UNIVERSIDADE ESTADUAL PAULISTA - UNESP

Instituto de Biociências, Letras e Ciências Exatas- campus de São José do Rio Preto

VINICIUS AUGUSTO TORELI BORGUE

Associação de Redes Adversárias Generativas e Inteligência Artificial Explicável para Análise de Imagens H&E via Vision Transformers

São José do Rio Preto

2025

Vinicius Augusto Toreli Borgue

Associação de Redes Adversárias Generativas e Inteligência Artificial Explicável para Análise de Imagens H&E via Vision Transformers

Monografia, apresentada à Universidade Estadual Paulista (UNESP), Instituto de Biociências, Letras e Ciências Exatas, São José do Rio Preto, para obtenção do título de Bacharealem Ciência da Computação.

Orientador: Profº Dr. Leandro Alves Neves

São José do Rio Preto
2025

B734a

Borgue, Vinicius Augusto Toreli
Associação de Redes Adversárias Generativas e Inteligência Artificial Explicável para Análise de Imagens H&E via Vision Transformers / Vinicius Augusto Toreli Borgue. -- São José do Rio Preto, 2025
51 f. : il., tabs., fotos

Trabalho de conclusão de curso (Bacharelado - Ciência da Computação) - Universidade Estadual Paulista (UNESP), Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto
Orientador: Leandro Alves Neves

1. Redes neurais (Computação). 2. Inteligência artificial Aplicações médicas. 3. Processamento de imagens. 4. Histopatologia. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

VINICIUS AUGUSTO TORELI BORGUE

ASSOCIAÇÃO DE REDES ADVERSÁRIAS GENERATIVAS E INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL PARA ANÁLISE DE IMAGENS H&E VIA VISION TRANSFORMERS

Monografia apresentado(a) à Universidade Estadual Paulista (UNESP), Instituto de Biociências, Letras e Ciências Exatas, São José do Rio Preto, para obtenção do título de Bachareal em Ciência da Computação.
Data de defesa: 18/11/2025

BANCA EXAMINADORA

Profº Dr. Leandro Alves Neves
UNESP – Instituto de Biociências, Letras e Ciências Exatas – Campus de São José do Rio Preto

---

Profº Dr. Wallace Correa de Oliveira Casaca
UNESP – Instituto de Biociências, Letras e Ciências Exatas – Campus de São José do Rio Preto

---

Prof. Dr. Eng. Rodrigo Capobianco Guido
UNESP – Instituto de Biociências, Letras e Ciências Exatas – Campus de São José do Rio Preto

Dedico este trabalho aos meus pais, pelo amor incondicional, pelo apoio em todos os momentos
e por sempre acreditarem no meu potencial.

AGRADECIMENTOS

Gostaria de expressar minha mais profunda gratidão aos meus pais, cuja presença constante, apoio incondicional e palavras de encorajamento foram fundamentais em todos os momentos desta jornada acadêmica. Seu amor, paciência e dedicação silenciosa foram pilares que sustentaram minha caminhada mesmo nos períodos mais difíceis. Muito do que conquistei até aqui é reflexo direto dos valores que me ensinaram.

Agradeço também ao meu orientador, Prof. Dr. Leandro Alves Neves, por sua orientação, disponibilidade e rigor técnico durante todo o desenvolvimento deste trabalho. Seus conselhos, críticas construtivas e incentivo constante foram essenciais para minha evolução como pesquisador e como estudante. Aos colegas do laboratório, com quem compartilhei ideias, dificuldades e conquistas, deixo meu sincero agradecimento.

Estendo meus agradecimentos à Universidade Estadual Paulista (UNESP), por oferecer uma boa formação, infraestrutura e um ambiente propício ao desenvolvimento científico e pessoal. Sou também grato ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo apoio financeiro durante a Iniciação Científica e o desenvolvimento deste trabalho.

Aos meus amigos de faculdade, que tornaram essa trajetória mais leve e divertida, agradeço pela parceria, pelas conversas, pelo companheirismo nos estudos e pela amizade construída ao longo dos anos. Por fim, agradeço a todos que, direta ou indiretamente, contribuíram para a concretização deste projeto.

“An idiot admires complexity. A genius admires simplicity”
(Terry A. Davis)

RESUMO

Este trabalho propôs uma abordagem para aprimorar a geração e o aumento artificial de dados em imagens histopatológicas utilizando *Generative Adversarial Networks* (GANs) integradas a técnicas de *Explainable AI* (XAI). Foram utilizadas arquiteturas de GANs como *DCGAN*, *RaGAN* e *WGAN-GP*, combinadas com métodos de explicabilidade, como *Saliency*, *DeepLIFT* e *InputXGradient*, com o objetivo de incorporar informações interpretáveis no processo de geração de imagens sintéticas. A proposta foi avaliada em dois conjuntos de dados histológicos: um com amostras de câncer colorretal (CR) e outro com imagens de câncer de mama (UCSB). A qualidade das imagens geradas foi avaliada por meio das métricas FID (*Frechet Inception Distance*) e IS (*Inception Score*), enquanto o desempenho dos classificadores foi medido pela acurácia de modelos *Transformers*, incluindo ViT, DeiT e PVT. Os resultados demonstraram que o uso de XAI no treinamento das GANs pode reduzir o FID em até 26% e melhorar a acurácia dos classificadores em até 3,83% em comparação com modelos sem aumento de dados, e até 3,44% em relação ao uso de GANs sem XAI. A combinação *DCGAN + Saliency + PVT* obteve a maior acurácia (92,24%), seguida de *WGAN-GP + InputXGrad + PVT* (92,13%). Estes achados destacam o potencial do uso conjunto de XAI e GANs na geração de dados sintéticos mais eficazes, contribuindo para avanços na classificação de imagens médicas em cenários com limitação de dados.

Palavras-chave: Redes Adversariais Generativas, Inteligência Artificial Explicável, Vision Transformers, Aumento de Dados, Classificação Histopatológica.

ABSTRACT

This work proposes an approach to enhance data augmentation and synthetic image generation in histopathological datasets by integrating *Generative Adversarial Networks* (GANs) with *Explainable AI* (XAI) techniques. The methodology employs GAN architectures such as *DCGAN*, *RaGAN*, and *WGAN-GP*, combined with interpretability methods, such as *Saliency*, *DeepLIFT*, and *InputXGradient*, aiming to embed explanation-driven features into the generator training process. The evaluation was conducted on two medical image datasets: one containing colorectal cancer (CR) samples and another with breast cancer specimens (UCSB). The quality of the generated images was assessed using the FID (*Frechet Inception Distance*) and IS (*Inception Score*) metrics, while the classification performance was measured using accuracy scores from Transformer-based models (ViT, DeiT, and PVT), with and without data augmentation. Results showed that integrating XAI into GAN training reduced FID by up to 26% and improved classification accuracy by up to 3.83% compared to models trained without data augmentation, and up to 3.44% over augmentation without XAI. The best performance was achieved by the combination of *DCGAN* + *Saliency* + *PVT* (92.24%), followed by *WGAN-GP* + *InputXGrad* + *PVT* (92.13%). These findings highlight the potential of combining XAI and GANs to generate more effective and interpretable synthetic data, offering relevant contributions to medical image classification in data-limited scenarios.

Keywords: Generative Adversarial Networks, Explainable Artificial Intelligence, Vision Transformers, Data Augmentation, Histopathological Classification.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ACC	Acurácia
CNN	*Convolutional Neural Network*
CR	Câncer Colorretal
ROI	*Region of interest*
DCGAN	*Deep Convolutional Generative Adversarial Network*
DeiT	*Data-efficient Image Transformer*
DL	*Deep Learning*
FID	*Fréchet Inception Distance*
GAN	*Generative Adversarial Network*
IS	*Inception Score*
PVT	*Pyramid Vision Transformer*
RaGAN	*Relativistic Average GAN*
UCSB	*University of California, Santa Barbara (Breast Cancer Dataset)*
ViT	*Vision Transformer*
WGAN-GP	*Wasserstein GAN with Gradient Penalty*
XAI	*Explainable Artificial Intelligence*
Saliency	*Saliency Map*
DeepLIFT	*Deep Learning Important FeaTures*
InputXGrad	*Input multiplied by Gradient*
H&E	Hematoxilina e Eosina
TP	*True Positive*
FP	*False Positive*
TN	*True Negative*
FN	*False Negative*

SUMÁRIO

1 INTRODUÇÃO

O avanço das técnicas de aprendizado profundo (*Deep Learning* - DL) (LeCun; Bengio; Hinton, 2015), impulsionado pelo aumento da capacidade computacional e pela popularização de dispositivos de processamento paralelo, como GPUs e TPUs, tem revolucionado a análise de dados em diversas áreas, com destaque para a medicina. Diferentemente dos métodos tradicionais baseados em descritores manuais (*handcrafted features*), que exigem conhecimento prévio para extrair características relevantes, os modelos de DL aprendem representações hierárquicas diretamente a partir de dados brutos, como imagens digitais. Essa capacidade os torna particularmente eficazes em tarefas complexas de visão computacional, como a classificação e o reconhecimento de padrões em imagens histológicas obtidas por coloração Hematoxilina & Eosina (H&E), amplamente utilizadas no diagnóstico de doenças oncológicas (Dargan et al., 2020).

No entanto, o treinamento de modelos de *deep learning* depende criticamente de grandes volumes de dados rotulados e balanceados, uma condição raramente atendida em aplicações médicas. Imagens H&E, por exemplo, enfrentam desafios como a escassez de amostras, o desbalanceamento entre classes (devido à prevalência desigual de diferentes tipos de câncer) e a alta variabilidade morfológica, que reflete a heterogeneidade citológica presente em tumores (Shen; Wu; Suk, 2017). Esses fatores podem levar ao *overfitting*, comprometendo a capacidade dos modelos de generalizar para novos dados (Wang; She; Ward, 2021). Além disso, a complexidade das imagens H&E, onde um único tumor pode exibir múltiplas características morfológicas, dificulta a identificação precisa de padrões patológicos, exigindo abordagens robustas para superar essas limitações.

Uma solução promissora para esses desafios é o uso de Redes Geradoras Adversariais (*Generative Adversarial Networks* - GANs) (Goodfellow et al., 2014), que consistem em duas redes neurais: uma geradora (G), que sintetiza imagens a partir da distribuição de probabilidade dos dados reais, e uma discriminadora (D), que diferencia imagens reais das sintéticas. Por meio de um processo competitivo, as GANs produzem amostras visualmente realistas, permitindo o aumento artificial de conjuntos de dados (Wang; She; Ward, 2021). Estudos recentes demonstram o sucesso dessa abordagem em aplicações médicas, como a geração de imagens sintéticas de tomografias computadorizadas (TC) de tumores de mama, que aumentaram a acurácia de classificadores de 69,85% para 94% (S et al., 2020), e de imagens histopatológicas, com melhorias de até 6,7% em classificadores baseados em ResNet34 (Xue et al., 2021a). Apesar de seu potencial, as GANs enfrentam desafios, como instabilidade no treinamento, que podem ser mitigados por técnicas como regularização (Radford; Metz; Chintala, 2015), condicionamento de classes (Mirza; Osindero, 2014; Odena; Olah; Shlens, 2017) e integração com métodos de Inteligência Artificial Explicável (*Explainable Artificial Intelligence* - XAI).

Os métodos XAI surgem para aumentar a transparência e a interpretabilidade dos modelos

de *deep learning*, permitindo que humanos compreendam as decisões tomadas pelos algoritmos (Adadi; Berrada, 2018). Em aplicações com GANs, o XAI pode guiar a geração de imagens, focando em regiões morfologicamente relevantes, o que melhora a qualidade das amostras sintéticas. Testes preliminares indicam que a combinação de GANs com XAI pode aumentar em até 23,18% a qualidade de imagens geradas em baixa dimensão (Nagisetty et al., 2020). Paralelamente, os *Vision Transformers* (ViTs) (Dosovitskiy et al., 2021a) emergem como uma alternativa inovadora às redes convolucionais tradicionais (CNNs). Baseados em mecanismos de *self-attention* e *tokenização*, os ViTs modelam dependências espaciais de longo alcance, capturando relações globais em imagens de forma mais eficiente. Essa característica os torna particularmente adequados para tarefas complexas, como a análise de imagens H&E, onde padrões patológicos podem estar distribuídos de maneira não local (Khan et al., 2022).

Apesar do potencial individual dessas tecnologias, sua integração em um *framework* unificado para a classificação de imagens H&E permanece pouco explorada. A combinação de GANs para aumento de dados, XAI para interpretabilidade e ViTs para classificação oferece uma abordagem inovadora, com potencial para superar as limitações atuais e melhorar a precisão de sistemas de apoio ao diagnóstico. Este trabalho propõe desenvolver e avaliar um pipeline que integre essas técnicas, investigando seu impacto conjunto no desempenho de classificadores para imagens histológicas H&E representativas de diferentes tipos de câncer.

1.1 JUSTIFICATIVAS E MOTIVAÇÃO

O câncer representa uma das principais causas de mortalidade mundial, sendo o diagnóstico precoce fundamental para o sucesso do tratamento (Organização Mundial da Saúde, 2025). Imagens histológicas desempenham papel essencial nesse processo, permitindo a análise detalhada de tecidos, porém exigindo interpretação especializada e sujeita à subjetividade (Kang et al., 2022). Nesse contexto, métodos de aprendizado profundo têm se mostrado promissores para apoiar o diagnóstico médico, oferecendo maior padronização e reprodutibilidade nos resultados. Entretanto, a escassez de bases de dados rotuladas e a alta variabilidade morfológica das amostras ainda impõem desafios à generalização dos modelos (Komura; Ishikawa, 2018).

Nesse contexto, o uso de Redes Adversárias Generativas apresenta-se como uma alternativa promissora para mitigar a limitação de dados, possibilitando a criação de imagens sintéticas realistas que ampliam a diversidade e a representatividade dos conjuntos de treinamento (Frid-Adar et al., 2018). Além disso, a incorporação de técnicas de Inteligência Artificial Explicável pode ser empregada para orientar o processo de geração das GANs, tornando-o mais controlado e direcionado a características relevantes das imagens histológicas. Complementarmente, os Vision Transformers destacam-se pela capacidade de capturar dependências espaciais globais, superando as limitações locais das redes convolucionais tradicionais e oferecendo um potencial significativo para a análise de padrões complexos em tecidos histológicos (Dosovitskiy et al., 2021b). Apesar dos avanços indicados previamente, ainda se observa uma lacuna significativa

na literatura quanto à aplicação integrada de GANs, técnicas de XAI e ViTs em conjuntos de dados histológicos. Dessa forma, este trabalho se justifica pela proposta de investigar, de maneira sistemática, a combinação dessas técnicas na análise de imagens histológicas H&E, buscando contribuir com o desenvolvimento de sistemas mais eficientes para apoio ao diagnóstico médico.

1.2 OBJETIVOS

O objetivo geral deste trabalho é investigar o uso de técnicas de aumento artificial baseadas em GANs e GAN + XAI para enriquecer conjuntos de dados de imagens histológicas H&E e analisar o impacto dessas técnicas no desempenho de classificadores baseados em *Vision Transformers*. Os objetivos específicos são:

- Realizar o aumento artificial de imagens H&E utilizando modelos GAN tradicionais e variantes com XAI;
- Treinar classificadores ViT em regime de aprendizado auto-supervisionado sobre conjuntos de dados reais, sintéticos e combinados;
- Avaliar o desempenho dos modelos por meio de métricas específicas para cada tarefa;
- Identificar as melhores combinações de técnicas para a tarefa de classificação em imagens histológicas H&E.

1.3 ORGANIZAÇÃO DO PROJETO

No Capítulo 2 são apresentados os conceitos fundamentais necessários para o entendimento das abordagens utilizadas ao longo do estudo, incluindo descrições sobre Redes Geradoras Adversariais, técnicas de Inteligência Artificial Explicável, *Vision Transformers* e métricas empregadas na avaliação de desempenho. No Capítulo 3 são reunidos os principais trabalhos relacionados ao tema, que servem como base teórica e motivação para a proposta desenvolvida, sendo destacadas aplicações recentes das técnicas exploradas. No Capítulo 4 é descrita detalhadamente a metodologia proposta, abrangendo a organização das etapas de geração de dados sintéticos, o treinamento dos classificadores ViT e as estratégias de validação adotadas. No Capítulo 5 são apresentados os resultados obtidos a partir da aplicação da metodologia, sendo discutido o desempenho das diferentes combinações, bem como a análise quantitativa baseada em métricas e a comparação com trabalhos correlatos. Por fim, no Capítulo 6 são expostas as considerações finais do trabalho, nas quais são destacadas as contribuições alcançadas, as limitações observadas e as sugestões para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

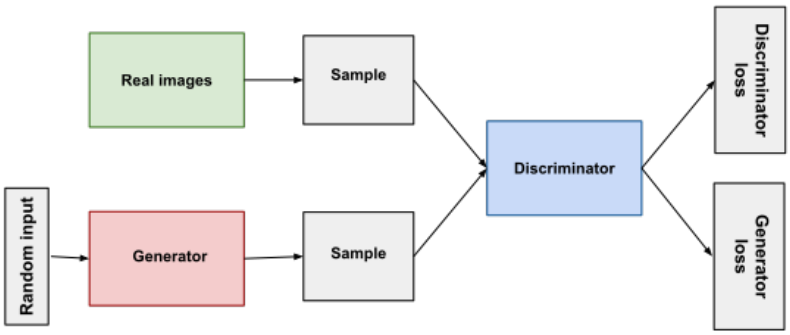
Neste capítulo são apresentados os conceitos fundamentais para o entendimento das abordagens utilizadas ao longo deste trabalho. São abordados alguns princípios básicos das Redes Geradoras Adversariais (GANs), incluindo variações arquiteturais relevantes; estratégias de explicabilidade em modelos de aprendizado profundo, por meio de técnicas de Inteligência Artificial Explicável; conceitos relacionados a *Vision Transformers* (ViTs); e, por fim, as principais métricas utilizadas na avaliação de modelos generativos e classificadores.

2.1 REDES GERADORAS ADVERSARIAIS (GANS)

As Redes Geradoras Adversariais, conhecidas como GANs (*Generative Adversarial Networks*), constituem uma classe de modelos de aprendizado profundo voltados à geração de dados sintéticos com alta similaridade estatística em relação a dados reais. As GANs introduziram um novo paradigma de aprendizado não supervisionado, baseado em um processo de competição entre duas redes neurais que aprendem de forma simultânea e adversarial: uma rede geradora (*generator*) e uma rede discriminadora (*discriminator*).

A rede geradora é responsável por mapear um vetor de entrada amostrado de uma distribuição latente (geralmente uma distribuição normal ou uniforme) em uma amostra sintética no domínio dos dados (por exemplo, uma imagem). Seu objetivo é produzir instâncias que se assemelhem o máximo possível às amostras reais, de modo que a discriminadora não consiga distinguir entre ambas. Por outro lado, a rede discriminadora atua como um classificador binário, tentando determinar se uma entrada provém do conjunto de dados real ou foi gerada artificialmente pela geradora. Na Figura 1 são ilustradas as estruturas básicas de uma GAN.

Figura 1 – Arquitetura básica de uma GAN.



Fonte: Adaptado de (Google Developers, 2023).

O treinamento das GANs é formulado como um jogo de soma zero, no qual ambas as redes são otimizadas simultaneamente com objetivos opostos: enquanto a geradora busca maximizar a

probabilidade de enganar a discriminadora, esta tenta minimizar o número de falsos positivos. Essa interação é formalizada pela seguinte função de valor min-max:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))], \tag{1}$$

em que x representa uma amostra real, z é um vetor latente amostrado de uma distribuição p_z, $G(z)$ é a amostra gerada, $D(x)$ é a probabilidade atribuída pela discriminadora de que x seja real, e $D(G(z))$ é a probabilidade de que a amostra gerada seja classificada como autêntica.

O equilíbrio ideal é alcançado quando a geradora produz amostras tão realistas que a discriminadora não consegue mais diferenciá-las, atribuindo probabilidade de 0,5 para ambas as classes. Entretanto, alcançar esse ponto de equilíbrio é um dos maiores desafios no treinamento de GANs, que é conhecido por ser instável. Problemas como colapso de modo (*mode collapse*), oscilações no gradiente, sobreajuste e sensibilidade à inicialização dos pesos são comuns (Goodfellow et al., 2014), tornando o processo de otimização delicado e fortemente dependente de ajustes de hiperparâmetros e funções de perda. Ao longo dos anos, diversas variantes de GANs foram propostas para mitigar essas limitações e melhorar a estabilidade do treinamento, a diversidade das amostras geradas e a qualidade visual dos resultados.

2.1.1 Funções de Ativação e Perda

A eficácia do treinamento de GANs está intimamente relacionada à escolha das funções de ativação e das funções de perda, pois ambas influenciam diretamente a propagação dos gradientes e a estabilidade da convergência. As funções de ativação introduzem não-linearidades que permitem às redes aprender representações complexas. Na geradora, é comum o uso de ReLU (Agarap, 2019) ou variantes em camadas intermediárias, combinadas com a função *tanh* (Dubey; Singh; Chaudhuri, 2022) na camada de saída, de modo a limitar o intervalo dos valores gerados (por exemplo, para o intervalo $[-1, 1]$). Na discriminadora, funções como *LeakyReLU* (Maas, 2013) são amplamente utilizadas para evitar a saturação dos gradientes e melhorar a capacidade de distinguir sutis diferenças entre amostras reais e sintéticas.

As funções de perda, por sua vez, definem a dinâmica de competição entre as redes. No modelo original de GAN, a função objetivo é formulada como um jogo min-max, conforme a Equação 1. Embora a estrutura conceitual das GANs permaneça consistente, a escolha cuidadosa dessas funções e de suas respectivas parametrizações tem impacto direto na qualidade e diversidade das amostras geradas. As próximas seções descrevem as arquiteturas utilizadas neste trabalho, destacando as particularidades adotadas para garantir estabilidade e qualidade na geração de imagens sintéticas.

2.1.2 Especificação da DCGAN

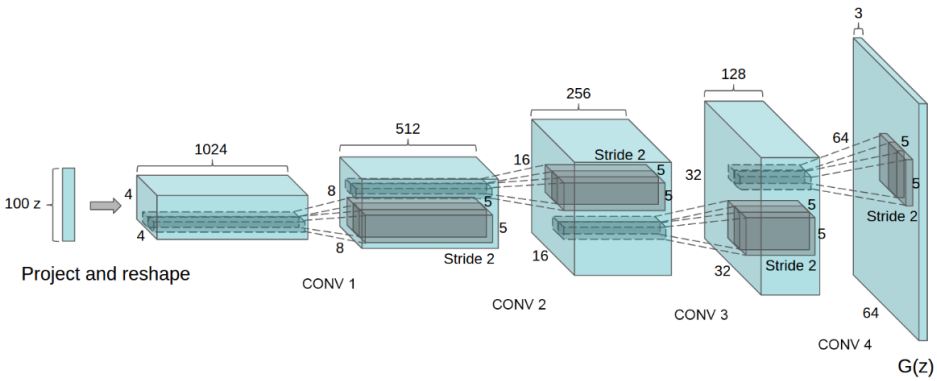
A *Deep Convolutional Generative Adversarial Network* (DCGAN) (Radford; Metz; Chintala, 2016) é uma das variações mais influentes e amplamente utilizadas das Redes Geradoras Adver-

sariais (GANs). A DCGAN teve como principal contribuição o estabelecimento de diretrizes arquiteturais que permitiram a aplicação bem-sucedida de redes convolucionais profundas no treinamento adversarial, resultando em maior estabilidade de convergência e imagens sintéticas de melhor qualidade. Ao contrário das primeiras implementações de GANs, que utilizavam camadas totalmente conectadas e eram propensas a instabilidade e colapso de modo, a DCGAN explora o poder das convoluções para capturar estruturas espaciais e padrões locais de forma mais eficiente. Essa reformulação tornou as GANs mais adequadas para tarefas de geração de imagens e estabeleceu uma base para diversas variantes posteriores.

Na arquitetura da DCGAN, a rede geradora é composta por uma sequência de camadas de convolução transposta (também conhecidas como *deconvolution* ou *Conv2D Transpose*), responsáveis por expandir progressivamente o vetor latente de entrada $z \in \mathbb{R}^{100}$ até atingir a resolução desejada da imagem de saída. Cada camada é seguida por normalização em lote (*Batch Normalization*) e função de ativação ReLU, promovendo estabilidade no fluxo de gradientes e acelerando a convergência. Na camada de saída, a função *tanh* é utilizada para restringir os valores dos pixels ao intervalo $[-1, 1]$, compatibilizando-os com a normalização dos dados reais.

O discriminador, por sua vez, é projetado de forma simétrica e inversa à geradora. Ele recebe como entrada uma imagem (real ou sintética) e aplica uma série de camadas convolucionais que reduzem gradualmente sua dimensionalidade, extraindo características discriminativas de alto nível. Cada camada é seguida por uma função de ativação *LeakyReLU*, que introduz uma pequena inclinação para valores negativos, melhorando a robustez do modelo. Diferentemente da geradora, a normalização em lote não é aplicada na primeira camada do discriminador, a fim de prevenir o vazamento de estatísticas entre amostras reais e falsas, o que poderia comprometer o aprendizado adversarial. Na Figura 2 é apresentada uma visão geral da arquitetura da DCGAN.

Figura 2 – Arquitetura da DCGAN.



Fonte: Adaptado de (Radford; Metz; Chintala, 2016).

Além de sua simplicidade estrutural, a DCGAN definiu um conjunto de boas práticas que se tornaram referência para o desenvolvimento de modelos generativos posteriores. Entre essas diretrizes estão o uso de *stride* em vez de *pooling* para redimensionamento, a remoção de

camadas totalmente conectadas, a aplicação consistente de normalização em lote e a escolha criteriosa das funções de ativação. Essas decisões arquiteturais contribuíram significativamente para reduzir a instabilidade típica das GANs e permitiram que modelos treinados em conjuntos de dados relativamente simples produzissem imagens realistas e coerentes. Devido à sua robustez e eficiência, a DCGAN é amplamente utilizada como modelo base em aplicações de geração de imagens sintéticas, reconstrução de dados e aprendizado não supervisionado de representações visuais.

2.1.3 Especificação da RaGAN

A *Relativistic Average GAN* (RaGAN) (Jolicoeur-Martineau, 2018) é uma variação das Redes Geradoras Adversariais que busca melhorar a estabilidade do treinamento e a qualidade perceptiva das imagens geradas por meio de uma modificação na função de perda da rede discriminadora. Diferentemente das GANs convencionais, nas quais o discriminador aprende a distinguir amostras reais de falsas de maneira absoluta, a RaGAN introduz uma perspectiva relativística: em vez de avaliar se uma amostra é real ou falsa isoladamente, o modelo aprende a estimar a probabilidade de uma imagem real parecer mais realista do que uma imagem gerada, e vice-versa.

Essa abordagem baseia-se na observação de que o treinamento adversarial deve refletir uma competição relativa entre as distribuições de dados reais e gerados, e não uma simples classificação binária. Assim, o discriminador da RaGAN atua como um comparador entre amostras, avaliando o grau de realismo relativo. Essa mudança conceitual permite que o gerador receba gradientes mais informativos, favorecendo um aprendizado mais estável e com menor propensão ao colapso de modo. A função de perda do discriminador é reformulada para capturar essa diferença relativa entre as previsões de realismo, sendo expressa como:

$$\mathcal{L}_D = -\mathbb{E}_{x_r}\left[\log(\sigma(C(x_r) - \mathbb{E}_{x_f}[C(x_f)]))\right] - \mathbb{E}_{x_f}\left[\log(1 - \sigma(C(x_f) - \mathbb{E}_{x_r}[C(x_r)]))\right], \quad (2)$$

em que x_r representa uma amostra real, x_f uma amostra gerada, $C(x)$ é a saída da rede discriminadora antes da aplicação da função sigmoide, e σ denota a função logística. As expectativas são calculadas sobre os respectivos conjuntos de amostras reais e falsas. Já a função de perda do gerador é obtida invertendo-se os sinais das diferenças, o que o incentiva a produzir amostras cuja pontuação média de realismo supere a das imagens reais. Essa reformulação faz com que o discriminador opere como um avaliador comparativo em vez de um classificador absoluto. Tal modificação reduz a tendência de saturação das funções de perda e proporciona gradientes mais consistentes ao gerador, mesmo em estágios avançados do treinamento. Como consequência, a RaGAN apresenta melhor estabilidade, convergência mais suave e imagens geradas com maior fidelidade visual, mantendo compatibilidade estrutural com arquiteturas clássicas, como a DCGAN.

2.1.4 Especificação da WGAN-GP

A *Wasserstein Generative Adversarial Network with Gradient Penalty* (WGAN-GP) (Gulrajani et al., 2017) é uma variação das Redes Geradoras Adversariais desenvolvida com o objetivo de aprimorar a estabilidade do treinamento e oferecer uma métrica mais consistente para medir a distância entre distribuições de probabilidade. Essa arquitetura é uma extensão direta da WGAN original (Arjovsky; Chintala; Bottou, 2017), que substitui a função de perda baseada na divergência de Jensen-Shannon, utilizada nas GANs clássicas, por uma função fundamentada na distância de Wasserstein (Panaretos; Zemel, 2019). Essa substituição busca resolver problemas recorrentes nas GANs tradicionais, como saturação da função de perda, gradientes pouco informativos e colapso de modos. A distância de Wasserstein é uma métrica contínua e diferenciável que expressa o "custo mínimo" de transformar uma distribuição em outra. Diferentemente das métricas baseadas em divergências, ela fornece uma medida mais estável e com gradientes mais suaves, mesmo quando as distribuições reais e geradas possuem pouco ou nenhum suporte em comum. No contexto da WGAN, essa propriedade resulta em um sinal de aprendizado mais informativo e correlacionado com a qualidade das amostras produzidas.

Para garantir que a distância de Wasserstein seja bem definida, é necessário impor que a função discriminadora, denominada *critic* nessa abordagem, seja 1-Lipschitz contínua, ou seja, que sua derivada em relação às entradas seja limitada por 1. Na versão original da WGAN, essa restrição era implementada por meio de *weight clipping*, limitando os valores dos pesos do critic a um intervalo fixo. Contudo, esse método frequentemente gerava redes mal condicionadas, restringindo a capacidade de representação do critic e comprometendo a estabilidade do treinamento. A WGAN-GP propõe uma alternativa mais eficaz: a introdução de uma penalidade no gradiente (*gradient penalty*) aplicada sobre amostras interpoladas entre dados reais e gerados. Essa penalidade encoraja a norma do gradiente da saída do critic em relação às amostras a permanecer próxima de 1, garantindo a condição de Lipschitz de forma mais estável. A função de perda do critic na WGAN-GP é dada por:

$$\mathcal{L}_D = \mathbb{E}_{\tilde{x}\sim\mathbb{P}_g}[D(\tilde{x})] - \mathbb{E}_{x\sim\mathbb{P}_r}[D(x)] + \lambda\,\mathbb{E}_{\hat{x}\sim\mathbb{P}_{\hat{x}}}\left[(\|\nabla_{\hat{x}}D(\hat{x})\|_2 - 1)^2\right], \tag{3}$$

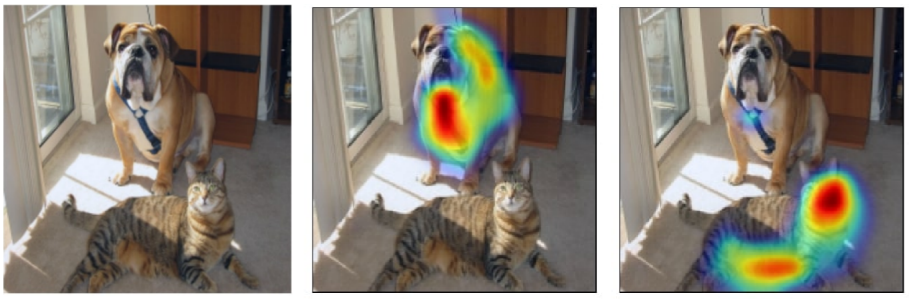
em que x representa amostras reais, $\tilde{x}$ amostras geradas e $\hat{x}$ são amostras interpoladas entre ambas. O termo λ é um hiperparâmetro que controla o peso da penalidade sobre o gradiente. Essa modificação garante que o critic mantenha derivadas próximas de 1 ao longo das amostras interpoladas, atendendo de forma robusta à condição de Lipschitz. A introdução da penalidade no gradiente tornou o treinamento das WGANs substancialmente mais estável, eliminando a necessidade de ajustes manuais de intervalos de *clipping* e proporcionando convergência mais suave. Além disso, a função de perda da WGAN-GP correlaciona-se melhor com a qualidade perceptiva das amostras, oferecendo um critério quantitativo mais confiável para avaliar o progresso do modelo durante o treinamento.

2.2 INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL (XAI)

Com o avanço dos modelos de aprendizado profundo nos últimos anos, tornou-se evidente a necessidade de mecanismos que possibilitem compreender e interpretar o processo decisório dessas arquiteturas. A área de *Inteligência Artificial Explicável* (XAI, do inglês *Explainable Artificial Intelligence*) surgiu como resposta à crescente demanda por transparência, interpretabilidade e confiabilidade em sistemas baseados em inteligência artificial, especialmente em domínios críticos como saúde, finanças, direito e segurança. Modelos de aprendizado profundo, como redes neurais convolucionais (CNNs), transformers e suas variantes, são frequentemente tratados como "caixas-pretas", devido à complexidade de suas estruturas internas e à dificuldade de compreender diretamente como uma determinada predição é realizada. Embora esses modelos alcancem altos níveis de desempenho, sua adoção em ambientes sensíveis exige a capacidade de justificar decisões, identificar falhas e mitigar riscos associados a vieses e inconsistências nos dados.

A XAI busca mitigar esses problemas por meio do desenvolvimento de métodos que ofereçam representações interpretáveis do comportamento do modelo. Tais métodos podem ser categorizados de várias formas, sendo uma das mais comuns a distinção entre abordagens pós-hoc e abordagens intrinsecamente interpretáveis. Métodos pós-hoc são aplicados após o treinamento do modelo, com o objetivo de explicar uma decisão já tomada. Já os modelos intrinsicamente interpretáveis são projetados desde o início com estruturas que facilitam a compreensão humana, como árvores de decisão ou modelos lineares. Na Figura 3 é apresento um exemplo visual da aplicação de técnicas XAI em imagens.

Figura 3 – Exemplo de uma explicação gerada por técnica XAI em classificação.



(a) Original Image (b) Cat Counterfactual exp (c) Dog Counterfactual exp

Fonte: Adaptado de (Selvaraju et al., 2019).

Além da explicabilidade individual, os métodos XAI também podem ser utilizados para auditoria de modelos, detecção de vieses, análise de robustez e orientação de novos conjuntos de treinamento. Em aplicações médicas, por exemplo, permitem validar se a rede está baseando suas decisões em estruturas anatômicas relevantes, e não em ruídos ou padrões aleatórios presentes nos dados. Outro ponto importante é que a explicabilidade contribui diretamente para a confiança e aceitação dos sistemas de IA por usuários humanos. Em ambientes clínicos, por exemplo, a

capacidade de justificar uma predição pode auxiliar profissionais na validação de resultados, melhorar o processo de tomada de decisão e ampliar a adoção dessas ferramentas em fluxos de trabalho reais.

2.2.1 Especificação do *Saliency*

Os mapas de saliência (*Saliency Maps*) (Simonyan; Vedaldi; Zisserman, 2014) constituem uma das abordagens mais fundamentais e amplamente empregadas em XAI, especialmente em tarefas de visão computacional baseadas em redes neurais convolucionais. O objetivo principal dessa técnica é identificar, para uma dada predição, quais regiões da imagem de entrada exerceram maior influência na decisão do modelo. O princípio de funcionamento baseia-se na análise do gradiente da saída da rede em relação à imagem de entrada. Em termos intuitivos, busca-se estimar como pequenas variações nos valores dos pixels impactariam a confiança atribuída à classe predita. Assim, os pixels cuja alteração provoca maior variação na saída do modelo são considerados mais relevantes, sendo evidenciados no mapa de saliência. Matematicamente, dado um modelo f e uma imagem de entrada I, o mapa de saliência S correspondente à classe de interesse c é definido como:

$$S = \left| \frac{\partial f_c(I)}{\partial I} \right|, \tag{4}$$

em que $f_c(I)$ representa a saída do modelo (tipicamente o logit ou a probabilidade) associada à classe c. O valor absoluto do gradiente indica a sensibilidade da saída em relação a cada pixel, refletindo o grau de importância local na decisão. Após o cálculo, o mapa é geralmente normalizado e sobreposto à imagem original sob a forma de um mapa de calor (*heatmap*), permitindo visualizar de maneira intuitiva as regiões de maior relevância para a predição. Essa representação fornece uma explicação visual direta do raciocínio interno da rede, sendo particularmente útil em contextos médicos, nos quais é fundamental verificar se o modelo está se baseando em estruturas anatômicas pertinentes.

2.2.2 Especificação do *DeepLIFT*

O *Deep Learning Important FeaTures* (DeepLIFT) (Shrikumar; Greenside; Kundaje, 2019) é uma técnica de *Explainable Artificial Intelligence* (XAI) desenvolvida para atribuir relevâncias às entradas de uma rede neural de acordo com sua contribuição para uma predição específica. Diferentemente de métodos baseados puramente em gradientes, o *DeepLIFT* introduz o conceito de comparação entre ativações, isto é, analisa como a ativação de cada neurônio em uma entrada real difere daquela obtida para uma entrada de referência (ou *baseline*). Essa diferença é então propagada ao longo da rede para estimar a importância relativa de cada característica de entrada.

A utilização de uma entrada de referência é um dos principais diferenciais do *DeepLIFT*. Ela pode ser definida como uma imagem neutra (por exemplo, completamente preta), a média do conjunto de dados ou outra entrada que represente ausência de informação relevante. Essa

abordagem permite capturar contribuições positivas e negativas de forma mais estável, evitando problemas comuns em métodos puramente diferenciais, como gradientes nulos em regiões saturadas da função de ativação. O princípio fundamental do método é decompor a variação da saída do modelo (Δy), ou seja, a diferença entre a predição para a entrada real e a predição para a entrada de referência, em termos das variações provocadas pelas entradas (Δx_i). O *DeepLIFT* assegura que essa decomposição obedeça a uma propriedade de conservação, expressa matematicamente como:

$$\sum_i C_{\Delta x_i} = \Delta y, \tag{5}$$

em que $C_{\Delta x_i}$ representa a contribuição atribuída à entrada x_i para a diferença total observada na saída. A propagação dessas contribuições é realizada de forma hierárquica, seguindo regras específicas para cada tipo de operação da rede (como ativações não lineares, somas e multiplicações). Essa estrutura de propagação torna o *DeepLIFT* capaz de gerar explicações mais consistentes e interpretáveis, especialmente em redes profundas, onde métodos baseados apenas em gradientes tendem a se tornar instáveis. Em aplicações médicas, por exemplo, o *DeepLIFT* permite identificar com precisão quais regiões da imagem mais contribuíram para uma classificação, mesmo em casos nos quais os gradientes locais são pouco informativos.

2.2.3 Especificação do *Input x Grad*

A técnica *Input* × *Gradient* (Hechtlinger, 2016) é uma abordagem simples, porém eficaz, para atribuição de relevância em modelos de aprendizado profundo. Sua proposta consiste em multiplicar, elemento a elemento, os valores da entrada do modelo pelos gradientes da saída em relação a essa mesma entrada, de modo a estimar a importância individual de cada característica na predição realizada. A intuição por trás do método é que o gradiente, isoladamente, expressa apenas a sensibilidade local da saída, ou seja, o quanto pequenas variações na entrada afetariam a predição, mas não considera a magnitude dos próprios valores de entrada. Ao multiplicar o gradiente pelos valores originais de entrada, o método pondera a sensibilidade pela intensidade da ativação, combinando assim informação sobre quanto a saída muda e onde a entrada é mais relevante. Essa formulação é expressa matematicamente como:

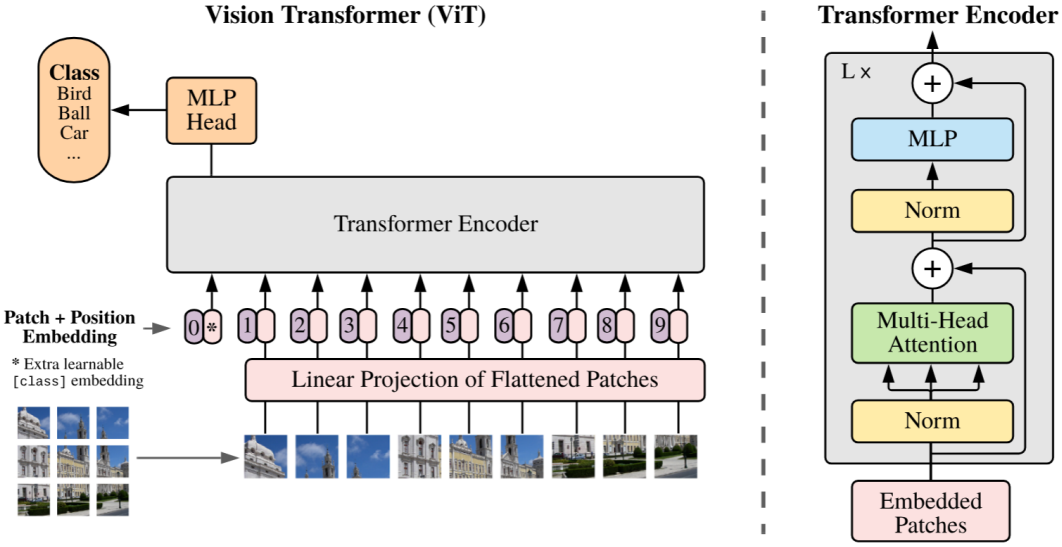
$$\text{Relevância}(x_i) = x_i \cdot \frac{\partial f(x)}{\partial x_i}, \tag{6}$$

em que x_i representa o valor do i-ésimo pixel (ou característica) da entrada x, e $\frac{\partial f(x)}{\partial x_i}$ é o gradiente da saída $f(x)$ em relação a x_i. O produto resultante fornece uma estimativa direta da contribuição de cada componente de entrada para a predição do modelo. Essa técnica apresenta baixo custo computacional e pode ser facilmente aplicada a qualquer rede diferenciável, o que a torna uma alternativa prática em cenários onde se busca interpretabilidade com eficiência.

2.3 VISION TRANSFORMERS (VITS)

Vision Transformers (ViTs) são uma classe de modelos de aprendizado profundo que adaptam a arquitetura de *Transformers*, originalmente desenvolvida para tarefas de processamento de linguagem natural, para o domínio da visão computacional. Essa adaptação representa uma mudança de paradigma em relação às tradicionais redes neurais convolucionais, ao substituir as operações de convolução por mecanismos de atenção que modelam relações globais entre as partes de uma imagem. A arquitetura básica dos ViTs consiste em dividir a imagem de entrada em pequenos blocos de tamanho fixo, chamados de *patches*, que são posteriormente achatados e projetados para um espaço vetorial por meio de uma camada linear. Cada *patch* projetado funciona como um token de entrada para o *Transformer Encoder*, de forma análoga às palavras em tarefas de linguagem. Além disso, um vetor de posição é adicionado a cada token, para preservar a informação espacial da imagem.

Figura 4 – Esquema simplificado de um ViT.



Fonte: (Dosovitskiy et al., 2021a).

A sequência de tokens é então processada por múltiplas camadas do *Transformer Encoder*, compostas por mecanismos de *Multi-Head Self-Attention* e camadas de *Feed-Forward*, ambas seguidas de normalização e conexões residuais. O mecanismo de *self-attention* permite que o modelo aprenda, em cada camada, a importância relativa de cada *patch* em relação aos demais, capturando padrões de longo alcance que podem ser difíceis de detectar com operações locais de convolução. Outra característica marcante dos ViTs é a presença do *class token*, um vetor especial concatenado à sequência de entrada, cuja representação final, ao fim do encoder, é usada como vetor de saída para tarefas de classificação. Essa abordagem dispensa o uso de camadas de agregação espacial, como *pooling*, comuns em CNNs. Entre as principais vantagens dos ViTs está a capacidade de capturar dependências globais desde as primeiras camadas, maior

flexibilidade para *transfer learning* e escalabilidade com grandes volumes de dados. Além disso, o uso de uma arquitetura uniforme baseada em atenção facilita a adaptação do modelo para diferentes tarefas além da classificação, como segmentação ou detecção, sem a necessidade de projetar blocos convolucionais especializados.

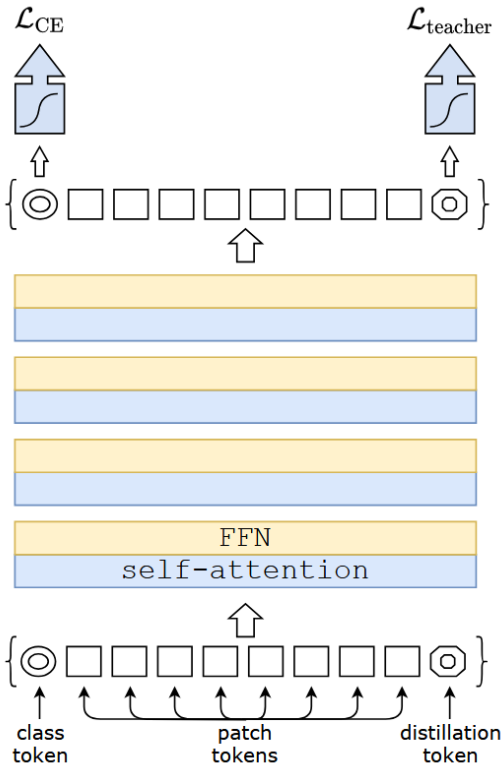
Por outro lado, os ViTs também apresentam desafios. Em comparação com CNNs, exigem maior quantidade de dados para treinamento eficaz, já que não possuem o viés indutivo espacial forte das convoluções. Esse viés ajuda as CNNs a generalizar melhor em cenários de poucos dados, enquanto os ViTs dependem fortemente de pré-treinamento em bases massivas (como o ImageNet-21k) para alcançar bom desempenho. Além disso, a complexidade quadrática do mecanismo de autoatenção em relação ao número de *patches* aumenta o custo computacional, tornando o treinamento de imagens de alta resolução caro em termos de memória e tempo de processamento.

2.3.1 Especificação do DeiT

O *Data-efficient Image Transformer* (DeiT) (Touvron et al., 2021) é uma variante dos ViTs com o objetivo de reduzir a dependência de grandes volumes de dados para o treinamento eficaz de modelos baseados em *Transformers* no domínio da visão computacional. Enquanto os ViTs originais demonstraram desempenho competitivo com redes convolucionais apenas quando treinados em grandes conjuntos de dados (como o ImageNet-21k), o DeiT foi projetado para alcançar bons resultados mesmo em conjuntos menores e com menor custo computacional, democratizando o uso de *Transformers* para cenários em que não há disponibilidade de dados em larga escala. Uma das principais contribuições do DeiT é a introdução de um mecanismo de treinamento supervisionado baseado em *knowledge distillation*. Essa técnica consiste no uso de um modelo professor, geralmente uma rede convolucional previamente treinada, como uma ResNet, para guiar o treinamento do *transformer* aluno (DeiT). O modelo aluno aprende não apenas a partir dos rótulos reais, mas também da saída suavizada do professor, que contém informações sobre as relações entre as classes, fornecendo um sinal de aprendizado mais rico e estável.

Além do *token* de classificação tradicional usado nos ViTs, o DeiT inclui um novo *token* chamado *distillation token*, que interage com os demais tokens durante o processo de autoatenção. Esse *token* é responsável por capturar o conhecimento transferido do modelo professor. Ao final do processo de codificação, o modelo produz duas saídas: uma baseada no *token* de classificação e outra no *token* de distilação. Ambas podem ser utilizadas para tomada de decisão, dependendo do modo de inferência adotado. Essa abordagem garante que o modelo não dependa exclusivamente da supervisão direta dos rótulos, mas também incorpore o viés indutivo do professor, que já sintetiza padrões relevantes aprendidos em conjuntos de dados maiores.

Figura 5 – Arquitetura geral do DeiT.



Fonte: (Touvron et al., 2021).

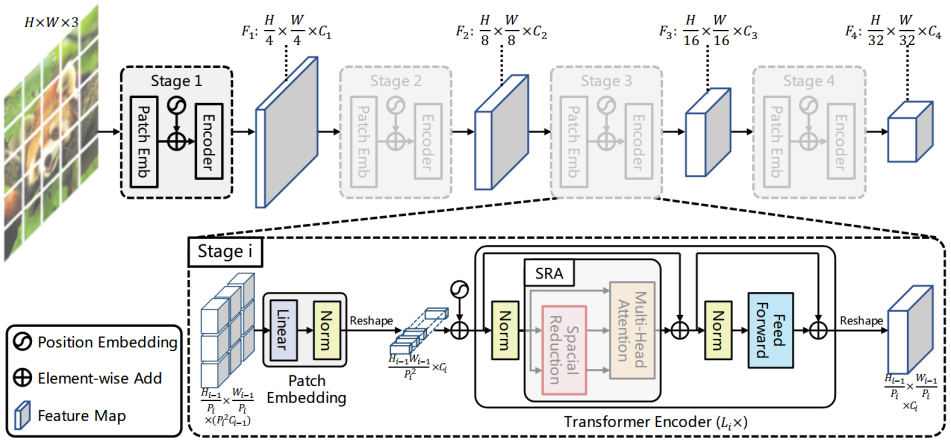
2.3.2 Especificação do PVT

O *Pyramid Vision Transformer* (PVT) (Wang et al., 2021) é uma variação dos *Vision Transformers* projetada para lidar com tarefas de visão computacional densa, como segmentação semântica e detecção de objetos, onde a preservação de hierarquias espaciais e a eficiência computacional são fundamentais. O PVT introduz uma estrutura piramidal semelhante às redes convolucionais tradicionais, permitindo capturar informações em múltiplas escalas ao longo da arquitetura. Diferentemente dos ViTs clássicos, que mantêm uma única resolução ao longo de toda a rede, o PVT adota uma arquitetura hierárquica dividida em estágios, com resoluções decrescentes progressivamente. Cada estágio processa uma versão mais compacta da imagem de entrada, semelhante ao funcionamento das CNNs com *pooling*. Isso permite uma representação mais eficiente de informações espaciais e reduz significativamente o custo computacional, o que é essencial em aplicações práticas, especialmente em dispositivos com recursos limitados.

Além da estrutura piramidal, o PVT incorpora o mecanismo de *Spatial-Reduction Attention* (SRA), uma modificação do mecanismo de autoatenção padrão. Enquanto o *self-attention* tradicional opera com complexidade quadrática em relação ao número de patches de entrada, o SRA reduz essa complexidade ao aplicar uma operação de redução espacial antes de calcular as atenções, tornando o modelo escalável para imagens de alta resolução. Essa adaptação é especialmente relevante em tarefas como detecção de objetos em imagens grandes, nas quais o

custo de memória do ViT tradicional se tornaria proibitivo.

Figura 6 – Arquitetura geral do PVT.



Fonte: (Wang et al., 2021).

2.4 MÉTRICAS DE AVALIAÇÃO

A definição de métricas apropriadas para avaliação de desempenho é um aspecto essencial no desenvolvimento de sistemas baseados em aprendizado de máquina, especialmente quando se lidam com tarefas distintas, como a geração e a classificação de imagens. A escolha adequada dessas métricas permite quantificar com precisão a qualidade dos resultados, orientar ajustes metodológicos e conferir maior rigor às análises experimentais. Entre as métricas mais amplamente utilizadas nesse contexto, destacam-se a acurácia, o *Inception Score* (IS) (Salimans et al., 2016) e o *Fréchet Inception Distance* (FID) (Heusel et al., 2018). A acurácia é tradicionalmente empregada em problemas de classificação supervisionada. Essa métrica mede a proporção de predições corretas em relação ao total de amostras avaliadas, fornecendo uma indicação direta da performance global do modelo. Por sua simplicidade e fácil interpretação, é frequentemente utilizada em cenários com classes balanceadas. Sua formulação matemática é expressa por:

$$\text{Acurácia} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{7}$$

em que TP (verdadeiros positivos) e TN (verdadeiros negativos) correspondem às classificações corretas, enquanto FP (falsos positivos) e FN (falsos negativos) representam os erros cometidos pelo modelo. Embora eficaz em muitos casos, a acurácia pode ser insuficiente para conjuntos de dados desbalanceados, exigindo, nesse caso, o uso de métricas complementares. Em contrapartida, para a avaliação de modelos geradores de imagens, são necessárias métricas específicas que levem em conta aspectos visuais e estatísticos. O *Inception Score* (IS) é uma das métricas mais utilizadas nesse cenário e foi concebido para avaliar dois critérios essenciais: a qualidade individual das imagens e a diversidade do conjunto gerado. Ele utiliza uma rede *Inception-v3* pré-treinada para calcular a distribuição de probabilidade $p(y|x)$ associada à classi-

ficação da imagem x. Espera-se que uma boa imagem apresente alta confiança em uma única classe (baixa entropia condicional) e que o conjunto apresente alta diversidade de classes (alta entropia marginal). O IS é calculado pela fórmula:

$$\text{IS} = \exp\left(\mathbb{E}_x\left[D_{\text{KL}}\left(p(y|x) \parallel p(y)\right)\right]\right), \tag{8}$$

em que D_{KL} representa a divergência de Kullback-Leibler entre a distribuição condicional $p(y|x)$ e a marginal $p(y)$. Valores mais altos de IS indicam imagens de boa qualidade, fáceis de classificar, e um conjunto globalmente diverso. Apesar de não considerar diretamente imagens reais como referência, o IS continua sendo amplamente utilizado devido à sua simplicidade e sensibilidade a propriedades relevantes na geração de imagens. Complementando essa análise, o *Fréchet Inception Distance* (FID) oferece uma abordagem mais robusta para comparar diretamente as distribuições de características de imagens reais e geradas. Assim como o IS, o FID utiliza uma rede Inception-v3 para extrair representações semânticas das imagens, mas, diferentemente dele, modela essas representações como distribuições gaussianas multivariadas. A distância entre essas distribuições é então medida pela fórmula:

$$\text{FID}(x, g) = \|\mu_x - \mu_g\|^2 + \text{Tr}\left(\Sigma_x + \Sigma_g - 2(\Sigma_x\Sigma_g)^{1/2}\right), \tag{9}$$

em que μ_x e Σ_x são a média e a matriz de covariância extraídas das imagens reais, e μ_g e Σ_g são as estatísticas correspondentes para as imagens geradas. A norma euclidiana entre as médias representa a diferença central entre os conjuntos, enquanto o termo envolvendo as matrizes de covariância captura a divergência em suas dispersões. Uma das principais vantagens do FID é sua sensibilidade tanto à fidelidade visual quanto à variabilidade estrutural das amostras. Ele penaliza modelos que geram imagens repetitivas ou pouco diversificadas e é capaz de refletir pequenas diferenças semânticas nos dados. Além disso, sua natureza contínua permite comparações graduais entre modelos. Valores mais baixos de FID indicam maior similaridade entre o conjunto gerado e o conjunto real, sendo, portanto, desejáveis. Em suma, as três métricas desempenham papéis complementares na análise de modelos de aprendizado profundo. Enquanto a acurácia permite avaliar o desempenho de classificadores em termos objetivos e diretos, o IS e o FID fornecem ferramentas robustas para examinar a qualidade e a diversidade das imagens produzidas por modelos generativos, viabilizando uma avaliação mais completa e confiável nesse contexto.

3 REVISÃO BIBLIOGRÁFICA

Neste capítulo, são apresentados trabalhos relacionados ao aumento artificial e à classificação de imagens histológicas por meio de técnicas de aprendizado profundo. A revisão foca principalmente em abordagens que exploram o uso de Redes Geradoras Adversariais para geração de imagens sintéticas e o emprego de modelos de classificação, como Redes Neurais Convolucionais e *Vision Transformers*, em tarefas de análise de imagens médicas. A análise dessas contribuições visa fornecer embasamento para a metodologia proposta neste trabalho, destacando os principais avanços na área, além de evidenciar limitações e oportunidades de pesquisa ainda pouco exploradas.

3.1 AUMENTO ARTIFICIAL VIA GANS E CLASSIFICADOR CNN

A aplicação de Redes Geradoras Adversariais como estratégia de aumento artificial tem se consolidado como uma alternativa eficaz para mitigar limitações inerentes a conjuntos de dados médicos, especialmente em contextos histopatológicos. Além disso, entre os classificadores mais utilizados nesse cenário, destacam-se as Redes Neurais Convolucionais, que se mostraram altamente eficazes na extração de padrões locais em imagens médicas. A combinação entre GANs e CNNs tem sido amplamente explorada na literatura, com estudos demonstrando que o uso de imagens geradas pode melhorar significativamente o desempenho de modelos classificadores.

Xue et al. (Xue et al., 2021b) propuseram o *HistoGAN*, um modelo de GAN condicional especialmente desenhado para gerar *patches* de imagens histopatológicas rotuladas com alta fidelidade visual e aderência morfológica às classes reais. O diferencial dessa abordagem não está apenas na geração sintética, mas também em um mecanismo de *augmentação seletiva*, que seleciona apenas as imagens geradas com alta confiança de rótulo e grande similaridade com amostras reais. Essa estratégia garante que apenas imagens sintéticas de qualidade sejam incorporadas ao conjunto de treinamento, o que mitiga os riscos associados à inclusão de amostras irrelevantes ou de baixa qualidade. Os experimentos foram realizados em dois conjuntos de dados histopatológicos: um de lesões cervicais e outro de metástase em linfonodos. O classificador adotado foi uma arquitetura ResNet34, aplicada para distinguir entre classes patológicas nas imagens. Quando treinado apenas com dados reais, o modelo alcançou acurácias de 90,4% (coluna cervical) e 86,3% (linfonodos). Com a introdução das imagens sintéticas selecionadas pelo HistoGAN, as acurácias aumentaram para 97,1% e 89,1%, resultando em ganhos de 6,7 e 2,8 pontos percentuais, respectivamente.

Benito-Del-Valle et al. (Benito-Del-Valle et al., 2024) investigaram a eficácia de diferentes abordagens de geração de imagens sintéticas, com foco especial em GANs, para tarefas de classificação histopatológica. O estudo foi conduzido utilizando o conjunto de dados PatchCamelyon (PCam), composto por *patches* de imagens histológicas rotulados como positivos ou

negativos para metástase. Os autores implementaram múltiplos modelos generativos, incluindo GANs condicionais, para criar novas amostras sintéticas a partir das classes reais. Como modelo discriminativo, foi utilizada a arquitetura ResNet18, com treinamento supervisionado sobre diferentes composições de conjuntos de dados: apenas imagens reais, imagens reais com augmentações tradicionais e conjuntos expandidos com imagens sintéticas filtradas. Os resultados demonstraram que a inclusão das imagens GAN filtradas proporcionou um ganho de acurácia, elevando o desempenho do classificador de 85,2% para 88,4%, superando também abordagens com augmentação geométrica padrão.

Kumar et al. (Kumar et al., 2023) propuseram o *XM-GAN*, um método de geração *few-shot* de imagens histopatológicas colorretais, ideal para cenários com escassez de dados. A arquitetura do modelo inclui um bloco de fusão controlada (Controllable Fusion Block) que agrega regiões locais de imagens de referência, garantindo coesão visual e diversidade nas amostras geradas. Imagens sintéticas foram avaliadas por especialistas (patologistas), que só conseguiram distinguir real de gerado cerca de 55% das vezes. A aplicação principal do XM-GAN foi o aumento de dados em tarefas de classificação com CNN (ResNet18). Foi observado um ganho medio de acurácia de 4,4 pontos (de 68,1% para 72,5%), em comparação com classificadores treinados apenas com os poucos exemplos reais disponíveis.

Kausar et al. (Kausar et al., 2022) propuseram o modelo *SA-GAN* (Stain Acclimation GAN), inicialmente desenvolvido para normalização de coloração (stain transfer) em imagens histológicas, mas também utilizado como estratégia de aumento. O SA-GAN gera novas imagens com coloração coerente e estruturada, preservando informações morfológicas relevantes do tecido. O classificador utilizado para avaliação nos experimentos foi uma arquitetura ResNet50, treinada para distinguir diferentes tipos de câncer no conjunto ICIAR 2018, que inclui imagens de neoplasias mamárias com variações de coloração. Quando treinado apenas com dados reais, o classificador alcançou uma acurácia de 86,1%. Com a inclusão das imagens sintéticas geradas pelo SA-GAN, essa taxa foi elevada para 93%, apresentando um ganho de 6,9 pontos percentuais.

Ruiz-Casado et al. (Ruiz-Casado; Molina-Cabello; Luque-Baena, 2024) avaliaram o uso de ProGAN e sua variante regularizada (ReGAN) como técnica de aumento artificial em imagens histopatológicas de câncer de mama. As imagens sintéticas foram incorporadas ao treinamento de classificadores CNN (Inception-v3, ResNet e VGG16), substituindo o aumento tradicional. O conjunto de dados incluiu imagens de tumores benignos e malignos, previamente rotuladas por especialistas. Os resultados mostraram que o aumento com ProGAN apresentou melhor desempenho do que métodos tradicionais ou *downsampling*, com precisão de classificação alta. A acurácia da CNN ResNet com ProGAN, por exemplo, teve um aumento de 92,6% (imagens originais) para 96,4% (imagens sintéticas), indicando um acréscimo de aproximadamente 3,8 pontos.

Zhou et al. (Jiang et al., 2023) desenvolveram o modelo *SMSG-GAN* (Multi-Scale Gradient GAN), voltado à geração de imagens histopatológicas colorretais com alta fidelidade visual e menor quantidade de artefatos em comparação a outras variantes de GAN. O modelo também

inclui um mecanismo de seleção de amostras sintéticas com base em similaridade de embeddings e pontuação de confiança. O classificador utilizado foi uma arquitetura InceptionV3 treinada sobre o dataset NCT-CRC-HE-100K, contendo nove classes de tecido colorretal. Quando treinado apenas com imagens reais, o classificador alcançou acurácia de 86,87%. Com a inclusão das imagens geradas pelo SMSG-GAN, essa acurácia subiu para 89,54%, representando um ganho de aproximadamente 2,7 pontos.

3.2 AUMENTO ARTIFICIAL VIA GANS E CLASSIFICADOR VIT

Apesar dos avanços significativos obtidos com a combinação de GANs e CNNs, esses modelos ainda apresentam limitações no que diz respeito à captura de relações espaciais globais e à capacidade de generalização em domínios com alta variabilidade morfológica, como o das imagens histopatológicas. Nesse cenário, os *Vision Transformers* surgem como uma alternativa promissora, substituindo as convoluções locais por mecanismos de atenção que permitem uma modelagem global da imagem desde as primeiras camadas da arquitetura. Embora ainda menos explorada na literatura, a integração entre GANs e ViTs tem demonstrado resultados encorajadores, apontando para um novo caminho na construção de classificadores mais robustos.

Li et al. (Li et al., 2024) propuseram um framework unificado para geração sintética e classificação simultânea de imagens histopatológicas, integrado por um modelo de GAN condicional baseado em Vision Transformer (ViT). Essa abordagem elimina a separação entre etapas: o mesmo modelo é responsável tanto pela síntese quanto pela inferência de rótulos, por meio da incorporação de uma cabeça de classificação diretamente no gerador. A arquitetura introduz técnica de *conditional class projection* para preservar a separação entre as classes na geração das imagens e um mecanismo de ponderação dinâmica de múltiplas funções de perda para equilibrar o treinamento adversarial e o processo supervisonado. Além disso, é implementado um sistema de seleção ativa de imagens sintéticas, utilizando critérios de confiança e realismo para decidir quais amostras acrescentar ao ciclo de treinamento. Os experimentos foram realizados utilizando o dataset PCam (PatchCamelyon), com classificadores integrados no próprio modelo ViT. Os resultados mostraram ganho consistente de acurácia em comparação a abordagens tradicionais que separam geração e classificação. Em cenários de avaliação, o modelo unificado alcançou acurácia até 94,5%, enquanto modelos treinados separadamente com CNN obtiveram acurácias da ordem de 91–92% em configurações similares.

3.3 CONSIDERAÇÕES SOBRE OS TRABALHOS RELACIONADOS

Neste capítulo, foram apresentados estudos que exploram o uso de Redes Geradoras Adversariais (GANs) como técnica de aumento artificial de dados, em conjunto com classificadores baseados em redes neurais profundas. As acurácias de todas as abordagens foram organizadas na Tabela 1.

Tabela 1 – Resumo dos principais trabalhos relacionados

Referência	**Abordagem**	**Acurácia (%)**
(Xue et al., 2021b)	HistoGAN + ResNet34	89,1
(Benito-Del-Valle et al., 2024)	GAN + ResNet18	88,4
(Kumar et al., 2023)	XM-GAN + ResNet18	72,5
(Kausar et al., 2022)	SA-GAN + ResNet50	93,0
(Ruiz-Casado; Molina-Cabello; Luque-Baena, 2024)	ProGAN/ReGAN + CNN	96,4
(Jiang et al., 2023)	SMSG-GAN + InceptionV3	89,5
(Li et al., 2024)	GAN condicional + ViT	94,5

Observando os resultados da Tabela 1, percebe-se que os modelos que integram GANs com CNNs apresentam variações significativas de desempenho conforme a arquitetura e a estratégia de treinamento utilizada. Por exemplo, abordagens como HistoGAN + ResNet34 (Xue et al., 2021b) e ProGAN/ReGAN + CNN (Ruiz-Casado; Molina-Cabello; Luque-Baena, 2024) demonstraram acurácias superiores a 96%, indicando que a escolha da arquitetura do gerador, bem como do classificador, tem impacto direto na qualidade das representações aprendidas e, consequentemente, na performance do modelo.

Por outro lado, abordagens como XM-GAN + ResNet18 (Kumar et al., 2023), que obtiveram acurácias mais modestas (72,5%), indicam que a simples utilização de dados sintéticos não garante melhoria no desempenho. Aspectos como o balanceamento entre classes, a proporção entre dados reais e gerados, a regularização das perdas adversariais e o controle da diversidade das amostras são fatores determinantes para evitar sobreajuste e garantir a representatividade dos dados. Além disso, estratégias que incorporam mecanismos de atenção, como SA-GAN + ResNet50 (Kausar et al., 2022), tendem a melhorar a capacidade do modelo de focar em regiões discriminativas das imagens, refletindo-se em um ganho expressivo de acurácia. O uso de Vision Transformers em combinação com GANs, embora ainda recente, revela-se promissor. O trabalho de Li et al. (Li et al., 2024) alcançou 94,5% de acurácia, demonstrando que a habilidade dos ViTs em modelar dependências espaciais globais pode compensar limitações intrínsecas das CNNs em lidar com variações morfológicas complexas, como as presentes em tecidos histológicos.

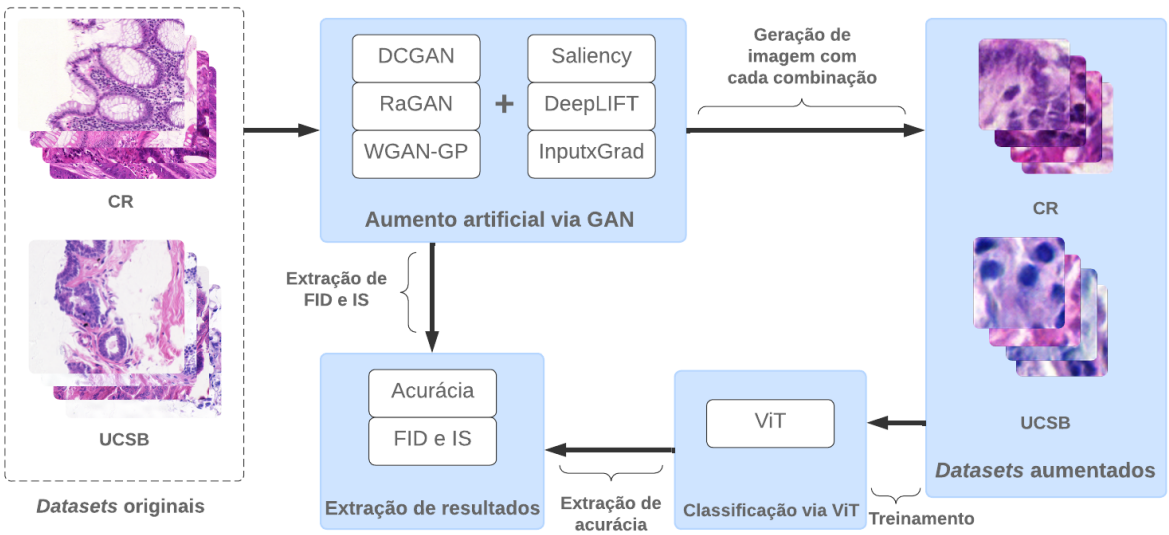
Em síntese, a análise da literatura evidencia que o desempenho dos modelos depende não apenas da arquitetura do GAN e do classificador, mas também de fatores complementares, como o volume e a diversidade dos dados sintéticos, a estratégia de treinamento empregada e a adequação das técnicas de regularização. Assim, observa-se um movimento progressivo no sentido de combinar arquiteturas mais expressivas, como os ViTs, a estratégias de geração controlada e guiada, o que pode representar um avanço significativo no desenvolvimento de sistemas mais precisos e confiáveis para análise de imagens histológicas.

4 METODOLOGIA

Neste capítulo, apresentam-se as estratégias e metodologias que orientaram a proposta desenvolvida, com ênfase no uso de Redes Geradoras Adversariais (GANs) e técnicas de Inteligência Artificial Explicável (XAI) aplicadas à tarefa de classificação de imagens histológicas H&E. Complementarmente, exploram-se os benefícios de arquiteturas baseadas em *Vision Transformers* (ViTs), destacando seu potencial na extração e representação de padrões visuais complexos presentes nas imagens. Também são discutidas as principais métricas utilizadas para avaliar a eficácia das abordagens adotadas.

A metodologia foi estruturada em três etapas principais. Na primeira etapa, denominada aumento artificial de dados, investigou-se a utilização de diferentes variantes de GANs, bem como de combinações entre GANs e técnicas de XAI, com o objetivo de gerar imagens sintéticas a partir de amostras reais de tecido corado (H&E). Na segunda etapa, as imagens geradas artificialmente foram utilizadas como entrada para modelos baseados em ViTs, os quais foram empregados na tarefa de classificação, buscando-se avaliar sua capacidade de generalização e identificação de padrões discriminativos. Por fim, a terceira etapa consistiu na avaliação quantitativa dos modelos envolvidos, por meio de métricas clássicas e específicas, como Acurácia, *Fréchet Inception Distance* (FID) e *Inception Score* (IS), visando identificar as combinações de modelos e estratégias com melhor desempenho na tarefa proposta.

Figura 7 – Esquema do método proposto.



Fonte: Elaborado pelo autor.

4.1 ETAPA 1: EXPLORAÇÃO DE TÉCNICAS DE AUMENTO ARTIFICIAL

Nesta etapa do trabalho, o aumento artificial do conjunto de imagens histológicas H&E foi realizado por meio de modelos generativos baseados em Redes Geradoras Adversariais (GANs), incluindo as variantes DCGAN (Radford; Metz; Chintala, 2016), RaGAN (Jolicoeur-Martineau, 2018) e WGAN-GP (Gulrajani et al., 2017). Cada um desses modelos foi treinado em dois regimes distintos: uma abordagem tradicional (sem interpretabilidade) e uma abordagem orientada por explicabilidade, utilizando métodos de Inteligência Artificial Explicável (XAI), como *Saliency* (Simonyan; Vedaldi; Zisserman, 2014), *DeepLIFT* (Shrikumar; Greenside; Kundaje, 2019) e *Input* × *Gradient* (Hechtlinger, 2016).

4.1.1 Hiperparâmetros

Com o objetivo de garantir a padronização dos experimentos e possibilitar comparações consistentes, todos os modelos foram treinados com os mesmos hiperparâmetros. O número de *epochs* (épocas), correspondente a ciclos completos sobre o conjunto de treinamento, foi fixado em 200. O *batch size*, que define o número de amostras por atualização de gradiente, foi mantido em 32. As capacidades das redes geradora e discriminadora foram controladas pelos parâmetros G_h_size e D_h_size, ambos definidos como 128, e o vetor de entrada latente (z_size) também foi fixado em 128 dimensões. A taxa de aprendizado utilizada tanto para a geradora quanto para a discriminadora foi estabelecida como 0.0001.

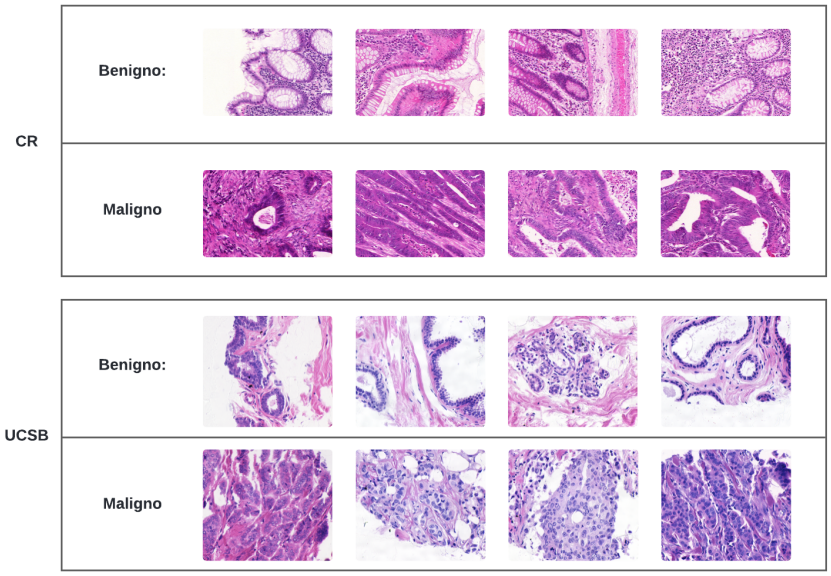
4.1.2 Conjunto de dados

Os experimentos de treinamento foram conduzidos utilizando dois conjuntos de dados de imagens histológicas coradas com Hematoxilina e Eosina, compostos por amostras de câncer colorretal e de câncer de mama. A Figura 8 apresenta exemplos visuais das imagens utilizadas.

1. **Câncer Colorretal (CR):** composto por imagens obtidas a partir de 16 seções histológicas de câncer colorretal nos estágios T3 ou T4, digitalizadas com o scanner Zeiss MIRAX MIDI em resolução de 0,465 µm/pixel. As imagens foram rotuladas como benignas ou malignas, totalizando 165 amostras (74 benignas e 91 malignas), com dimensões variando entre 775×522 e 581×442 pixels.

2. **Câncer de Mama (UCSB):** formado por 58 imagens histológicas oriundas de biópsias, fornecidas pela University of California, Santa Barbara. As imagens têm resolução de 896×768 pixels, em RGB, com profundidade de cor de 24 bits.

Figura 8 – Exemplos de imagens H&E utilizadas.



Fonte: Elaborado pelo autor.

Devido à limitação do número de imagens disponíveis, foi implementada uma etapa de extração de regiões de interesse (ROI, *Region of Interest*) (Jan; Zainal; Jamaludin, 2020), com o objetivo de uniformizar e maximizar o aproveitamento das imagens originais. Essa extração permitiu a segmentação e padronização das amostras para o tamanho fixo de 64×64 pixels, facilitando o treinamento dos modelos e acelerando o tempo de inferência.

Com essa abordagem, o volume total de amostras foi significativamente ampliado: o conjunto colorretal passou a conter 10.400 imagens (4664 benignas e 5736 malignas), enquanto o conjunto de câncer de mama passou a conter 8352 imagens (4608 benignas e 3744 malignas). Essas imagens foram então utilizadas para o treinamento dos diferentes modelos generativos propostos neste trabalho.

4.2 ETAPA 2: EXPLORAÇÃO DE MODELOS VIT

Nesta etapa, as imagens geradas por meio das abordagens de aumento artificial com GANs tradicionais e orientadas por XAI foram utilizadas como entrada para modelos Vision Transformers (ViTs), com o objetivo de realizar a tarefa de classificação de amostras histológicas H&E. Foram empregados três modelos distintos: o ViT Base (Dosovitskiy et al., 2021a), o *Data-efficient Image Transformer* (DeiT) (Touvron et al., 2021) e o *Pyramid Vision Transformer* (PVT) (Wang et al., 2021).

Os modelos adotados foram implementados com o auxílio de uma biblioteca que fornece versões pré-treinadas e otimizadas das arquiteturas ViT, permitindo a aplicação direta do *transfer learning*. Nesse contexto, os pesos das camadas iniciais (*patch embedding* e blocos Transformer)

foram preservados, enquanto as camadas finais de classificação foram adaptadas para refletir o número de classes do problema em questão.

Com o intuito de aumentar a variabilidade dos dados disponíveis e melhorar a capacidade de generalização dos modelos ViT, foi aplicado um aumento artificial de 100% sobre cada um dos conjuntos de dados descritos em 4.1.2. Isso significa que, para cada conjunto original, foi adicionado um volume igual de imagens sintéticas geradas pelas diferentes combinações de modelos GAN e GAN + XAI. Essa duplicação do volume de dados foi especialmente relevante considerando que arquiteturas Transformer, por sua natureza, demandam grandes quantidades de dados para treinamento eficaz, sob risco de *overfitting* quando submetidas a bases limitadas (Zhu; Chen; Yang, 2023).

O processo de treinamento foi realizado utilizando a técnica de *transfer learning*, com os parâmetros ajustados por um máximo de 10 épocas, taxa de aprendizado de 0.001 e otimizador *Adamax* (Wan et al., 2021). Apenas a última camada totalmente conectada de cada arquitetura foi atualizada durante o treinamento, preservando os pesos das camadas anteriores para maximizar o aproveitamento do conhecimento prévio.

4.2.1 Divisão dos Dados e Avaliação com Validação Cruzada

A avaliação dos modelos foi conduzida de forma rigorosa, combinando divisão estratificada dos dados e validação cruzada. Inicialmente, os dados foram divididos em três conjuntos: 60% para treinamento, 20% para validação e 20% para teste. A etapa de treinamento (60%) foi submetida à técnica de validação cruzada k-fold, com $k = 5$, a fim de assegurar maior robustez na estimativa de desempenho dos modelos.

Durante cada iteração da validação cruzada, o conjunto de treinamento é dividido em cinco subconjuntos: quatro utilizados para o ajuste dos pesos da rede e um para validação. Esse processo é repetido cinco vezes, garantindo que todos os dados sejam utilizados tanto para treinamento quanto para validação, em momentos distintos. Após a escolha do melhor modelo com base nos resultados médios da validação cruzada, a performance final é aferida no conjunto de teste, não exposto durante o processo de ajuste. Essa estratégia proporciona uma avaliação abrangente e confiável da capacidade de generalização dos classificadores baseados em Vision Transformers.

4.3 ETAPA 3: ANÁLISES E EXTRAÇÃO DE CONHECIMENTO

A última etapa deste trabalho consistiu na avaliação quantitativa dos modelos geradores e classificadores, com o objetivo de identificar as combinações mais eficazes entre as arquiteturas de GAN e ViT empregadas. Para isso, foram utilizadas métricas específicas e adequadas ao propósito de cada tipo de modelo.

No caso dos modelos generativos, a análise da qualidade das imagens sintéticas produzidas foi realizada por meio das métricas *Fréchet Inception Distance* (FID) e *Inception Score* (IS),

previamente detalhadas em 2.4. Essas métricas permitiram aferir a similaridade estatística entre as imagens reais e geradas (FID) e a diversidade e realismo visual das amostras geradas (IS), oferecendo subsídios objetivos para comparar o desempenho entre as diferentes combinações de GANs e métodos XAI.

Para os modelos classificadores baseados em *Vision Transformers*, a métrica primária utilizada foi a acurácia, também abordada anteriormente em 2.4. Essa métrica possibilitou avaliar a capacidade de generalização dos modelos diante de amostras anteriormente não vistas, sendo aplicada tanto nos testes finais quanto nos ciclos internos de validação cruzada.

4.4 BIBLIOTECAS E FERRAMENTAS DE DESENVOLVIMENTO

Tanto os algoritmos utilizados para o treinamento dos modelos generativos baseados em GANs quanto os métodos de classificação de imagens com *Vision Transformers* (ViTs) foram desenvolvidos com o uso da linguagem de programação *Python*. Para a implementação dos modelos generativos, utilizou-se o framework *PyTorch* (Paszke et al., 2019), que oferece suporte eficiente ao desenvolvimento e treinamento de redes neurais profundas. A visualização das imagens geradas foi facilitada pelo uso da biblioteca *TorchVision* (maintainers; contributors, 2016), integrada ao ecossistema do *PyTorch*.

A aplicação das técnicas de Inteligência Artificial Explicável, como *Saliency*, *DeepLIFT* e *Input* × *Gradient*, foi realizada com o auxílio da biblioteca *Captum* (Kokhlikyan et al., 2020), projetada especificamente para explicar modelos desenvolvidos em PyTorch. Para a etapa de classificação, foi empregada a biblioteca *Timm* (*PyTorch Image Models*) (Wightman, 2019), que disponibiliza uma ampla variedade de arquiteturas de *Vision Transformers* pré-treinadas, otimizando a experimentação e reprodutibilidade dos modelos. A avaliação dos modelos foi conduzida utilizando as bibliotecas *scikit-learn* (Pedregosa et al., 2011), para o cálculo de métricas de classificação como acurácia, e *TorchMetrics*, para a computação dos índices específicos para modelos generativos, como o *Fréchet Inception Distance* (FID) e o *Inception Score* (IS).

Por fim, os experimentos foram realizados em um ambiente local com as seguintes especificações de hardware: processador AMD Ryzen 7 5700X (3.4 GHz), 16 GB de memória RAM, placa de vídeo NVIDIA RTX 3060 e sistema operacional Linux.

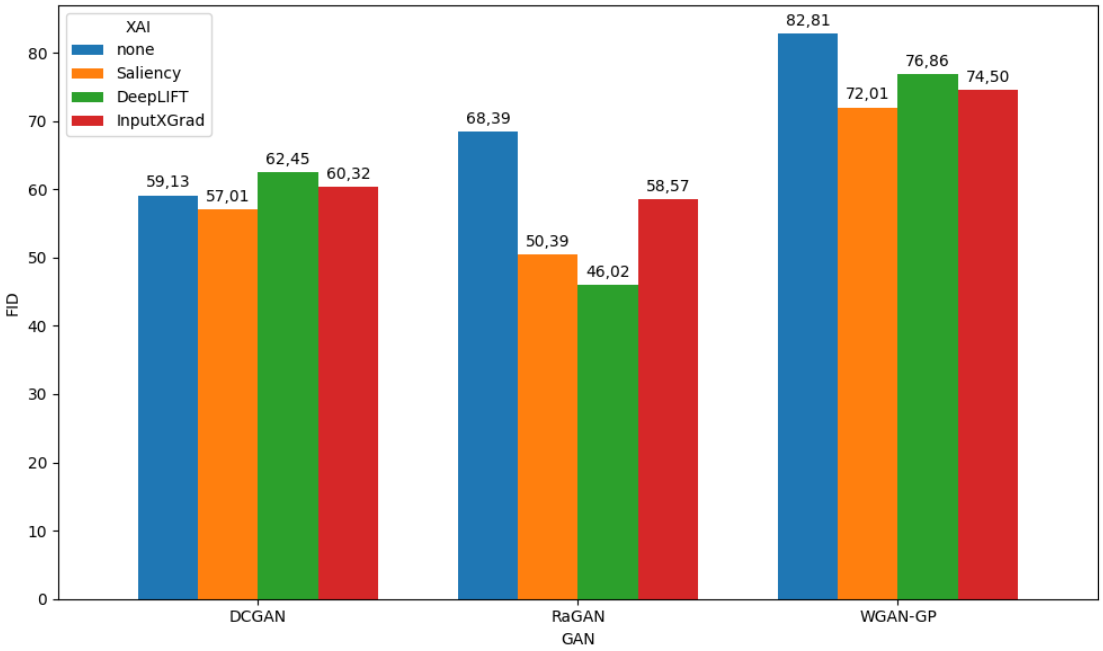
5 RESULTADOS

Neste capítulo estão os resultados conquistados a partir do estudo proposto para o contexto de aumento artificial com imagens H&E e classificadores ViT. Os destaques são as informações sobre os desempenhos obtidos com as diferentes composições de GAN tradicional e GAN + XAI para a melhora na classificação das imagens. Por fim, os desempenhos calculados via aplicação dos métodos FID e IS também são indicados com o propósito de fornecer uma justificativa para os resultados obtidos por cada uma das melhores combinações identificadas.

5.1 PRINCIPAIS PADRÕES E COMBINAÇÕES DE MODELOS GAN E XAI

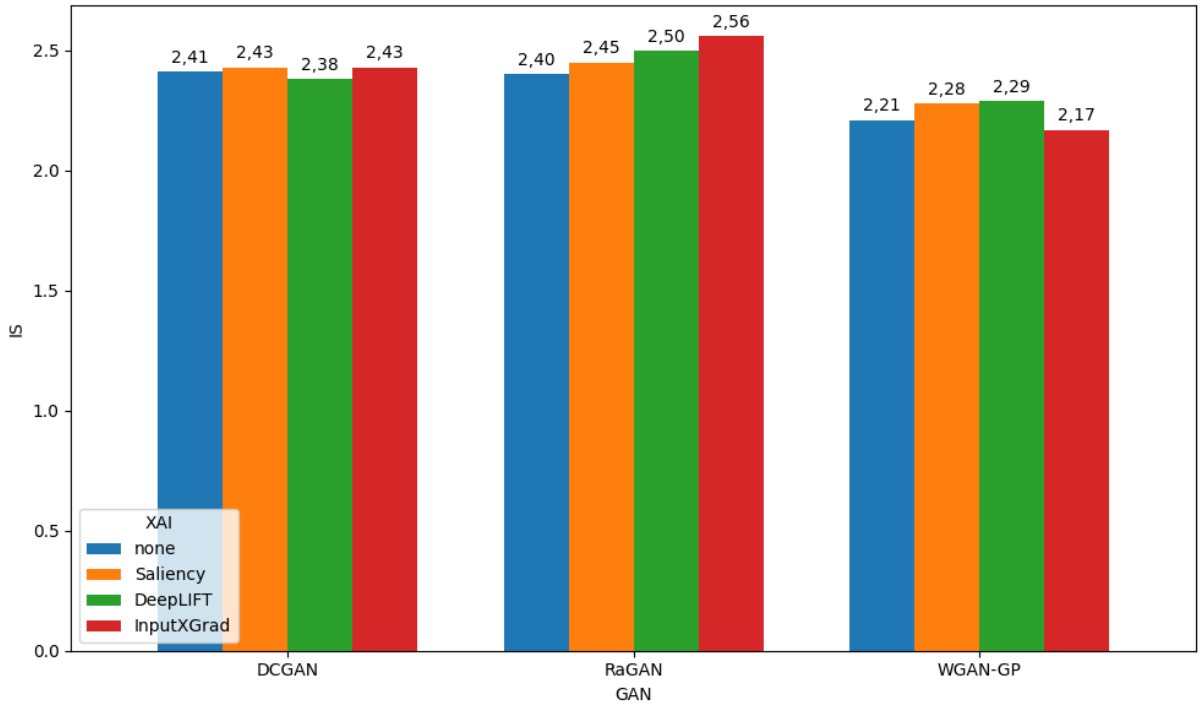
Os resultados apresentados nesta seção sintetizam as principais combinações obtidas ao longo da execução da proposta, contemplando as variações de desempenho decorrentes da aplicação dos métodos de XAI aos modelos generativos DCGAN, RaGAN e WGAN-GP. As métricas FID e IS foram empregadas para avaliar a qualidade e a diversidade das imagens sintéticas produzidas, bem como para identificar padrões de comportamento entre as diferentes configurações testadas, conforme as metodologias descritas no Capítulo 4. As distribuições dos resultados obtidos são apresentadas nas Figuras 9 a 12.

Figura 9 – Gráfico com valores de FID para CR.



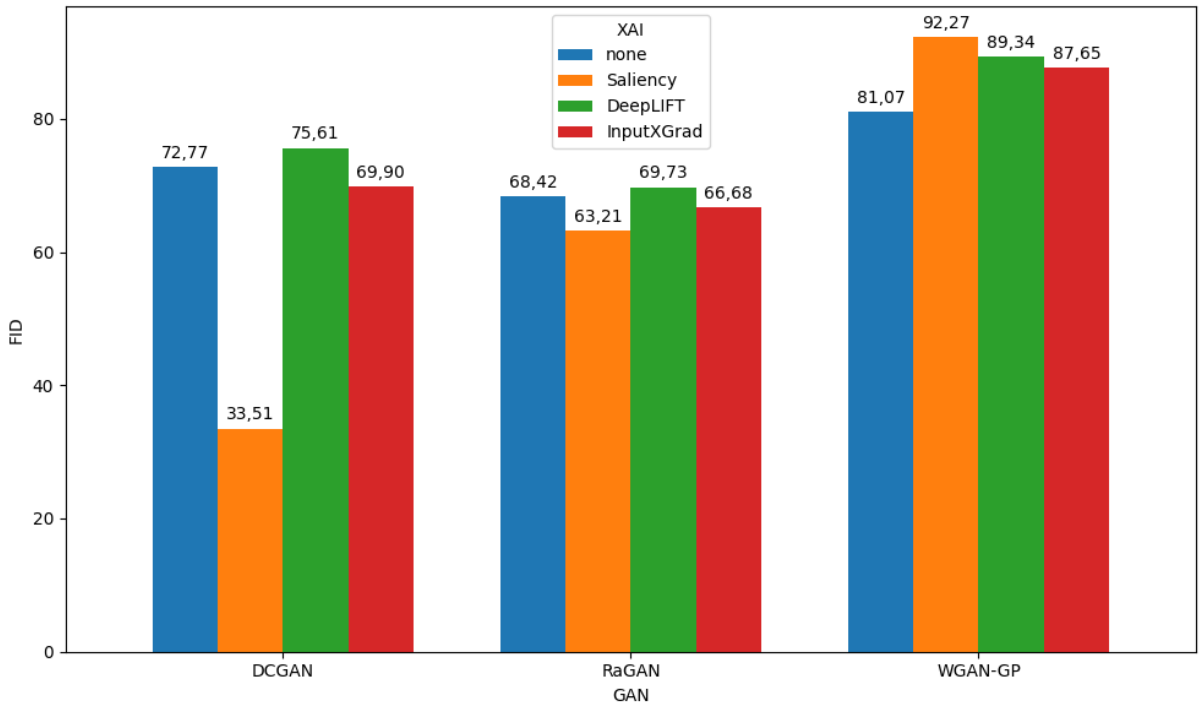
Fonte: Elaborado pelo autor.

Figura 10 – Gráfico com valores de IS para CR.



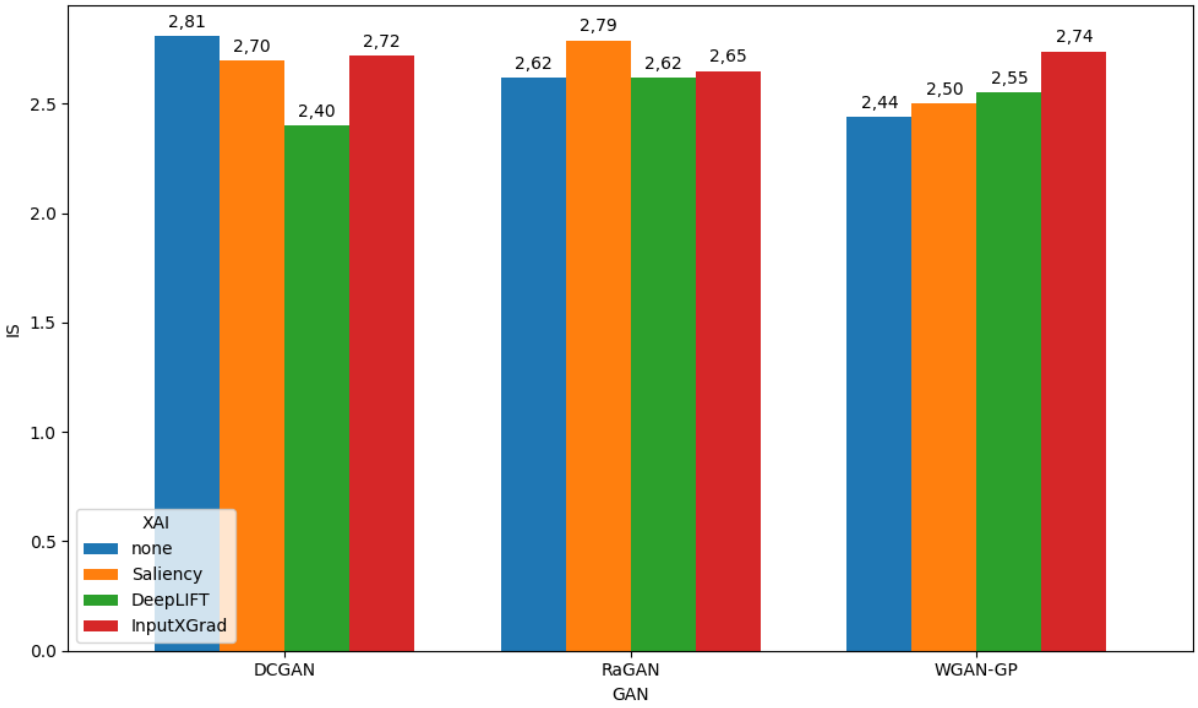
Fonte: Elaborado pelo autor.

Figura 11 – Gráfico com valores de FID UCSB.



Fonte: Elaborado pelo autor.

Figura 12 – Gráfico com valores de IS para UCSB.



Fonte: Elaborado pelo autor.

Para o dataset CR, observa-se que entre as variações do DCGAN, o melhor resultado de FID foi obtido com o uso de *Saliency* (FID = 57,01), indicando imagens sintéticas mais próximas das reais. Já os maiores valores de IS foram registrados com *Saliency* e InputXGrad (ambos IS = 2,43), sugerindo uma leve melhora na diversidade e qualidade perceptiva. A combinação com *DeepLIFT* apresentou desempenho inferior (FID = 62,45; IS = 2,38), ficando atrás das demais, enquanto o DCGAN puro teve um FID de 59,13, também abaixo da versão com *Saliency*. No caso do RaGAN, a melhor performance em FID foi obtida com *DeepLIFT* (FID = 46,02), representando uma redução considerável em relação ao modelo base (FID = 68,39).

Já o maior IS foi registrado com *InputXGrad* (IS = 2,56), seguido por *DeepLIFT* (IS = 2,50), o que demonstra melhora na variabilidade perceptiva das imagens geradas. A combinação com *Saliency* também obteve bons resultados (FID = 50,39; IS = 2,45), ficando à frente da versão sem XAI. No modelo WGAN-GP, a configuração com *Saliency* novamente obteve o melhor FID (72,01), seguida de *InputXGrad* (74,50) e *DeepLIFT* (76,86), todas com desempenho superior ao baseline (FID = 82,81). O maior valor de IS foi alcançado com *Saliency* (IS = 2,29), indicando ganhos na diversidade perceptiva em relação à versão sem XAI (IS = 2,21). O uso de *InputXGrad* teve o pior IS nesse modelo (2,17), apesar de ainda apresentar melhor FID que o baseline.

De maneira análoga, os resultados do dataset UCSB também revelam padrões consistentes. No DCGAN, a combinação com *Saliency* se destacou fortemente com o menor FID (33,51), refletindo excelente qualidade de geração. O maior IS também foi registrado nessa combinação

(IS = 2,81), superando o modelo base (FID = 72,77; IS = 2,81). Já as combinações com *DeepLIFT* (FID = 75,61; IS = 2,40) e *InputXGrad* (FID = 69,90; IS = 2,72) mostraram desempenho inferior ao método *Saliency*, embora o IS da última ainda tenha se mantido alto. Para o RaGAN, a melhor combinação foi novamente com *Saliency*, obtendo o menor FID (63,21) e o maior IS (2,79), superando claramente o modelo sem XAI (FID = 68,42; IS = 2,62).

As demais variações apresentaram resultados razoáveis, com *InputXGrad* (FID = 66,68; IS = 2,65) e *DeepLIFT* (FID = 69,73; IS = 2,62) mostrando pequenas melhorias no FID em relação ao modelo base, porém sem destaque significativo nos valores de IS. No caso do WGAN-GP, a melhor combinação foi com *InputXGrad*, que obteve o menor FID (87,65) e o maior IS (2,74), demonstrando melhora tanto na fidelidade quanto na diversidade das imagens geradas. *DeepLIFT* também superou o baseline (FID = 89,34; IS = 2,55), enquanto a versão com *Saliency* obteve desempenho intermediário (FID = 92,27; IS = 2,50). O modelo sem XAI teve o pior FID (81,07), mas seu IS (2,44) foi inferior ao das versões com XAI.

5.2 PRINCIPAIS PADRÕES E COMBINAÇÕES DE AUMENTO ARTIFICIAL E VITS

Nesta seção, os resultados das métricas FID e IS são comparados às acurácias obtidas pelos classificadores, com o objetivo de identificar as combinações de técnicas que proporcionaram ganhos significativos de desempenho no modelo ViT. Busca-se, assim, destacar as configurações mais eficazes de aumento artificial de dados e integração com métodos de XAI para os conjuntos CR e UCSB. Os valores de acurácia, juntamente com informações complementares sobre cada experimento, encontram-se apresentados nas Tabelas 2 e 3.

Tabela 2 – Acurácia com diferentes modelos de aumento artificial no dataset CR.

Combinação	**ViT**	**DeiT**	**PVT**
Sem aumento	81,29%	87,89%	91,81%
DCGAN	82,67%	87,48%	91,80%
DCGAN + Saliency	**84,14%**	**87,70%**	**92,24%**
DCGAN + DeepLIFT	79,97%	87,63%	91,48%
DCGAN + InputXGrad	82,89%	87,25%	91,98%
RaGAN	83,53%	**88,44%**	91,02%
RaGAN + Saliency	81,87%	88,33%	90,98%
RaGAN + DeepLIFT	**84,39%**	87,28%	**91,92%**
RaGAN + InputXGrad	80,56%	87,28%	91,86%
WGAN-GP	81,60%	86,52%	91,15%
WGAN-GP + Saliency	**84,41%**	**88,09%**	91,98%
WGAN-GP + DeepLIFT	78,25%	87,91%	91,54%
WGAN-GP + InputXGrad	83,34%	87,96%	**92,13%**

Fonte: Elaborado pelo autor.

Tabela 3 – Acurácia com diferentes modelos de aumento artificial no dataset UCSB.

Combinação	**ViT**	**DeiT**	**PVT**
Sem aumento	**71,30%**	73,36%	77,27%
DCGAN	72,21%	74,59%	77,84%
DCGAN + Saliency	**72,36%**	74,78%	77,14%
DCGAN + DeepLIFT	72,04%	75,01%	76,50%
DCGAN + InputXGrad	71,44%	**75,14%**	**78,36%**
RaGAN	71,30%	73,38%	77,00%
RaGAN + Saliency	**72,91%**	**75,88%**	75,82%
RaGAN + DeepLIFT	71,75%	73,08%	76,98%
RaGAN + InputXGrad	71,61%	75,12%	**77,92%**
WGAN-GP	72,86%	75,03%	77,14%
WGAN-GP + Saliency	70,91%	74,54%	76,98%
WGAN-GP + DeepLIFT	71,66%	75,20%	76,63%
WGAN-GP + InputXGrad	**73,51%**	**77,42%**	**77,77%**

Fonte: Elaborado pelo autor.

Para o dataset *CR*, o modelo sem aumento apresentou uma acurácia de 81,29% com o ViT, 87,89% com o DeiT e 91,81% com o PVT, servindo como ponto de partida para avaliar o impacto do aumento de dados com GANs combinados (ou não) com métodos XAI. No caso do DCGAN, a combinação sem explicabilidade apresentou desempenho competitivo, com acurácias próximas ao baseline (82,67% em ViT, 87,48% em DeiT e 91,80% em PVT). Ao incorporar técnicas XAI, os resultados foram mistos.

A combinação com *Saliency* obteve os melhores resultados dentro dessa arquitetura, superando as demais variantes nas três arquiteturas analisadas e atingindo 84,14% em ViT, 87,70% em DeiT e 92,24% em PVT. Já o método *DeepLIFT* obteve acurácias inferiores ao modelo base em duas das três arquiteturas (79,97% em ViT e 91,48% em PVT), indicando um possível impacto negativo dependendo da configuração. O *InputXGrad* manteve resultados consistentes e próximos ao modelo sem XAI (82,89% em ViT, 87,25% em DeiT e 91,98% em PVT). Em termos absolutos, o melhor resultado em ViT no CR foi obtido por WGAN-GP + *Saliency* (84,41%), o melhor em DeiT foi RaGAN sem XAI (88,44%) e o melhor em PVT foi DCGAN + *Saliency* (92,24%).

Para o RaGAN, o desempenho sem explicabilidade foi elevado, com 83,53% em ViT e 88,44% em DeiT, este último representando a maior acurácia observada para DeiT no conjunto CR. Entre as variantes com XAI, a combinação RaGAN + *DeepLIFT* apresentou picos de desempenho individuais, alcançando 84,39% em ViT e 91,92% em PVT, ambos superiores ao RaGAN sem XAI em suas respectivas arquiteturas. O método *Saliency* manteve estabilidade (81,87%, 88,33%, 90,98%), enquanto o *InputXGrad* mostrou boa consistência sem superar os melhores valores absolutos. Esses resultados indicam que, no RaGAN, certas técnicas XAI

podem complementar a geração sintética de forma a melhorar representações úteis para o classificador, em especial o DeepLIFT em configurações específicas.

No caso do WGAN-GP, o modelo base sem XAI apresentou acurácias discretas (81,60% em ViT, 86,52% em DeiT e 91,15% em PVT). A introdução dos métodos XAI trouxe melhorias relevantes em alguns cenários. WGAN-GP + *Saliency* atingiu 84,41% em ViT e 88,09% em DeiT, sendo este um dos melhores desempenhos observados para ViT no CR. WGAN-GP + *InputXGrad* alcançou 83,34% em ViT, 87,96% em DeiT e 92,13% em PVT, este último sendo um dos maiores valores absolutos no conjunto CR, embora ligeiramente inferior ao pico do DCGAN + *Saliency* em PVT. O *DeepLIFT* aplicado ao WGAN-GP apresentou desempenho inferior em ViT (78,25%), sugerindo sensibilidade dessa técnica ao tipo de gerador e arquitetura de classificador.

Para o dataset *UCSB*, o comportamento foi distinto em relação ao CR. O modelo sem aumento apresentou acurácias de 71,30% em ViT, 73,36% em DeiT e 77,27% em PVT. Em ViT, várias combinações superaram o baseline, com destaque para WGAN-GP + *InputXGrad* (73,51%), que obteve a maior acurácia em ViT no UCSB. Para DeiT, as combinações com *InputXGrad* e *DeepLIFT* na linha DCGAN foram competitivas, mas o maior valor absoluto em DeiT surgiu com WGAN-GP + *InputXGrad* (77,42%). Em PVT, a melhor combinação foi DCGAN + *InputXGrad* (78,36%), seguida por RaGAN + *InputXGrad* (77,92%) e WGAN-GP + *InputXGrad* (77,77%), indicando que o *InputXGrad* foi particularmente benéfico para PVT no UCSB.

A média das acurácias das combinações com XAI no UCSB varia entre os geradores, com algumas linhas mostrando leve melhoria média e outras não. Especificamente, várias combinações com *InputXGrad* e algumas com *Saliency* superaram o baseline em arquiteturas como DeiT e PVT, enquanto o efeito no ViT foi mais heterogêneo. Esses resultados sugerem que o aproveitamento das imagens sintéticas explicáveis depende fortemente da arquitetura do classificador e da combinação gerador-técnica XAI.

Complementando a análise com os resultados de FID e IS das melhores combinações, observa-se correlação entre maior qualidade e variabilidade das imagens geradas e melhor desempenho do classificador em vários casos. Por exemplo, no CR, RaGAN + *DeepLIFT*, que apresentou os picos em ViT e PVT para o RaGAN, teve FID = 46,02 e IS = 2,50, os melhores valores dentro do RaGAN nesse dataset. De forma similar, WGAN-GP + *Saliency*, com excelente desempenho em ViT e DeiT, apresentou FID = 72,01 e IS = 2,28, melhores que o WGAN-GP puro (FID = 82,81; IS = 2,17).

No UCSB, o padrão de correlação também aparece: RaGAN + *Saliency*, melhor em algumas configurações de DeiT, teve FID = 63,21 e IS = 2,79, enquanto DCGAN + *InputXGrad*, destaque em PVT, apresentou FID = 69,90 e IS = 2,72. Esses padrões indicam que a integração entre GANs e métodos XAI pode tanto melhorar a interpretabilidade quanto produzir amostras sintéticas cuja qualidade favorece a generalização dos classificadores, especialmente quando a escolha do gerador e da técnica XAI é alinhada com a arquitetura do classificador.

De modo geral, os resultados obtidos evidenciam três tendências principais. Primeiro, a arquitetura PVT demonstrou desempenho consistentemente superior, sugerindo maior capacidade de generalização e robustez frente às variações introduzidas pelas imagens sintéticas. Segundo, entre as técnicas de explicabilidade, o *InputXGrad* e *Saliency* apresentaram o comportamento mais estável e vantajoso, com ganhos recorrentes em diferentes arquiteturas e datasets, enquanto o *DeepLIFT* mostrou maior variabilidade e sensibilidade à configuração de treinamento. Por fim, a correlação observada entre menores valores de FID, maiores valores de IS e melhores acurácias reforça que a qualidade e a diversidade das imagens geradas influenciam diretamente o desempenho do classificador.

5.3 DISCUSSÃO SOBRE OS RESULTADOS OBTIDOS

Por fim, a relevância dos dois melhores resultados foi verificada em relação aos trabalhos correlatos, a fim de conhecer se a solução obtida enquadra-se dentre as principais abordagens realizadas com o processo de classificação e aumento artificial de dados via GANs. É importante observar que as técnicas utilizadas para essa visão geral foram desenvolvidas com estratégias e modelos distintos, além de terem sido testadas em bases diferentes das que foram exploradas neste estudo. Essa visão geral está disponível na Tabela 4.

Tabela 4 – Visão geral das acurácias de classificadores treinados (ordenado por acurácia).

Referência	**Abordagem**	**ACC (%)**
(Xue et al., 2021b)	HistoGAN + ResNet34	97,1
(Ruiz-Casado; Molina-Cabello; Luque-Baena, 2024)	ProGAN/ReGAN + CNN	96,4
(Li et al., 2024)	GAN condicional + ViT	94,5
(Kausar et al., 2022)	SA-GAN + ResNet50	93,0
Proposto	DCGAN + Saliency + PVT	92,24
(Jiang et al., 2023)	SMSG-GAN + InceptionV3	89,5
(Benito-Del-Valle et al., 2024)	GAN + ResNet18	88,4
(Kumar et al., 2023)	XM-GAN + ResNet18	72,5

Fonte: Elaborado pelo autor.

Analisando os resultados apresentados na Tabela 4, observa-se que a metodologia proposta alcançou resultados com alguns destaques, mesmo diante de condições e arquiteturas distintas empregadas nos trabalhos correlatos. Embora algumas abordagens, como (Xue et al., 2021b) (97,1%) e (Li et al., 2024) (94,5%), tenham apresentado acurácias superiores, é importante considerar que essas comparações envolvem diferentes volumes de dados, classificadores e estratégias de geração sintética, o que limita a equivalência direta entre os experimentos.

O principal aspecto inovador deste trabalho está na integração entre métodos de explicabilidade (XAI) e modelos generativos (GANs), aplicados de forma a guiar o processo de geração sintética e refinar a diversidade das amostras produzidas. Diferentemente dos trabalhos correlatos,

que utilizam XAI apenas na interpretação pós-treinamento, aqui a explicabilidade foi incorporada de maneira ativa no *pipeline* de geração, orientando o modelo a produzir imagens mais coerentes com as características morfológicas reais. Essa estratégia contribuiu para a redução do FID e aumento do IS, indicando melhor qualidade perceptual e maior proximidade entre distribuições reais e sintéticas. Outro diferencial relevante é a adoção de *Vision Transformers* em um contexto de escassez de dados, o que tradicionalmente representa uma limitação para esse tipo de arquitetura.

A metodologia proposta demonstrou que, com o uso de dados aumentados por GANs e orientados por XAI, é possível viabilizar o treinamento eficaz de modelos transformer mesmo em bases reduzidas, mantendo resultados relevantes em relação às abordagens convencionais. Dessa forma, ainda que o desempenho quantitativo não tenha superado todos os trabalhos correlatos, os avanços metodológicos apresentados representam uma contribuição significativa para o campo.

6 CONCLUSÃO

Neste estudo, foi proposta uma análise detalhada do aumento artificial de imagens via redes generativas adversariais (GANs), aliada ao uso de técnicas de explicabilidade (XAI), com o objetivo de melhorar a acurácia de classificadores baseados em *Vision Transformers* (ViT) em tarefas de classificação de imagens médicas H&E. Os experimentos foram conduzidos em duas bases de dados distintas e permitiram avaliar o impacto de diferentes combinações entre modelos de GAN, métodos XAI e arquiteturas de classificadores, considerando métricas como acurácia, FID e IS.

Os resultados obtidos mostraram que a combinação de *DCGAN* com *Saliency* e *PVT* obteve a melhor acurácia do estudo, alcançando 92,24%, seguida de *WGAN-GP* com *InputXGrad* e *PVT*, com 92,13%. Tais resultados são competitivos em relação aos principais trabalhos da literatura, mesmo considerando que foram obtidos com um número reduzido de imagens, uma limitação comum em cenários biomédicos e um desafio para modelos ViT, que geralmente demandam grandes volumes de dados.

Adicionalmente, observou-se que o uso de métodos de explicabilidade, como *Saliency* e *InputXGrad*, contribuiu de forma significativa não apenas para o incremento da interpretabilidade dos modelos, mas também para a melhoria na qualidade das imagens geradas, refletindo positivamente nos valores de FID e IS. Isso evidencia que a integração de XAI ao pipeline de geração e classificação pode agregar valor não só na compreensão do modelo, mas também na eficácia preditiva.

Portanto, este trabalho demonstra que é possível obter resultados expressivos em tarefas de classificação de imagens médicas, mesmo com dados limitados, por meio da sinergia entre GANs, XAI e ViTs. A principal contribuição está na demonstração de que a explicabilidade pode ser utilizada como um mecanismo não apenas interpretativo, mas também funcional para o aprimoramento do desempenho em classificações biomédicas com uso de dados sintéticos.

Como trabalhos futuros, pretende-se explorar arquiteturas mais recentes de GANs e técnicas XAI mais robustas, bem como investigar a hibridização entre *Vision Transformers* e redes convolucionais. Além disso, será avaliado o impacto de estratégias para aumentar a diversidade e qualidade das imagens sintéticas, visando elevar ainda mais o desempenho e a confiabilidade dos modelos de classificação.

REFERÊNCIAS

ADADI, A.; BERRADA, M. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). **IEEE Access**, v. 6, p. 52138–52160, 2018. Disponível em: https://doi.org/10.1109/ACCESS.2018.2870052.

AGARAP, A. F. **Deep Learning using Rectified Linear Units (ReLU)**. 2019. Disponível em: https://arxiv.org/abs/1803.08375.

ARJOVSKY, M.; CHINTALA, S.; BOTTOU, L. **Wasserstein GAN**. 2017. Disponível em: https://arxiv.org/abs/1701.07875.

BENITO-DEL-VALLE, L. et al. **Unleashing the Potential of Synthetic Images: A Study on Histopathology Image Classification**. 2024. Disponível em: https://arxiv.org/abs/2409.16002.

DARGAN, S. et al. A survey of deep learning and its applications: a new paradigm to machine learning. **Archives of Computational Methods in Engineering**, Springer, v. 27, n. 4, p. 1071–1092, 2020.

DOSOVITSKIY, A. et al. **An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale**. 2021. Disponível em: https://arxiv.org/abs/2010.11929.

DOSOVITSKIY, A. et al. **An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale**. 2021. Disponível em: https://arxiv.org/abs/2010.11929.

DUBEY, S. R.; SINGH, S. K.; CHAUDHURI, B. B. **Activation Functions in Deep Learning: A Comprehensive Survey and Benchmark**. 2022. Disponível em: https://arxiv.org/abs/2109.14545.

FRID-ADAR, M. et al. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. **Neurocomputing**, Elsevier BV, v. 321, p. 321–331, dez. 2018. ISSN 0925-2312. Disponível em: http://dx.doi.org/10.1016/j.neucom.2018.09.013.

GOODFELLOW, I. et al. Generative adversarial nets. In: **Advances in neural information processing systems**. [S.l.: s.n.], 2014. p. 2672–2680.

Google Developers. **Estrutura de uma GAN**. 2023. https://developers.google.com/machine-learning/gan/gan_structure?hl=pt-br. Acesso em: 11 jul. 2025.

GULRAJANI, I. et al. **Improved Training of Wasserstein GANs**. 2017.

HECHTLINGER, Y. **Interpretation of Prediction Models Using the Input Gradient**. 2016.

HEUSEL, M. et al. **GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium**. 2018. Disponível em: https://arxiv.org/abs/1706.08500.

JAN, M. M.; ZAINAL, N.; JAMALUDIN, S. Region of interest-based image retrieval techniques: a review. v. 9, p. 520, 05 2020.

JIANG, L. et al. An improved multi-scale gradient generative adversarial network for enhancing classification of colorectal cancer histological images. **Frontiers in Oncology**, Volume 13 - 2023, 2023. ISSN 2234-943X. Disponível em: https://www.frontiersin.org/journals/oncology/articles/10.3389/fonc.2023.1240645.

JOLICOEUR-MARTINEAU, A. **The relativistic discriminator: a key element missing from standard GAN**. 2018.

KANG, C. et al. **Variability Matters : Evaluating inter-rater variability in histopathology for robust cell detection**. 2022. Disponível em: https://arxiv.org/abs/2210.05175.

KAUSAR, T. et al. Sa-gan: Stain acclimation generative adversarial network for histopathology image analysis. **Applied Sciences**, v. 12, n. 1, 2022. ISSN 2076-3417. Disponível em: https://www.mdpi.com/2076-3417/12/1/288.

KHAN, S. et al. Transformers in vision: A survey. **ACM computing surveys (CSUR)**, ACM New York, NY, v. 54, n. 10s, p. 1–41, 2022.

KOKHLIKYAN, N. et al. **Captum: A unified and generic model interpretability library for PyTorch**. 2020.

KOMURA, D.; ISHIKAWA, S. Machine learning methods for histopathological image analysis. **Computational and Structural Biotechnology Journal**, Elsevier, v. 16, p. 34–42, 2018.

KUMAR, A. et al. **Cross-modulated Few-shot Image Generation for Colorectal Tissue Classification**. 2023. Disponível em: https://arxiv.org/abs/2304.01992.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature**, v. 521, n. 7553, p. 436–444, 2015. Disponível em: https://doi.org/10.1038/nature14539.

LI, M. et al. **Unified Framework for Histopathology Image Augmentation and Classification via Generative Models**. 2024. Disponível em: https://arxiv.org/abs/2212.09977.

MAAS, A. L. **Rectifier Nonlinearities Improve Neural Network Acoustic Models**. 2013. Disponível em: https://api.semanticscholar.org/CorpusID:16489696.

MAINTAINERS, T.; CONTRIBUTORS. **TorchVision: PyTorch's Computer Vision library**. [S.l.]: GitHub, 2016. https://github.com/pytorch/vision.

MIRZA, M.; OSINDERO, S. Conditional generative adversarial nets. **arXiv preprint arXiv:1411.1784**, 2014.

NAGISETTY, V. et al. xai-gan: Enhancing generative adversarial networks via explainable ai systems. **arXiv preprint arXiv:2002.10438**, 2020.

ODENA, A.; OLAH, C.; SHLENS, J. Conditional image synthesis with auxiliary classifier gans. In: PMLR. **International conference on machine learning**. [S.l.], 2017. p. 2642–2651.

Organização Mundial da Saúde. **Câncer**. 2025. Acesso em: 11 out. 2025. Disponível em: https://www.who.int/news-room/fact-sheets/detail/cancer.

PANARETOS, V. M.; ZEMEL, Y. Statistical aspects of wasserstein distances. **Annual Review of Statistics and Its Application**, Annual Reviews, v. 6, n. 1, p. 405–431, mar. 2019. ISSN 2326-831X. Disponível em: http://dx.doi.org/10.1146/annurev-statistics-030718-104938.

PASZKE, A. et al. Pytorch: An imperative style, high-performance deep learning library. In: **Advances in Neural Information Processing Systems 32**. Curran Associates, Inc., 2019. p. 8024–8035. Disponível em: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in python. **Journal of machine learning research**, v. 12, n. Oct, p. 2825–2830, 2011.

RADFORD, A.; METZ, L.; CHINTALA, S. Unsupervised representation learning with deep convolutional generative adversarial networks. **arXiv preprint arXiv:1511.06434**, 2015.

RADFORD, A.; METZ, L.; CHINTALA, S. **Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks**. 2016.

RUIZ-CASADO, J. L.; MOLINA-CABELLO, M. A.; LUQUE-BAENA, R. M. Enhancing histopathological image classification performance through synthetic data generation with generative adversarial networks. **Sensors**, v. 24, n. 12, 2024. ISSN 1424-8220. Disponível em: https://www.mdpi.com/1424-8220/24/12/3777.

S, D. et al. Gan based data augmentation for enhanced tumor classification. In: **2020 4th International Conference on Computer, Communication and Signal Processing (ICCCSP)**. [S.l.: s.n.], 2020. p. 1–5.

SALIMANS, T. et al. **Improved Techniques for Training GANs**. 2016. Disponível em: https://arxiv.org/abs/1606.03498.

SELVARAJU, R. R. et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. **International Journal of Computer Vision**, Springer Science and Business Media LLC, v. 128, n. 2, p. 336–359, out. 2019. ISSN 1573-1405. Disponível em: http://dx.doi.org/10.1007/s11263-019-01228-7.

SHEN, D.; WU, G.; SUK, H.-I. Deep learning in medical image analysis. **Annual Review of Biomedical Engineering**, v. 19, n. 1, p. 221–248, 2017. Disponível em: https://doi.org/10.1146/annurev-bioeng-071516-044442.

SHRIKUMAR, A.; GREENSIDE, P.; KUNDAJE, A. **Learning Important Features Through Propagating Activation Differences**. 2019.

SIMONYAN, K.; VEDALDI, A.; ZISSERMAN, A. **Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps**. 2014.

TOUVRON, H. et al. **Training data-efficient image transformers & distillation through attention**. 2021. Disponível em: https://arxiv.org/abs/2012.12877.

WAN, Z. et al. Densenet model with radam optimization algorithm for cancer image classification. In: **2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE)**. [S.l.: s.n.], 2021. p. 771–775.

WANG, W. et al. **Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions**. 2021. Disponível em: https://arxiv.org/abs/2102.12122.

WANG, Z.; SHE, Q.; WARD, T. E. Generative adversarial networks in computer vision: A survey and taxonomy. **ACM Computing Surveys (CSUR)**, ACM New York, NY, USA, v. 54, n. 2, p. 1–38, 2021.

WIGHTMAN, R. **PyTorch Image Models**. [S.l.]: GitHub, 2019. https://github.com/rwightman/pytorch-image-models.

XUE, Y. et al. Selective synthetic augmentation with histogan for improved histopathology image classification. **Medical Image Analysis**, v. 67, p. 101816, 2021. ISSN 1361-8415. Disponível em: https://www.sciencedirect.com/science/article/pii/S1361841520301808.

XUE, Y. et al. Selective synthetic augmentation with histogan for improved histopathology image classification. **Medical Image Analysis**, Elsevier BV, v. 67, p. 101816, jan. 2021. ISSN 1361-8415. Disponível em: http://dx.doi.org/10.1016/j.media.2020.101816.

ZHU, H.; CHEN, B.; YANG, C. **Understanding Why ViT Trains Badly on Small Datasets: An Intuitive Perspective**. 2023. Disponível em: https://arxiv.org/abs/2302.03751.