ADVERSARIAL ATTACKS AND DEFENCE STRATEGIES FOR IMPROVING ACCURACY IN MEDICAL IMAGE DIAGNOSIS

M Ajay Kumar Naidu (B210742EC)
M Veera Abhi Nanda (B210708EC)
N Vijay Praneeth (B210658EC)
Pranav Kandukuru (B210612EC)
Liyana N K (B210761EC)

Guide: Dr. Ameer P M, NIT Calicut
09 September 2024

B.Tech, Electronics and Communication Engineering
National Institute of Technology, Calicut

Overview

Introduction

- Deep learning significantly enhances medical image analysis by improving diagnostic accuracy using neural networks for complex images like MRI, CT scans, and X-rays.
- However, adversarial attacks pose risks by corrupting pixel semantics, leading to potential clinical errors.
- This study aims to develop a defense mechanism against such attacks by training a ResNet50 model on the Kvasir[1] and Chest X-ray[2] datasets, targeting high-frequency perturbations.

Motivation

- The reliance on deep learning in medical diagnostics requires addressing risks from adversarial attacks.
- A recently discovered frequency-based adversarial attack is claimed to be imperceptible and more efficient than traditional adversarial attacks[3].
- Effective defense mechanisms for this type of attacks are urgently needed to ensure the reliability of AI diagnostic tools.
- Safeguarding AI-driven systems is crucial to enhance the quality and safety of patient care.

Problem Statement

- Develop a robust defense mechanism to protect deep learning models in medical diagnostics from frequency-based adversarial attacks.
- Address the challenge of high-frequency perturbations that can cause dangerous misdiagnoses[4].
- Ensure the mechanism effectively mitigates adversarial perturbations while preserving the integrity of original medical images.
- Implement a frequency-based strategy to enhance the security and reliability of AI-driven diagnostic systems in clinical settings.

Objectives

- Develop a deep-learning defense mechanism to counter frequency-based adversarial attacks in medical image analysis.
- Implement frequency-constrained attacks on Kvasir dataset and develop a strategy targeting high-frequency perturbations.
- Evaluate and fine-tune the defense mechanism by comparing prediction accuracy of original and defended images.
- Enhance the robustness of AI-driven diagnostic systems, ensuring secure and accurate healthcare delivery.

Background

Traditional Adversarial Attacks

- **Fast Gradient Sign Method (FGSM):** FGSM generates adversarial examples by computing the gradient of the loss concerning the input image and adding a small perturbation in the direction that maximizes the loss[5].
- **Projected Gradient Descent (PGD):** An iterative refinement of FGSM that applies multiple small perturbations, projecting the result back onto a feasible set after each step to maintain imperceptibility[5].
- **Carlini & Wagner (C&W) Attack:** A powerful optimization-based attack that minimizes the perturbation added while ensuring mis-classification[5].

Background

Traditional Adversarial Attacks

- **Directed Information Maximization (DIM):** It involves manipulating inputs to maximize the amount of information transferred from the input to the model's output, aiming to fool the model into making incorrect predictions or classifications[6].
- **DeepFool:** It works by iteratively finding the smallest perturbation that can be added to an image to push it across the decision boundary, causing the model to misclassify the image[6].

Background

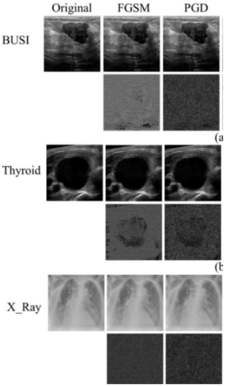

Figure 1: The adversarial examples and perturbations generated by different attack methods on 2D medical image datasets with different modals against ResNet18[5]

Background

Frequency Constraint-Based Adversarial Attack

- **High-Frequency Perturbations:** By constraining perturbations to highfrequency domains, the attack maintains visual similarity, making alterations imperceptible to human observers and less detectable by standard defense mechanisms[7].
- **Applicability Across Modalities:** The attack has been tested on various medical imaging modalities and dimensionalities, including 2D chest X-rays, breast and thyroid ultrasound images, and 3D CT scans, demonstrating its versatility.

Background

Feature Space Attack:

- **Objective**: Modify image features to deceive the model by altering its representation in the feature space.
- **Optimization**:
  - Find adversarial images x_{adv} that minimize similarity to the original class and maximize similarity to target features.
  - **Formula**:

    $$x_{\text{adv}} = \arg\min_{x_i'} \left(\max\left[0, h_{i,i} - \min(h_{i,j} | j \neq i)\right]\right)$$

    where $h_{i,i}$ is the similarity to the original image, and $h_{i,j}$ is the similarity to other classes.
- **Targeted vs. Non-Targeted Attacks**:
  - **Targeted Attack**: Increase similarity to a target class.
  - **Non-Targeted Attack**: Reduce similarity to the original class[5].

Background

Frequency Domain Transformation:

- **Fourier Transform**: Converts spatial image to frequency space to separate low and high-frequency components.
- **Low vs. High Frequency**:
  - Low frequencies hold main structural information.
  - High frequencies contain textures and details (targeted for attack).
- **Formula for Separation**:

$$z_L(i,j) = \begin{cases} z(i,j), & \text{if } d((i,j),(c_i,c_j)) \leq r \\ 0, & \text{otherwise} \end{cases}$$

$$z_H(i,j) = \begin{cases} 0, & \text{if } d((i,j),(c_i,c_j)) \leq r \\ z(i,j), & \text{otherwise} \end{cases}$$

- **Low-Frequency Distance Constraint**:

$$D_L(z_{\text{ori},L}, z_{\text{adv},L}) = d((a_{\text{ori}}, b_{\text{ori}}), (a_{\text{adv}}, b_{\text{adv}}))$$

Background

Overall Attack Objective:

- Balances feature similarity and frequency constraints to ensure effectiveness and imperceptibility.
- **Formula**:

$$x_{\text{adv}} = \arg\min_{x_i'} \{\max[0, \alpha_i h_{i,i} - \beta_i \min(h_{i,j}|j \neq i)] + \lambda D_L\}$$

- **Parameters**:
  - α_i, β_i: Weights for feature space similarity.
  - λ: Controls importance of low-frequency constraint.

Conclusion:

- This attack framework strategically perturbs high-frequency components, deceiving models while preserving realistic appearance.

Background

Algorithm 1 Adversarial attack by using the proposed method

Input: The map function M(·)(feature exactor); a batch of original images $\{x_i^{ori}\}_{i=1}^{N}$; the number of iterations K.

Output: A batch of adversarial images $\{x_i^{adv}\}_{i=1}^{N}$

Initialize: $\{x_i^{adv}\}_{i=1}^{N} \neg \{x_i^{ori}\}_{i=1}^{N}$

for $i = 1$ to N:

  $r_i \neg \operatorname{arctanh}(2x_i^{adv} - 1)$

  for $k = 1$ to K:

    $h_{i,i} \neg$ Calcuate $sim(M(r_i'), M(r_i))$ from i to N

    $h_{i,j} \neg$ Calcuate $sim(M(r_i'), M(r_j))(j^{1}\ i)$

    z_L^{ori}, $z_L^{adv} \neg$ Use Fourier transform get low frequency from x_i^{ori}, x_i^{adv}

    $D_L \neg d(z_L^{ori}, z_L^{adv})$

    $r_i \neg$ Optimize variable r_i by $\arg\min_{x_i'}\{\max[0, \alpha_i h_{i,j} - \beta_i \min(h_{i,j} \mid j^{1}\ i)]\} + \lambda D_L$

    $x_i^{adv} \neg \frac{1}{2}\tanh(r_i)$

  end for

end for

return $\{x_i^{adv}\}_{i=1}^{N}$

Figure 2: Algorithm for generating adversarial examples using feature space attack and frequency domain constraint [7].

Background

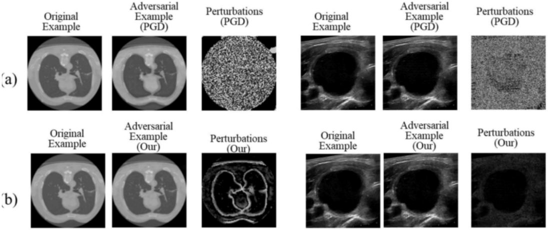

Figure 3: (a) Adversarial examples and perturbations generated by PGD method for CT and ultrasound medical images; (b) Adversarial examples and perturbations generated by the proposed method for CT and ultrasound medical images. For the visualization, perturbations are regularized by taking their absolute value and multiplying by 20 [7].

Background

Performance Metrics

- **Attack Success Rate (ASR):** The method consistently achieved high ASR across all datasets, indicating its effectiveness in deceiving various deep learning models.
- **Frechet Inception Distance (FID):** Low FID scores were observed, reflecting high similarity between original and adversarial images and confirming the imperceptibility of perturbations.
- **Low-Frequency Component Distortion (LF):** Minimal LF values indicated that essential structural information was preserved, ensuring that adversarial examples remained realistic and clinically plausible

Methodology

Data Collection and Preprocessing

- **Data Sources:** Used Kvasir,Chest X-ray and BUSI datasets to cover diverse medical conditions and imaging types.
- **Data Preprocessing:** Applied normalization and augmentation techniques to enhance data variability.

Model Implementation

- **Model Architecture:** Utilized the ResNet50 architecture on Kvasir dataset(RGB images) and ResNet 18 Architecture on Chest X-ray and BUSI datasets(Grayscale images) for its effectiveness in medical image classification tasks.
- **Framework:** Implemented the model using the PyTorch framework for flexibility and ease of development.
- **Training:** Trained the model on the preprocessed datasets to accurately classify medical images and learn feature representations.

Methodology - RPCA-Based High-Frequency Component Filtering

RPCA Defense Strategy:

- **Purpose**: Use Robust Principal Component Analysis (RPCA) to filter high-frequency components containing adversarial perturbations.
- **Process Overview**:
  - Decompose the image into low-rank (structural) and high-frequency (sparse) components.
  - Apply Gaussian noise to neutralize adversarial effects in high-frequency areas while preserving diagnostic information.
- **RPCA Advantages**:
  - Focuses on high-frequency components to remove adversarial noise without significantly altering critical image content.

Methodology - Working of the RPCA Mechanism

RPCA Steps:

1. **Adding Gaussian Noise**: Iteratively applies Gaussian noise to high-frequency components to weaken adversarial effects.
2. **Frobenius Norm Calculation**:
   - Calculates the matrix magnitude to monitor convergence during RPCA decomposition.
3. **Convergence Check**:
   - Stops iterations once the relative error between original and decomposed matrices is minimal, ensuring effective filtering.

Filtering Approach:

- Separates key image structures from noise using a threshold, preserving main diagnostic features and filtering adversarial perturbations.

RPCA Defense - Detailed Process and Components

RPCA Component Decomposition:

- **Soft Thresholding**: Shrinks matrix values toward zero to isolate small noise components.
- **SVD-Based Soft Thresholding**:
  - Uses Singular Value Decomposition (SVD) to distinguish low-rank components (important structural information) from high-frequency noise.

Final Image Reconstruction:

- **Low-Rank Component** J: Retains main diagnostic structures[8].
- **Sparse Component** S: Captures high-frequency details, potentially including adversarial noise[8].
- **Noise Component** W: High-frequency noise, filtered out to reduce adversarial impact[8].

Outcome:

- A defended image with high diagnostic integrity and reduced adversarial influence.

Model Evaluation and Fine-Tuning

Evaluation of RPCA-Based Defense Mechanism:

- The effectiveness of the RPCA-based defense is evaluated by comparing the classification accuracy on original and defended images.
- Ensures that the RPCA method effectively filters adversarial noise while preserving diagnostic information[9].

Fine-Tuning the Defense Strategy:

- Fine-tuning is performed to optimize defense performance against diverse adversarial attacks.
- Goal: Maintain high accuracy and minimize impact on diagnostic quality[9].

Results - Model Implementation on Datasets

Kvasir Dataset (ResNet50):

- Achieved a high classification accuracy of **93.8%**.

Chest X-ray Dataset (ResNet18):

- Obtained an accuracy of **94.36%** for chest X-ray classification.

BUSI Dataset (ResNet18):

- Recorded an accuracy of **92.18%** for breast ultrasound images.

Insight: These results demonstrate the strong performance of the models on clean (unaltered) datasets.

Results of Traditional Adversarial Attacks

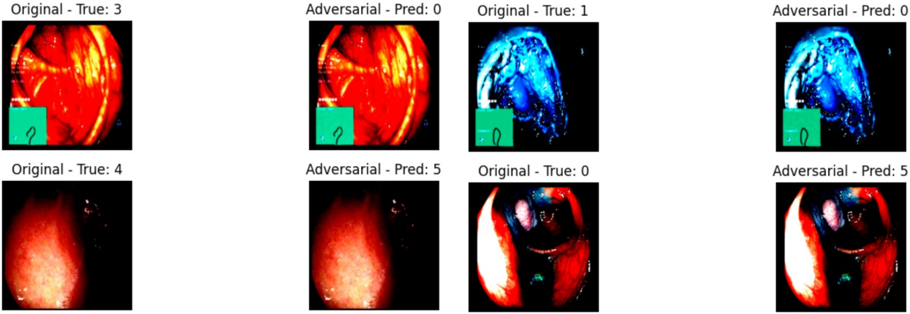

Figure 4: Original images with predicted labels and corresponding Attacked images with prediction by the model

Attack Evaluation

```
FGSM Attack:
Success Rate: 89.38%
Image 0: Original: 3, Adversarial: 5
Image 1: Original: 4, Adversarial: 5
Image 2: Original: 1, Adversarial: 1
Image 3: Original: 0, Adversarial: 0
Image 4: Original: 2, Adversarial: 5
.
.
.
Image 635: Original: 1, Adversarial: 0
Image 636: Original: 3, Adversarial: 0
Image 637: Original: 7, Adversarial: 0
Image 638: Original: 5, Adversarial: 2
Image 639: Original: 1, Adversarial: 5
```

```
PGD Attack:
Success Rate: 99.84%
Image 0: Original: 3, Adversarial: 0
Image 1: Original: 4, Adversarial: 5
Image 2: Original: 1, Adversarial: 0
Image 3: Original: 0, Adversarial: 5
Image 4: Original: 2, Adversarial: 1
.
.
.
Image 635: Original: 1, Adversarial: 0
Image 636: Original: 3, Adversarial: 0
Image 637: Original: 7, Adversarial: 0
Image 638: Original: 5, Adversarial: 2
Image 639: Original: 1, Adversarial: 5
```

```
BIM Attack:
Success Rate: 98.12%
Image 0: Original: 3, Adversarial: 0
Image 1: Original: 4, Adversarial: 5
Image 2: Original: 1, Adversarial: 0
Image 3: Original: 0, Adversarial: 1
Image 4: Original: 2, Adversarial: 5
.
.
.
Image 635: Original: 1, Adversarial: 0
Image 636: Original: 3, Adversarial: 0
Image 637: Original: 7, Adversarial: 6
Image 638: Original: 5, Adversarial: 2
Image 639: Original: 1, Adversarial: 0
```

```
DeepFool Attack:
Success Rate: 75.94%
Image 0: Original: 3, Adversarial: 0
Image 1: Original: 4, Adversarial: 5
Image 2: Original: 1, Adversarial: 1
Image 3: Original: 0, Adversarial: 0
Image 4: Original: 2, Adversarial: 2
.
.
.
Image 635: Original: 1, Adversarial: 0
Image 636: Original: 3, Adversarial: 0
Image 637: Original: 7, Adversarial: 0
Image 638: Original: 5, Adversarial: 2
Image 639: Original: 1, Adversarial: 1
```

```
C&W Attack:
Success Rate: 99.53%
Image 0: Original: 3, Adversarial: 0
Image 1: Original: 4, Adversarial: 5
Image 2: Original: 1, Adversarial: 0
Image 3: Original: 0, Adversarial: 1
Image 4: Original: 2, Adversarial: 1
.
.
.
Image 635: Original: 1, Adversarial: 5
Image 636: Original: 3, Adversarial: 0
Image 637: Original: 7, Adversarial: 0
Image 638: Original: 5, Adversarial: 2
Image 639: Original: 1, Adversarial: 5
```

Figure 5: (a) FGSM Attack (b) PGD Attack (c) BIM Attack (d) Deepfool Attack (e) C&W Attack

RPCA Results on Original and PGD-Attacked Images

```
Original Labels | Noisy Predictions | RPCA Predictions
Image 1: Original: 3 | Noisy: 4 | RPCA: 3
Image 2: Original: 2 | Noisy: 5 | RPCA: 2
Image 3: Original: 2 | Noisy: 5 | RPCA: 2
Image 4: Original: 4 | Noisy: 5 | RPCA: 4
Image 5: Original: 6 | Noisy: 1 | RPCA: 0
Image 6: Original: 7 | Noisy: 5 | RPCA: 7
Image 7: Original: 7 | Noisy: 5 | RPCA: 7
Image 8: Original: 5 | Noisy: 5 | RPCA: 5
Image 9: Original: 6 | Noisy: 5 | RPCA: 6
Image 10: Original: 4 | Noisy: 5 | RPCA: 4
Image 11: Original: 4 | Noisy: 5 | RPCA: 4
Image 12: Original: 6 | Noisy: 5 | RPCA: 6
Image 13: Original: 6 | Noisy: 5 | RPCA: 6
Image 14: Original: 1 | Noisy: 1 | RPCA: 1
Image 15: Original: 1 | Noisy: 1 | RPCA: 1
```

(a) RPCA results on original images with original predictions (retention = 92.57%).

```
Original Labels | Attacked Predictions | RPCA Predictions
Image 1: Original: 3 | Attacked: 0 | RPCA: 7
Image 2: Original: 2 | Attacked: 5 | RPCA: 2
Image 3: Original: 2 | Attacked: 5 | RPCA: 2
Image 4: Original: 4 | Attacked: 0 | RPCA: 0
Image 5: Original: 6 | Attacked: 0 | RPCA: 6
Image 6: Original: 7 | Attacked: 0 | RPCA: 6
Image 7: Original: 7 | Attacked: 0 | RPCA: 0
Image 8: Original: 5 | Attacked: 2 | RPCA: 2
Image 9: Original: 6 | Attacked: 5 | RPCA: 7
Image 10: Original: 4 | Attacked: 5 | RPCA: 7
Image 11: Original: 4 | Attacked: 5 | RPCA: 7
Image 12: Original: 6 | Attacked: 0 | RPCA: 6
Image 13: Original: 6 | Attacked: 0 | RPCA: 6
Image 14: Original: 1 | Attacked: 0 | RPCA: 0
Image 15: Original: 1 | Attacked: 5 | RPCA: 6
```

(b) RPCA results on images attacked with PGD.

Figure 6: RPCA results on original and PGD-attacked images

RPCA Results on FGSM and BIM Attacks

```
Original Labels | Attacked Predictions | RPCA Predictions
Image 1: Original: 3 | Attacked: 5 | RPCA: 7
Image 2: Original: 2 | Attacked: 4 | RPCA: 2
Image 3: Original: 2 | Attacked: 5 | RPCA: 2
Image 4: Original: 4 | Attacked: 0 | RPCA: 0
Image 5: Original: 6 | Attacked: 0 | RPCA: 6
Image 6: Original: 7 | Attacked: 0 | RPCA: 6
Image 7: Original: 7 | Attacked: 0 | RPCA: 0
Image 8: Original: 5 | Attacked: 2 | RPCA: 2
Image 9: Original: 6 | Attacked: 5 | RPCA: 7
Image 10: Original: 4 | Attacked: 4 | RPCA: 7
Image 11: Original: 4 | Attacked: 5 | RPCA: 7
Image 12: Original: 6 | Attacked: 0 | RPCA: 6
Image 13: Original: 6 | Attacked: 0 | RPCA: 6
Image 14: Original: 1 | Attacked: 0 | RPCA: 0
Image 15: Original: 1 | Attacked: 0 | RPCA: 6
```

(a) RPCA results of images attacked with FGSM.

```
Original Labels | Attacked Predictions | RPCA Predictions
Image 1: Original: 3 | Attacked: 6 | RPCA: 7
Image 2: Original: 2 | Attacked: 5 | RPCA: 2
Image 3: Original: 2 | Attacked: 5 | RPCA: 2
Image 4: Original: 4 | Attacked: 5 | RPCA: 0
Image 5: Original: 6 | Attacked: 0 | RPCA: 6
Image 6: Original: 7 | Attacked: 2 | RPCA: 6
Image 7: Original: 7 | Attacked: 0 | RPCA: 0
Image 8: Original: 5 | Attacked: 2 | RPCA: 2
Image 9: Original: 6 | Attacked: 7 | RPCA: 7
Image 10: Original: 4 | Attacked: 3 | RPCA: 7
Image 11: Original: 4 | Attacked: 5 | RPCA: 7
Image 12: Original: 6 | Attacked: 0 | RPCA: 6
Image 13: Original: 6 | Attacked: 2 | RPCA: 6
Image 14: Original: 1 | Attacked: 0 | RPCA: 0
Image 15: Original: 1 | Attacked: 1 | RPCA: 6
```

(b) RPCA results of images attacked with BIM.

Figure 7: RPCA results on images attacked with FGSM and BIM

RPCA Results on Deep Fool and C&W Attacks

```
Original Labels | Attacked Predictions | RPCA Predictions
Image 1: Original: 3 | Attacked: 5 | RPCA: 7
Image 2: Original: 2 | Attacked: 5 | RPCA: 2
Image 3: Original: 2 | Attacked: 5 | RPCA: 2
Image 4: Original: 4 | Attacked: 0 | RPCA: 0
Image 5: Original: 6 | Attacked: 0 | RPCA: 6
Image 6: Original: 7 | Attacked: 0 | RPCA: 6
Image 7: Original: 7 | Attacked: 0 | RPCA: 0
Image 8: Original: 5 | Attacked: 2 | RPCA: 2
Image 9: Original: 6 | Attacked: 5 | RPCA: 7
Image 10: Original: 4 | Attacked: 5 | RPCA: 7
Image 11: Original: 4 | Attacked: 1 | RPCA: 7
Image 12: Original: 6 | Attacked: 0 | RPCA: 6
Image 13: Original: 6 | Attacked: 0 | RPCA: 6
Image 14: Original: 1 | Attacked: 5 | RPCA: 0
Image 15: Original: 1 | Attacked: 0 | RPCA: 6
```

(a) RPCA results of images attacked with Deep Fool.

```
Original Labels | Attacked Predictions | RPCA Predictions
Image 1: Original: 3 | Attacked: 0 | RPCA: 7
Image 2: Original: 2 | Attacked: 5 | RPCA: 2
Image 3: Original: 2 | Attacked: 5 | RPCA: 2
Image 4: Original: 4 | Attacked: 0 | RPCA: 0
Image 5: Original: 6 | Attacked: 0 | RPCA: 6
Image 6: Original: 7 | Attacked: 0 | RPCA: 6
Image 7: Original: 7 | Attacked: 1 | RPCA: 0
Image 8: Original: 5 | Attacked: 2 | RPCA: 2
Image 9: Original: 6 | Attacked: 5 | RPCA: 7
Image 10: Original: 4 | Attacked: 4 | RPCA: 7
Image 11: Original: 4 | Attacked: 4 | RPCA: 7
Image 12: Original: 6 | Attacked: 0 | RPCA: 6
Image 13: Original: 6 | Attacked: 0 | RPCA: 6
Image 14: Original: 1 | Attacked: 5 | RPCA: 0
Image 15: Original: 1 | Attacked: 5 | RPCA: 6
```

(b) RPCA results of images attacked with C&W.

Figure 8: RPCA results on images attacked with Deep Fool and C&W

Frequency-Based Adversarial Attack Results

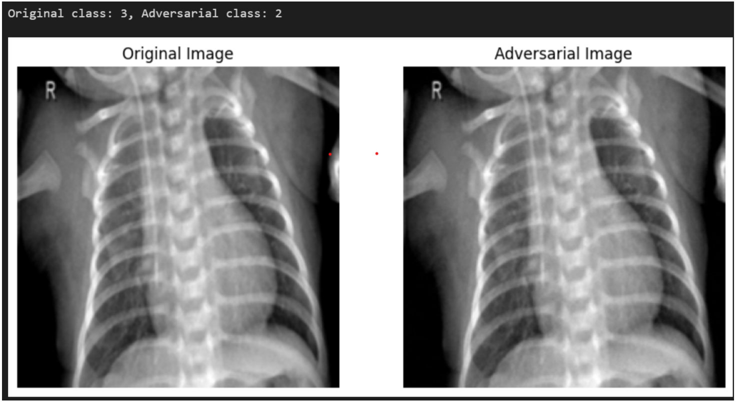

Figure 9: Comparison between original image and frequency-based adversarial attacked image from the chest X-ray dataset.

Frequency-Based Attack Success Rate and RPCA results on Chest X-ray Images

Attack Success Rate: 92.50%
RPCA Success Rate: 59.46%

	Original Class	Adversarial Class	RPCA Class
0	0	3	0
1	0	1	0
2	0	3	0
3	0	0	0
4	0	3	0
5	0	2	1
6	0	0	0
7	0	2	0
8	0	3	0
9	0	2	0
10	2	0	0
11	2	0	2
12	0	2	2
13	2	0	0
14	2	0	0
15	2	0	0
16	2	0	2
17	2	0	1
18	2	0	3
19	0	0	0
20	1	0	1
21	1	0	1
22	1	0	0
23	1	0	1
24	1	0	0
25	1	0	0
26	1	0	0
27	1	0	1
28	1	0	1
29	1	0	1
30	3	0	3
31	3	0	3
32	3	0	0
33	3	2	3
34	3	0	3
35	3	2	3
36	3	0	3
37	3	0	0
38	3	0	3
39	3	0	0

Figure 10: Attack and RPCA results on Chest X-ray images using frequency-based attack with an attack success rate of 92.5% and original prediction retention of 59.46%.

Work Plan - Ongoing Tasks

RPCA Implementation and Testing:

- Implementing RPCA defense on frequency-based adversarial attacked images.
- Test the defense on multiple datasets to ensure its effectiveness.

Fine-Tuning:

- Optimize the RPCA parameters to enhance defense performance while maintaining high diagnostic accuracy.

Performance Evaluation:

- Compare the accuracy of defended images with that of original images to confirm the defense's reliability in various scenarios.

Work Plan - Future Directions

Development of Advanced Defense Mechanisms:

- Research and create more resilient defense techniques to counter complex frequency-based adversarial attacks.

Exploration of New Attack Types:

- Study and test emerging adversarial attacks that could potentially bypass existing defenses.

Real-World Application Testing:

- Validate the RPCA-based defense mechanism's robustness and reliability in clinical settings to ensure practical viability for medical diagnostics.

References I

- C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826.
- K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," *IEEE Access*, vol. 6, pp. 14 410–14 430, 2018.
- Y. Guo, Y. Liu, T. Georgiou, and M. Lew, "A review of semantic segmentation using deep neural networks," *International Journal of Multimedia Information Retrieval*, vol. 7, 06 2018.

References II

G. Bortsova, C. González-Gonzalo, S. C. Wetstein, F. Dubost, I. Katramados, L. Hogeweg, B. Liefers, B. van Ginneken, J. P. Pluim, M. Veta, C. I. Sánchez, and M. de Bruijne, "Adversarial attack vulnerability of medical image analysis systems: Unexplored factors," *Medical Image Analysis*, vol. 73, p. 102141, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841521001870

X. Ma, Y. Niu, L. Gu, Y. Wang, Y. Zhao, J. Bailey, and F. Lu, "Understanding adversarial attacks on deep learning based medical image analysis systems," *Pattern Recognition*, vol. 110, p. 107332, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320320301357

F. Chen, J. Wang, H. Liu, W. Kong, Z. Zhao, L. Ma, H. Liao, and D. Zhang, "Frequency constraint-based adversarial attack on deep neural networks for medical image classification," *Computers in Biology and Medicine*, vol. 164, p. 107248, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482523007138

References III

A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "Unetr: Transformers for 3d medical image segmentation," in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022, pp. 1748–1758.

X. Wei, R. Yu, and J. Sun, "View-gcn: View-based graph convolutional network for 3d shape analysis," 06 2020, pp. 1847–1856.

Thank You