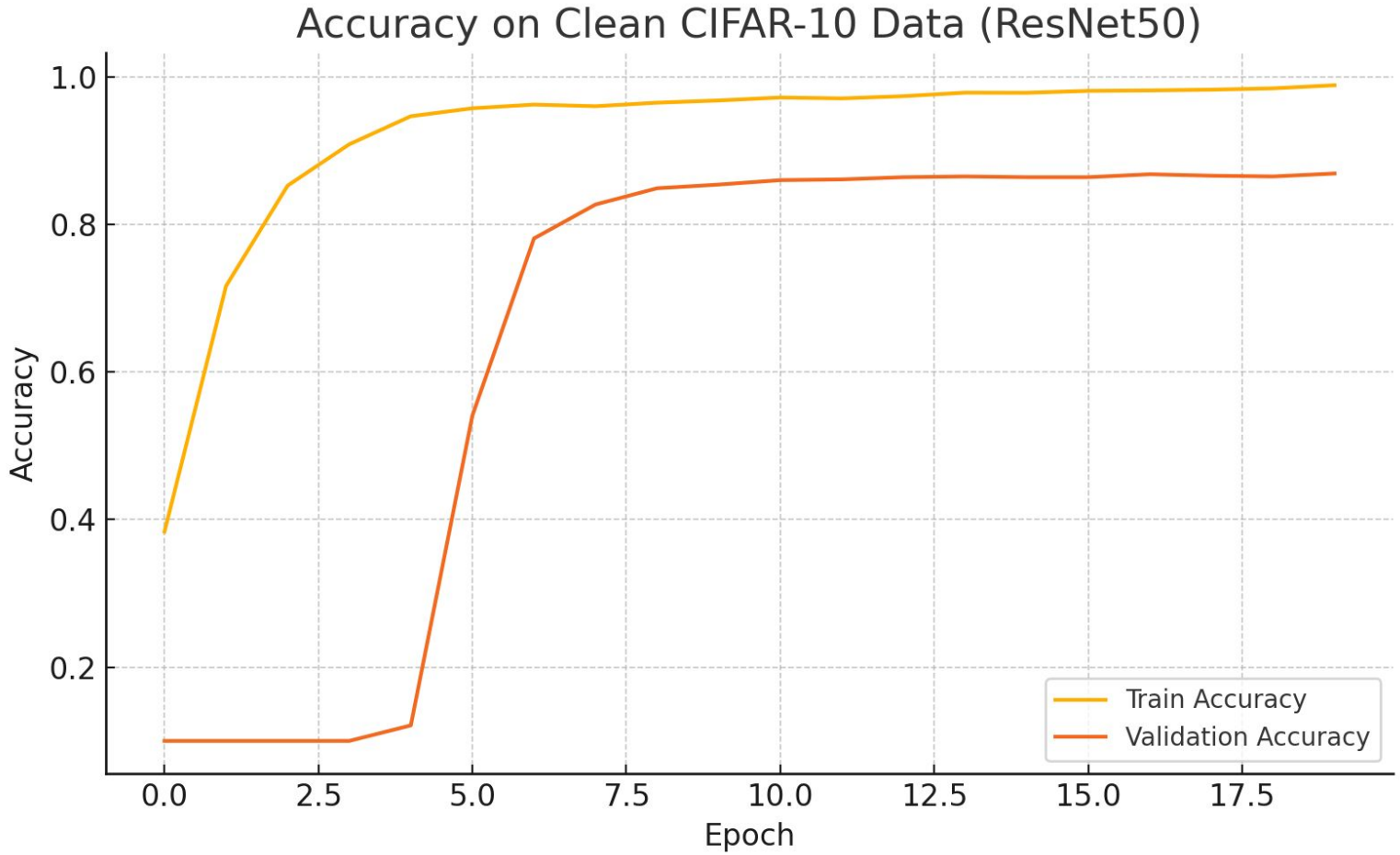
Carnegie Mellon University

How Robust Are Pretrained Image Models to Adversarial Attacks?

Chijioke Ugwuanyi

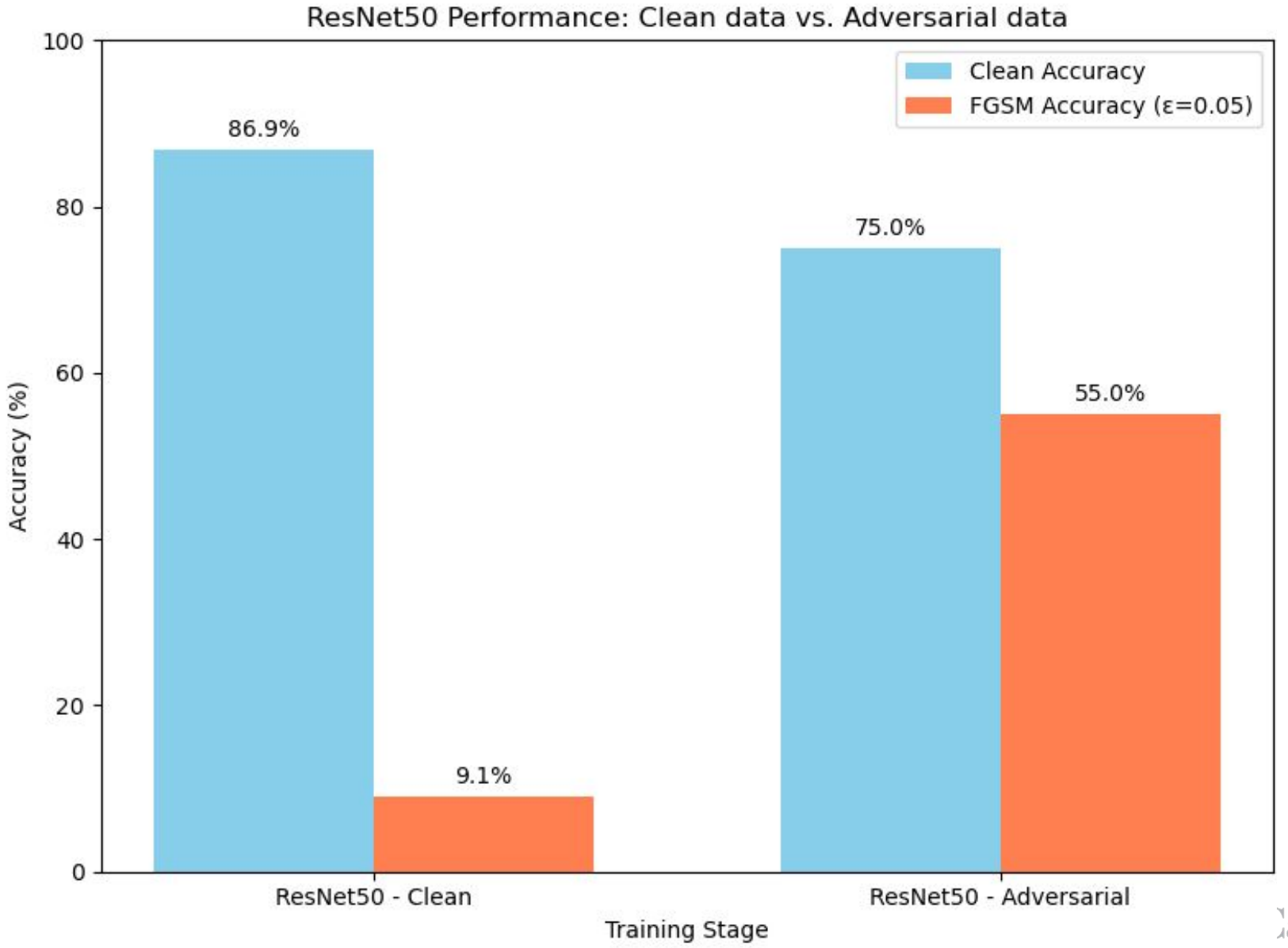
22/04/2025

Insights (ResNet50 - Baseline training)



Clean data accuracy: 86.90%

Insights (ResNet50 - Mixed training)

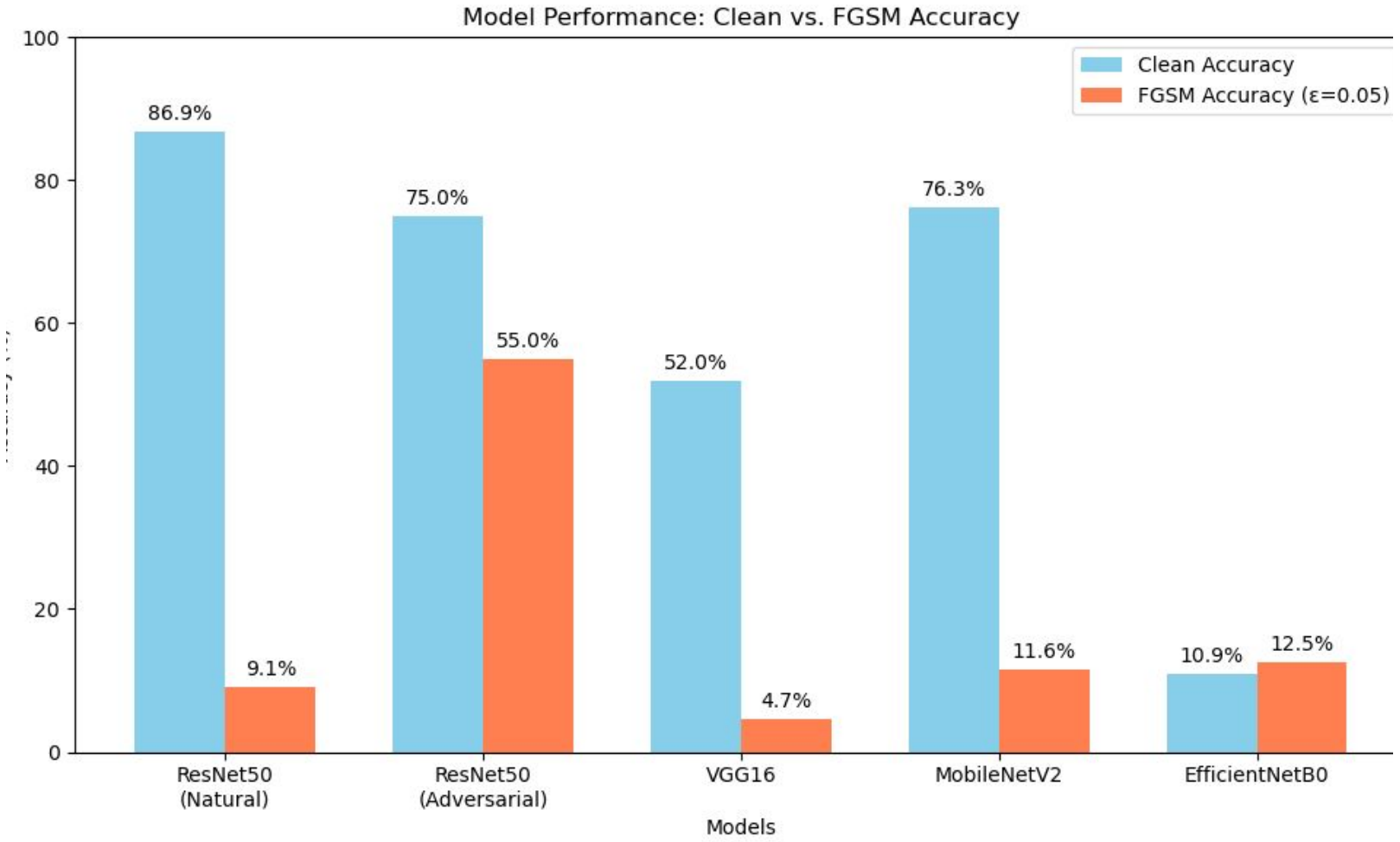


Epsilon: 0.05

Before adversarial training, FGSM accuracy: 9.06%.
Fooling rate: 89%.

After adversarial training FGSM accuracy: 55.01%

Insights (ResNet50, VGG16, MobileNetV2, EfficientNetB0)



EfficientNetB0 seem to have the highest robustness to adversarial attack.

MobileNetV2 is potentially the most robust given EfficientNetB0's poor clean accuracy.

Conclusion

Pretrained models are not robust by default. Defenses like adversarial training are essential to ensure its robustness.

Thank you!