Recovering Medical Images from Adversarial Attacks: Genetic Algorithm-based Adaptive Compression (GA-AC)

Paulo Vitor C. Lima[1], Silvio E. Quincozes[1,2], Marcelo Z. do Nascimento[1], Juliano F. Kazienko[3], Daniel Welfer[3], and Shigueo Nomura[1]

[1]PPGCO – Universidade Federal de Uberlândia (UFU)
[2]Horizon AI Labs/PPGES – Universidade Federal do Pampa (UNIPAMPA)
[3]Universidade Federal de Santa Maria (UFSM)

{paulo.lima1,marcelo.nascimento,shigueonomura}@ufu.br
silvioquincozes@unipampa.edu.br
kazienko@redes.ufsm.br, daniel.welfer@ufsm.br

Abstract. *The Fast Gradient Sign Method (FGSM) has become such a critical threat of adversarial attacks on deep learning models for processing medical images. The problem has led to prediction errors with non-satisfactory results on diagnosis and patient safety. We address this challenge by proposing a novel approach named Genetic Algorithm-based Adaptive Compression (GA-AC) for recovering images perturbed by the FGSM attacks. The GA-AC optimize PNG and WebP compression methods to maximize Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM), thereby preserving essential diagnostic features in the restored images. Experimental results on multiple X-ray images demonstrate the effectiveness of GA-AC, which was able to restore the model's F1-score from 24.14% to 98.10% after FGSM attacks.*

1. Introduction

Deep learning has become a great tool for medical images, detecting and diagnosing critical diseases [Aggarwal et al. 2021]. However, nowadays, the incidence of cyber attacks in all areas has also grown considerably, and the healthcare sector has certainly suffered from this, especially with the impact caused on models that classify images in order to diagnose diseases [Bortsova et al. 2021]. One of the main offenders in these systems are the so-called adversarial attacks, which insert small disturbances and images to the point of causing errors in models, generating incorrect diagnoses [Yao Li and Lee 2022]. In the context of Healthcare 5.0, where artificial intelligence plays a fundamental role in evolving the sector [Gomathi et al. 2023], concerns grow as the technology advances.

It is critical to be capable of recovering medical images from adversarial attacks. Firstly, because they induce to improper model predictions which accordingly leads to wrong diagnoses directly affecting patient treatment. Moreover, for situations in which images do not have corresponding copies, recovering techniques become even more relevant in order to restore their original properties.

While medical systems are generally protected by secure acquisition and transmission protocols, real-world scenarios still expose them to adversarial vulnerabilities. For instance, black-box attacks do not require access to the target model. Instead, they

exploit shared input-output behavior by training substitute models using available image-label pairs, enabling gradient-free or transfer-based adversarial examples to succeed even across architectures[Inkawhich et al. 2020].

Additionally, recent studies have demonstrated clinically-oriented adversarial attacks such as CoRPA (Clinically-Oriented Perturbations for Radiology Applications), which generate domain-aware perturbations aligned with pathological features. These perturbations may obscure or simulate findings like nodules or infiltrates, simultaneously affecting the image and its associated clinical report without being perceptible to the human eye [Ma et al. 2021].

Therefore, despite robust security practices, these threats highlight that adversarially manipulated medical images can still be introduced into diagnostic pipelines particularly in remote consultations, AI-as-a-service platforms, or mobile-based acquisition systems.

Adversarial attacks exploit the intrinsic vulnerabilities of deep neural networks by introducing subtle, human-imperceptible perturbations into medical images, severely degrading the performance of diagnostic classifiers, even under extensive training [Goodfellow et al. 2014, Carlini and Wagner 2017]. To counteract such threats, traditional image compression techniques such as JPEG and WebP have been explored as low-level defense mechanisms. While early works [Dziugaite et al. 2016, Das et al. 2018] provided initial evidence of their potential in reducing adversarial impact, their effectiveness remains limited by fixed compression parameters, often resulting in suboptimal restoration and compromised diagnostic reliability. Notably, the cited studies reflect the early stage of research in this domain, and may not capture recent advancements in adaptive and learning-based recovery methods. This gap underscores the need for more robust and tunable defense strategies. In this context, our study proposes a novel adaptive recovery approach that integrates genetic algorithms with image compression, aiming to dynamically optimize restoration quality against adversarial perturbations, as further motivated by promising outcomes in [Yin et al. 2020].

This work proposes the GA-AC (Genetic Algorithm-based Adaptive Compression) approach to optimize several parameters for medical images compression — including PNG and WebP compression, and parameters such as quality, flip intensity, rotation angle, scaling factor, and filtering operations is produced. This way, GA-AC efficiently recover images from Fast Gradient Sign Method (FGSM) adversarial attack. For that, a chest X-ray images dataset is adopted [Kermany et al. 2018] for training, testing, and validating. We use the Genetic Algorithms technique to further improve compression to increase the classifier model's success rates. Experiments reveal that WebP compression optimized by GA-AC improves classification accuracy from 28.85% to 97.92% and F1-score from 24.14% to 98.10%.

This paper is structured as follows. In Section 2, we introduce foundational concepts regarding genetic algorithms and adversarial attacks. Related works are discussed in Section 3, highlighting the effectiveness of our proposed approach. The proposed GA-AC in Section 4 presents a detailed description of our novel algorithm, including its operational mechanisms and implementation. Experimental evaluation and results discussion are presented in the Section 5. Section 6 describes the conclusions and future works.

2. Background

We first present the fundamental ideas guiding this work: *Genetic Algorithms (GAs)* for optimization of image recovery and *adversarial attacks*–in particular, the FGSM [Chen et al. 2020]–used to compromise machine learning models in image classification.

2.1. Genetic Algorithms: Image Recovery and Compression

Inspired by the process of natural selection, where a population of candidate solutions (individuals) develops toward better solutions over numerous generations [Alam et al. 2020] genetic algorithms (GAs) are optimization heuristics. Their capacity to explore difficult search spaces and avoid getting caught in local minima makes them extensively applied in artificial intelligence, optimization, and machine learning. Below, in in Table 1, the steps used for a general GA configuration are shown.

Step	**Description**
Initial population	A population of individuals is created, where each individual represents a potential solution to the problem.
Fitness Function	Each individual is evaluated using a metric (or combination of metrics) that captures its quality or suitability. This metric is often referred to as the *fitness*.
Selection	Individuals with higher fitness are more likely to be selected to pass on their traits.
Crossover	Selected individuals are "mated" by swapping sections of their representation to produce children with traits from both parents.
Mutation	Some offspring receive small random modifications to maintain diversity and prevent premature convergence.
Replacements	The new generation, consisting of offspring (and possibly some existing fit individuals), replaces the previous one until a stopping criterion is met.

Table 1. Summary of Genetic Algorithm Steps.

GAs have been effectively used in image processing to minimize distortions [Ople et al. 2023] and maximize transformations preserving particular characteristics of images. For instance, in this work a GA-AC is used to identify the optimal combination of parameters—such as compression method, rotation, or other transformations—that produce high Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) values when recovering an image [Horé and Ziou 2010].

PSNR compares the quality of a changed or compressed image to the original; higher values suggest fewer error. SSIM measure brightness, contrast, and structure, evaluates the perceived quality of two images; higher SSIM scores suggest more similarity [Kumar et al. 2013].

To thus direct the evolutionary search toward parameter sets that best restore adversarially altered images, the *fitness function* in the GA combines PSNR and SSIM.

2.2. Adversarial Attacks: The Fast Gradient Sign Method (FGSM)

Adversarial assaults are methods meant to trick machine learning models by including tiny, usually undetectable perturbations into input data, especially photographs. These

perturbations may be invisible to the human eye, but they can significantly change the prediction of a model, therefore exposing weaknesses in many deep learning architectures [Yao Li and Lee 2022].

The FGSM is among the most often used techniques to create adversarial instances. Using the gradient of the loss function concerning the input, FGSM seeks a perturbation that maximizes the error of the model.

This attack makes use of the direction of the gradient to find the perturbation that maximizes the loss, where x' is the generated adversarial example, x is the original input, ϵ is the parameter that controls the magnitude of the perturbation (attack size), $(\nabla_x J(\theta, x, y))$ is the gradient of the loss function, where θ represents the model parameters and y the true label, $sign$ is the function that returns the sign (positive, negative or zero) of the gradient, indicating the direction in which the input should be perturbed.

$$x' = x + \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y))$$

These disturbances can produce notable misclassifications even if they may be aesthetically invisible. An adversarially assaulted image displays no obvious difference to a human viewer, yet greatly influences the output of a classifier.

We can understand why combining these two ideas is advantageous by knowing both the mechanics of *Genetic Algorithms*—as strong search heuristics for optimizing complex parameter spaces—and *Adversarial Attacks*—which show how delicate small perturbations can be in image classification tasks. We investigate in the next sections how GAs might be used to repair images affected by adversarial perturbations, so boosting their quality metrics (PSNR and SSIM) and increasing general model robustness.

3. Related Work

In this section, we review the main contributions in the literature. We present a short literature overview in Section 3.1, highlighting existing limitations and how our proposal overcomes these challenges, and present more details about related works in Section 3.2.

3.1. Literature Overview

While prior studies have addressed specific aspects such as robustness to simple perturbations, image quality preservation, or domain-specific challenges, most fall short in offering a unified, scalable, and automated solution. Table 2 synthesizes these works based on seven critical dimensions.

Work	Efficiency Against Simple Attacks	Efficiency Against Complex Attacks	Generalization for New Scenarios	Preserves Quality of Image	Scalability	Focus on Medical Images	Automation with Genetic Algorithm
Yin et al. (2020)	Yes	No	No	No	Yes	No	No
Madry et al. (2019)	No	Yes	No	Yes	No	No	No
Mahfuz et al. (2021)	Yes	No	No	No	Yes	No	No
Dong et al. (2024)	No	No	No	Yes	No	No	No
Yao et al. (2022)	No	No	No	No	No	Yes	No
Jogani et al. (2022)	No	No	No	No	No	Yes	No
Yao et al. (2023)	No	No	No	No	No	Yes	No
Zhang et al. (2021)	Yes	Yes	No	Yes	No	No	No
Our Proposal	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**

Table 2. Comparison between related works and our proposal.

Overall, several existing works address individual aspects of adversarial defense but fall short of delivering a comprehensive solution. Basic transformations like flipping or noise injection [Yin et al. 2020, Zhang et al. 2021] offer lightweight protection against simple attacks but lack robustness against stronger, iterative ones. Methods like adversarial training [Madry et al. 2019] improve resilience to complex attacks yet suffer from high computational cost and poor generalization. Other strategies, such as compression and denoising [Mahfuz et al. 2021], may reduce attack impact but often compromise image quality—an unacceptable trade-off in medical imaging.

Domain-specific efforts in medical AI [Yao et al. 2022, Jogani et al. 2022, Yao et al. 2023] emphasize security, but many remain static, lack scalability, or are not adaptable to new threats. Importantly, most approaches do not integrate automation, limiting their practical use in dynamic or large-scale clinical environments.

Our proposal addresses these gaps by introducing a unified, adaptive framework that combines image transformations with genetic algorithms to dynamically optimize defense parameters. Unlike static methods, our GA-AC approach balances robustness, image quality, and generalization, while remaining scalable and tailored for medical applications. Moreover, our work departs from the traditional use of genetic algorithms in adversarial contexts—typically focused on crafting attacks—by positioning GAs as a central component of a defensive strategy. GA-AC reframes adversarial image recovery as a combinatorial optimization task, adaptively tuning transformation parameters to restore both visual fidelity and semantic integrity. This integration of adaptive transformations with evolutionary search enables a robust, automated, and clinically viable response to adversarial threats.

3.2. Details of Related Works

Early works, such as [Yin et al. 2020], demonstrated that simple image transformations—like flipping, noise addition, or brightness adjustments—can mitigate adversarial perturbations without requiring adversarial training. These lightweight techniques showed potential for enhancing robustness but offered limited protection against stronger, more adaptive attacks. In a different context, [Roberto et al. 2025] explored polynomial-based image processing methods to improve medical image analysis. Although they used the same dataset as in our study [Kermany et al. 2018], their focus was on enhancing detection rather than defending against adversarial inputs.

Many studies have demonstrated that machine learning models can be misled by adversarial examples. A notable approach introduced a non-deterministic component to detect such perturbed images, aiming to make it harder for attackers to predict the system's behavior [Machado et al. 2018]. However, this method was later bypassed by systematic adversarial attacks capable of abstracting the defense strategy. In contrast, our work focuses on recovering adversarially attacked images by leveraging a genetic algorithm to identify optimal compression parameters (e.g., PNG and WebP), aiming to restore classification accuracy and image integrity.

The FGSM, introduced in [Goodfellow et al. 2014], revealed how minimal perturbations can significantly mislead deep learning models. To address this, adversarial training was proposed, exposing models to adversarial examples during training to increase resilience. This idea was extended in [Madry et al. 2019], which combined adversarial

training with Projected Gradient Descent (PGD) for stronger robustness. However, despite their effectiveness, such approaches are computationally expensive and can degrade accuracy on clean inputs.

Alternative defenses have included randomization and non-determinism, as seen in [Machado et al. 2018], where unpredictability was used to confuse attackers. However, this strategy was later circumvented by more sophisticated adversaries. Studies like [Mahfuz et al. 2021] adopted compression and denoising to suppress gradient-based attacks, which proved efficient against FGSM but compromised image fidelity—an unacceptable drawback in clinical scenarios. Similarly, [Zhang et al. 2021] combined compression and geometric transformations (e.g., flipping) to reduce adversarial effects but lacked adaptiveness across diverse inputs and attack types.

While prior works have explored JPEG compression [Das et al. 2018], adversarial training [Madry et al. 2019], and detection techniques [Machado et al. 2018], these often entail fixed parameters or significant retraining overhead. In contrast, GA-AC enables adaptive, image-specific defense without retraining.

The work [Bortsova et al. 2021] builds on these basic ideas and helps a lot with understanding vulnerabilities by looking at things that aren't always looked at in medical imaging settings, like the characteristics of datasets and how easy it is to understand models. Understanding these traits is important for coming up with smart defenses, especially when you think about what they could mean for patient safety and accurate diagnosis.

[Hirano et al. 2021] also talks about how hard it is to deal with general adversarial perturbations that can break multiple models at the same time. This shows how important it is to make defense systems that can be changed and adjusted, instead of relying only on set, model-specific methods.

[Dong et al. 2024] review attacks and adversarial defenses in medical imaging, highlighting existing challenges and strategies. However, they lack practical experimentation and do not present a unified framework for direct application in clinical settings.

Em [Yao et al. 2022] proposes "Medical Aegis", a robust system to protect medical images. Although innovative, it faces limitations of scalability and applicability outside the medical domain due to the high specialization of the approach.

In case of [Jogani et al. 2022] this is a study on adversarial attacks in skin cancer classifiers. Although relevant to the domain, it lacks robustness against small variations and practical evaluation in real clinical settings.

For the [Yao et al. 2023] a hierarchical concealment of features is proposed to protect medical images. Although innovative, it presents challenges of interpretability and limited effectiveness against adaptive attacks, compromising practical applicability.

Our adaptive compression method is very important in the medical imaging field, where keeping diagnostic accuracy high is very important. The GA-AC approach is better at keeping diagnostic quality high and reducing adversarial effects compared to other adversarial defense methods. We learned more about adversarial attacks like FGSM and their possible effects from more research by [Paul et al. 2020] and [Gandhi and Jain 2020]. However, basic studies like [Wu 2014] and [Omari and Yaichi 2015] gave us important information about how genetic algorithms can

be used in image processing.

4. System Architecture

In this section, we present the complete system architecture in which our mechanism is integrated. The overview of components and pipeline stages of the architecture are discussed in Section 4.1 and the GA-AC algorithm is introduced in Section 4.2.

4.1. Overview: Components and Pipeline

In this section we present each phase of the proposed pipeline and their respective components. Figure 1 illustrates the entire pipeline—from the preprocessing of original images, through the generation of adversarial examples, to the recovery process using our compression-based optimization strategy. The components associated to this pipeline are divided into three groups: algorithms, content (*i.e.,* images) and evaluation (*i.e.,* detection performance metrics). Below we details them in the context of their respective phases.

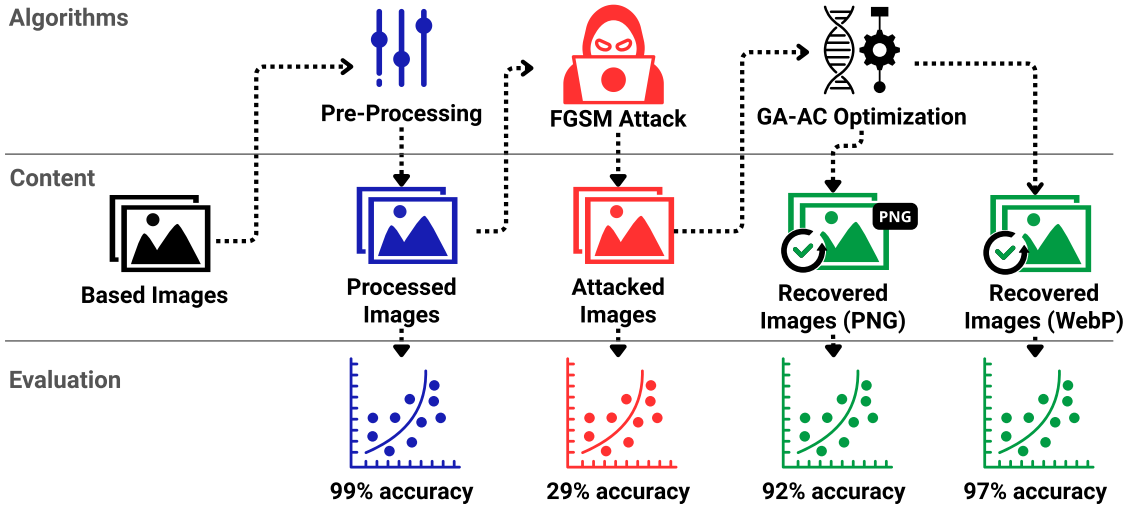


Figure 1. Proposal flowchart.

The first phase comprises the picture processing to create the basic set, applying standard procedures such as resizing all images to a fixed resolution, normalizing pixel values, and converting them to grayscale when needed. After that, our classifier was trained using these pictures. This classifier is expected to reach a high classification accuracy due to the absence of attacks in this phase (i.e., over 99%, as reported in Section 5.2).

In the second phase, adversarial examples were generated using the FGSM method, which introduces subtle, imperceptible perturbations to the input data. These modifications led to a substantial degradation in the classifier's performance. As anticipated, such perturbations impaired the model's ability to correctly classify the images into their original categories—for instance, distinguishing between healthy chest X-rays and those indicating specific diseases—resulting in a significant accuracy drop (e.g., 29%, as reported in Section 5.2).

To counteract this effect, we applied our GA-AC optimization method. This method selects compression configurations aimed at restoring the predictive performance

of the model. We evaluated two formats: PNG and WebP. As reported in Section 5.2, the classifier achieved 92% accuracy on PNG-recovered images and 97% on WebP-recovered ones, highlighting the effectiveness of the proposed recovery mechanism.

4.2. Proposed Genetic Algorithm

The proposed GA-AC recovery method is a population-based approach designed to search for optimal transformation parameters to restore adversarially perturbed images. The key parameters and fitness components of GA-AC are summarized in Table 3. As outlined in Algorithm 1, the process begins by receiving as input an adversarial image, its corresponding original version (for reference), a classifier model, and the algorithm configuration parameters such as population size and the number of generations (Line 1).

Table 3. Key Parameters and Fitness Components of GA-AC

Category	Element	Description
Chromosome Parameters	QF (Compression Quality)	WebP compression factor ranging from low to high quality (ignored for PNG format).
	$flip$ (Flip Intensity)	Degree and axis of image flipping, used to disrupt localized perturbations.
	$angle$ (Rotation Angle)	Rotation in degrees applied to the image (e.g., $[-10°, 10°]$).
	$scale$ (Scaling Factor)	Resizing factor applied to simulate geometric transformations.
	$filter$ (Filter Parameters)	Parameters defining spatial filters (e.g., Gaussian blur, sharpening).
Fitness Function Components	PSNR	Peak Signal-to-Noise Ratio, measures pixel-level similarity to original image.
	SSIM	Structural Similarity Index, evaluates perceptual quality based on structural features.
	Classifier Confidence	Confidence score or correctness of classification produced by clf on the recovered image.
GA Configuration	$popSize$	Number of individuals in the population.
	$genMax$	Maximum number of generations for evolution.
	$mutRate$ / $cxRate$	Mutation and crossover rates used during genetic operations.

The initial population is randomly generated (Line 2), with each individual representing a candidate solution composed of a combination of image transformation parameters. These include WebP compression factor (excluded for PNG format), flip intensity, rotation angle, scaling factor, and filter settings. This initialization step defines a rich and heterogeneous search space that supports both discrete and continuous transformations.

Preliminary ablation analysis suggests compression has the strongest individual effect, followed by rotation and flip. However, their combination learned adaptively by the GA offers superior results due to synergy across transformations.

From Line 3 onward, the algorithm proceeds through iterative refinement across a fixed number of generations. At each generation, individuals are evaluated using a multi-objective fitness function (Line 4), which considers three key criteria: PSNR, to ensure visual fidelity; SSIM, to assess perceptual quality; and classification reliability, to verify

Algorithm 1 GA-AC - Genetic Algorithm for Optimized Image Recovery

1: **Input:** Adversarial image img_{adv}, Original image img_{orig}, Classifier clf, Population size $popSize$, Generations $genMax$
2: **Initialize** population P with random parameters:
- WebP Compression Factor (QF) - Not applicable for PNG
- Flip Intensity ($flip$)
- Rotation Angle ($angle$)
- Scaling Factor ($scale$)
- Filter Parameters ($filter$)

3: **for** $generation \leftarrow 1$ **to** $genMax$ **do**
4: Evaluate fitness of each individual in P based on:
- Peak Signal-to-Noise Ratio (PSNR)
- Structural Similarity Index (SSIM)
- Classification reliability by clf

5: Select parents from P using tournament selection
6: Generate offspring using blend crossover ($cxBlend$)
7: Mutate offspring with Gaussian mutation ($mutGaussian$)
8: Replace least fit individuals in P with offspring
9: Update best individual parameters if improved
10: **end for**
11: **Output:** Optimal recovery parameters for img_{adv}

that the restored image aligns with the expected output of the classifier. These objectives guide the optimization process, even when they are in conflict.

Parents are selected using tournament selection (Line 5), a method that balances exploitation of good solutions with exploration of new areas in the search space. Offspring are generated via blend crossover (`cxBlend`, Line 6), which interpolates between parent parameters to create smooth variations. These offspring are then perturbed by Gaussian mutation (Line 7), a mechanism that injects controlled randomness and prevents premature convergence.

The updated offspring population replaces the least fit individuals in the current population (Line 8), and the algorithm keeps track of the best individual found so far, updating it whenever improvements are detected (Line 9). This ensures that high-quality solutions are preserved throughout the evolutionary process. After the maximum number of generations is reached (Line 10), the algorithm returns the optimal parameter configuration found for the recovery task (Line 11).

The GA-AC formulation stands out for its flexibility. It supports the integration of new transformation operators or evaluation criteria without modification to the core loop. This makes it not only effective in recovering adversarially attacked images but also extensible to broader tasks requiring adaptive, non-gradient-based optimization strategies.

5. Evaluation and Results

To evaluate the effectiveness of the proposed GA-AC recovery strategy, we conducted experiments on a deep learning model trained to classify chest X-ray images. The evaluation comprised four scenarios: (1) the original dataset without perturbations, (2) ad-

versarially attacked images using FGSM, and two recovery approaches based on image compression—(3) PNG and (4) WebP. Additionally, we analyzed the impact of the attacks themselves on the model's performance to contextualize the recovery results.

5.1. Materials and Methods

This study employed a publicly available chest X-ray dataset[1] comprising three diagnostic categories: Normal, Bacterial Pneumonia, and Viral Pneumonia. All images were converted to grayscale and resized to 128×128 pixels for consistency. A CNN-based classifier [Kesim et al. 2019] was trained using this dataset, divided into training (80%), validation, and test subsets. The training set included 1,341 normal and 3,875 pneumonia cases, while the test set contained 234 normal and 390 pneumonia cases. The model was trained over 10 epochs with a batch size of 32 using the Adam optimizer, which dynamically adjusts the learning rate to accelerate convergence. To enhance generalization and reduce overfitting, data augmentation techniques such as flipping and contrast adjustment were applied. Model performance was monitored via validation data to ensure robust learning without overfitting. After training, the model was tested under adversarial conditions using images perturbed by the FGSM, simulating real-world threats in medical AI applications through imperceptible perturbations that induce misclassification.

The classifier used in this study is a small CNN based on [Kesim et al. 2019], consisting of three convolutional layers followed by two dense layers and softmax output. Training was conducted for 10 epochs using the Adam optimizer (learning rate = 0.001, batch size = 32). For GA-AC, the genetic algorithm was configured with a population size of 50, 25 generations, a crossover rate of 0.7 (cxBlend), and mutation rate of 0.2 (mutGaussian).

To counter these adversarial effects, we applied the proposed GA-AC method, which uses a genetic algorithm to search for optimal combinations of image recovery parameters—such as compression level, rotation, scaling, brightness, and contrast adjustments. These transformations aim to restore classification performance without compromising diagnostic quality. After applying GA-AC, the recovered images were re-evaluated by the original classifier. Performance was compared across four experimental conditions as outlined in Table 4.

Table 4. Evaluation scenarios used in the methodology.

Scenario	Description
Baseline	Original, unperturbed test images
Adversarial	Test images perturbed by FGSM
PNG Recovery	GA-AC recovery using PNG compression
WebP Recovery	GA-AC recovery using WebP compression

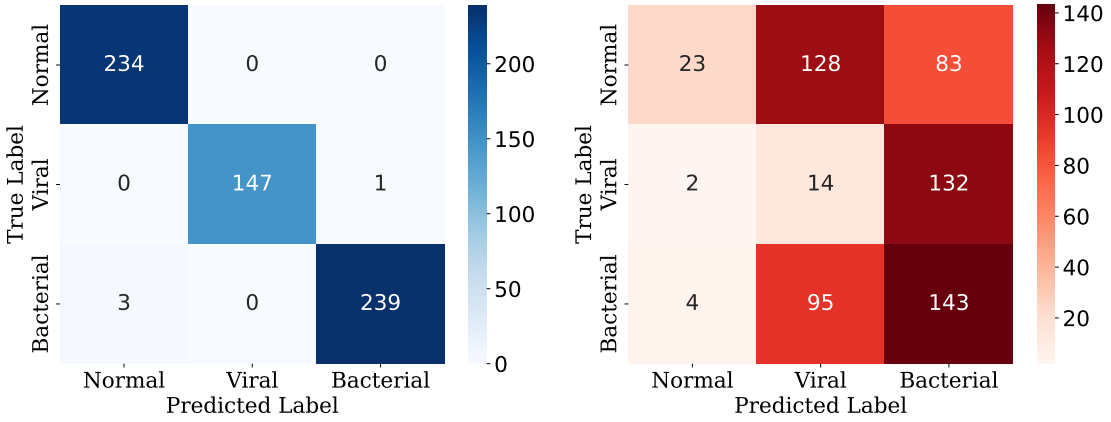
Effectiveness was measured through confusion matrices, accuracy, and F1-score, ensuring that GA-AC not only mitigates adversarial perturbations but also preserves critical diagnostic information. All scripts used in this study are open source and publicly available[2]. Therefore, this work is completely reproducible.

[1]Dataset available at `https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia/data`

[2]`https://anonymous.4open.science/r/adversarial-recover-ga-ac-61BD/`

5.2. Confusion Matrices Results

Our results reveal that, in the unaltered images, the classifier demonstrated strong detection performance across all three classes, with minimal misclassifications. As shown in Figure 2(a), it correctly classified all 234 normal cases, 147 of 148 viral pneumonia cases, and 239 of 242 bacterial pneumonia cases, indicating a well-generalized and balanced model. However, the application of adversarial perturbations using FGSM led to a dramatic degradation in classification accuracy. Figure 2(b) highlights how the model became heavily biased toward the bacterial class, misclassifying 128 normal and 95 bacterial cases as viral, while only 23 normal and 14 viral cases were correctly predicted (bacterial cases were less affected). This shift emphasizes how adversarial examples can severely distort model predictions, especially in sensitive domains like medical diagnostics. The high misclassification of normal and viral cases as bacterial raises serious concerns, as it could lead to overtreatment or mismanagement in real clinical settings.



(a) Original Images (b) Attacked Images

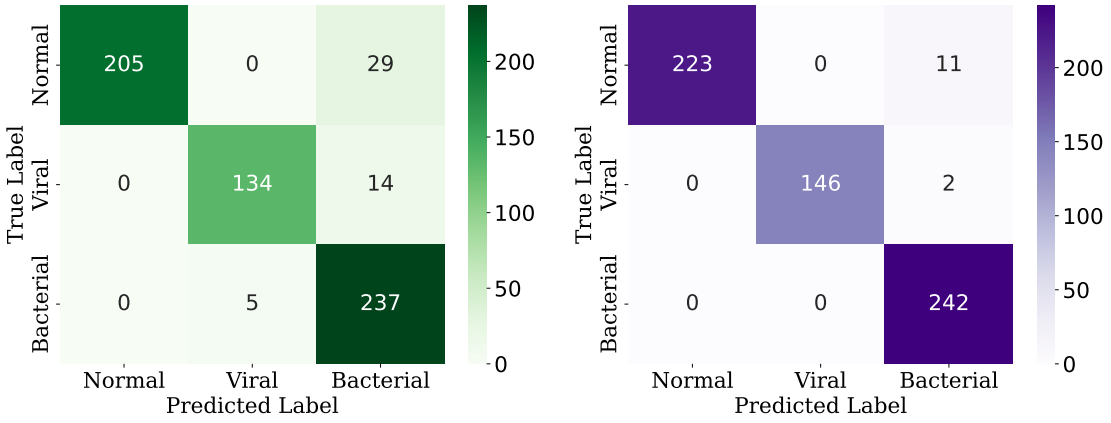
Figure 2. Confusion matrices comparing classification results for (a) original and (b) FGSM-attacked images.

Following the attack, the adversarially perturbed images were subjected to a recovery process using compression-based transformations. As shown in Figure 3, both PNG and WebP recovery strategies substantially improved the classification performance when compared to the attacked scenario. The WebP-based method achieved the most consistent recovery, correctly classifying 223 of 234 normal cases, 146 of 147 viral cases, and all the 242 bacterial cases, with only minor residual confusion. In contrast, the PNG-based recovery, while effective, exhibited a slightly higher misclassification rate–most notably with 29 normal cases and 14 viral cases predicted as bacterial (misclassified). Nonetheless, it still managed to recover 205 normal, 134 viral, and 237 bacterial cases correctly. These results confirm that while both techniques can mitigate the effects of adversarial perturbations, WebP compression offers a more robust and precise restoration of the model's diagnostic capabilities.

5.3. Performance Metrics

To further validate the effectiveness of our recovery strategies, we evaluated standard classification metrics—accuracy, F1-score, precision, and recall—across all scenarios,

`README.md`



(a) PNG-Recovered Images (b) WebP-Recovered Images

Figure 3. Confusion matrices for recovered images using (a) PNG and (b) WebP compression strategies.

as presented in Figure 4. The original, unperturbed images yielded near-perfect performance, with all metrics exceeding 0.99. In stark contrast, the FGSM-attacked images showed a dramatic collapse in classification capability, with accuracy falling to 0.2885 and recall reaching only 0.2250, underscoring the vulnerability of deep learning models to adversarial noise. Following recovery, the PNG method partially restored performance, achieving accuracy and F1-score values of 0.9231 and 0.9253, respectively. However, WebP-based recovery significantly outperformed PNG, pushing accuracy to 0.9792 and recall to 0.9800—approaching the original model's baseline. These results confirm that our GA-AC recovery method, especially when leveraging WebP compression, effectively counters adversarial degradation while preserving diagnostic reliability.

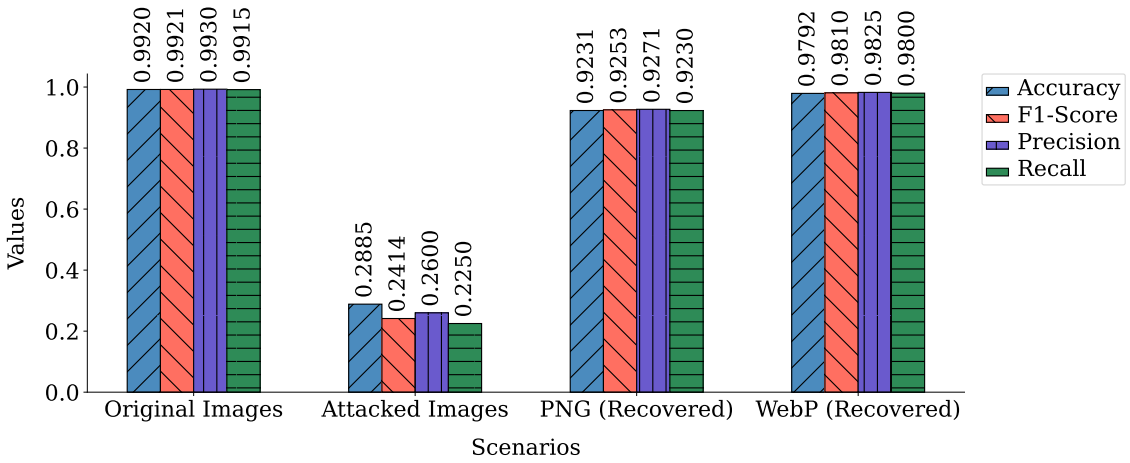


Figure 4. Classification metrics.

5.4. Results in the Practice

Figure 5 shows how different processing techniques affect hostile environment chest X-ray images. First row of a normal chest X-ray shows. The first image (A) shows the undirected, natural lung anatomy. The little noise the FGSM adversarial assault on the

second image (B) introduces can fool an artificial intelligence classifier. The third image (C) employs a PNG-based recovery strategy to somewhat lower adversarial noise and improve contrast and brightness. When combined with a 180° rotation, the WebP-based recovery shown in the fourth image (D) significantly removes hostile artifacts while maintaining the integrity of diagnostic characteristics.

Second row shows an X-ray displaying a viral infection. One of the features of viral pneumonia, widespread lung inflammation is illustrated in the original image shown in figure number five (E). Image (F) number six shows the adversarially attacked form whereby noise warps the original pattern and might lead to misdiagnosis. The PNG recovery technique reduces noise, therefore improving clarity, and preserves important elements in the seventh photo (G). Recovering the eighth image (H) from WebP compression enhances the contrast and brightness, so improving the identification of the infection indications and more effective defense against the hostile attack.

The third row illustrates an instance of bacterial pneumonia. Number nine (I), the original picture, reveals localized lung consolidation—a symptom of bacterial infection. The tenth image (J), which has been adversarially altered with distortions that obscure significant parts, can confound a classifier. Though some residual noise remains, the PNG-recovered eleventh image (K) largely replicates the original characteristics of bacterial pneumonia. The twelfth image (L) was recovered using WebP compression; it provides more exact contrast and brightness changes, so counteracting hostile disturbances while preserving the visibility of significant lung illness.

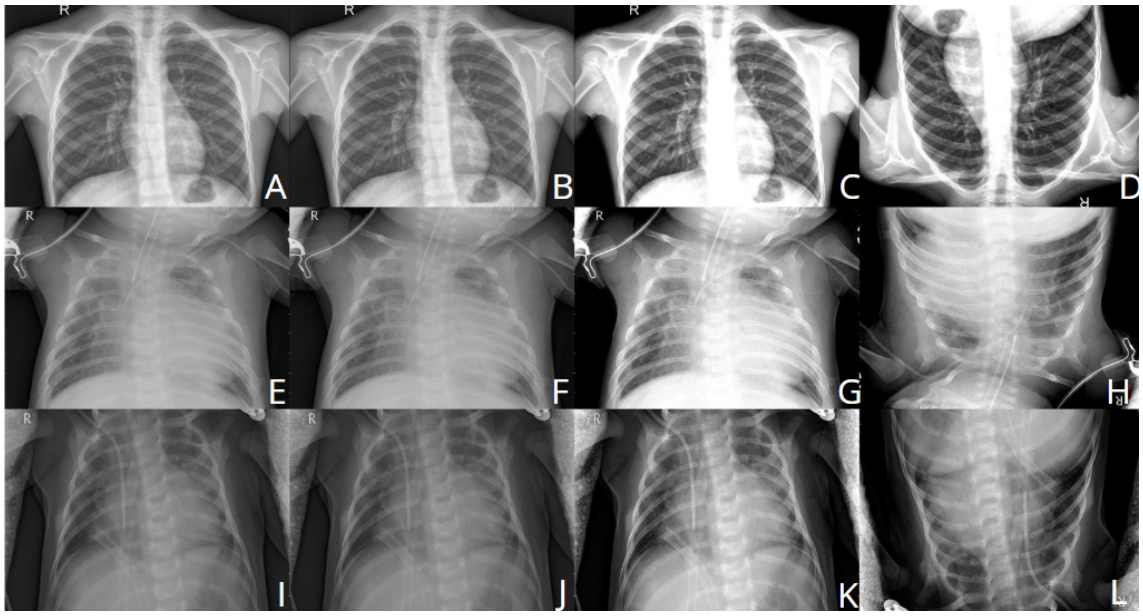


Figure 5. Adversarial Attack and Recovery Comparison (Chest X-ray).

These results demonstrate the effectiveness of the WebP-based recovery technique in preserving critical diagnostic features and suppressing noise, while also significantly mitigating the impact of FGSM adversarial attacks. Specifically, the recovery process retains high-frequency structural information, such as edges and contrast variations, which are essential for distinguishing between different thoracic conditions. The WebP compression, optimized through our genetic algorithm, filters out adversarial perturbations

without overly degrading medically relevant content. This balance is reflected in the improved classification performance and image quality metrics (e.g., PSNR and SSIM), indicating that diagnostic fidelity is maintained after recovery.

5.5. Scalability and Runtime Considerations

Despite the iterative nature of evolutionary algorithms, our GA was optimized for practical deployment. We measured the execution time of the GA-AC approach across 100 adversarial examples. The average execution time per image was approximately 3.2 seconds on a system equipped with an Intel Core i7 processor (3.2GHz) and 16GB RAM. This performance indicates that the algorithm is feasible for batch recovery tasks, though it may require optimization for real-time deployment scenarios such as clinical workflows or embedded diagnostic systems.

Though sufficient for offline or batch processing, GA-AC is not currently suited for real-time scenarios. Future work will explore parallelization, GPU acceleration, and hybrid GA strategies for real-time applicability.

5.6. Qualitative Aspects

GAs are particularly suited for this domain due to their ability to explore high-dimensional, non-differentiable, and noisy search spaces. These properties are essential when dealing with transformations that introduce nonlinear effects—such as compression artifacts or geometric distortions—where gradient-based methods fail.

GA-AC encodes image transformations as chromosomes, combining operations such as compression, flipping, rotation, scaling, and filtering. This enables the discovery of synergistic parameter sets that would be overlooked by isolated or sequential applications. The algorithm employs key evolutionary mechanisms: *Selection:* Tournament selection preserves high-fitness individuals while maintaining population diversity; *Crossover:* Blend crossover (`cxBlend`) allows fine-grained recombination of parental parameter sets; *Mutation:* Gaussian mutation introduces controlled stochasticity to prevent premature convergence; *Elitism:* Best-performing individuals are preserved across generations to guarantee continual improvement.

A multi-objective fitness function guides the search process, jointly optimizing three conflicting goals: i) Visual fidelity (measured by PSNR); ii) Perceptual similarity (measured by SSIM); and, Classification reliability (based on model output). These objectives often conflict—e.g., increasing image sharpness may improve SSIM but inadvertently reactivate adversarial artifacts, harming classification. GA-AC mediates such trade-offs by evolving candidate solutions that strike a balanced compromise. Moreover, GA-AC is inherently modular. New transformation operators (e.g., brightness correction, denoising) can be added to the chromosome encoding, and the fitness function can be adjusted to task-specific criteria (e.g., bounding-box alignment in object detection). This extensibility makes GA-AC a customizable framework for adaptive adversarial defense, rather than a rigid, one-size-fits-all solution.

Therefore, our GA-AC method is both empirically effective and theoretically grounded. It leverages principles of evolutionary computation, multi-objective optimization, and image-specific adaptation to counter adversarial perturbations. This represents

a shift from static preprocessing to dynamic, adaptive defense strategies aligned with the evolving nature of adversarial threats.

6. Conclusions and Future Works

We introduced the GA-AC method to recover perturbed X-ray images, with a focus on the FGSM algorithm. Our method optimizes parameters such as compression level, rotation, scaling, and filtering. By reframing image recovery as a multi-objective optimization problem, GA-AC seeks to preserve visual quality and semantic consistency while restoring diagnostic reliability in deep learning models.

Our results demonstrated the effectiveness of the GA-AC approach, particularly when applied with WebP compression, in mitigating the impact of adversarial noise. Classification accuracy improved from 28.85% to 97.92%, and F1-score rose from 24.14% to 98.10%, approaching baseline performance on clean images. The methodology also proved computationally feasible, with an average recovery time of 3.2 seconds per image.

The current study focuses on FGSM for controlled evaluation. Future work includes extending GA-AC to tackle iterative attacks such as PGD and Carlini-Wagner, to evaluate its robustness under more sophisticated adversarial scenarios.

Future directions also can include reducing the runtime of the algorithm through parallelization or GPU acceleration and integrating GA-AC with other defensive techniques such as adversarial detectors and reinforcement learning. Additionally, exploring its applicability in real-time clinical workflows and extending the framework to defend against other types of attacks may further enhance the resilience and practicality of AI systems in safety-critical domains.

References

Aggarwal, R., Sounderajah, V., Martin, G., Ting, D. S. W., Karthikesalingam, A., King, D., Ashrafian, H., and Darzi, A. (2021). Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *npj Digital Medicine*, 4(65):1.

Alam, T., Qamar, S., Dixit, A., and Benaida, M. (2020). Genetic algorithm: Reviews, implementations, and applications. *CoRR*, abs/2007.12673.

Bortsova, G., González-Gonzalo, C., Wetstein, S. C., Dubost, F., Katramados, I., Hogeweg, L., Liefers, B., van Ginneken, B., Pluim, J. P., Veta, M., Sánchez, C. I., and de Bruijne, M. (2021). Adversarial attack vulnerability of medical image analysis systems: Unexplored factors. *Medical Image Analysis*, 73:102141.

Carlini, N. and Wagner, D. (2017). Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE.

Chen, L., Li, H., Zhu, G., Li, Q., Zhu, J., Huang, H., Peng, J., and Zhao, L. (2020). Attack selectivity of adversarial examples in remote sensing image scene classification. *IEEE Access*, 8:137477–137489.

Das, N., Shanbhogue, M., Chen, S.-T., Hohman, F., Chen, L., Kounavis, M. E., and Chau, D. H. (2018). Shield: Fast, practical defense and vaccination for deep learning using jpeg compression. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 196–204.

Dong, J., Chen, J., Xie, X., Lai, J., and Chen, H. (2024). Survey on adversarial attack and defense for medical image analysis: Methods and challenges. *ACM Computing Surveys*, 57(3):1–38.

Dziugaite, G. K., Ghahramani, Z., and Roy, D. M. (2016). A study of the effect of jpeg compression on adversarial images. In *arXiv preprint arXiv:1608.00853*.

Gandhi, A. and Jain, S. (2020). Adversarial perturbations fool deepfake detectors. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Gomathi, L., Mishra, A. K., and Tyagi, A. K. (2023). Industry 5.0 for healthcare 5.0: Opportunities, challenges and future research possibilities. In *2023 7th International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 204–213.

Goodfellow, I., Shlens, J., and Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv 1412.6572*.

Hirano, H., Minagi, A., and Takemoto, K. (2021). Universal adversarial attacks on deep neural networks for medical image classification. *BMC Medical Imaging*, 21(1):9.

Horé, A. and Ziou, D. (2010). Image quality metrics: Psnr vs. ssim. In *2010 20th International Conference on Pattern Recognition*, pages 2366–2369.

Inkawhich, N., Wang, Y., Chen, J., and et al. (2020). Transferable adversarial attacks for image and video classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1184–1193.

Jogani, V., Purohit, J., Shivhare, I., Attari, S., and Surtkar, S. (2022). Adversarial attacks and defences for skin cancer classification.

Kermany, D., Zhang, K., and Goldbaum, M. (2018). Labeled optical coherence tomography (oct) and chest x-ray images for classification.

Kesim, E., Dokur, Z., and Olmez, T. (2019). X-ray chest image classification by a small-sized convolutional neural network. In *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*, pages 1–5.

Kumar, B., Kumar, S. B., and Kumar, C. (2013). Development of improved ssim quality index for compressed medical images. In *2013 IEEE Second International Conference on Image Information Processing (ICIIP-2013)*, pages 251–255.

Ma, X., Niu, C., Lee, H. K., and et al. (2021). Understanding adversarial attacks on deep learning-based medical image analysis systems. *npj Digital Medicine*, 4(1):1–13.

Machado, G. R., Silva, E., and Goldschmidt, R. R. (2018). Multimagnet: Uma abordagem não determinística na escolha de múltiplos autoencoders para detecção de imagens contraditórias. In *Anais do XVIII Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais*, pages 281–294, Porto Alegre, RS, Brasil. SBC.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2019). Towards deep learning models resistant to adversarial attacks.

Mahfuz, R., Sahay, R., and Gamal, A. E. (2021). Mitigating gradient-based adversarial attacks via denoising and compression.

Omari, M. and Yaichi, S. (2015). Image compression based on genetic algorithm optimization. In *2015 2nd World Symposium on Web Applications and Networking (WSWAN)*, pages 1–5.

Ople, J. J. M., Huang, T.-M., Chiu, M.-C., Chen, Y.-L., and Hua, K.-L. (2023). Adjustable model compression using multiple genetic algorithm. *IEEE Transactions on Multimedia*, 25:1125–1132.

Paul, R., Schabath, M., Gillies, R., Hall, L., and Goldgof, D. (2020). Mitigating adversarial attacks on medical image understanding systems. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1517–1521.

Roberto, G. F., Pereira, D. C., Martins, A. S., Tosta, T. A., Soares, C., Lumini, A., Rozendo, G. B., Neves, L. A., and Nascimento, M. Z. (2025). Exploring percolation features with polynomial algorithms for classifying covid-19 in chest x-ray images. *Pattern Recognition Letters*, 189:248–255.

Wu, M.-S. (2014). Genetic algorithm based on discrete wavelet transformation for fractal image compression. *Journal of Visual Communication and Image Representation*, 25(8):1835–1841.

Yao, Q., He, Z., Li, Y., Lin, Y., Ma, K., Zheng, Y., and Zhou, S. K. (2023). Adversarial medical image with hierarchical feature hiding.

Yao, Q., He, Z., and Zhou, S. K. (2022). Medical aegis: Robust adversarial protectors for medical images.

Yao Li, Minhao Cheng, C.-J. H. and Lee, T. C. M. (2022). A review of adversarial attack and defense for classification methods. *The American Statistician*, 76(4):329–345.

Yin, Z., Wang, H., Wang, J., Tang, J., and Wang, W. (2020). Defense against adversarial attacks by low-level image transformations. *International Journal of Intelligent Systems*, 35.

Zhang, S., Gao, H., and Rao, Q. (2021). Defense against adversarial attacks by reconstructing images. *IEEE Transactions on Image Processing*, PP:1–1.