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

1. Introduction

SynthShield-AI is a deep learning framework designed to detect AI-generated (synthetic) images, evaluate model vulnerabilities using adversarial attacks, and improve robustness through adversarial defense mechanisms.

The system follows a structured Build → Break → Improve methodology to systematically analyze and strengthen image classification models against adversarial manipulation.

The CIFAKE dataset (Real vs AI-generated images) was used for experimentation.

2. Phase 1 — Build (Baseline Detector)

Model Architecture

- Backbone: ResNet-18 (pretrained on ImageNet)
- Modified final fully connected layer for binary classification (FAKE vs REAL)

Training Strategy

Two-stage fine-tuning was performed:

1. Stage 1 — Classifier Head Training

    - All convolutional layers frozen
    - Only final FC layer trained
    - Learning rate: 1e-3

2. Stage 2 — Full Fine-Tuning

    - All layers unfrozen
    - Learning rate: 1e-5
    - End-to-end fine-tuning

Evaluation

The model was evaluated on a held-out test set using:

- Accuracy
- Classification report
- Confusion matrix

This established a strong baseline detector for clean images.

3. Phase 2 — Break (Adversarial Vulnerability Analysis)

To evaluate robustness, a targeted Projected Gradient Descent (PGD) attack was implemented.

Attack Configuration

- Norm constraint: L_∞
- Epsilon (ε): 0.03
- Step size (α): 0.003
- Iterations: 20
- Target: Flip high-confidence FAKE → REAL

Observations

1. High-confidence FAKE predictions were successfully flipped to REAL.
2. Perturbations were visually indistinguishable.
3. Model confidence toward the target class increased iteratively.
4. The attack remained effective even after subtle Gaussian blur refinement.

Visualization & Analysis

1. Confidence Growth Plot

Tracked the increase in REAL class confidence during attack iterations.

2. Perturbation Heatmap

Visualized pixel-level adversarial noise.

3. Grad-CAM Analysis

Grad-CAM was applied:

- Before attack (FAKE class)
- After attack (REAL class)

Observation:

- Attention shifted significantly after adversarial manipulation.
- The model’s focus changed despite minimal perceptual differences.

4. Frequency Domain Analysis (FFT)
FFT magnitude spectrum was computed for:

- Original image
- Adversarial image

Observation:

- Increased high-frequency components
- Structured frequency artifacts introduced by PGD

This confirms adversarial perturbations alter frequency characteristics.

4. Phase 3 — Improve (Adversarial Defense & Robust Training)

To improve robustness, PGD-based adversarial training was applied.

Defense Strategy
For each batch:

1. Generate adversarial samples using PGD.
2. Combine clean + adversarial images.
3. Train model on both simultaneously.

PGD Training Parameters

- Epsilon: 0.03
- Step size: 0.003
- Steps: 5
- Combined clean and adversarial loss

Evaluation Metrics

Comparison:
Before robust training: High clean accuracy

After robust fine-tuning:
Slight trade-off in clean accuracy

This demonstrates the robustness–accuracy trade-off common in adversarial training.

5. Key Insights

- High-performing detectors are vulnerable to small structured perturbations.
- Targeted PGD can reliably flip predictions without perceptual changes.

- Adversarial perturbations alter both spatial attention (Grad-CAM) and frequency spectrum (FFT).
- Adversarial training significantly improves robustness.
- There exists a trade-off between clean performance and adversarial resilience.

6. Conclusion

SynthShield-AI demonstrates a complete robustness evaluation pipeline:

- Baseline detector construction
- Adversarial vulnerability assessment
- Interpretability analysis
- Frequency-domain investigation
- Robust adversarial fine-tuning

The Build → Break → Improve framework provides a practical and systematic approach to strengthening AI-based synthetic media detection systems.