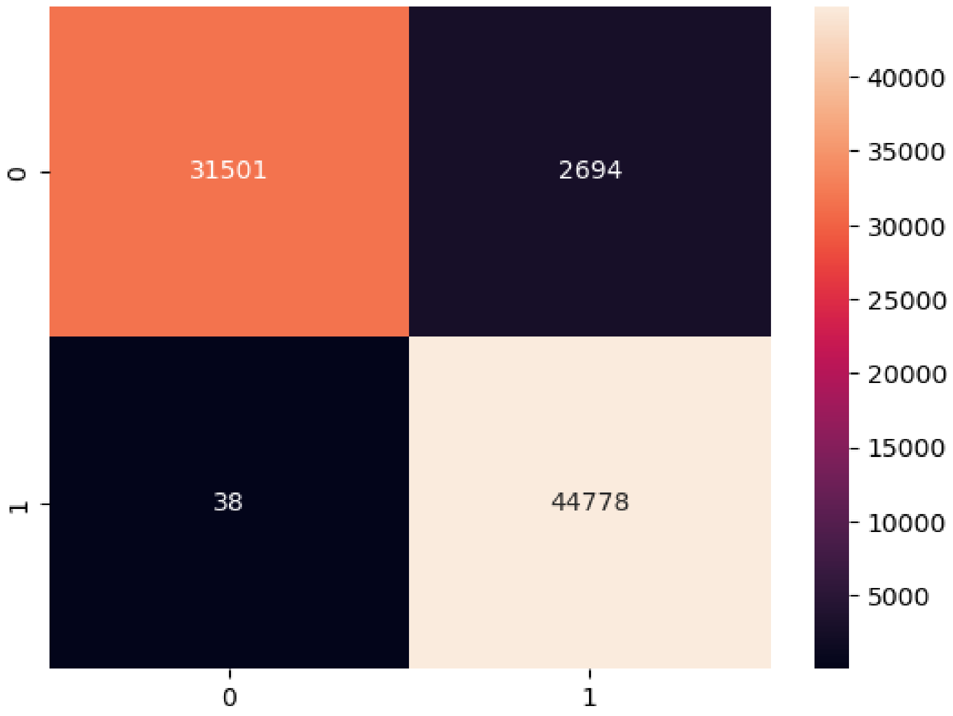

Figure 14: Confusion Matrix (ANN) [DDoS Cybersecurity Test Dataset (After Feature Selection)]

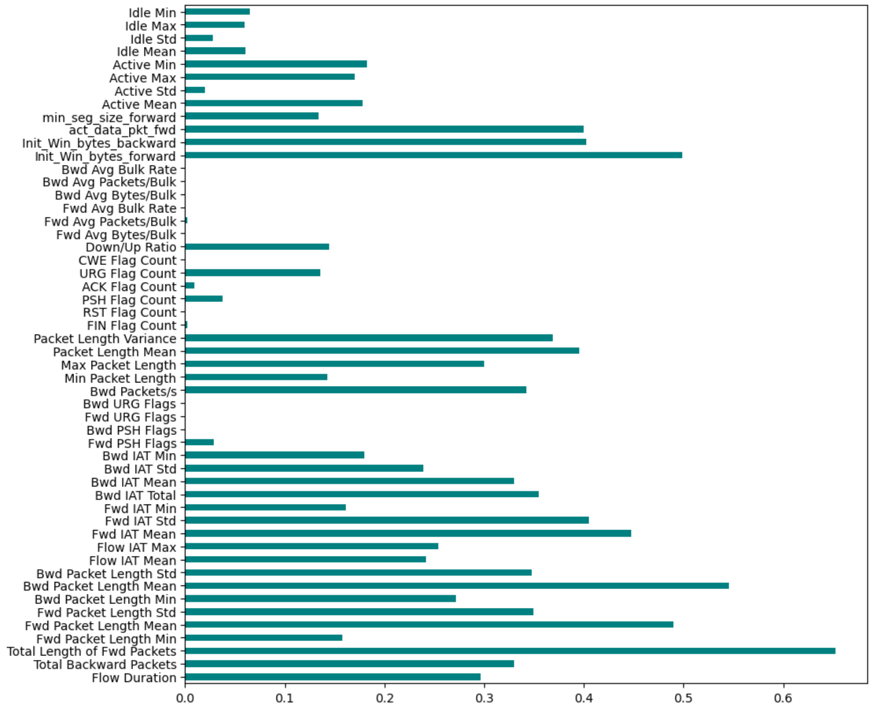

Figure 15: Feature Selection (Information Gain) [DDoS Cybersecurity Test Dataset]

Features	Description
Total Backward Packets	Quantity of data travelling in the opposite direction
Total Length of Fwd Packets	Forward package size in its entirety
Fwd Packet Length Mean	The mean of outgoing packets
Bwd Packet Length Min	The minimum packet size in the reverse direction
Bwd Packet Length Std	Variation in the size of packets sent in reverse
Packet Length Mean	Standard packet length
FIN Flag Count	Total FIN-containing packets count
RST Flag Count	Total RST-containing packets count
URG Flag Count	Total URG-containing packets count
Down/Up Ratio	The ratio of downloads to uploads
Init_Win_bytes_forward	Initial window bytes transmitted forward
act_data_pkt_fwd	Forward-direction TCP data packet count
min_seg_size_forward	The smallest forward-going segments recorded to date

Table 11: Selected Features of DDoS Cybersecurity Data

The following is a comparison of the features that are required to distinguish a DDoS attack and one that is not a DDoS attack:

Features	DDoS Attack	Not a DDoS Attack
Unauthorized access	No	Yes
Data theft or manipulation	No	Yes
Malware infections	No	Yes
Suspicious network activity	Yes	Yes
Unusual traffic patterns	Yes	No
High packet rates	Yes	No
Bandwidth consumption	Yes	No
Protocol anomalies	Yes	No
Source IP addresses	Yes	No
Phishing emails	No	Yes
Social engineering	No	Yes
Ransomware	No	Yes
SQL injection	No	Yes
Cross-site scripting (XSS)	No	Yes
Reflection attacks	Yes	No
Amplification attacks	Yes	No
Botnets	Yes	No
Distributed sources	Yes	No
Duration	Yes	No

Table 12: Comparison of the features required to identify a DDoS attack and not a DDoS attack

4.2 Conclusion

The tables above allow comparison of models across three cybersecurity test datasets. Logistic Regression, K-Nearest Neighbours (KNN), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and Artificial Neural Network (ANN) models were evaluated on each dataset using accuracy, precision, recall, and F1-Score. After feature selection, each model was compared to its baseline. The findings of this study lead to the following conclusions:

1. The Artificial Neural Network (ANN) model exhibited the most uniform and superior efficacy across the entirety of the three datasets. Hence, it can be deemed as the most efficacious framework for detecting and analysing cybersecurity threats.
2. For the UNSW-NB15 dataset, the best features for detecting cybersecurity threats are sttl, dttl, dwin, synack, dmean, trans_depth, and ct_dst_sport_ltm.
3. For the IoT-Botnet dataset, the best features for detecting cybersecurity threats are min, state_number, mean, N_IN_Conn_P_DstIP, drate, and max.
4. Similarly, for the DDoS Evaluation Dataset, the best features for detecting cybersecurity threats are Total Backward Packets, Total Length of Fwd Packets, Fwd Packet Length Mean, Bwd Packet Length, Min, Bwd Packet Length Std, Packet Length Mean, FIN Flag Count, RST Flag Count, URG Flag Count, Down/Up Ratio, Init_Win_bytes_forward, act_data_pkt_fwd, and min_seg_size_forward.
5. The identification of a DDoS attack necessitates the detection of various features, including but not limited to anomalous traffic patterns, high packet rates, bandwidth consumption, protocol anomalies, distributed sources, server response time, system vulnerabilities, and source IP addresses.

4.3 Future Work

The following is a list of potential work that could be done in the future based on the outcomes that were achieved:

1. Evaluate the performance of the ANN model on a wider variety of datasets related to cybersecurity in order to evaluate its efficacy and robustness in the identification of cybersecurity threats.
2. Carry out additional in-depth study about identifying and detecting emerging cybersecurity threats, such as those associated with cloud computing, the Internet of Things (IoT), and artificial intelligence.
3. Assess how well the produced models work on real-world cybersecurity datasets, and compare how well those models perform to the performance of any current commercial tools or solutions.
4. Create visualisation tools that will help cybersecurity experts analyse the output of machine learning models more accurately and determine which risks are the most significant.

Bibliography

[1] T. Friedman and M. Smith, "Measuring the Business Value of Data Quality," *Gartner,* no. G00218962, 2011.

[2] A. Labrinidis and H. V. Jagadish, "Challenges and opportunities with big data," *Proceedings of the VLDB Endowment,* vol. 5, no. 12, pp. 2032-2033, 2012.

[3] J. A. Aloysius, H. Hoehle, S. Goodarzi and V. Venkatesh, "Big data initiatives in retail environments: Linking service process perceptions to shopping outcomes," *Annals of Operations Research,* pp. 1-27, 2016.

[4] L. Cai and Y. Zhu, "The Challenges of Data Quality and Data Quality Assessment in the Big Data Era," *Data Science Journal,* vol. 14, no. 0, p. 2, 2015.

[5] Fischer EA. Cybersecurity issues and challenges: In brief. Congressional Research Service, 2014.

[6] Aftergood S. Cybersecurity: the cold war online. Nature. 2017;547(7661):30.

[7] Sun N, Zhang J, Rimba P, Gao S, Zhang LY, Xiang Y. Data-driven cybersecurity incident prediction: a survey. IEEE Commun Surv Tutor. 2018;21(2):1744–72.

[8] M. Albala, "cognizant," June 2011. [Online]. Available: https://www.cognizant.com/InsightsWhitepapers/Making-Sense-of-Big-Data-in-the-PetabyteAge.pdf.

[9] W. Eckerson, "Data Warehousing Special Report: Data quality and the bottom line," *Applications Development Trends May,* 2002.

[10] M. Zahedi Nooghabi and A. Fathian Dastgerdi, "Proposed metrics for data accessibility in the context of linked open data," *Program,* vol. 50, no. 2, pp. 184-194, 2016.

[11] J. Laitio, "Semantic Web Data Quality Control," Aalto University, 2011.

[12] Y. W. Lee, D. M. Strong, B. K. Kahn and R. Y. Wang, "AIMQ: A methodology for information quality assessment," *Information and Management,* vol. 40, no. 2, pp. 133-146, 2002.

[13] Interaction-Design, "Information Overload, Why it Matters and How to Combat It," 2017. [Online]. Available: https://www.interaction-design.org/literature/article/information-overloadwhy-it-matters-and-how-to-combat-it.

[14] C. S. M. Batini, Data and Information Quality, Springer, 2016.

[15] J. J. Geiger, "Data quality management: the most critical initiative you can implement," *SUGI 29 Proceedings,* pp. 1-14, 2004.

[16] N. Askham, D. Cook, M. Doyle, H. Fereday, M. Gibson, U. Landbeck, R. Lee, C. Maynard, G. Palmer and J. Schwarzenbach, "The Six Primary Dimensions for Data Quality Assessment," 2013.

[17] Y. Wand and R. Y. Wang, "Anchoring data quality dimensions in ontological foundations," *Communications of the ACM,* vol. 39, no. 11, pp. 86-95, 1996.

[18] Rainie L, Anderson J, Connolly J. Cyber-attacks likely to increase. Digital Life in. 2014, vol. 2025.

[19] Fischer EA. Creating a national framework for cybersecurity: an analysis of issues and options. LIBRARY OF CONGRESS WASHINGTON DC CONGRESSIONAL RESEARCH SERVICE, 2005.

[20] Craigen D, Diakun-Thibault N, Purse R. Defining cybersecurity. Technology Innovation. Manag Rev. 2014;4(10):13–21.

[21] Khraisat A, Gondal I, Vamplew P, Kamruzzaman J. Survey of intrusion detection systems: techniques, datasets and challenges. Cybersecurity. 2019;2(1):20.

[22] Johnson L. Computer incident response and forensics team management: conducting a successful incident response, 2013.

[23] Rizk A, Elragal A. Data science: developing theoretical contributions in information systems via text analytics. J Big Data. 2020;7(1):1–26.

[24] R. Y. Wang and D. M. Strong, "Beyond Accuracy: What Data Quality Means to Data Consumers," *Journal of Management Information Systems,* vol. 12, no. 4, pp. 5-33, 1996.

[25] T. Peng, L. Li and J. Kennedy, "A Comparison of Techniques for Name Matching," *GSTF Journal on Computing (JoC),* vol. 2, no. 1, pp. 55-61, 2012.

[26] Hagos DH, Yazidi A, Kure O, Engelstad PE (2017) Enhancing security attacks analysis using regularized machine learning techniques. In: 2017 IEEE 31st international conference on advanced information networking and applications (AINA). IEEE, pp 909–918.

[27] Jo S, Sung H, Ahn B (2015) A comparative study on the performance of intrusion detection using decision tree and artificial neural network models. J Korea Soc Digit Ind Inf Manag 11(4):33–45.

[28] Mitchell R, Chen R (2014) Behavior rule specification-based intrusion detection for safety critical medical cyber physical systems. IEEE Trans Depend Secure Comput 12(1):16–30.

[29] Hoang DH, Nguyen HD (2018) A PCA-based method for IoT network traffic anomaly detection. In: 2018 20th international conference on advanced communication technology (ICACT). IEEE, pp 381–386.

[30] Rathore S, Park JH (2018) Semi-supervised learning based distributed attack detection framework for IoT. Appl Soft Comput 72:79–89.

[31] Alrawashdeh K, Purdy C (2016) Toward an online anomaly intrusion detection system based on deep learning. In: 2016 15th IEEE international conference on machine learning and applications (ICMLA). IEEE, pp 195–200.

[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.

[33] Kim J,Kim J, ThuHLT,KimH(2016) Long short termmemory recurrent neural network classifier for intrusion detection. In: 2016 International conference on platform technology and service (PlatCon). IEEE, pp 1–5.

[34] Almiani M, AbuGhazleh A, Al-Rahayfeh A, Atiewi S, Razaque A (2019) Deep recurrent neural network for IoT intrusion detection system. Simul Model Pract Theory 101:102031.

[35] Prokofiev AO, Smirnova YS, Surov VA (2018) A method to detect internet of things botnets. In: 2018 IEEE conference of Russian young researchers in electrical and electronic engineering (EIConRus). IEEE, pp 105–108.

[36] Breiman L (2001) Random forests. Mach Learn 45(1):5–32.

[37] ChangY, LiW,Yang Z (2017) Network intrusion detection based on random forest and support vector machine. In: 2017 IEEE international conference on computational science and engineering (CSE) and IEEE international conference on embedded and ubiquitous computing (EUC), vol 1. IEEE, pp 635–638.

[38] Primartha R, Tama BA (2017) Anomaly detection using random forest: a performance revisited. In: 2017 International conference on data and software engineering (ICoDSE). IEEE, pp 1–6.

[39] Doshi R, Apthorpe N, Feamster N (2018) Machine learning DDOS detection for consumer internet of things devices. In: 2018 IEEE security and privacy workshops (SPW). IEEE, pp 29–35.

[40] Resende PAA, DrummondAC(2018)Asurvey of random forest basedmethods for intrusion detection systems. ACM Comput Surv (CSUR) 51(3):1–36.

[41] Mohamed TA, Otsuka T, Ito T (2018) Towardsmachine learning based IoT intrusion detection service. In: International conference on industrial, engineering and other applications of applied intelligent systems. Springer, pp 580–585.

[42] Jaganathan V, Cherurveettil P, Muthu SP (2015) Using a prediction model to manage cyber security threats. Sci World J.

[43] Kotpalliwar MV, Wajgi R (2015) Classification of attacks using support vector machine (SVM) on KDD cup'99 IDS database. In: 2015 Fifth international conference on communication systems and network technologies. IEEE, pp 987–990.

[44] Pervez MS, Farid DM (2014) Feature selection and intrusion classification in NSL-KDD cup 99 dataset employing SVMs. In: The 8th international conference on software, knowledge, information management and applications (SKIMA 2014). IEEE, pp 1–6.

[45] Yan M, Liu Z (2010) A new method of transductive SVM-based network intrusion detection. In: International conference on computer and computing technologies in agriculture. Springer, pp 87–95.

[46] Li Y, Xia J, Zhang S, Yan J, Ai X, Dai K (2012) An efficient intrusion detection system based on support vector machines and gradually feature removal method. Expert Syst Appl 39(1):424–430.

[47] Gauthama RamanMR, Somu N, Jagarapu S, Manghnani T, Selvam T, Krithivasan K, Shankar Sriram VS (2019) An efficient intrusion detection technique based on support vector machine and improved binary gravitational search algorithm. Artif Intell Rev 53:3255–3286.

[48] Saxena H, Richariya V (2014) Intrusion detection in KDD99 dataset using SVM-PSO and feature reduction with information gain. Int J Comput Appl 98(6).

[49] Chandrasekhar AM, Raghuveer K (2014) Confederation of FCM clustering, ANN and SVM techniques to implement hybrid NIDS using corrected KDD cup 99 dataset. In: 2014 International conference on communication and signal processing. IEEE, pp 672–676.

[50] Shapoorifard H, Shamsinejad P (2017) Intrusion detection using a novel hybrid method incorporating an improved KNN. Int J Comput Appl 173(1):5–9.

[51] Vishwakarma S, Sharma V, Tiwari A (2017) An intrusion detection system using KNN-ACO algorithm. Int J Comput Appl 171(10):18–23.

[52] Meng W, Li W, Kwok L-F (2015) Design of intelligent KNN-based alarm filter using knowledge-based alert verification in intrusion detection. Secur Commun Netw 8(18):3883–3895.