CHAPTER I
INTRODUCTION

1.1 Background

The Internet of Things (IoT) technology has been extensively used in many wireless telecommunications applications [1], including smart transportation, smart health services, and smart cities, due to its significant advancements. IoT is continuously transforming our everyday lives and work practices. However, it also confronts significant security vulnerabilities [2]. Intrusion detection systems (IDS) are capable of detecting atypical or malevolent behavior in networks and hosts, making them a crucial security solution for ensuring the security of IoT [3].

Based on various monitoring data sources, Intrusion Detection Systems (IDS) may be categorized into two types [4]: network-based IDS (NIDS) and host-based IDS (HIDS). HIDS is often used on commercial to safeguard against unauthorized access to system records and sensitive data. NIDS is deployed on either hosts or switches and mainly monitors network traffic data, including sessions, flows, and packets. IDS may be categorized into two types based on their detection principles [5]: anomaly detection and misuse detection. Misuse detection involves the extraction of network traffic characteristics and their comparison with a pre-established characteristic database. If the matching is successful, network traffic is classified as an attack activity. Misuse detection has a low incidence of false alarms (FAR). Nevertheless, it lacks the capability to identify unfamiliar assaults and requires the upkeep of an extensive feature database. Anomaly detection involves the pre-learning of a standard behavior pattern, which is then used to identify any network traffic that deviates from this pattern as abnormal behavior. Anomaly detection has a slightly higher False Alarm Rate (FAR) compared to misuse detection. However, it has gained significant attention due to its capability to detect unexpected assaults and its tremendous adaptability.

Conventional machine learning (ML) methods, including, logistic regression (LR), decision tree (DT) [6], random forest (RF) [7], K-nearest neighbor (KNN) [8], AdaBoost [9], etc, have been extensively used for detecting abnormal network intrusions. IDS that use machine learning technology provide several benefits, including simple deployment and robust technical analysis. Boosting is a prevalent machine learning approach used to enhance the accuracy of classification and re-

gression models [10]. The algorithm operates by progressively constructing a sequence of weak models, whereby each model is trained on the leftover data from the preceding model, with an emphasis on rectifying notable misclassifications. Subsequently, the feeble models are amalgamated into a more robust model via the utilization of procedures that provide suitable weights [11]. Nevertheless, because of the surge in network traffic and the proliferation of attack methods, relying just on shallow machine learning technology is insufficient to fulfill the demands of large-scale NIDS.

Prior NIDS methods disregarded the significance of new data sets and the issue of class imbalance. Data sets are crucial in acquiring knowledge about the behavioral patterns of anomalies in order to identify them. The ongoing advancements in attack techniques need the use of more advanced network traffic in order to accurately assess the effectiveness of NIDS in contemporary network settings. The majority of network traffic consists of regular data, whereas anomalous data represents just a tiny fraction of the overall data. Typically, intrusion detection models that are trained directly from data with class imbalance issues tend to have low performance when it comes to recognizing certain attack classes. Hence, it is essential to address the issue of data class imbalance or enhance the intrusion detection model to enable IDS to effectively face the escalating network security threats.

This research proposes a machine learning-based technique employing boosting algorithms to address the issue of class imbalance in large-scale network flow data. The focus is on developing a double layer machine learning approach for network intrusion detection model specifically designed for IoT contexts. This research assessed the CICIoT2023 dataset, which is a more recent and extensive collection of network traffic data. This research also utilizes Principal Component Analysis (PCA) to decrease the number of features in the dataset.

1.2 Problem Identification and Objective

As explained previously in the background, although there have been various approaches using machine learning methods for network intrusion detection, previous studies often do not consider the significant class imbalance problem in network traffic data, which can result in poor detection performance against certain attack classes. In addition, as attack techniques evolve and network traffic volumes increase, the use of more representative and relevant datasets such as CICIoT2023 becomes critical to ensure that intrusion detection models remain effective in changing network environments. Traditional approaches that rely solely on shallow machine

learning are often insufficient to handle this complexity.

Therefore, this research attempts to address two major gaps: first, by addressing the problem of class imbalance in network traffic data, and second, by utilizing a more recent dataset that better matches current real-world conditions. By developing a machine learning technique based on boosting algorithms and a two-layer approach specifically designed for the IoT context, this study aims to improve the accuracy and resilience of intrusion detection against various types of attacks, including previously unknown attacks. The use of PCA is also carried out to reduce data complexity without losing important information, thereby increasing the efficiency and effectiveness of the detection model.

The proposed approach has the potential not only to improve the overall intrusion detection performance but also to provide a more adaptive and responsive solution to the growing IoT security threats. Thus, this research not only contributes to the improvement of intrusion detection technology but also provides a stronger foundation for the development of more intelligent and effective network security systems in the future.

In this research, double layer machine learning system with Categorical Boosting (CAB) and eXtreme Gradient Boosting (XGB) algorithm will be used to classify network traffic dataset. CAB and XGB are machine learning algorithms that work by classifying tabular data. CAB and XGB has higher accuracy in classify label, but require moderate to high computation to use it. There are some methods that has been applied to reduce computation for CAB and XGB. Based on the explanation above, there are some objectives of this research:

1. Offers a two-layer network intrusion detection model that combines the CAB and XGB methods for detecting intrusions in large-scale stream data. This model is anticipated to enhance the efficiency of detection by effectively using extensive data. This study will also provide a comparison of the assessment outcomes of different boosting algorithms used.
2. Assess and compare the performance of the CAB and XGB algorithms using the most recent IoT IDS dataset named CICIoT2023.
3. Utilizing Principal Component Analysis (PCA) to convert the data into a lower-dimensional space, with the aim of decreasing computational time.
4. Analyze the learning techniques for a double layer machine learning system utilizing the CAB and XGB algorithms, and evaluate their performance based on accuracy, F1-score, precision, and recall on simulation Google Colab and emulation on Raspberry Pi4.

1.3 Related Research

This section provides an overview and analysis of existing research on network intrusion detection in IoT contexts. Based on research [12], the researcher proposes a machine learning (ML)based Wifi Network Intrusion Detection System (WNIDS) for Wi-Fi networks to efficiently detect attacks. The proposed WNIDS consists of two stages that work together sequentially. The ML model was developed for each stage to classify the network records into normal or one of the specific attack classes. Researchers trained and validated ML models for WNIDS using the publicly available Aegean Wi-Fi Intrusion Dataset (AWID). Several feature selection techniques have been considered to identify the best feature set for WNIDS. The research proves that it can improve detection accuracy for certain attacks on the AWID dataset.

In research [13], the researcher explored the use of a stack architecture to combine several classifier ensembles, namely gradient boosting machine (GBM), random forest (RF), and extreme gradient boosting machine (XGB) to detect anomalies in Web application scenarios using the CICIDS2017 dataset. The results obtained indicate that several algorithms used such as random forest and XGB have a higher detection accuracy rate than other algorithms for certain attack classifications. Eman et al. [14] introduced the concept of a Federated Intrusion Detection BlockChain (FIDC) that utilizes lightweight Artificial Neural Networks (ANN) using Federated Learning (FL) to safeguard the privacy of healthcare data. This is achieved by leveraging the capabilities of blockchain technology, which offers a decentralized ledger. Researchers evaluated the performance of the Artificial Neural Network (ANN) and eXtreme Gradient Boosting (XGBoost) models using the BoT-IoT dataset. The results indicate that the Artificial Neural Network (ANN) model exhibits superior accuracy and performance while dealing with data heterogeneity on IoT devices. The researchers conducted tests on FIDChain using various datasets, including Bot-Net-IoT, CSE-CIC-IDS2018, and KDDCup99. The findings indicated that the BoT-IoT dataset yielded the most dependable and accurate results when evaluating IoT applications, particularly those utilized in Healthcare systems. The accuracy improvements achieved were 99,99% with ANN and 98,40% with XGBoost. Sandeep et al. [15] present a hierarchical ML-based hyperparameter-optimization approach for classifying network intrusions. The CICIDS-2017 standard dataset was used, and the CatBoost algorithm demonstrated 99.62% accuracy in classifying DDoS attacks, demonstrating its effectiveness in enhancing cybersecurity.

In research [16], researchers have used different feature selection algorithms

and two classification algorithms with the WEKA tool to detect intrusions using the CICIDS2017 dataset which consists of seven different types of attacks. According to the results, feature selection reduced the dataset size and time and provides the high performance. In research [17], researchers have used various classification algorithms and feature selection algorithms using Pearson correlation to detect brute force attacks on the network with the help of the WEKA tool on the CICIDS-2017 dataset and have shown the best performance in terms of accuracy, precision, Recall, F-measure and time to build the model. The simulation results show that the use of the various algorithm and feature selection using the Pearson correlation reduces dataset dimensions, time to build the model, false alarms and induces high performance results that obtain 99% accuracy.

Based on previous research, the reduction of dataset features and the selection of the correct boosting algorithm for certain types of attacks can affect the accuracy of the attack detection process. In addition, the CAB and XGB methods are expected to have higher accuracy results than other learning methods for the CICIoT2023 dataset.

1.4 Scope of Work

To limit the discussion on this topic, we describe several scopes of work for this research:

1. The data balancing methods used in this research are Random Under Sampler (RUS) and Synthetic Minority Oversampling Technique (SMOTE),
2. Principal Component Analysis (PCA) will be used as a dimension reduction technique on the dataset,
3. The machine learning method used in the first layer in this research is CAB,
4. The machine learning method used in the second layer in this research is XGB,
5. Comparing the CAB and XGB algorithms used with other boosting algorithms such as Gradient Boosting (GAB), Hist Gradient Boosting (HGB), Ada Boosting (ADB), and Light Gradient Boosting (LGB),
6. All parameters used in the algorithms are default parameters,
7. Evaluation of computing time.
8. Emulation is performed on a Raspberry Pi4 model B device.

1.5 Methodology

These are some methode to complete the research:

1. Study Literature
   This process is learning stage about the theories of the internet of things, boosting algorithms, and IDS from the newest sources such as papers, journals, and books.
2. Design System Model
   Design a system model based on the study literature. The model consists of dataset pre-processing, feature reduction dataset, CAB and XGB training process, and the output of the learning process.
3. Simulation and Analysis
   This process start from getting the dataset of CICIoT2023. After that the simulation will be doing in Colab Google with Python language and emulation also will be doing on Raspberry Pi4. This simulation process consists of data preprocessing, feature reduction in dataset, and double layer system with CAB and XGB training process. After that, the output of simulation will be analyzed.
4. Conclusion
   The analysis result is used to conclude the research problems and purposes that have been stated before.

1.6 Research Timeline

The research timeline can be seen on Table 1.1 and the output will be published on June to July.

Table 1.1 Research timeline.

Timeline / Activity	March 2024	April 2024	May 2024	June 2024	July 2024	August 2024	September 2024
Create general system model							
Build data preprocessing							
Build double layer system							
Train & test double layer system							
Create the paper/journal							
Publish paper							
Paper review and acceptance							

1.7 Structure of The Thesis

The rest of this thesis is organized as follows:

CHAPTER 1: INTRODUCTION
This chapter discusses the background of this research, from the problem in the field, related research, the scope of work, and the research methodology that is used in this research

CHAPTER 2: BASIC CONCEPTS
This chapter provides basic concepts used in this thesis. The explanation focuses on IDS, IoT, machine learning, CICIoT2023 Dataset, and the feature reduction method with PCA that will be applied to this research.

CHAPTER 3: SYSTEM MODEL AND RESEARCH DESIGN
This chapter describes the system model including parameters and variables used in the thesis, research flow, and how the simulation works in the algorithm.

CHAPTER 4: RESULT & ANALYSIS
This chapter discusses the thesis results starting from the simulation output

consisting of classification metrics, confusion matrices, and learning and testing process times.

CHAPTER 5: CONCLUSIONS AND FUTURE WORKS

This chapter provides the conclusion of this thesis and notifies the future works.