National College of Ireland

National College of Ireland

Project Submission Sheet

Student Name:	Abinash Mishra		
Student ID:	X23153903		
Programme:	Msc Cybersecurity	**Year:**	2024
Module:	Msc Research Practicum		
Lecturer:	Eugene Mclaughlin		
Submission Due Date:	16-09-2024		
Project Title:	DEEP LEARNING-BASED INTRUSION DETECTION SYSTEM IN THE INTERNET OF THINGS		
Word Count:	6832		

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: Abinash Mishra

Date: 16-09-2024

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

DEEP LEARNING-BASED INTRUSION DETECTION SYSTEM IN THE INTERNET OF THINGS

Abinash Mishra
X23153903

Abstract

This study aims to create and assess a deep learning-based intrusion detection system (IDS) for enhancing security in Internet of Things (IoT) environments. Goals include investigating DL algorithms to enhance IDS efficiency and accuracy in IoT, identifying effective approaches for detecting and mitigating various cyberattacks, and addressing how deep learning-based IDS architecture handles dynamic and heterogeneous IoT network traffic and devices. The study also explores strategic planning for robust DL approaches to mitigate threats in IoT devices and systems and examines architectural frameworks of DL models that can be implemented to identify and mitigate threats in IoT systems and devices.

Chapter 1: Introduction

Background

The “Internet of Things (IoT)” is a network of gadgets that can join and swap data with each other and the cloud. IoT instruments can contain “mechanical and digital machines”, customer objects, and electronics. They are typically implanted with “sensors and software” technology and can have amazing IDs. IoT machines can work together to collect, communicate, and interpret data to provide understanding to end users. The leading associates of IoT are “Sensors, Actuators, Connectivity, Data processing, and User interface”. In terms of, these appliances sense knowledge from the surroundings and transform it into data. Connectivity especially authorises devices to connect to a network and communicate with each other. Data processing is mainly a mechanism for gathering, transmitting, and analyzing data. The user interface allows users to interact with the IoT system. It is also notable that IoT has multiple applications across various domains, such as “homes, cities, agriculture, healthcare”, and more. The adoption of IoT has rapidly risen over the last decade. The IHS calculated a seated base of 15.41 billion instruments in 2015, almost tripling to “42.62 billion” by 2022. It is predicted to increase at an even more quick pace, reaching “75.44 billion” by 2025.

IoT has multiple unique security challenges such as “Weak authentication, Lack of standardization, Lack of encryption, Insecure protocols” and many more. Manufacturers can assist in making authentication more protected by directing multiple stages, utilising powerful default passwords, and establishing parameters that lead to assured user-generated passwords. With a lot of additional “devices, protocols, and platforms”, it is quite challenging to provide “compatibility and interoperability” between them (HaddadPajouh *et al*., 2021). IoT can also show exposures which can be used by attackers. It has also been identified that Multiple IoT devices transmit operating protocols that lack built-in protection components which offers a considerable challenge for securing the IoT. These uncertain protocols can uncover data for interception, tampering and unauthorized credentials during communication.

IoT also faces different security threats and vulnerabilities like weak credentials, insecure networks, insecure updates, DDoS attacks, malware, data breaches and many more. Weak, guessable or hardcoded passwords are typically the exposures that

malware manipulates (Mukhtar *et al*., 2023). The data information without encryption can be introduced by hackers by revealing certificates and other details. An intrusion detection system (IDS) monitors network traffic for suspicious movement and transmits warnings when such movements are uncovered. Anomaly detection and reporting are the direct processes of an IDS but some of the strategies also takes action when a negative movement or anomalous traffic is noticed.

Signature-based and anomaly-based IDS are two primary techniques for recognizing and precautioning against threats. The primary distinction between the two is that signature-based techniques depend on a database of learned attack signatures, while anomaly-based approaches recognise abnormal behaviour. Compares network packets to a database of known attack signatures. It is compelling at detecting common attacks, but it can not detect unknown attacks or variations.

Aim & Objective

AIM

This study aims to create and assess a deep learning-based "intrusion detection system (IDS)" for enhancing security in the "Internet of Things" (IoT) environments.

Objectives:

- To investigate the algorithms of DL assist to increase the efficiency, and accuracy of IDS in the section of IoT.
- To identify the effective approaches for the identification and mitigation of different cyber attacks in IoT.
- To address the deep learning-based architecture of IDS handles dynamic, and heterogeneous IoT network traffic and devices.
- To explore strategical planning can be implemented to handle robust deep-learning approach for the mitigation of the threats in IoT devices, and systems.
- To explore the architectural frameworks of the deep-learning model that can be implemented to identify and mitigate the threats in the IoT systems, and devices.

Research Question

1. How can the algorithm of deep learning assist in increasing the efficiency, and accuracy of intrusion detection systems in the section of IoT?
2. What are the effective approaches for the identification, as well as mitigation of different cyber-attacks in IoT?
3. How the deep learning-based architecture of IDS handles dynamic and heterogeneous IoT network traffic, and devices?
4. What strategical planning can be implemented to handle a robust deep-learning approach for the mitigation of the threats in IoT devices, and systems?
5. What are the architectural frameworks of the deep-learning model that can be implemented to identify and mitigate the threats in the IoT systems, and devices?

Research Significance

Deep learning is a variety of "artificial intelligence (AI)" that introduces computers to process data in a manner that's motivated by the human brain. DL standards can recognise complicated patterns in data, such as "images, text, and sounds", to construct forecasts and understandings. These models can also automate tasks that usually demand human brains, such as transcribing audio to text or representing images (Alzahrani, and Alenazi, 2021). DL can be a cutting-edge solution for intrusion detection in IoT. DL algorithms can handle high-dimensional characteristics and identify unwanted attacks accurately and quickly. They can also automatically adjust security systems when malware or breaches are detected, even with low computational power. Benefits of DL-based IDS in IoT such as "Data-driven and anomaly, Real-time monitoring Predictive insights" (Ahmad *et al*., 2021). In terms of Data-driven and anomaly-based systems can detect new and unknown attacks. Real-time monitoring can help to analyze video streams to address threats and anomalies.

Research Rational

DL can be used to develop resilient and adaptive network IDS that can notice and organise network attacks. DL techniques can assist IDSs in determining both known and new network behavioral attributes, which can assist remove intruders and decreasing the chance of compromise (Taye, 2023). It also identifies that DL can also be utilized to interpret input data to indicate normal and abnormal behaviours, as well as new, anonymous attacks. DL models that have been used in IDSs such as "CNN, RNN, LSTM" and many more. Some latest research trends in DL such as ML techniques specifically allow an agent to learn through trial and error in an interactive enviroment (Mallick *et al*., 2023). On the other hand, Embedded ML is a more efficient option to cloud-based systems that permit the reduction of cyber threats, data theft, and bandwidth and network resource costs.

Chapter 2: Literature Review

Introduction

IDS based on DL can address cyberattacks in IoT networks. IoT systems are distributed and complex, with multiple devices and sensors connected to a network. It is also identified that DL-based IDS mainly monitors network traffic for unauthorized access or malicious activity (Tharewal (Tharewal *et al*.,2022). Leading benefits of using a Deep learning approach for IDS such as DL-based IDS can achieve high "accuracy, precision, recall, and F-score". It is also identified that ML-based IDS can adapt as well as it can enhance detection capabilities over time through understanding new data. It is also identified that, Anomaly-based IDS can manage new zero-day intrusions.

Algorithms of DL assist in increasing the efficiency, and accuracy of IDS in the section IoT

Deep learning architectures are models that can assist in learning complex patterns in data and contain. In terms of "CNN, GANs, LSTM, MLPs" and many more. As per the view of Li *et al*., 2023, CNN networks are made up of convolutional layers, activation layers, and fully connected layers, and are good for image-based problems. Also recognized is that in terms of generative adversarial networks, it especially uses two neural networks to grow, and can be utilized for the epoch of images and also other assignments with minimal human intervention. These recurrent neural networks are good for modelling sequential data. Another research paper published by Zhang *et al*., 2023, this research explores that LSTM recurrent neural networks are good for modelling sequential data. In terms of MLPs, this network can handle complex computations and develop the accuracy of training models.

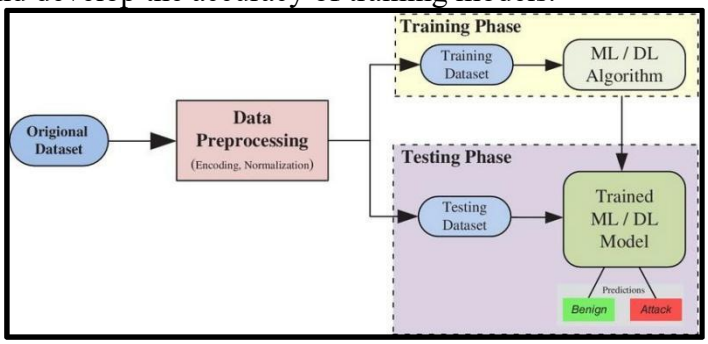


Figure 2.1: DL approach

(Source: Ahmad *et al*., 20221)

According to Mishra, and Pandya, 2021, cybersecurity for IoT devices and networks faces several challenges in intrusion detection. It has been identified that IDS can

mistakenly activity as suspicious, wasting time and resources. In terms of Noise, bad packets, corrupt DNS data, and local packets can develop a high false-alarm rate.

Effective approaches for the identification and mitigation of different cyber-attacks in IoT

Signature-based detection is a well-established technique for determining malicious movement by corresponding network traffic to a database of understood attack signatures. These signatures can designate known marks or fingerprints of network episodes or suspicious process behaviour. For instance, a signature could be established on the number of bytes, 1s, or 0s in network traffic, or on an available malicious instruction series employed by malware. When a match is made, the system causes an alert and may take detailed actions against the code, such as forewarning the user (Robinette, *et al*., 2024). According to corelight.com Signature-based detection was originally used by antivirus designers to scan system files for evidence of malicious movement. It can observe endpoints and traffic from a cloud background for anything matching a special attack signature, which may be located in packet headers, application code, or data stores. As per the view of Kime, (2024), Network security architecture is the creation and performance of security standards to guard an organization's computer networks from cyberattacks, data violations, and undesirable entrance. It involves checking security components to network segments that is established on essential goals or principles, such as data integrity, data confidentiality, risk, data availability, measurable actions, and useful management.

The deep learning-based architecture of IDS handles dynamic and heterogeneous IoT network traffic and devices

"Recurrent neural networks (RNNs), Long Short-Term Memory (LSTM), and Gated Recurrent Units (GRUs)" are all well-suited for handling time-series data and engaged traffic practices. In terms of RNNs, they Can tolerate irregularly spaced time intervals and adjust to additional forecasting tasks. Apart from these, RNNs can have problems charging long-term dependencies due to disappearing or flashing gradient problems, or memory limitations when dealing with very long successions. On the other hand, LSTMs Can capture long-term dependencies better than RNNs, which is important for time series with long-term cycles or activities. LSTMs are a well-known deep-learning model that has been confirmed as correct in holding material dependencies and making accurate forecasts. CNNs are powerful tools for feature extraction and classification, creating them useful for intrusion detection systems (IDS). They can process different data types, including "network traffic data, log files, and system calls", to recognise practices and anomalies expressive of malicious movement. CNNs can automatically learn elements from the data without manual feature engineering. Apart from these, CNNs have demonstrated high accuracy in determining different intrusion types. Also, CNNs can be used for real-time investigation of network traffic and system training.

Strategical planning can be implemented to handle a robust deep-learning approach for the mitigation of the theatres in IoT devices, and systems

As per the view of Akhtar, and Feng, (2022), the LSTM algorithm is a considerably effective deep-learning approach for detecting malware. The CNN-LSTM approach has also offered a 99% detection accuracy, which is more elevated than other methods such as "SVM and DT. "

Deep Convolutional Neural Networks (DCNNs)" are also considered robust and efficient for malware detection. According to Wu *et al*., (2023), the deep learning model is examined and combined with differential solitude to supply the strategy framework for the deep differential solitude data protection algorithm. The comparable model of the productive adversarial network is utilized to allow the attacker to receive the optimal fake selections. Another research paper published by Alwahedi *et al*.,(2024), ML plays a vital role in IoT security by examining data from

interconnected instruments and networks to catch and stop threats. ML models can investigate normal and attack traffic in real time to recognise unusual patterns that show security breaches. They can also acclimate to new threats utilising historical data. Apart from these, ML can also be utilized to improve intellect by ingesting images, video, and audio. For instance, IoT contraptions and sensors can deliver further details and context to video management systems, such as “motion, sound, temperature, and humidity”. This can allow the techniques to produce more insightful data and bring out cultivated analytics.

Theoretical Underpinning

Deep Learning Theory

Deep learning is a subset of artificial intelligence (AI) and machine learning that teaches computers to process data in a way that is inspired by the human brain. Deep learning models are made up of numerous layers of connected nodes, reached artificial neurons, that function together to process and learn from data. These examples are prepared on large quantities of unlabeled data to determine and classify sensations, identify patterns and relationships, evaluate opportunities, and make forecasts and conclusions (Ni *et al*.,2020). Apart from these, Deep learning standards can recognize complicated patterns in images, text, sounds, and other data to deliver objective understandings and forecasts. Different kinds of deep learning models can be utilized to perform specific solutions, such as CNN, and RNN.

Anomaly Detection Theory

Anomaly detection in IoT security is the method of determining data points, events or statements that are extremely different from that of the expected patterns in a dataset. It mainly helps to monitor the health, implementation and security of devices and methods in the IoT, and it can also help in detecting problems in the previous stage. This can contain security breaches, equipment malfunctions and inefficiencies, which can assist decrease the risk of major failures. It is also identified that Anomaly detection uses data analytics, AI, and ML to identify observations that break the norm, such as those that exceed or lag behind usual behaviours (Chatterjee, and Ahmed, 2022). These abnormalities can be damaging, like a device that's not functioning correctly, or they can be positive, like a mobile application noticing a strong uptick in use during a specific duration.

Game Theory

Game theory can model the engaged relations between attackers and an “intrusion detection system (IDS) in the Internet of Things (IoT)”. The objective is to determine the best access control system for the IDS. It is also determined that IDS and detractors are sported as participants in a two-player game. The exchanges are dynamic, with comprehensive knowledge, but the IDS is unsure about the attacker's version (Esposito *et al*., 2020). The interaction is sported as a nonparticipatory game, suggesting players can't deal with or form partnerships. Mathematical formulations and the Nash compensation are utilized to create a dedicated IoT-based IDS system. The IDS seller employs detection measures, such as the number of receptacles received and created, to provide the “intrusion ratio (IR)”.

Conceptual Framework

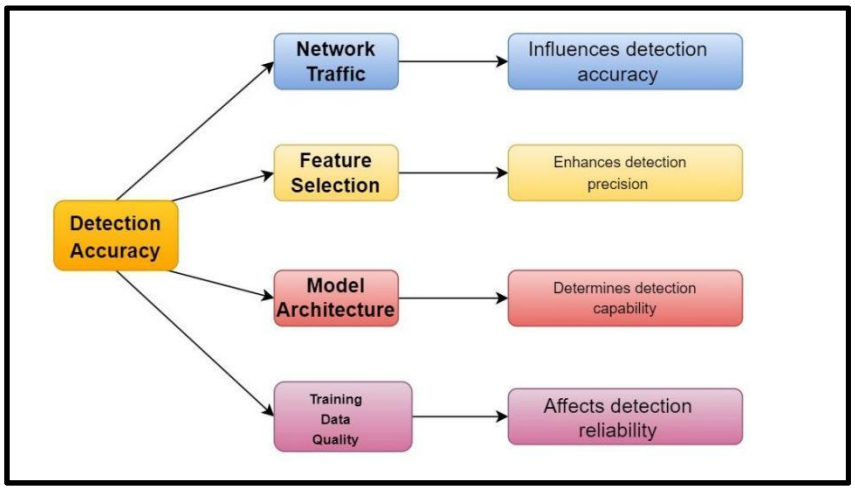


Figure 2.2: Conceptual Framework

(Source: Self-Created)

Chapter 3: Methodology

3.1 Introduction

IDS for the IoT utilizes different methodologies to detect and react to security hazards. These methods contain signature-based, anomaly-based, stateful protocol research, and hybrid methods. It is also determined that Hybrid systems can take benefit of the details of selected processes while minimizing the effect of their weaknesses. It is found that some research papers represented hybrid-based NIDS for IoT strategies that contain anomaly-based NIDS with signature and specification-based detection to reduce false information while keeping the benefits of the other methods. This section is vital as it allows readers to evaluate the study's validity and reliability. It also offers the understanding of the researcher's research theory.

3.2 Research Method

The below flowchart illustrates that the data is first loaded, where the dataset is imported for analysis.Then the data is preprocessed, where the data is cleaned and formatted.This ensures that the data is free from any type of inconsistencies and is suitable for analysis.Then Exploratory Data Analysis (EDA) is conducted for gaining initial insights into the data structure,distributions and variables relationships.After this process, Feature Extraction is carried out for identification and selection of relevant features which is used in model building.Then model is trained by applying machine learning algorithms to that of the preprocessed data for development of predictive models.After this, models are Evaluated for determining their effectiveness using various performance metrics.Finally the Calculating Metrics provides the detailed quantitative assessments of the models by comparing them to identify the most effective model for the given dataset.This methodology gives a systematic approach to building,evaluating and selecting of the best ML model for the research objective.

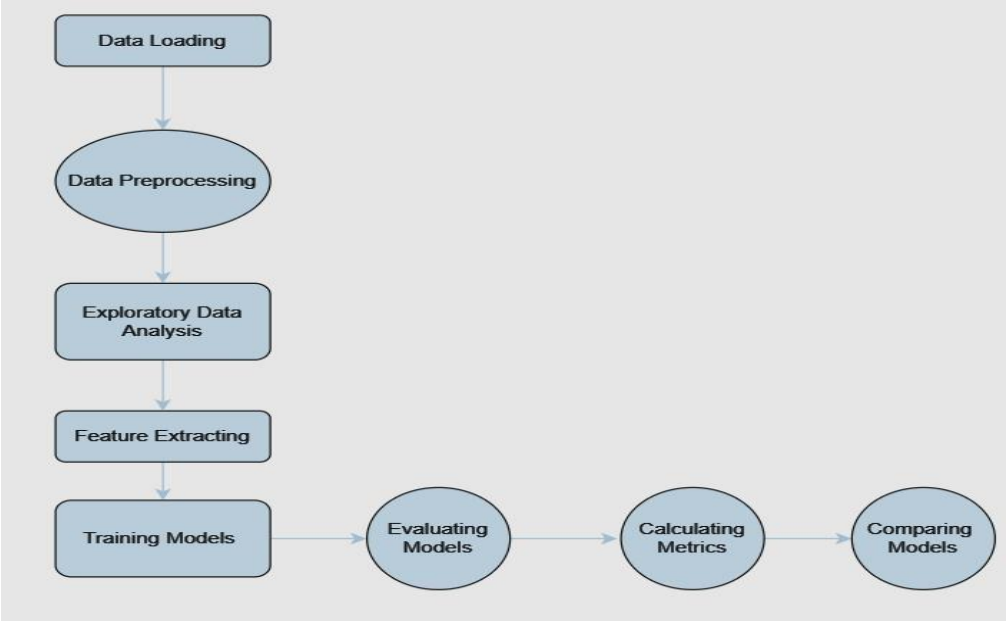


Figure 3.1: Flowchart
(Source: Self-Created)

3.3 Data collection

The dataset that was chosen, simulated various normal and attack scenarios across different applications and protocols.With the feature that was extracted using CICFlowmeter, network traffic was captured using tools like Wireshark and Netflow. This dataset specially focuses on brute force attacks and normal traffic that is suitable for testing and training intrusion detection system (IDS).

This is a secondary type data cluster that was orginally collected by (CIC and CSE).A secondary data cluster is the process of retrieving information from sources that already exist. This data is usually collected by other people or associations for a different purpose. Although primary data was gathered raw data from primary sources in this research primary data does not work accurately for this significant reason this research paper uses a secondary data collection process. It is also notable that secondary data is usually cheaper and faster to collect than primary data, and it is already been pre-processed and managed (Sharma, 2022). This particularly allows researchers to focus on their research purposes and proceed instantly to the analysis method. Also identified that secondary data can enable researchers and data analysts to create extensive, high-quality databases that specifically support cracking problems.

3.4 Data Pre-Processing

Initially the data has some missing values, which are handled by either dropping the missing values or filling them with the mean value of the respective column. This ensures that the dataset is complete and does not have any null values that could introduce the error while training the model. All the rows with missing data were dropped to ensure the integrity of the dataset. After that the categorical variables are encoded, specifically the values of variable ‘labels’ are encoded using techniques like ‘LabelEncoder’. Categorical variables ate encoded to convert it into a numerical values which are usually easy to process for machine learning models.

After encoding the variables features were scaled using ‘standardScaler’ which normalizes the data so that each features has a mean of 0 and a standard deviation of 1.

Scaling the features is crucial for models like SVM as models like this are sensitive to the scale of input features.

3.5 Data Resampling:

After converting the categorical values to numerical values further dataset was resampled using the 'resample' method to create a balanced dataset as it was found to be imbalanced with majority of the samples labeled as benign. The dataset was split into two sets which are testing and training datasets by using "train_test_split". This was done for evaluating the model's performance on unseen data that ensures the model generalizes it. The dataset was split into training and testing set using 80-20 split. The split was done to ensure the proportional representation of each class in both the training and testing dataset.

3.6 Model Development:

Once both the training dataset as well as testing dataset was ready, further step is to create or build a model. Here, the sequential model with multiple layers was constructed using the Keras which includes the three Dense layers which are input layer, hidden layer and output layer. The model was compiled with 'categorical_crossentropy' loss, 'adam' optimizer and 'accuracy' metrics and trained for 50 epochs with a batch size of 32. Further a support vector machine model with a linear kernel was developed using 'SVC' class from scikitlearn. Along with this Linear regression model was also developed and evaluated as a baseline model for comparison. All the models was trained on the scaled training data and evaluated on the test data. Linear regression was applied to the classification problem to observe its effectiveness.

3.7 Model Evaluation:

The models were evaluated using accuracy scores to measure their performance on the test dataset. Also cross-validation was performed during the training process to validate the models' performance and avoid overfitting. Along with this confusion matrices were generated to understand the distribution of true positives, false positives, true negatives, and false negatives across the different classes.

3.8 Data Visualization:

The last step is to visualize the data to gain insights into the distribution of features and the relationships between different variables. Visualization helps in understanding the underlying patterns and in guiding the further analysis and model development. There are different visualization techniques used in machine learning, in this project the techniques used are scatter plot, pie chart, and count plot. By visualizing the data it helps to identify class imbalance and relationship between features that could influence the model performance.

Chapter 4: Design Specification

The design is centered around the development of a machine learning based intrusion detection system that can classify the IoT network traffic as benign or malicious. The system inolves multiple machine learning models which includes support vector machines(SVM), Linear regression to analyze the network traffic data and detect potential threats. The design integrates data preprocessing, model training and evaluation. The execution of the code leverages several key frameworks and libraries such as pandas, numpy, scikit-learn, tensorflow. The system architecture includes following key componentes;

1. Data Ingestion which involves data loading, data cleaning and feature scaling
2. Feature engineering which involves feature selection, aggregation and correlation analysis
3. Model design and training which involves training the models
4. Evaluation and Validation which involves measuring the model performance using several metrics.

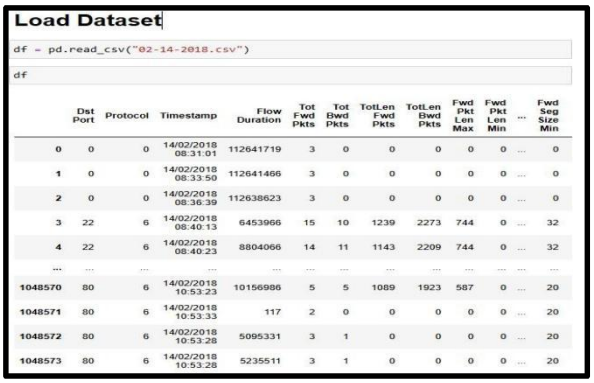
Chapter 5: Data Analysis & Result

Introduction

“Deep learning is a type of machine learning” that uses “artificial neural networks (ANNs)” to teach computers to process data in a way that mimics the human brain. The goal of deep learning is to create accurate predictive models from large amounts of unlabeled data. Deep learning models are made up of multiple layers of interconnected nodes, or neurons, that work together to perform nonlinear transformations on input data. The term "deep" usually refers to the number of hidden layers in the neural network. Deep learning models can determine complicated patterns in “pictures, text, sounds”, and other data to produce insights and predictions. They can be used to automate tasks that typically require human intellect, such as describing images or transcribing a sound file into text. It is mainly beneficial for developing and assessing DL-based IDS for improving security in the IoT environment.

5.1 Data Analysis

Necessary libraries especially pandas, numpy and matplotlib etc, were imported for execution.There are numerous amount of IDS datasets that are availble online.But mostly every dataset that is provided has a less success rate.



Load Dataset

```
df = pd.read_csv("02-14-2018.csv")
```

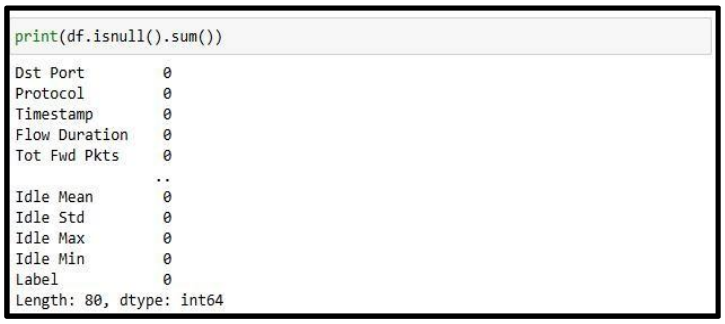
```
df
```

	Dst Port	Protocol	Timestamp	Flow Duration	Tot Fwd Pkts	Tot Bwd Pkts	TotLen Fwd Pkts	TotLen Bwd Pkts	Fwd Pkt Len Max	Fwd Pkt Len Min	...	Fwd Seg Size Min
0	0	0	14/02/2018 08:31:01	112641719	3	0	0	0	0	0	...	0
1	0	0	14/02/2018 08:33:50	112641466	3	0	0	0	0	0	...	0
2	0	0	14/02/2018 08:36:39	112638623	3	0	0	0	0	0	...	0
3	22	6	14/02/2018 08:40:13	6453966	15	10	1239	2273	744	0	...	32
4	22	6	14/02/2018 08:40:23	8804066	14	11	1143	2209	744	0	...	32
...	...	...	...	...	...	...	...	...	...	...	...	...
1048570	80	6	14/02/2018 10:53:23	10156986	5	5	1089	1923	587	0	...	20
1048571	80	6	14/02/2018 10:53:33	117	2	0	0	0	0	0	...	20
1048572	80	6	14/02/2018 10:53:28	5095331	3	1	0	0	0	0	...	20
1048573	80	6	14/02/2018 10:53:28	5235511	3	1	0	0	0	0	...	20

Figure 5.1: Load Dataset

(Source: Self-created)

Datasets are vital for ML due to they provide the data to build predictive models. The quantity, quality, and relevance of this dataset. It determined the performance of this ML model.



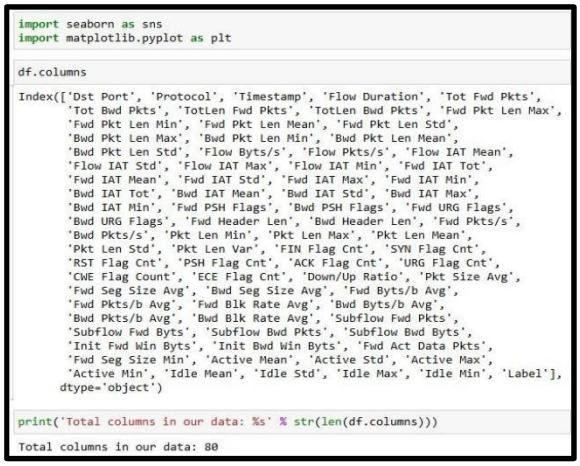
```
print(df.isnull().sum())

Dst Port         0
Protocol         0
Timestamp        0
Flow Duration    0
Tot Fwd Pkts     0
                ..
Idle Mean        0
Idle Std         0
Idle Max         0
Idle Min         0
Label            0
Length: 80, dtype: int64
```

Figure 5.2: Null Check

(Source: Self-created)

Null checks are vital in programming because they can help prevent null-reference errors, which can cause programs to crash. It is important to address and remove the Null value due to it represents missing values, such as when a user leaves an optional field blank in a form.



```
import seaborn as sns
import matplotlib.pyplot as plt

df.columns

Index(['Dst Port', 'Protocol', 'Timestamp', 'Flow Duration', 'Tot Fwd Pkts',
       'Tot Bwd Pkts', 'TotLen Fwd Pkts', 'TotLen Bwd Pkts', 'Fwd Pkt Len Max',
       'Fwd Pkt Len Min', 'Fwd Pkt Len Mean', 'Fwd Pkt Len Std',
       'Bwd Pkt Len Max', 'Bwd Pkt Len Min', 'Bwd Pkt Len Mean',
       'Bwd Pkt Len Std', 'Flow Byts/s', 'Flow Pkts/s', 'Flow IAT Mean',
       'Flow IAT Std', 'Flow IAT Max', 'Flow IAT Min', 'Fwd IAT Tot',
       'Fwd IAT Mean', 'Fwd IAT Std', 'Fwd IAT Max', 'Fwd IAT Min',
       'Bwd IAT Tot', 'Bwd IAT Mean', 'Bwd IAT Std', 'Bwd IAT Max',
       'Bwd IAT Min', 'Fwd PSH Flags', 'Bwd PSH Flags', 'Fwd URG Flags',
       'Bwd URG Flags', 'Fwd Header Len', 'Bwd Header Len', 'Fwd Pkts/s',
       'Bwd Pkts/s', 'Pkt Len Min', 'Pkt Len Max', 'Pkt Len Mean',
       'Pkt Len Std', 'Pkt Len Var', 'FIN Flag Cnt', 'SYN Flag Cnt',
       'RST Flag Cnt', 'PSH Flag Cnt', 'ACK Flag Cnt', 'URG Flag Cnt',
       'CWE Flag Count', 'ECE Flag Cnt', 'Down/Up Ratio', 'Pkt Size Avg',
       'Fwd Seg Size Avg', 'Bwd Seg Size Avg', 'Fwd Byts/b Avg',
       'Fwd Pkts/b Avg', 'Fwd Blk Rate Avg', 'Bwd Byts/b Avg',
       'Bwd Pkts/b Avg', 'Bwd Blk Rate Avg', 'Subflow Fwd Pkts',
       'Subflow Fwd Byts', 'Subflow Bwd Pkts', 'Subflow Bwd Byts',
       'Init Fwd Win Byts', 'Init Bwd Win Byts', 'Fwd Act Data Pkts',
       'Fwd Seg Size Min', 'Active Mean', 'Active Std', 'Active Max',
       'Active Min', 'Idle Mean', 'Idle Std', 'Idle Max', 'Idle Min', 'Label'],
      dtype='object')

print('Total columns in our data: %s' % str(len(df.columns)))

Total columns in our data: 80
```

Figure 5.3: Total Column

(Source: Self-created)

The above figure shows the total columns in the using dataset, it contains a total of 80 columns.

```
df['Label'].value_counts()

Label
Benign             665355
FTP-BruteForce     193354
SSH-Bruteforce     187589
Name: count, dtype: int64
```

Figure 5.4: Label

(Source: Self-created)

This figure shows the label of “Benign, FTP-BruteForce, SSH-BruteForce”. It has been identified that Benign is 665355, FTP-BruteForce 193354, SSH-Bruteforce 187589.

```
cleaned_data = df.dropna()
cleaned_data.isna().sum().to_numpy()
```

```
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], dtype=int64)
```

```
label_encoder = LabelEncoder()
cleaned_data['Label']= label_encoder.fit_transform(cleaned_data['Label'])
cleaned_data['Label'].unique()
```

```
array([0, 1, 2])
```

Figure 5.5: label Encoder

(Source: Self-created)

Label Encoder is a vital process in Python for deep learning and ML tasks, it also identified that most ML algorithms require numerical input. Label Encoder transforms categorical data such as colours, genders, or countries. It mostly uses numerical labels, making them compatible with these algorithms.

Labeling

```
X = df.drop(columns=non_numeric_columns)
y = df['Label']
```

```
X.replace([np.inf, -np.inf], np.nan, inplace=True)
```

```
X.fillna(X.mean(), inplace=True)
```

```
label_encoder = LabelEncoder()
y_encoded = label_encoder.fit_transform(y)
```

```
y_categorical = to_categorical(y_encoded)
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y_categorical, test_size=0.2, random_state=42)
```

```
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
```

```
model = Sequential()
model.add(Dense(64, input_dim=X_train.shape[1], activation='relu'))
model.add(Dense(32, activation='relu'))
model.add(Dense(y_train.shape[1], activation='softmax'))
```

Figure 5.6: labeling

(Source: Self-created)

The above figure shows labelling, deep learning a subset of ML that can use supervised, unsupervised, or semi-supervised learning methods. Labelling supports supervised learning algorithms that learn from labelled data, where each data point has a corresponding target variable. This labelling allows the algorithm to comprehend the connection between input elements and the desired output.

5.2 Result

```
sns.set(rc={'figure.figsize':(12, 6)})
plt.xlabel('Attack Type')
sns.set_theme()
ax = sns.countplot(x='Label', data=df)
ax.set(xlabel='Attack Type', ylabel='Number of Attacks')
plt.show()
```

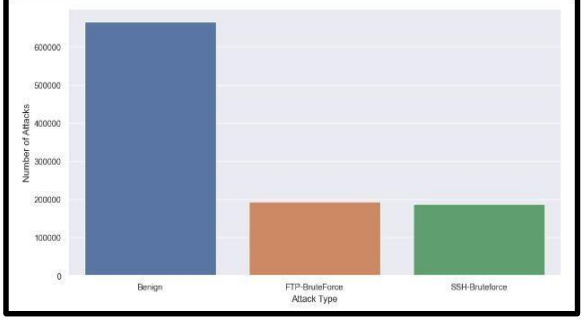


Figure 5.7: Attack type

(Source: Self-created)

The above figure shows the attack type with several attacks, in this bar graph X-axis represents the attack type, and the Y-axis represents the number of attacks. Through this figure, it has been comprehended that a total of three types of attacks are shown “Benign, FTP-BruteForce, SSH-BruteForce”. Benign has the highest attack number, and the other two have a closely similar number of attacks.

Figure 5.8: Scatter Plot

(Source: Self-created)

The above figure shows a scatter plot where the x-axis represents Bwd Pkts/s and the y-axis represents Fwd Seg Size Min. Notably,c Scatter plots are necessary because they can assist this research imagine and statistically explore relationships between two numeric variables. They can also assist in determining outliers and practices in the data.

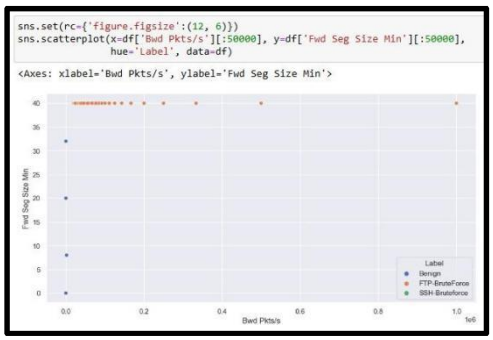


Figure 5.9: Scatter Plot

(Source: Self-created)

The above figure illustrates a scatter plot where the x-axis denotes backward Pkts/s and the y-axis designates Forward Segmentation Size Min. Notably,c Scatter plots are essential because they can assist this research imagine and statistically analysing relationships between two numeric variables. They can also assist in determining outliers and practices in the data.

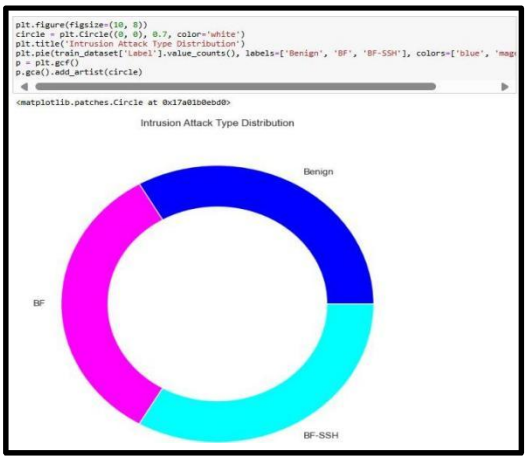


Figure 5.10: Pie Chart

(Source: Self-created)

The pie chart shows the different attack percentages, this Pie chart is particularly useful for analysis because it can help to analyse data and understand the relationship between parts of a whole, especially when this study visualizes a small number of categories. Pie charts are simple and also easy to understand, even for people who aren't familiar with statistics or data science.

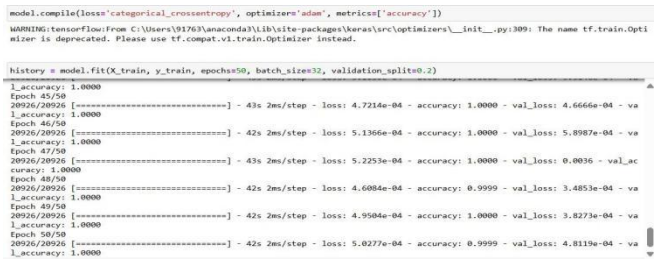


Figure 5.11: Epoch

(Source: Self-created)

The above figure shows 50 Epoch, it is determined that epoch is a vital image in deep learning and ML that particularly helps model learning and extends its performance over time. It essentially determines how numerous times the learning algorithm will run via this entire training dataset. It is also determined that during each epoch the model especially updates it is parameters based on the errors it produces when specifically indicating the output for each activity model.

Figure 5.12: Report

(Source: Self-created)

The above figure shows the accuracy of this epoch, it has been identified that the accuracy score is 1, which means that the accuracy score of this report is showing

100% accuracy. This passes 100 times on 100. Therefore it has been concluded that this report completely fit with this study.

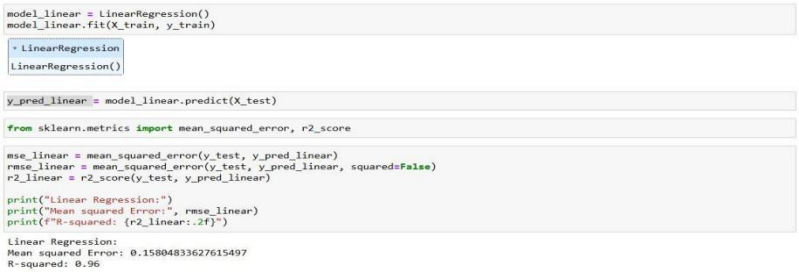


Figure 5.13: Linear Regression

(Source: Self-created)

The above figure shows established Linear regression, it is essentially a statistical technique that can be used to analyze data and reveal relationships between variables. It is used in different fields such as “business, science, and data science”. During the creation and assess a deep learning-based IDS for developing IoT environments using linear regression to convert raw data into actionable insights, such as measuring attacks and their effects.

5.3 Discussion

The study analyzed the deep learning model‘s elements to develop an Intrusion Detection System for providing security to Internet of Things devices to prevent and mitigate the threats of cyber attacks. Here Python language along with Jupyter Notebook in deep learning model has been implemented to provide an accurate result and develop an effective system. The deep learning model is a Machine learning algorithm which is able to learn patterns and its hybrid approaches can tackle various cybersecurity issues (Sarker., 2021). Here deep learning has been used in many cybersecurity tasks in the Internet of Things such as instruction detection, detection of malware or botnets, phishing, cyber attack prediction process and development of security techniques.

Here it implements deep learning models in the dataset to learn the complex patterns and predict the cyber-attacks to evaluate mitigating strategy. In the first step, it trained the dataset to remove any errors and preprocessed it through encoding and normalization with Deep learning models and artificial neural networks (ANNs). The data preprocessing is important to enhance the quality of raw data for further analysis (Fan et al.,2021). The preprocessed dataset is concerned with training dataset where the ML model Deep learning have been immense to predict the cyber-attacks as benign or attack for improving the Intrusion Detection System and securing the IOT devices.

Here it takes pandas and numpy libraries to have accurate coding which can impact the outcome of cyber attack detection. It loaded the data and checked the null elements in the dataset to explain further steps. It also defined the cyber attacks as ‘Benign’, ‘FTP-BruteForce’ and ‘SSH-Bruteforce’ to effectively detect the cyber threats as well as malware and attacks. Through this classification of cyber attacks, it can identify the specific attacks which have occurred in the dataset and based on that it can develop a reverse process to mitigate its impact. In the data analysis process, it discusses that the dataset has been attacked by Benign 670000 times. Comparatively, FTP-BruteForce has attacked 190000 times and SSH-Bruteforce attacked 180000 times. The number showed that Benign can implement the largest impact on IOT devices and has a negative outcome.

Here the scatter plot matrix has been used the represent the relationship among the examined features such as Benign, FTP-BruteForce and SSH-Bruteforce with the Internet of Things devices to empower the Intrusion Detection Systems (Karaman., 2022). By exploring the relationships between two numerical numbers in a scatter plot, it statistically anysed the co-relationship between the examined elements to visualize an accurate outcome. It also use linear regression to analyze the dataset and explain the relationship between variables such as classification of cyber attacks, IDS and IOT. Linear regression is a statistical tool which provide predictive techniques and relationship detection techniques in cybersecurity and antivirus application development processes. Here a deep learning model with linear regression is used to convert the dataset into a normalized version for measuring the cyber attacks. It also used to detect and analyze the cyber attack effect for performing a development process to develop a mitigating system in IOT devices.

Chapter 6: Conclusion & Recommendation

6.1 Conclusion

The entire research strives to develop and evaluate a deep learning-based IDS for enhancing protection in IoT settings. This investigation will shed light on earlier research on the creation and performance of a deep learning model tailored for noticing different kinds of “cyber threats” and intrusions within IoT networks. Also consider the implementation of the generated IDS in terms of accuracy, precision, recall, and across-the-board effectiveness in identifying and mitigating potential security infringements in IoT systems. Also, the proposed deep learning-based IDS with standard intrusion detection techniques highlight progress in detection abilities and reaction time.

The literature part recreates a significant role in completing this research the literature review part examines additional earlier analysis on the algorithm of deep learning to improve the efficiency and accuracy of IDS in the region of IoT. This study again investigates different academic papers on the practical approaches for the designation, as well as the alleviation of different cyber-attacks in IoT. This research paper also examines the deep learning-based architecture of IDS manages dynamic and heterogeneous IoT network traffic, and appliances. This section also examines earlier journals on the architectural frameworks of the deep learning instance that can be executed to address and mitigate the threats in IoT strategies, and appliances.

The methodology section covers the overall method which mainly used in this study. This study observes interpretivism due to it is a qualitative research philosophy that can help in achieving a more in-depth knowledge of the details of individuals and social life. This study also uses an inductive research approach which specifically supports to generation of innovative ideas for developing and assessing deep learning approaches in IDS for developing secure IDS.

The finding analysis section explores the data analysis entire section concentrates on a deep learning system for data analysis. During this study, it has been identified that deep learning is a vital AI technology that specifically allows to revolution of different initiatives and also helps get more wisdom form from this dataset. This study also analyses data by utilizing Linear regression, this province shows all data analysis methods and shows various graphs which allow the analysis of data.

6.2 Recommendation

- Based on the performance of the ANN, it can be recommended that deep learning models should be integrated into the Intrusion Detection Systems (IDS) for IoT devices. These models have held a superior accuracy and can handle complex, non-linear patterns in network traffic.
- As there is resource constraints in many IoT device, deployment of lightweight versions of the ANN or by using more computing power to

distribute the computational load can ensure that the IDS remains effective without overwhelming the IoT devices.

- Focus should be on optimizing the training process by using transfer learning, where the pre-trained data models can adapt to new IoT environments with a minimal retraing can improve the real time detection capabilities.
- By the combination of deep learning models with traditional ML models like the SVMs can help in improving the detection rates by increasing strength of each approach. This hybrid model can help in effective diversification of IoT ecosystems.
- Continuous learning should be implemented to update the IDS as new threats emerge. This involves periodic retraining of the models by using any latest network traffic data thus by maintaing high detection accuracy.
- Ensuring that the data used for training the IDS is handled securely for preventing any potential data breaches. Encryption and privacy should be the most important part to be considered when deployment of IDS across the IoT networks.
- As part of the future work, the model could be trained with different datasets to produce better results that can impact the analysis in a positive way.

6.3 Limitation

This study focused on deep learning-based IDS to improve secure IoT, although deep learning has several advantages still it has multiple limitations such as DL requiring large amounts of data to generalize well, which mainly limits its use in areas where data is scarce or hard to obtain.

The limitations also being the use of high computational complexity of deep learning models like the ANN, which is not suitable for resource constrained IoT devices.The models may struggle with continuous and high velocity data streams, so this remains challeging for real time applications. The data resampling that has helped in class imbalance can also lead to overfitting.The ANN models interpretability is limited which makes it difficult to understand decision making processes. So the project mostly focuses on Sequential Neural Network , which akso leaves other deep learning models unexplored. Also during using DL when training on difficult or high-dimensional data, deep learning models can overfit. It mostly means it learns too much from the training data and also it can not generalize to new data.The scalability of these models for millions of IoT devices could pose significant challenges as they require more computational cost and there is a need for distributed of processing solutions. DL can be expensive to compute and it is difficult to analyse. Deep learning can perpetuate existing biases in society if the data it is trained on reflects those biases. Another limitation is DL may not be ideal for deterministic difficulties that rely on computational modeling, such as weather forecasts. Also determined that Parallelized deep learning can be determined by memory problems, as all medium states and input mini-batches need to fit into the GPU's unreasonable memory. This study uses linear regression which is an effective approach, but it faces several limitations such as when two or more independent variables are highly correlated with each other. It can impact the stability and precision of the coefficients. In terms of linearity, the assumption of a straight-line relationship between variables is often false and may lead to inaccurate results.

Reference List

Ahmad, Z., Shahid Khan, A., Wai Shiang, C., Abdullah, J. and Ahmad, F., 2021. Network intrusion detection system: A systematic study of machine learning and deep learning approaches. Transactions on Emerging Telecommunications Technologies, 32(1), p.e4150.

Akhtar, M.S. and Feng, T., 2022. Detection of Malware by Deep Learning as CNN-LSTM Machine Learning Techniques in Real Time., Symmetry 2022, 14, 2308.
Alharahsheh, H.H. and Pius, A., 2020. A review of key paradigms: Positivism VS interpretivism. Global Academic Journal of Humanities and Social Sciences, 2(3), pp.39-43.

Alwahedi, F., Aldhaheri, A., Ferrag, M.A., Battah, A. and Tihanyi, N., 2024. Machine learning techniques for IoT security: Current research and future vision with generative AI and large language models. Internet of Things and Cyber-Physical Systems.

Alzahrani, A.O. and Alenazi, M.J., 2021. Designing a network intrusion detection system based on machine learning for software defined networks. Future Internet, 13(5), p.111.

Chatterjee, A. and Ahmed, B.S., 2022. IoT anomaly detection methods and applications: A survey. Internet of Things, 19, p.100568.

Esposito, C., Tamburis, O., Su, X. and Choi, C., 2020. Robust decentralised trust management for the internet of things by using game theory. Information Processing & Management, 57(6), p.102308.

HaddadPajouh, H., Dehghantanha, A., Parizi, R.M., Aledhari, M. and Karimipour, H., 2021. A survey on internet of things security: Requirements, challenges, and solutions. Internet of Things, 14, p.100129.

Khalil, M., McGough, A.S., Pourmirza, Z., Pazhoohesh, M. and Walker, S., 2022. Machine Learning, Deep Learning and Statistical Analysis for forecasting building energy consumption—A systematic review. Engineering Applications of Artificial Intelligence, 115, p.105287.

Kime, C. (2024) Network security architecture: Best practices & tools, eSecurity Planet. Available at: https://www.esecurityplanet.com/networks/network-security-architecture/#:~:text=Network%20security%20architecture%20matches%20security,effective%20controls%2C%20and%20measurable%20efforts. (Accessed: 22 July2024).

Li, D., Shi, X., Zhang, Y., Cheung, K.C., See, S., Wang, X., Qin, H. and Li, H., 2023. A simple baseline for video restoration with grouped spatial-temporal shift. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9822-9832).

Mallick, C., Mishra, S. and Senapati, M.R., 2023. A cooperative deep learning model for fake news detection in online social networks. Journal of Ambient Intelligence and Humanized Computing, 14(4), pp.4451-4460.

Mishra, N. and Pandya, S., 2021. Internet of things applications, security challenges, attacks, intrusion detection, and future visions: A systematic review. IEEE Access, 9, pp.59353-59377.

Mukhtar, B.I., Elsayed, M.S., Jurcut, A.D. and Azer, M.A., 2023. IoT vulnerabilities and attacks: SILEX malware case study. Symmetry, 15(11), p.1978.

Ni, J., Chen, Y., Chen, Y., Zhu, J., Ali, D. and Cao, W., 2020. A survey on theories and applications for self-driving cars based on deep learning methods. Applied Sciences, 10(8), p.2749.

Robinette, D. (2024) What are the detection methods of ids?, Network Threat Detection and Response. Available at: https://www.stamus-networks.com/blog/what-are-the-detection-methods-of-ids#:~:text=Signature%2DBased%20IDS:%20Signature%2D,attacks%20or%20suspicious%20system%20behavior. (Accessed: 22 July 2024).

Sharma, D.N.K., 2022. Instruments used in the collection of data in research. Available at SSRN 4138751.

State of IOT 2023: Number of connected IOT devices growing 16% to 16.7 billion globally (2024) IoT Analytics. Available at: https://iot-analytics.com/number-connected-iot-devices/ (Accessed: 22 July 2024).

Taye, M.M., 2023. Theoretical understanding of convolutional neural network: Concepts, architectures, applications, future directions. Computation, 11(3), p.52.

Tharewal, S., Ashfaque, M.W., Banu, S.S., Uma, P., Hassen, S.M. and Shabaz, M., 2022. Intrusion detection system for industrial Internet of Things based on deep reinforcement learning. Wireless Communications and Mobile Computing, 2022(1), p.9023719.

Vears, D.F. and Gillam, L., 2022. Inductive content analysis: A guide for beginning qualitative researchers. Focus on Health Professional Education: A Multi-Professional Journal, 23(1), pp.111-127.

von Rueden, L., Mayer, S., Sifa, R., Bauckhage, C. and Garcke, J., 2020. Combining machine learning and simulation to a hybrid modelling approach: Current and future directions. In Advances in Intelligent Data Analysis XVIII: 18th International Symposium on Intelligent Data Analysis, IDA 2020, Konstanz, Germany, April 27–29, 2020, Proceedings 18 (pp. 548-560). Springer International Publishing.

What is signature-based detection? Corelight. Available at: https://corelight.com/resources/glossary/signature-based-detection#:~:text=Signature%2Dbased%20detection%20is%20one%20of%20the%20most%20direct%20and,the%20specifics%20of%20the%20rule (Accessed: 22 July 2024).

Wintjen, M., 2020. Practical Data Analysis Using Jupyter Notebook: Learn how to speak the language of data by extracting useful and actionable insights using Python. Packt Publishing Ltd.

Wu, W., Qi, Q. and Yu, X., 2023. Deep learning-based data privacy protection in software-defined industrial networking. Computers and Electrical Engineering, 106, p.108578.

Zhang, X., Cui, J., Jia, Y., Zhang, P., Song, F., Cao, X., Zhang, J., Zhang, L. and Zhang, G., 2023. Image restoration for blurry optical images caused by photon diffusion with deep learning. JOSA A, 40(1), pp.96-107.

'MCIDS-Multi Classifier Intrusion Detection system for IoT Cyber Attack using Deep Learning algorithm | IEEE Conference Publication | IEEE Xplore'. https://ieeexplore.ieee.org/document/9388579 (accessed Aug. 01, 2024).

N. Balakrishnan, A. Rajendran, D. Pelusi, and V. Ponnusamy, 'Deep Belief Network enhanced intrusion detection system to prevent security breach in the Internet of Things', Internet of Things, vol. 14, p. 100112, Jun. 2021, doi: 10.1016/j.iot.2019.100112. (accessed: Aug 02,2024).