# Project 2: Improving Robustness of Deepfake Detectors through Gradient Regularization

**Course of Computer Vision - 2025**


| | |
|---|---|
| Di Santo Francesca | 1998148 |
| Navarra Arianna | 2195207 |
| Tiberia Valentina | 2027897 |

# Outline

- ❑ Problem Statement
- ❑ State of the Art
- ❑ Dataset
- ❑Proposed Method
- ❑ Experimental Setup
- ❑ Model Evaluation
- ❑ Conclusions and Future Work
- ❑ References

# Problem Statement

## ➢ Introduction

Deepfake generation techniques have rapidly improved, making it increasingly difficult to distinguish between real and manipulated faces.

This poses a serious threat in many domains, such as:


Politics

Cybersecuruty


Social Media

## ➢ Deepfake detector problems

### ❑ Poor generalization

Perform well on known manipulation method but struggle with **unseen fake generation techniques**

### ❑ Vulnerable to Adversarial Attacks

Small imperceptible perturbations can cause **misclassification**, and thus **lack of robustness** against adversarial attacks compromise reliability in real-world scenarios

### ❑ Overfitting to Dataset Bias

Models often rely on visual clues that are specific to the dataset used for training , so they **perform poorly** when tested on new datasets or in real-world conditions

## ➢ Project Goal

Improve the **robustness** and **generalization** of deepfake detectors by applying **gradient regularization** on shallow feature statistics.

# State of the Art

- **Gradient regularization** improves the robustness of deepfake detectors by reducing their sensitivity to superficial texture cues, thus enhancing generalization to unseen manipulations
  *(Guan et al., 2024 – IEEE TIFS)*
- **EfficientNet** provides an excellent balance between accuracy and computational efficiency, making it ideal for large-scale deepfake detection tasks
  *(Tan & Le, 2019 – ICML)*
- The **DFFD** dataset offers a broad spectrum of synthetic face manipulations, enabling consistent evaluation across various deepfake detection methods
  *(Dang et al., 2020 – CVPR)*
- Recent evaluations show that many models struggle under adversarial attacks, highlighting the need to design detectors that remain reliable even when facing intentionally manipulated inputs
  *(Abbasi et al., 2024 – Applied Sciences)*

# Dataset

## ➢ Dataset Description

**DFFD (Diverse Fake Face Dataset)**

It contains:

- Real images from the FFHQ dataset
- Fake images generated by multiple methods: FaceApp, StyleGAN, PG-GAN, StarGAN, etc.

# ➢ Dataset Split & Augmentation

The dataset is splitted into:

- **Training set**: ~10k real, ~56k fake (handled with class-weighted loss)
- **Validation set**: for tuning hyperparameters
- **Test set**: for final evaluation

**Data Augmentation (training):**

- Random Crop
- Horizontal Flip
- Color Jitter
- Random Rotation

Handling Class Imbalance

Used **weighted cross-entropy loss**: CrossEntropyLoss(weight=class_weights)

# Proposed Method

## ➢Two Models, Two Training Strategies

- ❑ Baseline Model
  - Built on **EfficientNet-B0**, pretrained on ImageNet
  - Final layer replaced with a binary classifier: *real* vs *fake*
  - Trained using **class-weighted cross-entropy loss** to address class imbalance
  - Focus on **high accuracy** on known manipulations
- ❑ Regularized Model
  - Based on the same EfficientNet-B0 architecture.
  - Trained with **gradient regularization** to reduce overfitting to texture patterns.
  - Enhances generalization and robustness to **unseen manipulations** and **attacks**.

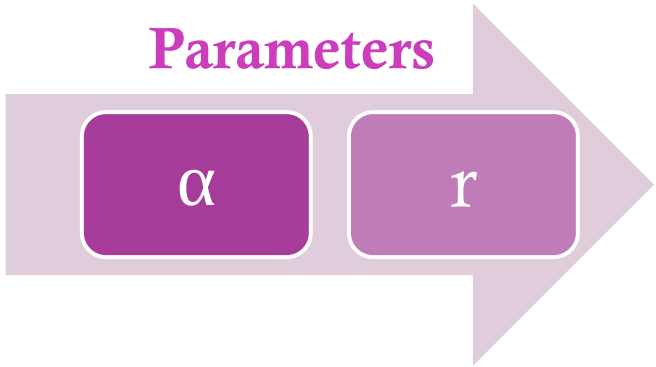
# ➢ Improving Robustness: Regularization

## Gradient Regularization:

- Acts on **shallow feature statistics** (μ and σ) from early convolutional layers
- Uses a **Perturbation Injection Module (PIM)**:
  i. First pass: clean forward propagation
  ii. Second pass: shallow features are slightly perturbed
- Final loss combines both:

$$\mathcal{L}_{\text{total}} = (1-\alpha)\mathcal{L}_{\text{clean}} + \alpha\mathcal{L}_{\text{perturbed}}$$

PIM Transformation: $x_{\text{recon}} = x_{\text{norm}} \cdot (\sigma + \Delta\sigma) + (\mu + \Delta\mu)$




α → Balances accuracy on clean inputs vs. robustness to perturbed inputs

r → Controls the strength of the noise added to the shallow feature statistics simulating invariance to small variations in low-level texture patterns

# Experimental Setup

## ➢Training Configuration

### ○ Ablation study

•We systematically varied the **regularization weight α** and the **perturbation strength r** to understand their effect on performance.

•Increasing r too much harms performance, confirming the importance of controlled perturbation.

•The study helped **identify the best hyperparameter configuration** to improve generalization without overfitting.

- Optimizer: Adam → Learning rate: 3e-5
- Batch Size: 4
- Epochs: 10
- Loss Function: Class-weighted Cross-entropy
- Early Stopping: based on validation performance
- Gradient Regularization:
  - α = 0.5
  - r = 0.05
- Reproducibility: Fixed random seed

# Model Evaluation

## ➢Metrics

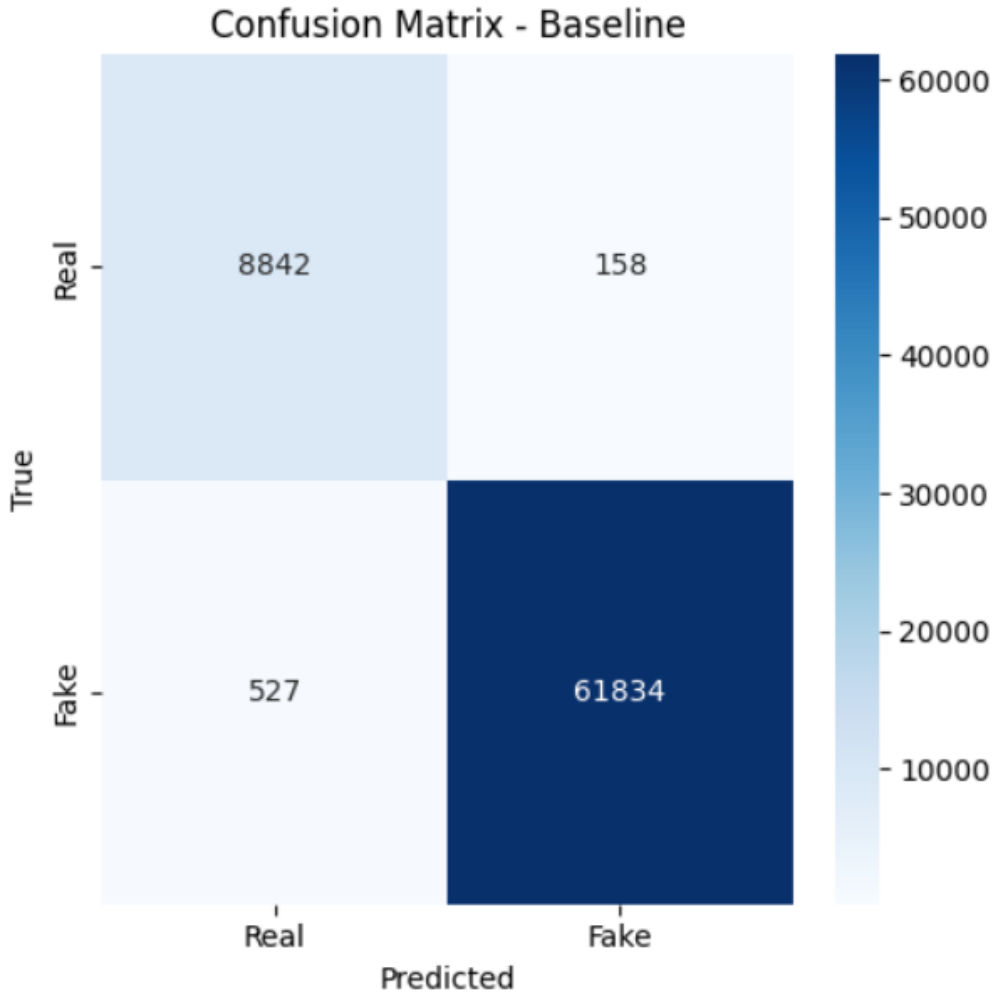
**Accuracy**
→ Measures the overall percentage of correct predictions out of all samples

**Macro F1 Score**
→ Harmonic mean of precision and recall, averaged across both classes (real and fake). Especially useful for imbalanced datasets

**AUC (Area Under the Curve)**
→ Represents the area under the ROC curve, indicating the model's ability to distinguish between real and fake samples

**IINC** (*Interpretability-Informed Consistency*)
→ Quantifies the stability of Grad-CAM attention maps when input are perturbed (for example by adversarial attacks)

# ➢ Baseline Model Performance (No Regularization)




**Overall Results** (Test Set):
- **Accuracy:** 99.04%
- **F1 Score:** 97.86%
- Dataset size: 17,841 samples

The low false positive rate and the moderate false negative rate confirm that the **baseline overfits** to seen textures and performs best on **known manipulation types**, but this **strong performance does not generalize** to adversarial attacks

## ➢ Adversarial Attacks

- The **Fast Gradient Sign Method (FGSM):**

  Single-step attack that perturbs the input image in the direction of the gradient sign of the loss with respect to the input.

- **Projected Gradient Descent (PGD):**

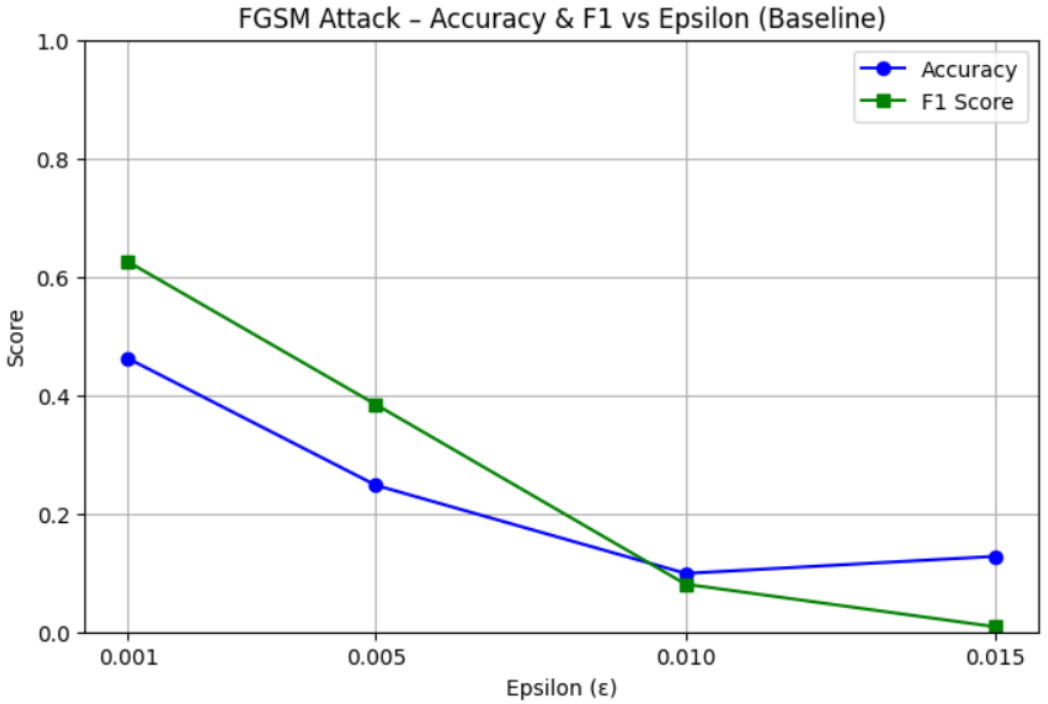
  Iterative extension of FGSM. It applies multiple small perturbations, and after each step, the perturbed image is projected back into an ε-ball around the original input.

- **Momentum Iterative FGSM (MIFGSM):**

  Introduces a momentum term to the iterative process, accumulating gradients across steps to stabilize the direction of the perturbation.

- FGSM Attack (Baseline)

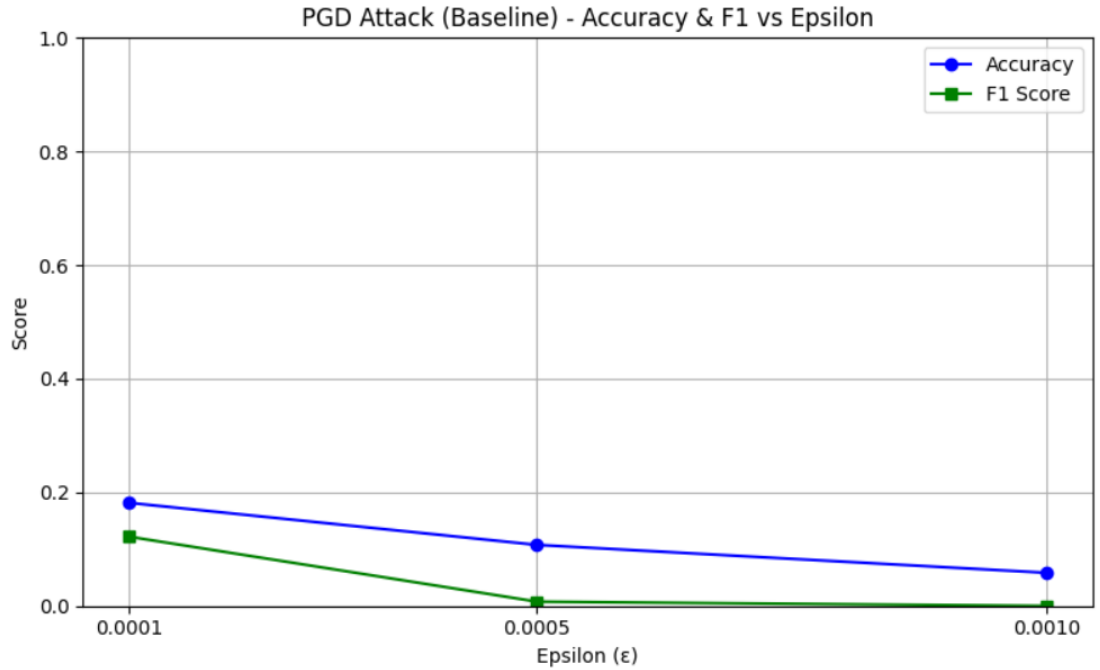
| Epsilon (ε) | Accuracy | F1 Score |
|---|---|---|
| 0.001 | 46.2% | 62.5% |
| 0.005 | 24.8% | 38.4% |
| 0.010 | 9.9% | 8.1% |
| 0.015 | 12.8% | 0.9% |




- The model fails to maintain correct predictions under perturbations that are visually imperceptible to the human eye.
- The F1 Score drops close to zero, indicating a severe inability to distinguish between real and fake classes when under attack.

- PGD Attacks (Baseline)

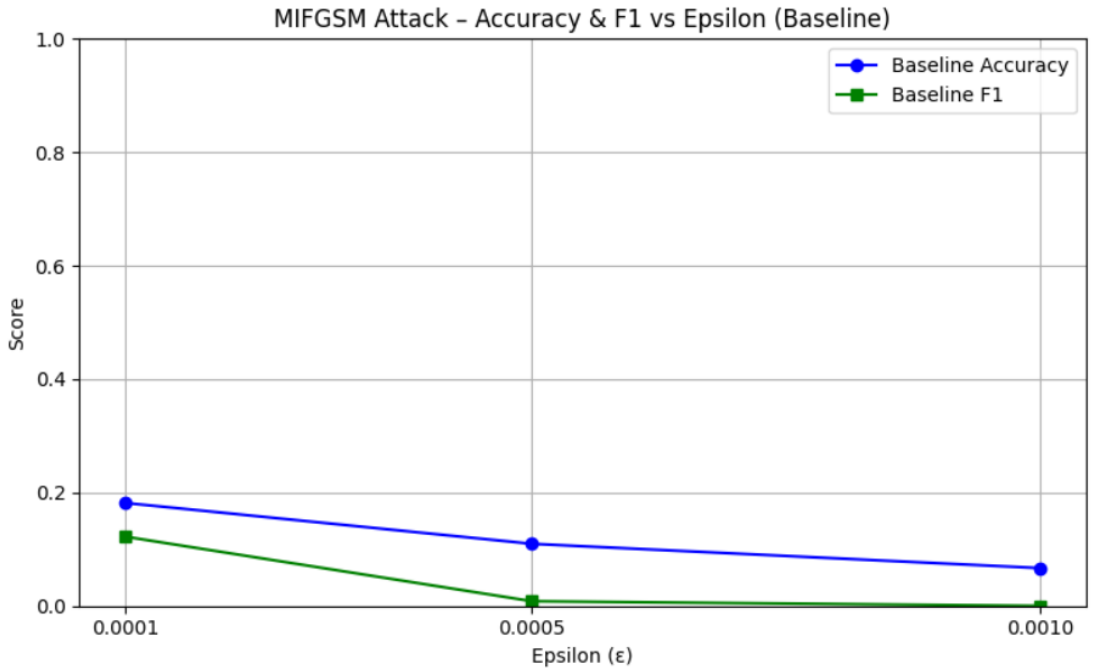
| Epsilon (ε) | Accuracy | F1 Score |
|---|---|---|
| 0.0001 | 18.17% | 12.22% |
| 0.0005 | 10.76% | 0.77% |
| 0.0010 | 5.85% | 0.01% |




The model is highly **non-robust** under iterative attacks like PGD. Even imperceptible perturbations lead to **complete failure in classification**, confirming its over-reliance on fragile features.

- MIFGSM Attacks (Baseline)

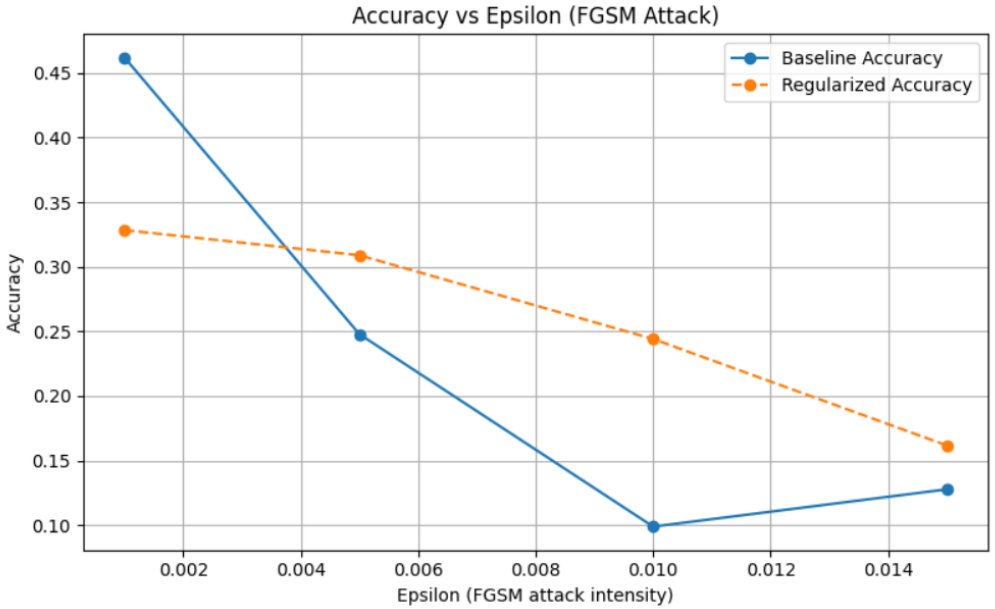
| Epsilon (ε) | Accuracy | F1 Score |
|---|---|---|
| 0.0001 | 18.18% | 12.24% |
| 0.0005 | 10.97% | 0.85% |
| 0.0010 | 6.69% | 0.02% |

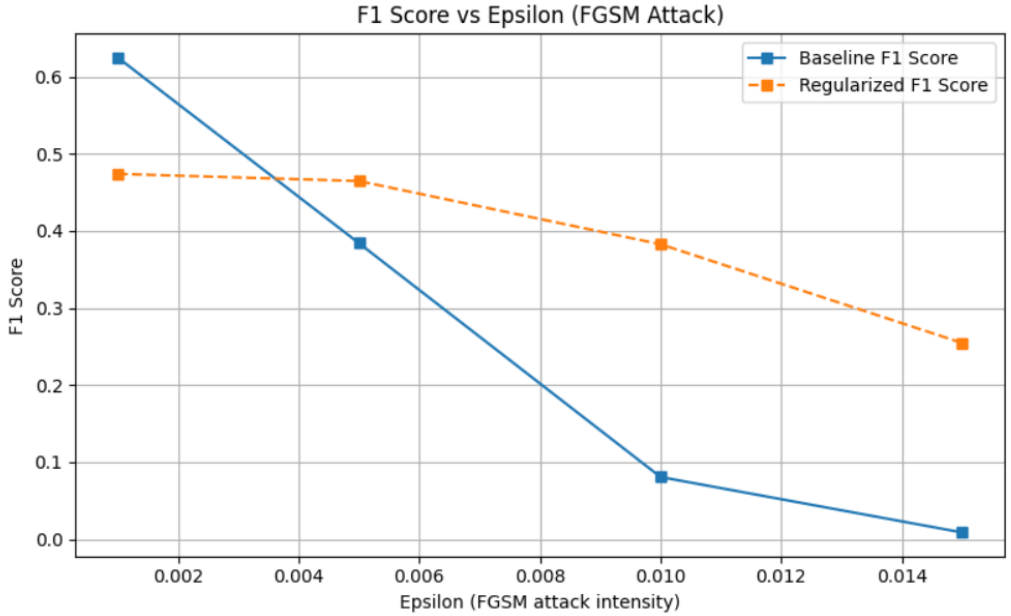



The baseline model demonstrates **poor robustness** under MIFGSM attacks. These results confirm that the baseline model lacks resilience to adversarial examples—even when these perturbations are visually imperceptible.

- FGSM Comparison – Tabular and Graphical

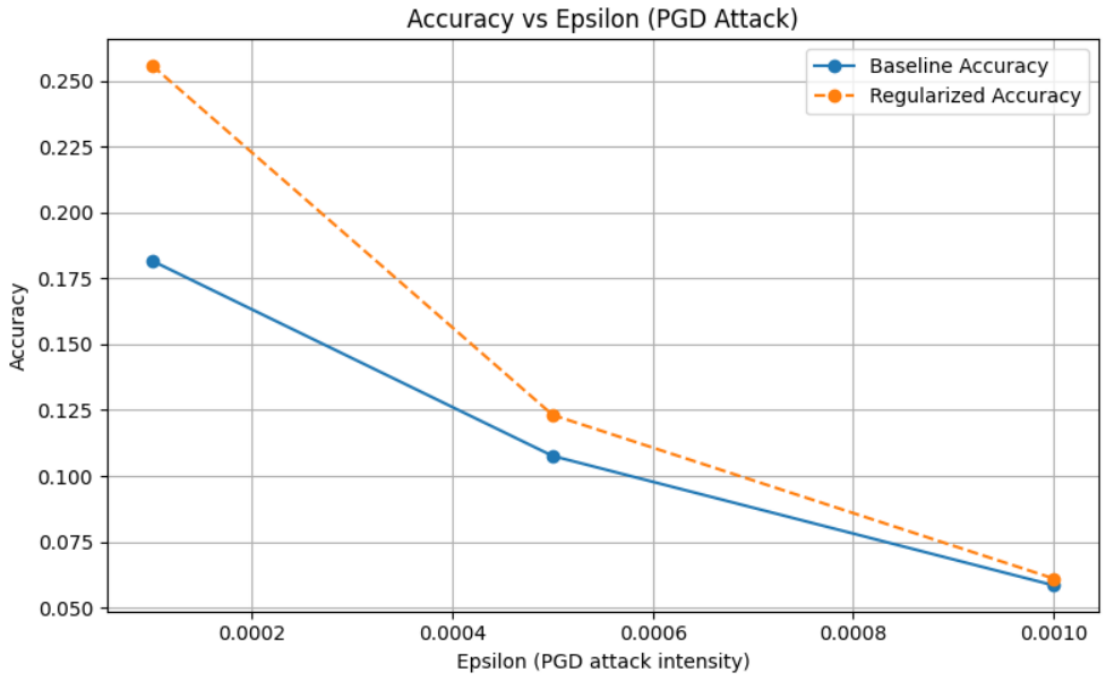
| Epsilon (ε) | Accuracy (Baseline) | Accuracy (Regularized) | F1 Score (Baseline) | F1 Score (Regularized) |
|---|---|---|---|---|
| 0.001 | 46.2% | 32.84% | 62.5% | 47.41% |
| 0.005 | 24.8% | 30.89% | 38.4% | 46.48% |
| 0.010 | 9.9% | 24.41% | 8.1% | 38.28% |
| 0.015 | 12.8% | 16.17% | 0.9% | 25.47% |

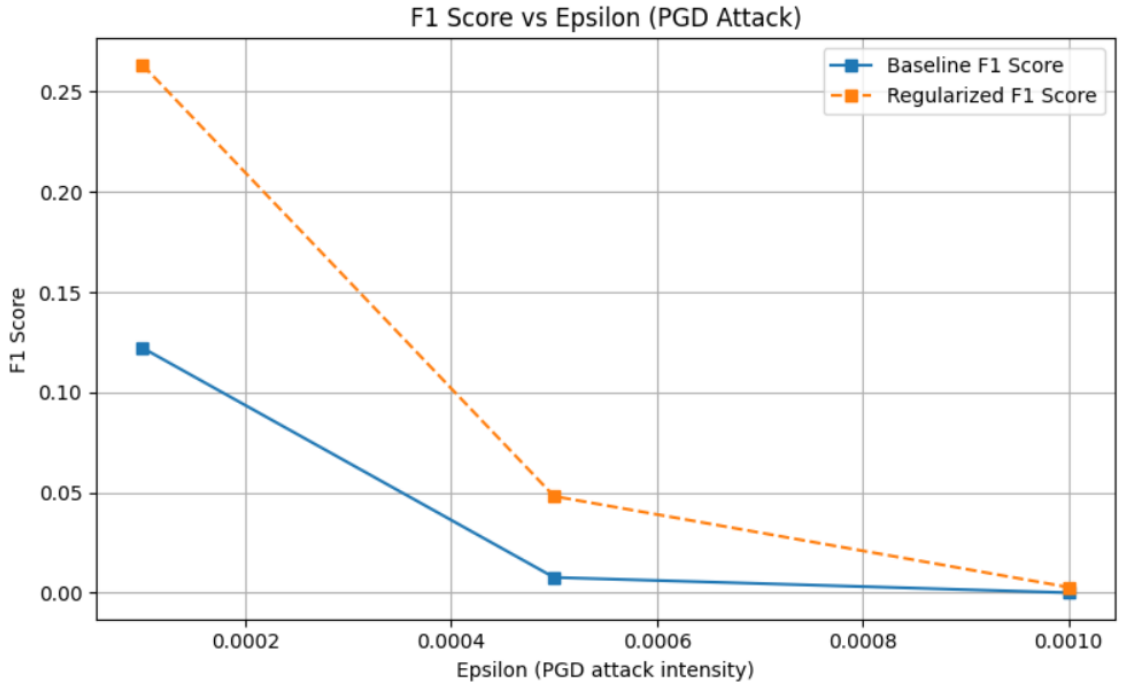


- PGD Comparison – Tabular and Graphical

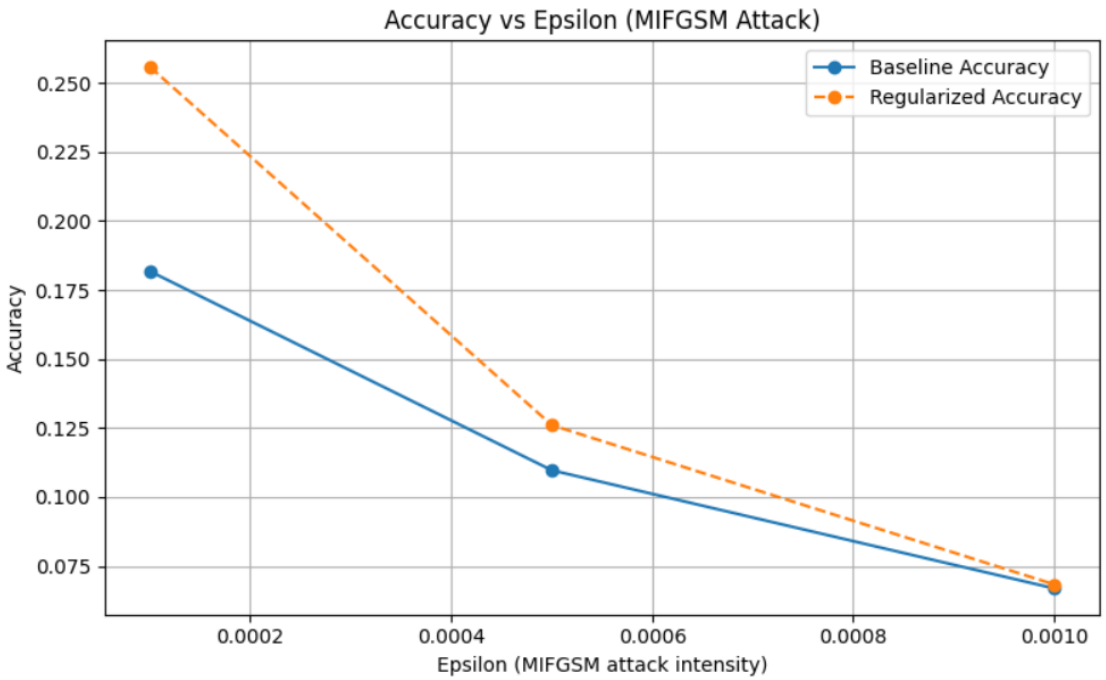
| Epsilon (ε) | Accuracy (Baseline) | Accuracy (Regularized) | F1 Score (Baseline) | F1 Score (Regularized) |
|---|---|---|---|---|
| 0.0001 | 18.17% | 25.59% | 12.22% | 26.34% |
| 0.0005 | 10.76% | 12.31% | 0.77% | 4.82% |
| 0.0010 | 5.85% | 6.11% | 0.01% | 0.28% |

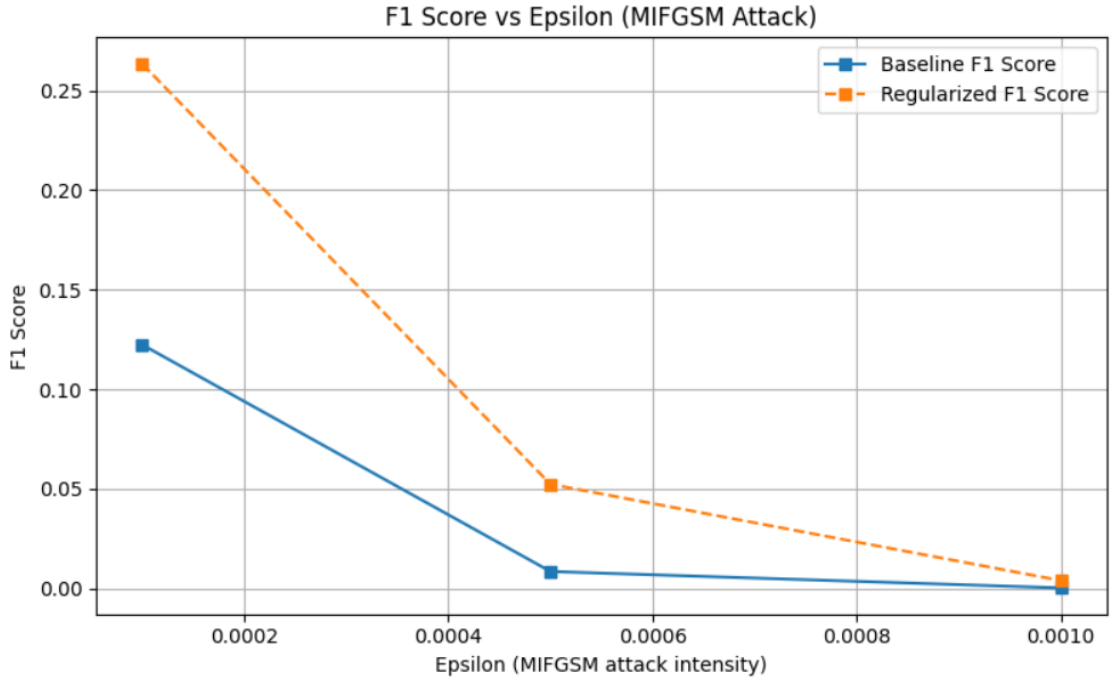


- MIFGSM Comparison – Tabular and Graphical

| Epsilon (ε) | Accuracy (Baseline) | Accuracy (Regularized) | F1 Score (Baseline) | F1 Score (Regularized) |
|---|---|---|---|---|
| 0.0001 | 18.18% | 25.59% | 12.24% | 26.35% |
| 0.0005 | 10.97% | 12.60% | 0.85% | 5.23% |
| 0.0010 | 6.69% | 6.83% | 0.02% | 0.39% |



- AUC Comparison

**FGSM**

| Normalized AUC Accuracy (Baseline) | Normalized AUC Accuracy (Regularized) | Normalized AUC F1 Score (Baseline) | Normalized AUC F1 Score (Regularized) |
|---|---|---|---|
| 0.1902 | 0.2448 | 0.2270 | 0.3727 |

**PGD**

| Normalized AUC Accuracy (Baseline) | Normalized AUC Accuracy (Regularized) | Normalized AUC F1 Score (Baseline) | Normalized AUC F1 Score (Regularized) |
|---|---|---|---|
| 0.0994 | 0.1219 | 0.0279 | 0.0751 |

**MIFGSM**

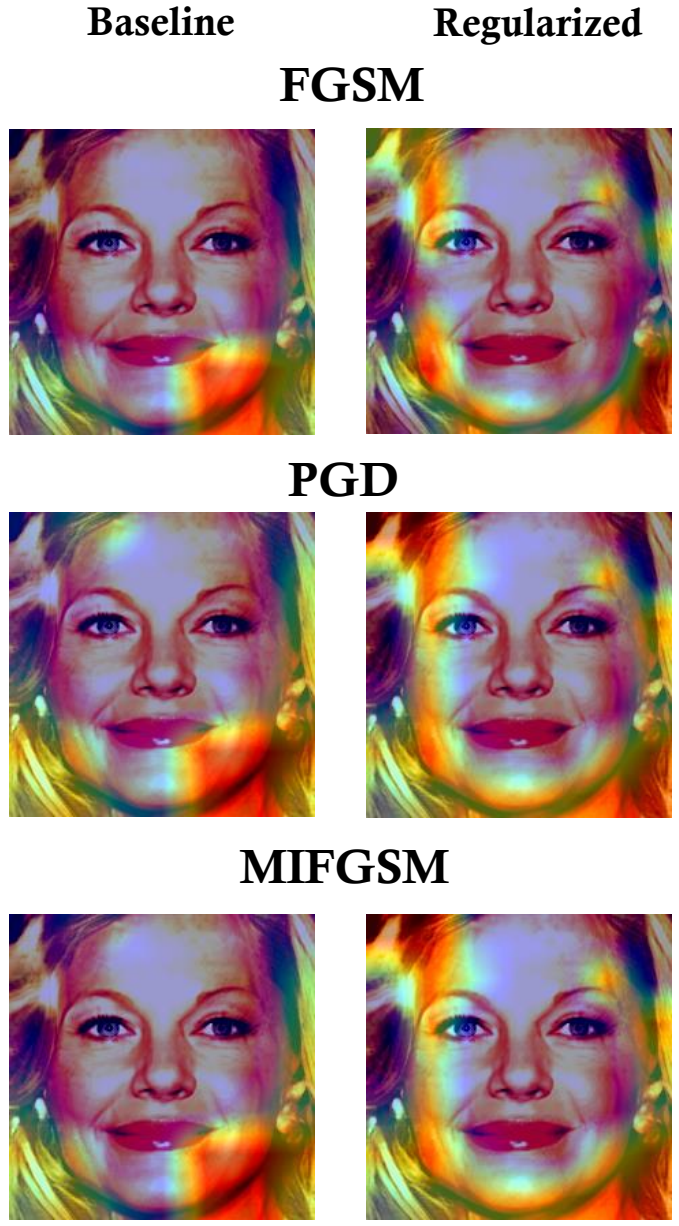
| Normalized AUC Accuracy (Baseline) | Normalized AUC Accuracy (Regularized) | Normalized AUC F1 Score (Baseline) | Normalized AUC F1 Score (Regularized) |
|---|---|---|---|
| 0.1024 | 0.1250 | 0.0284 | 0.0772 |

- Grad-CAM Visual Analysis – Attention under Attack

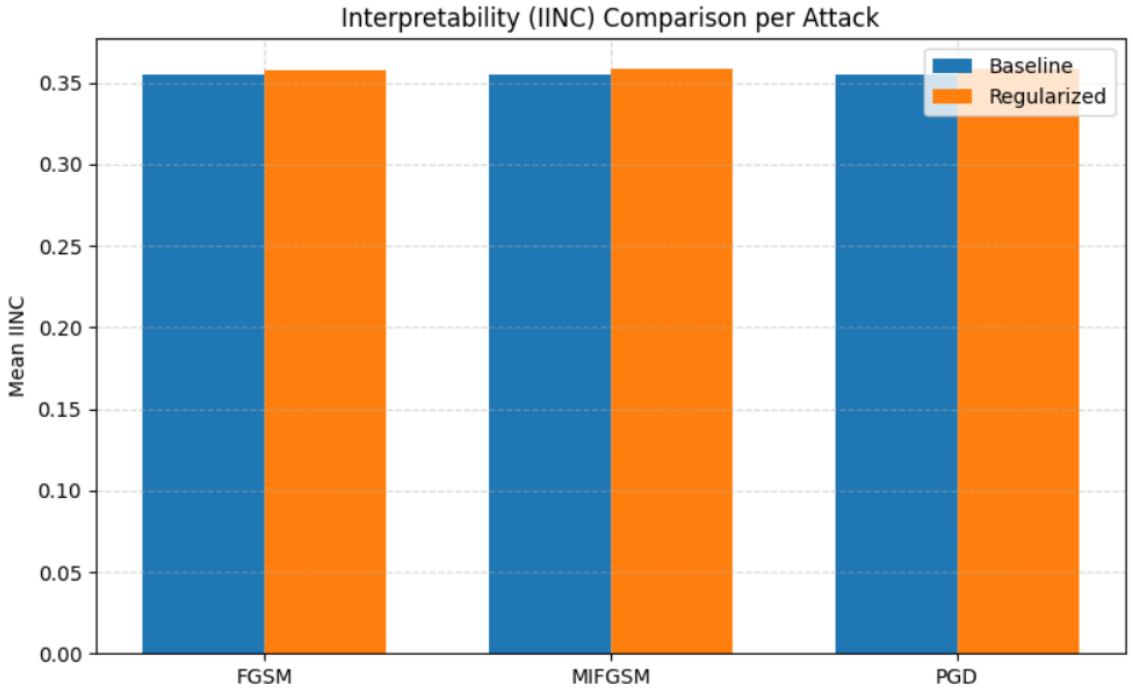
Grad-CAM visualizations highlight the model's focus areas during classification.

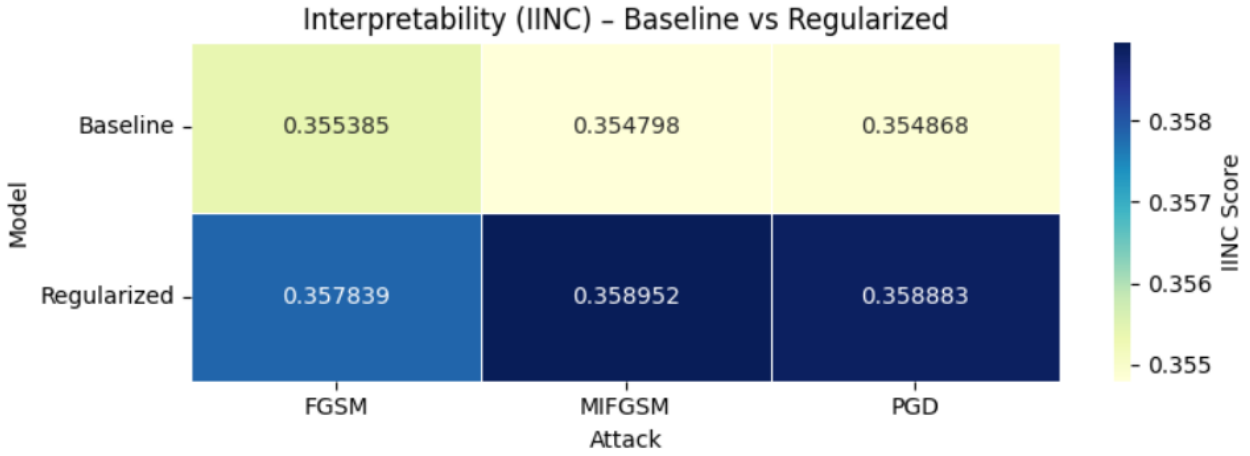
Even under adversarial attacks (FGSM, PGD, MIFGSM), the regularized model maintains a more stable and coherent attention map compared to the baseline.

This indicates higher **interpretability** and **feature consistency**, especially for fake images.



- IINC Comparison




The IINC metric evaluates how stable the model's attention (Grad-CAM) remains under adversarial perturbations.
Across all attacks, the regularized model consistently shows **slightly higher IINC scores**, indicating **improved interpretability and robustness** of internal representations.

# Conclusions and Future Work

## ➢ Conclusions

•We proposed a **deepfake detection model enhanced with gradient regularization**, targeting both **robustness** and **generalization**
•The regularized model maintained **higher accuracy, F1 scores** and **AUC** under adversarial attacks (FGSM, PGD, MIFGSM) and also showed **higher interpretability consistency** (IINC) across all attack types, indicating **stable attention maps**.

## ➢ Future Work

• Incorporate **adversarial training** (e.g., FGSM, PGD) directly into the training loop.
• Test on **additional datasets** (e.g., DFDC, Celeb-DF) to evaluate **cross-domain generalization**.
• Explore **advanced regularization strategies** (e.g., consistency loss, Mixup, feature denoising).

# References


- 1. W. Guan, W. Wang, J. Dong and B. Peng, (2024). Improving Generalization of Deepfake Detectors by Imposing Gradient Regularization, In IEEE Transactions on Information Forensics and Security, vol. 19, pp. 5345-5356.
- 2. M. Tan and Q. Le, (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In Proc. Int. Conf. Mach. Learn., pp. 6105–6114.
- 3. On the Detection of Digital Face Manipulation Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, Anil Jain, (2020), In: Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR 2020), Seattle, WA, Jun. 2020
- 4. Abbasi, M., V´az, P., Silva, J. and Martins, P. (2025). Comprehensive Evaluation of Deepfake Detection Models: Accuracy, Generalization, and Resilience to Adversarial Attacks. Applied Sciences, 15(3), 1225.