IoT Network Intrusion Detection and Classification using Explainable Machine Learning Algorithms

ISE 5194 – Human-Centered Machine Learning
Project Proposal
Spring 2021

Harshil S. Patel

1. Research Question(s)

IoT networks have become an increasingly valuable target of malicious attacks due to the increased amount of valuable user data they contain. In response, network intrusion detection systems(IDS) based on machine learning algorithms/models have been developed to detect suspicious network activity where They monitor network traffic for suspicious activities and issue alerts in case of detected attack types. ML algorithm are being investigated as potential IDS frameworks as: IoT traditional network security solutions may not be directly applicable due to the differences in IoT structure and behavior, Low operating energy and minimal computational capabilities; hence, traditional security mechanism such as encryption protocols and authentication cannot be directly applied, and the lack of a single standard for IoT architecture, policies, and connectivity domains.

But current research indicates these ML-based IDSs are mainly “black boxes” where users of the systems in cybersecurity services are lacking the ability to explain why the system arrived at the classification of attack, which is important if cybersecurity and information assurance planning.

The research objective of this proposal is to apply a survey of ML algorithms that have been modified (Explainable AI (XAI)) to explain its decisions and increase transparency to solve the problem of the need of ML-based IDSs to be understood to a human analyst in the IoT cybersecurity domain, while measuring performance metrics of accuracy and false alarm rate.

2. Data and Software

UNSW-NB15 is an IoT-based network traffic data set with different categories for normal activities and malicious attack behaviors from botnets (through classification of attack type including Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode and Worms). The raw network packets of the UNSW-NB 15 dataset were created by the IXIA PerfectStorm tool in the Cyber Range Lab of the Australian Centre for Cyber Security (ACCS) for generating a hybrid of real modern normal activities and synthetic contemporary attack behaviors on IoT based networks.

UNSW-NB15 dataset has different categories for normal activities and malicious attack behaviors. UNSW-NB15 botnet datasets with IoT sensors' data can be used to obtain results that show that the proposed features have the potential characteristics of identifying and classifying normal and malicious activity for application of ML algorithms.

The details of the UNSW-NB15 dataset are published in following the papers:

Mustafa, Nour, and Jill Slay. "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)." Military Communications and Information Systems Conference (MilCIS), 2015. IEEE, 2015.

Zoghi, Zeinab, and Gursel Serpen. "UNSW-NB15 Computer Security Dataset: Analysis through Visualization." arXiv preprint arXiv:2101.05067 (2021).

The dataset can also be found at https://www.kaggle.com/mrwellsdavid/unsw-nb15/code

In terms of software used to conduct and answer the research question raised, Python will be used to develop interpretable machine learning models based on classic ML algorithms based on the UNSW-NB15 dataset to detect botnets' attacks effectively and efficiently in a IoT network traffic. The important Python packages to use and investigate to modify traditional ML algorithms for explainability will be:

1.) XAI - An eXplainability toolbox for machine learning (https://github.com/EthicalML/xai)
2.) Yellowbricks, extension of the scikit-learn library and provides some explainability tools and visualizations for machine learning models.
3.) ELI5 is a visualisation library that is useful for debugging machine learning models and explaining the predictions they have produced.
4.) LIME (local interpretable model-agnostic explanations) is a package for explaining the predictions made by machine learning algorithms.

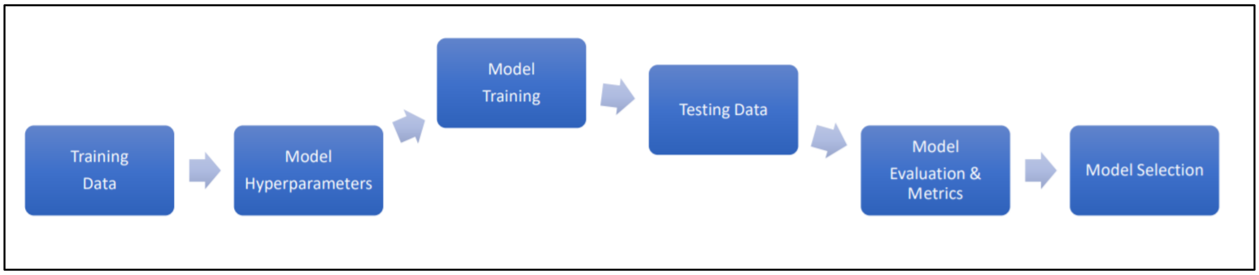
3. ML Algorithm(s)

The survey of ML algorithms used to create modified explainable ML models using the UNSW-NB15 will be:

1.) Decision Tree Classifier
2.) Random Forests Classifier
3.) Multi-Layer Perceptron Classifier
4.) XGBoost Classifier

Additionally, determining which explainable model should be used to classify category attacks and network intrusions on IoT networks based on transparency(explainability) and performance will be evaluated.

The overall process proposed is given by the flowchart below:



4. Planned Team Contributions

I will be conducting the project individually based on previous work for Dr. Allen’s class where a binary classification project for applying multi-linear and logistic regression and seeing results on computer security data was conducted.