CHAPTER I
INTRODUCTION

1.1. Background

The Internet of Things (IoT) is a paradigm that connects various devices so they can communicate and exchange data over the internet [1]. In an industrial context, this concept has evolved into the Industrial Internet of Things (IIoT), which leverages communication and automation to make manufacturing processes more monitorable and efficient [2]. However, increased connectivity also expands the attack surface and heightens security risks in network-dependent systems. As a result, security breaches not only have the potential to cause data confidentiality issues but can also impact service continuity and operations [3].

An Intrusion Detection System (IDS) is a security mechanism for monitoring network or system activity to detect signs of an attack. Based on the data sources being monitored, IDSs are classified into host-based and network-based intrusion detection systems (NIDS) [4]. This study focuses on developing a machine learning model for a network traffic-based NIDS. As the complexity of cyber threats in IoT and IIoT environments increases, machine learning-based IDS are being developed more frequently because they can automatically learn network traffic patterns and detect various types of attacks more adaptively than rule-based approaches [5]. However, their performance is highly dependent on the alignment of the training data distribution with operational conditions. When a shift in the data distribution occurs, detection performance may decline even if the model initially shows good results [6].

In IoT and IIoT environments, attack patterns and network traffic characteristics may vary across different deployment scenarios due to differences in the types of attacks encountered and the characteristics of the participating endpoints [7]. Such variations can lead to distribution shift, where the distribution of operational data differs from that of the training data. As a result, models that perform well during training may become less effective when deployed in different environments, increasing the risk of both missed attacks and false alarms [8]. Previous studies on IoT intrusion detection systems have identified distribution

shift as one of the primary causes of performance degradation in real-world deployments. Consequently, there is a need for model designs and evaluation frameworks that explicitly assess robustness against data distribution variations rather than relying solely on performance under a single training scenario [9].

Based on these considerations, this study evaluates the robustness of intrusion detection models under various data distribution shifts. In the context of IDS, strong generalization capability is essential because variations in attack patterns and network traffic characteristics can increase the risk of false alarms and undetected attacks [8]. To improve the stability of detection decisions under these conditions, the proposed approach employs a stacking ensemble model. Previous studies have shown that stacking ensembles can produce more stable classification performance than individual models when handling data with diverse characteristics [10], [11]. Accordingly, the proposed model is evaluated under multiple distribution shift scenarios to assess its detection performance consistency and error behavior.

XGBoost and LightGBM were selected as base learners because both are gradient boosting algorithms that are effective for tabular data but have different learning characteristics [12], [13], [14]. These differences have the potential to produce error patterns that complement each other, making them suitable for combination through ensemble stacking. By learning from the outputs of multiple models, stacking can improve generalization ability and reduce errors caused by data variations [15], [16]. Therefore, the XGBoost–LightGBM ensemble stacking approach has the potential to produce more stable detection results under varying data distribution conditions.

To ensure that the testing is measurable and relevant to the IoT/IIoT context, this study uses the ToN-IoT dataset's network data subset as the experimental data source [17]. The use of a benchmark dataset helps ensure the reproducibility of the research and facilitates objective comparison of results [18]. Additionally, this study proposes the development of a stacking ensemble-based intrusion detection model to improve the quality of detection decisions compared to single models [10], [19], and applies hyperparameter optimization using Bayesian Optimization, which

is known to be efficient for finding good model configurations with a limited number of evaluations [20].

In addition to conventional evaluations based on data files, this study also considers how IDSs operate in practice, particularly NIDSs, which monitor network traffic on specific segments or devices and issue notifications when suspicious activity is detected [4]. Previous research has noted that machine learning-based IDS evaluation should not be limited to static classification metrics but must also consider operational aspects, such as inference time, throughput, and computational efficiency when the model is run in an environment that approximates real-world conditions [21]. Therefore, the testing in this study is supplemented with a simple streaming inference simulation to obtain an initial understanding of the model's operational behavior when test data is streamed sequentially.

Based on the above, this study uses an XGBoost–LightGBM ensemble stacking model optimized using Bayesian Optimization on the ToN-IoT "Network" subset dataset to evaluate intrusion detection capabilities under various data distribution conditions. The evaluation was conducted through unseen attack and unseen endpoint scenarios, as well as streaming inference simulations, to obtain a more comprehensive picture of the model's performance consistency and operational costs compared to a single model in an IoT/IIoT environment.

1.2. Problem Statements

Based on the background described above, the following four research questions can be formulated:

1. How does the XGBoost–LightGBM stacking ensemble-based intrusion detection model perform against distribution shifts in scenarios involving unseen attacks and unseen endpoints, compared to a single model, based on the Macro-F1, PR-AUC, and Recall metrics, as well as an analysis of false positives and false negatives?
2. How does the performance consistency and operational characteristics of the XGBoost–LightGBM stacking ensemble model compare to those of a single model in streaming inference simulations, in terms of detection performance, latency, and throughput?

1.3. Research Objectives

Based on the research questions outlined above, this study has several objectives, namely:

1. To evaluate the performance of the XGBoost–LightGBM stacking ensemble-based intrusion detection model against distribution shifts in scenarios involving unseen attacks and unseen endpoints, compared to a single model, using the Macro-F1, PR-AUC, and Recall metrics, as well as an analysis of false positives and false negatives.
2. Evaluate the consistency of performance and operational characteristics of the XGBoost–LightGBM stacking ensemble-based intrusion detection model compared to a single model in streaming inference simulations, in terms of detection performance, latency, and throughput.

1.4. Research Contributions

The results of this study are expected to provide the following benefits:

1. For academics, this study is expected to enrich the study of machine learning-based intrusion detection models in IoT/IIoT environments, particularly in the face of changes in data distribution. This study not only evaluates classification performance but also examines the model's robustness to unseen attack and unseen endpoint scenarios, detection error characteristics through false positives and false negatives, as well as operational aspects such as inference latency and throughput. Thus, this study can serve as a reference in developing more comprehensive evaluation methods for assessing a model's generalization ability and practical applicability.
2. For practitioners, this study is expected to serve as a reference in selecting and evaluating machine learning-based intrusion detection models before implementing them in IoT/IIoT environments. The research results provide insights into the impact of changes in data distribution on detection performance, detection error characteristics, and the operational characteristics of the models in streaming inference simulations, and can thus be used as a basis for determining which model best suits implementation needs.

1.5. Scope and Limitations

To maintain the focus of the study and limit the scope of the discussion, this study has the following limitations:

1. This study is limited to the development and evaluation of a machine learning-based intrusion detection model intended for use in NIDS within the IoT/IIoT context.
2. The dataset used is the ToN-IoT secondary dataset (Network subset), obtained from the official ToN-IoT website. The study uses the `train_test_network.csv` file, which is the official subset provided by UNSW Canberra Cyber.
3. The focus of this study is to compare ensemble stacking models—using XGBoost and LightGBM as base learners and Logistic Regression as the meta-learner—against a single model (baseline) for the intrusion detection task.
4. This study limits its analysis of robustness to novelty to two forms of data distribution changes: unseen attacks and unseen endpoints. Other forms of novelty, such as temporal drift, concept drift, and cross-dataset evaluation, are not discussed in this study.
5. The analysis of distribution shifts in this study is limited to an empirical approach based on the distribution of key features derived from Permutation Feature Importance (PFI) and the Jensen-Shannon Divergence (JSD) metric. Therefore, the analysis results are used as an aid in interpreting model behavior and not as a comprehensive representation of all characteristics of the data distribution.
6. Testing was conducted through conventional evaluation and streaming inference simulations without online model updates.
7. This study does not address the implementation of intrusion detection models on edge devices or production infrastructure. Simulations were conducted using a Python-script-based computational environment as a controlled evaluation approach.