Machine Learning Project Proposal - CSAI 801

Project Title:

Securing IoT Networks with Robust and Explainable Machine Learning

Group Number: 29

1. Motivation

The rapid expansion of Internet of Things (IoT) devices in critical infrastructure (spanning healthcare, smart grids, and industrial automation) has expanded the attack surface for malicious actors. While Intrusion Detection Systems (IDS) are essential for security, deploying complex Deep Learning models on resource-constrained IoT edge devices is often impractical due to high computational overhead and lack of interpretability.

This project focuses on **Classical Machine Learning** approaches, which offer a superior balance of efficiency and performance for real-time IoT applications. However, a key challenge remains: ensuring these models are both **robust** against diverse attack types and **transparent** in their decision-making. Our goal is to develop an optimized, explainable IDS using ensemble learning techniques, addressing the critical need for "trustworthy AI" in cybersecurity without relying on black-box neural networks.

2. Dataset

We will utilize the **CICIoT2023** dataset, a comprehensive benchmark developed by the Canadian Institute for Cybersecurity (CIC) at the University of New Brunswick.

- **URL:** https://www.unb.ca/cic/datasets/iotdataset-2023.html
- **Description:** This dataset represents a realistic IoT environment consisting of 105 devices (including cameras, sensors, and hubs) and captures 33 different types of attacks (e.g., DDoS, DoS, Reconnaissance) alongside benign traffic.
- **Suitability:** Its large scale and diverse attack vectors make it an ideal testbed for evaluating the limits of classical ML classifiers in handling high-dimensional network data.

3. Related Work

Our research builds upon recent studies in IoT security and adversarial machine learning:

- **Benchmarking:** Dadkhah et al. (2023) introduced the CICIoT2023 dataset and benchmarked various algorithms, establishing high baseline accuracy for Random Forest models [1].
- **Adversarial Risks:** Alotaibi & Rassam (2023) demonstrated that standard classifiers are highly vulnerable to adversarial evasion attacks, where subtle perturbations can deceive the model into misclassifying malicious traffic as benign [2].

Existing work often focuses on maximizing accuracy in benign environments. Our project addresses the gap identified by [2] by explicitly testing robustness and adding explainability layers to standard ensemble models.

4. Intended Experiments & Evaluation

We propose a four-phase experimental pipeline to develop a high-performance, interpretable IDS:

Phase 1: Baseline & Preprocessing

- **Objective:** Reproduce baseline results and prepare data.
- **Method:** Train standard **Decision Tree** and **Random Forest** classifiers. Implement rigorous data preprocessing (feature scaling, encoding) suitable for classical algorithms.
- **Evaluation:** Establish baseline F1-Scores and Training/Inference time.

Phase 2: Advanced Classical Models

- **Objective:** Enhance detection accuracy and efficiency.
- **Method:** Implement and fine-tune Gradient Boosting machines: **XGBoost** and **LightGBM**. We will use **Grid Search** or **Bayesian Optimization** to tune hyperparameters (learning rate, tree depth) for optimal performance.
- **Hypothesis:** Gradient Boosting methods will outperform standard Random Forest in classifying rare attack types while maintaining lower latency.

Phase 3: Robustness Analysis & Adversarial Training

- **Objective:** Evaluate and improve stability.
- **Method:** Test the models against noisy data and feature perturbations (simulating network jitter or evasion attempts). We will identify weak points and adopt an **adversarial training strategy**, retraining the model on these challenging examples to harden it against future attacks.
- **Evaluation:** Robustness score (accuracy retention under perturbation) before and after retraining.

Phase 4: Explainability & Analysis

- **Objective:** Validate feature relevance (The "White Box" approach).
- **Method:** Apply **SHAP** values to interpret the best-performing model (e.g., XGBoost). We will identify the top 5-10 features driving detection (e.g., packet header length, inter-arrival time) and verify they align with known network security principles.
- **Evaluation:** Qualitative analysis ensuring the model is not learning spurious correlations (artifacts), thereby validating its scientific reliability.

5. References

[1] S. Dadkhah et al., "CICIoT2023: A Real-Time Dataset and Benchmark for Large-Scale Attacks in IoT Environment," *Sensors*, vol. 23, no. 13, p. 5941, 2023.

[2] F. Alotaibi and M. A. Rassam, "Adversarial Machine Learning Attacks against Intrusion Detection Systems: A Survey," *Future Internet*, vol. 15, no. 2, p. 62, 2023.