SynthShield-AI: Build → Break → Improve Framework for Robust AI-Generated Image Detection

1. Introduction

SynthShield-AI is a deep learning pipeline designed to detect AI-generated (fake) images, evaluate model vulnerabilities through adversarial attacks, and improve robustness using defense strategies. The project follows a structured Build → Break → Improve methodology to systematically analyze and strengthen image detection systems.

2. Phase 1 — Build (Baseline Model)

A pretrained ResNet18 model was fine-tuned for binary classification (FAKE vs REAL) using the CIFAKE dataset. The model achieved high accuracy on clean images and was evaluated using accuracy metrics and confusion matrices. This phase established a strong baseline detector.

3. Phase 2 — Break (Adversarial Attack)

A targeted iterative FGSM attack was implemented to flip high-confidence FAKE predictions into REAL. Small, bounded perturbations were applied iteratively while tracking confidence degradation. Results showed that visually indistinguishable perturbations could successfully alter model predictions, demonstrating vulnerability to adversarial manipulation.

4. Phase 3 — Improve (Defense & Robustness)

Adversarial training was introduced to enhance model robustness. The model was retrained using adversarial samples and compared against the baseline. Grad-CAM visualizations were applied to analyze attention regions before and after defense. The defended model showed improved stability and resistance to adversarial attacks.

5. Conclusion

SynthShield-AI demonstrates that high-performing AI detectors can be vulnerable to adversarial perturbations, but structured robustness techniques significantly improve resilience. This framework provides a practical approach for evaluating and strengthening AI-based media verification systems.