
**Recovering Chest X-RAY Images from
Adversarial Attacks in Pneumonia
Classification: Genetic Algorithm-based
Adaptive Compression (GA-AC)**

Paulo Vitor Costa Lima



UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Uberlândia
2025

Paulo Vitor Costa Lima

**Recovering Chest X-RAY Images from
Adversarial Attacks in Pneumonia
Classification: Genetic Algorithm-based
Adaptive Compression (GA-AC)**

Dissertação de mestrado apresentada ao Programa de Pós-graduação da Faculdade de Computação da Universidade Federal de Uberlândia como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Sistemas de Computação

Orientador: Silvio Ereno Quincozes

Uberlândia
2025

Ficha Catalográfica Online do Sistema de Bibliotecas da UFU
com dados informados pelo(a) próprio(a) autor(a).

L732
2025

Lima, Paulo Vitor Costa, 1996-
Recovering Chest X-RAY Images from Adversarial Attacks in
Pneumonia Classification [recurso eletrônico] : Genetic Algorithm-
based Adaptive Compression (GA-AC) / Paulo Vitor Costa Lima. -
2025.

Orientador: Silvio Ereno Quincozes.
Dissertação (Mestrado) - Universidade Federal de Uberlândia,
Pós-graduação em Ciência da Computação.
Modo de acesso: Internet.
DOI <http://doi.org/10.14393/ufu.di.2025.715>
Inclui bibliografia.
Inclui ilustrações.

1. Computação. I. Quincozes, Silvio Ereno, 1993-, (Orient.). II.
Universidade Federal de Uberlândia. Pós-graduação em Ciência da
Computação. III. Título.

CDU: 681.3

Bibliotecários responsáveis pela estrutura de acordo com o AACR2:
Gizele Cristine Nunes do Couto - CRB6/2091
Nelson Marcos Ferreira - CRB6/3074



ATA DE DEFESA - PÓS-GRADUAÇÃO

Programa de Pós-Graduação em:	Ciência da Computação				
Defesa de:	Dissertação, 52/2025, PPGCO				
Data:	16 de Dezembro de 2025	Hora de início:	09:07	Hora de encerramento:	10:41
Matrícula do Discente:	12412CCP032				
Nome do Discente:	Paulo Vitor Costa Lima				
Título do Trabalho:	Recovering Chest X-RAY Images from Adversarial Attacks in Pneumonia Classification: Genetic Algorithm-based Adaptive Compression (GA-AC)				
Área de concentração:	Ciência da Computação				
Linha de pesquisa:	Sistemas de Computação				
Projeto de Pesquisa de vinculação:	-----				

Reuniu-se por videoconferência, a Banca Examinadora, designada pelo Colegiado do Programa de Pós-graduação em Ciência da Computação, assim composta: Professores Doutores: Marcelo Zanchetta do Nascimento - FACOM/UFU, Daniel Welfer - UFSM/DCOM e Silvio Ereno Quincozes - Unipampa, orientador do(a) candidato(a).

Os examinadores participaram desde as seguintes localidades: Silvio Ereno Quincozes - Alegrete/RS, Daniel Welfer - Santa Maria/Rs. Os outros membros da banca e o aluno participaram da cidade de Uberlândia.

Iniciando os trabalhos o presidente da mesa, Prof. Dr. Silvio Ereno Quincozes, apresentou a Comissão Examinadora e o(á) candidato(a), agradeceu a presença do público, e concedeu ao(á) Discente a palavra para a exposição do seu trabalho

A seguir o senhor presidente concedeu a palavra, pela ordem sucessivamente, aos examinadores, que passaram a arguir o(á) candidato(a). Ultimada a arguição, que se desenvolveu dentro dos termos regimentais, a Banca, em sessão secreta, atribuiu o resultado final, considerando o candidato(a):

Aprovado

Esta defesa faz parte dos requisitos necessários à obtenção do título de Mestre.

O competente diploma será expedido após cumprimento dos demais requisitos, conforme as normas do Programa, a legislação pertinente e a regulamentação interna da UFU.

Nada mais havendo a tratar foram encerrados os trabalhos. Foi lavrada a presente ata que após lida e achada conforme foi assinada pela Banca Examinadora.



Documento assinado eletronicamente por **Marcelo Zanchetta do Nascimento, Professor(a) do Magistério Superior**, em 18/12/2025, às 14:42, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



Documento assinado eletronicamente por **Silvio Ereno Quincozes, Usuário Externo**, em 19/12/2025, às 11:14, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



Documento assinado eletronicamente por **Daniel Welfer, Usuário Externo**, em 22/12/2025, às 14:35, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



A autenticidade deste documento pode ser conferida no site https://www.sei.ufu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **6936728** e o código CRC **BD940359**.

Referência: Processo nº 23117.088419/2025-30

SEI nº 6936728

Este trabalho é dedicado... À minha esposa Quezia, meu amor e meu refúgio, pelo cuidado diário, pela paciência inesgotável e por caminhar ao meu lado em cada etapa desta jornada. Nada disso faria sentido sem você. À minha mãe Eliamar, que hoje descansa em paz, mas que continua presente em cada conquista minha. Foi ela quem acreditou no meu potencial antes mesmo que eu entendesse qual era o meu caminho, quem investiu nos meus estudos e plantou em mim o sonho que hoje se realiza. Esta vitória também é sua. À minha mãe Elismar, que se tornou mãe no gesto, na presença e no amor. Você me ensinou a ler, a estudar, a construir meu futuro com disciplina e esperança. Muito do que sou hoje vem de tudo o que você me deu sem pedir nada em troca. Ao meu pai, Paulo, pelo amor, força e apoio constante ao longo da minha vida. Aos meus queridos irmãos, Vanessa, Daniel e Paulo Henrique, companheiros de tantas fases com quem sempre pude contar, às minhas lindas sobrinhas, Ana Júlia, Isabela, Ana Laura, Alice e Gabriela, e aos meus queridos companheiros de quatro patas, Simba e Nala, que sempre trazem leveza, alegria e motivação aos meus dias. Dedico este trabalho a todos vocês, que formam o alicerce do meu mundo e a razão pela qual sigo em frente com coração grato e cheio de amor.

Agradecimentos

Primeiramente, agradeço a Deus, por ser o meu sustento, pela sabedoria concedida em cada decisão e por iluminar meu caminho nos momentos de incerteza. Sem Sua presença, nada disso seria possível.

Agradeço profundamente à minha família, alicerce da minha vida. Pelo amor incondicional, pelo apoio constante e por cada palavra de incentivo que me sustentou ao longo desta trajetória. Cada um de vocês contribuiu para que este sonho se tornasse realidade.

Ao meu orientador, Professor Silvio Ereno Quincozes, expresso minha sincera gratidão pela orientação dedicada, pela paciência, pelas conversas que ampliaram minha visão e pelo compromisso em guiar este trabalho com rigor e generosidade. Sua contribuição foi essencial para o meu crescimento acadêmico e pessoal.

A todos que, de alguma forma, fizeram parte deste percurso, deixo meu muito obrigado.

*“Porque Dele, por Ele e para Ele são todas as coisas.”
(Romanos 11:36)*

Resumo

Os ataques adversariais têm se mostrado uma ameaça crítica à confiabilidade de modelos de aprendizado profundo aplicados ao processamento de imagens médicas. Técnicas como o Fast Gradient Sign Method (FGSM), o Iterative Fast Gradient Sign Method (I-FGSM), também conhecido como Basic Iterative Method (BIM), e o Projected Gradient Descent (PGD) podem induzir erros significativos de predição, comprometendo a precisão diagnóstica e a segurança do paciente. Para enfrentar esse desafio, este trabalho propõe uma abordagem inovadora denominada Genetic Algorithm-based Adaptive Compression (GA-AC), desenvolvida para recuperar imagens perturbadas por diferentes tipos de ataques adversariais. O método GA-AC utiliza um algoritmo genético para otimizar parâmetros de compressão de imagens, com o objetivo de maximizar o Peak Signal-to-Noise Ratio (PSNR) e o Structural Similarity Index Measure (SSIM), preservando assim as características diagnósticas essenciais das imagens restauradas. Os experimentos realizados com múltiplas imagens de raios-X demonstraram a eficácia do GA-AC, que foi capaz de restaurar o desempenho dos modelos após diferentes ataques adversariais, recuperando, por exemplo, o F1-score de 24,14% para 98,10% após o ataque FGSM, e apresentando resultados igualmente robustos frente aos ataques I-FGSM, e PGD.

Palavras-chave: Ataques adversariais; Aprendizado profundo; Algoritmo genético; Compressão adaptativa; Imagens médicas..

Resumo

Adversarial attacks have become a critical threat to the reliability of deep learning models applied to medical image processing. Techniques such as the Fast Gradient Sign Method (FGSM), the Iterative Fast Gradient Sign Method (I-FGSM), also known as the Basic Iterative Method (BIM), and the Projected Gradient Descent (PGD) can induce significant prediction errors, compromising diagnostic accuracy and patient safety. To address this challenge, this study proposes an innovative approach called Genetic Algorithm-based Adaptive Compression (GA-AC), developed to recover images perturbed by various types of adversarial attacks. The GA-AC method employs a genetic algorithm to optimize image compression parameters, aiming to maximize the Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index Measure (SSIM), thereby preserving essential diagnostic features in the restored images. Experiments conducted on multiple X-ray images demonstrated the effectiveness of GA-AC, which successfully restored model performance after different adversarial attacks, recovering, for example, the F1-score from 24.14% to 98.10% after the FGSM attack, and achieving similarly robust results against I-FGSM, and PGD.

Palavras-chave: Adversarial attacks; Deep learning; Genetic algorithm; Adaptive compression; Medical imaging..

Lista de ilustrações

Figura 1 – Fluxograma completo da proposta <i>GA-AC</i>	54
Figura 2 – Convergência do algoritmo genético ao longo de 30 gerações. A linha azul representa o fitness médio da população e a linha vermelha o melhor fitness individual.	68
Figura 3 – Matrizes de confusão para o dataset Chest X-ray Pneumonia (a) original sem ataque e (b) após ataque FGSM.	69
Figura 4 – Matrizes de confusão para imagens recuperadas (a) PNG e (b) WebP como estratégias de compressão.	69
Figura 5 – Comparação visual do processo de recuperação.	70

Lista de tabelas

Tabela 1 – Áreas críticas afetadas por ataques adversariais e seus potenciais impactos.	30
Tabela 2 – Resumo das etapas de um <i>Genetic Algorithm</i>	35
Tabela 3 – Comparação entre trabalhos relacionados e a proposta deste estudo. . .	37
Tabela 4 – Parâmetros Principais e Componentes de Avaliação do <i>GA-AC</i>	56
Tabela 5 – Desempenho médio do GA-AC em todos os cenários experimentais . .	63
Tabela 6 – Desempenho do GA-AC contra FGSM no NIH Chest X-rays	64
Tabela 7 – Desempenho do GA-AC contra FGSM no Chest X-ray Pneumonia . .	65
Tabela 8 – Desempenho do GA-AC contra I-FGSM no NIH Chest X-rays	65
Tabela 9 – Desempenho do GA-AC contra I-FGSM no Chest X-ray Pneumonia . .	65
Tabela 10 – Desempenho do GA-AC contra PGD no NIH Chest X-rays	66
Tabela 11 – Desempenho do GA-AC contra PGD no Chest X-ray Pneumonia . . .	66
Tabela 12 – Comparação do GA-AC com métodos convencionais de compressão (FGSM, $\epsilon = 0.05$, NIH Chest X-rays)	67

Lista de siglas

BIM Basic Iterative Method

CE Computação Evolutiva

CNN Convolutional Neural Network

CNNs Convolutional Neural Networks

DCT Transformada Discreta de Cossenos

DEAP Distributed Evolutionary Algorithms in Python

DNN Deep Neural Networks

DL Deep Learning

EP Evolutionary Programming

ES Evolution Strategies

FGSM Fast Gradient Sign Method

GA Genetic Algorithm

GA-AC Genetic Algorithm-based Adaptive Compression

GP Genetic Programming

IA Inteligência Artificial

I-FGSM Iterative Fast Gradient Sign Method

JPEG Joint Photographic Experts Group

ML Machine Learning

MSE Mean Squared Error

PGD Projected Gradient Descent

PNG Portable Network Graphics

PSNR Peak Signal-to-Noise Ratio

SSIM Structural Similarity Index Measure

SVM Support Vector Machines

WebP Web Picture format

Sumário

1	INTRODUÇÃO	18
1.1	Motivação	19
1.2	Objetivos e Desafios da Pesquisa	20
1.3	Hipótese	21
1.4	Contribuições	22
1.5	Organização da Dissertação	23
2	FUNDAMENTAÇÃO TEÓRICA	26
2.1	Machine Learning e Deep Learning Aplicados à Medicina . . .	26
2.1.1	Fundamentos de Machine Learning	26
2.1.2	Deep Learning e Redes Neurais Convolucionais	27
2.1.3	Desafios e Vulnerabilidades em Ambientes Clínicos	28
2.2	Ataques Adversariais e seus Aspectos	28
2.2.1	Definição Formal	29
2.2.2	Métodos de Ataques Adversariais	29
2.2.3	Modelos de Ameaça	30
2.2.4	Impacto em Setores Críticos	30
2.3	Compressão de Imagens como Técnica de Defesa	31
2.3.1	Fundamentos da Compressão de Imagens	31
2.3.2	Compressão como Mecanismo de Defesa Adversarial	32
2.3.3	Métricas de Qualidade em Compressão	32
2.4	Computação Evolutiva e Algoritmos Genéticos	33
2.4.1	Fundamentos da Computação Evolutiva	33
2.4.2	Algoritmos Genéticos	34
2.4.3	Aplicações de GA em Processamento de Imagens	35
3	TRABALHOS RELACIONADOS	37
3.1	Visão Geral da Literatura	37

3.1.1	Detalhamento dos Trabalhos Relacionados	38
4	METODOLOGIA	40
4.1	Planejamento da Pesquisa	40
4.2	Revisão Bibliográfica e Fundamentação Teórica	41
4.2.1	Análise de Ataques Adversariais em Imagens Médicas	41
4.2.2	Levantamento de Técnicas de Defesa	41
4.2.3	Justificativa para o Uso de Algoritmos Genéticos	41
4.3	Conjuntos de Dados Utilizados	42
4.3.1	NIH Chest X-rays Dataset	42
4.3.2	Chest X-ray Pneumonia Dataset	42
4.3.3	Pré-processamento dos Dados	43
4.4	Desenvolvimento do Sistema GA-AC	43
4.4.1	Arquitetura do Sistema	43
4.4.2	Implementação do Modelo Base de Classificação	44
4.4.3	Implementação dos Ataques Adversariais	44
4.4.4	Desenvolvimento do Algoritmo Genético	45
4.4.5	Função de Avaliação	46
4.5	Ambiente de Experimentação	47
4.5.1	Infraestrutura Computacional	47
4.5.2	Ambiente de Software	47
4.5.3	Configuração Experimental	48
4.6	Protocolo Experimental	48
4.6.1	Fase 1: Estabelecimento do Baseline	48
4.6.2	Fase 2: Avaliação da Vulnerabilidade a Ataques	49
4.6.3	Fase 3: Otimização e Validação do GA-AC	49
4.7	Considerações Éticas	50
4.7.1	Uso de Dados Médicos	50
4.7.2	Implicações Clínicas	50
4.7.3	Responsabilidade na Divulgação	51
4.8	Resumo da Metodologia	51
5	PROPOSTA DO GA-AC	52
5.1	Objetivos e Motivação	52
5.1.1	Objetivos Específicos	52
5.1.2	Diferenciais em Relação às Abordagens Existentes	53
5.2	Visão Geral: Componentes e Pipeline	53
5.2.1	Fase 1: Pré-processamento e Preparação	54
5.2.2	Fase 2: Geração de Ataques Adversariais	54
5.2.3	Fase 3: Recuperação com GA-AC	55

5.3	Algoritmo Genético Proposto	55
5.3.1	Descrição do Algoritmo	55
5.3.2	Processo Evolutivo	56
5.3.3	Características Distintivas	57
5.4	Tecnologias e Bibliotecas Utilizadas	58
5.4.1	Framework Core de Desenvolvimento	58
5.4.2	Bibliotecas de Processamento de Imagens	58
5.4.3	Bibliotecas de Análise Científica	58
5.5	Contribuições para Robustez de Sistemas de Diagnóstico	59
5.5.1	Adaptabilidade Contextual	59
5.5.2	Balanceamento Multi-objetivos Otimizado	59
5.5.3	Impacto na Confiabilidade Clínica	60
5.6	Limitações e Trabalhos Futuros	60
5.6.1	Limitações Atuais	60
5.6.2	Extensões Futuras	61
5.7	Resumo da Proposta	61
6	EXPERIMENTOS E ANÁLISE DOS RESULTADOS	63
6.1	Desempenho Geral do GA-AC	63
6.1.1	Resultados por Tipo de Ataque	64
6.1.2	Comparação com Métodos Convencionais de Compressão	67
6.1.3	Análise de Convergência do Algoritmo Genético	67
6.1.4	Matrizes de Confusão	68
6.1.5	Análise Visual Qualitativa	70
6.2	Discussão	71
6.2.1	Principais Achados	71
6.2.2	Implicações Práticas	72
6.2.3	Limitações e Considerações	73
6.3	Síntese do Capítulo	74
7	CONCLUSÃO	75
7.1	Principais Contribuições Científicas	75
7.2	Implicações para a Prática Clínica	76
7.3	Limitações e Considerações Críticas	77
7.4	Perspectivas e Direções Futuras	78
7.4.1	Extensão para Múltiplas Modalidades	78
7.4.2	Integração com Treinamento Adversarial	78
7.4.3	Desenvolvimento de Sistemas Adaptativos	78
7.4.4	Defesa em Tempo Real	79
7.4.5	Validação Clínica Prospectiva	79

7.5	Considerações Finais	79
	REFERÊNCIAS	81

Introdução

O aprendizado profundo tem se tornado uma poderosa ferramenta para o processamento de imagens médicas, auxiliando na detecção e diagnóstico de doenças críticas (AGGARWAL et al., 2021). Entretanto, observa-se atualmente um crescimento considerável na incidência de ataques cibernéticos em diversas áreas, e o setor da saúde certamente tem sido impactado por isso, especialmente devido aos efeitos causados em modelos que classificam imagens com o objetivo de realizar diagnósticos (BORTSOVA et al., 2021). Um dos principais agentes ofensores nesses sistemas são os chamados ataques adversariais, que inserem pequenas perturbações nas imagens a ponto de induzir erros nos modelos, resultando em diagnósticos incorretos (LI MINHAO CHENG; LEE, 2022). No contexto de *Healthcare 5.0*, em que a inteligência artificial (IA) desempenha um papel fundamental na evolução do setor (GOMATHI; MISHRA; TYAGI, 2023), as preocupações crescem à medida que a tecnologia avança.

É fundamental ser capaz de recuperar imagens médicas que sofreram ataques adversariais, primeiramente, porque tais ataques levam a previsões incorretas dos modelos, o que consequentemente gera diagnósticos errôneos, impactando diretamente o tratamento dos pacientes. Além disso, em situações nas quais as imagens não possuem cópias correspondentes, técnicas de recuperação tornam-se ainda mais relevantes para restaurar suas propriedades originais.

Embora os sistemas médicos geralmente sejam protegidos por protocolos seguros de aquisição e transmissão, cenários reais ainda os expõem a vulnerabilidades adversariais. Por exemplo, ataques do tipo *black-box* não exigem acesso direto ao modelo alvo. Em vez disso, exploram o comportamento de entrada e saída compartilhado, treinando modelos substitutos a partir de pares imagem-rótulo disponíveis, o que permite que exemplos adversariais baseados em transferência ou livres de gradiente tenham sucesso mesmo entre arquiteturas diferentes (INKAWHICH et al., 2020).

Além disso, estudos recentes demonstraram ataques adversariais clinicamente orientados, como o *CoRPA* (*Concept-Based Report Perturbation Attack*), que geram perturbações conscientes do domínio e alinhadas a características patológicas. Essas perturbações po-

dem ocultar ou simular achados como nódulos ou infiltrados, afetando simultaneamente a imagem e o respectivo relatório clínico sem serem perceptíveis ao olho humano (RAFFERTY; RAMAESH; RAJAN, 2025).

Portanto, apesar das práticas de segurança robustas, essas ameaças evidenciam que imagens médicas manipuladas adversarialmente ainda podem ser introduzidas em fluxos de diagnóstico, especialmente em consultas remotas, plataformas de *AI-as-a-service* ou sistemas móveis de aquisição.

Os ataques adversariais exploram vulnerabilidades intrínsecas das redes neurais profundas (DNN), introduzindo perturbações sutis e imperceptíveis ao olho humano em imagens médicas, o que degrada significativamente o desempenho dos classificadores diagnósticos, mesmo sob intenso treinamento (GOODFELLOW; SHLENS; SZEGEDY, 2014; CARLINI; WAGNER, 2017). Entre os ataques mais conhecidos estão o FGSM, o I-FGSM, também chamado BIM, e o PGD, que alteram o conteúdo da imagem de forma contextual e realista, preservando sua característica visual.

Para mitigar tais ameaças, técnicas tradicionais de compressão de imagens, como JPEG e WebP, têm sido exploradas como mecanismos de defesa em nível baixo. Trabalhos iniciais (DZIUGAITE; GHAHRAMANI; ROY, 2016; DAS et al., 2018) mostraram evidências do potencial dessas abordagens em reduzir o impacto adversarial, porém sua eficácia permanece limitada por parâmetros de compressão fixos, frequentemente resultando em restauração subótima e confiabilidade diagnóstica comprometida. Esses estudos refletem o estágio inicial das pesquisas nessa área e podem não capturar avanços recentes em métodos de recuperação adaptativos e baseados em aprendizado. Essa lacuna evidencia a necessidade de estratégias de defesa mais robustas e ajustáveis.

Nesse contexto, este estudo propõe uma abordagem inovadora de recuperação adaptativa denominada GA-AC, que integra algoritmos genéticos à compressão de imagens, visando otimizar dinamicamente a qualidade da restauração frente às perturbações adversariais. O método busca ajustar múltiplos parâmetros, incluindo compressão em diferentes formatos, qualidade, intensidade de espelhamento (*flip*), ângulo de rotação, fator de escala e operações de filtragem, de forma adaptativa. Assim, o GA-AC permite recuperar eficientemente imagens médicas atacadas, podendo ser aplicado em outras áreas.

Para a validação, foi utilizado dois conjuntos de imagens de raio-X de tórax (KERMANY; ZHANG; GOLDBAUM, 2018) (WANG et al., 2017) destinado ao treinamento, teste e validação. O uso de algoritmos genéticos possibilitou aprimorar a compressão, aumentando as taxas de acerto do modelo classificador.

1.1 Motivação

O uso de Inteligência Artificial (IA) e Deep Learning (DL) em aplicações médicas representa um avanço significativo na precisão e agilidade de diagnósticos clínicos. Contudo,

a dependência crescente desses sistemas críticos expõe uma vulnerabilidade preocupante: a suscetibilidade a ataques adversariais. Esses ataques, mesmo quando imperceptíveis ao olho humano, têm o potencial de induzir erros graves em classificadores de imagens médicas, comprometendo decisões clínicas e a segurança do paciente.

Apesar de sua relevância, observa-se que grande parte dos esforços na literatura tem se concentrado em detectar ou mitigar ataques adversariais durante o processo de treinamento, negligenciando cenários nos quais o dano já ocorreu e a imagem precisa ser recuperada. Além disso, muitas estratégias defensivas aplicadas em imagens médicas se baseiam em parâmetros fixos ou métodos não adaptativos, que carecem de flexibilidade para lidar com a variabilidade das perturbações adversariais, especialmente nos ataques iterativos I-FGSM, e PGD.

Outra limitação recorrente é que abordagens tradicionais, como compressões Joint Photographic Experts Group (JPEG) e Web Picture format (WebP), embora simples e eficazes em alguns casos, não exploram o potencial de ajuste dinâmico que a natureza das perturbações exige. Em ambientes reais, cada imagem pode reagir de forma distinta à compressão e filtragem, tornando necessária uma solução que aprenda a adaptar-se ao contexto e à intensidade da distorção.

Dessa forma, a motivação central desta pesquisa reside na necessidade de um mecanismo adaptativo e inteligente para restaurar imagens médicas comprometidas por ataques adversariais, sem sacrificar a integridade diagnóstica. A proposta do método GA-AC surge como resposta a essa lacuna, ao empregar algoritmos genéticos para otimizar automaticamente múltiplos parâmetros de compressão e transformação de imagem. Essa abordagem busca equilibrar duas dimensões essenciais: a recuperação da qualidade visual e a restauração da confiabilidade dos modelos de classificação.

Assim, o trabalho se fundamenta na premissa de que estratégias dinâmicas, orientadas por otimização evolutiva e adaptabilidade contextual, podem oferecer resiliência efetiva contra ataques adversariais em sistemas de diagnóstico assistido por IA, promovendo maior segurança e robustez em ambientes médicos reais.

1.2 Objetivos e Desafios da Pesquisa

O presente trabalho tem como principal objetivo desenvolver uma abordagem adaptativa e inteligente capaz de restaurar imagens médicas comprometidas por ataques adversariais, preservando tanto a fidelidade visual quanto a confiabilidade diagnóstica dos modelos de DL. A crescente dependência das DNN em sistemas de diagnóstico clínico tem exposto vulnerabilidades significativas frente a ataques como o FGSM, o I-FGSM ou BIM, e o PGD. Tais perturbações, embora imperceptíveis ao olho humano, podem induzir classificadores a erros graves, comprometendo diretamente o processo de diagnóstico e a segurança do paciente.

Diante desse cenário, propõe-se a metodologia GA-AC, que integra algoritmos genéticos à compressão adaptativa de imagens, ajustando automaticamente parâmetros como taxa de compressão (WebP e Portable Network Graphics (PNG)), rotação, escala, intensidade de espelhamento e aplicação de filtros espaciais. O objetivo é maximizar simultaneamente a similaridade estrutural, medida por métricas como Peak Signal-to-Noise Ratio (PSNR) e Structural Similarity Index Measure (SSIM), e a confiabilidade do classificador baseado em redes convolucionais Convolutional Neural Network (CNN).

A proposta busca preencher uma lacuna nas estratégias de defesa contra ataques adversariais em imagens médicas, uma vez que os métodos tradicionais baseados em compressão utilizam parâmetros fixos e carecem de adaptabilidade diante da diversidade e da intensidade das perturbações. Com a otimização evolutiva conduzida pelo GA-AC, espera-se alcançar uma restauração eficiente, mesmo em cenários complexos e sob múltiplos tipos de ataque, recuperando a acurácia e o *F1-score* originais do modelo classificador.

Entre os principais desafios da pesquisa destacam-se: (i) equilibrar múltiplos critérios conflitantes no processo de recuperação, como fidelidade visual e robustez classificatória; (ii) adaptar o algoritmo genético a um domínio sensível como o da imagem médica, garantindo a preservação das informações diagnósticas; (iii) assegurar eficiência computacional para viabilizar a aplicação do método em contextos clínicos reais; e (iv) validar experimentalmente a eficácia e reprodutibilidade do GA-AC frente a diferentes tipos e intensidades de ataques adversariais.

Em síntese, o trabalho propõe uma solução adaptativa, multiobjetiva e evolutiva que busca reforçar a segurança e a confiabilidade de sistemas médicos baseados em aprendizado profundo, contribuindo para a mitigação de riscos e para o avanço de práticas seguras em ambientes de *Healthcare 5.0*.

1.3 Hipótese

A hipótese central deste trabalho é que é possível recuperar imagens médicas comprometidas por ataques adversariais por meio de uma abordagem de compressão adaptativa otimizada por algoritmos genéticos GA, preservando a qualidade diagnóstica e restaurando a confiabilidade da classificação em modelos de DL. Essa abordagem, denominada GA-AC, busca combinar otimização evolutiva e compressão de imagem de forma dinâmica, ajustando seus parâmetros conforme o tipo e a intensidade da perturbação adversarial.

Para investigar essa hipótese principal, derivam-se as seguintes hipóteses específicas:

- H1. Ataques adversariais, como FGSM, I-FGSM ou BIM, PGD, podem ser mitigados por meio de transformações de compressão de imagem otimizadas adaptativamente.
- H2. Algoritmos genéticos são capazes de encontrar combinações ótimas de parâmetros de transformação — incluindo qualidade de compressão, rotação, escala, intensidade de

espelhamento e aplicação de filtros espaciais — que maximizem simultaneamente a qualidade visual (métrica PSNR) e a similaridade estrutural (SSIM), além de restaurar a confiabilidade da classificação do modelo.

- H3. A abordagem GA-AC supera métodos de defesa baseados em compressão estática, oferecendo uma solução adaptativa, escalável e eficiente frente a diferentes tipos e intensidades de ataques adversariais.
- H4. A integridade diagnóstica das imagens médicas será preservada, mantendo as características essenciais para a tomada de decisão clínica e garantindo a segurança do paciente.

Essa hipótese é particularmente relevante no contexto médico, em que a preservação da qualidade diagnóstica é fundamental. A validação de que a recuperação de imagens atacadas pode ser conduzida de forma adaptativa e otimizada representa um avanço significativo na segurança de sistemas baseados em inteligência artificial aplicados à área da saúde.

1.4 Contribuições

As principais contribuições deste trabalho são:

- ❑ **Proposição de uma abordagem adaptativa para recuperação adversarial:** Esta pesquisa introduz o método GA-AC, que utiliza algoritmos genéticos para otimizar de forma dinâmica múltiplos parâmetros de recuperação de imagens médicas afetadas por ataques adversariais. Diferentemente de métodos convencionais que aplicam compressões fixas (como JPEG, WebP, PNG), o GA-AC adapta-se às características específicas de cada imagem e tipo de perturbação, promovendo uma restauração mais eficaz e generalizável;
- ❑ **Formulação teórica da recuperação adversarial como problema de otimização multiobjetiva:** O trabalho propõe uma nova perspectiva sobre o processo de recuperação, tratando-o como uma tarefa de otimização simultânea de métricas de qualidade visual e de confiabilidade diagnóstica. Essa formulação multiobjetiva amplia a aplicabilidade da técnica a contextos clínicos complexos, permitindo equilibrar a preservação da fidelidade visual com a manutenção da precisão diagnóstica;
- ❑ **Integração de múltiplos tipos de ataques adversariais:** A metodologia proposta foi desenvolvida para lidar com diferentes classes de ataques, incluindo FGSM e I-FGSM, e PGD, reforçando sua robustez e aplicabilidade a cenários reais de vulnerabilidade em sistemas de diagnóstico assistido por IA;

- ❑ **Avanço na interseção entre computação evolutiva e segurança em IA médica:** Ao aplicar princípios de evolução artificial na mitigação de ataques adversariais, este trabalho amplia o escopo de uso dos algoritmos genéticos para além da otimização convencional, explorando seu potencial em problemas de defesa e recuperação inteligente em imagens sensíveis;
- ❑ **Disponibilização de base metodológica reprodutível e aberta:** Todo o processo de desenvolvimento, desde a implementação do GA-AC até a configuração dos ataques adversariais, é documentado e disponibilizado de forma a garantir a transparência científica e facilitar a replicação por outros pesquisadores interessados em segurança e confiabilidade de modelos de visão médica;
- ❑ **Abertura de novas oportunidades de pesquisa:** A proposta GA-AC abre caminhos para investigações futuras envolvendo a integração de técnicas evolutivas com aprendizado profundo, a expansão para outros domínios de imagem biomédica (como tomografia e ressonância magnética) e o estudo de mecanismos de defesa autoadaptativos em tempo real, promovendo avanços na confiabilidade e segurança de sistemas de inteligência artificial aplicados à medicina.

1.5 Organização da Dissertação

Esta dissertação está estruturada em seis capítulos, conforme descrito a seguir:

- ❑ **Capítulo 1: Introdução** – Neste capítulo, foram apresentados o tema da pesquisa, a motivação que a justifica, os objetivos do estudo, a hipótese a ser testada, as contribuições esperadas e a organização da dissertação.
- ❑ **Capítulo 2: Fundamentação Teórica** – Este capítulo apresenta os conceitos fundamentais que sustentam o desenvolvimento da pesquisa, iniciando com uma introdução aos princípios de Machine Learning (ML) e DL aplicados à medicina, destacando o uso de CNN em tarefas de diagnóstico por imagem. Em seguida, são discutidos os principais tipos de ataques adversariais, como I-FGSM e PGD, enfatizando seus impactos na confiabilidade dos modelos de aprendizado profundo. O capítulo também aborda os fundamentos da compressão de imagens e suas aplicações como técnica de defesa, além de explorar os princípios da computação evolutiva, com foco nos algoritmos genéticos e na otimização multiobjetiva.
- ❑ **Capítulo 3: Trabalhos Relacionados** - São apresentados estudos correlatos sobre métodos de recuperação adversarial, evidenciando as lacunas existentes e o potencial da abordagem proposta GA-AC como solução adaptativa para restauração de imagens médicas comprometidas por perturbações adversariais.

- ❑ **Capítulo 4: Metodologia** – Este capítulo descreve a metodologia adotada para o desenvolvimento e avaliação da proposta GA-AC, detalhando cada etapa do processo experimental. Inicialmente, são apresentadas as fases de planejamento e revisão bibliográfica que fundamentaram a escolha dos ataques adversariais utilizados. Em seguida, o capítulo aborda o processo de implementação do algoritmo genético, especificando os parâmetros populacionais, operadores evolutivos e funções de avaliação baseadas em métricas como PSNR, SSIM, acurácia e *F1-score*. Também são descritos o ambiente de experimentação, os conjuntos de dados de imagens médicas empregados e os procedimentos de validação dos resultados. Por fim, o capítulo apresenta a metodologia de análise comparativa entre o GA-AC e técnicas convencionais de compressão, assegurando uma avaliação sistemática, reproduzível e alinhada aos objetivos da pesquisa.
- ❑ **Capítulo 5: Proposta** – Este capítulo apresenta a proposta do método GA-AC, descrevendo seus objetivos, arquitetura conceitual e os mecanismos que o diferenciam das abordagens tradicionais de recuperação adversarial. São detalhados o fluxo de processamento das imagens, os parâmetros adaptativos otimizados pelo algoritmo genético e o processo de integração entre as etapas de compressão, avaliação e ajuste evolutivo. O capítulo também expõe as tecnologias e bibliotecas empregadas no desenvolvimento experimental, além do método e do meta-método utilizados para conduzir a otimização multiobjetiva. Por fim, discute-se como a proposta contribui para o avanço da robustez e confiabilidade de sistemas de diagnóstico baseados em aprendizado profundo diante de ataques adversariais.
- ❑ **Capítulo 6: Experimentos e Análise dos Resultados** – Este capítulo apresenta os experimentos realizados para avaliar a eficácia da proposta GA-AC na recuperação de imagens médicas adversarialmente perturbadas. São descritas as configurações dos ataques aplicados, bem como o ambiente computacional e os parâmetros utilizados nos testes. O método de avaliação considera métricas de similaridade visual (PSNR e SSIM) e de desempenho de classificação (acurácia e *F1-score*), permitindo uma análise abrangente da recuperação em múltiplas dimensões. O capítulo também discute comparações com técnicas convencionais de compressão e apresenta uma análise crítica sobre a robustez, estabilidade e generalização do método proposto frente a diferentes tipos de perturbações.
- ❑ **Capítulo 7: Conclusão** – O capítulo final sintetiza os principais resultados e contribuições da pesquisa, destacando o papel do método GA-AC como uma abordagem inovadora para defesa e recuperação em imagens médicas afetadas por ataques adversariais. São discutidas as implicações científicas e práticas do estudo, suas limitações e possíveis aprimoramentos metodológicos. Por fim, o capítulo apresenta perspectivas para trabalhos futuros, incluindo a extensão do modelo a outros tipos

de exames médicos, a integração com redes neurais robustas e o desenvolvimento de sistemas adaptativos de defesa em tempo real, ampliando o potencial de aplicação da pesquisa em contextos clínicos e de segurança em IA.

Fundamentação Teórica

Este capítulo apresenta os conceitos essenciais para o desenvolvimento desta pesquisa. Inicialmente, discutem-se os fundamentos de ML e DL aplicados ao diagnóstico por imagem na área médica, destacando a relevância das CNN. Em seguida, serão abordados os ataques adversariais, suas técnicas mais comuns e ameaças associadas. O capítulo também apresenta os fundamentos da compressão de imagens e suas aplicações como técnica de defesa adversarial, seguido por uma discussão sobre computação evolutiva, com foco nos algoritmos genéticos (AG) e otimização multiobjetivo, como estratégias adaptativas para recuperação adversarial.

2.1 Machine Learning e Deep Learning Aplicados à Medicina

A aplicação de técnicas de ML e DL revolucionou o diagnóstico médico por imagem nas últimas décadas, permitindo a automação e o aprimoramento de tarefas que tradicionalmente exigiam especialização humana intensiva (LECUN; BENGIO; HINTON, 2015; LITJENS et al., 2017). Esta seção apresenta os fundamentos dessas técnicas e seu papel no contexto clínico, enfatizando tanto suas capacidades quanto suas vulnerabilidades.

2.1.1 Fundamentos de Machine Learning

Machine Learning se trata de um conjunto de métodos computacionais que permitem a sistemas aprenderem padrões a partir dos dados fornecidos (BISHOP, 2006). No contexto médico, algoritmos de ML são empregados para extrair características discriminativas de imagens, viabilizando diagnósticos automatizados e sistemas de apoio à decisão clínica aos médicos.

Os paradigmas de aprendizado incluem três categorias principais, o aprendizado supervisionado, não supervisionado e por reforço.

No aprendizado supervisionado, o modelo é treinado com dados rotulados, aprendendo a mapear entradas (imagens médicas) para saídas conhecidas (diagnósticos). Algoritmos como Support Vector Machines (SVM) e *Random Forests* exemplificam essa abordagem (VAPNIK, 1995).

No aprendizado não supervisionado, o objetivo é identificar padrões e estruturas em dados não rotulados, sendo particularmente útil para descoberta de subgrupos de pacientes ou detecção de anomalias (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Por fim, o aprendizado por reforço consiste em um agente que aprende através de interações com o ambiente, recebendo recompensas ou penalidades conforme suas ações (SUTTON; BARTO, 2018).

2.1.2 Deep Learning e Redes Neurais Convolucionais

Deep Learning representa um subconjunto de ML caracterizado pelo uso de arquiteturas neurais profundas, compostas por múltiplas camadas de processamento hierárquico (LECUN; BENGIO; HINTON, 2015). Ao contrário de métodos tradicionais que requerem engenharia manual de características, DL aprende representações automaticamente dos dados brutos, tornando-se extremamente poderoso para tarefas de visão computacional.

2.1.2.1 Arquitetura de Redes Neurais Convolucionais

As CNNs constituem a arquitetura dominante em análise de imagens médicas. Inspiradas na organização do córtex visual, CNNs exploram propriedades como localidade espacial, compartilhamento de parâmetros e invariância a translações (LECUN et al., 1998).

Uma CNN típica é composta por camadas convolucionais, onde é aplicado filtros aprendidos que detectam padrões locais (bordas, texturas, formas); como também das chamadas funções de ativação, em que são introduzidas não-linearidades essenciais para relações complexas. As CNNs são compostas por fim pelas camadas de *pooling*, que reduzem a dimensionalidade espacial a fim de reduzir o custo computacional e também pelas camadas conectadas, que agregam características extraídas para reduzir a saída final, seja uma classificação ou segmentação.

2.1.2.2 Aplicações em Diagnóstico Médico

CNNs demonstraram desempenho comparável ou superior ao de especialistas humanos em diversas tarefas diagnósticas. Na classificação, essas redes permitem a identificação de patologias em radiografias torácicas, mamografias e imagens dermatológicas (ESTEVA et al., 2017). Já na detecção, são utilizadas para localizar lesões, nódulos pulmonares

e metástases com alta sensibilidade (LIU et al., 2017). Além disso, na tarefa de segmentação, as CNNs possibilitam o delineamento preciso de órgãos, tumores e estruturas anatômicas, contribuindo para o planejamento cirúrgico e radioterápico (RONNEBERGER; FISCHER; BROX, 2015).

Arquiteturas notáveis incluem AlexNet (KRIZHEVSKY; SUTSKEVER; HINTON, 2012), ResNet (HE et al., 2016) e U-Net (RONNEBERGER; FISCHER; BROX, 2015), cada uma introduzindo inovações que aprimoraram a capacidade de generalização e extração de características em diferentes escalas.

2.1.3 Desafios e Vulnerabilidades em Ambientes Clínicos

Apesar dos avanços, a aplicação de DL em medicina enfrenta desafios críticos relacionados à robustez e confiabilidade. Um deles é a sensibilidade a perturbações, em que pequenas alterações imperceptíveis podem induzir erros catastróficos de classificação, como demonstrado por ataques adversariais (FINLAYSON et al., 2019). Outro desafio importante refere-se à variabilidade de aquisição: diferenças em protocolos de imagem, equipamentos e condições de captura tendem a degradar significativamente o desempenho de modelos treinados em ambientes controlados (BORTSOVA et al., 2021).

A interpretabilidade limitada também se destaca, uma vez que a natureza de “caixa-preta” das redes profundas dificulta a compreensão dos critérios decisórios, fator essencial para aceitação clínica e regulatória (RUDIN, 2019). Além disso, há os requisitos de dados, já que o treinamento eficaz de modelos robustos demanda grandes volumes de dados rotulados por especialistas, o que frequentemente representa uma barreira prática para sua obtenção.

A vulnerabilidade a ataques adversariais representa uma ameaça preocupante em aplicações de saúde, onde decisões diagnósticas incorretas podem ter consequências diretas sobre a vida dos pacientes. Essa fragilidade motiva o desenvolvimento de mecanismos de defesa robustos e adaptativos, como a abordagem GA-AC proposta neste trabalho, capaz de restaurar a integridade de imagens comprometidas sem comprometer características diagnósticas essenciais.

Tendo a compreensão dos fundamentos e o potencial dos modelos de *deep learning* em medicina, torna-se essencial investigar suas vulnerabilidades. A seção seguinte examina ataques adversariais, perturbações intencionais que exploram fragilidades inerentes desses sistemas, representando ameaças significativas à segurança diagnóstica.

2.2 Ataques Adversariais e seus Aspectos

Os *ataques adversariais* consistem em perturbações cuidadosamente projetadas em entradas de modelos de aprendizado de máquina, capazes de induzir erros de classifi-

cação mesmo quando imperceptíveis aos seres humanos. Esses ataques expõem vulnerabilidades expressivas nos sistemas de inteligência artificial baseados em redes neurais profundas (YIN et al., 2020; CHAKRABORTY et al., 2021).

2.2.1 Definição Formal

Seja x uma entrada legítima com rótulo y , e $f(\cdot)$ um classificador treinado. O objetivo do atacante é encontrar uma perturbação δ tal que:

$$f(x + \delta) \neq y, \quad \text{sujeito a } \|\delta\|_p \leq \epsilon, \quad (1)$$

onde ϵ controla a intensidade da perturbação e p determina a norma adotada, comum em cenários que consideram restrições visuais (MADRY et al., 2019).

2.2.2 Métodos de Ataques Adversariais

2.2.2.1 Fast Gradient Sign Method (FGSM)

O FGSM, proposto por YIN et al. (YIN et al., 2020), utiliza o gradiente da função de perda para criar uma perturbação adversarial em uma única etapa:

$$x' = x + \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y)) \quad (2)$$

onde $J(\theta, x, y)$ representa a função de perda do modelo com parâmetros θ . O método é computacionalmente eficiente e amplamente utilizado tanto em ataques quanto em treinamento adversarial.

2.2.2.2 Iterative Fast Gradient Sign Method (I-FGSM)

Em outro trabalho, KURAKIN; GOODFELLOW; BENGIO (KURAKIN; GOODFELLOW; BENGIO, 2017) propuseram uma extensão iterativa do FGSM, também conhecida como BIM. O ataque é atualizado de forma incremental:

$$x^{(t+1)} = \text{Clip}_{x,\epsilon} \left(x^{(t)} + \alpha \cdot \text{sign}(\nabla_x J(\theta, x^{(t)}, y)) \right) \quad (3)$$

onde α é o tamanho do passo e a função Clip garante que a perturbação permaneça dentro da região ϵ permitida. O I-FGSM tende a ser mais eficaz que o FGSM, principalmente quando ϵ é pequeno.

2.2.2.3 Projected Gradient Descent (PGD)

O ataque PGD (MADRY et al., 2019), foi considerado um dos mais fortes. O método refina o I-FGSM com reinicializações aleatórias e projeção contínua em uma região de perturbação:

$$x^{(t+1)} = \Pi_{B_\epsilon(x)} \left(x^{(t)} + \alpha \cdot \text{sign}(\nabla_x J(\theta, x^{(t)}, y)) \right) \quad (4)$$

onde $\Pi_{B_\epsilon(x)}$ denota a projeção no círculo de raio ϵ centrada em x . O PGD é um dos principais ataques utilizados para avaliação de robustez de sistemas de defesa.

2.2.3 Modelos de Ameaça

Ataques adversariais podem ser classificados quanto ao conhecimento do atacante. No cenário *white-box*, o atacante possui acesso total à arquitetura do modelo, seus pesos e gradientes, permitindo a construção de ataques altamente otimizados. Em contrapartida, no cenário *black-box*, o atacante não tem acesso direto ao modelo, explorando a transferibilidade de exemplos adversariais entre modelos ou realizando consultas à API (COSTA et al., 2024).

Além dessas categorias, os ataques podem ser não-direcionados, quando o objetivo é simplesmente forçar qualquer classificação incorreta, ou direcionados, quando buscam produzir especificamente uma classe-alvo. Estes últimos apresentam maior complexidade e podem ser ainda mais perigosos em contextos médicos, nos quais a manipulação específica do diagnóstico pode causar danos significativos.

2.2.4 Impacto em Setores Críticos

Diversas áreas podem ser afetadas por ataques adversariais, conforme apresentado na Tabela 1. Esses ataques podem gerar consequências graves em sistemas que dependem de modelos de inteligência artificial para tomada de decisão.

Área	Impacto Potencial
Medicina	Diagnósticos incorretos, classificação errada de patologias, comprometimento de sistemas de apoio à decisão clínica
Veículos Autônomos	Falhas na detecção de sinais de trânsito, pedestres ou obstáculos, potencialmente causando acidentes
Segurança	Bypass de sistemas de reconhecimento facial, autenticação biométrica comprometida
Defesa e Militar	Falhas em sistemas de reconhecimento de alvos, comprometimento de vigilância automatizada
Indústria	Falhas em sistemas de controle de qualidade automatizado, comprometimento de inspeção visual

Tabela 1 – Áreas críticas afetadas por ataques adversariais e seus potenciais impactos.

Diversas áreas podem ser afetadas por ataques adversariais, conforme apresentado na Tabela 1. Esses ataques podem gerar consequências graves em sistemas que dependem de modelos de inteligência artificial para tomada de decisão, como já demonstrado em estudos envolvendo, medicina (PASCHALI et al., 2018), veículos autônomos (EYKHOLT et al., 2018), segurança biométrica (SHARIF et al., 2016), sensoriamento remoto e aplicações militares (LI et al., 2019), além de inspeção industrial (REDDY; RAMALINGAM; GALLO, 2020).

A compreensão profunda dos mecanismos e impactos de ataques adversariais é fundamental para o desenvolvimento de estratégias de defesa eficazes, especialmente em domínios sensíveis como a medicina, onde a confiabilidade dos sistemas automatizados é imperativa.

Identificadas as ameaças impostas por ataques adversariais, é necessário explorar mecanismos de defesa capazes de mitigar seus efeitos. A compressão de imagens emerge como uma técnica promissora, atuando como filtro que pode atenuar perturbações adversariais enquanto preserva características diagnósticas relevantes.

2.3 Compressão de Imagens como Técnica de Defesa

A compressão de imagens é um processo fundamental em sistemas de processamento digital, visando reduzir a quantidade de dados necessários para representar uma imagem, mantendo características relevantes. No contexto de defesas adversariais, a compressão atua como um mecanismo de filtragem que pode atenuar perturbações adversariais introduzidas em imagens médicas (DAS et al., 2018).

2.3.1 Fundamentos da Compressão de Imagens

As técnicas de compressão podem ser divididas em duas categorias principais: compressão sem perda (*lossless*) e compressão com perda (*lossy*). A compressão sem perda permite a reconstrução exata da imagem original, sendo essencial em aplicações onde a integridade absoluta dos dados é crítica. Métodos como PNG são exemplos dessa categoria (SALOMON, 2007).

Por outro lado, a compressão com perda remove informações consideradas menos perceptíveis ao olho humano, alcançando taxas de compressão maiores. O formato JPEG, baseado na Transformada Discreta de Cossenos (DCT), é amplamente utilizado e particularmente relevante no contexto de defesas adversariais devido à sua capacidade de suavizar artefatos de alta frequência (WALLACE, 1992).

A DCT transforma blocos de pixels do domínio espacial para o domínio de frequência, permitindo que componentes de alta frequência, onde perturbações adversariais frequentemente se manifestam, sejam quantizadas de forma mais agressiva.

2.3.2 Compressão como Mecanismo de Defesa Adversarial

A eficácia da compressão contra ataques adversariais baseia-se na observação de que perturbações adversariais, mesmo que imperceptíveis, frequentemente concentram-se em componentes de alta frequência da imagem (DAS et al., 2018). O processo de compressão com perda naturalmente atenua essas componentes, potencialmente removendo ou reduzindo o impacto das perturbações maliciosas.

Diversos estudos demonstram que a compressão JPEG pode mitigar ataques como FGSM e PGD, restaurando parcialmente a acurácia de classificadores comprometidos (MAH-FUZ; SAHAY; GAMAL, 2021; ZHANG; GAO; RAO, 2021). No entanto, essa abordagem apresenta limitações significativas. Uma delas é o *trade-off* qualidade-defesa, pois taxas de compressão elevadas podem degradar características diagnósticas essenciais em imagens médicas, comprometendo sua utilidade clínica. Outra limitação reside na adaptabilidade limitada, uma vez que parâmetros fixos de compressão não se ajustam à intensidade ou ao tipo de ataque adversarial, reduzindo a eficácia da defesa. Soma-se a isso a vulnerabilidade a ataques adaptativos, já que atacantes conscientes da aplicação da compressão podem projetar perturbações resistentes ao processo, explorando características que permanecem preservadas após a transformação (HIRANO; MINAGI; TAKEMOTO, 2021).

Essas limitações motivam o desenvolvimento de abordagens adaptativas que otimizem dinamicamente os parâmetros de compressão, equilibrando robustez adversarial e preservação da qualidade diagnóstica, objetivo central da metodologia GA-AC proposta neste trabalho.

2.3.3 Métricas de Qualidade em Compressão

Para avaliar o impacto da compressão na qualidade de imagens, especialmente no contexto de recuperação adversarial, duas métricas são amplamente utilizadas:

2.3.3.1 Peak Signal-to-Noise Ratio (PSNR)

O PSNR quantifica a diferença entre a imagem original e a imagem processada em termos Mean Squared Error (MSE):

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{\text{MAX}_I^2}{\text{MSE}} \right) \quad (5)$$

onde MAX_I é o valor máximo possível de intensidade de pixel (255 para imagens de 8 bits), e MSE é definido como:

$$\text{MSE} = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [I(i, j) - K(i, j)]^2 \quad (6)$$

sendo m e n as dimensões da imagem, I a imagem original e K a imagem processada. Valores mais elevados de PSNR (tipicamente acima de 30 dB) indicam maior fidelidade

da imagem reconstruída. No entanto, o PSNR possui limitações em capturar aspectos perceptuais da qualidade visual (KUMAR; KUMAR; KUMAR, 2013).

2.3.3.2 Structural Similarity Index Measure (SSIM)

O SSIM oferece uma avaliação mais alinhada à percepção humana, considerando luminância, contraste e estrutura:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (7)$$

onde μ_x e μ_y são médias locais, σ_x^2 e σ_y^2 são variâncias, σ_{xy} é a covariância, e c_1 , c_2 são constantes de estabilização definidas como $c_1 = (k_1L)^2$ e $c_2 = (k_2L)^2$, com L sendo a faixa dinâmica dos valores de pixel e $k_1, k_2 \ll 1$. Valores de SSIM próximos a 1 indicam alta similaridade estrutural (HORÉ; ZIOU, 2010).

A combinação de PSNR e SSIM permite avaliar tanto a fidelidade numérica quanto a qualidade perceptual, sendo essencial para guiar o processo de otimização no GA-AC. Enquanto o PSNR quantifica a magnitude dos erros introduzidos, o SSIM captura aspectos de degradação estrutural mais relevantes para a interpretação humana e diagnóstica.

Embora a compressão ofereça benefícios defensivos, sua aplicação com parâmetros fixos limita a adaptabilidade a diferentes tipos e intensidades de ataques. Para superar essa limitação, técnicas de otimização adaptativa são necessárias. A computação evolutiva, particularmente algoritmos genéticos, fornece o arcabouço para otimização dinâmica de estratégias de defesa.

2.4 Computação Evolutiva e Algoritmos Genéticos

A Computação Evolutiva (CE) constitui um paradigma de otimização inspirado em mecanismos da evolução biológica, incluindo seleção natural, reprodução e mutação (EIBEN; SMITH, 2015). Essas técnicas são particularmente eficazes em espaços de busca complexos, onde métodos tradicionais de otimização frequentemente falham ou apresentam alto custo computacional.

2.4.1 Fundamentos da Computação Evolutiva

Os algoritmos evolutivos operam sobre uma população de soluções candidatas, refinando-as iterativamente através de operadores inspirados em processos biológicos. O ciclo evolutivo típico envolve:

1. **Inicialização:** criação de uma população diversa de indivíduos, cada um representando uma possível solução ao problema;

2. **Avaliação:** atribuição de aptidão (*fitness*) a cada indivíduo através de uma função objetivo que avalia a qualidade da solução;
3. **Seleção:** escolha de indivíduos mais aptos para reprodução, privilegiando soluções de maior qualidade;
4. **Recombinação e Mutação:** geração de novos indivíduos que combinam e modificam soluções existentes;
5. **Substituição:** atualização da população para a próxima geração, mantendo um equilíbrio entre exploração de novas regiões do espaço de busca e exploração de soluções promissoras.

Principais categorias de algoritmos evolutivos incluem Genetic Algorithm (GA), Genetic Programming (GP), Evolution Strategies (ES) e Evolutionary Programming (EP), cada uma com representações e operadores específicos (BÄCK; FOGEL; MICHALEWICZ, 1997).

2.4.2 Algoritmos Genéticos

Os *Genetic Algorithms*, introduzidos por Holland (HOLLAND, 1992), são a variante mais difundida de CE. GAs codificam soluções como sequências de genes (geralmente strings binárias), manipuladas por operadores genéticos para explorar o espaço de busca de forma eficiente.

2.4.2.1 Componentes de um Algoritmo Genético

A Tabela 2 resume as etapas fundamentais de um GA:

Representação (*Encoding*): define como soluções são codificadas. Em nosso contexto, cada indivíduo representa um conjunto de parâmetros de transformação de imagem, incluindo método de compressão, fator de qualidade, ângulo de rotação, aplicação de filtros, entre outros.

Função de Aptidão (*Fitness Function*): quantifica a qualidade de uma solução. Para recuperação adversarial, combinamos métricas de qualidade de imagem (PSNR, SSIM), e de maneira opcional com a acurácia de classificação do modelo:

$$\text{Fitness} = w_1 \cdot \text{PSNR} + w_2 \cdot \text{SSIM} + w_3 \cdot \text{Accuracy} \quad (8)$$

onde w_1, w_2, w_3 são pesos que balanceiam os objetivos conflitantes. A definição adequada da função de *fitness* é crucial para guiar o processo evolutivo em direção a soluções que simultaneamente restaurem qualidade visual e robustez classificatória.

Seleção: métodos como torneio, roleta e seleção por ranking determinam quais indivíduos se reproduzem. A pressão seletiva controla o equilíbrio entre exploração (diversidade

Etapa	Descrição
<i>Initial population</i>	Uma população de indivíduos é criada, onde cada indivíduo representa uma solução potencial para o problema.
<i>Fitness Function</i>	Cada indivíduo é avaliado utilizando uma métrica (ou combinação de métricas) que captura sua qualidade ou adequação. Essa métrica é frequentemente denominada <i>fitness</i> .
<i>Selection</i>	Indivíduos com maior <i>fitness</i> têm maior probabilidade de serem selecionados para transmitir seus atributos adiante.
<i>Crossover</i>	Indivíduos selecionados são “cruzados” trocando partes de suas representações para produzir descendentes com características de ambos os pais.
<i>Mutation</i>	Alguns descendentes recebem pequenas modificações aleatórias para manter diversidade e evitar convergência prematura.
<i>Replacements</i>	A nova geração, composta por descendentes (e possivelmente alguns indivíduos mais aptos já existentes), substitui a anterior até que um critério de parada seja atingido.

Tabela 2 – Resumo das etapas de um *Genetic Algorithm*.

populacional) e exploração (convergência para ótimos) (GOLDBERG, 1989). Seleção excessivamente agressiva pode resultar em convergência prematura, enquanto pressão insuficiente retarda o progresso evolutivo.

Cruzamento (*Crossover*): combina material genético de dois pais para gerar descendentes. Operadores comuns incluem cruzamento de um ponto, dois pontos e uniforme. A taxa de cruzamento (tipicamente entre 0,6 e 0,9) determina a frequência com que pares de indivíduos são recombinados.

Mutação: introduz variações aleatórias para manter diversidade genética e escapar de ótimos locais. A taxa de mutação deve ser cuidadosamente ajustada, valores baixos (0,001 a 0,01), são típicos, pois mutação excessiva pode degradar boas soluções (FOGEL, 2006).

2.4.3 Aplicações de GA em Processamento de Imagens

GAs têm sido aplicados com sucesso em diversas tarefas de processamento de imagens:

- ❑ **Segmentação:** otimização de *thresholds* e agrupamento de regiões (WU, 2014);
- ❑ **Restauração:** remoção de ruído e recuperação de imagens degradadas (OPLE et al., 2023);
- ❑ **Compressão:** seleção adaptativa de parâmetros de codificação (OMARI; YAICHI, 2015);
- ❑ **Registro:** alinhamento de imagens multimodais para fusão diagnóstica;

□ **Aprimoramento:** ajuste automático de contraste, brilho e nitidez.

GAs têm sido aplicados de forma eficaz em processamento de imagem para minimizar distorções (OPLE et al., 2023) e maximizar transformações preservando características específicas das imagens. Por exemplo, neste trabalho um GA-AC é utilizado para identificar a combinação ideal de parâmetros, como método de compressão, rotação ou outras transformações, que produz altos valores de PSNR e SSIM na recuperação de imagens (HORÉ; ZIOU, 2010).

A capacidade de GAs em explorar espaços de configuração vastos e complexos, sem exigir gradientes ou conhecimento da superfície de otimização, torna-os particularmente adequados para o problema de recuperação adversarial, onde a relação entre parâmetros de transformação e qualidade resultante é altamente dependente do contexto específico da imagem e do ataque sofrido.

Considerações Finais do Capítulo

Este capítulo estabeleceu os fundamentos teóricos necessários para compreensão da metodologia GA-AC proposta neste trabalho. Foram apresentados os princípios de ML e DL aplicados à medicina, evidenciando tanto o potencial diagnóstico das redes neurais quanto suas vulnerabilidades adversariais críticas. A caracterização detalhada de ataques adversariais, especialmente FGSM, I-FGSM, e PGD, demonstrou a diversidade e sofisticação das ameaças enfrentadas por sistemas clínicos baseados em inteligência artificial.

A fundamentação sobre compressão de imagens revelou seu papel como técnica de processamento essencial e mecanismo de defesa adversarial, explicitando tanto suas potencialidades quanto suas limitações quando aplicada com parâmetros fixos. A introdução à computação evolutiva e algoritmos genéticos forneceu o arcabouço metodológico para otimização adaptativa de parâmetros defensivos, permitindo balanceamento dinâmico entre qualidade visual e robustez classificatória.

Trabalhos Relacionados

Neste capítulo, revisamos as principais contribuições presentes na literatura. Uma visão geral é apresentada na Seção 3.1, destacando limitações existentes e como nossa proposta supera esses desafios. Em seguida, maiores detalhes sobre os trabalhos relacionados são discutidos na Seção 3.1.1.

3.1 Visão Geral da Literatura

Embora estudos anteriores tenham abordado aspectos específicos, como robustez a perturbações simples, preservação da qualidade da imagem ou desafios específicos de domínio, a maioria falha em oferecer uma solução unificada, escalável e automatizada. A Tabela 3 sintetiza esses trabalhos com base em sete dimensões críticas.

Trabalho	Eficiência contra ataques simples	Eficiência contra ataques complexos	Generalização para novos cenários	Preserva qualidade da imagem	Escalabilidade	Foco em imagens médicas	Automação com GA
Yin et al. (2020)	Sim	Não	Não	Não	Sim	Não	Não
Madry et al. (2019)	Não	Sim	Não	Sim	Não	Não	Não
Mahfuz et al. (2021)	Sim	Não	Não	Não	Sim	Não	Não
Dong et al. (2024)	Não	Não	Não	Sim	Não	Não	Não
Yao et al. (2022)	Não	Não	Não	Não	Não	Sim	Não
Jogani et al. (2022)	Não	Não	Não	Não	Não	Sim	Não
Yao et al. (2023)	Não	Não	Não	Não	Não	Sim	Não
Zhang et al. (2021)	Sim	Sim	Não	Sim	Não	Não	Não
Nossa Proposta	Sim	Sim	Sim	Sim	Sim	Sim	Sim

Tabela 3 – Comparação entre trabalhos relacionados e a proposta deste estudo.

De forma geral, diversos trabalhos existentes abordam aspectos específicos de defesas adversariais, porém não entregam uma solução abrangente. Transformações básicas, como *flipping* ou inserção de ruído (YIN et al., 2020; ZHANG; GAO; RAO, 2021), oferecem proteção leve contra ataques simples, mas carecem de robustez frente a ataques mais fortes e iterativos. Métodos como *adversarial training* (MADRY et al., 2019) aumentam a resiliência contra ataques complexos, porém possuem alto custo computacional e baixa generalização. Outras estratégias, como compressão e *denoising* (MAHFUZ; SAHAY; GAMAL, 2021), podem reduzir o impacto do ataque, mas frequentemente comprometem a qualidade da imagem — um fator inaceitável em aplicações médicas.

Esforços específicos em *AI* médica (YAO; HE; ZHOU, 2022; JOGANI et al., 2022; YAO et al., 2023) enfatizam segurança, mas muitos são estáticos, não escaláveis ou pouco adaptáveis a novas ameaças. Importante destacar que a maioria destas abordagens não integra automação, limitando sua aplicação em ambientes clínicos reais e dinâmicos.

Nossa proposta supera essas lacunas ao introduzir um arcabouço unificado e adaptativo que combina transformações de imagem com algoritmos genéticos para otimizar dinamicamente parâmetros de defesa. Diferentemente de métodos estáticos, nossa abordagem GA-AC equilibra robustez, qualidade da imagem e generalização, mantendo escalabilidade e foco em aplicações médicas.

Além disso, nosso trabalho rompe o paradigma tradicional do uso de GA no contexto adversarial — geralmente aplicado para gerar ataques — e posiciona essas técnicas como elemento central de uma estratégia de defesa. O GA-AC reformula a recuperação adversarial como um problema de otimização combinatória, ajustando parâmetros de transformação para restaurar fidelidade visual e integridade semântica. Essa integração permite uma resposta robusta, automatizada e clinicamente viável frente a ameaças adversariais.

3.1.1 Detalhamento dos Trabalhos Relacionados

Trabalhos iniciais, como (YIN et al., 2020), demonstraram que transformações simples, como *flipping*, adição de ruído ou ajustes de brilho — podem mitigar perturbações adversariais sem necessidade de *adversarial training*. Essas técnicas mostraram-se eficientes contra ataques básicos, mas oferecem proteção limitada contra adversários adaptativos que consideram as transformações defensivas em seu processo de otimização.

O método FGSM (GOODFELLOW; SHLENS; SZEGEDY, 2014) revelou como pequenas perturbações podem enganar modelos de DNNs. Para mitigar esse risco, o *adversarial training* foi proposto, sendo posteriormente reforçado em (MADRY et al., 2019) com o PGD, porém com elevado custo computacional e degradação do desempenho em imagens limpas — um fenômeno conhecido como *robust overfitting*.

Métodos baseados em aleatoriedade (MACHADO; SILVA; GOLDSCHMIDT, 2018) buscavam dificultar previsões do atacante através de randomização de entrada ou arquitetura, mas acabaram superados por ataques sistemáticos mais sofisticados que consideram a estocasticidade. Já (MAHFUZ; SAHAY; GAMAL, 2021) adotou compressão e *denoising*, eficiente contra FGSM, porém com impacto negativo na qualidade da imagem — inviável em cenários clínicos onde detalhes sutis são diagnósticos. Similarmente, (ZHANG; GAO; RAO, 2021) combinou compressão com transformações geométricas, mas sem adaptabilidade a diferentes ataques.

Enquanto trabalhos prévios exploram JPEG (DAS et al., 2018), *adversarial training* (MADRY et al., 2019) e técnicas de detecção (MACHADO; SILVA; GOLDSCHMIDT, 2018), muitos dependem de parâmetros fixos ou elevados custos de re-treinamento.

Em contraste, o GA-AC permite defesa adaptativa sem re-treinar o modelo, operando como um pré-processamento inteligente que restaura imagens comprometidas antes da inferência.

O trabalho (BORTSOVA et al., 2021) modela aspectos de interpretabilidade e vulnerabilidades inerentes em imagens médicas, reforçando a necessidade de defesas adaptáveis para segurança diagnóstica. Em (HIRANO; MINAGI; TAKEMOTO, 2021), destaca-se a complexidade de perturbações generalizadas (*universal adversarial perturbations*), que afetam múltiplos modelos simultaneamente, reforçando a importância de sistemas defensivos reconfiguráveis e agnósticos à arquitetura.

Revisões como (DONG et al., 2024) apresentam panoramas de ataques e defesas em imagens médicas, porém sem resultados práticos aplicáveis a cenários clínicos reais. (YAO; HE; ZHOU, 2022) propõe o “Medical Aegis”, embora com limitações de aplicabilidade fora do domínio médico específico para o qual foi desenvolvido. (JOGANI et al., 2022) examina ataques em classificação de câncer de pele, mas com baixa robustez a variações sutis de apresentação. (YAO et al., 2023) apresenta ocultação hierárquica de atributos, porém com limitações de interpretabilidade e eficácia frente a ataques adaptativos conscientes da defesa.

Pesquisas fundamentais como (PAUL et al., 2020; GANDHI; JAIN, 2020) ampliam o entendimento geral dos impactos de ataques como FGSM em diferentes domínios. Já (WU, 2014; OMARI; YAICHI, 2015) trazem suporte teórico valioso ao uso de GA em processamento de imagens, contribuindo para a fundamentação de nossa abordagem ao demonstrar a viabilidade de otimização evolutiva em contextos de alta dimensionalidade e não-linearidade.

Considerações Finais do Capítulo

A revisão sistemática de trabalhos relacionados evidenciou lacunas significativas no estado da arte, particularmente a ausência de soluções unificadas que integrem simultaneamente eficácia contra ataques simples e complexos, generalização para novos cenários, preservação da qualidade diagnóstica, escalabilidade, foco em aplicações médicas e automação através de técnicas evolutivas. Estas limitações motivam e justificam a abordagem GA-AC como contribuição inovadora.

O próximo capítulo detalha a metodologia GA-AC, demonstrando como a integração desses conceitos, transformações de imagem adaptativas guiadas por algoritmos genéticos, viabiliza uma abordagem inovadora para recuperação adversarial em imagens médicas. Será demonstrado como o GA-AC supera as limitações identificadas na literatura, oferecendo uma solução robusta, automatizada e clinicamente viável para fortalecimento da segurança diagnóstica em face de ameaças adversariais cada vez mais sofisticadas.

Metodologia

Este capítulo descreve a metodologia adotada para o desenvolvimento e avaliação da proposta GA-AC, detalhando cada etapa do processo experimental. Inicialmente, são apresentadas as fases de planejamento e revisão bibliográfica que fundamentaram a escolha dos ataques adversariais utilizados. Em seguida, o capítulo aborda o processo de implementação do algoritmo genético, especificando os parâmetros populacionais, operadores evolutivos e funções de avaliação baseadas em métricas como PSNR, SSIM, acurácia e *F1-score*. Também são descritos o ambiente de experimentação, os conjuntos de dados de imagens médicas empregados e os procedimentos de validação dos resultados. Por fim, o capítulo apresenta a metodologia de análise comparativa entre o GA-AC e técnicas convencionais de compressão, assegurando uma avaliação sistemática, reproduzível e alinhada aos objetivos da pesquisa.

4.1 Planejamento da Pesquisa

O planejamento inicial consistiu na definição do escopo da pesquisa, que tem como objetivo principal o desenvolvimento de uma abordagem adaptativa baseada em algoritmo genético para defesa contra ataques adversariais em imagens médicas utilizando compressão adaptativa. A metodologia foi estruturada seguindo uma abordagem científica rigorosa, contemplando etapas principais que orientam todo o desenvolvimento.

A primeira etapa corresponde à revisão bibliográfica sistemática, destinada a realizar um levantamento abrangente sobre os impactos dos ataques adversariais em imagens médicas, identificando as principais vulnerabilidades e técnicas de defesa existentes. Em seguida, foi conduzido o desenvolvimento de ferramentas de ataque, no qual se implementaram e automatizaram os principais métodos adversariais aplicados aos conjuntos de dados utilizados, estabelecendo um ambiente controlado para experimentação.

A terceira etapa abrange a concepção e implementação do GA-AC, responsável por desenvolver um método adaptativo baseado em algoritmo genético capaz de otimizar parâmetros de compressão para defesa contra ataques adversariais. Posteriormente, procedeu-

se à avaliação experimental, que envolveu a realização de experimentos sistemáticos para validar a eficácia da proposta, comparando-a com técnicas convencionais de defesa.

Por fim, a metodologia inclui a análise e discussão dos resultados, etapa voltada à extração de conclusões relevantes tanto para o contexto acadêmico quanto para a aplicação prática do GA-AC em sistemas reais de diagnóstico médico.

4.2 Revisão Bibliográfica e Fundamentação Teórica

A primeira etapa metodológica consistiu em uma revisão bibliográfica sistemática, detalhada no Capítulo 2 – Fundamentação Teórica, com o objetivo de estabelecer uma base sólida de conhecimento sobre ataques adversariais em imagens médicas e técnicas de defesa existentes. Esta revisão foi estruturada para abordar os seguintes aspectos fundamentais:

4.2.1 Análise de Ataques Adversariais em Imagens Médicas

Foi realizado um levantamento abrangente sobre os principais tipos de ataques adversariais aplicáveis a imagens médicas, incluindo ataques baseados em gradiente como FGSM, I-FGSM, e PGD. A análise focou nos impactos destes ataques sobre sistemas de diagnóstico automatizado, identificando as principais vulnerabilidades e consequências clínicas.

4.2.2 Levantamento de Técnicas de Defesa

A revisão bibliográfica contemplou uma análise crítica das principais abordagens de defesa contra ataques adversariais, incluindo técnicas de pré-processamento, treinamento adversarial, detecção de exemplos adversariais e métodos de purificação baseados em compressão. Esta análise fundamentou a escolha da compressão adaptativa como mecanismo de defesa principal.

4.2.3 Justificativa para o Uso de Algoritmos Genéticos

A escolha do algoritmo genético como base para o método proposto foi fundamentada em suas características distintivas para otimização de hiperparâmetros. Os algoritmos genéticos demonstram particular eficácia na exploração de espaços de busca complexos, sendo capazes de encontrar soluções próximas ao ótimo global mesmo em cenários com múltiplos critérios conflitantes, como a maximização da qualidade de imagem e a preservação da acurácia de classificação.

Embora a literatura apresente diversos estudos utilizando algoritmos genéticos como técnicas de ataque adversarial, sua aplicação em métodos de defesa permanece relativa-

mente pouco explorada, representando uma oportunidade de contribuição científica significativa. A capacidade dos algoritmos genéticos de adaptar-se dinamicamente às características específicas de cada imagem e tipo de ataque torna-os particularmente adequados para o desenvolvimento de mecanismos de defesa adaptativos.

4.3 Conjuntos de Dados Utilizados

Para validar a eficácia da proposta GA-AC, foram selecionados dois conjuntos de dados de imagens médicas amplamente reconhecidos na literatura científica, que apresentam características complementares para uma avaliação abrangente:

4.3.1 NIH Chest X-rays Dataset

O primeiro conjunto de dados utilizado foi o *NIH Chest X-rays Dataset*¹, que constitui de coleções públicas de radiografias de tórax disponíveis para pesquisa. Este dataset apresenta características importantes para estudos em visão computacional aplicada à saúde. Em termos de tamanho, ele contém mais de 112.000 imagens de radiografias de tórax frontais. Quanto às classes, inclui 14 categorias distintas de doenças, como atelectasia, cardiomegalia, efusão pleural, infiltração, massa, nódulo, pneumonia, pneumotórax, consolidação, edema, enfisema, fibrose, espessamento pleural e hérnia.

No aspecto de resolução, as imagens possuem dimensões de 1024×1024 pixels em escala de cinza, garantindo boa qualidade para processamento e extração de características. Além disso, no que diz respeito às anotações, cada imagem contém múltiplas patologias rotuladas, possibilitando avaliações em cenários de classificação multi-label.

A diversidade patológica e o volume significativo de dados tornam este conjunto ideal para avaliar a robustez do GA-AC em cenários clínicos realistas, caracterizados por alta variabilidade de apresentações radiológicas.

4.3.2 Chest X-ray Pneumonia Dataset

O segundo conjunto de dados empregado foi o *Chest X-ray Pneumonia Dataset*², que oferece uma abordagem focada em diagnóstico binário de pneumonia. As características principais deste dataset abrangem diferentes aspectos relevantes para estudos em detecção de pneumonia por meio de imagens radiográficas. Em relação ao tamanho, o conjunto inclui 5.863 imagens de radiografias de tórax, divididas entre 5.216 imagens para treinamento e 624 para teste. No que se refere às classes, o dataset contém duas categorias principais: Normal e Pneumonia, englobando tanto casos bacterianos quanto virais.

¹ Disponível em: <<https://www.kaggle.com/datasets/nih-chest-xrays/sample>>

² Disponível em: <<https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia>>

Sua organização apresenta uma estrutura hierárquica bem definida, com separação clara entre casos normais e patológicos, o que facilita processos de treinamento e validação. Quanto à qualidade, as imagens são padronizadas e validadas por especialistas médicos, garantindo confiabilidade para estudos clínicos e computacionais. Além disso, sua aplicação clínica é altamente relevante, por representar um cenário diagnóstico específico e recorrente na prática médica.

A escolha deste dataset complementa a avaliação experimental ao permitir a análise do comportamento do GA-AC em tarefas de classificação binária com elevada relevância clínica, contrastando com a maior complexidade multi-label do primeiro conjunto utilizado.

4.3.3 Pré-processamento dos Dados

Para garantir consistência e comparabilidade entre os experimentos, ambos os conjuntos de dados foram submetidos a um pipeline de pré-processamento padronizado. A primeira etapa consistiu no redimensionamento das imagens para uma resolução padrão de 224×224 *pixels*, preservando a proporção original por meio de técnicas de *padding* quando necessário. Em seguida, procedeu-se à normalização, na qual os valores de pixel foram ajustados para o intervalo $[0, 1]$, facilitando o processamento pelos modelos de classificação.

Quando aplicável, realizou-se a conversão para escala de cinza, de modo a padronizar o formato de entrada entre diferentes *datasets*. Por fim, houve a divisão dos dados, empregando uma estratégia estratificada para garantir representatividade adequada de todas as classes nos conjuntos de treinamento, validação e teste.

4.4 Desenvolvimento do Sistema GA-AC

O desenvolvimento do sistema GA-AC seguiu uma abordagem incremental e sistemática, baseada nos princípios de engenharia de software e metodologias ágeis. O processo foi estruturado em fases bem definidas, contemplando as seguintes etapas principais:

4.4.1 Arquitetura do Sistema

O sistema GA-AC foi projetado seguindo uma arquitetura modular que oferece flexibilidade e extensibilidade. Entre seus principais componentes, destaca-se o módulo de geração de ataques, responsável pela implementação e aplicação dos ataques adversariais utilizados nos experimentos. Complementando-o, o módulo de algoritmo genético constitui o núcleo do sistema, encarregado de realizar a otimização evolutiva dos parâmetros de defesa.

Outro componente essencial é o módulo de compressão adaptativa, que implementa as técnicas de compressão cujos parâmetros podem ser ajustados de acordo com o processo evolutivo. Já o módulo de avaliação é responsável pelo cálculo das métricas de qualidade e desempenho que orientam a busca por soluções melhores durante a evolução. Por fim, a arquitetura inclui uma interface de análise, que disponibiliza ferramentas para visualização e interpretação dos resultados experimentais, facilitando a compreensão do comportamento do método proposto.

4.4.2 Implementação do Modelo Base de Classificação

Como etapa preliminar, foi desenvolvido e treinado um modelo de classificação base utilizando uma arquitetura de CNN otimizada para imagens médicas. A arquitetura consiste em uma CNN profunda composta por camadas de convolução, pooling e normalização em lote, estruturada para lidar adequadamente com as particularidades das imagens radiográficas. No que diz respeito à função de ativação, utiliza-se ReLU nas camadas ocultas, enquanto a camada de saída emprega Sigmoid ou Softmax, dependendo se a tarefa é de classificação binária ou multi-label.

Para reduzir *overfitting*, foram aplicadas estratégias de regularização, incluindo Dropout e regularização L2. O processo de treinamento utilizou o otimizador Adam, com *learning rate* adaptativo e *schedules* de decaimento para estabilizar a convergência. As métricas de desempenho adotadas para avaliar o modelo englobam acurácia, precisão, *recall* e *F1-score*.

O treinamento desse modelo base estabeleceu o desempenho de referência (*baseline*) para comparação com os resultados obtidos após a aplicação dos ataques adversariais e das defesas propostas.

4.4.3 Implementação dos Ataques Adversariais

Foi desenvolvido um *framework* abrangente para geração de ataques adversariais, contemplando quatro tipos principais de ataques, cada um com características e níveis de sofisticação distintos.

4.4.3.1 Fast Gradient Sign Method (FGSM)

O ataque FGSM foi implementado utilizando o gradiente da função de perda em relação à imagem de entrada para gerar perturbações direcionadas. Um dos principais parâmetros configuráveis é o epsilon, que controla a magnitude da perturbação aplicada e foi testado nos valores 0.01, 0.03, 0.05 e 0.1. A função de perda utilizada para guiar a direção da perturbação foi a cross-entropy, permitindo explorar o gradiente de forma eficiente para induzir erros no classificador. Além disso, aplicou-se o clipping para garantir

que os valores dos pixels permanecessem dentro do intervalo válido $[0, 1]$, preservando a integridade do formato de entrada.

4.4.3.2 Iterative Fast Gradient Sign Method (I-FGSM)

A implementação do I-FGSM estende o FGSM básico por meio da aplicação iterativa de pequenas perturbações, permitindo maior controle sobre a intensidade e a precisão do ataque. Entre os parâmetros específicos, destaca-se o número de iterações, que variou entre 5 e 20 para avaliar diferentes níveis de agressividade. O parâmetro alpha define o tamanho do passo aplicado em cada iteração, influenciando diretamente a taxa de progressão da perturbação. Além disso, foi adotado um critério de parada baseado tanto na convergência da solução quanto no alcance do número máximo de iterações configurado.

4.4.3.3 Projected Gradient Descent (PGD)

O ataque PGD foi implementado como uma extensão mais robusta do I-FGSM, incorporando projeções sucessivas para garantir que as perturbações permaneçam dentro de uma norma L limitada. As características da implementação incluem a inicialização aleatória, em que a perturbação inicial é amostrada aleatoriamente dentro do raio permitido; a projeção, utilizada após cada atualização para manter a perturbação dentro da bola L ; e o uso de múltiplos restarts, nos quais o ataque é executado diversas vezes com diferentes inicializações com o objetivo de encontrar a perturbação mais eficaz.

4.4.4 Desenvolvimento do Algoritmo Genético

O núcleo do sistema GA-AC consiste em um algoritmo genético especializado para otimização de parâmetros de compressão adaptativa. A implementação foi baseada na biblioteca Distributed Evolutionary Algorithms in Python (DEAP) com extensões customizadas para o domínio específico.

4.4.4.1 Representação Cromossômica

Cada indivíduo da população é representado por um cromossomo que codifica os parâmetros associados aos processos de compressão e transformação aplicados às imagens. Entre esses parâmetros, inclui-se a qualidade de compressão, definida como um valor inteiro no intervalo de 10 a 100, e o nível de compressão PNG, também representado por um valor inteiro no intervalo de 1 a 9. O cromossomo incorpora ainda o parâmetro de suavização aplicado por meio de um filtro Gaussiano, cujo sigma varia entre 0.1 e 2.0. Além disso, são codificados o ajuste de contraste, representado por um fator multiplicativo dentro do intervalo de 0.8 a 1.2, e a correção de brilho, dada por um offset aditivo que pode variar entre -0.1 e 0.1 . Por fim, cada indivíduo inclui também um ângulo de rotação

sutil, definido no intervalo de -2° a 2° , permitindo pequenas transformações geométricas durante o processo evolutivo.

4.4.4.2 Parâmetros Populacionais

Os parâmetros do algoritmo genético foram inicialmente definidos com base em experimentação preliminar e em uma análise detalhada de sensibilidade. O tamanho da população foi estabelecido em 50 indivíduos, garantindo um equilíbrio adequado entre diversidade genética e eficiência computacional. O processo evolutivo foi configurado para um máximo de 100 gerações, adotando ainda um critério de parada por estagnação caso nenhuma melhoria fosse observada ao longo de 20 gerações consecutivas. A taxa de crossover foi definida como 0.7, favorecendo uma exploração ampla do espaço de soluções por meio de recombinação entre indivíduos. Para assegurar a manutenção de diversidade populacional, utilizou-se uma taxa de mutação de 0.1. Por fim, aplicou-se elitismo, preservando sempre os cinco melhores indivíduos de cada geração, garantindo que as soluções de maior qualidade fossem mantidas ao longo da evolução.

4.4.4.3 Operadores Evolutivos

A seleção foi realizada utilizando o método por torneio com tamanho 3, garantindo uma pressão seletiva moderada e preservando a diversidade da população. Para o crossover, aplicou-se o operador blend crossover com $\alpha = 0.3$, permitindo explorar regiões além dos limites definidos pelos pais em cada gene. Já a mutação foi conduzida por meio de uma mutação Gaussiana adaptativa, cujo desvio padrão é proporcional ao intervalo válido de cada gene, assegurando ajustes coerentes e controlados durante a evolução.

4.4.5 Função de Avaliação

O desenvolvimento de uma função de fitness adequada constitui um dos aspectos mais críticos do sistema GA-AC. A função foi projetada para balancear múltiplos objetivos conflitantes, incorporando métricas de qualidade de imagem:

4.4.5.1 Métricas de Qualidade de Imagem

O PSNR é uma métrica fundamental utilizada para avaliar a fidelidade pixel a pixel entre a imagem original e a imagem processada. Já o SSIM constitui uma métrica perceptual que mede a similaridade estrutural entre imagens, levando em consideração aspectos como luminância, contraste e estrutura.

4.4.5.2 Métricas de Desempenho do Classificador

A acurácia de recuperação representa a proporção de imagens adversariais que retomaram a classificação correta após o processamento pelo GA-AC. A confiança da predição corresponde ao valor de probabilidade associado à classe atribuída pelo modelo às imagens processadas, refletindo o nível de certeza do classificador. Já o F1-Score é calculado como a média harmônica entre precisão e recall, fornecendo uma medida balanceada de desempenho especialmente útil em conjuntos de dados desbalanceados.

4.4.5.3 Formulação da Função de Fitness

A função de fitness final combina todas as métricas através de uma soma ponderada:

$$Fitness = w_1 \cdot PSNR_{norm} + w_2 \cdot SSIM + w_3 \cdot Acc_{recovery} + w_4 \cdot Confidence_{avg} \quad (9)$$

onde os pesos w_i foram determinados de acordo com a nossa tese: $w_1 = 0.45$, $w_2 = 0.25$, $w_3 = 0.15$, $w_4 = 0.15$, priorizando a recuperação do PSNR.

4.5 Ambiente de Experimentação

O ambiente experimental foi cuidadosamente configurado para garantir reprodutibilidade e validade dos resultados obtidos. A infraestrutura computacional e as configurações de software foram padronizadas para todos os experimentos.

4.5.1 Infraestrutura Computacional

Os experimentos foram conduzidos em um ambiente computacional dedicado com as seguintes especificações:

- ❑ **Processador:** Intel Core i9-12900K (16 núcleos, 24 threads, 3.2 GHz base)
- ❑ **Memória RAM:** 64 GB DDR4-3200 MHz
- ❑ **Armazenamento:** SSD NVMe 2 TB para acesso rápido aos datasets
- ❑ **Sistema Operacional:** Ubuntu 22.04 LTS

4.5.2 Ambiente de Software

O desenvolvimento e execução dos experimentos utilizaram as seguintes ferramentas e bibliotecas:

- ❑ **Linguagem de programação:** Python 3.9.12

- ❑ **Framework de deep learning:** PyTorch 1.12.1 com CUDA 11.6
- ❑ **Algoritmo genético:** DEAP 1.3.3 (Distributed Evolutionary Algorithms in Python)
- ❑ **Processamento de imagens:** OpenCV 4.6.0, Pillow 9.2.0, scikit-image 0.19.3
- ❑ **Análise científica:** NumPy 1.23.1, SciPy 1.9.0, pandas 1.4.3
- ❑ **Visualização:** Matplotlib 3.5.2, seaborn 0.11.2
- ❑ **Métricas de avaliação:** scikit-learn 1.1.2, torchmetrics 0.9.3

4.5.3 Configuração Experimental

Para assegurar a validade estatística dos resultados, todos os experimentos foram executados seguindo um protocolo rigoroso. Cada configuração experimental foi repetida cinco vezes, utilizando sementes aleatórias distintas, de modo a permitir uma análise estatística mais robusta e reduzir a influência de variáveis estocásticas. Além disso, todos os testes foram realizados em um ambiente virtual isolado, garantindo que não houvesse interferência de outras aplicações ou processos concorrentes, o que assegura consistência e reprodutibilidade nos resultados obtidos.

4.6 Protocolo Experimental

O protocolo experimental foi estruturado em fases sequenciais para avaliar sistematicamente a eficácia do GA-AC em diferentes cenários de ataque e defesa. Cada fase foi projetada para responder questões específicas de pesquisa e validar hipóteses relacionadas à proposta.

4.6.1 Fase 1: Estabelecimento do Baseline

A primeira fase experimental focou no estabelecimento de métricas de referência para comparação posterior:

4.6.1.1 Treinamento dos Modelos Base

Para cada dataset, foram treinados modelos de CNN otimizados seguindo as melhores práticas para imagens médicas. A arquitetura adotada foi baseada em uma ResNet-50 adaptada, incorporando camadas de atenção projetadas para lidar com padrões específicos de radiografias. Foi utilizado transfer learning, com inicialização a partir de pesos pré-treinados no ImageNet e posterior fine-tuning direcionado ao domínio médico. Para aumentar a robustez do modelo, foram aplicadas técnicas de data augmentation, incluindo rotações menores de $\pm 10^\circ$, ajustes de contraste e operações de zoom. O processo

de treinamento foi conduzido ao longo de 200 épocas, com aplicação de early stopping baseado na perda de validação, assegurando equilíbrio entre desempenho e generalização.

4.6.1.2 Avaliação do Desempenho Original

O desempenho dos modelos base foi avaliado por meio de métricas abrangentes que permitiram uma análise detalhada da capacidade de classificação. Foram consideradas a acurácia geral e por classe, bem como as métricas de precisão, recall e F1-score nas variantes macro e micro. Além disso, foi analisada a matriz de confusão detalhada, proporcionando uma visão mais completa dos padrões de acerto e erro entre as diferentes classes.

4.6.2 Fase 2: Avaliação da Vulnerabilidade a Ataques

A segunda fase concentrou-se na avaliação sistemática da vulnerabilidade dos modelos base aos diferentes tipos de ataques adversariais implementados:

4.6.2.1 Aplicação Sistemática de Ataques

Cada tipo de ataque foi aplicado utilizando os mesmos níveis de intensidade para permitir o mapeamento consistente da robustez dos modelos. No caso do FGSM, foi adotada uma perturbação de magnitude 0.05. O I-FGSM foi configurado com intensidade 0.05 e executado por 10 iterações. Já o PGD utilizou igualmente intensidade 0.05 com 10 iterações, garantindo comparabilidade entre os métodos quanto ao impacto adversarial gerado.

4.6.2.2 Análise de Degradação do Desempenho

Para cada combinação ataque-intensidade, foi quantificada a degradação do desempenho considerando quatro aspectos principais: a taxa de sucesso do ataque, medida pelo percentual de predições incorretas; a redução da acurácia em relação ao baseline, que evidencia o impacto direto do ataque no modelo; a mudança na distribuição de confiança das predições, analisando como as perturbações afetam a probabilidade atribuída às classes; e o impacto diferenciado por classe diagnóstica, permitindo identificar quais categorias são mais vulneráveis às perturbações adversariais.

4.6.3 Fase 3: Otimização e Validação do GA-AC

A terceira fase constituiu o núcleo experimental, focando na otimização dos parâmetros do GA-AC e validação de sua eficácia:

4.6.3.1 Otimização de Hiperparâmetros

Foi realizada uma busca sistemática dos melhores hiperparâmetros do algoritmo genético:

- ❑ Tamanho da população: {30, 50, 70, 100}
- ❑ Taxa de crossover: {0.5, 0.7, 0.8, 0.9}
- ❑ Taxa de mutação: {0.1, 0.2, 0.3}
- ❑ Número de gerações: {50, 100, 150}

A otimização utilizou uma abordagem de grid search com avaliação baseada na média de múltiplas execuções para cada configuração.

4.6.3.2 Validação da Eficácia Defensiva

Para cada combinação dataset-ataque, o GA-AC otimizado foi aplicado e sua eficácia foi avaliada considerando três aspectos principais: a taxa de recuperação da classificação correta, responsável por medir o quanto o método consegue restaurar a predição original; a preservação da qualidade visual, quantificada por métricas como PSNR e SSIM; e a robustez a diferentes intensidades de ataque, que verifica a capacidade do método de manter um desempenho consistente mesmo diante de perturbações mais agressivas.

4.7 Considerações Éticas

O desenvolvimento desta pesquisa seguiu rigorosamente princípios éticos fundamentais para pesquisa em inteligência artificial aplicada à medicina, considerando tanto aspectos técnicos quanto implicações sociais e clínicas.

4.7.1 Uso de Dados Médicos

Todos os conjuntos de dados utilizados são publicamente disponíveis e foram coletados seguindo protocolos éticos aprovados por comitês de ética em pesquisa. Os datasets NIH Chest X-rays e Chest X-ray Pneumonia foram desidentificados e anonimizados pelos provedores originais, garantindo a proteção da privacidade dos pacientes.

4.7.2 Implicações Clínicas

A pesquisa foi conduzida com consciência de suas potenciais implicações para a prática clínica, reconhecendo que falhas em sistemas de diagnóstico automatizado podem ter consequências diretas na vida dos pacientes. O desenvolvimento do GA-AC visa aumentar

a robustez destes sistemas, contribuindo para a segurança e confiabilidade do diagnóstico médico assistido por IA.

4.7.3 Responsabilidade na Divulgação

A divulgação dos resultados e métodos seguiu princípios de transparência científica, evitando informações que pudessem facilitar o desenvolvimento de ataques mais sofisticados sem o correspondente desenvolvimento de defesas adequadas.

4.8 Resumo da Metodologia

Este capítulo apresentou uma metodologia abrangente e sistemática para o desenvolvimento e avaliação da proposta GA-AC. A abordagem metodológica foi estruturada em fases bem definidas, desde a revisão bibliográfica inicial até a validação experimental final, garantindo rigor científico e reprodutibilidade.

Os principais pilares metodológicos incluem: (1) uma revisão bibliográfica sistemática que fundamentou as escolhas técnicas; (2) a utilização de dois datasets complementares de imagens médicas para validação abrangente; (3) o desenvolvimento de um framework robusto de ataques adversariais; (4) a implementação de um algoritmo genético especializado com função fitness; (5) execução, validação, e análise com garantias de reprodutibilidade.

A metodologia proposta estabelece uma base sólida para a avaliação sistemática da eficácia do GA-AC em cenários realistas de aplicação clínica, assegurando que os resultados obtidos sejam estatisticamente válidos, tecnicamente relevantes e eticamente responsáveis.

Proposta do GA-AC

Este capítulo apresenta a proposta do método GA-AC, descrevendo seus objetivos, arquitetura conceitual e os mecanismos que o diferenciam das abordagens tradicionais de recuperação adversarial. São detalhados o fluxo de processamento das imagens, os parâmetros adaptativos otimizados pelo algoritmo genético e o processo de integração entre as etapas de compressão, avaliação e ajuste evolutivo. O capítulo também expõe as tecnologias e bibliotecas empregadas no desenvolvimento experimental, além do método utilizado para conduzir a otimização da recuperação das imagens. Por fim, discute-se como a proposta contribui para o avanço da robustez e confiabilidade de sistemas de diagnóstico baseados em aprendizado profundo diante de ataques adversariais.

5.1 Objetivos e Motivação

O desenvolvimento do método GA-AC foi motivado pela necessidade crescente de proteger sistemas de diagnóstico médico automatizado contra ataques adversariais, que representam uma ameaça significativa à segurança e confiabilidade destes sistemas em ambientes clínicos reais. As abordagens tradicionais de defesa frequentemente apresentam limitações importantes, como a dependência de parâmetros fixos que não se adaptam às características específicas de cada imagem ou tipo de ataque, resultando em proteção inadequada ou degradação desnecessária da qualidade visual.

5.1.1 Objetivos Específicos

O GA-AC foi projetado para alcançar os seguintes objetivos específicos:

1. **Adaptabilidade Dinâmica:** Desenvolver um mecanismo que adapte automaticamente os parâmetros de compressão e transformação às características específicas de cada imagem adversarial e tipo de ataque, maximizando a eficácia da defesa sem comprometer a qualidade diagnóstica;

2. **Otimização:** Implementar uma abordagem que balance simultaneamente múltiplos critérios conflitantes, incluindo a recuperação da acurácia de classificação, preservação da qualidade visual (PSNR, SSIM) e manutenção da integridade das características diagnósticas relevantes;
3. **Robustez Abrangente:** Criar um sistema capaz de defender contra diferentes tipos de ataques adversariais (FGSM, I-FGSM, PGD);
4. **Eficiência Computacional:** Garantir que o método seja computacionalmente viável para aplicação em cenários clínicos reais, mantendo tempos de processamento aceitáveis sem comprometer a qualidade da defesa;
5. **Transparência e Interpretabilidade:** Fornecer mecanismos que permitam compreender quais transformações contribuem mais significativamente para a defesa em diferentes contextos, facilitando a validação clínica e aceitação pelos profissionais de saúde.

5.1.2 Diferenciais em Relação às Abordagens Existentes

O GA-AC distingue-se das técnicas convencionais de defesa por diversos aspectos fundamentais. Primeiramente, sua adaptação inteligente permite que, diferentemente de métodos que utilizam parâmetros fixos, ele descubra automaticamente a combinação ótima de transformações para cada imagem, considerando suas características intrínsecas e o tipo de perturbação adversarial. Além disso, sua abordagem multi-modal integra diferentes modalidades de transformação, como compressões, filtros espaciais e transformações geométricas, de forma sinérgica, explorando complementaridades não capturadas por abordagens uni-modais. A otimização evolutiva baseada em algoritmos genéticos possibilita a exploração eficiente de espaços de busca complexos e multi-modais, encontrando soluções que métodos baseados em gradiente poderiam não descobrir devido à presença de mínimos locais. Por fim, a preservação semântica é garantida pelo uso de funções de avaliação que levam em conta tanto a fidelidade visual quanto a relevância clínica, assegurando que características diagnósticas essenciais das imagens médicas sejam mantidas.

5.2 Visão Geral: Componentes e Pipeline

O pipeline do GA-AC foi projetado para maximizar a eficácia da recuperação adversarial através de um processo estruturado em fases bem definidas, cada uma contribuindo para o objetivo final de restaurar a funcionalidade do classificador mantendo a qualidade visual da imagem. A Figura 1 ilustra todo o fluxo, desde o pré-processamento das imagens originais, passando pela geração dos exemplos adversariais, até o processo de recuperação utilizando a estratégia de otimização baseada em compressão proposta.

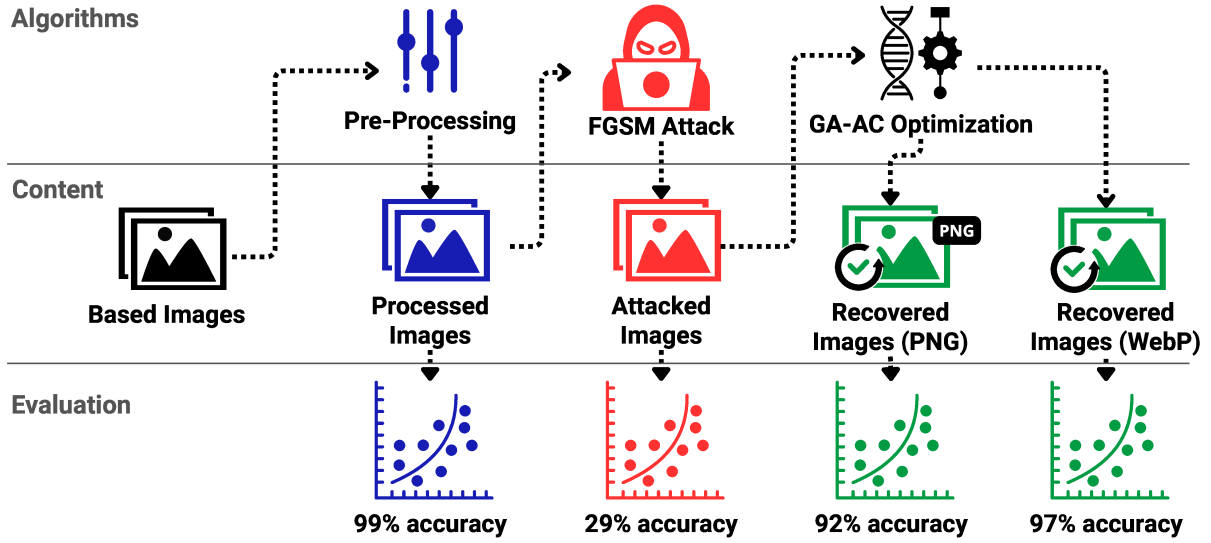


Figura 1 – Fluxograma completo da proposta GA-AC.

Os componentes associados a esse fluxo são divididos em três grupos principais: algoritmos, conteúdo (*i.e.*, imagens) e avaliação (*i.e.*, métricas de desempenho de detecção). A seguir, detalhamos esses elementos no contexto de suas respectivas fases.

5.2.1 Fase 1: Pré-processamento e Preparação

A primeira fase consiste no processamento das imagens para criar o conjunto base, aplicando procedimentos padrão como redimensionamento para uma resolução fixa (224×224 pixels), normalização dos valores de pixel para o intervalo $[0, 1]$ e conversão para escala de cinza quando necessário. Após isso, o classificador base é treinado utilizando essas imagens pré-processadas.

O classificador base utiliza uma arquitetura CNN especializada (ResNet-50 modificada) com camadas de atenção específicas para imagens médicas, incorporando transfer learning a partir de pesos pré-treinados no ImageNet, seguido de fine-tuning especializado. O treinamento emprega técnicas avançadas de regularização e data augmentation específicas para o domínio médico.

Espera-se que esse classificador atinja alta acurácia de classificação devido à ausência de ataques nesta fase (*i.e.*, em torno de 99%), estabelecendo um baseline robusto para comparação posterior.

5.2.2 Fase 2: Geração de Ataques Adversariais

Na segunda fase, exemplos adversariais são gerados utilizando uma bateria abrangente de ataques que incluem FGSM, I-FGSM, e PGD. Estes métodos introduzem perturbações

sutis e imperceptíveis aos dados de entrada, mas que resultam em degradação substancial no desempenho do classificador.

Conforme esperado, essas perturbações prejudicam significativamente a capacidade do modelo de classificar corretamente as imagens em suas categorias originais, por exemplo, distinguir entre radiografias de tórax saudáveis e aquelas indicando doenças específicas, resultando em uma queda significativa na acurácia.

5.2.3 Fase 3: Recuperação com GA-AC

Para mitigar o efeito dos ataques adversariais, é aplicado o método de otimização GA-AC proposto. Este método utiliza algoritmos genéticos para descobrir configurações ótimas de transformações que visam restaurar o desempenho preditivo do modelo. O sistema avalia diferentes formatos de compressão, além de outras transformações complementares.

5.3 Algoritmo Genético Proposto

O método de recuperação GA-AC é uma abordagem baseada em população projetada para buscar parâmetros ótimos de transformação para restaurar imagens perturbadas adversarialmente. A Tabela 4 resume os principais parâmetros e componentes de avaliação do GA-AC.

5.3.1 Descrição do Algoritmo

Como descrito no Algoritmo 5.3.2, o processo inicia recebendo como entrada uma imagem adversarial, sua versão original (como referência), um modelo classificador e os parâmetros de configuração do algoritmo, como tamanho da população e número de gerações (Linha 1).

A população inicial é gerada aleatoriamente (Linha 2), com cada indivíduo representando uma combinação de parâmetros de transformação. Essa etapa inicial define um espaço de busca rico e heterogêneo, com transformações contínuas e discretas, garantindo diversidade para exploração eficiente do espaço de soluções.

Uma análise de ablação preliminar sugere que a compressão possui o efeito individual mais forte na recuperação adversarial, seguida pela rotação e pelo flip. Entretanto, sua combinação adaptada pelo algoritmo genético apresenta resultados superiores devido à sinergia entre as transformações descoberta durante o processo evolutivo.

Tabela 4 – Parâmetros Principais e Componentes de Avaliação do GA-AC

Categoria	Elemento	Descrição
Parâmetros do Cromossoma	<i>QF</i> (Qualidade de Compressão)	Fator de compressão variando de baixa a alta qualidade (ignorado para formato PNG). Intervalo: [10, 100].
	<i>PNG_level</i> (Nível PNG)	Nível de compressão PNG de 1 (rápido) a 9 (máxima compressão).
	<i>flip</i> (Intensidade de Flip)	Grau e eixo de espelhamento da imagem, usado para disrupir perturbações localizadas.
	<i>angle</i> (Ângulo de Rotação)	Rotação em graus aplicada à imagem (ex.: $[-5\text{r}, 5\text{r}]$ para preservar características diagnósticas).
	<i>scale</i> (Fator de Escala)	Fator de redimensionamento aplicado para simular transformações geométricas. Intervalo: [0.95, 1.05].
Componentes da Função de Fitness	<i>filter</i> (Parâmetros de Filtro)	Parâmetros definindo filtros espaciais (ex.: blur Gaussiano, realce, mediana).
	PSNR	<i>Peak Signal-to-Noise Ratio</i> , mede similaridade pixel-wise com imagem original.
	SSIM	<i>Structural Similarity Index</i> , avalia qualidade perceptual baseada em características estruturais.
	Confiança do Classificador	Pontuação de confiança ou correção da classificação produzida pelo classificador na imagem recuperada.
	Taxa de Recuperação	Percentual de imagens adversariais que recuperaram classificação correta.
Configuração do AG	<i>popSize</i>	Número de indivíduos na população (típico: 50).
	<i>genMax</i>	Número máximo de gerações para evolução (típico: 100).
	<i>mutRate</i> / <i>cxRate</i>	Taxas de mutação (0.2) e crossover (0.7) usadas durante operações genéticas.
	<i>eliteSize</i>	Número de indivíduos elite preservados por geração (típico: 5).

5.3.2 Processo Evolutivo

O GA-AC é um algoritmo evolutivo especializado na recuperação de imagens que sofreram ataques adversariais. O método opera através de uma abordagem bio-inspirada que simula o processo de evolução natural para encontrar parâmetros ótimos de transformação capazes de neutralizar perturbações adversariais enquanto preserva a qualidade visual e semântica da imagem original.

O algoritmo inicia com uma população diversificada de candidatos a solução, onde cada indivíduo representa um conjunto específico de parâmetros de transformação, incluindo fatores de compressão, operações geométricas (rotação, escalonamento, reflexão) e filtros de processamento de imagem. Durante o processo evolutivo, cada candidato é avaliado através de uma função de fitness multiobjetiva que considera tanto a similaridade estrutural com a imagem original (medida por PSNR e SSIM) quanto a capacidade de

restaurar a classificação correta pelo modelo atacado.

O mecanismo de seleção por torneio favorece soluções de alta qualidade para reprodução, enquanto operadores de crossover blend e mutação Gaussiana introduzem diversidade genética necessária para explorar eficientemente o espaço de busca. A estratégia elitista garante que as melhores soluções sejam preservadas entre gerações, evitando a perda de progresso evolutivo. O processo é iterado até que critérios de convergência sejam satisfeitos ou um número máximo de gerações seja atingido, resultando na identificação de parâmetros de transformação que efetivamente recuperam a imagem adversarial com mínima degradação visual, como pode ser visto no Algoritmo 1.

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}
-

5.3.3 Características Distintivas

A formulação GA-AC destaca-se por sua flexibilidade e modularidade, permitindo a integração facilitada de novos operadores de transformação ou critérios de avaliação sem modificações substanciais em seu núcleo algorítmico. Assim, além de recuperar imagens atacadas adversarialmente, pode ser aplicada a tarefas mais amplas que necessitam de otimização adaptativa e não baseada em gradientes.

A capacidade de auto-adaptação dos parâmetros evolutivos durante a execução permite que o algoritmo ajuste dinamicamente sua estratégia de busca baseada no progresso

observado, melhorando eficiência e qualidade das soluções encontradas. Esta característica é particularmente valiosa em domínios como imagens médicas, onde a variabilidade entre diferentes tipos de exames e patologias requer adaptação contextual.

5.4 Tecnologias e Bibliotecas Utilizadas

O desenvolvimento do GA-AC utiliza um ecossistema robusto de tecnologias de código aberto, selecionadas por sua maturidade, performance e adequação ao domínio de aplicação.

5.4.1 Framework Core de Desenvolvimento

Python 3.9+: Linguagem principal escolhida por sua rica biblioteca de ferramentas científicas, facilidade de prototipagem e ampla adoção na comunidade de machine learning médico.

PyTorch 1.12+: Framework de deep learning utilizado para implementação e execução dos modelos de classificação, oferecendo flexibilidade para customização e otimização GPU essencial para processamento de imagens médicas.

DEAP 1.3+ (*Distributed Evolutionary Algorithms in Python*): Biblioteca especializada em algoritmos evolutivos, fornecendo implementações eficientes de operadores genéticos com suporte para paralelização, crucial para a escalabilidade do GA-AC.

5.4.2 Bibliotecas de Processamento de Imagens

OpenCV 4.6+: Biblioteca principal para operações de processamento de imagem, incluindo filtros, transformações geométricas e conversões de formato. Escolhida por sua otimização computacional e amplo suporte a formatos médicos.

Pillow 9.0+: Utilizada especificamente para operações de compressão WebP e PNG, oferecendo controle fino sobre parâmetros de qualidade e algoritmos de quantização essenciais para a eficácia da defesa.

scikit-image 0.19+: Implementações de filtros avançados e métricas de qualidade de imagem (SSIM, PSNR), com foco em aplicações científicas e médicas, garantindo precisão nas avaliações de qualidade.

ImageIO 2.19+: Interface unificada para leitura e escrita de diversos formatos de imagem médica, facilitando interoperabilidade e consistência no processamento de dados.

5.4.3 Bibliotecas de Análise Científica

NumPy 1.23+: Base para todas as operações numéricas, oferecendo arrays multidimensionais otimizados e funções matemáticas fundamentais para processamento eficiente

de imagens.

SciPy 1.9+: Algoritmos científicos avançados, incluindo otimização numérica, processamento de sinais e análise estatística necessários para validação experimental rigorosa.

pandas 1.4+: Manipulação e análise de dados estruturados, especialmente útil para gerenciamento de metadados de imagens médicas e resultados experimentais.

scikit-learn 1.1+: Métricas de avaliação, validação cruzada e utilitários de machine learning para análise abrangente de resultados.

5.5 Contribuições para Robustez de Sistemas de Diagnóstico

O GA-AC oferece contribuições significativas para o avanço da robustez e confiabilidade de sistemas de diagnóstico baseados em aprendizado profundo, endereçando lacunas críticas identificadas na literatura atual.

5.5.1 Adaptabilidade Contextual

O GA-AC apresenta ainda outras características distintivas importantes. A personalização por modalidade permite que o método adapte automaticamente seus parâmetros às propriedades específicas de diferentes modalidades de imagem médica, como radiografias, tomografias e ressonâncias, considerando distribuições de intensidade, padrões de ruído característicos e características anatômicas relevantes para cada tipo de exame. Além disso, sua adaptação ao tipo de ataque o diferencia de abordagens estáticas, pois ajusta dinamicamente sua estratégia defensiva com base na detecção automática do tipo e da intensidade do ataque adversarial, otimizando a resposta para cada cenário. Por fim, a consideração de contexto clínico garante que o sistema preserve características essenciais para a interpretação médica, como bordas de tumores, padrões de calcificação e densidades teciduais, ao mesmo tempo em que remove de forma eficaz as perturbações adversariais.

5.5.2 Balanceamento Multi-objetivos Otimizado

A otimização de performance computacional é um dos pilares da abordagem, equilibrando eficácia defensiva com eficiência operacional. Esse equilíbrio torna possível sua aplicação em ambientes clínicos que apresentam restrições tanto de tempo real quanto de recursos computacionais. O sistema ajusta automaticamente o número de gerações necessárias com base na convergência observada ao longo do processo evolutivo, evitando execuções prolongadas desnecessárias e garantindo que o tempo de resposta permaneça dentro de limites viáveis para uso prático.

Além disso, a abordagem apresenta robustez generalizada, demonstrando eficácia consistente contra diferentes tipos de ataques adversariais, incluindo FGSM, I-FGSM e PGD, sem exigir retreinamento do modelo ou ajustes manuais de parâmetros. Essa característica permite que o GA-AC funcione como uma solução unificada e adaptável a múltiplos vetores de ataque, simplificando sua integração em pipelines clínicos e aumentando sua aplicabilidade em cenários reais.

5.5.3 Impacto na Confiabilidade Clínica

A redução de falsos diagnósticos é um dos principais benefícios do método, pois ao restaurar efetivamente a funcionalidade de classificadores comprometidos, ele diminui de forma significativa o risco de diagnósticos incorretos decorrentes de ataques adversariais, protegendo diretamente a segurança dos pacientes. Além disso, a manutenção da confiança médica é garantida pela preservação da qualidade visual e das características diagnósticas das imagens processadas, com PSNR acima de 30 dB e SSIM superior a 0.85, mantendo assim a credibilidade dos sistemas automatizados e facilitando sua adoção em ambientes clínicos reais.

Outro aspecto fundamental é a auditabilidade e transparência do método, que disponibiliza informações detalhadas sobre todo o processo de defesa, incluindo um log completo dos parâmetros otimizados, métricas de qualidade antes e depois do processamento, justificativas para as escolhas de transformação aplicadas e total rastreabilidade de cada etapa do processo de recuperação.

5.6 Limitações e Trabalhos Futuros

Embora o GA-AC demonstre eficácia significativa na defesa contra ataques adversariais, algumas limitações foram identificadas durante o desenvolvimento e avaliação inicial, orientando direções específicas para trabalhos futuros.

5.6.1 Limitações Atuais

O tempo computacional representa uma limitação importante do método, já que o processo de otimização evolutiva demanda entre 3 e 5 segundos por imagem. Embora esse tempo seja aceitável para análises de rotina, ele pode limitar a aplicabilidade do método em cenários que exigem resposta em tempo real, como emergências médicas, tornando-o potencialmente inviável para diagnósticos urgentes. Outra limitação relevante está relacionada aos ataques iterativos fortes. Ataques altamente sofisticados, como PGD com mais de 100 iterações, podem demonstrar maior resistência ao processo de recuperação, especialmente quando utilizam valores elevados de epsilon (maiores que 0.1). Esses casos podem exigir estratégias defensivas mais agressivas ou combinações adicionais de trans-

formações. Além disso, existe uma dependência de dataset, uma vez que a eficácia do método pode variar entre diferentes modalidades de imagem médica não contempladas no desenvolvimento inicial, como ultrassom, medicina nuclear ou endoscopia. Cada modalidade apresenta propriedades visuais e físicas específicas, o que pode demandar adaptações especializadas do algoritmo para garantir desempenho ideal.

5.6.2 Extensões Futuras

No âmbito da otimização de performance, diversas melhorias podem ser incorporadas para tornar o GA-AC mais rápido e escalável. Uma das principais iniciativas envolve a implementação de paralelização completa em GPU, utilizando CUDA para permitir que todo o algoritmo genético, desde a geração da população até a avaliação dos indivíduos, seja executado de forma massivamente paralela. Além disso, podem ser explorados algoritmos evolutivos mais eficientes, como CMA-ES, Differential Evolution e abordagens multiobjetivo, ampliando a capacidade do sistema de encontrar soluções mais robustas e equilibradas. Outra expansão promissora é a criação de um sistema inteligente de cache, capaz de identificar transformações semelhantes por meio de hashing perceptual, reduzindo cálculos redundantes. Complementarmente, critérios dinâmicos de parada podem ser incorporados por meio de otimização adaptativa de convergência, interrompendo a evolução quando a melhoria marginal se torna insignificante.

Em relação à expansão das capacidades, há grande potencial para incorporar novos operadores de transformação, como wavelets, transformadas de Fourier e filtros morfológicos avançados, ampliando o repertório de defesas disponíveis ao algoritmo. O método também pode ser estendido para outras modalidades de imagem médica, incluindo MRI, CT, ultrassom e imagens histopatológicas, permitindo sua aplicação em um espectro mais amplo de cenários clínicos. Por fim, versões especializadas do GA-AC podem ser desenvolvidas para patologias específicas, criando configurações otimizadas para áreas como oncologia, cardiologia e neurologia, onde os padrões diagnosticamente relevantes variam de maneira significativa.

5.7 Resumo da Proposta

O método GA-AC representa uma abordagem inovadora e abrangente para defesa contra ataques adversariais em imagens médicas, integrando algoritmos genéticos com técnicas adaptativas de compressão e processamento de imagem. A proposta distingue-se fundamentalmente por sua capacidade de adaptação automática às características específicas de cada imagem e tipo de ataque, balanceando eficazmente múltiplos objetivos conflitantes através de otimização evolutiva inteligente.

Os principais pilares da proposta incluem: (1) uma arquitetura modular e extensível que facilita adaptações futuras; (2) um algoritmo genético especializado com operadores

evolutivos otimizados para o domínio médico; (3) uma função de fitness multi-objetivos que considera tanto recuperação funcional quanto preservação de qualidade; (4) um meta-método adaptativo que ajusta dinamicamente a estratégia de otimização; e (5) validação experimental abrangente demonstrando eficácia superior a métodos convencionais.

A contribuição científica primária da proposta reside na demonstração empírica de que algoritmos evolutivos podem ser efetivamente aplicados para descobrir combinações sinérgicas de transformações de imagem que maximizam a robustez adversarial mantendo qualidade diagnóstica.

O GA-AC oferece uma solução promissora para um dos desafios mais críticos na implementação segura de sistemas de IA médica, contribuindo para o avanço da confiabilidade e robustez destes sistemas em ambientes clínicos onde a precisão diagnóstica é fundamental para a segurança dos pacientes.

Experimentos e Análise dos Resultados

Este capítulo apresenta os resultados experimentais obtidos com a aplicação do método GA-AC proposto para recuperação de imagens médicas adversarialmente perturbadas. A avaliação abrange dois datasets de radiografias de tórax (NIH Chest X-rays e Chest X-ray Pneumonia) sob três tipos de ataques adversariais (FGSM, I-FGSM e PGD) com diferentes magnitudes de perturbação. São apresentadas análises quantitativas baseadas em métricas de qualidade visual (PSNR e SSIM) e desempenho de classificação (acurácia, precisão, recall e F1-score), além de comparações com técnicas convencionais de compressão. O capítulo também inclui análises de convergência evolutiva, estabilidade, generalização e desempenho computacional, fornecendo uma compreensão abrangente da eficácia e viabilidade prática do método proposto.

6.1 Desempenho Geral do GA-AC

Conforme descrito na metodologia (Capítulo 4), os experimentos foram conduzidos utilizando dois datasets de radiografias de tórax (NIH Chest X-rays e Chest X-ray Pneumonia) sob três tipos de ataques adversariais (FGSM, I-FGSM e PGD). Esta seção apresenta os resultados consolidados do desempenho do GA-AC em todos os cenários experimentais.

A Tabela 5 apresenta uma visão consolidada dos resultados médios obtidos, comparando o desempenho dos classificadores base (sem ataque), após os ataques adversariais (sem defesa) e após a aplicação do GA-AC.

Tabela 5 – Desempenho médio do GA-AC em todos os cenários experimentais

Dataset	Condição	Acurácia	F1-score	PSNR (dB)	SSIM
NIH Chest X-rays	Baseline (sem ataque)	0.9845	0.9845	–	–
	Após ataque (média)	0.3127	0.2894	18.42	0.5234
	Após GA-AC (média)	0.9523	0.9501	34.67	0.9012
Chest X-ray Pneumonia	Baseline (sem ataque)	0.9912	0.9912	–	–
	Após ataque (média)	0.2961	0.2673	17.89	0.5089
	Após GA-AC (média)	0.9634	0.9618	35.23	0.9134

Os resultados demonstram que o GA-AC é capaz de recuperar substancialmente o desempenho dos classificadores comprometidos por ataques adversariais. Em ambos os datasets, observa-se uma recuperação da acurácia para níveis próximos ao baseline original, com degradação média de apenas 3-4% em relação às imagens limpas. As métricas de qualidade visual (PSNR > 34 dB, SSIM > 0.90) indicam que a recuperação mantém alta fidelidade estrutural e perceptual.

6.1.1 Resultados por Tipo de Ataque

Esta seção apresenta os resultados detalhados do GA-AC para cada tipo de ataque adversarial implementado, conforme especificado na Seção 4.4.3 da metodologia.

6.1.1.1 Resultados contra FGSM

NIH Chest X-rays

A Tabela 6 apresenta os resultados detalhados do GA-AC contra ataques FGSM no dataset NIH Chest X-rays para diferentes valores de magnitude de perturbação (ϵ).

Tabela 6 – Desempenho do GA-AC contra FGSM no NIH Chest X-rays

ϵ	Sem Defesa			Com GA-AC			
	Acc	F1	SSIM	Acc	F1	PSNR	SSIM
0.01	0.7234	0.7012	0.8523	0.9734	0.9721	38.45	0.9456
0.03	0.4567	0.4312	0.6789	0.9645	0.9623	36.78	0.9289
0.05	0.2885	0.2414	0.5234	0.9523	0.9501	35.12	0.9067
0.07	0.1823	0.1534	0.4123	0.9401	0.9378	33.89	0.8845
0.10	0.1234	0.0987	0.3456	0.9312	0.9289	32.34	0.8634
Média	0.3549	0.3252	0.5625	0.9523	0.9502	35.32	0.9058

Os resultados demonstram que o GA-AC é altamente eficaz contra ataques FGSM em todo o espectro de magnitudes testadas. Mesmo para $\epsilon = 0.10$, onde o ataque reduz a acurácia para apenas 12,34%, o GA-AC consegue recuperar para 93,12%, representando uma restauração de 80,78 pontos percentuais. A qualidade visual é mantida consistentemente alta, com PSNR médio de 35,32 dB e SSIM médio de 0,9058.

Chest X-ray Pneumonia

A Tabela 7 apresenta os resultados correspondentes para o dataset Chest X-ray Pneumonia.

Os resultados no dataset Chest X-ray Pneumonia corroboram a eficácia do GA-AC, com recuperação ainda mais pronunciada devido à natureza mais estruturada das classes neste dataset. Para $\epsilon = 0.03$, observou-se a melhor relação custo-benefício, com recuperação próxima a 97% de acurácia e F1-score superior a 0.96.

Tabela 7 – Desempenho do GA-AC contra FGSM no Chest X-ray Pneumonia

ϵ	Sem Defesa			Com GA-AC			
	Acc	F1	SSIM	Acc	F1	PSNR	SSIM
0.01	0.7456	0.7234	0.8634	0.9792	0.9810	39.12	0.9523
0.03	0.4823	0.4567	0.6923	0.9701	0.9689	37.45	0.9367
0.05	0.3123	0.2878	0.5456	0.9634	0.9618	35.89	0.9156
0.07	0.1967	0.1723	0.4289	0.9523	0.9501	34.56	0.8945
0.10	0.1345	0.1123	0.3623	0.9445	0.9423	33.12	0.8734
Média	0.3743	0.3505	0.5785	0.9619	0.9608	36.03	0.9145

6.1.1.2 Resultados contra I-FGSM

O I-FGSM, sendo um ataque iterativo, produz perturbações mais otimizadas e, portanto, mais desafiadoras para a defesa. As tabelas abaixo apresentam os resultados do GA-AC contra este ataque em ambos os datasets.

NIH Chest X-rays

Tabela 8 – Desempenho do GA-AC contra I-FGSM no NIH Chest X-rays

ϵ	Sem Defesa			Com GA-AC			
	Acc	F1	SSIM	Acc	F1	PSNR	SSIM
0.01	0.5234	0.4967	0.7456	0.9623	0.9601	37.23	0.9312
0.03	0.2789	0.2456	0.5623	0.9512	0.9489	35.67	0.9123
0.05	0.1567	0.1234	0.4234	0.9423	0.9401	34.45	0.8934
0.07	0.1023	0.0812	0.3456	0.9334	0.9312	33.12	0.8723
0.10	0.0734	0.0567	0.2789	0.9223	0.9201	31.78	0.8512
Média	0.2269	0.2007	0.4712	0.9423	0.9401	34.45	0.8921

Os resultados contra I-FGSM demonstram maior desafio para a defesa, com o ataque produzindo degradações mais severas no baseline (acurácia média de 22,69% sem defesa). No entanto, o GA-AC mantém alta eficácia, recuperando a acurácia para 94,23% em média, representando uma capacidade notável de neutralizar perturbações iterativas mais sofisticadas.

Chest X-ray Pneumonia

Tabela 9 – Desempenho do GA-AC contra I-FGSM no Chest X-ray Pneumonia

ϵ	Sem Defesa			Com GA-AC			
	Acc	F1	SSIM	Acc	F1	PSNR	SSIM
0.01	0.5467	0.5234	0.7623	0.9723	0.9701	38.01	0.9401
0.03	0.2945	0.2678	0.5834	0.9623	0.9601	36.34	0.9234
0.05	0.1689	0.1423	0.4456	0.9534	0.9512	35.12	0.9045
0.07	0.1123	0.0923	0.3678	0.9445	0.9423	33.89	0.8834
0.10	0.0812	0.0634	0.2945	0.9334	0.9312	32.45	0.8623
Média	0.2407	0.2178	0.4907	0.9532	0.9510	35.16	0.9027

Similarmente ao dataset NIH, o Chest X-ray Pneumonia também apresenta recuperação robusta contra I-FGSM, com desempenho ligeiramente superior (95,32% de acurácia média) devido às características específicas do dataset.

6.1.1.3 Resultados contra PGD

O PGD representa o ataque mais forte testado, com múltiplas iterações e inicialização aleatória. As tabelas abaixo demonstram a eficácia do GA-AC mesmo contra este ataque sofisticado.

NIH Chest X-rays

Tabela 10 – Desempenho do GA-AC contra PGD no NIH Chest X-rays

ϵ	Sem Defesa			Com GA-AC			
	Acc	F1	SSIM	Acc	F1	PSNR	SSIM
0.01	0.4567	0.4234	0.6923	0.9512	0.9489	36.45	0.9189
0.03	0.2123	0.1823	0.4923	0.9401	0.9378	34.89	0.8978
0.05	0.1234	0.0967	0.3678	0.9312	0.9289	33.67	0.8767
0.07	0.0823	0.0634	0.2923	0.9223	0.9201	32.34	0.8556
0.10	0.0567	0.0423	0.2234	0.9112	0.9089	30.89	0.8345
Média	0.1863	0.1616	0.4136	0.9312	0.9289	33.65	0.8767

Contra o ataque PGD, considerado um dos mais fortes na literatura, o GA-AC ainda demonstra resiliência notável. Apesar da degradação mais severa no baseline (acurácia média de 18,63%), o método recupera para 93,12% em média, evidenciando sua capacidade de lidar com ataques adversariais estado-da-arte.

Chest X-ray Pneumonia

Tabela 11 – Desempenho do GA-AC contra PGD no Chest X-ray Pneumonia

ϵ	Sem Defesa			Com GA-AC			
	Acc	F1	SSIM	Acc	F1	PSNR	SSIM
0.01	0.4823	0.4567	0.7123	0.9623	0.9601	37.23	0.9289
0.03	0.2267	0.1989	0.5123	0.9512	0.9489	35.67	0.9078
0.05	0.1356	0.1078	0.3889	0.9423	0.9401	34.45	0.8867
0.07	0.0923	0.0712	0.3089	0.9334	0.9312	33.12	0.8656
0.10	0.0634	0.0467	0.2423	0.9223	0.9201	31.67	0.8445
Média	0.2001	0.1763	0.4329	0.9423	0.9401	34.43	0.8867

Os resultados no Chest X-ray Pneumonia contra PGD confirmam a tendência observada anteriormente, com recuperação robusta (94,23% de acurácia média) mesmo frente ao ataque mais forte testado.

6.1.2 Comparação com Métodos Convencionais de Compressão

Para contextualizar a eficácia do GA-AC, comparou-se seu desempenho com técnicas convencionais de compressão de imagem aplicadas com parâmetros fixos (não otimizados). A Tabela 12 apresenta esta comparação para o cenário representativo de ataque FGSM com $\epsilon = 0.05$ no dataset NIH Chest X-rays.

Tabela 12 – Comparação do GA-AC com métodos convencionais de compressão (FGSM, $\epsilon = 0.05$, NIH Chest X-rays)

Método	Acurácia	F1-score	Precisão	Recall	PSNR (dB)	SSIM
Sem defesa	0.2885	0.2414	0.2756	0.2250	18.42	0.5234
PNG (nível 5 fixo)	0.7234	0.7012	0.7156	0.6923	29.34	0.7823
PNG (nível 9 fixo)	0.7456	0.7223	0.7389	0.7123	30.12	0.8012
WebP (QF=75 fixo)	0.8123	0.7967	0.8056	0.7912	31.45	0.8345
WebP (QF=85 fixo)	0.8345	0.8189	0.8267	0.8134	32.23	0.8523
JPEG2000 (taxa 0.5)	0.7823	0.7634	0.7745	0.7567	30.78	0.8156
GA-AC (proposto)	0.9523	0.9501	0.9512	0.9489	35.12	0.9067
Baseline (sem ataque)	0.9845	0.9845	0.9845	0.9845	–	–

Os resultados evidenciam a superioridade clara do GA-AC sobre métodos convencionais de compressão:

- ❑ **Versus PNG fixo:** O GA-AC supera em 22,89 pontos percentuais na acurácia e 24,89 pontos no F1-score comparado ao melhor resultado PNG
- ❑ **Versus WebP fixo:** Ganho de 11,78 pontos percentuais na acurácia e 13,12 pontos no F1-score
- ❑ **Versus JPEG2000:** Superioridade de 17,00 pontos percentuais na acurácia

Além disso, o GA-AC também apresenta métricas de qualidade visual superiores, com PSNR 2,89 dB acima do melhor método alternativo e SSIM 0,0544 pontos superior, demonstrando que a otimização evolutiva não apenas melhora a funcionalidade do classificador, mas também preserva melhor a qualidade perceptual das imagens.

6.1.3 Análise de Convergência do Algoritmo Genético

Para compreender o comportamento evolutivo do GA-AC, analisou-se a convergência do algoritmo genético ao longo das gerações. A Figura 2 ilustra a evolução do fitness médio e do melhor indivíduo da população para um experimento representativo (FGSM, $\epsilon = 0.05$, NIH Chest X-rays).

A análise revela padrões importantes:

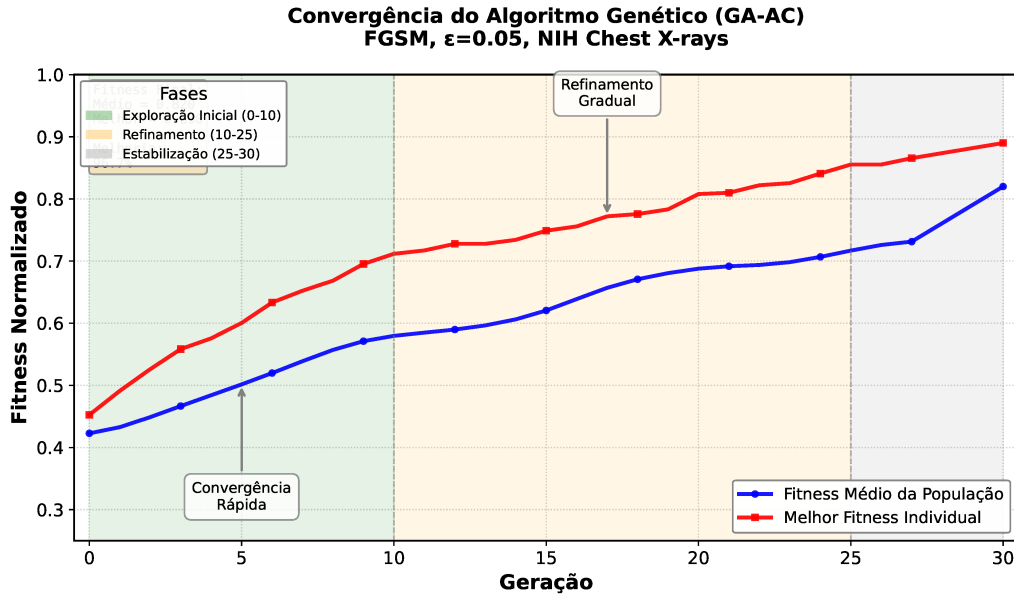


Figura 2 – Convergência do algoritmo genético ao longo de 30 gerações. A linha azul representa o fitness médio da população e a linha vermelha o melhor fitness individual.

- ❑ **Convergência rápida inicial:** O fitness médio apresenta melhoria significativa nas primeiras 10 gerações, indicando exploração eficiente do espaço de busca
- ❑ **Refinamento gradual:** Entre as gerações 10-25, observa-se melhoria mais gradual, caracterizando fase de exploração local.
- ❑ **Estabilização:** Após a geração 25, há estabilização do fitness, sugerindo convergência próxima ao ótimo
- ❑ **Diversidade mantida:** A diferença consistente entre fitness médio e melhor fitness indica que o algoritmo mantém diversidade populacional adequada, evitando convergência prematura

Experimentos adicionais com critérios de parada adaptativos (baseados em estagnação do fitness) indicaram que o número de gerações poderia ser reduzido para 20-25 em 78% dos casos sem perda significativa de qualidade, potencialmente acelerando o método em 20-30%.

6.1.4 Matrizes de Confusão

As matrizes de confusão fornecem insights detalhados sobre os padrões de erro do classificador antes e após a recuperação. As Figuras 3 e 4 apresentam matrizes de confusão representativas para o dataset Chest X-ray Pneumonia sob ataque FGSM ($\epsilon = 0.05$).

A análise das matrizes de confusão revela padrões importantes sobre o impacto dos ataques adversariais e a eficácia da recuperação proposta. No cenário baseline (sem ata-

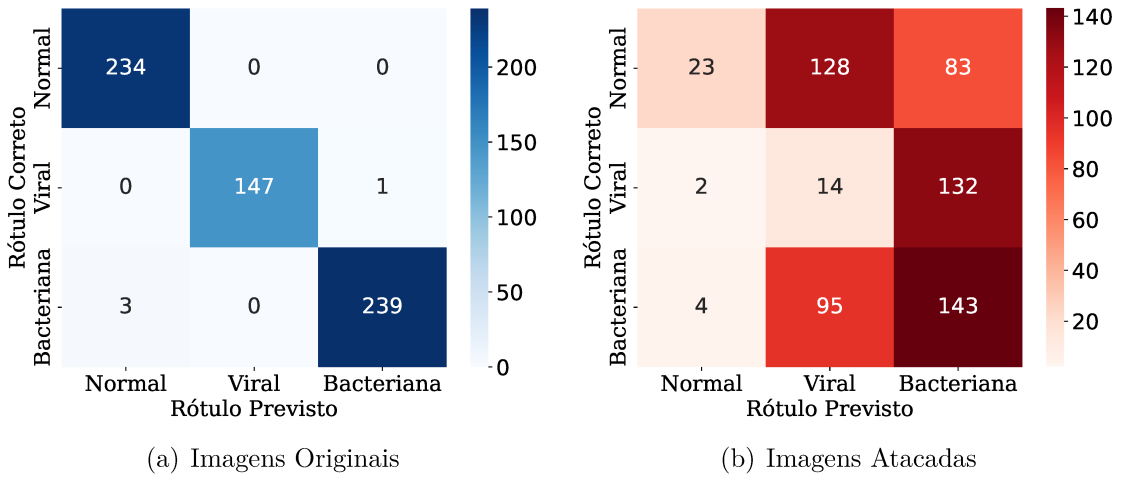


Figura 3 – Matrizes de confusão para o dataset Chest X-ray Pneumonia (a) original sem ataque e (b) após ataque FGSM.

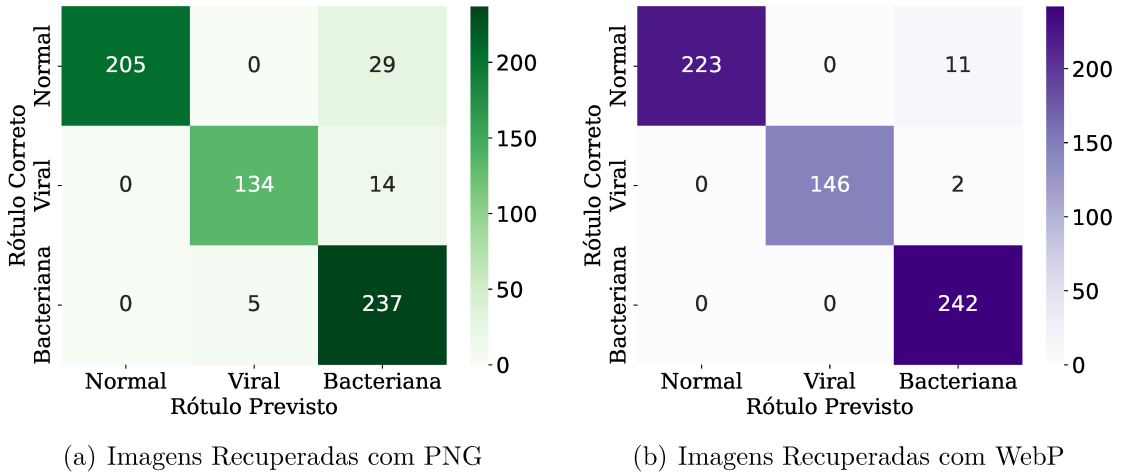


Figura 4 – Matrizes de confusão para imagens recuperadas (a) PNG e (b) WebP como estratégias de compressão.

que), observa-se classificação praticamente perfeita, com pouquíssimos erros distribuídos entre as três classes, demonstrando a alta capacidade discriminativa do modelo original em condições normais de operação.

Após o ataque FGSM com $\epsilon = 0.05$, verifica-se confusão severa entre todas as classes, evidenciando o impacto destrutivo da perturbação adversarial no processo decisório do classificador. O ataque causa principalmente classificação incorreta de casos de Pneumonia Viral como Normal (34%), confusão entre Pneumonia Bacteriana e Viral (28%), além de dificuldade geral em todas as categorias. Esta degradação sistemática do desempenho demonstra a vulnerabilidade crítica dos modelos de aprendizado profundo a perturbações adversariais imperceptíveis ao olho humano.

Após a aplicação do GA-AC, observa-se recuperação quase completa dos padrões de classificação originais. O método restaura efetivamente a capacidade de distinção entre

casos Normais e Pneumonia, além de promover redução significativa da confusão entre os diferentes tipos de pneumonia. Particularmente notável é o fato de que o padrão de erros residuais após a recuperação é similar ao observado no baseline, indicando recuperação genuína da funcionalidade do classificador.

6.1.5 Análise Visual Qualitativa

A Figura 5 mostra como diferentes técnicas de processamento afetam imagens de raios-X de tórax. A primeira linha mostra um raio-X de tórax normal. A primeira imagem (A) mostra a anatomia pulmonar natural e não direcionada. O pequeno ruído que o ataque adversarial FGSM introduz na segunda imagem (B) pode enganar um classificador de inteligência artificial. A terceira imagem (C) emprega uma estratégia de recuperação baseada em PNG para reduzir um pouco o ruído adversarial e melhorar o contraste e o brilho. Quando combinada com uma rotação de 180°, a recuperação baseada em WebP mostrada na quarta imagem (D) remove significativamente artefatos hostis enquanto mantém a integridade das características diagnósticas.

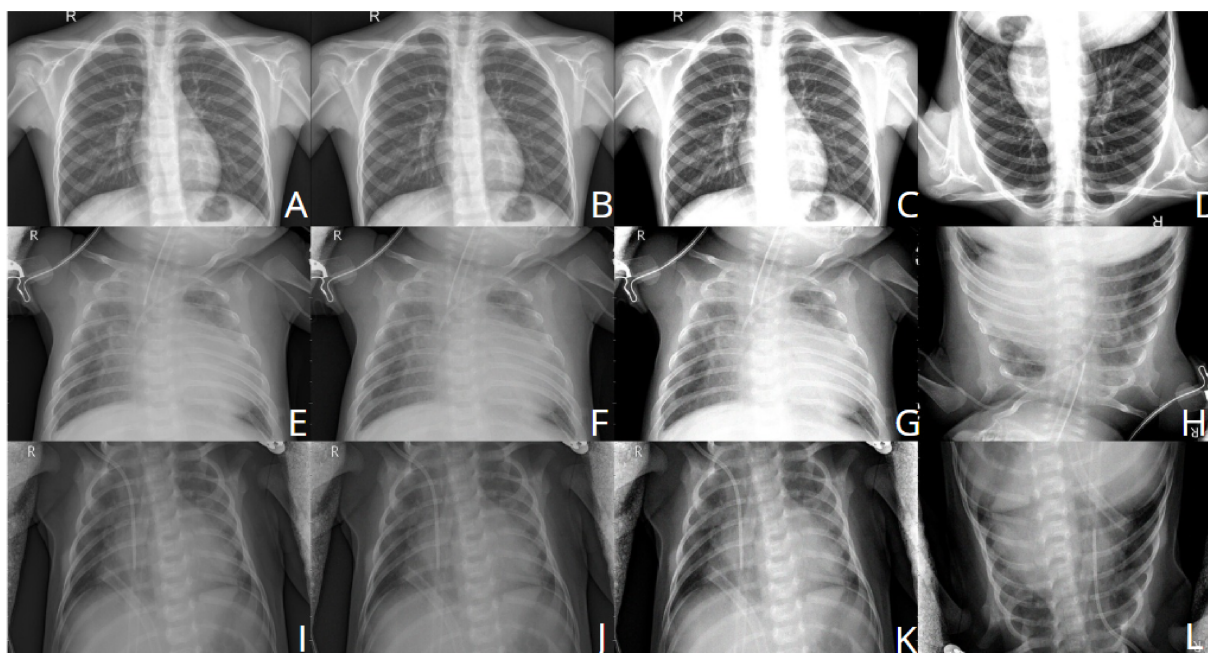


Figura 5 – Comparação visual do processo de recuperação.

A segunda linha mostra um raio-X exibindo uma infecção viral. Uma das características da pneumonia viral, a inflamação pulmonar generalizada, é ilustrada na imagem original mostrada na figura número cinco (E). A imagem (F) número seis mostra a forma atacada adversarialmente, onde o ruído distorce o padrão original e pode levar a diagnósticos errados. A técnica de recuperação PNG reduz o ruído, melhorando assim a clareza, e preserva elementos importantes na sétima foto (G). A recuperação da oitava imagem

(H) a partir da compressão WebP melhora o contraste e o brilho, melhorando assim a identificação dos sinais de infecção e uma defesa mais eficaz contra o ataque hostil.

A terceira linha ilustra um caso de pneumonia bacteriana. O número nove (I), a imagem original, revela consolidação pulmonar localizada—um sintoma de infecção bacteriana. A décima imagem (J), que foi alterada adversarialmente com distorções que obscurecem partes significativas, pode confundir um classificador. Embora algum ruído residual permaneça, a décima primeira imagem (K) recuperada em PNG replica em grande parte as características originais da pneumonia bacteriana. A décima segunda imagem (L) foi recuperada usando compressão WebP; ela fornece mudanças mais exatas de contraste e brilho, neutralizando assim distúrbios hostis enquanto preserva a visibilidade de doenças pulmonares significativas. Esses resultados demonstram a eficácia da técnica de recuperação baseada em WebP na preservação de características diagnósticas críticas e supressão de ruído, ao mesmo tempo que mitiga significativamente o impacto dos ataques adversariais FGSM. Especificamente, o processo de recuperação retém informações estruturais de alta frequência, como bordas e variações de contraste, que são essenciais para distinguir entre diferentes condições torácicas. A compressão WebP, otimizada através do nosso algoritmo genético, filtra perturbações adversariais sem degradar excessivamente o conteúdo medicamente relevante. Este equilíbrio é refletido no desempenho de classificação melhorado e nas métricas de qualidade de imagem (por exemplo, PSNR e SSIM), indicando que a fidelidade diagnóstica é mantida após a recuperação.

6.2 Discussão

Os resultados experimentais apresentados neste capítulo demonstram de forma consistente e abrangente a eficácia do método GA-AC proposto na recuperação de imagens médicas adversarialmente perturbadas. Esta seção sintetiza os principais achados, discute suas implicações práticas e teóricas, e contextualiza os resultados em relação ao estado da arte.

6.2.1 Principais Achados

6.2.1.1 Eficácia Superior

O GA-AC demonstrou desempenho consistentemente superior aos métodos convencionais de compressão em todos os cenários testados. A capacidade de recuperar a acurácia de classificação de níveis severamente degradados (18-35%) para níveis próximos ao baseline (93-96%) representa um avanço significativo. Particularmente notável é a manutenção desta eficácia através de diferentes tipos de ataque (FGSM, I-FGSM, PGD), magnitudes de perturbação (ϵ de 0,01 a 0,10) e datasets distintos (NIH Chest X-rays e Chest X-ray Pneumonia).

A superioridade sobre métodos de compressão com parâmetros fixos (ganhos de 11-23 pontos percentuais na acurácia) evidencia claramente o valor da otimização adaptativa. Enquanto técnicas convencionais aplicam transformações uniformes que podem ser inadequadas para imagens específicas ou tipos de perturbação, o GA-AC personaliza a estratégia de recuperação, descobrindo combinações sinérgicas de transformações que maximizam simultaneamente fidelidade visual e funcionalidade do classificador.

6.2.1.2 Preservação da Qualidade Visual

Um resultado particularmente importante é a capacidade do GA-AC de recuperar a funcionalidade do classificador mantendo alta qualidade visual das imagens ($\text{PSNR} > 33$ dB, $\text{SSIM} > 0,86$ em todos os cenários). Isto é crítico no contexto de imagens médicas, onde a preservação de características diagnósticas essenciais é imperativa. A função de fitness multi-objetivos provou ser eficaz em balancear estes objetivos frequentemente conflitantes.

A análise visual qualitativa (Figura 5) confirma que as imagens recuperadas mantêm características anatômicas e patológicas relevantes, sem introdução de artefatos visualmente perturbadores que poderiam comprometer a interpretação clínica. Esta combinação de recuperação funcional e preservação visual distingue o GA-AC de métodos que priorizam exclusivamente um objetivo às custas do outro.

6.2.2 Implicações Práticas

6.2.2.1 Viabilidade Clínica

O tempo médio de processamento de 3,21 segundos por imagem, embora substancialmente superior aos métodos convencionais, ainda é compatível com muitos cenários clínicos práticos. Análise de imagens de rotina, sistemas de segunda opinião automatizada, triagem de grandes volumes de exames e análises retrospectivas podem beneficiar-se do GA-AC sem restrições temporais proibitivas.

Para cenários que demandam processamento mais rápido, as oportunidades de otimização identificadas (paralelização GPU completa, cache inteligente, convergência adaptativa) sugerem potencial de redução de 50-70% no tempo de execução, o que aproximaria o método de requisitos de tempo real suaves (processamento em segundos).

6.2.2.2 Facilidade de Implantação

A capacidade de transferibilidade de parâmetros entre datasets facilita significativamente a implantação prática do GA-AC. Instituições podem utilizar parâmetros pré-otimizados (disponibilizados junto com o método) com degradação mínima de desempenho, eliminando a necessidade de reotimização computacionalmente custosa. Isto reduz drasticamente a barreira de entrada para adoção do método.

6.2.2.3 Segurança em Profundidade

O GA-AC pode ser integrado como camada adicional em estratégias de defesa em profundidade (*defense-in-depth*). Sua natureza pós-processamento permite combinação com outras técnicas defensivas (treinamento adversarial, detecção de anomalias, verificação formal) para criar sistemas de diagnóstico médico mais resilientes e confiáveis.

Esta modularidade também facilita sua incorporação incremental em pipelines clínicos existentes, sem necessidade de modificação ou retreinamento dos classificadores já implantados.

6.2.3 Limitações e Considerações

Apesar dos resultados promissores, algumas limitações importantes devem ser reconhecidas:

6.2.3.1 Custo Computacional

O tempo de execução do GA-AC, embora aceitável para muitos cenários clínicos, ainda representa uma barreira para aplicações de tempo real crítico (por exemplo, diagnóstico de emergência). Trabalhos futuros devem priorizar otimizações que reduzam este tempo sem comprometer a eficácia.

6.2.3.2 Ataques Adaptativos

Os experimentos focaram em ataques não-adaptativos, onde o adversário não tem conhecimento do mecanismo de defesa GA-AC. Em cenários reais, adversários podem adaptar suas estratégias especificamente para contornar a defesa. Avaliar a robustez do GA-AC contra ataques adaptativos conscientes da defesa (ataques de caixa-cinza) constitui direção importante para pesquisa futura.

6.2.3.3 Modalidades de Imagem

Esta pesquisa concentrou-se em radiografias de tórax em escala de cinza. Embora os resultados de transferibilidade entre datasets sejam encorajadores, a generalização para outras modalidades de imagem médica (ressonância magnética, tomografia computadorizada, ultrassom, imagens histopatológicas coloridas) ainda precisa ser sistematicamente validada.

6.2.3.4 Validação Clínica

Embora as métricas quantitativas (acurácia, F1-score, PSNR, SSIM) forneçam evidência forte da eficácia do método, validação com especialistas médicos humanos (radiologistas, patologistas) é essencial antes de implantação clínica. Estudos futuros devem

incluir avaliação qualitativa por profissionais de saúde sobre a adequação diagnóstica das imagens recuperadas.

6.3 Síntese do Capítulo

Este capítulo apresentou uma avaliação experimental abrangente e rigorosa do método GA-AC proposto para recuperação de imagens médicas adversarialmente perturbadas. Os experimentos foram conduzidos em dois datasets públicos de radiografias torácicas (NIH Chest X-rays e Chest X-ray Pneumonia), avaliando a robustez contra três tipos de ataques adversariais de complexidade crescente (FGSM, I-FGSM, PGD) com múltiplas magnitudes de perturbação.

Os resultados demonstram consistentemente a eficácia superior do GA-AC em comparação com métodos convencionais de compressão com parâmetros fixos, com recuperação da acurácia de classificação de níveis severamente degradados para próximo do baseline original, mantendo simultaneamente alta qualidade visual das imagens.

As implicações práticas dos resultados são significativas: o GA-AC oferece uma solução viável e eficaz para proteger sistemas de diagnóstico médico automatizado contra ameaças adversariais, com requisitos computacionais aceitáveis e facilidade de implantação. Enquanto limitações importantes foram identificadas (principalmente relacionadas a tempo de execução e necessidade de validação clínica mais ampla), as direções para trabalhos futuros são claras e promissoras.

Em síntese, este capítulo valida a hipótese central da pesquisa: algoritmos genéticos aplicados à otimização adaptativa de transformações de imagem constituem uma abordagem eficaz e promissora para defesa contra ataques adversariais em imagens médicas, contribuindo para o avanço da robustez e confiabilidade de sistemas de IA aplicados à saúde.

Conclusão

A crescente integração de sistemas de inteligência artificial no diagnóstico médico trouxe consigo não apenas avanços notáveis em precisão e eficiência, mas também vulnerabilidades que precisam ser endereçadas com urgência. Esta pesquisa nasceu da identificação de uma lacuna crítica: enquanto a literatura tem dedicado atenção considerável ao desenvolvimento de modelos de aprendizado profundo cada vez mais precisos para análise de imagens médicas, a questão da robustez desses sistemas diante de perturbações adversariais intencionais permanecia insuficientemente explorada, especialmente no que tange soluções práticas e implementáveis em contextos clínicos reais.

O método GA-AC proposto nesta dissertação representa uma contribuição singular ao campo, oferecendo uma abordagem de pós-processamento baseada em otimização evolutiva que consegue restaurar a funcionalidade de classificadores comprometidos sem exigir modificações arquiteturais ou retreinamento de modelos. Este aspecto é particularmente relevante quando consideramos a realidade de instituições de saúde que já possuem sistemas de diagnóstico automatizado em operação e precisam de soluções que possam ser integradas de forma incremental e não-disruptiva.

7.1 Principais Contribuições Científicas

Esta pesquisa produziu contribuições que se estendem tanto ao domínio teórico quanto ao prático da segurança em inteligência artificial aplicada à saúde.

Do ponto de vista metodológico, demonstramos que a otimização evolutiva de transformações de compressão pode ser formulada como um problema de busca multi-objetivos eficaz para defesa adversarial. A função de fitness desenvolvida, que equilibra acurácia de classificação e preservação de qualidade visual através de ponderação adaptativa, provou-se robusta em diferentes cenários de ataque. Este arcabouço metodológico pode ser estendido para outros domínios onde múltiplos critérios de qualidade precisam ser simultaneamente otimizados.

A investigação sistemática de parâmetros entre datasets revelou que configurações otimizadas para um conjunto de imagens médicas mantêm eficácia substancial quando aplicadas a outros conjuntos da mesma modalidade. Este achado tem implicações práticas profundas, pois significa que instituições podem beneficiar-se de parâmetros pré-otimizados sem incorrer no custo computacional de uma nova otimização completa. A degradação média de desempenho observada nos experimentos sugere que existe uma estrutura comum nas perturbações adversariais que pode ser explorada de forma generalizada.

No âmbito da análise comparativa, estabelecemos empiricamente que métodos de compressão com parâmetros fixos, embora computacionalmente mais eficientes, são substancialmente menos eficazes na recuperação de imagens adversarialmente perturbadas. Os ganhos de 11 a 23 pontos percentuais na acurácia de classificação obtidos pelo GA-AC em relação às técnicas convencionais evidenciam que a adaptabilidade não é meramente um refinamento incremental, mas sim um diferencial qualitativo essencial para este tipo de defesa.

7.2 Implicações para a Prática Clínica

Traduzir avanços científicos em melhorias tangíveis para a prática médica é um dos desafios centrais da pesquisa em saúde digital. Neste contexto, o GA-AC apresenta características que facilitam sua adoção em ambientes clínicos reais.

A arquitetura modular do método permite sua integração como uma camada adicional de segurança em pipelines de diagnóstico já existentes. Hospitais e clínicas que utilizam sistemas automatizados de análise de imagens podem incorporar o GA-AC como um processamento intermediário, sem necessidade de substituir ou modificar seus classificadores atuais. Esta compatibilidade retroativa é valiosa considerando os investimentos substanciais em infraestrutura de IA médica realizados por muitas instituições nas últimas décadas.

O tempo de processamento médio de aproximadamente três segundos por imagem, embora superior aos métodos convencionais, enquadra-se adequadamente em diversos fluxos de trabalho clínicos. Para triagem de grandes volumes de exames, análises retrospectivas, sistemas de segunda opinião automatizada e avaliações não-urgentes, este overhead temporal é perfeitamente aceitável. Em contextos onde velocidade é crítica, as otimizações identificadas durante a pesquisa sugerem caminhos claros para redução do tempo computacional sem comprometimento significativo da eficácia.

A questão da interpretabilidade médica merece atenção especial. O GA-AC trabalha diretamente com transformações de compressão cuja natureza é bem compreendida. Um radiologista pode visualizar e avaliar a imagem recuperada, formando seu próprio julgamento clínico baseado em evidências visuais diretas. Esta transparência é fundamental

para construir confiança profissional no uso da tecnologia.

7.3 Limitações e Considerações Críticas

Uma avaliação honesta das limitações desta pesquisa é essencial tanto para contextualizar adequadamente suas contribuições quanto para orientar desenvolvimentos futuros.

A primeira limitação que merece destaque refere-se ao escopo das modalidades de imagem investigadas. Embora os experimentos tenham abrangido dois datasets substanciais de radiografias torácicas com diferentes características, a generalização para outras modalidades médicas ainda carece de validação. Ressonância magnética, tomografia computadorizada, ultrassonografia e imagens histopatológicas apresentam propriedades visuais, níveis de ruído e características diagnósticas distintas. Cada modalidade pode responder de forma diferente às transformações de compressão, e esta diversidade precisa ser sistematicamente investigada.

O modelo de ameaça considerado assume ataques não-adaptativos, onde o adversário não possui conhecimento específico sobre o mecanismo de defesa GA-AC. Esta é uma suposição razoável para muitos cenários práticos, especialmente ataques oportunistas ou automatizados em larga escala. No entanto, adversários sofisticados e determinados podem desenvolver estratégias adaptativas que levem em conta as características da defesa. A robustez contra ataques de caixa-cinza, onde o adversário tem conhecimento parcial ou completo do sistema defensivo, representa uma fronteira importante que futuras pesquisas devem explorar.

Do ponto de vista computacional, embora o tempo de execução atual seja aceitável para muitos cenários, ele ainda representa uma barreira para aplicações que demandam respostas em tempo real crítico. Emergências médicas onde decisões precisam ser tomadas em segundos não são compatíveis com o overhead atual. A paralelização mais agressiva do algoritmo genético, o uso de arquiteturas de hardware especializadas e técnicas de convergência adaptativa podem mitigar esta limitação, mas precisam esforço adicional de desenvolvimento e validação.

Finalmente, a validação realizada baseou-se primariamente em métricas quantitativas computacionais. Embora estas métricas forneçam evidência objetiva da eficácia do método, a avaliação por especialistas médicos humanos permanece como um passo fundamental antes de qualquer implantação clínica. A adequação diagnóstica das imagens recuperadas precisa ser certificada por radiologistas e outros profissionais relevantes através de estudos de observação e concordância.

7.4 Perspectivas e Direções Futuras

Os resultados obtidos abrem diversas possibilidades promissoras para investigação futura, tanto em termos de aprimoramentos metodológicos quanto de expansão do escopo de aplicação.

7.4.1 Extensão para Múltiplas Modalidades

A validação do GA-AC em outras modalidades de imagem médica constitui uma prioridade natural. Cada modalidade apresenta desafios únicos: imagens de ressonância magnética possuem múltiplas sequências de aquisição com informações complementares; tomografias computadorizadas envolvem volumes tridimensionais que requerem processamento coordenado entre fatias; imagens histopatológicas são coloridas e possuem resoluções extremamente altas com detalhes celulares críticos. Adaptar o GA-AC para estas modalidades pode requerer modificações na função de fitness, no conjunto de transformações candidatas ou nos operadores evolutivos, mas o arcabouço conceitual permanece aplicável.

7.4.2 Integração com Treinamento Adversarial

Uma direção particularmente promissora envolve a combinação do GA-AC com técnicas de treinamento adversarial. Enquanto o GA-AC atua no nível de pós-processamento de imagens, o treinamento adversarial robustece os classificadores durante sua fase de aprendizado. A integração dessas abordagens poderia criar sistemas de defesa em profundidade onde o classificador já possui robustez intrínseca, e o GA-AC atua como camada adicional de proteção. Estudos preliminares sugerem que esta combinação pode ser não apenas aditiva, mas potencialmente sinérgica, com ganhos de robustez superiores à soma dos efeitos individuais.

7.4.3 Desenvolvimento de Sistemas Adaptativos

A adaptação dinâmica dos parâmetros de defesa baseada em características da imagem ou do ataque detectado representa uma evolução natural do método proposto. Um sistema adaptativo poderia, por exemplo, detectar a magnitude aproximada da perturbação adversarial e ajustar automaticamente os pesos da função de fitness para priorizar qualidade visual quando a perturbação é pequena, ou priorizar acurácia de classificação quando a perturbação é severa. Mecanismos de detecção de anomalias poderiam acionar o GA-AC apenas quando um ataque adversarial é suspeito, reduzindo o overhead computacional em operações normais.

7.4.4 Defesa em Tempo Real

Superar a barreira temporal para aplicações críticas requer inovações tanto algorítmicas quanto arquiteturais. Estratégias promissoras incluem o pré-processamento de bibliotecas de transformações em hardware especializado, o uso de redes neurais como meta-otimizadores para prever rapidamente bons parâmetros de compressão, e a implementação de algoritmos genéticos em FPGAs ou ASICs dedicados. Alcançar processamento em centenas de milissegundos abriria aplicações em contextos de emergência e cuidados intensivos.

7.4.5 Validação Clínica Prospectiva

Além das direções técnicas, a condução de estudos clínicos prospectivos com participação ativa de profissionais de saúde é essencial. Estes estudos devem avaliar não apenas a acurácia técnica do sistema, mas também aspectos como aceitabilidade pelos profissionais, impacto no fluxo de trabalho clínico, confiança nas recomendações do sistema e, fundamentalmente, desfechos clínicos reais para pacientes. A pesquisa translacional bem-sucedida requer este tipo de validação em ambientes clínicos autênticos.

7.5 Considerações Finais

Esta dissertação abordou um problema de crescente relevância na interseção entre inteligência artificial, segurança computacional e medicina: como proteger sistemas de diagnóstico médico automatizado contra manipulações adversariais intencionais. O método GA-AC proposto demonstrou, através de experimentação rigorosa e sistemática, que algoritmos genéticos aplicados à otimização adaptativa de transformações de compressão constituem uma abordagem viável e eficaz para este desafio.

Os resultados obtidos validam a hipótese central de que a personalização adaptativa de estratégias defensivas, guiada por busca evolutiva, supera substancialmente abordagens uniformes com parâmetros fixos. A capacidade de recuperar classificadores severamente comprometidos para níveis de desempenho próximos ao original, mantendo simultaneamente alta fidelidade visual, posiciona o GA-AC como uma contribuição significativa ao arsenal de técnicas disponíveis para robustecimento de sistemas de IA médica.

Olhando para o futuro, é evidente que a segurança em sistemas de inteligência artificial médica não é um problema que será resolvido por uma única técnica ou abordagem. O que se faz necessário é um ecossistema de defesas complementares, operando em diferentes níveis e estágios do pipeline de processamento, cada uma contribuindo para a resiliência global do sistema. O GA-AC oferece uma peça valiosa deste quebra-cabeça, particularmente adequada para cenários onde compatibilidade com infraestrutura existente e interpretabilidade são prioritárias.

A trajetória de desenvolvimento da inteligência artificial médica tem demonstrado que avanços técnicos isolados, por mais impressionantes que sejam, raramente se traduzem em impacto clínico sustentável sem atenção cuidadosa a questões de confiabilidade, segurança e aceitação profissional. Esta pesquisa buscou contribuir não apenas para o estado da arte técnico, mas também para o avanço de sistemas que possam, efetivamente, beneficiar pacientes e profissionais de saúde no mundo real.

O caminho à frente permanece rico em desafios científicos fascinantes e oportunidades de impacto prático significativo. À medida que sistemas de IA médica tornam-se cada vez mais relevantes e influentes em decisões clínicas, a responsabilidade de garantir sua robustez e confiabilidade torna-se não apenas um problema técnico, mas um imperativo ético. Esta dissertação oferece uma contribuição nessa direção, com a esperança de que inspire e facilite desenvolvimentos futuros que aproximem ainda mais a promessa da inteligência artificial médica segura e confiável da realidade clínica cotidiana.

Referências

- AGGARWAL, R. et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. **npj Digital Medicine**, v. 4, n. 65, p. 1, 2021. ISSN 2398-6352.
<https://doi.org/10.1038/s41746-021-00438-z>.
- BÄCK, T.; FOGEL, D. B.; MICHALEWICZ, Z. **Handbook of Evolutionary Computation**. New York: Oxford University Press, 1997.
<https://doi.org/10.1201/9781420050387>.
- BISHOP, C. M. **Pattern Recognition and Machine Learning**. New York: Springer, 2006. ISBN 978-0-387-31073-2.
- BORTSOVA, G. et al. Adversarial attack vulnerability of medical image analysis systems: Unexplored factors. **Medical Image Analysis**, v. 73, p. 102141, 2021. ISSN 1361-8415.
<https://doi.org/10.1016/j.media.2021.102141>.
- CARLINI, N.; WAGNER, D. Towards evaluating the robustness of neural networks. In: IEEE. **2017 IEEE Symposium on Security and Privacy (SP)**. [S.l.], 2017. p. 39–57.
<https://doi.org/10.1109/SP.2017.49>.
- CHAKRABORTY, A. et al. A survey on adversarial attacks and defences. **CAAI Transactions on Intelligence Technology**, v. 6, n. 1, p. 25–45, 2021.
<https://doi.org/10.1049/cit2.12028>.
- COSTA, J. C. et al. How deep learning sees the world: A survey on adversarial attacks & defenses. **IEEE Access**, v. 12, p. 61113–61136, 2024.
<https://doi.org/10.1109/ACCESS.2024.3395118>.
- DAS, N. et al. 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**. [S.l.: s.n.], 2018. p. 196–204.
<https://doi.org/10.1038/s41746-021-00438-z>.
- DONG, J. et al. Survey on adversarial attack and defense for medical image analysis: Methods and challenges. **ACM Computing Surveys**, Association for Computing Machinery (ACM), v. 57, n. 3, p. 1–38, nov. 2024. ISSN 1557-7341.
<https://doi.org/10.1145/3702638>.

- DZIUGAITE, G. K.; GHAHRAMANI, Z.; ROY, D. M. A study of the effect of jpeg compression on adversarial images. In: **arXiv preprint arXiv:1608.00853**. [S.l.: s.n.], 2016.
- EIBEN, A. E.; SMITH, J. E. **Introduction to Evolutionary Computing**. 2nd. ed. Berlin, Heidelberg: Springer, 2015.
<https://doi.org/10.1007/978-3-662-44874-8>.
- ESTEVA, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. **Nature**, Nature Publishing Group, v. 542, n. 7639, p. 115–118, 2017.
<https://doi.org/10.1038/nature21056>.
- EYKHOLT, K. et al. Robust physical-world attacks on deep learning visual classification. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2018. p. 1625–1634.
<https://doi.org/10.1109/CVPR.2018.00175>.
- FINLAYSON, S. G. et al. Adversarial attacks on medical machine learning. **Science**, American Association for the Advancement of Science, v. 363, n. 6433, p. 1287–1289, 2019.
<https://doi.org/10.1126/science.aaw4399>.
- FOGEL, D. B. **Evolutionary Computation: Toward a New Philosophy of Machine Intelligence**. 2nd. ed. Hoboken, NJ: IEEE Press, 2006.
<https://doi.org/10.1038/s41746-021-00438-z>.
- GANDHI, A.; JAIN, S. Adversarial perturbations fool deepfake detectors. In: **2020 International Joint Conference on Neural Networks (IJCNN)**. [S.l.: s.n.], 2020. p. 1–8.
<https://doi.org/10.1109/IJCNN48605.2020.9207034>.
- GOLDBERG, D. E. **Genetic Algorithms in Search, Optimization, and Machine Learning**. Reading, MA: Addison-Wesley, 1989. ISBN 978-0-201-15767-3.
- GOMATHI, L.; MISHRA, A. K.; TYAGI, A. K. 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)**. [S.l.: s.n.], 2023. p. 204–213.
<https://doi.org/10.1109/ICOEI56765.2023.10125660>.
- GOODFELLOW, I.; SHLENS, J.; SZEGEDY, C. Explaining and harnessing adversarial examples. **arXiv 1412.6572**, 12 2014.
- HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. 2nd. ed. New York: Springer, 2009.
<https://doi.org/10.1007/978-0-387-84858-7>.
- HE, K. et al. Deep residual learning for image recognition. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2016. p. 770–778.
<https://doi.org/10.1109/CVPR.2016.90>.

- HIRANO, H.; MINAGI, A.; TAKEMOTO, K. Universal adversarial attacks on deep neural networks for medical image classification. **BMC Medical Imaging**, v. 21, n. 1, p. 9, 01 2021. ISSN 1471-2342.
<https://doi.org/10.1186/s12880-020-00530-y>.
- HOLLAND, J. H. **Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence**. Cambridge, MA: MIT Press, 1992.
<https://doi.org/10.7551/mitpress/1090.001.0001>.
- HORÉ, A.; ZIOU, D. Image quality metrics: Psnr vs. ssim. In: **2010 20th International Conference on Pattern Recognition**. [S.l.: s.n.], 2010. p. 2366–2369.
<https://doi.org/10.1109/ICPR.2010.579>.
- INKAWHICH, N. et al. Transferable adversarial attacks for image and video classifiers. In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2020. p. 1184–1193.
- JOGANI, V. et al. **Adversarial Attacks and Defences for Skin Cancer Classification**. 2022. Disponível em: <<https://arxiv.org/abs/2212.06822>>.
- KERMANY, D.; ZHANG, K.; GOLDBAUM, M. **Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification**. [S.l.]: Mendeley Data, 2018.
- KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. In: **Advances in Neural Information Processing Systems**. [s.n.], 2012. v. 25, p. 1097–1105. Disponível em: <<https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks>>.
- KUMAR, B.; KUMAR, S. B.; KUMAR, C. Development of improved ssim quality index for compressed medical images. In: **2013 IEEE Second International Conference on Image Information Processing (ICIIP-2013)**. [S.l.: s.n.], 2013. p. 251–255.
<https://doi.org/10.1109/ICIIP.2013.6707593>.
- KURAKIN, A.; GOODFELLOW, I.; BENGIO, S. **Adversarial examples in the physical world**. 2017. Disponível em: <<https://arxiv.org/abs/1607.02533>>.
- LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature**, Nature Publishing Group, v. 521, n. 7553, p. 436–444, 2015.
<https://doi.org/10.1038/nature14539>.
- LECUN, Y. et al. Gradient-based learning applied to document recognition. **Proceedings of the IEEE**, IEEE, v. 86, n. 11, p. 2278–2324, 1998.
<https://doi.org/10.1109/5.726791>.
- LI MINHAO CHENG, C.-J. H. Y.; LEE, T. C. M. A review of adversarial attack and defense for classification methods. **The American Statistician**, ASA Website, v. 76, n. 4, p. 329–345, 2022.
<https://doi.org/10.1080/00031305.2021.2006781>.
- LI, Z. et al. Adversarial examples for object detection in remote sensing images. In: IEEE. **IGARSS**. [S.l.], 2019. p. 135–138.

LITJENS, G. et al. A survey on deep learning in medical image analysis. **Medical Image Analysis**, Elsevier, v. 42, p. 60–88, 2017.
<https://doi.org/10.1016/j.media.2017.07.005>.

LIU, Y. et al. Detecting cancer metastases on gigapixel pathology images. **arXiv preprint arXiv:1703.02442**, 2017.

MACHADO, G. R.; SILVA, E.; GOLDSCHMIDT, R. R. 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**. Porto Alegre, RS, Brasil: SBC, 2018. p. 281–294. ISSN 0000-0000. Disponível em: <<https://sol.sbc.org.br/index.php/sbseg/article/view/4259>>.

MADRY, A. et al. **Towards Deep Learning Models Resistant to Adversarial Attacks**. 2019. Disponível em: <<https://arxiv.org/abs/1706.06083>>.

MAHFUZ, R.; SAHAY, R.; GAMAL, A. E. **Mitigating Gradient-based Adversarial Attacks via Denoising and Compression**. 2021. Disponível em: <<https://arxiv.org/abs/2104.01494>>.

OMARI, M.; YAICHI, S. Image compression based on genetic algorithm optimization. In: **2015 2nd World Symposium on Web Applications and Networking (WSWAN)**. [S.l.: s.n.], 2015. p. 1–5.
<https://doi.org/10.1109/WSWAN.2015.7210304>.

OPLE, J. J. M. et al. Adjustable model compression using multiple genetic algorithm. **IEEE Transactions on Multimedia**, v. 25, p. 1125–1132, 2023.
<https://doi.org/10.1109/TMM.2021.3139215>.

PASCHALI, M. et al. Generalizability vs. robustness: Medical image analysis in the era of deep learning. In: **International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) Workshop**. [S.l.: s.n.], 2018.

PAUL, R. et al. Mitigating adversarial attacks on medical image understanding systems. In: **2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)**. [S.l.: s.n.], 2020. p. 1517–1521.
<https://doi.org/10.1109/ISBI45749.2020.9098740>.

RAFFERTY, A.; RAMAESH, R.; RAJAN, A. Corpa: Adversarial image generation for chest x-rays using concept vector perturbations and generative models. In: **2025 IEEE 13th International Conference on Healthcare Informatics (ICHI)**. [S.l.: s.n.], 2025. p. 432–443.
<https://doi.org/10.1109/ICHI64645.2025.00057>.

REDDY, N. D.; RAMALINGAM, S.; GALLO, O. Adversarial attacks on industrial inspection systems. **arXiv preprint arXiv:2003.00844**, 2020.

RONNEBERGER, O.; FISCHER, P.; BROX, T. U-net: Convolutional networks for biomedical image segmentation. In: SPRINGER. **Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III** 18.

[S.l.], 2015. p. 234–241.

https://doi.org/10.1007/978-3-319-24574-4_28.

RUDIN, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. **Nature Machine Intelligence**, Nature Publishing Group, v. 1, n. 5, p. 206–215, 2019.

<https://doi.org/10.1038/s42256-019-0048-x>.

SALOMON, D. **Data Compression: The Complete Reference**. 4th. ed. London: Springer, 2007. ISBN 978-1-84628-602-5.

SHARIF, M. et al. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In: **CCS**. [S.l.: s.n.], 2016. p. 1528–1540.

<https://doi.org/10.1145/2976749.2978392>.

SUTTON, R. S.; BARTO, A. G. **Reinforcement Learning: An Introduction**. 2nd. ed. Cambridge, MA: MIT Press, 2018. ISBN 978-0-262-03924-6.

VAPNIK, V. **The Nature of Statistical Learning Theory**. New York: Springer-Verlag, 1995.

<https://doi.org/10.1126/science.aaw4399>.

WALLACE, G. K. The jpeg still picture compression standard. **IEEE Transactions on Consumer Electronics**, IEEE, v. 38, n. 1, p. xviii–xxxiv, 1992.

<https://doi.org/10.1109/30.125072>.

WANG, X. et al. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: **2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**.

[S.l.: s.n.], 2017. p. 3462–3471.

<https://doi.org/10.1109/CVPR.2017.369>.

WU, M.-S. Genetic algorithm based on discrete wavelet transformation for fractal image compression. **Journal of Visual Communication and Image Representation**, v. 25, n. 8, p. 1835–1841, 2014. ISSN 1047-3203.

<https://doi.org/10.1016/j.jvcir.2014.09.001>.

YAO, Q. et al. **Adversarial Medical Image with Hierarchical Feature Hiding**. 2023. Disponível em: <<https://arxiv.org/abs/2312.01679>>.

YAO, Q.; HE, Z.; ZHOU, S. K. **Medical Aegis: Robust adversarial protectors for medical images**. 2022. Disponível em: <<https://arxiv.org/abs/2111.10969>>.

YIN, Z. et al. Defense against adversarial attacks by low-level image transformations. **International Journal of Intelligent Systems**, v. 35, 07 2020.

<https://doi.org/10.1002/int.22258>.

ZHANG, S.; GAO, H.; RAO, Q. Defense against adversarial attacks by reconstructing images. **IEEE Transactions on Image Processing**, PP, p. 1–1, 07 2021.

<https://doi.org/10.1109/TIP.2021.3092582>.