Adversarial Robustness of Traffic Sign Recognition:
Evaluating FGSM Attack on Models

DuckAI

AAI 595-B

Jingxuan Zhu, Yasin Hasanpoor, Shotitouch Tuangcharoentip
AAI 595 Applied Machine Learning

Introduction

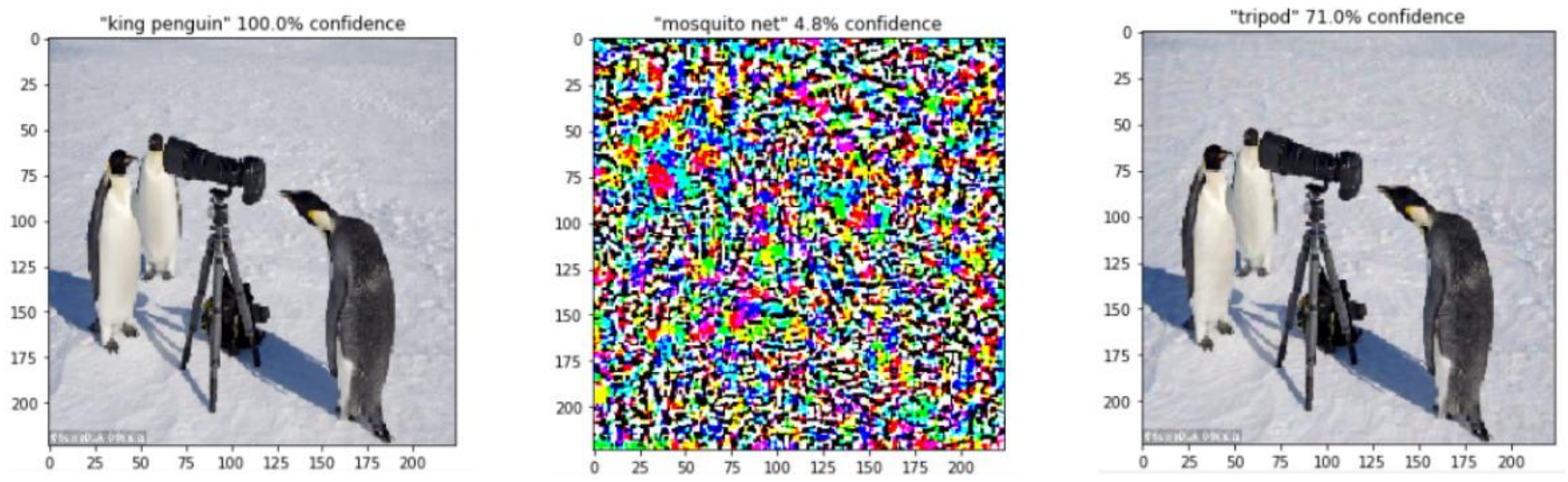
- Traffic sign recognition is critical for autonomous driving and public road safety.
- Machine learning (ML) models, while accurate, are vulnerable to adversarial attacks like Fast Gradient Sign Method(FGSM).
- **Core question**: Are modern ML models more robust against FGSM attacks without explicit defenses compared to older architectures?

FGSM Attack

(Fast Gradient Sign Method)

Perturbs input images to trick models
ε values used: 0.001, 0.005, 0.01, 0.02, 0.03, 0.05, 0.1

$$x \quad + \ \varepsilon \ \text{sign}(\nabla_x L(\theta, x, y)) \quad = \quad x_{\text{adv}}$$


"king penguin" 100.0% confidence
"mosquito net" 4.8% confidence
"tripod" 71.0% confidence

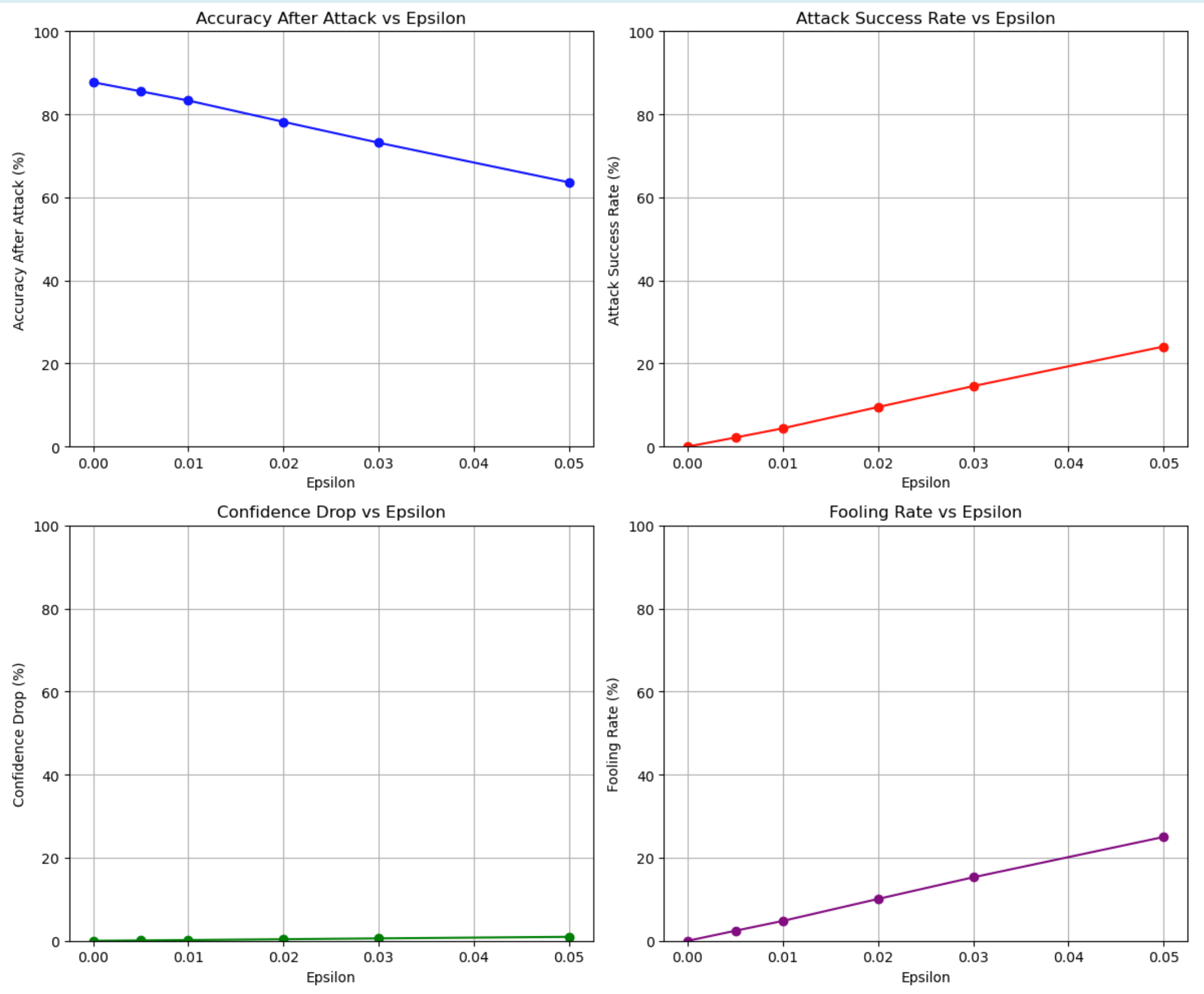
Data

- **Dataset**: German Traffic Sign Recognition Benchmark (GTSRB)
  - 43 traffic sign classes
  - High-quality labeled images
- **Preprocessing**:
  - Greyscale conversion
  - Data normalization
  - Split: 80% training / 20% testing
- GTSRB is widely used in both classification and adversarial robustness research.

AlexNet (Old)

Optimizer: SGD | Epochs: 50

FGSM attack result



- Worst Accuracy on Clean Data
  - Accuracy: 87.72%
- **Best Robustness to FGSM** ☆
  - At **ε** = 0.01 Accuracy drops to 83.32%
  - At **ε** = 0.05 Accuracy drops to 63.63%

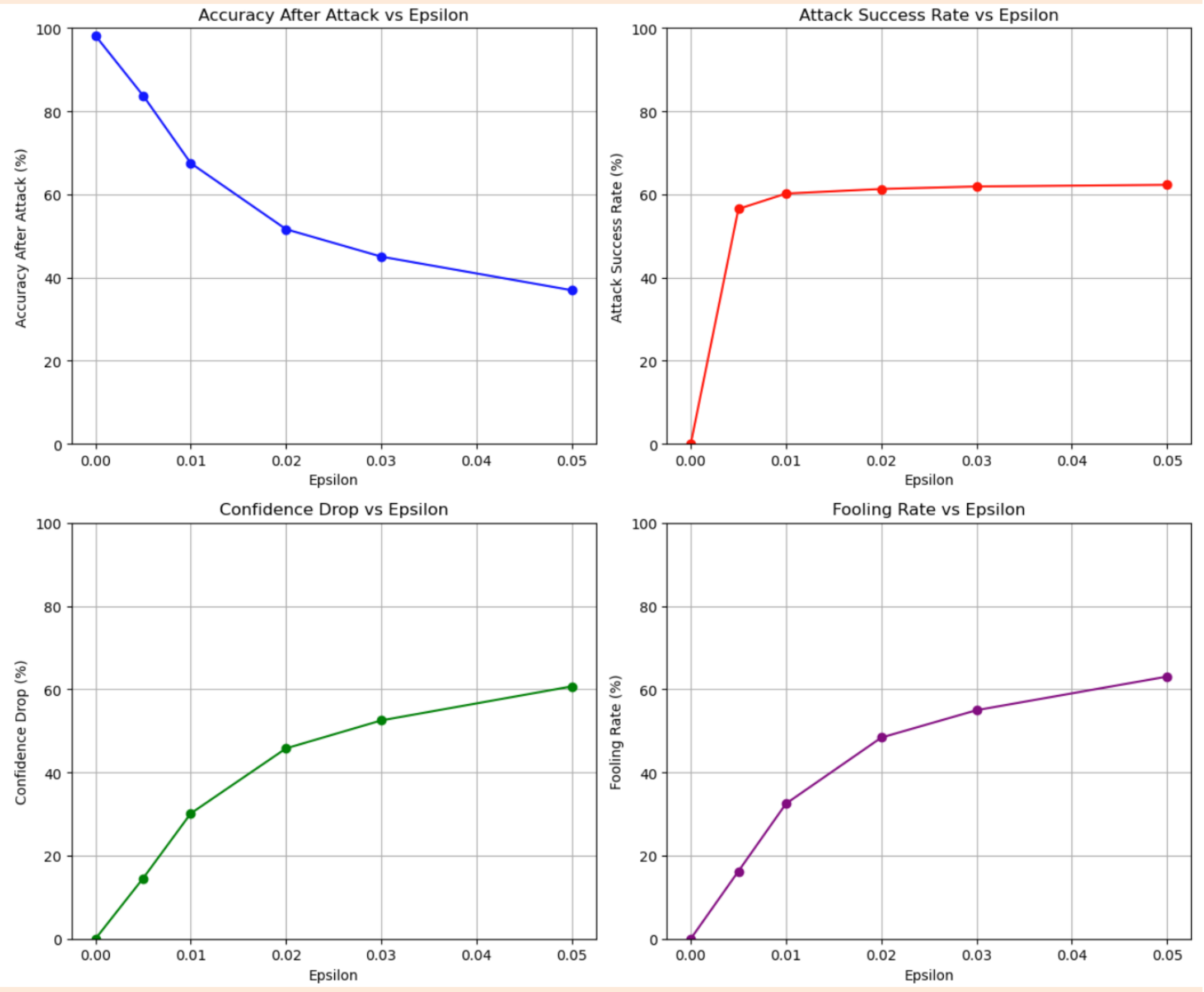
Why?
- Deep-architecture model
- Outdated design

- Quite heavy
  - Take long to train,
  - High memory usage
  - Slow inference

CNN-VGG based (Mid)

Optimizer: Adam | Epochs: 50

FGSM attack result



- Good Accuracy on Clean Data
  - Accuracy: 98.15%
- Bad Robustness to FGSM
  - At **ε** = 0.01 Accuracy drops to 67.50%
  - At **ε** = 0.05 Accuracy drops to 36.95%

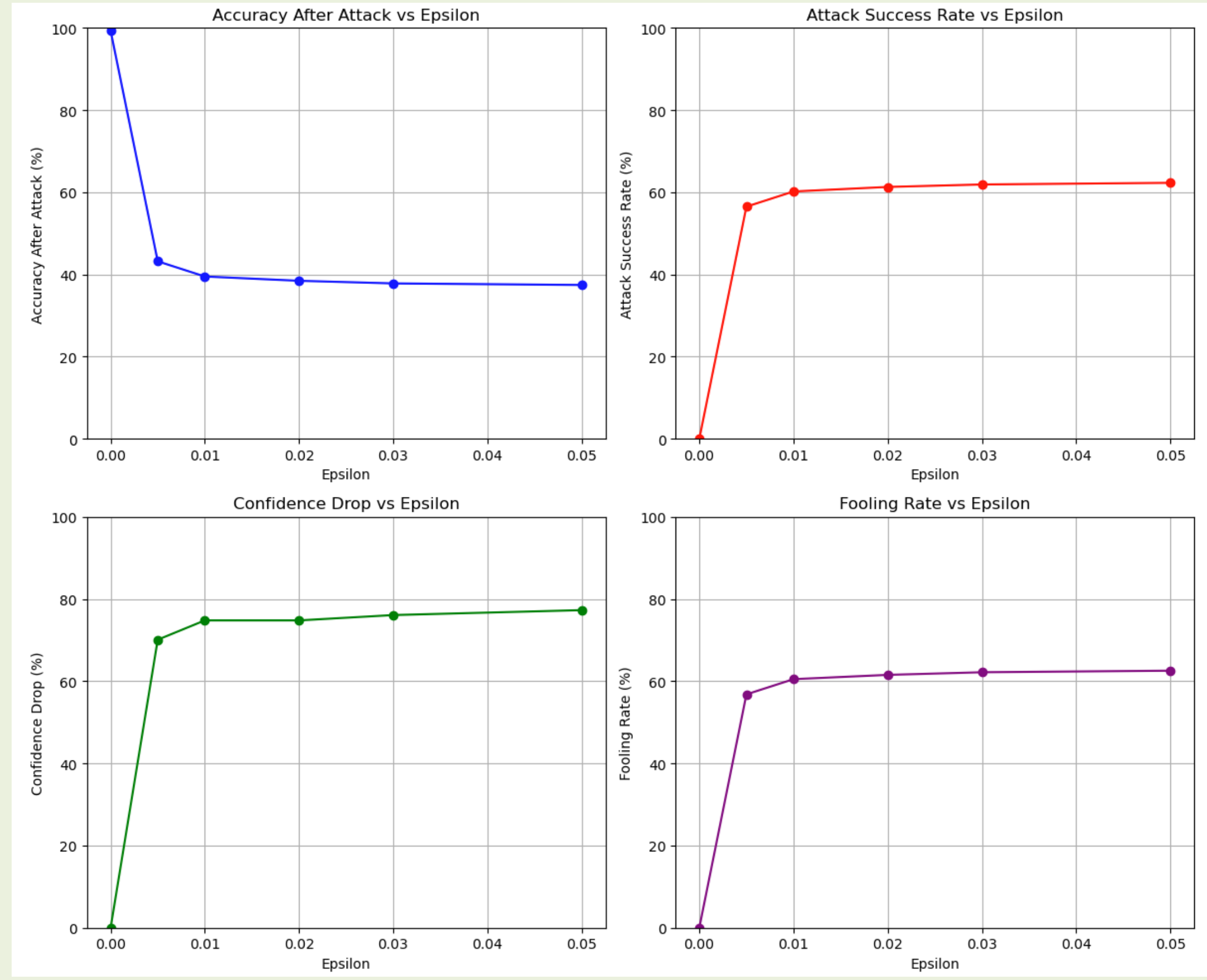
Why?
- Shallow-architecture Model
- Compact design

- **Lightweight** ☆
  - Less time to train,
  - Low memory usage
  - Fast inference

EfficientNet B0(Modern)

Optimizer: Adam | Pretrained B0 | Epochs: 10

FGSM attack result



- **Best Accuracy on Clean Data** ☆
  - Accuracy: 99.30%
- Bad Robustness to FGSM
  - At **ε** = 0.01 Accuracy drops to 43.20%
  - At **ε** = 0.05 Accuracy drops to 37.44%
  - But low drop rate after

Why?
- Very Deep architecture model
- Modern design

- Heavy
  - Moderate training time
  - High memory usage
  - Slowest inference
- **Transfer learning -> Sensitive to FGSM**
- Accuray 88% after Adversarial Training

STEVENS
INSTITUTE OF TECHNOLOGY

I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations (ICLR)*, 2015.

H. Singh, "Acc 97.4% sign classification (alexnet/vggnet/cnn)." `https://www.kaggle.com/code/harbhajansingh21/acc-97-4-sign-classification-alexnet-vggnet-cnn/`, 2022. Accessed: 2025-05-14.