CHAPTER I
INTRODUCTION

1.1 Background

In the current era of digital transformation, internet connectivity has become the vital backbone for the operations of critical infrastructure, governments, and the global business sector. This massive exchange of data, unfortunately, corresponds directly to an increase in the volume and sophistication of cyber threats that compromise information confidentiality, integrity, and availability [1]. Attacks such as Distributed Denial of Service (DDoS), ransomware, and session hijacking are no longer executed manually; instead, they have become highly automated and large-scale. This situation establishes network security as a non-negotiable, top priority for maintaining the continuity of digital services. Consequently, there is an urgent need for proactive and responsive defense mechanisms to mitigate the risks of severe financial and reputational losses [2].

The urgency of cyber defense becomes increasingly evident when examining the escalating financial impact and frequency of global attacks. According to projections by Cybersecurity Ventures, global damages from cybercrime are expected to reach a staggering 10.5 trillion USD annually by 2025, a dramatic increase from 3 trillion USD in 2015 [3]. The financial impact per incident also continues to rise; IBM's *Cost of a Data Breach 2024* report notes that the average global cost of a single data breach has reached an all-time high of 4.88 million USD, representing a 10% increase from the previous year [2]. In terms of threat volume, data from Check Point Research indicates a 30% surge in global cyberattacks during the second quarter of 2024 compared to the same period last year, with organizations facing an average of over 1,600 attacks per week [4]. These statistics indicate that cyberattacks are no longer mere operational or technical disruptions; rather, they have evolved into macro-economic threats that demand far more robust defense strategies.

Serving as the frontline of cybersecurity, the Network Intrusion Detection System (NIDS) monitors network traffic and identifies suspicious activities in real-

time. Historically, conventional NIDS heavily relied on signature-based methods, which match data packets against a database of known threats. However, this approach possesses a fundamental flaw: the inability to detect novel attacks (zero-day attacks) or modified attack variants [5]. This limitation has triggered a paradigm shift in research toward anomaly-based NIDS that leverages artificial intelligence, specifically Machine Learning (ML) [1]. The ML-based approach offers superior generalization capabilities, as it can learn normal behavioral patterns and detect deviations without requiring specific signatures.

Although the application of Machine Learning in NIDS promises more adaptive detection, its real-world implementation faces significant technical challenges regarding computational resources. Highly complex ML models, such as multi-layered Deep Learning architectures, are often capable of achieving remarkably high accuracy in recognizing attacks [6]. However, this complexity incurs heavy computational costs, directly impacting processing duration or inference time. In high-speed network environments with throughput reaching gigabits per second, even a minor processing delay can cause packet queues to pile up (bottlenecks) or allow attacks to bypass the system before it can react [7]. Consequently, relying solely on accuracy metrics without considering processing speed is an unrealistic approach for practical deployment.

Unfortunately, the majority of recent academic research in the NIDS domain remains trapped in an evaluation bias oriented exclusively toward detection performance. Numerous studies compete to achieve F1-scores or accuracy metrics approaching 100% by employing heavy ensemble models or deep neural network architectures [8]. This singular focus on accuracy frequently overlooks operational efficiency, despite the fact that NIDS often needs to be deployed on resource-constrained hardware, such as edge routers or Internet of Things (IoT) devices. This imbalance in research focus creates a gap between theoretically superior laboratory results and industrial demands for lightweight, high-speed systems. The lack of in-depth analysis regarding the trade-off between accuracy and inference time represents a critical research gap that needs to be addressed.

To address these efficiency challenges, this study utilizes the eXtreme Gradient Boosting (XGBoost) algorithm integrated with a Cost-Sensitive Learning approach. XGBoost is selected due to its architecture supporting parallel processing and high scalability [9], while Cost-Sensitive Learning is applied to mitigate the class imbalance problem prevalent in network data, where normal traffic heavily dominates the volume of attacks. Without this cost-sensitive approach, models tend to be biased toward predicting the majority class (normal) and fail to detect rare attacks [10]. By assigning higher penalty weights to misclassified attacks, the model is compelled to learn threats more aggressively without excessively increasing computational complexity.

However, configuring XGBoost to balance detection accuracy (particularly regarding minority classes) and inference speed is a non-trivial task. The use of default parameters is often sub-optimal, and manual tuning methods are highly inefficient and prone to human error [11], [12]. The complexity of this issue indicates that finding an ideal NIDS model is a Multi-Objective Optimization Problem. A systematic search method is required to navigate the extensive hyperparameter space to identify the optimal trade-off between two conflicting objectives: maximizing detection quality while minimizing processing time.

This study proposes a comprehensive comparative approach by implementing a Multi-Objective Optimization framework. To evaluate the effectiveness of searching for optimal solutions, this study compares three distinct search strategies. First, the Non-dominated Sorting Genetic Algorithm II (NSGA-II) is used to represent evolutionary methods that excel at maintaining solution diversity along the Pareto Front [13]. Second, the Tree-structured Parzen Estimator (TPE) algorithm is applied as a Bayesian probability-based method known for its convergence efficiency [14]. Third, an XGBoost model with default parameters serves as the baseline to evaluate the benefits of hyperparameter optimization achieved by TPE and NSGA-II.

Beyond the complexities of finding an optimal model, the validation of intrusion detection systems frequently faces a gap between experimental evaluation and operational deployment [15]. The majority of studies evaluate models using static

metrics on randomly partitioned data (train-test split), which often fails to capture the continuous data stream characteristics of real-world networks. To bridge this gap without risking physical infrastructure, a trace-driven simulation approach serves as a crucial solution for precisely modeling system performance [16]. This method allows for the sequential replay of historical traffic data to mimic pseudo-real-time behavior. By implementing this simulation, this study not only demonstrates model accuracy theoretically but also visualizes the stability of inference latency when the model processes incoming data packet queues. This provides a more compelling proof-of-concept regarding the system's feasibility, both in terms of performance efficiency on constrained devices and real-time security monitoring capabilities [17], [18].

All experiments are executed using the NF-UNSW-NB15 dataset, a modern standard for intrusion datasets in a compact yet representative NetFlow format [19]. By combining the efficiency of the Cost-Sensitive Learning-enhanced XGBoost algorithm with a rigorous evaluation among NSGA-II, TPE, and the XGBoost Baseline, this study is expected to provide robust empirical evidence. The ultimate outcome is not merely a high-accuracy model, but an in-depth analysis of which optimization method is most effective in producing a balanced, efficient, and statistically validated NIDS framework for deployment in real-world network infrastructures.

In this study, the term "detection accuracy" does not solely refer to the general accuracy metric; rather, it is operationally represented using the Macro F1-Score. The Macro F1-Score was selected because the NF-UNSW-NB15 dataset exhibits an imbalanced class distribution, meaning that an evaluation based strictly on accuracy could potentially present a performance depiction biased toward the majority class.

1.2 Problem Formulation

Based on the background described above, the core problems to be addressed in this study are formulated into the following research questions:

1. How can the XGBoost algorithm be integrated with a Cost-Sensitive Learning approach to handle class imbalance in intrusion detection using the NF-UNSW-NB15 dataset?

2. How do the multi-objective optimization performances of the NSGA-II (Non-dominated Sorting Genetic Algorithm II) and Tree-structured Parzen Estimator (TPE) algorithms compare against the baseline XGBoost model in finding the optimal balance between the Macro F1-Score and inference time?
3. What are the characteristics of the Pareto Front produced by the optimization process, and which hyperparameter configuration delivers the non-dominated (optimal) solution for NIDS scenarios that demand high accuracy as well as rapid response times?
4. How can a trace-driven intrusion detection simulation be designed and implemented to visualize the performance of the optimized XGBoost model in monitoring historical traffic streams in a pseudo-real-time manner?

1.3 Research Objectives

In alignment with the research questions outlined above, the specific objectives to be achieved in this study are:

1. To construct an XGBoost-based intrusion detection model capable of handling imbalanced data within the NF-UNSW-NB15 dataset through a class weighting mechanism.
2. To evaluate and compare the effectiveness of two optimization algorithms, namely NSGA-II and TPE, against the XGBoost Baseline, in navigating the hyperparameter search space to achieve two conflicting objectives: maximizing the Macro F1-Score and minimizing inference time.
3. To analyze the Pareto-optimal solution set to determine the ideal trade-off point, as well as to empirically test whether the NSGA-II and TPE-based optimization methods provide significant performance differences compared to the XGBoost Baseline.
4. To develop a trace-driven detection simulation prototype capable of visualizing the model's operational mechanism in threat detection and displaying inference latency metrics as a proof-of-concept for system efficiency.

1.4 Research Benefit

This research is expected to yield positive impacts across multiple dimensions, including theoretical contributions, practical applications, and personal development for the author:

1. Theoretical Benefits (Scientific Contributions)
    a) Empirical Validation of Time Efficiency: Providing empirical evidence that the "Inference Time" parameter can be treated as an objective function equivalent in weight to accuracy within cybersecurity model optimization.
    b) Comparative Analysis of Optimization Algorithms: Contributing a comprehensive performance comparison between evolutionary methods (NSGA-II) and probabilistic methods (TPE) in searching for Pareto-optimal solutions on high-dimensional datasets.
    c) Enrichment of NetFlow Literature: Expanding the academic literature regarding the effectiveness of modern flow-based data, specifically the NF-UNSW-NB15 dataset, as a more efficient alternative to raw packet data (pcap).
2. Practical Benefits (Operational Applications)
    a) Deployment on Resource-Constrained Devices: Producing lightweight NIDS model candidates suitable for deployment on low-resource hardware, such as IoT gateways, without over-allocating CPU resources.
    b) Operational Flexibility for Network Administrators: Offering adaptable model options through the Pareto Front curve, which enables network administrators to select models based on dynamic, real-time network priorities (such as extreme processing speed versus maximum detection precision).
    c) Visual Integration Insights: Providing a visual dashboard framework that demonstrates how Machine Learning-based NIDS models can be

seamlessly integrated into security monitoring systems to detect threats and measure inference latency concurrently.

3. Personal Benefits for the Author
    a) Practical Application of Knowledge: Serving as a vehicle to apply academic theories acquired throughout the degree program particularly in Machine Learning, Data Mining, and Network Security to resolve real-world operational challenges.
    b) Technical Skill Enhancement: Sharpening technical competencies in utilizing advanced optimization frameworks (such as Optuna) and managing large-scale data analysis within cloud computing environments.

1.5 Research Limitation

To ensure that the discussion remains focused and the research outcomes are interpreted within the appropriate context, the author defines the scope and limitations based on the methods and environments utilized in the experimental setup:

1. Specific Dataset (NetFlow Focus)

    This study utilizes the NF-UNSW-NB15-v3 dataset. The data consists of flow statistics (NetFlow) rather than raw packet captures (pcap). Consequently, this research does not analyze payload content (such as email bodies or message text), but focuses exclusively on statistical communication patterns (e.g., duration, bytes, ports, etc.).

2. Focus on the XGBoost Algorithm

    The classification algorithm employed is limited to XGBoost (eXtreme Gradient Boosting). This study does not perform cross-algorithmic comparisons (e.g., comparing XGBoost against Random Forest or Support Vector Machines); instead, it focuses on optimizing the internal mechanics of XGBoost itself using Cost-Sensitive Learning techniques to handle data imbalance.

3. Optimization Methods and Baseline

The hyperparameter tuning process is executed using the Optuna library. The search algorithms under comparison are restricted to three methods:

a) NSGA-II (Non-dominated Sorting Genetic Algorithm II), representing multi-objective evolutionary algorithms.
b) TPE (Tree-structured Parzen Estimator).
c) XGBoost Baseline (serving as the baseline or fundamental benchmark).

4. Scope of Experimentation (Offline Simulation)

This research is designed as an offline simulation utilizing historical data. This implies that:

a) The model is trained and tested using a static dataset consisting of recorded historical network traffic.
b) The intrusion detection system is implemented using a trace-driven simulation approach, wherein the test set is streamed sequentially in a pseudo-real-time manner to mimic actual network traffic behavior, rather than via live packet capture from physical network hardware.

5. Computing Environment (Hardware and Software)

All experiments are executed within the cloud-based Kaggle Notebooks environment.

a) Hardware: The setup utilizes an NVIDIA Tesla P100 GPU accelerator from Kaggle Notebooks.
b) Implication: The measured "Inference Time" (in milliseconds or seconds) is highly dependent on the specifications of the P100 GPU. Although absolute time values may vary when executed on alternative hardware platforms, the relative comparative trends remain fully valid as an efficiency reference.