

Machine Learning and Deep Learning Approaches for Class-Imbalance-Aware and Adversarially-Robust Botnet Detection in IoT Networks

A Research Proposal

Student Name: Mohammed Hussein Bagharib

Student ID: 252000080

Course: Research Methodology in Computer Science (AI800)

Lecturer: Dr. Mohammed Anbar

Abstract

The rapid growth of Internet of Things (IoT) devices has expanded the cyberattack surface, and botnet malware families such as Mirai and Bashlite can compromise hundreds of thousands of resource-constrained devices. Machine learning (ML) and deep learning (DL) detectors are the main practical defense and routinely report near-perfect accuracy on IoT botnet traffic. However, two weaknesses undermine the trustworthiness of these reported figures. First, most detectors are evaluated on clean, unperturbed traffic and ignore adversarial evasion, under which undefended classifiers have been shown to collapse from above 99% accuracy to below 44%. Second, the severe class imbalance between normal and botnet traffic is frequently left uncorrected, inflating apparent performance. Crucially, class-imbalance handling and adversarial robustness have been studied in isolation so far: imbalance correction on IoT traffic without adversarial testing, and adversarial robustness without IoT botnet traffic or imbalance correction. This research proposes to close that gap by building an imbalance-aware IoT botnet detector, exposing it to gradient-based evasion attacks (FGSM, BIM, DeepFool, JSMA), designing an adversarial-training defense, and verifying that class-imbalance handling and adversarial robustness hold simultaneously. Experiments will use the Bot-IoT and N-BaIoT-2021 datasets and report F1-score, AUC-ROC, Matthews Correlation Coefficient, and accuracy retained under attack. The expected outcome is an empirically validated answer to whether IoT botnet detection can be made reliable under class imbalance and resilient under adversarial evasion at the same time.

Chapter 1: Introduction

1.1 Introduction

The Internet of Things (IoT) refers to the development and deployment of interconnected devices, such as smart-home systems, sensors, and healthcare equipment, that collect, transmit, and act on data with minimal human involvement. The rapid spread of these devices has greatly expanded the cyberattack surface, especially for botnet-based threats that exploit the limited security capabilities of IoT endpoints. Botnet malware families such as Mirai and Bashlite have compromised hundreds of thousands of devices, enabling distributed denial-of-service (DDoS) attacks, data theft, and large-scale network infiltration. Because IoT devices are resource-limited and often deployed by non-expert users, automated machine learning-based detection has become the main practical defense against botnet infection, making the reliability of these detectors' reported accuracy a matter of direct security importance.

1.2 Background

Machine learning (ML) and deep learning (DL) intrusion detection systems (IDSs) learn to separate benign IoT traffic from malicious botnet traffic using labeled network-flow or behavioral features. Reported results are impressive: SMOTE-DRNN reaches a 99.62% F1-score (Popoola et al., 2021), the ELBA-IoT boosted decision-tree ensemble reaches 100% accuracy on binary classification (Abu Al-Haija & Al-Dala'ien, 2022), and a comparative CNN reaches 99.94% accuracy (Idrissi et al., 2021). Two methodological issues, however, threaten the validity of such figures. The first is class imbalance: in realistic IoT traffic, benign flows vastly outnumber botnet flows, so a model can appear highly accurate while failing on the rare attack class. Oversampling methods such as SMOTE are commonly used to correct this. The second issue is adversarial evasion: an attacker can add small, carefully crafted payloads to fool a classifier. Def-IDS shows that an undefended DL classifier drops from above 99% accuracy to between 32.8% and 43.4% under gradient-based attacks, while an ensemble adversarial defense can recover accuracy to about 97.9% (Wang et al., 2021). Recent work

confirms that these attacks are not merely theoretical for IoT: adversarial botnet traffic can be crafted to evade network IDSs while preserving malicious functionality (Debicha et al., 2023), and an IoT botnet detector can be driven to 0% detection by altering a single feature guided by model explainability (Alani et al., 2023).

1.3 Problem Statement

Bridging the gap between reported accuracy and adversarial robustness in class-imbalanced IoT botnet detection.

Several ML and DL methods for IoT botnet detection report near-perfect results, but these results are almost always reported on clean, unaltered test traffic, and only a minority of studies correct for the severe class imbalance that characterizes real IoT networks. Classifiers that ignore this imbalance likely overstate their true detection capability. Separately, the reviewed studies that evaluate adversarial robustness show that undefended DL intrusion detectors drop from above 99% accuracy on clean traffic to between 32.8% and 43.4% under gradient-based evasion attacks (FGSM, BIM, DeepFool, JSMA), while an ensemble defense recovers accuracy to about 97.9% (Wang et al., 2021). Yet that adversarial evaluation was conducted on an enterprise intrusion dataset that contains no IoT botnet traffic and no class-imbalance correction. More recent IoT-specific studies show that evasion attacks succeed against IoT and botnet detectors in realistic, black-box settings (Ibitoye et al., 2019; Debicha et al., 2023; Alani et al., 2023; Debicha et al., 2026), but none apply class-imbalance correction.

This exposes a precise, unexamined gap: no existing study has tested whether an IoT botnet detection model that already corrects for class imbalance remains accurate when confronted with adversarial evasion, nor whether an adversarial defense remains effective once class-imbalance correction is applied. Because imbalance handling and adversarial robustness have so far been studied separately, it is unknown whether imbalance-correction techniques preserve or worsen a model's vulnerability to evasion. Resolving this question is feasible within a single study because it builds on established, publicly available techniques and datasets: SMOTE

oversampling, ensemble and DL classifiers, the Bot-IoT and N-BaIoT-2021 datasets, and the FGSM/BIM/DeepFool/JSMA attacks with multi-source adversarial-training defense already validated in the literature (Popoola et al., 2021; Abu Al-Haija & Al-Dala'ien, 2022; Wang et al., 2021).

1.4 Research Objectives

This research aims to determine whether class-imbalance-corrected IoT botnet detection models retain their reported accuracy under adversarial evasion and to develop a defense that preserves both properties. The specific objectives are:

1. To establish an imbalance-aware baseline for IoT botnet detection by applying SMOTE oversampling with an ensemble or deep learning classifier on the Bot-IoT and N-BaIoT-2021 datasets, measuring baseline performance using F1-score, AUC-ROC, and Matthews Correlation Coefficient (MCC) to confirm reliable detection of both majority and minority traffic classes.
2. To generate adversarial IoT botnet traffic samples using FGSM, BIM, DeepFool, and JSMA evasion techniques and measure the resulting accuracy drop of the imbalance-aware baseline model, establishing a measurable robustness ceiling for comparison with later defenses.
3. To design and implement an adversarial defense mechanism, adapted from multi-source adversarial training, for the class-imbalance-corrected IoT botnet detector, and to measure the percentage of detection accuracy recovered under the same evasion attacks used in Objective 2.
4. To determine whether the defended model maintains its class-imbalance-aware detection quality (F1-score, AUC-ROC, MCC on minority botnet classes) together with its adversarial robustness (accuracy retained under attack), rather than improving one property at the cost of the other.

5. To benchmark the resulting class-imbalance-aware, adversarially robust detection model against the undefended baseline and against the accuracy figures reported in the reviewed literature, reporting the quantified improvement margin in both clean and adversarial conditions.

1.5 Research Scope

To keep the study feasible within the available time and resources, its scope is bounded as follows:

- Datasets: the Bot-IoT (UNSW Canberra) and N-BaIoT-2021 IoT botnet datasets.
- Threat model: evasion (test-time) attacks only, generated with FGSM, BIM, DeepFool, and JSMA in a white-box setting, with black-box transfer considered where feasible.
- Imbalance handling: SMOTE-based oversampling, with cost-sensitive learning considered as an alternative.
- Defense: multi-source adversarial training adapted to the imbalance-corrected detector.
- Metrics: F1-score, AUC-ROC, Matthews Correlation Coefficient, and accuracy retained under attack.
- Out of scope: physical deployment on IoT hardware; data-poisoning and backdoor attacks; non-botnet attack families.

Chapter 2: Literature Review

2.1 Summary of the research papers

Study 1: SMOTE-DRNN: A Deep Learning Algorithm for Botnet Detection in the Internet-of-Things Networks

SMOTE-DRNN is a hybrid framework that combines SMOTE with DRNN to tackle class imbalance issues in IoT botnet detection. SOMTE solves the class imbalance problem by generating additional artificial attack samples, and the DRNN is used to learn the representation of what an attack looks like. They use the Bot-IoT dataset (UNSW Canberra) with 73,370,320 records of 11 different traffic classes, then apply information gain to reduce the feature space to the 10 most relevant features. SMOTE is used on the training data to augment the synthetic minority class samples before training the DRNN, and the evaluation is done with F1-score, AUC-ROC, and MCC metrics to account for imbalance. The main contributions are the system design and an in-depth ablation study that tested the necessity of both individual SMOTE and DRNN components by comparing the results. Results show that the model performed 99.62% F1-score, 99.87% AUC-ROC, and 99.62% MCC, which outperformed the RNN, BLSTM, CNN, and MLP baselines.

The primary drawbacks are:

1. The model required approximately 1,147 seconds to train, making real-time retraining impractical.
2. DRNN has high computing cost. It is very difficult to apply it on the resource constrained IoT edge nodes.
3. The work is not generalizable, as it is evaluated only on one dataset. (Segun I. Popoola, Bamidele Adebisi, Ruth Ande, Mohammad Hammoudeh, Kelvin Anoh and Aderemi A. Atayero, 2021)

Study 2: ELBA-IoT: An Ensemble Learning Model for Botnet Attack Detection in IoT Networks

In this paper, the researchers present ELBA-IoT, an anomaly-based intrusion detection system (IDS) for IoT networks, that is based on an ensemble of three decision tree classifiers (AdaBoosted, RUSBoosted, and Bagged). The sizes of the classifiers are small enough to deploy on IoT devices, and the final decision is made based on their combined result. The system is tested on the N-BaIoT-2021 dataset of 619,096 samples (611,359 normal, 7,737 botnet) that represent Mirai and Bashlite attacks on five IoT devices. The data

preparation steps are Pearson Correlation Coefficient Score feature selection, Z-score normalization, and 5-fold cross-validation. The model classifies traffic at three granularities: binary (normal vs. anomaly), ternary (normal, Mirai, Bashlite), and 10-class. Main findings: ELBA-IoT reached 100% accuracy on both binary and ternary classification, 99.6% on the 10-class task with 97.7% F1-score, and 40 microseconds of inference overhead per sample (lowest of all compared methods).

The primary drawbacks are:

1. Our dataset only includes five IoT device types and two botnet families, and so it is not generalizable.
2. Real deployment on IoT hardware is not tested.
3. The large class imbalance between normal and botnet traffic is not explicitly addressed. (*Qasem Abu Al-Haija and Mu'awya Al-Dala'ien, 2022*)

Study 3: A Collaborative Approach to Early Detection of IoT Botnet

Contrary to most of the above papers, this paper's work concerns early indicators of a botnet attack. The paper proposes a cooperative machine learning model for early-stage detection of IoT botnet behavior, which could be effective for detection before the malware's complete execution. The authors execute IoT malware samples within a V-Sandbox environment, which is a virtual environment used to safely extract early behavior-based features including system calls, network connections, and file access activities. A late fusion ensemble of KNN, Random Forest, and KNN classifier used the probabilistic output of each model and combine it to reach the final decision regarding the classification. The dataset included in this paper contains a total of 8,911 samples that included benign IoT behavior as well as malicious behavior. The important results included in this paper include the model reaching an accuracy of 99.37%

with an AUC of 0.9896. The overall classification time required was 6 seconds, and the model was able to detect botnet behavior in an early stage before it was completely executed.

The primary drawbacks are:

1. The V-Sandbox environment introduces approximately 3 minutes of overhead per sample, which is not realistic for real-time detection.
2. The dataset of 8,911 samples is small compared to large-scale traffic datasets.
3. Sophisticated botnets and malware actively try to detect if they are running inside a sandbox; these evasion techniques can be easily defeated. *(Giang L. Nguyen, Braulio Dumba, Quoc-Dung Ngo, Hai-Viet Le and Tu N. Nguyen, 2022)*

Study 4: Toward a Deep Learning-Based Intrusion Detection System for IoT Against Botnet Attacks

The researchers present in this paper BotIDS, a comparison of 4 deep learning architectures. Convolutional Neural Network (CNN), SimpleRNN, LSTM, and GRU are trained and evaluated for 11-class botnet detection on the Bot-IoT dataset with standard preprocessing techniques of normalization and one-hot encoding, and all four architectures are trained and tested under the same conditions. Performance is evaluated using accuracy, loss, and prediction time per inference. Results show that the CNN outperforms all compared recurrent architectures, achieving 99.94% accuracy, 0.58% loss, and 0.34 ms prediction time. This indicates that local feature correlation using convolutional filters is better suited than sequential modeling for network traffic classification.

The primary drawbacks are:

1. Class imbalance in the Bot-IoT dataset is not addressed. Given that the majority of the dataset is normal traffic, this might lead to unrealistic results if the class imbalance is not handled using techniques like SMOTE.
2. An attacker can know that CNN is used and accordingly design his attack to bypass the CNN. Adversarial robustness is not assessed.
3. The proposed CNN architecture is not well-documented; this could make it difficult to reproduce the experiment to verify the results. *(Idriss Idrissi, Mohammed Boukabous, Mostafa Azizi, Omar Moussaoui and Hakim El Fadili, 2021)*

Study 5: Def-IDS: An Ensemble Defense Mechanism Against Adversarial Attacks for Deep Learning-Based Network Intrusion Detection

This paper introduces Def-IDS, a defense framework to protect against adversarial attacks on deep learning-based intrusion detection systems (IDSs). Def-IDS combines a Multi-source Generative Adversarial Network (MGAN) of four generators, where each generator is trained to generate adversarial examples based on one of the following four methods: FGSM, BIM, DeepFool, or JSMA, with Multi-source Adversarial Training (MAT), which retrain the IDS classifier with an augmented dataset of both the original clean and generated adversarial traffic. Def-IDS is evaluated on a state-of-the-art CSE-CIC-IDS2018 large scale modern dataset. Results show that without a defense the performance of DL classifiers on the adversarial testing samples dropped to 32.8 - 43.6% accuracy while Def-IDS can maintain an accuracy of about 97.9% and perform better than all single-source adversarial training baselines with a total training cost of 111.51 seconds.

The primary drawbacks are:

1. The system is not evaluated on IoT-specific traffic or deployed on IoT hardware.
2. Due to the expensive training cost of GAN, its deployment on resource constrained IoT edge devices are prohibitive.
3. The dataset CSE-CIC-IDS2018 does not include IoT-specific botnet traffic such as Mirai or Bashlite. *(Jianyu Wang, Jianli Pan, Ismail AlQerm, and Yuanni Liu, 2021)*

Study 6: Deep Transfer Learning: A Novel Collaborative Learning Model for Cyberattack Detection Systems in IoT Networks

This paper proposes Federated Transfer Learning (FTL), a cyberattack detection framework combining both Federated Learning (FL) and Transfer Learning (TL). The combined use of FL and TL enables participating IoT networks, all heterogeneous in their feature spaces and non-IID data distributions, to collaboratively train a distributed detection model, while not having to share their raw data. Transfer Learning allows heterogeneous feature spaces from different networks to be connected through a common representation space by using only around 10% of the data from each network that is exchanged, in order to align their shared layers for representation. The new FTL framework has been instantiated and evaluated on two cyberattack detection networks, using the KDD Cup 1999, NSL-KDD and UNSW-NB15 datasets. The major results showed that FTL achieved as high as 99% accuracy in attack detection, while increasing the accuracy against a baseline unsupervised distributed learning (UDL) model by as much as 40%. Only a marginal improvement was seen in FL when the networks' feature spaces were heterogeneous, making TL the critical component.

The primary drawbacks are:

1. The framework requires approximately 10% of mutual data between networks, which may be difficult to obtain in practice.
2. Negative transfer can occur when participating networks have too dissimilar traffic patterns.

3. The framework is only validated for two networks, and the KDD/NSL-KDD datasets are not IoT-specific. *(Tran Viet Khoa, Dinh Thai Hoang, Nguyen Linh Trung, Cong T. Nguyen, Tran Thi Thuy Quynh, Diep N. Nguyen, Nguyen Viet Ha and Eryk Dutkiewicz, 2022)*

Study 7: Analyzing Adversarial Attacks Against Deep Learning for Intrusion Detection in IoT Networks

This study analyses adversarial attacks against deep-learning-based intrusion detection in IoT networks using the Bot-IoT dataset from the Cyber Range Lab of UNSW Canberra. The authors compare a Feedforward Neural Network (FNN) with a Self-normalizing Neural Network (SNN) for classifying intrusion attacks, evaluating accuracy, precision, recall, and the multi-class Cohen's Kappa score. They then test both networks' adversarial robustness against adversarial samples generated from the IoT dataset. The results show that the FNN outperforms the SNN on clean traffic across all reported metrics, but the SNN demonstrates better resilience against adversarial samples, indicating that the most accurate model on clean data is not necessarily the most robust under attack.

The primary drawbacks are:

1. Only two simple feedforward architectures (FNN and SNN) are compared, so the findings may not extend to deeper or ensemble models.
2. Class imbalance in the Bot-IoT dataset is not addressed.
3. Adversarial robustness is evaluated only for these two models and not under a defended setting. *(Olakunle Ibitoye, Omair Shafiq and Ashraf Matrawy, 2019)*

Study 8: Adv-Bot: Realistic Adversarial Botnet Attacks against Network Intrusion Detection Systems

Adv-Bot investigates the real-world feasibility of evasion attacks against machine-learning-based network intrusion detection systems (NIDS) using realistic botnet traffic traces. Unlike most prior work that perturbs extracted features in theory, this study assumes as many real implementation constraints as possible in a black-box setting and generates adversarial botnet traffic that evades detection while still performing its intended malicious functionality. The authors also present a defensive scheme to protect ML-based IDSs against such attacks. Their experiments demonstrate that it is entirely possible to fool ML-based IDSs with the proposed adversarial algorithm under realistic constraints.

The primary drawbacks are:

1. The study focuses on the attack and its feasibility, with less emphasis on a comprehensive, generalizable defense.
2. Class imbalance between benign and botnet traffic is not addressed.
3. The evaluation is limited to botnet traffic and does not cover other IoT attack families. *(Islam Debicha, Benjamin Cochez, Tayeb Kenaza, Thibault Debatty, Jean-Michel Dricot and Wim Mees, 2023)*

Study 9: Adversarial Explainability: Utilizing Explainable Machine Learning in Bypassing IoT Botnet Detection Systems

This study introduces a novel attack that uses a machine-learning model's own explainability to evade detection by botnet detection systems. By leveraging the information obtained from the model's explanations, the authors construct adversarial samples that bypass detection in a black-box setting. The attack was tested on a trained IoT botnet detection system and was capable of bypassing the detector with 0% detection by altering only one feature to generate the adversarial samples.

The primary drawbacks are:

1. The work presents an attack only and does not propose or evaluate a corresponding defense.
2. Class imbalance between normal and botnet traffic is not addressed.
3. The attack relies on access to model explainability outputs, which may not always be available to an attacker. *(Mohammed M. Alani, Atefeh Mashatan and Ali Miri, 2023)*

Study 10: Targeted Adversarial Traffic Generation: Black-box Approach to Evade Intrusion Detection Systems in IoT Networks

This recent study evaluates the feasibility of evasion attacks on IoT network-based intrusion detection systems using a novel black-box adversarial attack, aiming to bridge theoretical vulnerabilities with real-world applicability. The authors address practical implementation constraints that are often overlooked in theoretical adversarial research, and they also propose a defense scheme tailored to mitigate the impact of evasion attacks. Their findings demonstrate successful evasion attacks against IDSs, while the proposed defense effectively detects the majority of adversarial traffic and performs competitively with current state-of-the-art defenses.

The primary drawbacks are:

1. Although a defense is proposed, its robustness is demonstrated against the authors' own attack and may not generalize to unseen attack strategies.
2. Class imbalance between normal and attack traffic is not addressed.

3. The approach is validated in a simulated IoT environment rather than on live deployed IoT devices. (Islam Debicha, Tayeb Kenaza, Ishak Charfi, Salah Mosbah, Mehdi Sehaki and Jean-Michel Dricot, 2026)

2.2 Critical Analysis

The ten reviewed studies collectively show that ensemble and hybrid approaches consistently outperform single-model baselines and that class imbalance must be handled explicitly, since studies that ignore it and report only accuracy likely overstate their true detection capability. Two gaps become sharper once the new studies are added. First, adversarial evasion is no longer a purely enterprise or theoretical concern: it is now demonstrated directly on IoT and botnet detectors, including on the Bot-IoT dataset itself (Ibitoye et al., 2019), with realistic, functionality-preserving traffic (Debicha et al., 2023) and with attacks so efficient that a single altered feature yields 0% detection (Alani et al., 2023; Debicha et al., 2026). Yet none of these adversarial studies correct for class imbalance. Therefore, the central unexamined question is whether an IoT botnet detector can be simultaneously imbalance-aware and adversarially robust. Until adversarial robustness and rigorous imbalance handling are evaluated together rather than separately, near-perfect reported accuracy should be treated as promising but not yet trustworthy for deployment.

2.3 Connections

6. Shared objective: all studies aim to improve automated detection of malicious traffic in IoT or network settings.
7. Ensemble and hybrid design: Studies 1, 2, 3, 5, and 6 combine multiple models or techniques, reflecting consensus that single models are insufficient.

8. Common benchmark: Studies 1, 4, and 7 all use the Bot-IoT dataset, enabling direct comparison across clean and adversarial conditions.
9. Adversarial thread: Studies 5, 7, 8, 9, and 10 form a coherent line on evasion attacks and defenses, progressing from enterprise data (5) to IoT-specific and realistic botnet traffic (7-10).
10. Imbalance handling: Studies 1 (SMOTE-DRNN) and 2 (ELBA-IoT) are the only studies that directly address class imbalance, however they don't test adversarial robustness.

2.4 Conflicts Between the Findings

11. CNN versus recurrent architectures: Study 4 finds CNN best on Bot-IoT while Study 1 reports strong results with a deep recurrent model on the same data; the conflict is likely explained by differing class-imbalance handling rather than a true architectural difference.
12. Efficiency versus thoroughness: Study 2 targets 40-microsecond inference for gateway deployment, Study 1 accepts ~1,147 s training for rigorous imbalance correction, and Study 3 needs ~3 minutes per sample for behavioural analysis, reflecting incompatible deployment assumptions.
13. Clean versus adversarial evaluation: most studies assume clean test traffic, but Studies 5, 7, 8, 9, and 10 show that accuracy can collapse under evasion, meaning clean-traffic accuracy claims describe only non-adaptive attackers.

2.5 Synthesis Matrix

Study	Authors (Year)	Design	Key Results	Methods / Data	Limitations
SMOTE-DRNN	Popoola et al. (2021)	Experimental	99.62% F1; 99.87% AUC-ROC; 99.62% MCC	Bot-IoT; info-gain (10 feats); SMOTE; DRNN	High training time; single dataset; edge cost
ELBA-IoT	Abu Al-Haija & Al-Dala'ien (2022)	Experimental	100% (binary/ternary); 99.6% (10-class); 40us	N-BaIoT-2021; Pearson FS; boosted/bagged DTs	5 devices/2 families; no HW test; imbalance ignored
Collaborative Early Detection	Nguyen et al. (2022)	Experimental	99.37% acc; AUC 0.9896; pre-execution	V-Sandbox behavioural; KNN+RF+KNN fusion	~3-min overhead; small data; sandbox evasion
BotIDS	Idrissi et al. (2021)	Experimental	CNN 99.94% acc; 0.34 ms	Bot-IoT; CNN vs SimpleRNN/LSTM/GRU	Imbalance ignored; no adversarial test
Def-IDS	Wang et al. (2021)	Experimental	~97.9% under attack; undefended 32.8-43.6%	CSE-CIC-IDS2018; MGAN + MAT (4 attacks)	No IoT data/HW; high GAN cost
Deep Transfer Learning	Khoa et al. (2022)	Experimental	Up to 99% acc; +40% over UDL	KDD/NSL-KDD/UNSW-NB15; FL	~10% mutual data; 2

(FTL)				+ TL	networks; legacy data
Adversarial Attacks on IoT IDS	Ibitoye et al. (2019)	Experimental	FNN best clean; SNN more robust	Bot-IoT; FNN vs SNN; adversarial samples	No imbalance correction; 2 models only
Adv-Bot	Debicha et al. (2023)	Experimental	Realistic evasion succeeds; defense proposed	Botnet traces; black-box functionality-preserving	No imbalance handling
Adversarial Explainability	Alani et al. (2023)	Experimental	0% detection by altering 1 feature	IoT botnet detector; explainability-guided attack	No imbalance handling; attack-focused
Targeted Adversarial Traffic	Debicha et al. (2026)	Experimental	Successful black-box evasion; tailored defense	IoT IDS; black-box attack + defense	No imbalance handling

Chapter 3: Research Methodology

3.1 Research Methodology

This research adopts an experimental, quantitative methodology. The section below outlines the steps, procedures, and tools that will be used to collect and analyse data, together with the rationale behind their selection, and it explains how each activity contributes to answering the research problem rather than simply listing tasks. All experiments use the publicly available Bot-IoT and N-BaIoT-2021 datasets so that results remain comparable with the reviewed literature. The methodology directly addresses the four research objectives: it builds an imbalance-aware baseline (Objective 1), attacks it (Objective 2), defends it (Objective 3), and benchmarks the result.

3.1.1 Research Design

The study follows a comparative experimental design with two primary experimental conditions: an undefended imbalance-aware baseline detector and a defended, adversarially trained detector. Each condition is evaluated under two data settings, clean traffic and adversarially perturbed traffic, producing a two-by-two comparison that isolates the individual and combined effects of class-imbalance correction and adversarial defense. This design is chosen because it allows the central research question, whether the two properties can hold simultaneously, to be answered by direct measurement rather than inference, and it mirrors the evaluation logic validated by Wang et al. (2021) and Ibitoye et al. (2019).

3.1.2 Study Groups (Datasets and Experimental Groups)

Because this is a computational study, the study groups are datasets and model configurations rather than human participants. Two benchmark IoT datasets are used: Bot-IoT (UNSW Canberra), which contains large-scale, highly imbalanced botnet traffic across eleven classes, and N-BaIoT-2021, which contains Mirai and Bashlite traffic captured from real IoT devices. Each dataset is split into

stratified training, validation, and test partitions. The experimental groups are: (i) a baseline classifier trained on imbalance-corrected data, and (ii) the same classifier hardened with adversarial training. Both groups are evaluated on identical clean and adversarial test sets to ensure a fair comparison.

3.1.3 Access to Samples (Data Access)

Both datasets are publicly available for academic research from their respective providers, so no new data collection or human-subject sampling is required. The data will be obtained under the providers' published terms of use and cited accordingly. Using established public datasets improves reproducibility and removes the ethical and logistical difficulties of collecting live IoT traffic.

3.1.4 Infrastructure (Tools and Environment)

Experiments will be implemented in Python. Data handling and classical models will use pandas, NumPy, and scikit-learn; class-imbalance correction will use the imbalanced-learn library (SMOTE and related variants); deep learning models will use TensorFlow/Keras or PyTorch; and adversarial examples (FGSM, BIM, DeepFool, JSMA) together with adversarial training will be produced using an established adversarial robustness library such as the Adversarial Robustness Toolbox (ART). Training and evaluation will run on a GPU-enabled workstation or cloud instance. These tools are selected because they are open-source, widely validated in the reviewed literature, and reproducible.

3.1.5 Procedure (Experimental Steps Mapped to Objectives)

Stage 1: Data preparation and imbalance-aware baseline (Objective 1).

The datasets are cleaned, encoded, and normalized, and the most informative features are selected using information gain or correlation-based selection, consistent with prior work. SMOTE oversampling is applied to the training split only, to avoid data

leakage, and an ensemble or deep learning classifier is trained. Performance is reported with imbalance-aware metrics rather than raw accuracy.

Stage 2: Adversarial attack generation (Objective 2).

FGSM, BIM, DeepFool, and JSMA are used to generate adversarial versions of botnet test samples. The baseline model is evaluated on these adversarial samples, and the resulting accuracy drop establishes a measurable robustness ceiling, following the setups of Wang et al. (2021), Ibitoye et al. (2019), and Debicha et al. (2023).

Stage 3: Adversarial defense (Objective 3).

A defense adapted from multi-source adversarial training is applied to the imbalance-corrected model: the classifier is retrained on data augmented with adversarial samples from all four attack methods, and the accuracy recovered under the same attacks is measured relative to the undefended baseline.

Stage 4: Benchmarking.

Finally, the defended model is evaluated for both properties at once: class-imbalance-aware detection quality (F1-score, AUC-ROC, MCC on minority botnet classes) and adversarial robustness (accuracy retained under attack), confirming that the defense does not restore robustness at the expense of minority-class detection, or vice versa. Then, the model is benchmarked against the undefended baseline and against the accuracy figures reported in the reviewed literature, reporting the quantified improvement margin in both clean and adversarial conditions.

3.1.6 Data Analysis

Model performance is analysed using imbalance-aware metrics: F1-score, AUC-ROC, and the Matthews Correlation Coefficient (MCC), reported per class with emphasis on minority botnet classes, plus the accuracy retained under each attack. Stratified k-fold cross-validation is used to obtain stable estimates, and results across the baseline and defended conditions are compared to determine whether observed differences are consistent rather than incidental. Raw accuracy alone is deliberately avoided because it can mask poor minority-class detection in imbalanced data.

3.1.7 Ethical Considerations

Although the study involves no human participants or personal data, it generates offensive adversarial techniques and therefore carries dual-use responsibilities. All experiments are confined to public benchmark datasets in an isolated laboratory environment; no attacks are launched against live IoT devices or third-party networks. The purpose of generating adversarial samples is strictly defensive, namely to measure and improve detector robustness. Datasets are used in accordance with their licenses and are properly cited, and any released code will document its intended defensive use.

3.1.8 Limitations and Mitigation

- Oversampling reliability: SMOTE may synthesise unrepresentative minority samples; mitigated by validating with F1/AUC-ROC/MCC and comparing against a cost-sensitive alternative.
- Attack assumptions: white-box attacks may overstate vulnerability; mitigated by additionally testing black-box transferability, in line with Debicha et al. (2023, 2026).
- Computational cost: adversarial training and deep models are expensive; mitigated through feature selection, subsampling, and GPU acceleration.

- Generalisability: findings are bounded by two datasets; mitigated by using two complementary IoT datasets (Bot-IoT and N-BaIoT-2021) and reporting per-dataset results.

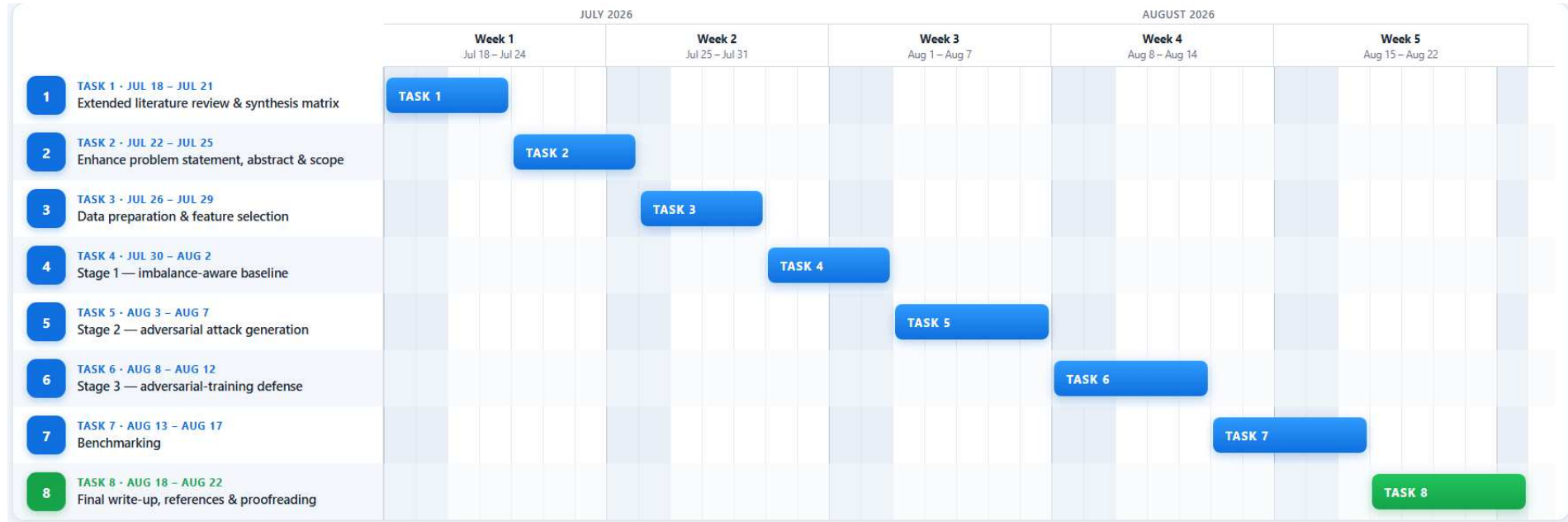
3.2 Expected Outcomes

- An empirical answer to whether class-imbalance-corrected IoT botnet detectors retain their reported accuracy under adversarial evasion.
- A quantified robustness ceiling showing how far accuracy falls under FGSM, BIM, DeepFool, and JSMA.
- An adversarially robust, imbalance-aware detector reporting both detection quality (F1/AUC-ROC/MCC) and accuracy retained under attack.
- Evidence on whether adversarial defense and imbalance correction can hold simultaneously, and a benchmarked improvement margin over the undefended baseline and prior literature.

3.3 Gantt Chart

The plan of work below shows the project's phases, their start and end points across the approximately five-week window (18 July to 22 August 2026), and the final submission milestone. Bars indicate the duration of each task, and later stages depend on the completion of earlier ones.

Figure 1. Gantt chart of the proposed plan of work.



References

- Abu Al-Haija, Q., & Al-Dala'ien, M. (2022). ELBA-IoT: An ensemble learning model for botnet attack detection in IoT networks. *Journal of Sensor and Actuator Networks*, 11(1), 18. <https://doi.org/10.3390/jsan11010018>
- Alani, M. M., Mashatan, A., & Miri, A. (2023). Adversarial explainability: Utilizing explainable machine learning in bypassing IoT botnet detection systems. *arXiv*. <https://arxiv.org/abs/2310.00070>
- Debicha, I., Cochez, B., Kenaza, T., Debatty, T., Dricot, J.-M., & Mees, W. (2023). Adv-Bot: Realistic adversarial botnet attacks against network intrusion detection systems. *Computers & Security*, 129, 103176. <https://doi.org/10.1016/j.cose.2023.103176>
- Debicha, I., Kenaza, T., Charfi, I., Mosbah, S., Sehaki, M., & Dricot, J.-M. (2026). Targeted adversarial traffic generation: Black-box approach to evade intrusion detection systems in IoT networks. *International Journal of Machine Learning and Cybernetics*, 17, 58. <https://doi.org/10.1007/s13042-025-02873-w>
- Ibitoye, O., Shafiq, O., & Matrawy, A. (2019). Analyzing adversarial attacks against deep learning for intrusion detection in IoT networks. In *2019 IEEE Global Communications Conference (GLOBECOM)*. IEEE. <https://arxiv.org/abs/1905.05137>
- Idrissi, I., Boukabous, M., Azizi, M., Moussaoui, O., & El Fadili, H. (2021). Toward a deep learning-based intrusion detection system for IoT against botnet attacks. *IAES International Journal of Artificial Intelligence*, 10(1), 110-120. <https://doi.org/10.11591/ijai.v10.i1.pp110-120>
- Khoa, T. V., Hoang, D. T., Trung, N. L., Nguyen, C. T., Quynh, T. T. T., Nguyen, D. N., Ha, N. V., & Dutkiewicz, E. (2022). Deep transfer learning: A novel collaborative learning model for cyberattack detection systems in IoT networks. *IEEE Internet of Things Journal*, 9(11), 8825-8836. <https://doi.org/10.1109/JIOT.2022.3202029>

- Nguyen, G. L., Dumba, B., Ngo, Q.-D., Le, H.-V., & Nguyen, T. N. (2022). A collaborative approach to early detection of IoT botnet. *Computers and Electrical Engineering*, 97, 107525. <https://doi.org/10.1016/j.compeleceng.2021.107525>
- Popoola, S. I., Adebisi, B., Ande, R., Hammoudeh, M., Anoh, K., & Atayero, A. A. (2021). SMOTE-DRNN: A deep learning algorithm for botnet detection in the Internet-of-Things networks. *Sensors*, 21(9), 2985. <https://doi.org/10.3390/s21092985>
- Wang, J., Pan, J., AlQerm, I., & Liu, Y. (2021). Def-IDS: An ensemble defense mechanism against adversarial attacks for deep learning-based network intrusion detection. In 2021 International Conference on Computer Communications and Networks (ICCCN). IEEE. <https://doi.org/10.1109/ICCCN52240.2021.9522215>