CHAPTER I
INTRODUCTION

1.1 Background

In recent years, the explosive growth of connected devices in the Internet of Things (IoT) ecosystem has led to an exponential increase in network traffic, much of which is vulnerable to various forms of cyberattacks. Intrusion Detection Systems (IDS) have emerged as a critical component for ensuring network security, particularly in large-scale environments. However, traditional IDS models often rely on fully supervised learning, which assumes the availability of extensive labeled datasets—a condition that is rarely met in real-world deployments. The lack of labeled data, coupled with severe class imbalance and evolving attack patterns, poses significant challenges to IDS development.

Intrusion Detection Systems (IDS) are of significant importance in the realm of modern network defense, as they are capable of facilitating the identification of malicious activities through the continuous monitoring of network traffic [1]. Conventional IDS methodologies often rely on signature-based detection, where incoming traffic is compared against a repository of known attack patterns, or on supervised learning models that require substantial amounts of labeled data for training. However, the rapid emergence of novel cyberattacks and the increasing complexity of network environments render these methods insufficient [2]. The limitations of signature-based systems in detecting zero-day exploits and the shortcomings of supervised models in generalizing beyond their labeled training data reveal critical vulnerabilities in network protection.

To address these limitations, semi-supervised learning (SSL) has emerged as a promising solution by combining a small amount of labeled data with a much larger pool of unlabeled instances. This paradigm reduces the reliance on costly manual annotation while improving model adaptability [3]. It has been demonstrated that techniques such as self-training, co-training, and graph-based label propagation can reveal latent structures in network traffic and generate high-confidence pseudo-labels that undergo refinement during iterative training [4], [5]. These capabilities render SSL particularly valuable in the realm of intrusion detection, where labeled data is scarce yet patterns of abnormal behavior must still be captured in dynamic, real-world scenarios.

However, the integration of SSL within the IDS framework poses certain challenges, most notably the high dimensionality of network flow data, which can include packet headers, flow statistics, and a multitude of protocol-specific features. In the absence of dimensionality reduction, SSL models face the risk of overfitting and incur significant computational overhead. Incremental Principal Component Analysis (IPCA) has been developed to address these issues by updating principal components in mini-batches, thereby enabling memory-efficient projection of large or streaming datasets [6]. The present study explores the potential of random projection as an alternative to IPCA, particularly in the context of mapping high-dimensional data into a lower-dimensional space [7]. This approach utilizes a simple random matrix to preserve pairwise distances with a high degree of probability. The employment of IPCA or random projection to reduce the original feature space by half enables the SSL pipeline to achieve accelerated training and inference while preserving the most informative variance for robust intrusion detection.

These strengths are built upon by cascaded semi-supervised co-training frameworks, which introduce an additional layer of structure to the learning process. Rather than addressing the entire classification task in a single step, the dataset is divided into major and minor attack types. This approach enables the model to handle easier decisions first and defer more ambiguous cases to a subsequent stage. The architecture is composed of two stages. The first stage refines predictions for minority classes. This refinement is achieved by using co-trained classifiers alongside graph-based agreement mechanisms and flexible confidence thresholds. The purpose of these mechanisms and thresholds is to minimize labeling errors. By employing this approach, the cascading structure not only facilitates the detection of multiple classes but also substantially enhances performance under conditions of limited supervision.

1.2 State of the Art

Recent advancements in semi-supervised learning (SSL) have demonstrated its effectiveness in handling imbalanced and label-scarce environments, particularly in security-sensitive domains such as intrusion detection. Wrapper-based SSL methods, including self-training and co-training, rely on confidence thresholds to determine the inclusion of pseudo-labeled samples. Vale *et al.* introduced a flexible confidence mechanism (FlexCon) to dynamically adjust thresholds across iterations, improving generalization over static cutoffs [8]. Medeiros *et al.* further enhanced this framework through FlexCon-CE, which integrates classifier ensembles into the

confidence estimation process, yielding superior performance on 20 public datasets when compared to fixed or single-model thresholds [9]. These methods highlight the benefit of adaptive labeling in SSL, especially for class-imbalanced problems.

The integration of ensemble deep learning in intrusion detection systems (IDS) for Internet-of-Things (IoT) networks has also gained traction. Lazzarini *et al.* developed DIS-IoT, a stacked ensemble of deep learning models, which achieved exceptional accuracy and extremely low false-positive rates across ToN-IoT, CICIDS2017, and SWaT datasets [10]. Similarly, hybrid CNN-LSTM architectures have achieved over 99% accuracy on binary and multi-class intrusion detection tasks, illustrating the robustness of feature extraction when combining spatial and temporal information [11]. These studies confirm the potential of deep ensembles in enhancing IoT attack detection performance.

Optimized machine learning pipelines tailored to intrusion detection have also shown promise. Amine *et al.* proposed a CNN model equipped with data augmentation and dropout regularization to mitigate overfitting on IoT-specific data, achieving perfect precision in certain binary classes and over 82% in multi-class detection [12]. Another hybrid framework combining XGBoost and a sequential neural network achieved 99.93% and 99.00% accuracy in binary and multi-class settings, respectively, showcasing how hyperparameter optimization and model diversity can significantly elevate IDS effectiveness [13].

Specialized IDS solutions for protocol-specific threats further underscore the need for adaptable detection strategies. Ahmad *et al.* presented a DNN-based IDS for MQTT-based IoT networks, achieving 99.92% accuracy on a Uni-flow representation of attack traffic [14]. In a complementary approach, Solanki and Gupta introduced a stacked ensemble combining Gaussian Naïve Bayes, kernel SVM, and MLP with logistic regression meta-classification, attaining up to 99.80% accuracy in MQTT multi-class scenarios [15]. These results emphasize the value of tailored detection frameworks for IoT-specific protocols and architectures.

In summary, the literature suggests that effective IoT intrusion detection demands a combination of: (1) semi-supervised learning with adaptive confidence controls; (2) multi-model and ensemble architectures; (3) tailored preprocessing and optimization pipelines; and (4) contextual awareness of underlying protocols. These trends motivate the need for research that bridges adaptive SSL with efficient classifier stacking, particularly in the context of large-scale IoT datasets such as CICIoT2023.

1.3 Problem Identification

1. **Limited and Costly Labeling:**
   The process of manually annotating network traffic is both time-consuming and expensive, which hinders the ability to maintain current, up-to-date labeled datasets that accurately reflect the latest attack variants. Supervised IDS models frequently become outdated as new threats and zero-day exploits emerge at a rate that outpaces the production of labels, leading to the failure to detect novel anomalies. The utilization of small annotated subsets can result in the phenomenon of overfitting and suboptimal generalization when employing purely supervised approaches. This is due to the limited availability of reliable labels, which restricts the diversity of patterns that can be learned. Consequently, methods capable of leveraging large volumes of unlabeled traffic to enhance model training—without assuming the ability to automatically identify entirely unseen attacks—are increasingly in demand.

2. **Complex, High-Dimensional Data:**
   High-dimensional datasets are created by incorporating various features, such as packet header fields, protocol flags, and flow-level statistics, which are frequently present in network traffic flows. The ”curse of dimensionality” can lead to computational overhead and degrade model performance when training classifiers on vast feature spaces, while also making it more difficult to divide these features into complementary views for co-training. Effective dimensionality reduction (e.g, PCA, IPCA) and careful feature-view design are necessary to streamline computation and improve classifier robustness.

3. **Risk of Noisy Pseudo-Labels:**
   In semi-supervised co–training, classifiers iteratively assign pseudo-labels to unlabeled samples based on prediction confidence, but small amounts of mislabeling may still occur during over-iterated sessions. Correct pseudo-labels can propagate through the training process if confidence thresholds are not carefully managed or correlated feature views among classifiers, which can harm the overall detection accuracy. This is particularly problematic for some statistical models. The risk is reduced by utilizing dynamic thresholding and feature-diverse co-training views, which ensure that only the most accurate predictions are taken into account. Despite this, controlling pseudo-label noise is still a key obstacle to high accuracy when dealing with limited labeled data.

1.4 Objective and Contributions

This research aims to design a two-stage semi-supervised learning (SSL) framework to enhance multi-class intrusion detection under conditions of limited labeled data, class imbalance, and high-dimensional feature spaces. Unlike prior approaches that process all traffic uniformly, the proposed system adopts a true cascaded architecture in which each stage is functionally distinct and sequential. Stage 1 performs coarse-grained classification using Naive Bayes to identify well-represented major attacks and benign traffic. Only traffic deemed ambiguous or associated with underrepresented classes is passed forward to Stage 2, which applies a more expressive Random Forest classifier for fine-grained detection of minor attacks. To achieve this goal, the following research objectives are addressed:

1. Design a cascaded co-training framework that processes network traffic hierarchically. Stage 2 is only activated if Stage 1 is uncertain or detects potential minority class samples.

2. Apply adaptive confidence thresholds using variants of Flexible Confidence (FlexCon) to govern pseudo-label propagation dynamically, reducing false positives while maximizing label utility.

3. Incorporate dimensionality reduction methods, including Incremental PCA and Random Projection, to enhance scalability, minimize redundancy, and support real-time deployment scenarios.

4. Evaluate the framework across multiple label budgets and feature variants, comparing accuracy, runtime, and class-wise performance under realistic labeling constraints.

To achieve these goals, this research proposes a two-stage semi-supervised learning framework that divides the intrusion detection task based on class dominance. By structuring the system in stages, the framework can separately optimize detection strategies for both well-represented and underrepresented attack types. In each stage, co-training is used to expand the training set using unlabeled samples, guided by adaptive confidence thresholds. This approach is designed to reduce dependence on labeled data, address class imbalance, and improve multi-class detection accuracy. The experiments are structured to evaluate this framework against single-stage and fixed-threshold baselines, ensuring that each component of the design contributes meaningfully to the final detection performance.

1.5 Scope of Work

The scope of this thesis is as follows:

1. **Dataset:** This research uses the CICIoT2023 dataset from the Canadian Institute for Cybersecurity. This dataset comprises 33 executed attacks on an IoT network, categorized into seven classes (DDoS, DoS, Recon, Web-based, Brute Force, Spoofing, Mirai). This research focuses on the network intrusion content (e.g., DoS, DDoS, Mirai) and treats it as a general anomaly detection dataset, not specifically on IoT device characteristics.

2. **Problem Domain:**
   The work is confined to multi-class classification of network traffic flows. This research does not address IoT device management, real-time streaming, or unsupervised clustering beyond the semi-supervised paradigm outlined.

1.6 Research Methodology

This thesis is divided into several work packages (WP).

- WP 1: Literature Review
  This WP involves reviewing existing research on semi-supervised learning (SSL), SSL co-training, and cascading intrusion detection systems. The goal is to understand the state-of-the-art and identify gaps in the SSL co-training implementation in high-dimensional network data.

- WP 2: Data Acquisition and Preprocessing
  This WP focuses on acquiring the CICIoT2023 dataset and pre-processing it. This includes feature extraction, cleaning, and normalization. Apply dimensionality reduction (Incremental PCA or Random Projection) to make the data manageable.

- WP 3: Stage 1 and Stage 2 Implementation
  In this WP, the two-stage co-training modules are implemented in parallel. In Stage 1, two Naive Bayes classifiers are developed, each trained on a different feature view, and co-trained by exchanging high-confidence pseudo-labels according to the FlexCon dynamic threshold mechanism. In Stage 2, two Random Forest classifiers are similarly co-trained—also using the FlexCon thresholds—but applied directly to the minor-attack subset of the training data rather than depending on Stage 1's outputs.

- WP 4: Training and Testing
  In this WP, execute training runs for both stages under each label proportion scenario. Iteratively refine the threshold policy based on preliminary results. After training, apply the full model to the 20% test sets withheld from the beginning.

- WP 5: Performance and Security Evaluation
  In this WP, evaluate the trained models using accuracy and execution time. Use k-fold cross-validation scoring to assess performance stability. Compare results across different label proportions and against baseline methods.

- WP 6: Documentation and Thesis Writing
  This WP focuses on documenting the research findings, including results, challenges, and proposed solutions. The outcome will be compiled into the thesis, along with recommendations for future research and practical implementation.

1.7 Expected Results

It is hypothesized that the proposed cascaded SSL framework will enhance detection accuracy under limited-label conditions. By leveraging unlabeled data in both stages and applying adaptive FlexCon thresholds, the model is expected to approach the accuracy of a purely supervised model on well-represented classes, while offering improved sensitivity to modification attacks and novel variants. It has been demonstrated in previous studies that confidence mechanisms that are dynamically adjusted can exhibit superior performance in comparison to static thresholds in the context of semi-supervised methods. For instance, Vale et al. have demonstrated that FlexCon variants consistently surpassed original self-training and co-training methods across a range of diverse datasets by automating the adaptation of thresholds [8]. It is anticipated that dynamic thresholding will minimize false pseudo-labeling, and the two-stage cascade will more effectively capture the multi-class structure than a single classifier. The objective of this design is to deliver robust multi-class intrusion detection with substantially fewer labeled samples, thereby validating the central hypothesis of the thesis.

1.8 Research Plan and Action Point

This thesis plans to follow the plans shown in Table 1.1:

Table 1.1 Research Plan

Month	Action Points
October 2024	Search for existing research on SSL, SSL co-training, and cascading intrusion detection system.
November 2024	Write literature review and identify gaps in the SSL co-training implementation in high-dimensional network data.
Desember 2024	Select and optimize CICIoT2023 as the dataset.
January 2025	Implement SSL co-training using k-fold cross-validation on CICIoT2023 dataset.
February 2025	Implement and continue to optimize stage 1 SSL co-training on CICIoT2023 dataset.
March 2025	Implement and continue to optimize stage 2 SSL co-training on CICIoT2023 dataset.
April 2025	Continue training and testing using different proportion of label.
May 2025	Evaluate the performance of the optimized SSL co-training model.
June 2025	Compile the research findings, including results, challenges, and proposed solution to the thesis.