**Project Information**

Project Type: Individual

Student Name: Raine Morrigan

Mentor Name: Dr. Hasan

Working Research Title: Exploring Coordinated Attacks on ML-Enabled IoT Devices.

**1. Problem Statement, Hypothesis, and Research Questions**

**Problem:** Detecting remaining vulnerabilities in IoT systems trained with Machine Learning Models.

**Hypothesis:** Coordinated adversarial attacks targeting TinyML-enabled IoT devices can achieve higher success rates and lower detectability than isolated attacks by exploiting the resource constraints of edge devices

**Research Questions:**

* How does the limited computational capacity of TinyML devices affect the detectability and mitigation of coordinated adversarial attacks compared to traditional ML systems?
* What metrics best capture the trade-offs between attack impact, resource consumption, and detection probability in coordinated TinyML attacks?

**2. Literature Review**

**MCUNet: Tiny Deep Learning on IoT Devices**

**Ji Lin, Wei-Ming Chen, Yujun Lin, John Cohn, Chuang Gan, and Song Han** | Publication Year: 2020-2022

This project creates a framework called MCUNet that combines the efficient neural architecture of TinyNAS and the lightweight inference engine of TinyEngine to allow for near ImageNet level inference on microcontrollers. They were able to have face vs facemask detection and wake words that acted on par with or faster than the SOTA. Memory Bottlenecking remained a limitation with this version. With the second version they were able to begin addressing the bottleneck issues caused by the limited memory available as well as allowing object detection on IOT devices which opens more applications for the model. They did this by moving to a patch-by-patch image processing plan. To combat overlapping regions, they move some of the computational complexity to the later layers that can rely on smaller patches. They also automate finding the optimal architecture by employing a Neural Architecture Search. The newest model is version 3 where they propose to turn the IOTs into devices capable of lifelong, privatized learning by employing Quantized Training Optimization for the gradient optimization issue and Sparce Update to address memory footprint. They also created a Tiny Training Engine which prunes the computation graph and pre-computes complex calculations at compile time while arranging data for minimal memory usage. At runtime, with Sparce Update, they only calculate what’s needed and reuses memory buffers when possible. Additionally, with Quantized operations, all computations are only 8-bit. I believe this could potentially be the model I intend to build upon in my research. It seems to be performing at or better than SOTA and with it’s lifelong learning capabilities, it could allow pathways compromise information in FL systems.

**FairGuard: A Fairness Attack and Defense Framework in Federated Learning**

**Xinyi Sheng , Zhengjie Yang , and Wei Bao** | Publication Year: 2025

Algorithmic fairness is the belief that machine learning algorithms shouldn’t show bias or discrimination to specific demographics. This paper states that the existing algorithmic fairness algorithms in Federated Learning (FL) are primarily focused on mitigating the inherit bias found in the training data and leave fairness largely undefended. They propose 4 innovative adversarial attack methods that compromise the fairness of FL Models: label-flipping, attribute-flipping, hybrid-flipping, and double-flipping. These each have an individual way to allow the Hacker to utilize the injected Malicious Users to manipulate the model. Label-flipping flipped the Y attribute. Attribute-flipping flipped the X attribute. Hybrid-flipping combined both using a ratio to determine to apply label or attribute-flipping. Double-flipping was their attempt to integrate their fairness attacks with the existing untargeted label-flipping accuracy attack to create a new strategy. Though they found the randomness of Double-flipping to be counterintuitive to the goal of the fairness attack and often didn’t skew the right way. After showing the existing flaws in FL models, the researchers then proposed their novel security framework, FairGuard. FairGuard defends against Malicious Clients using a Dual-Inference Algorithm. The algorithm generates random samples with undefined sensitive attributes, then creates an entry with a positive attribute and an entry with the negative attribute. This is to avoid accessing real data and violating FL. A "suspicious score" is calculated based on prediction conflicts when the clients model is run through both demographic groups. Then, they use clustering to identify malicious clients based on the score. Additionally, they attempt to account for the inherent non-IID nature of FL by asking each client to run local fairness checks called Local Debiasing. This is done on a volunteer basis as a trap for malicious users. Either they apply the centralized debiasing algorithm and make their attack less effective or avoid doing so and make their attack more obvious. This work identifies a critical gap in security while offering effective defenses against the attacks identified. The limitations remain that FairGuard is only effective against data tin tables and not images. Additionally, it is only set for classification tasks and requires clustering assumptions for malicious client detection. I plan to expand upon this work by adapting the attacks to a facial recognition visual model to poison the data of my TinyML FL model and gain unauthorized access to the model.

**Federated Learning Minimal Model Replacement Attack Using Optimal Transport: An Attacker Perspective**

**K. Naveen Kumar , C. Krishna Mohan , and Linga Reddy Cenkeramaddi** | Publication Year: 2025

This paper addresses the known existing vulnerabilities in Federated Learning’s (FL) decentralized nature. While certain adversarial tactics like data poisoning and model poisoning attacks are well documented, the literature often focuses only on the impact of the attack. The author’s take on the Attacker’s perspective to consider not only the impact, but also an attack’s budget and visibility. They show the significance of these factors by proposing a novel FL minimal Model Replacement (FL-MMR) attack that they intend to perform better than the existing total neuron replacement and randomly selected neuron replacement that only partially address attack visibility and cost. Model replacement attacks This is considered a three-fold minimal attack: (i) By focusing on weights and not biases the model is dataset-independent and offers more privacy, (ii) the zeta variable is used to determine how many neurons to replace per layer, and (iii) using a replacement map to determine the optimal number of parameters per layer per CLP. FL-MMR uses optimal transport to find the optimal neurons to replace after candidate neurons are randomly selected. The neurons are selected from both the benign model and the poisoned surrogate model in a way that maintains balance by creating a OT map to determine where the “mass” should be transported. This map is what marks the difference between FL-MMR approach and other unoptimized random neuron replacement attacks. The evaluation showed a low visibility, low budget, and high impact attack over various tests, but some limitations remain. In future work, they believe a dynamic psi (replacement map) would increase efficiency when calculating CLP rounds and they consider the task agnostic but have only tested on image classification. They intend to extend to object detection, which I believe would lead well into my project as I attempt to break the face classification model we plan to make. I think more than anything, the example of how to have an Adversarial mindset and to think like a hacker is helpful. Keeping attacks low cost and low visibility will allow them to be more easily adapted to TinyML models and maintain effectiveness. This study will be one that I return to as a guide on the Adversarial Mindset.

**Model Poisoning Attack Against Neural Network Interpreters in IoT Devices**

**Xianglong Zhang, Feng Li, Huanle Zhang, Haoxin Zhang, Zhijian Huang, Lisheng Fan, Xiuzhen Cheng, and Pengfei Hu |** Publication Year: 2024

This paper highlights the security flaws in Neural Network Interpreters by creating an attack that bypasses the inherent security of the constrained storage and low computational power. This attack is the first to not need extra datasets to perform data poisoning attacks. The adversary searches the last convolutional layer for information on the features that are most strongly tied to the prediction result and the interpretation result. By manipulating these results, the attacker can poison the model and conceal their identity. During the model poisoning training phase, the attacker manipulates the values of the sensitive features identified in the previous step to be able to calculate positive and negative impact parameters that will be further used via backpropagation to maximize the attacker’s goal while minimizing the likelihood of getting caught. They target top-k fooling, center-mass fooling and location fooling. Further limitations of the attack remain that it was more effective on CAM-based interpreters than gradient based interpreters. I could build upon this project by adapting it to smart home scenarios.

**Securing IoT-Enabled Smart Cities and Detecting Cyber Attacks in Smart Homes for a Greener Future**

**L. Zhou, A. Gaurav, R. W. Attar, V. Arya, A. Alhomoud and K. T. Chui** | Publication Year: 2025

This paper addresses a variety of cybersecurity challenges seen in smart homes in the effort to protect future smart cities. Modern smart homes consist of entire networks of IoT devices that span through the home, opening them to significant security vulnerabilities like DDoS attacks, botnet intrusions, and network scanning attacks. The authors present the model that attackers would use for targeting smart home: Reconnaissance, Exploitation of weak points, and execution of attack, most likely mirai of Gafgyt based DDoS attacks. The authors propose a CNN based deep learning model that would analyze the network traffic of the router. The model preprocesses date to remove missing and irrelevant values and uses feature selection to improve model accuracy. Then, following the rest of the process allows for classification between malicious and benign behavior. The results showed high promise when tesing a Kaggle dataset of diverse attack types with a 99% success rate, but MiraiUDP remains to be a difficult foe to identify. The main take aways are the CNN doesn’t need manual feature engineering, but it struggles against mirai UDP. My project will build upon this project by utilizing the attack model and database for future whitebox attacking.

**3. High-Level Proposed Approach**

* **AI models**
  + **TinyML**
  + **CNN**
* **Tools/Libraries**
  + **TensorFlow**
  + **MobileNetV2**
* **Cybersecurity Context**
  + **Threat Simulation**

**4. Experimental Design**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Reference** | **Dataset** | **Metric** | **Remarks** |
| **Enhanced CNN** |  |  |  |  |
| **MobileNetV2** |  |  |  |  |

**5. Reproduce SOTA**

**Harvard University Fundamentals of TinyML Course**

**Vijay Janapa Reddi, and Laurence Moroney |** Available April 2025 – April 2026

<https://pll.harvard.edu/course/fundamentals-tinyml>

This project covers several useful topics from the basics of machine learning to responsible AI creation and design. Building lesson over lesson, I recreated and finished CNNs and TinyMLs used for image identification. This would be particularly useful for building an understanding of what I was attempting to do with my own model. Following along the course teaches how to fit lines with linear regression and the concept of gradient descent before giving a coding assignment to break into Neural networks. Eventually, you are taught how to construct a basic CNN utilizing keras and tensorflow to create an image classifier. By working along with the program, I was able to better understand the task at hand of facial verification to protect IoT devices in a smart home capacity. I adopted the code to use a small human face dataset instead of the bean image classifier. The limitations remain that this is a basic visual classifier and I would need to test SOTA attacks against a more robust model, so I plan to build upon this by using transfer learning from the SOTA MobileNetV2 model and using the model I adapted from this course as a comparison.

**6. References**

* Lin, Ji, et al. "On-device training under 256kb memory." *Advances in Neural Information Processing Systems* 35 (2022): 22941-22954.
* X. Sheng, Z. Yang and W. Bao, "FairGuard: A Fairness Attack and Defense Framework in Federated Learning," in IEEE Transactions on Dependable and Secure Computing, vol. 22, no. 3, pp. 2519-2532, May-June 2025
* K. Naveen Kumar, C. Krishna Mohan and L. Reddy Cenkeramaddi, "Federated Learning Minimal Model Replacement Attack Using Optimal Transport: An Attacker Perspective," in IEEE Transactions on Information Forensics and Security, vol. 20, pp. 478-487, 2025
* Harvard University Fundamentals of TinyML Course, Vijay Janapa Reddi, and Laurence Moroney | Available April 2025 – April 2026 <https://pll.harvard.edu/course/fundamentals-tinyml>
* L. Zhou, A. Gaurav, R. W. Attar, V. Arya, A. Alhomoud and K. T. Chui, "Securing IoT-Enabled Smart Cities and Detecting Cyber Attacks in Smart Homes for a Greener Future," in IEEE Internet of Things Magazine, doi: 10.1109/MIOT.2025.3575924.keywords: {Smart homes;Internet of Things;Accuracy;Feature extraction;Security;Machine learning;Support vector machines;Smart cities;Scalability;Authentication;IoT;Security;DDoS Attack},
* X. Zhang *et al*., "Model Poisoning Attack Against Neural Network Interpreters in IoT Devices," in *IEEE Transactions on Mobile Computing*, vol. 24, no. 3, pp. 1715-1730, March 2025, doi: 10.1109/TMC.2024.3486218. keywords: {Computational modeling;Data models;Internet of Things;Training;Predictive models;Neural networks;Monitoring;Mobile computing;Dogs;Biomedical monitoring;Interpreter;IoT devices;model poisoning attack;neural networks},