Cyber Security Intrusion Detection and Bot Data Collection using Deep Learning in the IoT

Fahad Ali Alotaibi[1], Shailendra Mishra[2]
Department of Information Technology, Majmaah University[1]
Department of Computer Engineering, Majmaah University[2]

Abstract—**In the digital age, cybersecurity is a growing concern, especially as IoT continues to grow rapidly. Cybersecurity intrusion detection systems are critical in protecting IoT environments from malicious activity. Deep learning approaches have emerged as promising intrusion detection techniques due to their ability to automatically learn complex patterns and features from large-scale data sets. In this research, we give a detailed assessment of the use of deep learning algorithms for cybersecurity intrusion detection in IoT contexts. The study discusses the challenges of securing IoT systems, such as device heterogeneity, limited computational resources, and the dynamic nature of IoT networks. To detect intrusions in IoT environments, convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been used. The NF-UQ-NIDS and NF-Bot-IoT data sets are used for training and assessing deep learning-based intrusion detection systems. Our study also explores using deep learning approaches to identify botnets in IoT settings to counter the growing threat of botnets. Also, analyze representative bot data sets and explain their significance in understanding botnet behavior and effective defenses. The study evaluated IDS performance and traffic flow in the IoT context using various machine learning algorithms. For IoT environments, the results highlight the importance of selecting appropriate algorithms and employing effective data pre-processing techniques to improve accuracy and performance. Cyber-attack detection with the proposed system is highly accurate when compared with other algorithms for both NF-UQ-NIDS and NF-BoT-IoT data sets.**

Keywords—Internet of things; intrusion detection system; random neural networks; feed forward neural networks; convolutional neural networks

I. Introduction

In the present era, technological advancements have been in the scope of the Internet of Things (IoT), cloud computing, and cybersecurity. In the next ten years, it is projected that the Internet of Things will grow enormously, with users adopting billions of IoT devices. The growth and expansion clarify the influence of technology through the IoT on matters of vulnerability and businesses and people's daily lives. Most enterprises, institutions, and government facilities increasingly adopt IoT technology since it can create a large amount of information used to test the function of the Internet of Things network, thus increasing the quality of the services and experience. Consequently, the Internet of Things makes data communication between actual equipment and sensors possible [1]. Employing technology elements in the Internet of Things connections has improved communication, evaluation, and the value of data collecting and projection for future strategy. Numerous layers comprise an Internet of Things building design, which looks into, recognizes, and monitors the network's reliability. The basic configuration comprises three tiers: awareness, system, and implementation [2].

On the other hand, deep learning mechanisms have become famous and popular for determining network breaches. Numerous literatures evaluate the comparison of deep learning structures, particularly the new data components for detecting Intrusion. As a result, the definition of the Internet of Things Intrusion is any illegal activity or conduct that affects the confidentiality of the IoT network, data availability, and integrity in any way [3]. Using virtual private networks (VPNs), safe and protected communication channels are created to safeguard the privacy and integrity of transmitted data. When an intruder blocks entrance to a service, preventing legitimate users from using it, this is called an incursion. An intrusion detection system (IDS) is a tool that monitors systems and networks on computers using hardware, software, or both to spot malicious or dangerous activity and to maintain the network and system safe. In response to this, deep learning can be utilized to assist in determining dangerous attacks on IoT networks and connections while minimizing risks and enhancing active deterrence of future attacks. In retrospect, the paper offers insights into deep learning-based approaches for cyber security intrusion detection and bot data collected in the Internet of Things [1].

In this paper, we propose a novel intelligent intrusion detection system (IDS) that performs feature extraction, feature selection, and intelligent classification via efficient rule matching, also by performing deductive inference. This study also includes a complete assessment of IDSs in the IoT, which highlights the advantages, benefits, and limitations of the existing IDSs for the IoT environment and compares them to the proposed work. The comprehensive literature survey, the identification of suitable metrics for comparison, the measurement of various parameters more efficiently by identifying the granularity of the measurement, and finally the proposal of a new IDS using deep learning techniques are the major contributions of this work. Based on the results of the tests conducted in this paper, it is discovered that the suggested intelligent IDS is more successful in terms of intrusion detection rates as well as false positive rates reduction.

The proposed study primarily focuses on deep learning-based approaches for cyber security intrusion detection and bot data collected in the Internet of Things. Several data pre-processing techniques were used to increase IDS quality. The instance-based and feature-based techniques were specifically explored. Instance-based pre-processing is concerned with data cleansing and removal strategies. Feature-based pre-processing comprises feature transformation, normalization, and dimensionality reduction through correct feature selection. Feature transformation was applied to all of the categorical characteristics of the selected datasets.

The main objectives of the proposed study are:

1) *To* offer systematic insight into the scholarly articles on detecting Intrusion on the IoT.

2) *To* evaluate and offer a comprehension of the procedures and techniques used to analyze the effect of information and algorithm quality on the heightening network intrusion detection rates.

3) *Proposed* an Intrusion detection system (IDS) based on deep learning (DL), for effective security in the IoT environment

4) *To* evaluate the performance of the proposed IDS.

By focusing on these goals, the research seeks to improve the comprehension, efficiency, and expandability of IDS systems, thereby bolstering the security and dependability of IoT networks. This research provides more accurate and reliable forecasts, which bolster Internet of Things security by thwarting data breaches, unauthorized entry, and service denials. Utilizing Python algorithms to tackle discrepancies in class problems within IoT cybersecurity databases subsequently boosts the capabilities of the generated models. Eventually, the research guides the most efficient approaches for utilizing neural networks and deep learning as sensors to predict cybersecurity issues.

The organization of the paper is as follows; Section II shows the related work, Section III represents the methodology of the proposed work, Section IV includes experimental setup, Section V discusses results and analysis, and Section VI shows the conclusion and future work.

II. Related Work

This section provides an in-depth study of the relevance of cybersecurity in IoT infrastructure by evaluating previous research and examining the progress achieved using ML and DL approaches. This highlights the importance of IoT security and the challenges faced due to the lack of IDS and the need for IDS in IoT networks. This section discusses current research on deep learning-based intrusion detection systems for IoT applications, focusing on identifying the research needs of this topic. To prevent cyberattacks and provide security solutions for lightweight IoT networks, many research challenges need to be addressed. Cars, health monitoring, robots, and smart homes will generate large amounts of data, requiring new security measures. Although some researchers have used machine learning and deep learning methods to develop and deploy IoT intrusion detection systems (IDS) in recent years, further research on IoT intrusion detection is still needed.

The deep learning approaches can be effective in cyber security intrusion detection in the context of the Internet of Things (IoT).The use of a Deep Learning-based Intrusion Detection System (IDS) using Feed Forward Neural Networks (FFNN), Long Short Term Memory (LSTM), and Random Neural Networks (RandNN) to enhance IoT network security and reduce cyber threats [1]. Sarah Alkadi et al., [2] discuss an empirical impact analysis of machine learning (ML) in building intrusion detection systems (IDSs) for IoT networks. The study found that using quality data and models, such as data cleaning, transformation, normalization, and parameter tuning, significantly improves IDS detection accuracy. The intelligent detection system is proposed in [3], using SVM, SMOTE, machine learning, and deep learning algorithms. The model achieved good accuracy and reduced error rates.

The potential of machine learning and deep learning techniques in detecting malware in IoT networks is discussed in [4]. It evaluates the efficacy of ten models and their performance when combined with the SMOTE algorithm to counterbalance imbalanced data. The effectiveness of the Rules and Decision Tree-based Intrusion Prevention System RDTIPS is a new intrusion prevention system for the Internet of Things networks discussed in study [5], which combines rules and decision trees, it demonstrates a superior performance, accuracy, detection rate, time overhead, and false alarm rate.

A new approach, using the focal loss function, improves accuracy, precision, score, and MCC score compared to traditional methods discussed in study [6]. Researchers have used Machine Learning techniques for intrusion detection, but imbalanced datasets can lead to unsatisfactory results. In the paper [7], the authors propose a novel approach using deep learning and three-level algorithms to detect cyber-attacks in IoT networks, demonstrating significant improvements in detection performance and potential for other IoT applications.

Paper in [8], this article provides statistics and architectures for IoT botnets and analyses the attacks in depth, but it is also susceptible to cyberattacks. To counter this, a new method of detection for intruders is proposed in the GA-FR-CNN framework. This method employs Deep Learning and FR-CNN and has a high degree of success on the UNSW-NB 15 and BOT-IoT datasets. The rapid growth of IoT devices, including wearables and smart sensors, has led to an increase in cyberattacks [9]. To surmount this, authors in [9] utilize both machine and deep learning. The Internet of Things (IoT) is a growing market, leading to increased cyber-attacks. To combat this, researchers [10], propose a hybrid approach using Autoencoder and Modified Particle Swarm Optimization (HAEMPSO) for feature selection and deep neural network (DNN) for classification. The proposed HAEMPSO-DNN achieved high accuracy and detection rates compared to existing machine-learning schemes.

The fourth industrial revolution has led to the generation of large-scale data in Industrial Internet of Things (IIoT) platforms, increasing security risks and data analysis procedures. The paper in [11] proposes an ensemble deep learning model using Long Short Term Memory and Autoencoder architecture to identify out-of-norm activities in IIoT cyber threat hunting. The industrial Internet of Things (IIoT) generates sensitive data, making security mechanisms like intrusion detection systems impractical. Federated learning and Blockchain are promising advancements to address these challenges [12]. The study in [12], explores the role of

Blockchain and federated learning in IIoT, highlighting potential applications in monitoring network traffic for anomaly detection and providing recommendations for effective implementation.

In research [13], the authors discuss a network intrusion detection (NID) method for IoT using a lightweight deep neural network. The method uses the PCA algorithm for feature reduction, expansion and compression structures, and NID loss for effective feature extraction. In [14], the authors discussed the fundamental principles of deep learning and machine learning, with 80 studies selected between 2016 and 2021, and discussed about the effectiveness of support vector machines, random forests, XGBoost, neural networks, and recurrent neural networks.

In research [15], the authors discuss the Internet of Things (IoT)'s influence on intelligent objects by decreasing power consumption. However, these devices are susceptible to invasions because of their direct association with the perilous Internet. Intrusion detection systems (IDSs) have a significant role in addressing these weaknesses, studying their principles, and recognizing potential dangers. The vulnerability of IoT systems to cyber-attacks focuses on learning-based methods and their impact on devices [16]. It reviews various types of attacks, presents literature on these developments, and provides future research directions. In study [17], authors discussed traditional and machine learning NIDS techniques, discussing future directions and enabling security professionals to differentiate IoT NIDS from traditional ones. In [18], the authors explore the vulnerability detection methods in IOT environments using machine learning, they propose a framework for recognizing potential vulnerabilities and reviewing the current state of the art.

In study [19] authors proposed a method of deep learning that is federated to improve the security of cyberphysical systems in the context of IOT, the performance of this method is evaluated in real IOT datasets. It demonstrates that these approaches are more effective at preserving device data privacy and recognizing attacks. Paper [20], discussed the increasing number of internet-connected devices (IoT) that pose a threat to the safety of the network. Traditional solutions based on rules fail to recognize these attacks. Machine learning (ML) is utilized for the detection of IoT attacks, the focus of this approach is on botnet attacks that target multiple devices.

In study [21], authors proposed a Deep Intrusion Detection (PB-DID) architecture, which classifies non-anomalous, DoS, and DDoS traffic uniquely using deep learning techniques, achieving a high accuracy (96%). In [22], the authors, proposed the Hybrid Intrusion Detection System (HIDS), combining the C5 decision tree and the One-Class Support vector machine to identify intrusions with a high degree of accuracy and a low rate of false alarm. The HIDS is assessed using the Bot-IoT dataset, this demonstrates a higher degree of detection and a lower percentage of false positives. Authors in [23], proposed a CorrAUC that uses a feature selection metric and an algorithm to filter features accurately. The procedure is assessed using the dataset of the Bot-IoT and four different machine learning methods, the average accuracy of which is over 96%.In [24], authors proposed a method of intrusion detection for IoT devices that utilizes machine learning to identify anomalous traffic in the network. The system uses binary grey wolf optimizer, recursive feature elimination, synthetic minority oversampling technique, XGBoost, Bayesian, and classification optimization with a tree-structured Parzen estimator.

In study [25], the authors examine the increasing cybersecurity challenges in the context of IoT technologies, which are increasingly vulnerable to cyber threats. It compares the effectiveness of machine learning methods like Support Vector Machine (SVM), Artificial Neural Network (ANN), Decision Tree (DT), Logistic Regression (LR), and k-nearest Neighbours (k-NN) in detecting cyber anomalies in IoT systems. The results show that the neural network outperforms other models, providing valuable insights for cybersecurity experts and guiding the development of robust protection strategies for the IoT ecosystem. Yaras, Sami, and Murat Dener, study employs PySpark and Apache Spark to analyze network traffic data and detect attacks using a deep learning algorithm, achieving high accuracy rates [26].

In study [27], the others, Yesi Novaria Kunang et al., propose a hybrid deep learning model for an intrusion detection system (IDS) on the IoT platform, using unsupervised approaches for feature extraction and a neural network for classification. The model demonstrated high detection performance and improved recognition of attacks compared to previous approaches. In [28], the authors introduce a framework that suggests selecting a suitable source domain data set for transfer learning, ensuring the highest accuracy in small-scale environments like home networks. Amit Kumar Mishra et al.[29], introduce a weighted stacked ensemble model for IoT networks, enhancing performance and reducing generalization error.

Mohanad Sarhan et al. present five NIDS datasets with a popular NetFlow feature set to bridge the gap between academic research and real-world deployments [30]. As part of the experiments, four benchmark NIDS datasets were labeled for traffic and attack classification experiments, and the results were evaluated using an Extra Trees ensemble classifier. For the NF-UQ-NIDS dataset, accuracy and recall values are not provided. Precision is reported as 70.81%, and F1-score is reported as 79%. For the NF-UQ-NIDS dataset, this information indicates how well the classification model performed in terms of precision and F1 score. The F1-score is given as 77%, and the precision is recorded as 73.58%. Using benchmark Net-flow-based datasets and machine learning techniques to address security issues.

In study [31], authors proposed a deep neural network-based intrusion detection system for real-time attack detection of malicious packets in IoT networks. They presented their findings, reporting an accuracy of 91.7%, precision of 91%, recall of 91%, and an F1-score of 91%. For the NF-UQ-NIDS dataset, our research yielded an accuracy, precision, recall, and F1-score of 92%. An accuracy of 76%, precision of 76%, recall of 76%, and F1-score of 70% are reported for the NF-BoT-IoT dataset. Using the NF-BoT-IoT dataset.

Most of the existing intrusion detection systems (IDSs) discussed in the related work section are general in nature and focused on network security, and most of them do not concentrate on the application of deep learning-based computational intelligence for constructing a reliable intrusion detection system. As a result, they are ineffective in delivering effective security in the IoT environment. The present IoT communication requires the deployment of a more flexible and efficient security system capable of detecting both known and innovative forms of threats and preventing them more intelligently utilizing artificial intelligence (AI) and machine learning (ML) methods. A comprehensive evaluation of IDSs in the IoT environment is also included, which highlights the advantages, benefits, and limitations of existing IDSs.

A. Research Gaps

Many existing (IDS) published in the literature are generic in nature and focus on network security, and most do not use deep learning-based computer intelligence to create reliable systems. As such, they are ineffective at providing effective security in the IoT context. The current communication style for IoT necessitates the implementation of a more flexible and efficient security system that can recognize both known and innovative threats and prevent them more effectively using AI and ML methods.

This paper proposes a new intelligent detection system for intrusion (IDS) that extracts features, chooses features, and categorizes instances via efficient rule matching, additionally, it also involves deductive reasoning. This study also contains a thorough review of IDSs in the IoT, which emphasizes the benefits, advantages, and limitations of existing IDSs for the IoT context and compares them to the suggested approach. This research makes major contributions by conducting an exhaustive literature assessment, identifying acceptable metrics for comparison, efficiently measuring multiple parameters through the identification of their granularity, and proposing a novel IDS based on deep learning. The test results conducted in this paper show that the proposed intelligent IDS has a high success rate in both detecting intrusions and reducing the number of false alarms.

III. RESEARCH METHODS

The proposed intelligent intrusion detection system (IDS) is shown in Fig. 1. It includes selecting a dataset, pre-processing the data, selecting relevant features, splitting the dataset, labeling the data (if applicable), performing classification, and deriving results. The proposed intelligent intrusion detection system (IDS) utilizes efficient rule matching and deductive inference to perform feature extraction, selection, and intelligent classification.

A. Select Dataset

The NIDS dataset NF-UQ-NIDS [32], simulates a realistic network environment with both normal and abnormal traffic. The dataset includes DDoS, Reconnaissance, Injection, DoS, Brute Force, Password, XSS, Infiltration, Exploits, Scanning, Fuzzers, Backdoor, Bot, Generic, Analysis, Theft, Shellcode, MITM, Worms, and Ransomware attacks The data set NF-BoT-IoT [33] simulates a realistic network environment with both normal and botnet traffic. The dataset includes Reconnaissance, DDoS, DoS, and Theft.

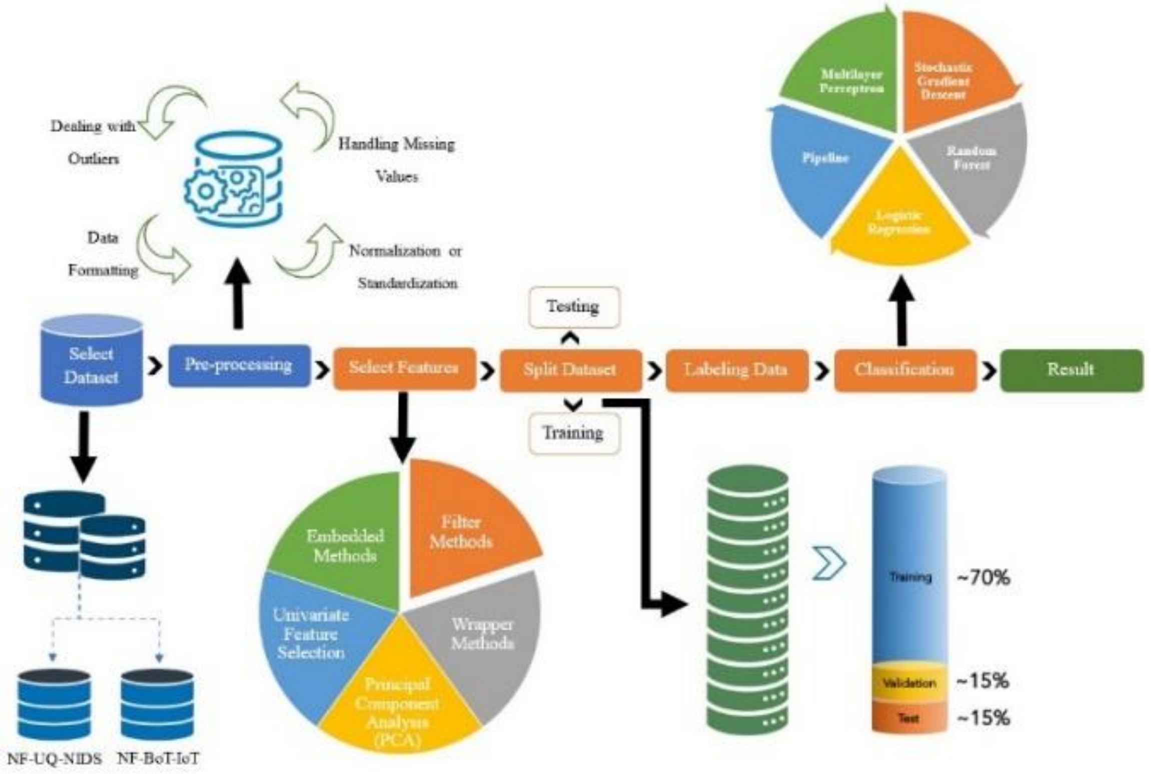

Fig. 1. Proposed intelligent intrusion detection system (IDS).

B. *Pre-processing*

Pre-processing is an essential process in cleaning and preparing the dataset for analysis. It consists of numerous duties, including:

1) Handling missing values: Identify and address any missing values in the dataset. This can be accomplished by either imputing missing values using statistical methods or removing instances or features with missing values, depending on the quantity of missing data and its impact on the study.

2) Dealing with outliers: Identify and address extreme values that deviate significantly from the normal distribution. Outliers can alter analytical results, which can be treated by deleting them, modifying the data, or employing strong statistical procedures.

3) Data formatting: Ensure data is properly formatted for analysis. Convert variables to the relevant data types (e.g., numerical, categorical, DateTime) and, if necessary, standardized units.

4) Normalization or standardization: Convert data to a consistent scale. This is particularly important when utilizing algorithms sensitive to variable magnitude and comparing variables on different scales.

C. *Select Features*

Selecting appropriate features from a dataset is crucial for achieving research objectives. There are several feature selection methods for the dataset, including:

1) Filter Methods:

a) Correlation-based feature selection: Use the correlation coefficient to determine the degree to which each feature is associated with the target variable and choose the feature with the greatest association.

2) Wrapper Methods:

a) Recursive Feature Elimination (RFE): Train the model repeatedly and remove the least significant feature at each iteration based on the model's performance, until the desired number of features is achieved.

b) Forward selection: This method builds a model by sequentially adding the most significant component that enhances the model's performance until a stopping rule is reached.

c) Backward elimination: It begins with all the features and then removes the least significant feature by using a defined criterion to stop when a stopping condition is met.

3) Embedded Methods:

a) LASSO (Least Absolute Shrinkage and Selection Operator): It employs regularization that is least absolute in nature, this type of regularization is used to penalize the coefficients of features, and thus, some of the features will become zero and the remaining will be selected.

b) Ridge Regression: It uses regularization via L2 to reduce the magnitude of the coefficients of less significant features to zero, this diminishes their influence on the model.

c) Elastic Net: A hybrid of L1 and L2 that enables feature selection and addresses multicollinearity.

4) Principal Component Analysis (PCA): Reduces the dataset's dimensionality by altering the original features into a new collection of uncorrelated variables known as principal components. The primary components have the greatest amount of variance in the data.

5) Univariate Feature Selection:

a) SelectKBest: Select the top features based on statistical tests such as ANOVA F-score or mutual information.

b) SelectPercentile: Selects the highest percentage of features based on a statistical test.

D. *Split Dataset*

Separate the dataset into training, validation, and testing subsets. The typical dividing ratio is 70% training, 15% validation, and 15% testing. The training dataset is used to train the classification model, the validation dataset is used to tune hyperparameters and choose models, and the testing dataset is used to perform the final evaluation.

E. *Labeling Data*

Label the data appropriately; this step entails adding class labels or categories to each data object. Human annotators can manually label items, or existing labels can be utilized to give reliable training data for the classification model.

F. *Classification*

We used appropriate classification techniques to train on the labeled training datasets. There are several classification methods available, including stochastic gradient descent (SGD), random forest, multilayer perceptron (MLP), pipeline, and logistic regression.

G. *Result*

We assessed the effectiveness of the trained classification model using the testing dataset, which included the dataset, such as accuracy, precision, recall, and F1 scores, to determine the model's effectiveness.

IV. Experimental Setup

The implementation was conducted on a Windows 10 operating system desktop, with hardware specifications that included 8 GB of RAM, an Intel(R) Core (TM) i7-10700 processor, Jupyter notebooks 7.0.6, and Python 3.12 as the programming language employed, with pandas, Scikit-Learn, NumPy, and Matplotlib that provided data processing and visualization functionality for our experiments. Pandas was used for data processing and preparation, Scikit-learn for machine learning methods and evaluation, Numpy for numerical computations, and Matplotlib for visualization. Python libraries are used to create machine learning models to identify intrusions in IoT networks, and with these modules, we were able to create a versatile and powerful analysis framework that can be readily extended and adjusted to meet my needs. To prepare the environment for studying the dataset, a VMware Workstation is installed, and then a Windows 10

VM and Python 3.12 are installed and run commands as shown in Table I, to prepare the environment.

TABLE I. COMMAND TO PREPARE THE ENVIRONMENT

Command	Description
pip install notebook	Installation Jupyter Notebook.
Pip install pandas	Installing pandas from PyPI.
pip install pandas numpy	Installing numpy from PyPI.
pip install matplotlib	Installing matplotlib from PyPI.
pip install -U scikit-learn	Installing scikit-learn from PyPI.
Jupyter Notebook	Running Jupyter Notebook

A. Evaluation Metrics

Accuracy, recall, precision, and F1 score are the assessment measures utilized in this research to analyze the performance of the suggested model. True positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) are the metrics used to construct these measures.

1) Accuracy: The accuracy of a model's predictions is calculated by dividing the number of successfully classified occurrences by the total number in the dataset as shown in Eq. (1).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

2) Precision: Precision measures the model's ability to correctly identify positive cases out of all those projected as positive. It is calculated using the Eq. (2):

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

3) Recall: Recall measures the model's ability to properly identify positive instances from among all positive examples in the dataset as shown in Eq. (3).

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

4) F1 Score: The F1 score is a numerical system that combines both precision and recall to have a single measurement. It provides a comprehensive evaluation of a model's effectiveness. The F1 rating is derived from the following Eq. (4):

$$F_1 = \frac{2*precision*recall}{precision+recall} \tag{4}$$

In the domain of intrusion detection, the following are the specific definitions of TP, FP, FN, and TN: The classification of an actual threat as a threat is called TP. The process of designating a typical normal behavior as a crime is called FP. The process of designating an actual category of crime as a normal counterpart is called FN. The formal designation of a typical normal category as a typical normal category is called TN.

Preprocessing is necessary following the collection of data. This stage involves, among other things, the cleaning of data, oversampling, selection of features, data normalizing, and partitioning of the dataset. In the cleaning stage, remove any duplicate values from the data set and replace any empty values with zeros. To mitigate the issue of uneven data distribution and reduce the impact of the problem on experimental results, we employ the SMOTE method to oversample the minority class. The dataset is normalized once all features have been converted to numerical types. The regularization approach is used to standardize the dataset, scaling the values between [0, 1]. This normalizing procedure improves the model's convergence speed and training effectiveness.

Machine learning relies heavily on data. When data is noisy and unpredictable, it can be incredibly difficult to analyze. Underfitting happens when training data cannot accurately establish a link between inputs and outputs. Overfitting occurs when a machine learning model performs poorly after being trained on a huge amount of data. As a result of the noisy and skewed data, the algorithm's performance will suffer. Machine learning is a very new and fast-expanding science. Learning is difficult since there are numerous opportunities for error because the process is always changing. The most important step in the machine learning process is data training. Predictions will be excessively biased or erroneous in the absence of sufficient training data. Slow implementation is one of the most common issues that machine learning specialists face. Machine learning models are quite good at providing proper results, even though it takes a long time. The algorithm may become flawed as the amount of data increases.

V. RESULTS AND DISCUSSION

A. NF-UQ-NIDS

By using Stochastic Gradient Descent, a machine learning optimization algorithm, to discover the model parameters that correspond to the best fit between expected and actual outputs. The result is shown in Table II and the confusion matrix for SGD is shown in Fig. 2.

TABLE II. SGD PERFORMANCE OF THE EVALUATED IN NF-UQ-NIDS

SGD	Accuracy	88%
	Precision	89%
	Recall	88%
	F1	87%

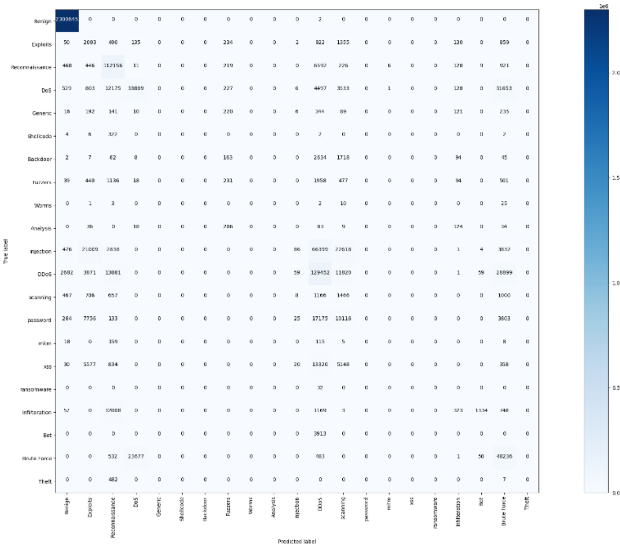

Fig. 2. SGD confusion matrix for NF-UQ-NIDS.

By utilizing a random forest classifier, it combines the votes of different decision trees to determine the final classification of the test object. The outcomes are listed in Table III, and a confusion matrix for RF is displayed in Fig. 3.

TABLE III. RF PERFORMANCE OF THE EVALUATED IN NF-UQ-NIDS

RF	Accuracy	92%
	Precision	92%
	Recall	92%
	F1	92%

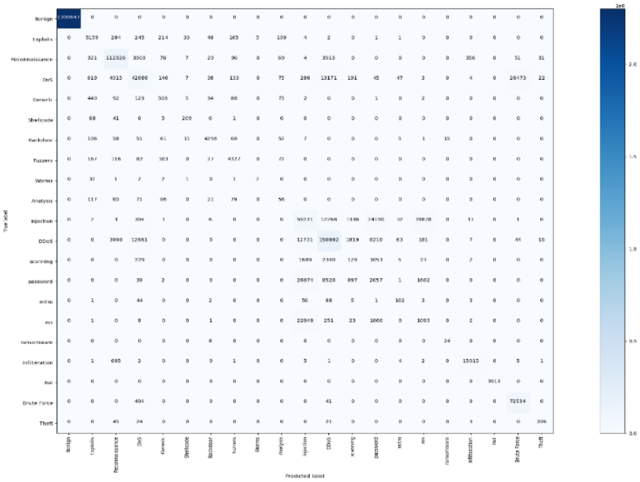

Fig. 3. RF confusion matrix for NF-UQ-NIDS.

By using an MLP Classifier that relies on an underlying Neural Network to perform the task of classification. The result is shown in Table IV, and the confusion matrix for MLP is shown in Fig. 4.

TABLE IV. MLP PERFORMANCE OF THE EVALUATED IN NF-UQ-NIDS

MLP	Accuracy	94%
	Precision	93%
	Recall	94%
	F1	93%

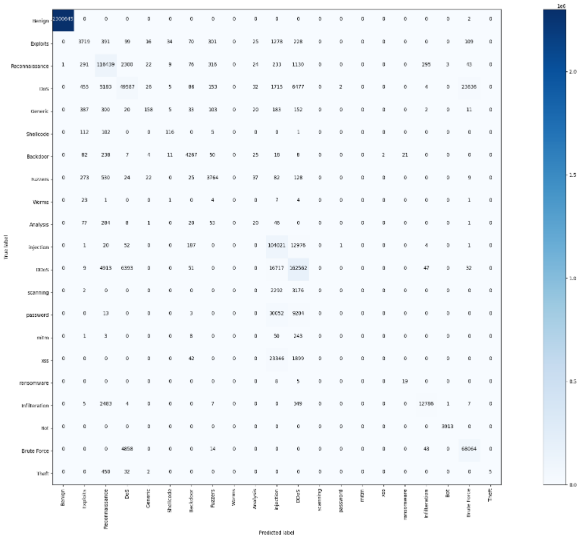

Fig. 4. MLP confusion matrix for NF-UQ-NIDS.

By using pipelines, each step is repeated to continuously increase the model's accuracy and establish a successful method. The result is shown in Table V and the Confusion matrix for PIPE is shown in Fig. 5.

TABLE V. PIPE PERFORMANCE OF THE EVALUATED IN NF-UQ-NIDS

PIPE	Accuracy	91%
	Precision	92%
	Recall	91%
	F1	90%

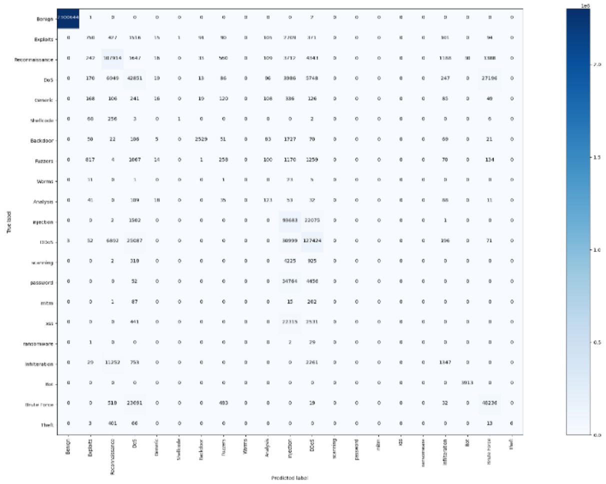

Fig. 5. Pipe confusion matrix for NF-UQ-NIDS.

By using logistic regression, we can classify the probability of certain classes based on some dependent variables. The result is shown in Table VI, and the Confusion matrix for LR is shown in Fig. 6.

TABLE VI. LR PERFORMANCE OF THE EVALUATED IN NF-UQ-NIDS

LR	Accuracy	90%
	Precision	91%
	Recall	90%
	F1	89%

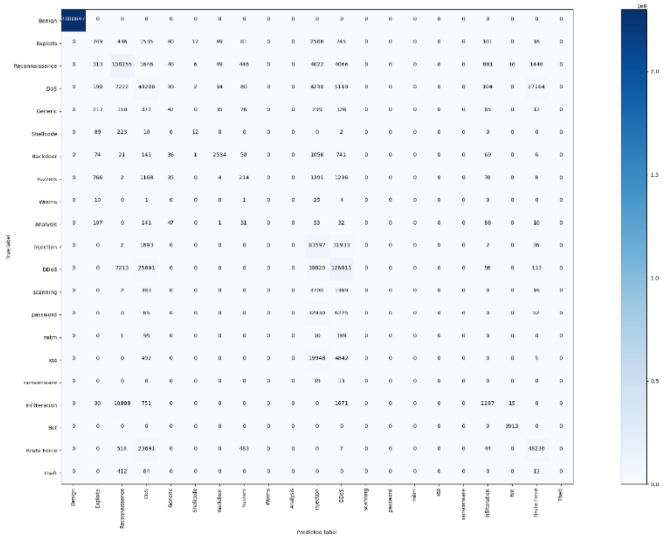

Fig. 6. Confusion matrix for NF-UQ-NIDS.

B. NF-BoT-IoT

By using Stochastic Gradient Descent, a machine learning optimization algorithm, to discover the model parameters that correspond to the best fit between expected and actual outputs. The results are shown in Table VII, and the confusion matrix for SGD is shown in Fig. 7.

TABLE VII. SGD PERFORMANCE OF THE EVALUATED IN NF-BOT-IOT

SGD	Accuracy	80%
	Precision	81%
	Recall	80%
	F1	75%

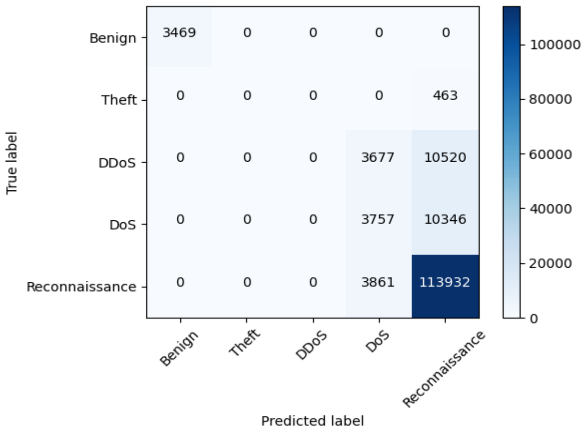

Fig. 7. SGD confusion matrix for NF-BoT-IoT.

By utilizing a random forest classifier, it combines the votes of different decision trees to determine the final classification of the test object. The outcomes are listed in Table VIII, and a confusion matrix for RF is displayed in Fig. 8.

TABLE VIII. RF PERFORMANCE OF THE EVALUATED IN NF-BOT-IOT

RF	Accuracy	77%
	Precision	77%
	Recall	77%
	F1	77%

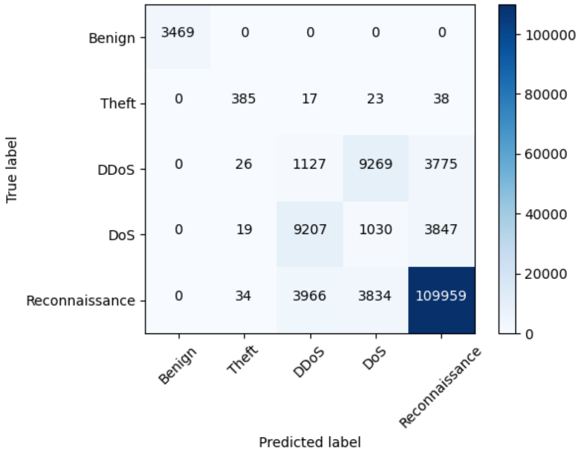

Fig. 8. RF confusion matrix for NF-BoT-IoT.

By using an MLP Classifier that relies on an underlying Neural Network to perform the task of classification. The result is shown in Table IX, and the confusion matrix for MLP is shown in Fig. 9.

TABLE IX. MLP PERFORMANCE OF THE EVALUATED IN NF-BOT-IOT

MLP	Accuracy	84%
	Precision	84%
	Recall	84%
	F1	82%

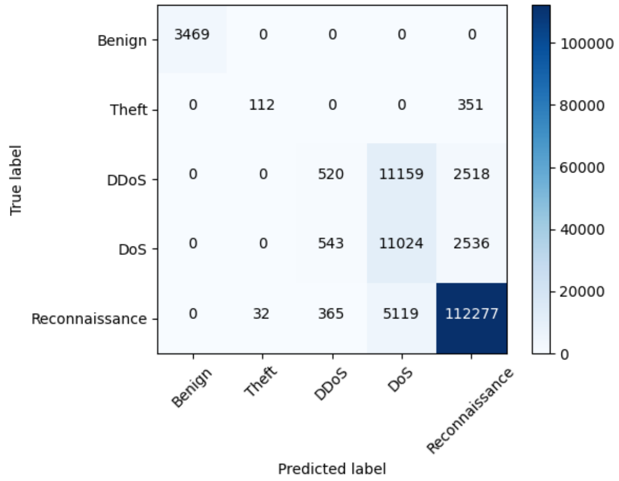

Fig. 9. MLP confusion matrix for NF-BoT-IoT.

By using pipelines, each step is repeated to continuously increase the model's accuracy and establish a successful method. The result is shown in Table X, and the Confusion matrix for pip is shown in Fig. 10.

TABLE X. PIPE PERFORMANCE OF THE EVALUATED IN NF-BOT-IOT

PIPE	Accuracy	80%
	Precision	75%
	Recall	80%
	F1	77%

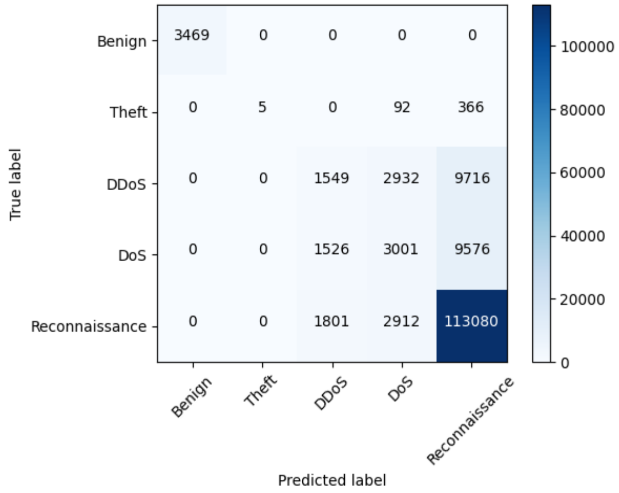

Fig. 10. PIPE confusion matrix for NF-BoT-IoT.

By using logistic regression, we can classify the probability of certain classes based on some dependent variables. The result is shown in Table XI, and the Confusion matrix for LR is shown in Fig. 11.

TABLE XI. LR PERFORMANCE OF THE EVALUATED IN NF-BOT-IOT

LR	Accuracy	80%
	Precision	75%
	Recall	80%
	F1	76%

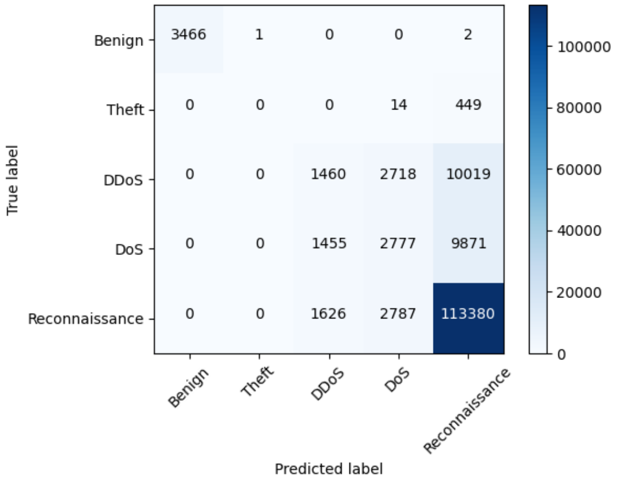

Fig. 11. LR confusion matrix for NF-BoT-IoT.

C. Multi-class Classification

In the multi-class classification task, several classification models were evaluated using two datasets: NF-UQ-NIDS and NF-BoT-IoT. The performance measurements of these models are summarized and compared in Table XII.

In terms of accuracy, precision, recall, and F1, the proposed IDS based on an RF and MLP outperforms better than other techniques reported in [30, 31]. Classifiers correctly classify attacks when the ability to define the class in which an attack is detected determines true positives and false positives. When a true negative was obtained, the classifier correctly discarded attacks. When a false negative was found, the classifier classed the attacks as normal traffic.

The accuracy rate for suggested and current strategies for detecting cyberattacks. The results show that applying algorithms to rank the features allows for the extraction of the desired data. The quantity of usable features in the proposed algorithms influences how well the jointly learned feature transformation operates, allowing for easier IDS fine-tuning. The algorithms do impact how well the jointly learned feature. Recall rate comparison displays the number of usable features in the proposed algorithms affects how well the jointly learned feature transformation performs. The F1 score for the number of features in the specified databases for the proposed algorithms affects how effectively the jointly learned feature transformation performs.

TABLE XII. MULTI-CLASS CLASSIFICATION

Classification Models		Measurements	Datasets	
			NF-UQ-NIDS	NF-BoT-IoT
Sarhan et.al.; [30]		Accuracy	-	-
		Precision	70.81%	73.58%
		Recall	-	-
		F1	79%	73.58%
Thirimanne et, al; (RF) [31]		Accuracy	91.7%	76%
		Precision	91%	76%
		Recall	91%	76%
		F1	91%	70%
Proposed IDS based on DL	SGD	Accuracy	88%	80%
		Precision	89%	81%
		Recall	88%	80%
		F1	87%	75%
	RF	Accuracy	92%	77%
		Precision	92%	77%
		Recall	92%	77%
		F1	92%	77%
	MLP	Accuracy	94%	84%
		Precision	93%	84%
		Recall	94%	84%
		F1	93%	82%
	PIPE	Accuracy	91%	80%
		Precision	92%	75%
		Recall	91%	80%
		F1	90%	77%
	LR	Accuracy	90%	80%
		Precision	91%	75%
		Recall	90%	80%
		F1	89%	76%

D. Discussion

The proposed investigation has listed several goals that are intended to contribute to the field of intrusion detection via the Internet of Things (IoT). To evaluate and offer a comprehension of the procedures and techniques used to analyze the effect of information and algorithm quality on the heightening network intrusion detection rates. Understanding the process and methodology used to evaluate the influence of data and model quality is fundamental to improving IDS detection rates. By studying existing approaches, the study can identify the factors that impact the performance of IDS, such as data pre-processing techniques, feature selection, and the choice of machine learning algorithms. This objective will help in determining effective strategies for enhancing the quality of data and models, ultimately leading to improved IDS performance.

Identifying and analyzing attacks on IoT networks is crucial for developing robust intrusion detection systems. By analyzing traffic data, the study can uncover patterns and anomalies that indicate the presence of attacks. The findings will enhance the ability to detect and mitigate these attacks effectively.

Analyzing the time and spatial complexity associated with IDS in IoT networks is crucial for understanding the scalability and performance limits of intrusion detection systems. By examining the computational requirements, resource utilization, and performance trade-offs, the study can provide insights into the feasibility and limitations of implementing IDS in large-scale IoT deployments. This objective will contribute to optimizing the resource allocation and efficiency of IDS in IoT environments. Several ML approaches were adopted in this study to assess IDS performance and model traffic flow in the IoT context. Stochastic Gradient Descent (SGD), Random Forest Classifier, MLPClassifier, Pipeline, and Logistic Regression are some of the candidate machine learning algorithms. They were used in a scenario involving multi-class classification on NF-UQ-NIDS and NF-BoT-IoT datasets.

Overall, the proposed study's objectives demonstrate a comprehensive and systematic approach to advancing the field of intrusion detection in the Internet of Things. By addressing these objectives, the study aims to enhance the understanding, performance, and scalability of IDS systems, ultimately contributing to the security and reliability of IoT networks.

Several data pre-processing techniques were used to increase IDS quality. The in-stance-based and feature-based techniques were specifically explored. Instance-based pre-processing is concerned with data cleansing and removal strategies. Feature-based pre-processing comprises feature transformation, normalization, and dimensionality reduction through correct feature selection. Feature transformation was applied to all of the categorical characteristics of the selected datasets. The results of the evaluation are summarized in Table 12. Along with dataset quality, machine learning plays an important function. By comparing these findings, it is clear that our RF model outperformed better than the RF model on both datasets [30, 31].

In conclusion, the study demonstrated the effectiveness of various ML algorithms in assessing IDS performance and modeling traffic flow in the IoT context. The results highlight the importance of selecting appropriate algorithms and employing effective data pre-processing techniques to enhance the accuracy and overall performance of IDS systems in IoT environments.

The Internet of Things is susceptible to attacks due to several reasons. First, IoT devices are typically left unattended, which makes it easy for an attacker to gain physical access to them. Additionally, the majority of data transmission is wireless, which makes it simpler to eavesdrop. Ultimately, the majority of IoT devices have a limited amount of storage and processing capacity, this implies that additional security software cannot be employed. While the NF-UQ-NIDS and NF-BoT-IoT datasets used in the study provided valuable insights into the performance of machine learning algorithms for intrusion detection systems (IDS) in the context of the Internet of Things (IoT), there are several limitations to consider i.e. the size of the datasets may affect the generalizability of the results.: The NF-UQ-NIDS and NF-BoT-IoT datasets may not fully represent the diverse range of IoT network traffic patterns and intrusion instances. The datasets might be biased towards specific types of attacks or IoT device behaviors, limiting the generalizability of the findings to different IoT environments.

Class imbalance in the datasets, where certain classes have significantly more or fewer instances than others, can affect the performance of the machine learning algorithms. Imbalanced data can lead to biased models that prioritize the majority class and perform poorly on the minority classes, which are often the ones of interest in intrusion detection. The quality and reliability of the labeled data in the datasets can impact the performance of the machine learning algorithms. Inaccurate or mislabeled instances can introduce noise and affect the training process, leading to suboptimal results. The study mentioned feature-based pre-processing techniques, including feature transformation and dimensionality reduction. However, the specific features selected or engineered for the analysis were not discussed. The choice of features can greatly influence the performance of the algorithms, and the study did not provide insights into the selection process or the relevance of the chosen features.

The study primarily focused on accuracy, precision, recall, and F1 score as evaluation metrics. While these metrics provide important information about the performance of the algorithms, they may not capture all aspects of IDS performance. Other metrics, such as false positive rate, false negative rate, or area under the ROC curve, could provide a more comprehensive assessment of the algorithms' effectiveness. The performance of machine learning algorithms can vary across different datasets and network environments. The results obtained from the NF-UQ-NIDS and NF-BoT-IoT datasets may not necessarily generalize to other datasets or real-world IoT scenarios.

Considering these limitations, further research and evaluation on larger, more diverse datasets, along with the incorporation of additional evaluation metrics, would be beneficial to gain a more comprehensive understanding of the performance of IDS in the IoT context.

E. Limitation of Research

The Internet of Things is susceptible to attacks due to several reasons. First, IoT devices are typically left unattended, which makes it easy for an attacker to gain physical access to them. Additionally, the majority of data transmission is wireless, which makes it simpler to eavesdrop. Ultimately, the majority of IoT devices have a limited amount of storage and processing capacity, which implies that additional security software cannot be employed.

While the NF-UQ-NIDS and NF-BoT-IoT datasets used in the study provided valuable insights into the performance of ML algorithms for IDS in the context of the IoT, there are several limitations to consider:

1) Dataset size: The size of the data set may affect the generalizability of the results. If the data set is small, the algorithm's performance may not accurately reflect its capabilities in real-world scenarios. Additionally, small data sets may increase the risk of overfitting. An algorithm may perform well on training data, but may not generalize to new, unknown data

2) Dataset representativeness: The NF-UQ-NIDS and NF-BoT-IoT datasets may not fully represent the diverse range of IoT network traffic patterns and intrusion instances. The datasets might be biased towards specific types of attacks or IoT device behaviors, limiting the generalizability of the findings to different IoT environments.

3) Data imbalance: Class imbalance in a dataset, where certain classes have significantly more or fewer instances than other classes, can affect the performance of machine learning algorithms. Imbalanced data can lead to erroneous models that focus on the majority of classes and perform poorly on the minority of classes that are typically of interest in intrusion detection.

4) Data quality: The quality and reliability of the labeled data in the datasets can impact the performance of the ML algorithms. Inaccurate or mislabeled instances can introduce noise and affect the training process, leading to suboptimal results.

5) Feature selection: The study mentioned feature-based pre-processing techniques, including feature transformation and dimensionality reduction. However, the specific features selected or engineered for the analysis were not discussed. The selection of features can have an important effect on the performance of the algorithms, the study did not provide information about the process of selecting features or the importance of the features chosen.

6) Evaluation metrics: As evaluation metrics, the study largely used accuracy, precision, recall, and the F1 score. While these metrics provide important information about the performance of the algorithms, they may not capture all aspects of IDS performance. Other metrics, such as false positive rate, false negative rate, or area under the ROC curve, could provide a more comprehensive assessment of the algorithms' effectiveness.

7) Generalizability: The performance of machine learning algorithms can vary across different datasets and network environments. The results obtained from the NF-UQ-NIDS and NF-BoT-IoT datasets may not necessarily generalize to other datasets or real-world IoT scenarios.

Considering these limitations, further research and evaluation on larger, more diverse datasets, along with the incorporation of additional evaluation metrics, would be beneficial to gain a more comprehensive understanding of the performance of IDS in the IoT context.

VI. Conclusions and Future Work

The study explores the use of deep learning approaches for cyber security intrusion detection in IoT networks. It highlights the need for advanced techniques to improve the accuracy and efficiency of intrusion detection in IoT networks. Traditional intrusion detection systems may not be suitable for handling the unique characteristics and complexities of IoT networks. The study recognized the necessity of more intricate approaches, such as deep learning, to augment the fidelity and efficiency of intrusion detection in the IoT. Insufficient datasets for training and evaluation are also a concern. For IoT networks, where data characteristics differ from conventional networks, there is a shortage of comprehensive and representative datasets that capture the specific challenges and attack patterns in IoT environments. The study also addresses botnet attacks in IoT networks, focusing on analyzing bot data collected from IoT networks to develop effective intrusion detection systems. By addressing these problems, the study aims to contribute to the development of more robust and accurate intrusion detection mechanisms for IoT networks. It emphasizes the utilization of deep learning approaches and the availability of suitable datasets, particularly for botnet-related attacks, to enhance the security and resilience of IoT systems. The findings in this research provide insights into the effectiveness and accuracy of machine learning evaluation metrics in detecting a wide spectrum of cyberattacks. The accuracy metric represents the percentage of correctly classified instances by the algorithms on both datasets. Higher accuracy values generally indicate better performance, although it's important to consider other evaluation metrics such as precision, recall, and F1 score for a more comprehensive assessment of the algorithms' effectiveness.

There may be several areas of future research that can improve the security of the IoT ecosystem. A promising future research direction is to explore the use of deep learning techniques to build more powerful and intelligent IDS for the IoT. By using deep learning algorithms such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Transformers, and block complex penetration attempts more accurately and effectively. Furthermore, using federated learning methods for collaborative and privacy-preserving model training on IoT devices can improve intrusion detection capabilities by combining scalability and diversity. To improve the detection of intrusions and complex network threats, integrated approaches and hybrid architectures can be explored. Furthermore, the development of adversarial defense methods is crucial to protect deep learning-based IDS against new attack vectors.

Acknowledgment

The authors would like to thank the Deanship of Scientific Research at Majmaah University for supporting this work under Project Number No-PGR-2024-1022.

References

[1] Bakhsh, Shahid Allah, et al.; Enhancing IoT network security through deep learning-powered Intrusion Detection System. Internet of Things, 2023, 24. Jg., S. 100936.

[2] ALKADI, Sarah; AL-AHMADI, Saad; BEN ISMAIL, Mohamed Maher.; Toward Improved Machine Learning-Based Intrusion Detection for Internet of Things Traffic. Computers, 2023, 12. Jg., Nr. 8, S. 148.

[3] SOLIMAN, Sahar; OUDAH, Wed; ALJUHANI, Ahamed.; Deep learning-based intrusion detection approach for securing industrial Internet of Things. Alexandria Engineering Journal, 2023, 81. Jg., S. 371-383.

[4] ALKHUDAYDI, Omar Azib; KRICHEN, Moez; ALGHAMDI, Ans D.; A deep learning methodology for predicting cybersecurity attacks on the internet of things. Information, 2023, 14. Jg., Nr. 10, S. 550.

[5] FERRAG, Mohamed Amine, et al.; Rdtids: Rules and decision tree-based intrusion detection system for internet-of-things networks. Future Internet, 2020, 12. Jg., Nr. 3, S. 44.

[6] DINA, Ayesha S.; SIDDIQUE, A. B.; MANIVANNAN, D.; A deep learning approach for intrusion detection in Internet of Things using focal loss function. Internet of Things, 2023, 22. Jg., S. 100699.

[7] ALOSAIMI, Shema; ALMUTAIRI, Saad M.; An Intrusion Detection System Using BoT-IoT. Applied Sciences, 2023, 13. Jg., Nr. 9, S. 5427.

[8] SINGH, N. J., HOQE, N., SINGH, K. R., & BHATTACHARYA, D. K. (2024). Botnet - based IoT network traffic analysis using deep learning. Security and Privacy, 2024,7(2), e355.

[9] ASHARF, Javed, et al.; A review of intrusion detection systems using machine and deep learning in internet of things: Challenges, solutions and future directions. Electronics, 2020, 9. Jg., Nr. 7, S. 1177.

[10] SAHEED, Yakub Kayode, et al.; A novel hybrid autoencoder and modified particle swarm optimization feature selection for intrusion detection in the internet of things network. Frontiers in Computer Science, 2023, 5. Jg., S. 997159.

[11] YAZDINEJAD, Abbas, et al.; An ensemble deep learning model for cyber threat hunting in industrial internet of things. Digital Communications and Networks, 2023, 9. Jg., Nr. 1, S. 101-110.

[12] ALI, Saqib; LI, Qianmu; YOUSAFZAI, Abdullah.; Blockchain and federated learning-based intrusion detection approaches for edge-enabled industrial IoT networks: A survey. Ad Hoc Networks, 2024, 152. Jg., S. 103320.

[13] ZHAO, Ruijie, et al.; A novel intrusion detection method based on lightweight neural network for internet of things. IEEE Internet of Things Journal, 2021, 9. Jg., Nr. 12, S. 9960-9972.

[14] ABDULLAHI, Mujaheed, et al.; Detecting cybersecurity attacks in internet of things using artificial intelligence methods: A systematic literature review. Electronics, 2022, 11. Jg., Nr. 2, S. 198.

[15] HAJIHEIDARI, Somayye, et al.; Intrusion detection systems in the Internet of things: A comprehensive investigation. Computer Networks, 2019, 160. Jg., S. 165-191.

[16] INAYAT, Usman, et al.; Learning-based methods for cyber attacks detection in IoT systems: A survey on methods, analysis, and future prospects. Electronics, 2022, 11. Jg., Nr. 9, S. 1502.

[17] CHAABOUNI, Nadia, et al.; Network intrusion detection for IoT security based on learning techniques. IEEE Communications Surveys & Tutorials, 2019, 21. Jg., Nr. 3, S. 2671-2701.

[18] HULAYYIL, Sarah Bin; LI, Shancang; XU, Lida.; Machine-learning-based vulnerability detection and classification in Internet of Things device security. Electronics, 2023, 12. Jg., Nr. 18, S. 3927.

[19] FERRAG, Mohamed Amine, et al.; Federated deep learning for cyber security in the internet of things: Concepts, applications, and experimental analysis. IEEE Access, 2021, 9. Jg., S. 138509-138542.

[20] KIM, Jiyeon, et al.; Intelligent detection of iot botnets using machine learning and deep learning. Applied Sciences, 2020, 10. Jg., Nr. 19, S. 7009.

[21] ZEESHAN, Muhammad, et al.; Protocol-based deep intrusion detection for dos and ddos attacks using unsw-nb15 and bot-iot data-sets. IEEE Access, 2021, 10. Jg., S. 2269-2283.

[22] KHRAISAT, Ansam, et al.; A novel ensemble of hybrid intrusion detection system for detecting internet of things attacks. Electronics, 2019, 8. Jg., Nr. 11, S. 1210.

[23] SHAFIQ, Muhammad, et al.; CorrAUC: A malicious bot-IoT traffic detection method in IoT network using machine-learning techniques. IEEE Internet of Things Journal, 2020, 8. Jg., Nr. 5, S. 3242-3254.

[24] XU, Bayi, et al. ;IoT Intrusion Detection System Based on Machine Learning. Electronics, 2023, 12. Jg., Nr. 20, S. 4289.

[25] Inuwa, Muhammad Muhammad, and Resul Das. “A Comparative Analysis of Various Machine Learning Methods for Anomaly Detection in Cyber Attacks on IOT Networks.” Internet of Things 26 (July 2024): 101162.

[26] Yaras, Sami, and Murat Dener. “IoT-Based Intrusion Detection System Using New Hybrid Deep Learning Algorithm.” Electronics, vol. 1053, no. 6, 12 Mar. 2024.

[27] Yesi Novaria Kunang, Siti Nurmaini, Deris Stiawan, and Bhakti Yudho Suprapto. “An End-to-end Intrusion Detection System With IoT Dataset Using Deep Learning With Unsupervised Feature Extraction.” International Journal of Information Security, 23 Jan. 2024.

[28] Haedam Kim, Suhyun Park, Hyemin Hong, Jieun Park, and Seongmin Kim. “A Transferable Deep Learning Framework for Improving the Accuracy of Internet of Things Intrusion Detection.” Future Internet, vol. 80, no. 3, 28 Feb. 2024.

[29] Amit Kumar Mishra, Shweta Paliwal, and Gautam Srivastava. “Anomaly Detection Using Deep Convolutional Generative Adversarial Networks in the Internet of Things.” ISA Transactions, vol. 493–504, 1 Feb. 2024.

[30] Sarhan, M., Layeghy, S., Moustafa, N., Portmann, M.;NetFlow Datasets for Machine Learning-Based Network Intrusion Detection Systems. In: Deze, Z., Huang, H., Hou, R., Rho, S., Chilamkurti, N. (eds) Big Data Technologies and Applications. BDTA WiCON 2020 2020. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 371. Springer, Cham. 2021,https://doi.org/10.1007/978-3-030-72802-1_9

[31] Thirimanne, Sharuka Promodya, et al.; Deep neural network based real-time intrusion detection system. SN Computer Science, 2022, 3. Jg., Nr. 2, S. 145.

[32] NF-UQ-NIDS-v2 - UQ eSpace. https://espace.library.uq.edu.au/view/UQ:631a24a (accessed on 3 Jan 2024).

[33] NF-BoT-IoT (kaggle.com). https://www.kaggle.com/datasets/dhoogla/nfbotiot (accessed on 3 Jan 2024).