

Machine Learning and Deep Learning Approaches for Botnet and Cyberattack Detection in IoT Networks

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

The Internet of Things (IoT) involves the development and deployment of interconnected devices, like 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 shown the ability to compromise hundreds of thousands of devices, enabling distributed denial-of-service 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.

Several machine learning and deep learning methods have been proposed for IoT botnet detection, reporting near-perfect results: SMOTE-DRNN combines synthetic oversampling with a deep recurrent network to reach a 99.62% F1-score (Popoola et al., 2021), ELBA-IoT's ensemble of boosted decision trees achieves 100% accuracy on binary classification (Abu Al-Haija & Al-Dala'ien, 2022), and a comparative CNN model reaches 99.94% accuracy (Idrissi et al., 2021). However, these results are all reported on clean, unaltered test traffic, and only one of the six reviewed studies (Popoola et al., 2021) attempts to correct for the severe class imbalance between regular traffic and botnet traffic that characterizes real IoT networks. The other classifiers, including the CNN, do not address this imbalance, which likely inflates their apparent performance.

Separately, the only reviewed study that evaluates adversarial robustness shows that an undefended deep learning intrusion detection classifier drops from above 99% accuracy on clean traffic to between 32.8% and 43.4% accuracy under gradient-based evasion attacks (FGSM, BIM, DeepFool, JSMA), while an ensemble adversarial defense mechanism recovers accuracy to about 97.9% (Wang et al., 2021). Yet, this adversarial evaluation was conducted on the CSE-CIC-IDS2018 enterprise intrusion dataset, which contains no IoT botnet traffic and no class imbalance correction. This means that no existing study has determined whether class-imbalance-corrected IoT botnet detectors maintain their reported accuracy when exposed to realistic adversarial evasion.

This problem statement addresses that exact 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 attacks. Since class imbalance handling and adversarial robustness have so far been studied separately, one on IoT traffic without adversarial testing, the other under adversarial testing without IoT traffic or imbalance correction, it remains unknown whether imbalance correction techniques preserve or worsen a model's vulnerability to evasion, and whether adversarial defenses remain effective once class imbalance correction is applied. Resolving this question is feasible within a single study because it builds directly on established, publicly available techniques and datasets: SMOTE oversampling, ensemble and deep learning classifiers, the Bot-IoT and N-BaIoT-2021 datasets, and the FGSM/BIM/DeepFool/JSMA attack and Multi-source Adversarial Training defense methods already validated in the reviewed literature (Popoola et al., 2021; Abu Al-Haija & Al-Dala'ien, 2022; Wang et al., 2021). It does not require new infrastructure or untested methods. This research area is justified by the inadequacy of current solutions and the pressing need to determine whether IoT botnet detection systems that report near-perfect accuracy would remain trustworthy against a realistic, adaptive adversary (Wang et al., 2021).

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 at once, as identified in the problem statement. The research objectives are structured around the main goal through which the tasks are carried out and measured. Each objective progresses sequentially from building an imbalance-aware baseline, to exposing it to adversarial attack, to defending it, and finally to verifying that both class-imbalance handling and adversarial robustness hold together.

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 (Popoola et al., 2021; Abu Al-Haija & Al-Dala'ien, 2022).
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 (Wang et al., 2021).
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, relative to the undefended baseline (Wang et al., 2021).

4. To determine whether the defended model maintains its class-imbalance-aware detection quality measured via F1-score, AUC-ROC, and MCC on minority botnet classes together with its adversarial robustness, measured via 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 on standardized IoT relevant benchmark datasets.

These research objectives align with the specific challenge identified in the problem statement, addressing the previously unexamined intersection of class imbalance and adversarial robustness in IoT botnet detection). Each objective is clear, specific, and measurable against a named metric or percentage improvement, and each builds directly on the one before it, ensuring systematic, feasible progress toward the research goal within the available scope and resources. The goal of this research is to determine whether IoT botnet detection systems can be made simultaneously reliable under class imbalance and resilient under adversarial evasion, and to contribute an empirically validated answer to a question the existing literature has left open.

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>

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 (IJ-AI)*, 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. *Proceedings of the 30th International Conference on Computer Communications and Networks (ICCCN 2021)*. IEEE.

<https://doi.org/10.1109/ICCCN52240.2021.9522215>