Cross-Model Universal Adversarial Perturbations in Medical Imaging

Ganza Belise Aloysie Isingizwe
The Dark Side of AI

Motivation

Can adversarial perturbations, crafted without access to any model's weights, fool multiple unseen classifiers trained on different data and architectures?

Project Overview

Federated learning requires shared model architectures across local and global models, increasing vulnerability to adversarial attacks crafted within the same environment. This project explores how such attacks transfer across models with shared or differing backbones. Architectural similarity appears to increase risk, while diversity may offer some protection, though it's limited by the constraints of federated learning.

Experimental Setup

I simulated a federated learning environment using:

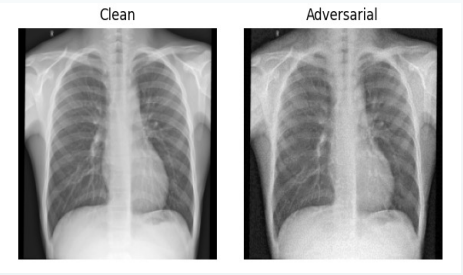
1. **Two Datasets:**
   - NIH Chest X-ray → split and used to train Sites A and B
   - RSNA Pneumonia Challenge → split and used to train Sites C and D
2. **Three Architectures:**
   - ResNet50
   - DenseNet121
   - InceptionV3
3. **Model Breakdown (n=15):**
   - 12 local models (3 architectures × 4 sites A–D)
   - 3 global models (1 per architecture) trained via Federated averaging over 3 rounds
   - Local models were fine-tuned post-aggregation

FGSM Attack

Adversarial attacks were crafted using FGSM (Fast Gradient Sign Method) on correctly classified pneumonia images

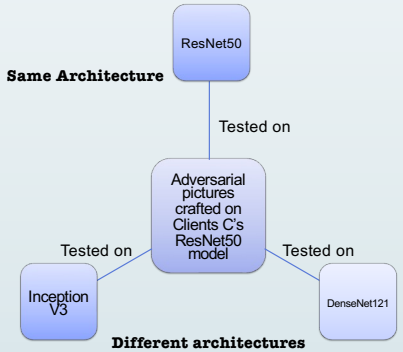
FGSM Formula: $x_{adv} = x + \varepsilon \cdot sign(\nabla_x \mathcal{L}(f(x), y))$

where x is the input image, f (x) is the model, L is the loss function, and ε = 0.02



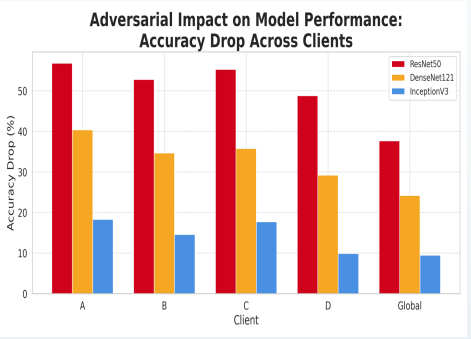
Attack Evaluation

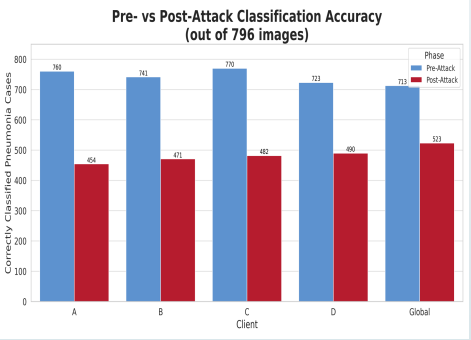
Transferability = $\frac{N_{misclassified\ after\ attack}}{N_{correct\ before\ attack}}$



Results

- ResNet50 models showed the highest transferability (mean drop: 54.6%, σ ≈ 4.4%).
- DenseNet121 models had moderate vulnerability (mean drop: 35.2%, σ ≈ 6.8%).
- InceptionV3 was the most robust (mean drop: 17.9%, σ ≈ 5.5%).
- Global models were consistently more resistant to adversarial transfer.





Conclusion

- Federated learning's reliance on shared architectures may introduce systemic risk.
- Although architectural diversity may reduce this risk, typical federated learning setups such as those using Federated Averaging (FedAvg) require shared model architectures, which can increase systemic vulnerability across sites.

Future work

- Test with other attack methods
- Explore real-world federated deployments and heterogeneity
- Evaluate the impact of defense strategies like adversarial training or model personalization
- Expand evaluation to include more model architectures
- Incorporate additional and more diverse medical imaging datasets during training

Glossary

Federated learning: A machine learning approach that allows multiple entities to collaboratively train a model while keeping their data on their own devices, ensuring privacy and security.

Architectural Similarity: The degree to which different models share structural components (e.g., ResNet50 and DenseNet121 have similar convolutional backbones).

FGSM (Fast Gradient Sign Method): An attack that adds tiny, human-imperceptible changes to an image, guided by the model's gradient to cause misclassification.