SynthShield-AI: Build → Break → Improve Framework

1. Introduction

SynthShield-AI is a deep learning framework designed to detect AI-generated images, evaluate vulnerabilities using adversarial attacks, and improve robustness via adversarial defense. The CIFAKE dataset (REAL vs FAKE images) was used for experimentation.

2. Phase 1 — Build (Baseline Detector)

A pretrained ResNet-18 model was fine-tuned for binary classification. Stage 1: Only classifier head trained (LR=1e-3). Stage 2: Full fine-tuning (LR=1e-5). Model evaluated using accuracy, classification report, and confusion matrix.

3. Phase 2 — Break (Adversarial Attack Analysis)

A targeted L∞ Projected Gradient Descent (PGD) attack was implemented. Parameters: epsilon=0.03, alpha=0.003, steps=20. High-confidence FAKE predictions were flipped to REAL with visually imperceptible perturbations. Grad-CAM and FFT analyses showed attention shifts and increased high-frequency components.

4. Phase 3 — Improve (Adversarial Training)

PGD-based adversarial training was applied. For each batch, adversarial examples were generated and combined with clean images for training. Parameters: epsilon=0.03, alpha=0.003, steps=5. Evaluation included both clean and adversarial accuracy metrics.

5. Key Insights

• High-performing detectors are vulnerable to small structured perturbations. • PGD effectively manipulates model predictions. • Adversarial noise affects spatial attention (Grad-CAM) and frequency domain (FFT). • Adversarial training improves robustness with a small trade-off in clean accuracy.

6. Conclusion

SynthShield-AI demonstrates a complete robustness evaluation pipeline: baseline training, adversarial breaking, interpretability analysis, frequency-domain inspection, and robust retraining. The Build → Break → Improve framework provides a systematic approach for strengthening AI-based media verification systems.