International Journal of Mathematical, Engineering and Management Sciences
Vol. 9, No. 1, 188-204, 2024
https://doi.org/10.33889/IJMEMS.2024.9.1.010

Ram Arti Publishers

A Novel Data Preprocessing Model for Lightweight Sensory IoT Intrusion Detection

Shahbaz Ahmad Khanday
Department of Computer Science & Engineering,
Sharda University, Knowledge Park III, Greater Noida, 201306, Uttar Pradesh, India.
E-mail: shahbazshaban10@gmail.com

Hoor Fatima
Department of Computer Science & Engineering,
Sharda University, Knowledge Park III, Greater Noida, 201306, Uttar Pradesh, India.
Corresponding author: hoor.iitd@gmail.com

Nitin Rakesh
Department of Computer Science and Engineering,
Symbiosis Institute of Technology, Nagpur Campus,
Symbiosis International (Deemed University), Pune, India.
E-mail: nitin.rakesh@gmail.com

(Received on August 15, 2023; Revised on November 2, 2023 & December 15, 2023; Accepted on December 17, 2023)

Abstract
IoT devices or sensor nodes are essential components of the machine learning (ML) application workflow because they gather abundant information for building models with sensors. Uncontrollable factors may impact this process and add inaccuracies to the data, raising the cost of computational resources for data preparation. Choosing the best method for this data pre-processing stage can lessen the complexity of ML models and wasteful bandwidth use for cloud processing. Devices in the IoT ecosystem with limited resources provide an easy target for attackers, who can make use of these devices to create botnets and spread malware. To repel attacks directed towards IoT, robust and lightweight intrusion detection systems are the need of an hour. Furthermore, data preprocessing remains the first step for modish machine learning models, ensemble techniques, and hybrid methods in developing anti-intrusion applications for lightweight IoT. This article proposes a novel data preprocessing model as a core structure using an Extra Tree classifier for feature selection and two classifiers LSTM and 1D-CNN for classification. The dataset used in this research is CIC IoT 2023 with 34 attack classes and SMOTE (Synthetic Memory Oversampling Technique) has been used for class balancing. The article evaluates the performance of 1D-CNN and LSTM on the CIC IoT 23 dataset using classification metrics. The proposed ensemble approach using LSTM has obtained 92% accuracy and with 1D-CNN the model obtained 99.87% accuracy.

Keywords- ML, CIC IoT 23, LSTM, 1D-CNN, SMOTE.

1. Introduction

The Internet of Things (IoT) is an infrastructure that will allow traditional networks and collaborative items to work together smoothly. The core concept behind the IoT is the connection of everything to the internet, formerly known as IOE. To collect information from tangible items, particular kinds of gadgets with sensors are employed. These Internet of Things (IoT) devices typically include an MCU unit, sensors, powered by batteries, and other wireless communication media, that may be set up to gather data in both confined and external settings. Because of their flexible evolution, Internet of Things devices can be put in challenging environments. Currently, data is being uploaded to the cloud from over 22 billion IoT devices. This figure grows exponentially yearly to keep gathering data from a wide range of sensors. The purpose of these data is to train machine learning (ML) models, which are potent instruments capable of uncovering hidden information inside data and logs that characterize decision-making. After the data has been analyzed, it is

saved locally and forwarded to a storage facility in the cloud, wherein the data is used to choose the best course of action. In IoT contexts (Rosero-Montalvo et al., 2022), the data-collecting process must manage uncontrollable circumstances such as ambient changes, MCU, and sensor manufacturing flaws that result in inaccurate measurements and disturbances in the workplace (Kalantar-zadeh, 2013). As a result, whether explaining an effect or analyzing human behavior, sophisticated models are often used to improve effectiveness. Preprocessing generates data that is dependable, precise, reproducible, and free of errors for model development (Dasgupta & Dey, 2013). Data from actual situations are larger, ambiguous, and noisier. They also need more information and are more chaotic. When the results of mining or modeling are obtained, several aspects lead to a decline in the integrity of the data. As a result, data must undergo refinement procedures before being mined or modeled. Contrarily, notwithstanding significant advantages, several issues must be resolved to ensure efficient and trustworthy functioning, such as seamless integration, protection, expectations, and server-client systems (Jane & Arockiam, 2021). The emergence of new technologies to enhance living standards might also change what the frameworks need to do (Shafique et al., 2020). For instance, the Internet of Vehicles (IoV) could have stricter reaction time requirements than typical IoT programs (Velarde-Alvarado et al., 2022).

Additional safeguarding challenges have been brought about by the widespread use and concomitant evolution of IoT technologies. The complex configurations and vagueness of new, often ad hoc settings make it difficult to comply with IoT security regulations. Unattended use is usual for the majority of Internet of Things (IoT) devices, which normally communicate wirelessly. Such circumstances frequently result in an intruder gaining easy physical or logical control over IoT devices or networks. An allegedly malevolent attacker might potentially inflict a significant, even lethal, effect upon a target by secretly employing countless IoT devices as bots (Albulayhi et al., 2021). In addition, constraint computing resources (which include software and embedded design choices including risky upgrade strategies, antiquated hardware, and outdated security rules) and a general paucity of setup concerns on the side of consumers are prevalent characteristics of the Internet of Things (Azimjonov & Kim, 2023). Regular malware deployments progressively emphasize compromising IoT networks as their primary objective because of the many and severe weaknesses in the IoT metasystem. Moreover, the rapid proliferation of unprotected Internet of Things devices and the ease with which assailants can locate them through web-based services have resulted in an ever-expanding pool of exploitable possibilities (Bovenzi et al., 2020). Today's cybercriminals use an immense amount of these susceptible IoT devices to launch enormous attacks targeting Web servers, such as scam URLs, emails that are spam, and Distributed Denial of Service (DDoS) strikes. The main goal of current malware operations is now IoT network infiltration due to numerous, severe protection vulnerabilities in IoT devices. Determining the security flaws brought about by malware developments and propagation in IoT and concentrating on efficient antimalware responses are crucial (McDermott et al., 2018). We provide a few instances of practical malware and botnet assaults using inconspicuous DDoS attacks to draw attention to safety and security issues within IoT (Ngo et al., 2020). Miria, BrickerBot, Torri, Hajime, and Bashlite are some of the real-world botnet examples (Vasan et al., 2020). Nonetheless, a great deal of work has gone into preparing IoT datasets for malware and DDoS attacks, and academics have suggested defensive strategies that make use of deep learning and modern machine learning techniques (Kolias et al., 2017; Costin et al., 2018; Su et al., 2018). To protect against IoT attacks, the researchers suggested implementing a significant number of intrusion detection systems. Table 1 of the manuscript contains the acronyms used throughout the manuscript

However, many machine learning-based intrusion detection models can benefit from extensible data preprocessing procedures to overcome noise and extract meaningful insights from large data chunks. Furthermore, several reasons make identifying and stopping attacks against IoT devices difficult (Khanday et al., 2021). For instance, scattered connections and lightweight devices having constrained resources and

without security features could make identifying and mitigating threats more difficult. (*A Survey on IoT Profiling, Fingerprinting, and Identification | ACM Transactions on Internet of Things*, n.d. (Erfani et al., 2021))

However, because of their widespread deployment, the current IoT device architecture leaves them susceptible to both physical as well as cyberattacks. In particular, harmful security concerns include user privacy leaks brought on by cyberattacks. To solve the security risks associated with the Internet of Things, it is imperative to build strong and flexible intrusion detection approaches and methods.

1.1 Motivation and Contribution

The design and development of defensive ways to protect IoT networks and sensory nodes for recognizing incursions continue to be among the critical goals for developers to improve their level of achievement in identifying harmful actions against the IoT ecosystem. Security measures for the IoT must recognize unwanted and illicit communication for surveillance and limit unsuitable data exchanges in the IoT network. Numerous academics have devised an assortment of intrusion detection strategies using machine learning and deep learning techniques to prevent illegitimate traffic from circulating within the structure of networks. Nevertheless, a few ML models frequently misidentify the most damaging traffic flows owing to the erroneous and impoverished selection of feature vectors from datasets. The main issue is that additional research must be carried out concerning the best attributes to pick to detect counterfeit traffic in these networks accurately. In that light, the article's primary objective is to create an IoT intrusion detection model, starting with a novel data pre-processing method to extract the best and most crucial features from the input data. The data pre-processing model was motivated by recent research publications (*CICIoT2023: A Real-Time Dataset and Benchmark for Large-Scale Attacks in IoT Environment[v1] | Preprints.Org*, n.d.), which showed that tree-based models had significantly improved efficiency and accuracy for the CIC IoT 2023 dataset. Secondly influenced by Khanday et al. (2023a) where binary classification is done using the BOT-IoT dataset. This study is noteworthy because it employs a customized pre-processing approach to tackle the unique troubles of IoT networks, focuses on an evolving cybersecurity demand associated with IoT gadgets and DDoS assaults, boosts the resilience of critical network infrastructure across multiple manufacturing sectors by successfully identifying DDoS incidents, and creates a lightweight intrusion detection system that is compatible alongside the constrained computational capabilities of IoT ecosystem. The article presents a novel technique for data pre-processing alongside deep learning models to create a portable and lightweight intrusion detection system. This article proposes a novel data preprocessing model as a core structure for developing a lightweight sensory IoT intrusion detection system. The datasets used in this research are CIC IoT 2023 and the classification metrics to test the performance of various classifiers. Some of the significant contributions of this research article are:

- A novel data pre-processing technique that uses tree-based feature importance (Extra Tree Classifier) to extract the best features based on a feature importance test. Henceforth to avoid noisy data features with the most minor importance are excluded.
- The study proposes hybrid models to test the performance of various deep-learning classifiers. Extra Tree Classifier with LSTM (ETLSTM) and Extra Tree Classifier with 1D-CNN (ETCNN) have been used for multiclass classification.
- Performance analysis of LSTM-CNN, and 1D-CNN classifiers for CIC IoT 23 dataset for multiple classification.

Table 1. The complete form of abbreviations used throughout the manuscript.

Acronym	Full Form
IoT	Internet of Things
DDoS	Distributed Denial of Services
IDS	Intrusion Detection System
LSTM	Long Short-Term Memory
CNN	Convolutional Neural Network
ELSTM	Extra Tree Classifier with Long Short-Term Memory
E1D-CNN	Extra Tree Classifier with 1D- Convolutional Neural Network
CIC IoT 23	Canadian Institute of Cybersecurity IoT 2023 dataset
MCU	Microcontroller Unit
IOE	Internet of Everything
NIDS	Network Intrusion Detection System
BLSTM	Bidirectional Long Short-Term Memory
SMOTE	Synthetic Minority Oversampling Technique`

2. Literature Review

Describing the various intrusions that can occur throughout an IoT context is crucial because intrusion detection systems are primarily focused on spotting assaults. To establish a cutting-edge intrusion detection system, numerous studies have been developed. The process is still ongoing to develop an effective method to recognize various types of intrusions towards IoT (Smys et al., 2020). Another requirement for the researchers to develop intrusion countermeasures remains to be realistic datasets with the modern class of attack vectors. We are reviewing some of the recent research articles in Table 2 of the manuscript to figure out some of the most successful defenses for being put to trial for malware transmission via covert DDoS assault in IoT networks. Recent efforts to implement IoT security solutions are also included. The literature review segment discusses the key features of the IoT and the dangers involving the incorrect use of bots and IoT sensory devices.

Table 2. Literature review of research articles.

Citation of the research article	Proposed IDS methodology	The Machine Learning Model used	Experimental Dataset
Susilo & Sari (2020)	DoS attack mitigation in IoT using deep learning methods.	Random Forrest, multilayer Perceptron, and Convolutional Neural Network.	BOT-IoT dataset by the University of New South Wales, Sydney.
Butt et al. (2022)	Lightweight real-time intrusion detection for intelligent home networks.	LSTM, AdaBoost, KNN, and Decision Tree.	CIC-IoT 2022 and UNSW-NB15.
Nguyen & Le (2023)	A novel method blends a soft-ordering convolutional neural network (SOCNN) structure paired with local outlier factor (LOF) and isolation-based anomaly detection utilizing nearest-neighbor ensembles (iNNE) approaches using the two types of learning techniques for DoS and DDoS attack detection.	Soft-ordering convolutional neural network (SOCNN) structure paired with local outlier factor (LOF) and isolation-based anomaly detection utilizing nearest-neighbor ensembles (iNNE).	CIC-IDS-2018, CIC-IDS-2017, and BOT-IoT dataset by the University of New South Wales, Sydney.
Vitorino et al. (2022)	The research article proposes an Adaptive perturbation pattern method. A2PM uses patterning cycles uniquely tailored to every group's unique features to provide accurate and cogent dataset perturbations.	Adaptive perturbation pattern method, formerly A2PM.	IoT datasets, namely the IoT 23 dataset and CIC-IDS2017.
Bowen et al. (2023)	In his study, BLoCNet, a combination deep learning (DL) paradigm consisting of bidirectional long short-term memory (BLSTM) and convolutional neural network (CNN) layers, is proposed.	BLSTM and CNN.	UNSW-NB15, IoT-23 and CIC-IDS2017.
Ahmad & Aziz (2019)	Particle Swarm Optimization and Correlation for feature selection.	Naïve Bayes, KNN and SVM classifiers are used.	KDD Cup99 dataset.

Table 2 continued…

Velarde-Alvarado et al. (2022)	The model created a fresh dataset with bot traffic to test the suggested approach using machine learning algorithms appropriate for data imbalances. Macro-average F1-Score and Mathews Correlation metrics are calculated.	Weighted Logistic Regression, Logistic regression with SMOTE, SVM alongside Sub-Sampling technique, XGBoost, and Weighted Decision Tree.	The UAN-12 dataset is prepared and used in various experiments.
Khanday et al. (2023b)	The authors propose an approach for data preparation modeling on the IoT-23 dataset and then use a set of classifiers for DDoS and other attack detection contained within IoT 23 dataset.	Gaussian Naïve Bayes, Linear SVC, and Ada Boost classifiers for binary and multiple classification.	IoT 23 dataset.
Qiu et al. (2021)	The authors suggested the model's extraction process to duplicate the representation for Adversarial Examples development. This method produces highly effective results with minimal data (just 10% of the initial training batch).	Saliency mapping technique for feature selection and Adversarial Examples generation.	The authors have upgraded the Kitsume NIDS.
Nimbalkar & Kshirsagar (2021)	To prevent DoS and DDoS assaults, this study suggests a technique for selecting features for an intrusion detection system (IDS) that uses Information Gain (IG) and Gain Ratio (GR) to extract the top 50% of features.	JRipclassifier.	BoT-IoT and KDD Cup99 datasets were used.
Shone et al. (2018)	The authors propose a Non-symmetric Deep Autoencoder model on KDD-CUP 99 and NSL-KDD datasets.	Non-symmetric Deep Autoencoder.	KDD-CUP ’99 and NSL-KDD datasets.
Vinayakumar et al. (2019)	This research explores the use of a deep neural network (DNN), a kind of deep learning algorithm, to develop an adaptable and efficient intrusion detection system (IDS) for identifying and categorizing unanticipated and unforeseen cyberattacks.	Deep Neural Network	CICIDS 2017, Kyoto, NSL-KDD, UNSW-NB15, and WSN-DS datasets are used in the manuscript.
Javaid et al. (2016)	The authors propose a deep learning model that uses Self Taught Learning approach to build a network intrusion detection system.	Sparse Auto-encoder and Soft-Max Regression for feature engineering.	NSL-KDD dataset
Yin et al. (2017)	The authors suggest employing RNN-IDS as a deep learning method for identifying intrusions. Additionally, the authors examine how well the model performs in binary and multiclass classification, as well as how the recommended model's efficiency is impacted by the value of neurons at each layer and different learning rates.	Recurrent Neural Network	KDD CUP ’99 dataset
Roy et al. (2017)	The possible use of Deep Neural Network as a tool for the classification of various incursion assault methods is examined in this research.	SVM and Deep Neural Network	-
Wang (2018)	The authors proposed a state-of-the-art adversarial method in comparison with various deep learning models.	MLP	NSL-KDD, JSMA generated dataset and FGSM generated dataset
Ramzan et al. (2023)	The present research uses algorithms based on deep learning, such as Gradient Recurrent Unit (GRU), Long Short-Term Memory (LSTM), and Recurrent Neural Network (RNN), to identify DDoS assaults on the most recently released dataset, CICDDoS2019. A comparison of these models is carried out utilizing the CICIDS 2017 dataset.	Gradient Recurrent Unit (GRU), Long Short-Term Memory (LSTM), and Recurrent Neural Network (RNN)	CICIDS2017 and CICDDoS2019
Zouhri et al. (2023)	The study compares the various feature selection techniques and uses L classifiers to test the performance of the best feature selection technique.	SVM, XG-Boost, and Random Forrest classifiers are used.	CIC-IDS 2018, CIC-IDS-2018 And TON-IoT
Wang et al. (2023)	Needs for research in ML resilience and potential future directions for NIDS are identified and discussed in this work. In conclusion, emphasizes the importance of developing and maintaining sturdiness over a machine learning (ML) NIDS's life span.	-	-

Table 2 continued…

Srivastava et al. (2023)	To lower the erroneous rate of detection of intrusions data classification, this research established an effective soft computing framework based on Grey Wolf Optimisation and Entropy-Based Graph (GWO-EBG).	Grey Wolf Optimization and Entropy-Based Graphs	KDD CUP'99
Wang et al. (2023)	We provide in this article a knowledge-distillation-based IoT intrusion detection framework called BT-TPF that can identify network threats that IoT gadgets might encounter in an IoT context with constrained computational power.		TON_IoT and CIC-IDS2017
Alrayes et al. (2023)	Using the ability of deep learning networks to acquire data arguments, the study has developed an approach in this study called a deep neural decision forest (DNDF), which enhances tree classification.	Deep Neural Decision Forrest	CIC2017 and CIC2018 datasets are used
Rodríguez et al. (2023)	To instruct the IDS strategy and produce understandable outcomes, the authors recommend a classification model called TabNet-IDS that takes advantage of attentiveness procedures to automatically choose important features from a given dataset.	TabNet model	CIC-IDS2017, CSE-CIC-IDS2018. and CICDDoS2019.
Zhang et al. (2023)	This study presents FS-DL, a feature selection with a deep learning-based data-driven network intrusion detection system. To enhance data quality, FS-DL employs techniques such as mining association rules and standard deviation to eliminate a significant number of superfluous features, lighten the processing burden, and increase detection precision.	-	USNW NB-15 and NSL-KDD
Wang et al. (2023)	To understand the actions and outcomes of assaults through the various kinds of data created in the varied IoT settings, this study suggests a transformer-based IoT network intrusion detection system technique.	Feature tokenizer-based transformer.	TON-IoT dataset by the University of New South Wales
Wu et al. (2022)	The autoencoder-based robust transformer method is used by the authors to build an IDS	Autoencoder	CICIoT2017 and CICIoT18 datasets are used.
Bakhsh et al. (2023)	The authors have used three different deep learning models for the development of a robust intrusion detection system. LSTM, Feed Forward Neural Network, and a Random Neural Network with flexible dynamics are used.	LSTM, FFNN, and RandNN	CIC IoT 2022
Li et al. (2020)	The article uses the state-of-the-art transformer and Random Forrest for developing Auto Encoder-based IDS for IoT meta-systems.	Auto Encoder and Random Forrest	CSE CIC-IDS-2018 dataset
Lopes et al. (2022)	To acquire the minimum dimensional depiction of features and compact the attack dataset, the unsupervised pre-learning strategy is applied using a denoising autoencoder. Next, a DNN classification algorithm that employs a multiclass supervised approach is trained using a subset of the compacted dataset.	Deep Neural Network and a denoising Auto-encoder	CICIDS2018
Thakkar & Lohiya (2023)	Featuring a deep neural network (DNN) serving as the foundational estimator, the bagging procedure classifier an ensemble learning technique is how the article attempts to tackle the issue of inequalities in target classes.	Deep Neural Network	CIC_IDS-2017, BOT-IoT, USNW-2017 and NSL-KDD

3. Proposed Methodology for IoT Intrusion Detection

We tested the proposed intrusion detection model for the most recent IoT dataset CIC IoT 2023. According to the author, only one research article has used various machine and deep learning classifiers to test the

CIC IoT 2023 dataset. The results obtained by tree-based classifiers (Random Forrest and Decision Tree classifiers) in the CIC IoT 2023 dataset inspired the preprocessing phase of the proposed methodology. (*CICIoT2023: A Real-Time Dataset and Benchmark for Large-Scale Attacks in IoT Environment[v1] | Preprints.Org*, n.d.) The various building blocks of the proposed preprocessing and, ultimately, intrusion detection model is depicted in Figure 1 of the manuscript.

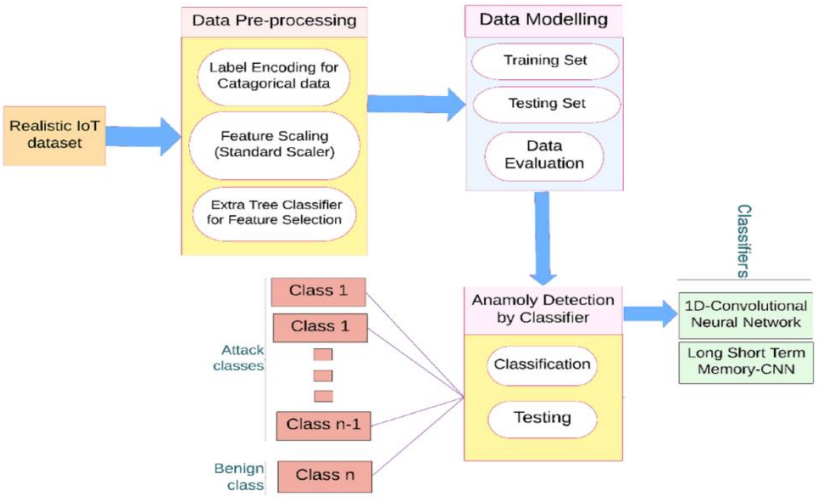

Figure 1. Intrusion detection model.

3.1 Realistic IoT Dataset

This part of the proposed model represents the CIC IoT 23 dataset, taken as input by the pre-processing unit. In this section, we have used the visualization and exploration plots of various IoT attack classes of the CIC IoT 23 dataset in Figure 2 of the manuscript.

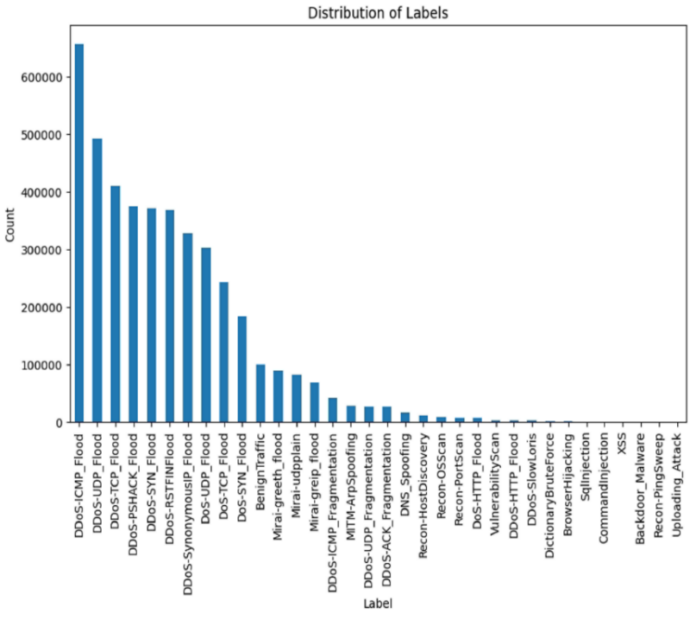

Figure 2. Attack class distribution of the CIC IoT 23 dataset.

There are 34 attack classes in the CIC IoT 23 dataset ranging from 656966 as the highest value to 131 as the least, and the names of attack classes are given on the horizontal axis while the total count of instances is on the vertical axis of Figure 2.

3.2 Data Pre-processing

The data preprocessing section of the proposed intrusion detection model starts with the investigation and visualization of dataset features. The categorical features of the CIC IoT 23 dataset are converted using label encoding. In the CIC IoT 23 dataset, there are 47 features, out of which 46 are float type, and only the label (target feature) is categorical and label encoded. After label encoding, data in each feature of the dataset are normalized using Standard Scaler before passing the datasets to the Extra Tree classifier to calculate the importance of each feature in the dataset. The usefulness of impurity-based variables can be assessed using tree-based cost estimators. This is then able to be utilized to remove needless features.

We put forward an innovative ensemble feature evaluation method called the Extra Tree classifier with the capability of feature importance and chose a total of twenty significant features spanning the CIC IoT 23 dataset. The dataset's leftover features are removed. An Extra Tree Classifier for feature importance is an approach based on models for selecting parameters that utilize an evaluation of a tree-based model to determine the significance of each feature. This estimation tool accommodates different data sources and smaller samples from an extensive variety of randomized decision trees. Each tree is also given a random assortment of numerous variables taken from the source data, from which it must select the feature that will divide the data into segments employing the Gini Index (Baby et al., 2021). Many de-correlated decision trees are produced by this random feature selection. The numerical criteria (Gini Index if the Gini Index is utilized in the forest design) that were implemented to determine which features ought to be split are scaled to give every feature an aggregate decline. The outcomes of the test are known as the Gini importance of features. After grading all features in terms of their Gini Importance in the lowest to the highest order, the user decides on the top number of features. (*Sklearn.Ensemble.ExtraTreesClassifier — Scikit-Learn 1.3.0 Documentation*, n.d.)
The determining element in this situation is going to be Information Gain. Start by calculating the Entropy of the data.

$$\text{Entropy(S)} = \sum_{i=1} c - \text{pilog2(pi)} \quad (1)$$

c - Represents the total number of unique target labels in the target variable
pi - Represents the Proportion of rows in a data frame
i- output label
Gain at the initial tree in the forest is calculated using:

$$Gain\ (S, A) = \sum v\varepsilon Values(A)\ |\text{Sv}| / |\text{S}|\ \text{Entropy (Sv)} \quad (2)$$

A - Individual feature in the data frame.
Every decision tree computes the gain of every feature within the data frame. For example
Gain (G1) = (Entropy (S) , A1) represents the gain of the feature named A1 at Decision tree D1.
Gain (G2) = (Entropy (S) , A2) represents the gain of the feature named A2 at Decision tree D2.

Gain (Gn) = (Entropy (S) , An) represents the gain of the feature named An at the Decision tree Dn.

Consequently, When the procedure is complete, each feature's total gain is calculated, and those features exhibiting the highest Gain are designated as important features. The importance of each feature in CIC IoT 23 is given in the Figure 3 plot of the article. In comparison, we are selecting the twenty most important features from the dataset.

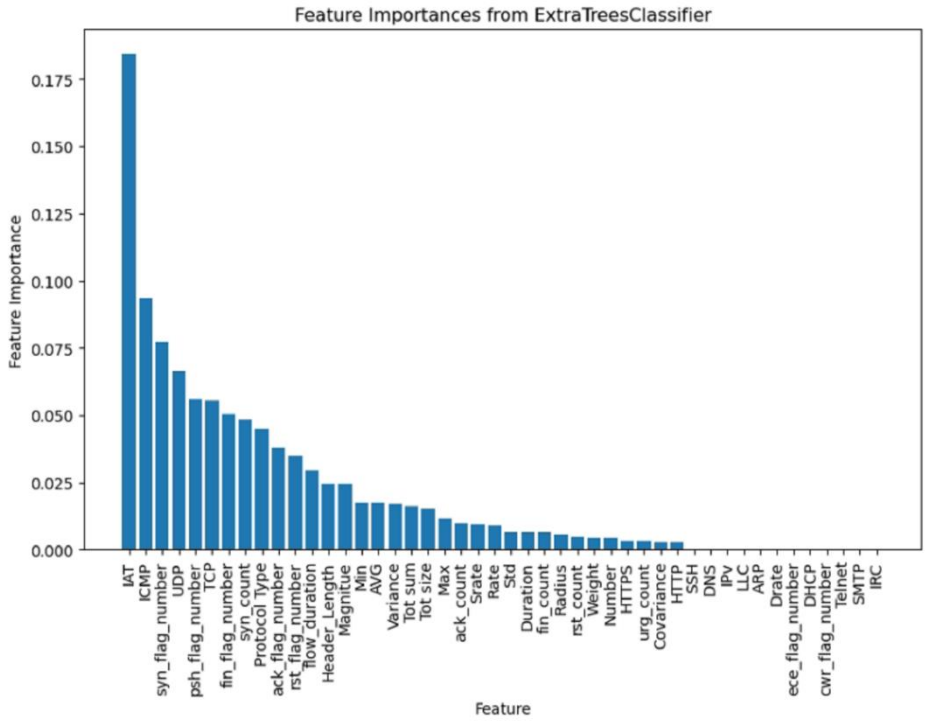

Figure 3. Feature Importance of CIC IoT 23 data frame.

The horizontal axis of the Figure 3 represents the name of every feature of the data frame and the vertical axis represents the feature importance ranging from the value 0.000 to 0.175. The twenty best features of the CIC IoT 23 dataset with feature importance ranging from 0.012 to 0.175 are "fl*ow_duration, Header_Length, Protocol Type, fin_flag_number, syn_flag_number, rst_flag_number, psh_flag_number, ack_flag_number, syn_count, TCP, ICMP, Tot sum, Min, Max, AVG, Tot size, IAT, Magnitue, variance and label.* We have selected the twenty best features from the dataset using tree-based feature importance for feature selection.

3.3 Data Modelling

We divided the data frames into training and testing sets, with training size making 80% of the data frame and testing set remaining 20% percent of the original data. The classes are balanced using the SMOTE (Synthetic Minority oversampling Technique) class balancing technique and the datasets are split into 10-fold validation sets for assessment.

3.4 Anomaly Detection by the Classifier

This section defines the performance of the classifier by various classification matrices. We use the Accuracy metric, Macro Average, Weighted Average of Precision, Recall, F1-Score, and Support of a classification test. The accuracy and loss plots of training and testing sets are also discussed in Section 4 of the article.

Accuracy = TP + TN/ TP + TN + FP + FN x 100
Precision = TP / TP + FN x 100
Recall = TP / TP + FP x 100
Support = 2 x Precision x Recall / Precision + Recall.

4. Results and Discussions

The results section is divided into two case studies for two different datasets. Case 1 and Case 2 contain the classification report, accuracy plots, and loss plots of both data frames.

Case 1: LSTM used on the CIC IoT 23 dataset.

Case 2: 1D-CNN used on CIC IoT 23 dataset.

4.1 Case 1: Long Short-Term Memory used on CIC IoT 23 Dataset

The LSTM approach is an enhanced edition of the RNN model that avoids the explosion of gradients or back-propagation gradient dispersion, making it a better choice for handling sequence data exhibiting long-term dependencies. The input gate, forget gate, and output gate are added to every neuron of the LSTM (Wang & Lu, 2020). Using this method, LSTM may resolve the gradient disappearing and gradient explosion issues during extended series training. Each forget gate takes in the outcome and input values, splicing them together, and then preserves knowledge under its control. The major internal processes of the LSTM algorithm are sigmoid, tanh, addition, and multiplication. We have initialized the LSTM unit with 128 dense units, optimizer as Adam, and categorical cross entropy for multiple attack classes within the datasets (*Neural Network Models (Supervised)*, n.d.). The classification report of LSTM is given in Table 3 below.

Table 3. Classification report LSTM.

Case	Accuracy	Precision		Recall		F1-Score		Support	
		Macro Average	Weighted Average	Macro Average	Weighted Average	Macro Average	Weighted Average	Macro Average	Weighted Average
LSTM for CIC IoT 23 dataset	92%	0.91	0.92	0.90	0.92	0.90	0.91	661013	661013

The LSTM model has achieved an overall 92% accuracy in the CIC IoT 23 data frame. The classification of testing and validation sets are represented in accuracy and loss plots of the LSTM model in Figure 4 and Figure 5 of the article.

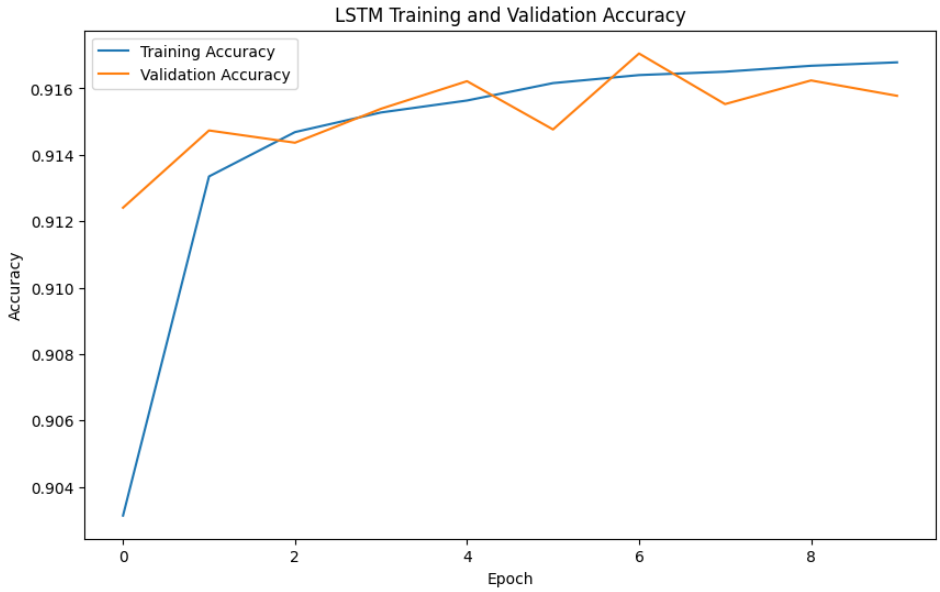

Figure 4. Training and Validation accuracy plots obtained using LSTM on the CIC IoT 23 dataset.

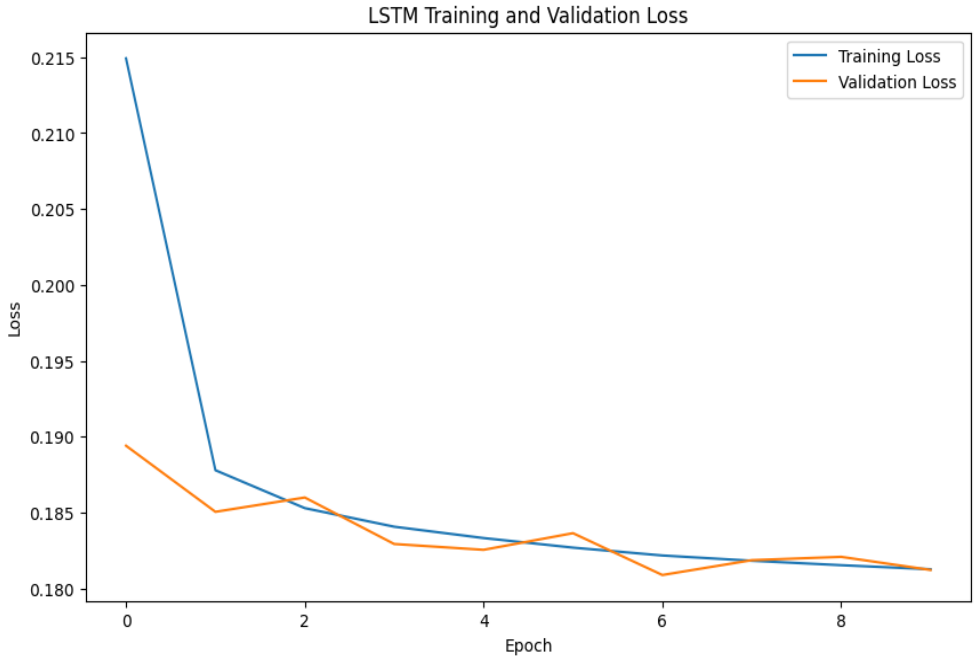

Figure 5. Training and Validation loss plots obtained using LSTM on the CIC IoT 23 dataset.

4.2 Case 2: 1D-Convolutional Neural Network

The advantage of employing CNNs for series classification is that they can be trained instantly using unprocessed time series data, negating the need for explicitly designing input characteristics and allowing for the use of the technology without specialized knowledge. The model should operate as well as models fitted on a variant of the information set with engineering amenities (Tang et al., 2022). This is possible because it can internalize the time series data and develop a representation. We are proposing a one-dimensional Convolutional Neural Network to eliminate the bias and loss by improving (Azizjon et al., 2020) classification metrics. For multiclass classification, the proposed 1D-CNN model is tuned with multiple dense layers (256, 128, and 64 dense units). The model is also adjusted with kernel size = 5, batch normalization, one dropout layer (dropout units = 0.5), and flatten layers. In addition to that, we have initialized the 1D-CNN model max pooling with a pooling size equal to 2 at each layer. The model's performance is evaluated with the help of a classification report in Table 4.

Table 4. Classification report of 1D-CNN.

Case	Accuracy	Precision		Recall		F1-Score		Support	
		Macro Average	Weighted Average	Macro Average	Weighted Average	Macro Average	Weighted Average	Macro Average	Weighted Average
1D-CNN for CIC IoT 23 dataset	99.87%	1.0	1.0	0.99	1.0	1.0	0.99	661013	661013

The 1D-CNN model has achieved an overall 99.87% accuracy in the CIC IoT 23 data frame. The classification of testing and validation sets are represented in accuracy and loss plots of the 1D-CNN model in Figure 6, and Figure 7 of the article.

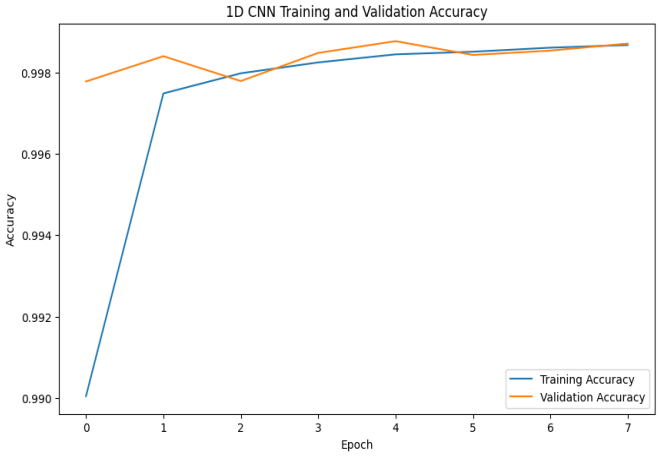

Figure 6. Training and Validation accuracy plots obtained using 1D-CNN on the CIC IoT 23 dataset.

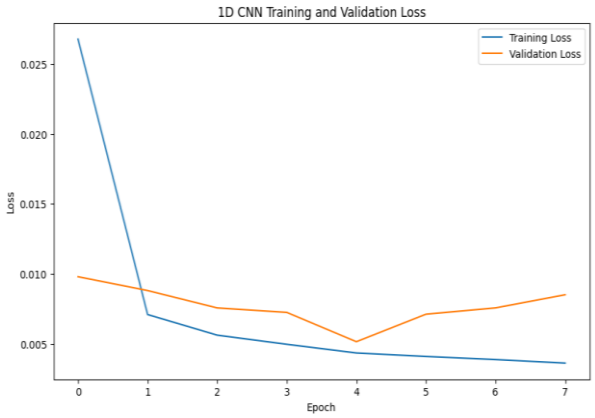

Figure 7. Training and Validation loss plots obtained using 1D-CNN on the CIC IoT 23 dataset.

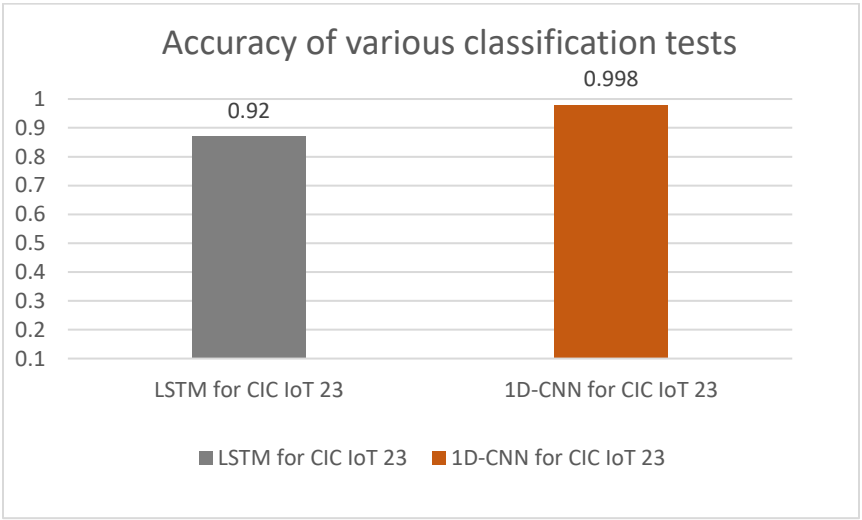

Figure 8. Accuracy graph.

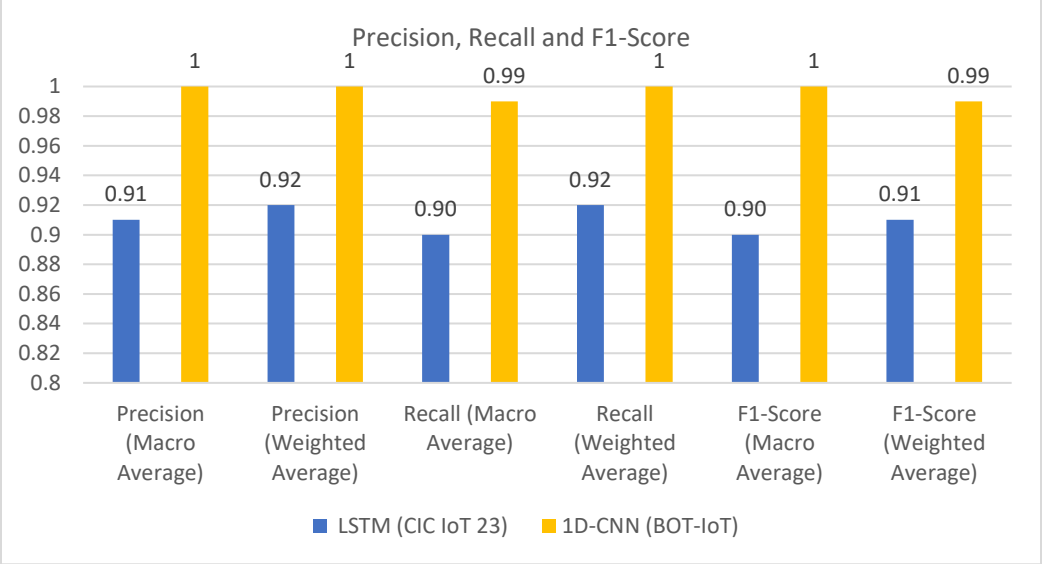

Figure 9. Precision, Recall, and F1-Score with macro and weighted average using LSTM and 1D-CNN.

4.3 Discussions

In this section, we are comparing various manuscripts with our proposed model. Comparatively, only one research article's preprint is available to this date, which has developed a testbed for the CIC IoT 23 dataset and made it available for research. The authors of that article have used Random Forrest, Ada Boost, MLP, Logistic Regression, and Deep ANN as classifiers for multiclass (34 Classes and 8 Classes) and binary classification. In our experiments, a class-balanced CIC IoT 23 dataset is used, and a novel intrusion detection model uses LSTM and 1D-CNN as classifiers. Figure 8 of the manuscript depicts the accuracy percentage achieved by two classifiers while Figure 9 contains the plots of macro and weighted average of precision, Recall and F1-Score. The deep learning models the authors of the published article used are Multilayer Perceptron and deep neural network model on the CIC IoT 23 dataset, while the proposed model uses two different fine-tuned deep learning architectures. In our proposed methodology, we obtained 99.87% accuracy higher than what was obtained by the previous researchers using Logistic Regression, Perceptron, DNN, random Forrest, and Ada Boost classifiers. Also, we have increased the model's precision from 0.70 to 1.0 in total. As far as robustness and limitations are concerned for the proposed model, selecting the most important features from the input data frame, and eliminating any unnecessary features maintains the model's resilience. The suggested models choose the maximum number of important features throughout the feature engineering portion of the model, as most machine learning-based IDS models overlook certain crucial features from the dataset. Further, it has been researched that choosing more than 45% of the features may impact the model's performance.

5. Conclusion and Future Scope

IoT is now becoming increasingly important to the community to advance the contemporary notion of a smart, safe, and sustainable community. In this regard, creating safety measures is essential to allowing effective, safe, and reliable IoT services. This study aimed to encourage the implementation of defense analytics tools for realistic IoT deployments by taking advantage of a fresh and substantial IoT attack dataset. The CICIoT2023 dataset expands on continuing Internet of Things security perspectives in comparison with the up-to-date state-of-the-art publications by employing a large structure alongside an

assortment of IoT nodes, carrying out numerous assaults that have never been seen in a single IoT stability dataset, and examining how commonly employed machine learning (ML) techniques perform in classifying benign samples over other thirty-three different variety of IoT assaults. So far, the dataset developers have only used a set of five classifiers. As per the author's, knowledge no hybrid or ensemble approach has been used on the CICIoT2023 dataset. Finally, the article explores the hybrid intrusion detection models ETLSTM and ET1D-CNN for IoT datasets leveraging tree-based feature importance for feature selection. The intrusion detection model presented in this work combines a novel method of data pre-processing alongside deep learning models in a compact form. Multiple deep learning method types are provided by the study, which makes them excellent for constructing inexpensive detection mechanisms that protect IoT devices from DDoS and other propagating malware attacks. The proposed state-of-the-art hybrid approach has achieved a higher success rate as compared to the other research articles, which used the CIC IoT 2023 dataset. The proposed ETCNN (Extra Tree classifier with 1D-CNN) model has achieved 99.87% accuracy with a precision of 0.99 and a recall of 1.0. Contrarily, the proposed IoT intrusion detection model has worked efficiently for two fresh and newly developed IoT datasets. We intend to generalize the proposed model on other IoT datasets. In addition, the proposed model can be optimized by bringing other machine-learning approaches into practice in the future.

Conflict of Interest

The authors confirm that there is no conflict of interest to declare for this publication.

Acknowledgments

This research did not receive any specific grants from funding agencies in the public, commercial, or not-for-profit sectors.

References

Neural network models (supervised). (n.d.). Scikit-Learn. Retrieved 9 April 2023, from https://scikit-learn/stable/modules/neural_networks_supervised.html

Ahmad, T., & Aziz, M.N. (2019). Data preprocessing and feature selection for machine learning intrusion detection systems (02). ICIC International, *13*(2), 93-101. https://doi.org/10.24507/icicel.13.02.93.

Albulayhi, K., Smadi, A.A., Sheldon, F.T., & Abercrombie, R.K. (2021). IoT intrusion detection taxonomy, reference architecture, and analyses. *Sensors, 21*(19), 6432. https://doi.org/10.3390/s21196432.

Alrayes, F.S., Zakariah, M., Driss, M., & Boulila, W. (2023). Deep neural decision forest (DNDF): A novel approach for enhancing intrusion detection systems in network traffic analysis. *Sensors, 23*(20), 8362. https://doi.org/10.3390/s23208362.

Azimjonov, J., & Kim, T. (2023). Stochastic gradient descent classifier-based lightweight intrusion detection systems using the most efficient feature subsets of datasets. *SSRN Scholarly Paper 4378339*. https://doi.org/10.2139/ssrn.4378339.

Azizjon, M., Jumabek, A., & Kim, W. (2020). 1D CNN based network intrusion detection with normalization on imbalanced data. In *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)* (pp. 218-224). Fukuoka, Japan. https://doi.org/10.1109/icaiic48513.2020.9064976.

Baby, D., Devaraj, S.J., Hemanth, J., & M., Anishin, R.M. (2021). Leukocyte classification based on feature selection using extra trees classifier: Atransfer learning approach. *Turkish Journal of Electrical Engineering and Computer Sciences, 29*(8), 2742-2757. https://doi.org/10.3906/elk-2104-183.

Bakhsh, S.A., Khan, M.A., Ahmed, F., Alshehri, M.S., Ali, H., & Ahmad, J. (2023). Enhancing IoT network security through deep learning-powered intrusion detection system. *Internet of Things, 24*, 100936.

Bovenzi, G., Aceto, G., Ciuonzo, D., Persico, V., & Pescapé, A. (2020). A Hierarchical hybrid intrusion detection approach in IoT scenarios. In *2020 GLOBECOM 2020 - 2020 Global Communications Conference* (pp. 1-7). IEEE. Taipei, Taiwan. https://doi.org/10.1109/globecom42002.2020.9348167.

Bowen, B., Chennamaneni, A., Goulart, A., & Lin, D. (2023). BLoCNet: A hybrid, dataset-independent intrusion detection system using deep learning. *International Journal of Information Security, 22*(4), 893-917. https://doi.org/10.1007/s10207-023-00663-5.

Butt, N., Shahid, A., Qureshi, K.N., Haider, S., Ibrahim, A.O., Binzagr, F., & Arshad, N. (2022). Intelligent deep learning for anomaly-based intrusion detection in IoT smart home networks. *Mathematics, 10*(23), 4598. https://doi.org/10.3390/math10234598.

Costin, A., Zaddach, J., & Antipolis, S. (2018). IoT malware: comprehensive survey, analysis framework and case studies. *1*(1), 1-9.

Dasgupta, R., & Dey, S. (2013). A comprehensive sensor taxonomy and semantic knowledge representation: Energy meter use case. In *2013 Seventh International Conference on Sensing Technology* (pp. 791-799). Wellington, New Zealand. https://doi.org/10.1109/icsenst.2013.6727761.

Erfani, M., Shoeleh, F., Dadkhah, S., Kaur, B., Xiong, P., Iqbal, S., Ray, S., & Ghorbani, A.A. (2021). A feature exploration approach for IoT attack type classification. In *2021 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)* (pp. 582-588). AB, Canada. https://doi.org/10.1109/DASC-PICom-CBDCom-CyberSciTech52372.2021.00101.

Jane, V.A., & Arockiam, L. (2021). Survey on IoT data preprocessing. *Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12*(9), 238-244. https://turcomat.org/index.php/turkbilmat/article/view/3001.

Javaid, A., Niyaz, Q., Sun, W., & Alam, M. (2016). A deep learning approach for network intrusion detection system.*EAI Endorsed Transactions on Security and Safety, 16*(9), e2. https://doi.org/10.4108/eai.3-12-2015.2262516.

Kalantar-zadeh, K. (2013). Sensors characteristics. In: Kalantar-zadeh, K. (ed.) *Sensors: An Introductory Course.* Springer US, pp. 11-28. https://doi.org/10.1007/978-1-4614-5052-8_2.

Khanday, S.A., Fatima, H., & Rakesh, N. (2023a). Implementation of intrusion detection model for DDoS attacks in lightweight IoT networks. *Expert Systems with Applications*, *215*, 119330. https://doi.org/10.1016/j.eswa.2022.11933.

Khanday, S. A., Fatima, H., & Rakesh, N. (2023b). Towards the Development of an Ensemble Intrusion Detection Model for DDoS and Botnet Mitigation using the IoT-23 Dataset. Harbin Gongcheng Daxue Xuebao/Journal of Harbin Engineering University, *44*(5), Article 5. https://harbinengineeringjournal.com/index.php/journal/article/view/255.

Khanday, S.A., Fatima, H., & Rakesh, N. (2021). Deep learning offering resilience from trending cyber-attacks, a review. In *2021 International Conference on Computational Performance Evaluation* (pp. 741-749). Shillong, India, https://doi.org/10.1109/ComPE53109.2021.9752099.

Kolias, C., Kambourakis, G., Stavrou, A., & Voas, J. (2017). DDoS in the IoT: Mirai and other botnets. *Computer, 50*(7), 80-84. https://doi.org/10.1109/mc.2017.201.

Li, X., Chen, W., Zhang, Q., & Wu, L. (2020). Building auto-encoder intrusion detection system based on random forest feature selection. *Computers & Security, 95*, 101851. https://doi.org/10.1016/j.cose.2020.101851.

Lopes, I.O., Zou, D., Abdulqadder, I.H., Ruambo, F.A., Yuan, B., & Jin, H. (2022). Effective network intrusion detection via representation learning: A denoising autoencoder approach. *Computer Communications, 194*, 55-65. https://doi.org/10.1016/j.comcom.2022.07.027.

McDermott, C.D., Majdani, F., & Petrovski, A.V. (2018). Botnet detection in the internet of things using deep learning approaches. In *2018 International Joint Conference on Neural Networks* (pp. 1-8). Rio de Janeiro, Brazil.

Neto, E.C.P., Dadkhah, S., Ferreira, R., Zohourian, A., Lu, R., & Ghorbani, A.A. (2023). CICIoT2023: A real-time dataset and benchmark for large-scale attacks in IoT environment. *Sensors 2023*, *23*, 5941. https://www.preprints.org/manuscript/202305.0443/v1.

Ngo, Q.-D., Nguyen, H.-T., Le, V.-H., & Nguyen, D.-H. (2020). A survey of IoT malware and detection methods based on static features. *ICT Express, 6*(4), 280-286. https://doi.org/10.1016/j.icte.2020.04.005.

Nguyen, X.-H., & Le, K.-H. (2023). Robust detection of unknown DoS/DDoS attacks in IoT networks using a hybrid learning model. *Internet of Things, 23,* 100851. https://doi.org/10.1016/j.iot.2023.100851.

Nimbalkar, P., & Kshirsagar, D. (2021). Feature selection for intrusion detection system in Internet-of-Things (IoT). *ICT Express, 7*(2), 177-181. https://doi.org/10.1016/j.icte.2021.04.012.

Qiu, H., Dong, T., Zhang, T., Lu, J., Memmi, G., & Qiu, M. (2021). Adversarial attacks against network intrusion detection in IoT systems. *IEEE Internet of Things Journal, 8*(13), 10327-10335. https://doi.org/10.1109/jiot.2020.3048038.

Ramzan, M., Shoaib, M., Altaf, A., Arshad, S., Iqbal, F., Castilla, Á.K., & Ashraf, I. (2023). Distributed denial of service attack detection in network traffic using deep learning algorithm. *Sensors, 23*(20), 8642. https://doi.org/10.3390/s23208642.

Rodríguez, D.Z., Okey, O.D., Maidin, S.S., Udo, E.U., & Kleinschmidt, J.H. (2023). Attentive transformer deep learning algorithm for intrusion detection on IoT systems using automatic Xplainable feature selection. *PLOS ONE, 18*(10), e0286652. https://doi.org/10.1371/journal.pone.0286652.

Rosero-Montalvo, P.D., López-Batista, V.F., & Peluffo-Ordóñez, D.H. (2022). A new data-preprocessing-related taxonomy of sensors for IoT applications. *Information, 13*(5), 241. https://doi.org/10.3390/info13050241.

Roy, S.S., Mallik, A., Gulati, R., Obaidat, M.S., & Krishna, P.V. (2017). A deep learning based artificial neural network approach for intrusion detection. In: Giri, D., Mohapatra, R.N., Begehr, H., Obaidat, M.S. (eds.) *Mathematics and Computing* (Vol. 655, pp. 44-53), Springer, Singapore. https://doi.org/10.1007/978-981-10-4642-1_5.

Shafique, K., Khawaja, B.A., Sabir, F., Qazi, S., & Mustaqim, M. (2020). Internet of things (IoT) for next-generation smart systems: A review of current challenges, future trends and prospects for emerging 5G-IoT scenarios. *IEEE Access, 8,* 23022-23040. https://doi.org/10.1109/access.2020.2970118.

Shone, N., Ngoc, T.N., Phai, V.D., & Shi, Q. (2018). A deep learning approach to network intrusion detection. *IEEE Transactions on Emerging Topics in Computational Intelligence, 2*(1), 41-50. https://doi.org/10.1109/tetci.2017.2772792.

sklearn.ensemble.ExtraTreesClassifier—Scikit-learn 1.3.0 documentation. (n.d.). Retrieved 9 August 2023, from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html

Smys, S., Basar, A., & Wang, H. (2020). Hybrid intrusion detection system for internet of things (IoT). *Journal of IoT in Social, Mobile, Analytics, and Cloud, 2*(4), 190-199. https://doi.org/10.36548/jismac.2020.4.002.

Srivastava, D., Singh, R., Chakraborty, C., Kumar, S., Makkar, A., & Sinwar, D. (2023). A framework for detection of cyber attacks by the classification of intrusion detection datasets. *Microprocessors and Microsystems*, 104964. https://doi.org/10.1016/j.micpro.2023.104964. (In press).

Su, J., Vasconcellos, D.V., Prasad, S., Sgandurra, D., Feng, Y., & Sakurai, K. (2018). Lightweight classification of IoT malware based on image recognition. In *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)* (Vol. 2, pp. 664-669). Tokyo, Japan. https://doi.org/10.1109/compsac.2018.10315.

Susilo, B., & Sari, R.F. (2020). Intrusion detection in IoT networks using deep learning algorithm. *Information, 11*(5), 279. https://doi.org/10.3390/info11050279.

Tang, W., Long, G., Liu, L., Zhou, T., Blumenstein, M., & Jiang, J. (2022). Omni-Scale CNNs: A simple and effective kernel size configuration for time series classification. *The Tenth International Conference on Learning Representations.* arXiv. https://doi.org/10.48550/arXiv.2002.10061.

Thakkar, A., & Lohiya, R. (2023). Attack classification of imbalanced intrusion data for IoT network using ensemble-learning-based deep neural network. *IEEE Internet of Things Journal, 10*(13), 11888-11895. https://doi.org/10.1109/jiot.2023.3244810.

Vasan, D., Alazab, M., Venkatraman, S., Akram, J., & Qin, Z. (2020). MTHAEL: Cross-architecture iot malware detection based on neural network advanced ensemble learning. *IEEE Transactions on Computers, 69*(11), 1654-1667. https://doi.org/10.1109/tc.2020.3015584.

Velarde-Alvarado, P., Gonzalez, H., Martínez-Peláez, R., Mena, L.J., Ochoa-Brust, A., Moreno-García, E., Félix, V.G., & Ostos, R. (2022). A novel framework for generating personalized network datasets for NIDS based on traffic aggregation. *Sensors, 22*(5), 1847. https://doi.org/10.3390/s22051847.

Vinayakumar, R., Alazab, M., Soman, K.P., Poornachandran, P., Al-Nemrat, A., & Venkatraman, S. (2019). Deep learning approach for intelligent intrusion detection system. *IEEE Access, 7*, 41525-41550. https://doi.org/10.1109/access.2019.2895334.

Vitorino, J., Oliveira, N., & Praça, I. (2022). Adaptative perturbation patterns: Realistic adversarial learning for robust intrusion detection. *Future Internet, 14*(4), 108. https://doi.org/10.3390/fi14040108.

Wang, M., Yang, N., & Weng, N. (2023). Securing a smart home with a transformer-based IoT intrusion detection system. *Electronics, 12*(9), 2100. https://doi.org/10.3390/electronics12092100.

Wang, M., Yang, N., Gunasinghe, D.H., & Weng, N. (2023). On the robustness of ML-based network intrusion detection systems: An adversarial and distribution shift perspective. *Computers, 12*(10), 209. https://doi.org/10.3390/computers12100209.

Wang, X., & Lu, X. (2020). A host-based anomaly detection framework using XGBoost and LSTM for IoT devices. *Wireless Communications and Mobile Computing, 2020*, e8838571. https://doi.org/10.1155/2020/8838571.

Wang, Z. (2018). Deep learning-based intrusion detection with adversaries. *IEEE Access, 6*, 38367-38384. https://doi.org/10.1109/access.2018.2854599.

Wang, Z., Li, J., Yang, S., Luo, X., Li, D., & Mahmoodi, S. (2024). A lightweight IoT intrusion detection model based on improved BERT-of-Theseus. *Expert Systems with Applications, 238*(F), 122045. https://doi.org/10.1016/j.eswa.2023.122045.

Wu, Z., Zhang, H., Wang, P., & Sun, Z. (2022). RTIDS: A Robust transformer-based approach for intrusion detection system. *IEEE Access, 10*, 64375-64387. https://doi.org/10.1109/access.2022.3182333.

Yin, C., Zhu, Y., Fei, J., & He, X. (2017). A deep learning approach for intrusion detection using recurrent neural networks. *IEEE Access, 5*, 21954-21961. https://doi.org/10.1109/access.2017.2762418.

Zhang, L., Liu, K., Xie, X., Bai, W., Wu, B., & Dong, P. (2023). A data-driven network intrusion detection system using feature selection and deep learning. *Journal of Information Security and Applications, 78*, 103606. https://doi.org/10.1016/j.jisa.2023.103606.

Zouhri, H., Idri, A., & Ratnani, A. (2023). Evaluating the impact of filter-based feature selection in intrusion detection systems. *International Journal of Information Security*. https://doi.org/10.1007/s10207-023-00767-y. (In press).