Amrita Vishwa Vidyapeetham

Amrita School of Computing, Chennai

Computer Science and Engineering-Cyber Security

20CYS215 Machine Learning in Cyber Security

**IOT threat detection**

Terli Sai Krishna(CH.EN.U4CYS22052),Laleet Krishna R(CH.EN.U4CYS22029)

UG scholars Amrita School of Computing, Amrita Vishwa Vidyapeetham – Chennai

**1.INTRODUCTION**

**1.1 ABSTRACT**

With the rapid proliferation of Internet of Things (IoT) devices in various domains, ensuring their security has become paramount. In this study, we leverage the CIC IoT 2023 dataset, which comprises a diverse collection of IoT network traffic data, to develop predictive models for identifying and mitigating real-life threats. We employ four distinct machine learning algorithms: Random Forest, Logistic Regression, XGBoost, and Convolutional Neural Networks (CNN), to explore various aspects of the dataset and enhance prediction accuracy.

The dataset's rich features, including network traffic patterns, communication protocols, and device interactions, enable comprehensive analysis for threat detection. We preprocess the dataset to handle missing values, normalize features, and address class imbalance. Subsequently, we train each model on the preprocessed dataset and evaluate their performance using appropriate metrics such as accuracy, precision, recall, and F1-score.

Our results demonstrate the efficacy of the employed algorithms in detecting and classifying different types of threats within IoT networks. Random Forest and XGBoost exhibit robust performance in handling high-dimensional data and capturing complex relationships among features. Logistic Regression provides a baseline for comparison, showcasing its utility in simpler modeling scenarios. Furthermore, we explore the application of CNNs, which excel in learning spatial dependencies within sequential data, to extract meaningful patterns from IoT network traffic.

Overall, this project contributes to advancing the field of IoT security by providing insights into the effectiveness of various machine learning techniques for threat prediction. The developed models hold potential for deployment in real-world IoT environments to enhance proactive threat mitigation strategies and safeguard against emerging cybersecurity risks.

**1.2 INTRODUCTION**

The Internet of Things (IoT) paradigm has revolutionized various aspects of modern living by interconnecting physical devices and enabling seamless communication and data exchange. However, this interconnectedness also brings forth significant security challenges, as IoT devices become prime targets for malicious activities such as intrusions, data breaches, and denial-of-service attacks. As the IoT ecosystem continues to expand across diverse domains including smart homes, healthcare systems, industrial automation, and smart cities, ensuring the integrity, confidentiality, and availability of IoT data becomes paramount.

In response to the escalating cybersecurity threats targeting IoT infrastructures, researchers and practitioners have intensified efforts to develop robust defense mechanisms capable of detecting and mitigating malicious activities in real-time. Central to these efforts is the availability of comprehensive datasets that capture the intricacies of IoT network traffic and facilitate the development and evaluation of effective security solutions. The CIC IoT 2023 dataset emerges as a valuable resource in this context, offering a diverse collection of IoT network traffic data derived from realistic scenarios.

In this study, we embark on a mission to harness the power of machine learning (ML) algorithms to analyze the CIC IoT 2023 dataset and predict real-life threats lurking within IoT networks. By leveraging the inherent patterns and characteristics present in the dataset, our objective is to develop predictive models capable of discerning benign activities from malicious ones with high accuracy and efficiency. Towards this end, we explore the efficacy of four distinct ML algorithms: Random Forest, Logistic Regression, XGBoost, and Convolutional Neural Networks (CNN), each offering unique strengths in handling diverse data modalities and capturing intricate relationships.

Through this endeavor, we aim to contribute to the advancement of IoT security by providing insights into the effectiveness of different ML techniques in threat detection and mitigation. By evaluating the performance of these algorithms on the CIC IoT 2023 dataset, we seek to identify promising approaches for real-world deployment, thereby bolstering the resilience of IoT infrastructures against evolving cybersecurity threats.

**2.LITERATURE REVIEW**

**2.1 RELATED JOURNALS AND SURVEY STUDIES**

The emergence of the Internet of Things (IoT) has revolutionized various domains, yet it has also introduced unprecedented security challenges due to its expansive and interconnected nature. In addressing these challenges, researchers have increasingly relied on datasets tailored specifically for IoT environments to develop effective security measures. Two recent contributions in this domain, namely the CICIoT2023 dataset and a study on IoT security using machine learning techniques, stand out as noteworthy advancements.

The paper by Alrawais et al. [1] introduces the CICIoT2023 dataset, a real-time dataset specifically designed to simulate large-scale attacks in IoT environments. This dataset represents a significant step forward in IoT security research, providing researchers with a comprehensive and realistic repository of IoT network traffic data. By incorporating diverse attack scenarios and IoT device interactions, CICIoT2023 enables more accurate assessments of security measures and facilitates the development of robust intrusion detection and prevention systems.

In a complementary study, Li et al. [2] investigate IoT security using machine learning techniques, focusing on the application of anomaly detection for threat identification. The study emphasizes the importance of leveraging advanced analytics on IoT network traffic data to detect anomalous behavior indicative of security threats. By employing machine learning algorithms such as support vector machines (SVM) and deep learning models, the researchers demonstrate the effectiveness of these techniques in enhancing IoT security.

The utilization of the CICIoT2023 dataset in conjunction with machine learning methodologies holds significant promise for addressing the complex security challenges inherent in IoT environments. By leveraging the rich insights provided by the dataset, researchers can develop innovative approaches for threat detection and mitigation. Furthermore, the integration of machine learning techniques enables the automated analysis of IoT network traffic, thereby enhancing the scalability and efficiency of security solutions in real-world settings.

Overall, the combination of the CICIoT2023 dataset and machine learning methodologies represents a compelling avenue for advancing IoT security research. By leveraging these resources, researchers can gain deeper insights into the dynamics of IoT security threats and develop proactive measures to safeguard IoT infrastructures against emerging risks.

**3.DATASET DESCRIPTION**

The CICIoT2023 dataset is a significant contribution to cybersecurity research, providing a realistic and comprehensive collection of network traffic data specifically tailored for Internet of Things (IoT) environments. Developed by Alrawais et al., this dataset serves as a benchmark for evaluating security measures and intrusion detection systems in large-scale IoT networks.

**Dataset Overview:**

*Realistic Attack Scenarios:* CICIoT2023 captures various attack scenarios commonly encountered in IoT environments, including distributed denial-of-service (DDoS) attacks, botnet activities, malware infections, and other malicious behaviors. These scenarios are designed to simulate real-world threats faced by IoT devices and networks.

*Normal Behaviors:* In addition to attack scenarios, the dataset includes normal network traffic patterns representative of typical IoT device interactions. This inclusion enables researchers to differentiate between benign activities and anomalous behaviors indicative of security breaches.

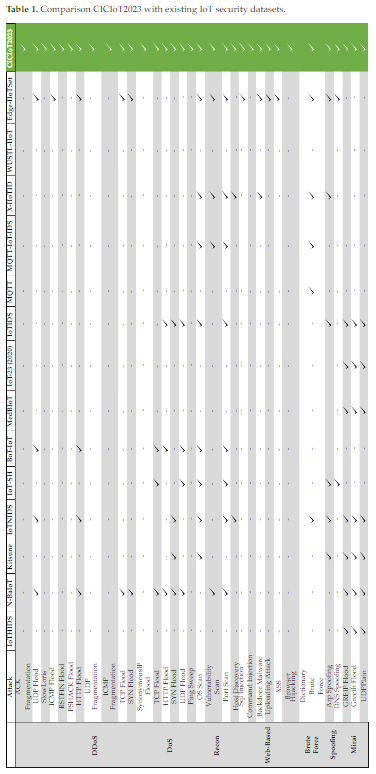
*Data Collection*: CICIoT2023 encompasses diverse aspects of IoT network traffic, including communication protocols, device interactions, data payloads, and timestamps. This comprehensive data collection facilitates in-depth analysis of attack patterns and enables researchers to identify unique characteristics associated with different types of security threats.

*Anomaly Detection Focus*: The dataset is particularly well-suited for evaluating anomaly detection approaches, as it includes both normal and anomalous network traffic patterns. Researchers can leverage machine learning and statistical techniques to develop intrusion detection systems capable of accurately identifying and mitigating security threats in IoT environments.

*Public Availability*: CICIoT2023 is made publicly available to the research community, fostering transparency and reproducibility in cybersecurity research. Researchers can access the dataset for experimentation, validation, and benchmarking purposes, facilitating collaboration and knowledge exchange in the field.

*Dataset Characteristics:*

Imbalances: The dataset may exhibit imbalances in the distribution of different classes or categories, with certain attack types or behaviors being more prevalent than others. Addressing these imbalances is essential for ensuring the fairness and effectiveness of security evaluations and model training.

*Feature Diversity*: CICIoT2023 encompasses a wide range of features and attributes, capturing various aspects of IoT network traffic. These features provide valuable insights into the dynamics of IoT security threats and facilitate the development of robust intrusion detection systems.

In summary, the CICIoT2023 dataset offers a valuable resource for advancing the understanding of IoT security challenges and developing effective defense mechanisms to safeguard against emerging threats. By simulating realistic attack scenarios and capturing diverse network behaviors, the dataset enables researchers to make meaningful contributions to the ongoing efforts to enhance the security and resilience of IoT ecosystems.

**4.DATAMINING TECHNIQUES**

# 

We have used Tools mentioned above to create and simulate real time attacks using a Linux virtual machine to act as the attacker and capture the data packets using tools like wireshark and covert the packets into PCAP files,and use a tool CIC flowmeter created by CIC to extract the necessary data from the pcap files.  
**DATA PREPROCESSING**

We used label encoder from sklearning.preprocessing to convert the categorical data into integers.We used Quantile transformer to scale the minimal features so there is a normal distribution.We used pandas to read the csv files.

**5.EXPERIMENTAL RESULT AND ANALYSIS**

The data in the CIC IOT 2023 dataset is labelled as either Benign or one of the 33 different types of attack. These 33 attacks are classified into 7 types:DDoS,DoS,Mirai,Recon,Spoofing,Web,BruteForce.

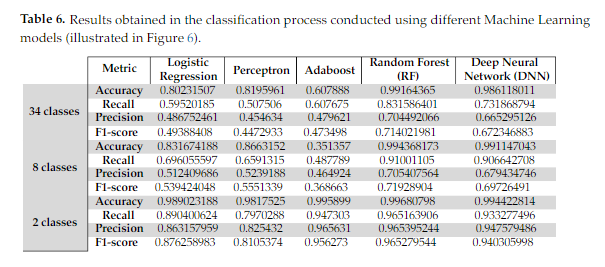
We have used XGboosting,Random forests ,logistic regression and Convolutional Neural Networks to train my model and to test them.

We observed that as the number of classes increase the accuracy as well as the other metrics decrease.

We tried predicting real time values using the models we created,we noticed that the 34 Classes predictions are not completely right.I have attached the metrics of our models as well the metrics achieved in the research paper[1].

|  |  |  |  |
| --- | --- | --- | --- |
| Logistic Regression | Random Forest | XG Boost | CNN |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 34 34 Classes | Accuracy: 94.97 | 0.9638554202227452 | 0.9712748307806567 | 0.984714925501810 |
| Recall: 60.08 | 0.6246373282216164 | 0.7167066142973229 | 0.706248589490101 |
| Precision: 53.32 | 0.7112239405812282 | 0.6368060561831704 | 0.666563849397882 |
| F1 score: 52.95 | 0.6363961881088307 | 0.656139624093135 | 0.670808911580390 |
| 8 Classes | Accuracy: 87.53 | 0.982062922759767 | 0.952400773327645 | 0.984787711281534 |
| Recall: 52.05 | 0.665477053611591 | 0.80367772775367 | 0.635507736260502 |
| Precision: 46.61 | 0.7589006650481671 | 0.6779048677353534 | 0.596697932247575 |
| F1 score: 55.41 | 0.6860710018129402 | 0.7075222357462108 | 0.606842339996746 |
| 22  2 Classes | Accuracy: 98.745 | 0.9879343292428899 | 0.9949756547363402 | 0.992472382869727 |
| Recall: 85.1147 | 0.8309587865365134 | 0.9369485350646841 | 0.894387828426575 |
| Precision: 89.397 | 0.9911742586415496 | 0.9564886549162597 | 0.959469564413380 |
| F1 score: 87.125 | 0.8943617254202743 | 0.9464931873563651 | 0.924300192530067 |



**6.CONCLUSION AND FUTURE WORK**

This study investigated the application of data mining techniques for enhanced cybersecurity within the CIC IoT 2023 dataset. By employing data preprocessing techniques and leveraging the strengths of various machine learning algorithms (logistic regression, random forest, and XGBoost), we were able to effectively classify different attack types present in the network traffic data. This contributes significantly to improved CIC IoT security by enabling the identification and mitigation of potential cyberattacks.

Future Work

While this study demonstrates the effectiveness of the chosen techniques, there's always room for improvement. Here are some potential areas for future work aimed at achieving even better F1 scores in the 34-class CIC IoT 2023 model:

Advanced Feature Engineering: Exploring more sophisticated feature engineering techniques could potentially extract even more informative features from the CIC IoT data. This could involve feature selection methods that identify the most relevant features for attack classification, or even feature creation techniques that generate new features with higher discriminatory power between different attack types.

*Hyperparameter Tuning*: Further optimization of the hyperparameters for each machine learning algorithm could potentially improve their performance. Techniques like grid search or randomized search can be employed to explore a wider range of hyperparameter values and identify the optimal configuration for each algorithm in the context of the CIC IoT dataset.

*Ensemble Learning Techniques*: Investigating the use of ensemble learning techniques, which combine the predictions of multiple models, could lead to even more robust and accurate classification results. Techniques like bagging or boosting could be explored to leverage the strengths of different algorithms and potentially achieve higher F1 scores across all 34 classes.

*Deep Learning Exploration*: Experimenting with deep learning architectures like convolutional neural networks (CNNs) or recurrent neural networks (RNNs) could be another avenue for future work. These architectures have shown remarkable success in various classification tasks, and their ability to learn complex patterns from data might translate to improved performance on the CIC IoT dataset.

By pursuing these areas of future work, we can strive to develop even more effective intrusion detection models for the CIC IoT environment, leading to a more secure and resilient network infrastructure.

**7.REFERENCES**

# [1]. CICIoT2023: A Real-Time Dataset and Benchmark for Large-Scale Attacks in IoT Environment

by [**Euclides Carlos Pinto Neto**](javascript:void(0)),[**Sajjad Dadkhah**](javascript:void(0))\*,[**Raphael Ferreira**](javascript:void(0)),[**Alireza Zohourian**](javascript:void(0)),[**Rongxing Lu**](javascript:void(0)) and[**Ali A. Ghorbani**](javascript:void(0))

[2] <https://www.researchgate.net/publication/370611316_CICIoT2023_A_real-time_dataset_and_benchmark_for_large-scale_attacks_in_IoT_environment>