UNIVERSIDADE ESTADUAL PAULISTA - UNESP

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

BIANCA LANÇONI DE OLIVEIRA GARCIA

GANS E *DIFFUSION MODELS* PARA O AUMENTO ARTIFICIAL DE *DATASETS* HISTOLÓGICOS H&E

São José do Rio Preto

2025

Bianca Lançoni de Oliveira Garcia

GANS E *DIFFUSION MODELS* PARA O AUMENTO ARTIFICIAL DE *DATASETS* HISTOLÓGICOS H&E

Trabalho de Conclusão de Curso, apresentado à Universidade Estadual Paulista (UNESP), Instituto de Biociências, Letras e Ciências Exatas, São José do Rio Preto, para obtenção do título de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Leandro Alves Neves

São José do Rio Preto
2025

G216g

Garcia, Bianca Lançoni de Oliveira
GANs e diffusion models para o aumento artificial de datasets histológicos H&E / Bianca Lançoni de Oliveira Garcia. -- São José do Rio Preto, 2025
63 p. : il., tabs.

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. 2. modelos de difusão. 3. redes adversárias generativas. 4. visão por computador. 5. histologia. I. Título.

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

BIANCA LANÇONI DE OLIVEIRA GARCIA

GANS E *DIFFUSION MODELS* PARA O AUMENTO ARTIFICIAL DE *DATASETS* HISTOLÓGICOS H&E

Trabalho de Conclusão de Curso apresentado(a) à Universidade Estadual Paulista (UNESP), Instituto de Biociências, Letras e Ciências Exatas, São José do Rio Preto, para obtenção do título de Bacharel em Ciência da Computação.

Data de defesa: 26/11/2025

BANCA EXAMINADORA

_____

Prof. Dr. Leandro Alves Neves
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

_____

Prof. Dr. Lucas Correia Ribas
UNESP – Instituto de Biociências, Letras e Ciências Exatas – Campus de São José do Rio Preto

AGRADECIMENTOS

A todos os meus familiares que me apoiaram ao longo dos anos de graduação. À minha mãe e ao meu pai, que moveriam montanhas por mim e nunca duvidaram da minha capacidade em nenhum momento, um dos meus maiores privilégios. À minha prima Mariana, minha amiga desde o primeiro segundo desta vida, que, por ser biomédica, me disse que as imagens artificiais geradas "parecem reais".

Aos meus amigos da Unesp, Gabriel, Pedro, Lucas, Júlia e Nathália, que me acompanharam em todos os momentos desde o primeiro dia e que espero ter por perto em todos os próximos.

Aos meus amigos Ana, Ana Peixe, Beatriz, Giulia, Laura, Lívia, Luan e Manuela, cuja presença resiste à distância e me lembram, há quinze anos, que amar é ter a coragem de olhar para trás.

Ao meu orientador, Prof. Dr. Leandro Alves Neves, pelo apoio e disponibilidade. A todos do Laboratório Interdisciplinar de Processamento e Análise de Imagens que contribuíram para este trabalho.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brasil, Código de Financiamento 001; e do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Bolsas nº 305386/2024-7 e nº 125460/2024-4.

RESUMO

A variabilidade morfológica dos tecidos e a complexidade dos padrões histológicos impõem desafios significativos à construção de modelos robustos de diagnóstico assistido por computador. Nesse contexto, a geração de imagens sintéticas surge como uma alternativa promissora para ampliar a diversidade e a qualidade das bases de treinamento, além de permitir análises controladas da sensibilidade dos modelos supervisionados. Neste trabalho, uma estrutura computacional voltada à geração, ao aumento e à avaliação quantitativa de imagens histológicas foi definida, investigando diferentes paradigmas de geração, incluindo GANs convencionais, variantes aprimoradas por explicabilidade (XGANs) e modelos de difusão, como DDIM e NCSN++. Nas XGANs, a estratégia foi fundamentada a partir de técnicas de *Explainable Artificial Intelligence* (XAI) incorporadas à função de perda do gerador, com o objetivo de aprimorar o realismo estrutural das imagens sintéticas. As imagens geradas foram utilizadas para expandir pequenos conjuntos de dados histológicos de tecidos colorretais, mamários e hepáticos, sendo posteriormente empregadas no treinamento de diferentes classificadores supervisionados. Foram avaliadas arquiteturas baseadas em *Convolutional Neural Networks* (CNNs) e *Vision Transformer*. No conjunto colorretal, por exemplo, a acurácia do ViT aumentou de 81,29% para 88,40% com o uso de imagens geradas pela StyleGAN3. Os modelos de difusão apresentaram os menores valores de FID, 33,36 (colorretal), 34,22 (hepático) e 40,86 (mamário), enquanto as XGANs e a StyleGAN3 obtiveram métricas KID e IS competitivas, indicando alta qualidade e diversidade na síntese. Esses resultados evidenciam a relevância da integração entre modelagem generativa e explicabilidade para aprimorar o aumento de dados e o desempenho dos classificadores em histopatologia, fornecendo subsídios importantes para o desenvolvimento e a avaliação de sistemas voltados à patologia computacional.

Palavras-Chave: redes neurais; modelos de difusão; redes adversárias generativas; visão por computador; histologia.

ABSTRACT

The morphological variability of tissues and the complexity of histological patterns represent significant challenges for the development of robust computer-aided diagnostic models. In this context, the generation of synthetic images emerges as a promising alternative to enhance the diversity and quality of training datasets while enabling controlled analyses of supervised model sensitivity. This work defines a computational framework for the generation, augmentation, and quantitative evaluation of histological images, exploring different paradigms of generative modeling, including conventional GANs, explainability-enhanced variants (XGANs), and diffusion models such as DDIM and NCSN++. In the XGAN formulation, strategies based on *Explainable Artificial Intelligence* (XAI) were incorporated into the generator loss function to improve the structural realism of synthetic images. The generated samples were used to expand small H&E datasets of colorectal, breast, and liver tissues, subsequently employed to train different supervised classifiers. Architectures based on *Convolutional Neural Networks* (CNNs) and the *Vision Transformer* (ViT) were evaluated. In the colorectal dataset, for instance, the ViT accuracy increased from 81.29% to 88.40% when trained with images generated by StyleGAN3. Diffusion models achieved the lowest FID scores, 33.36 (colorectal), 34.22 (hepatic), and 40.86 (breast), while XGANs and StyleGAN3 obtained competitive KID and IS metrics, indicating high quality and diversity in synthesis. These findings highlight the relevance of integrating generative modeling and explainability to improve data augmentation and classifier performance in histopathology, providing valuable insights for the development and evaluation of computational pathology systems.

Keywords: neural networks; diffusion models; generative adversarial networks; computer vision; histology.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A área de inteligência artificial generativa voltada à síntese de imagens evoluiu significativamente nos últimos anos, impulsionada pelo surgimento das Generative Adversarial Networks (GANs) (Goodfellow et al., 2014). Esses algoritmos baseiam-se em fundamentos da teoria dos jogos, em que duas redes neurais são treinadas de forma simultânea e competitiva: uma rede geradora, responsável por produzir amostras que se aproximem o máximo possível da distribuição real dos dados, e uma rede discriminadora, encarregada de distinguir entre amostras autênticas e sintéticas. Desde sua introdução, diversas estratégias e variações foram propostas para mitigar problemas de instabilidade e colapso de modo (Iglesias; Talavera; Díaz-Álvarez, 2023). Como resultado, surgiu uma ampla variedade de modelos generativos voltados à otimização tanto da arquitetura (Radford; Metz; Chintala, 2015; Karras et al., 2021a) quanto da função de perda (Gulrajani et al., 2017; Rozendo et al., 2024). Apesar dessas variações, as GANs consolidaram-se em múltiplas tarefas de geração de imagens (Iglesias; Talavera; Díaz-Álvarez, 2023), apresentando desempenho superior em métricas de variabilidade e qualidade aos *Variational Autoencoders* (VAEs) (Rezende; Mohamed; Wierstra, 2014), precursores das GANs nesse domínio. Contudo, esse cenário tem se transformado com o advento dos diffusion models, que se estabeleceram como uma nova e promissora classe de modelos generativos.

Apresentados pela primeira vez em Sohl-Dickstein et al. (2015), os diffusion models, também denominados Diffusion Generative Models (DGM) ou modelos de difusão, constituem uma classe de modelos generativos probabilísticos que passaram a ser amplamente explorados a partir de 2020, com os avanços introduzidos por Ho, Jain & Abbeel (2020). O princípio de funcionamento desses modelos baseia-se em verossimilhança (semelhante aos VAEs); ou seja, busca-se aproximar a distribuição dos dados reais. Esse processo ocorre em duas etapas complementares: o processo direto (*forward process*), responsável por adicionar ruído progressivamente às amostras, e o processo reverso (*reverse process*), cujo objetivo é remover esse ruído e reconstruir os dados originais. Esses modelos superam os GANs em diversos aspectos, oferecem maior estabilidade de treinamento, não sofrem de colapso de modo e têm melhor síntese (Dhariwal; Nichol, 2021). Entretanto, não são substitutos ideais para todas as tarefas: o elevado custo computacional de inferência torna sua aplicação em cenários de tempo real pouco viável. Ainda assim, os modelos de difusão têm apresentado desempenhos relevantes em diversas áreas, especialmente na geração de imagens (Ho; Jain; Abbeel, 2020; Song; Meng; Ermon, 2020).

No campo da patologia digital, em que os conjuntos de dados costumam ser limitados devido ao custo, ao tempo e à expertise necessários para anotações manuais, o potencial dos modelos generativos torna-se ainda mais relevante. A criação sintética de imagens permite equilibrar conjuntos de dados e ampliar a capacidade de generalização de modelos supervisionados treinados com amostras escassas. Diversos trabalhos têm explorado essa possibilidade, relatando

ganhos significativos de desempenho em classificadores quando empregam dados aumentados por modelos de difusão (Niehues et al., 2024; Harb; Pock; Müller, 2024; Thakkar et al., 2024). Como exemplo, Niehues et al. (2024) mostraram que os *Latent Diffusion Models* (LDMs) podem elevar em até 4% a acurácia de classificadores em bases de câncer colorretal, como o NCT-CRC-HE-100K. Além disso, esses modelos são capazes de sintetizar *whole-slide images* (WSIs) em escala gigapixel, utilizando abordagens *coarse-to-fine* que adicionam gradualmente detalhes estruturais e morfológicos (Harb; Pock; Müller, 2024). Essa flexibilidade em gerar conteúdo histológico em diferentes campos de visão (*fields of view* – FOVs) e resoluções faz dos modelos de difusão candidatos robustos à comparação com arquiteturas baseadas em GANs no contexto de aumento de dados em imagens médicas.

As abordagens híbridas expandem ainda mais esse cenário, como exemplifica a *Generative Adversarial Network with Explainable Artificial Intelligence* (XGAN). Proposta originalmente por Rozendo et al. (2024), essa arquitetura introduz mecanismos de explicabilidade diretamente no processo de geração, atuando não como ferramentas de análise posterior, mas como elementos ativos que aprimoram a percepção espacial e morfológica do gerador. Essa integração permite ao modelo direcionar o próprio processo de síntese, resultando em imagens mais coerentes e histologicamente fidedignas. Em sua formulação inicial, a XGAN foi aplicada à ampliação de conjuntos de dados de tecidos mamários, colorretais e hepáticos, alcançando reduções de até 32,70% no *Fréchet Inception Distance* (FID) e ganhos de até 3,81% em acurácia de classificadores baseados em *Transformers*.

Colaborando com os avanços da patologia digital, os modelos de classificação baseados em aprendizado profundo também têm se destacado em diferentes aplicações, empregando arquiteturas como redes neurais convolucionais (CNNs) (Huang et al., 2017), *Vision Transformers* (ViT) (Dosovitskiy et al., 2020) e modelos híbridos como o CoAtNet (Dai et al., 2024). Essas arquiteturas têm o poder de reconhecer padrões morfológicos complexos e determinar, de forma automatizada, a que classe uma imagem pertence com base nas distribuições aprendidas durante o treinamento. Ao extrair representações hierárquicas de tecidos e células, esses modelos conseguem identificar alterações sutis em amostras histológicas que, muitas vezes, passam despercebidas na avaliação humana. A relevância dessas abordagens em histopatologia é particularmente acentuada, pois, além de auxiliarem no diagnóstico por meio da detecção de classes-alvo, também oferecem suporte a tarefas de triagem, gradação de lesões e estratificação de risco. Desse modo, contribuem para a padronização de laudos, a priorização de casos e o aumento da eficiência em contextos clínicos de alta demanda e recursos limitados.

Neste contexto, entre as doenças de maior relevância para a histopatologia, câncer colorretal e câncer de mama são notórios, pois se tratam das segunda e terceira maiores causas de morte por câncer globalmente (Bray et al., 2024). O câncer colorretal compreende os tumores originados no intestino grosso, englobando tanto o cólon quanto o reto (porção terminal do trato digestivo anterior ao ânus), além de incluir, em menor frequência, neoplasias do próprio canal anal. Segundo Kim et al. (2025), estima-se um aumento de 63,30% no número de casos em vinte

anos, de 1.931.590 em 2020 para 3.154.674 em 2040, reforçando a necessidade de estratégias integradas de prevenção, diagnóstico precoce e tratamento personalizado. O câncer de mama, em particular, continua sendo o mais incidente entre as mulheres. Estima-se que, a cada minuto, quatro mulheres são diagnosticadas com a doença em escala global, e uma mulher morre a cada minuto em decorrência dela (International Agency for Research on Cancer (IARC), 2016). Em 2024, o câncer de mama foi o câncer com maior mortalidade feminina no Brasil (Instituto Nacional de Câncer (INCA), 2024). Dados recentes também indicam que, mesmo em países altamente desenvolvidos, as taxas de incidência de câncer de mama se mantêm estáveis ou em leve crescimento (Kim et al., 2025). Assim, a alta incidência e a expressiva mortalidade associadas a esses tipos de câncer reforçam a importância de desenvolver e aprimorar abordagens computacionais capazes de apoiar o diagnóstico histopatológico de forma mais precisa.

1.1 MOTIVAÇÃO E JUSTIFICATIVAS

O desenvolvimento de soluções de inteligência artificial aplicadas à medicina envolve um conjunto de desafios que vai além da obtenção de alto desempenho em métricas computacionais. Conforme discutem Chen, Liu & Peng (2019), a criação de modelos de aprendizado de máquina para a área da saúde requer uma abordagem translacional que integre rigor científico, ética e aplicabilidade clínica. O processo abrange etapas interdependentes: definição precisa do problema clínico, curadoria e padronização de dados, desenvolvimento e validação do modelo, avaliação de impacto real e implantação acompanhada de monitoramento contínuo. A eficácia técnica deve vir acompanhada de segurança, interpretabilidade e relevância clínica, princípios fundamentais para a IA contribuir de forma responsável para o avanço do diagnóstico médico. Ademais, o acesso restrito a dados sensíveis, a necessidade de garantir privacidade e consentimento informado e o risco de vieses algorítmicos que possam reproduzir desigualdades clínicas tornam a aplicação da IA particularmente delicada.

O cenário destacado previamente torna-se ainda mais desafiador no caso de imagens histológicas coradas com hematoxilina e eosina (H&E), consideradas o *padrão-ouro* para o diagnóstico de diversos tipos de câncer e outras doenças (Xu et al., 2025). Entretanto, o acesso a conjuntos de dados que contenham imagens histológicas de alta qualidade e devidamente rotuladas é reconhecidamente limitado na literatura especializada, o que impacta diretamente o desenvolvimento de modelos de aprendizado profundo realmente robustos (Xu et al., 2025). Essa limitação afeta especialmente as tarefas de classificação, uma vez que tais modelos demandam um número substancial de amostras e adequada representatividade entre as classes para alcançar desempenho satisfatório.

Diante dessas limitações e considerando a relevância do estudo de imagens histológicas H&E, observa-se uma necessidade crescente de estratégias capazes de mitigar a escassez de dados rotulados e, simultaneamente, preservar a variabilidade morfológica inerente aos tecidos biológicos. Nesse contexto, técnicas baseadas em modelos generativos, como as GANs e os Mo-

delos de Difusão, despontam como alternativas promissoras para a geração sintética de imagens histológicas. Ao produzirem amostras visuais e estatisticamente compatíveis com dados reais, esses métodos possibilitam a ampliação artificial dos conjuntos de treinamento, contribuindo potencialmente para o desenvolvimento de classificadores mais robustos e generalizáveis. Mais ainda, embora estudos recentes indiquem um desempenho superior dos modelos de difusão em relação às GANs na qualidade das imagens geradas (Dhariwal; Nichol, 2021), ainda são escassas as investigações que comparam, de forma sistemática, essas abordagens no domínio específico de tecidos histopatológicos, particularmente quanto ao impacto que cada uma pode exercer sobre a eficácia do aumento artificial de dados em tarefas de classificação.

Nesse cenário, a seguinte questão de pesquisa merece atenção: modelos GANs e de difusão apresentam desempenhos distintos na geração sintética de imagens histológicas, e em que medida tais diferenças influenciam o aumento artificial de dados e o desempenho de classificadores baseados em aprendizado profundo? A relevância dessa investigação reside no fato de que, embora os modelos generativos tenham demonstrado avanços notáveis na síntese de imagens realistas, ainda há incertezas quanto à adequação de cada paradigma ao contexto aqui explorado, envolvendo estruturas histológicas complexas e alta variabilidade morfológica. Além disso, a literatura carece de análises comparativas abrangentes que avaliem simultaneamente a qualidade das imagens geradas e o impacto prático dessas amostras no treinamento de classificadores supervisionados.

1.2 OBJETIVOS

O objetivo geral deste trabalho foi investigar e comparar o desempenho de diferentes arquiteturas de GANs e de modelos de difusão na geração de imagens histológicas coradas com H&E. Os objetivos específicos foram:

- Implementar, treinar e ajustar diferentes modelos generativos para a síntese de imagens histológicas H&E;
- Avaliar e comparar o desempenho dos modelos quanto à qualidade visual das imagens geradas, por meio de métricas quantitativas, e quanto ao impacto dessas imagens no treinamento de modelos classificadores baseados em CNN e ViT.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo são apresentados os conceitos que fundamentam o desenvolvimento deste trabalho. Na Seção 2.1, discutem-se as imagens utilizadas para validação nas etapas posteriores e sua relevância no contexto de inteligência artificial aplicada à medicina diagnóstica. Na Seção 2.2 são abordados os princípios do aumento artificial de conjuntos de dados, técnica comumente empregada para aprimorar o desempenho dos modelos de classificação descritos na Seção 2.3, os quais são avaliados por meio de métricas (Seção 2.4). Por fim, o processo de aumento é explorado a partir do uso de modelos generativos, detalhados na Seção 2.5.

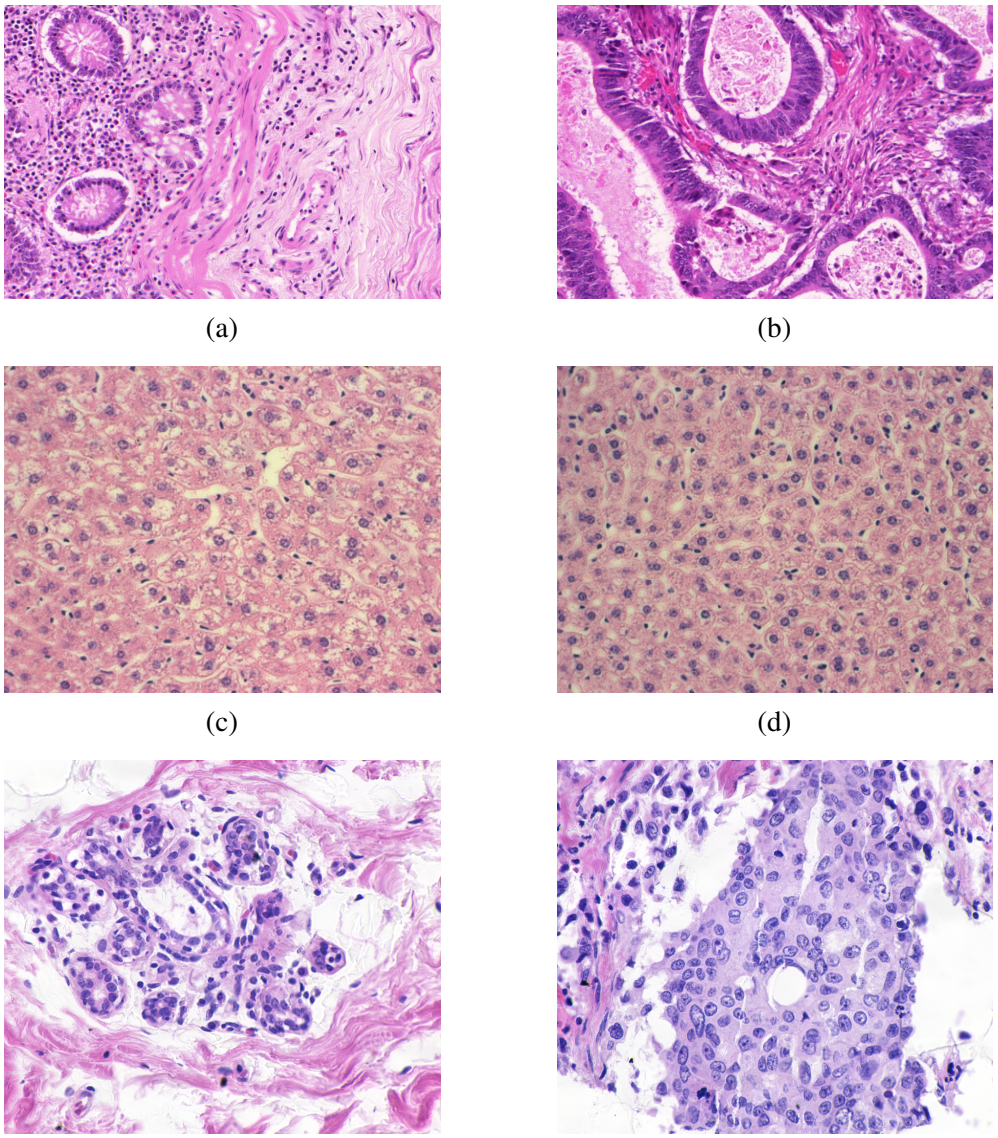
2.1 IMAGENS HISTOLÓGICAS H&E

A medicina diagnóstica tem papel fundamental na detecção precoce de doenças e na definição de estratégias terapêuticas eficazes. No contexto oncológico, o diagnóstico precoce é um dos principais determinantes da sobrevida do paciente. Segundo a Organização Mundial da Saúde (OMS) e o International Agency for Research on Cancer (IARC), o câncer de mama e o câncer colorretal estão entre os mais incidentes globalmente, representando causas significativas de morbimortalidade, especialmente em países de média e baixa renda (Roshandel; Ghasemi-Kebria; Malekzadeh, 2024; Kim et al., 2025). A histopatologia é uma etapa central da medicina diagnóstica moderna, consistindo na análise de diferentes componentes teciduais, muitas vezes provindos de crescimento celular elevado, como hiperplasia, metaplasia e displasia. O crescimento descontrolado pode levar à formação de massas teciduais anômalas, originando neoplasias (*in situ* ou invasivas), geralmente referidas como tumores. O câncer caracteriza-se, portanto, por um conjunto de doenças que compartilham o mesmo princípio biológico de proliferação celular desregulada e capacidade de invasão tecidual, sendo resultado de alterações genéticas acumuladas que conferem vantagens seletivas às células afetadas. No contexto histopatológico, tais processos se manifestam por alterações morfológicas detectáveis em lâminas que, coradas com hematoxilina e eosina, destacam, respectivamente, núcleos de azul e roxo, e citoplasma e tecido conectivo de rosa, permitindo visualização clara da morfologia dos tecidos. A produção digital dessas imagens é complexa, frequentemente apresentando diferenças na coloração (Xu et al., 2025). Além disso, laboratórios raramente documentam a concentração, lotes e temperatura da coloração. Essa ausência de registro dificulta a replicação da cor e da qualidade das imagens. Exemplos de imagens representativas desse processo encontram-se na Figura 1.

2.2 AUMENTO ARTIFICIAL DE *DATASETS*

A geração de dados sintéticos visa suprir lacunas éticas, logísticas e quantitativas associadas à coleta de amostras reais, além de melhorar a generalização dos modelos discriminativos de

Figura 1 – Exemplos de imagens histológicas coradas com hematoxilina-eosina; (a) e (b) de tecido colorretal, (c) e (d) de tecido hepático e (e) e (f) de tecido mamário.



(a) (b)

(c) (d)

(e) (f)

Fonte: Elaboração própria, com imagens de Gelasca et al. (2008), AGEMAP (2020), Sirinukunwattana et al. (2017).

aprendizado profundo. Muitas vezes esta etapa se faz necessária por certos comportamentos como *overfitting* (Shorten; Khoshgoftaar, 2019), quando um modelo de aprendizado profundo aprende a função de distribuição dos dados de treino com alta variância; ou seja, aprende a classificar apenas os dados de treino, mas não generaliza o suficiente para classificar os dados de validação ou teste. Lidar com esse problema é necessário em *datasets* limitados, comuns em cenários histopatológicos por conta das condições de produção dessas imagens.

O aumento artificial de *datasets* mitiga as consequências do *overfitting* ao lidar diretamente com o problema dos dados de treino, diferentemente de outras estratégias que tentam amenizá-lo através da arquitetura ou do método de treinamento (Shorten; Khoshgoftaar, 2019). Tradicio-

nalmente, o aumento é realizado por transformações geométricas ou de cor, que permitem ao modelo manter bom desempenho mesmo diante de variações translacionais. Entretanto, o uso de modelos generativos para criar instâncias artificiais dos dados é uma maneira de "desbloquear" informações adicionais de um *dataset* (Shorten; Khoshgoftaar, 2019), tornando o uso de GANs e modelos de difusão uma alternativa viável às técnicas tradicionais, contribuindo como mais uma métrica de comparação entre essas arquiteturas.

2.3 MODELOS DE CLASSIFICAÇÃO

A classificação de imagens é uma das tarefas mais consolidadas do aprendizado profundo, consistindo em atribuir uma categoria ou rótulo a uma imagem com base em seus padrões visuais. Essa tarefa depende fortemente da capacidade do modelo de extrair e hierarquizar características relevantes, separando representações discriminativas de alto e baixo nível. As arquiteturas de aprendizado profundo mais proeminentes nessa área incluem as Redes Neurais Convolucionais (CNNs), os *Transformers* e modelos híbridos, que combinam elementos de ambas as abordagens.

No contexto da histopatologia, a classificação automatizada de imagens é uma ferramenta essencial para auxiliar patologistas no diagnóstico, triagem e quantificação de anomalias celulares. Esses modelos permitem detectar padrões sutis de malignidade, quantificar infiltrados inflamatórios e identificar regiões de interesse com alta precisão. Especificamente neste trabalho, esses modelos são empregados em problemas de classificação binária, no qual cada imagem é atribuída a uma entre duas classes. No entanto, sua eficácia depende fortemente da disponibilidade e diversidade de dados de treinamento. Como conjuntos de dados histológicos são frequentemente limitados e heterogêneos, o desempenho dessas arquiteturas pode ser comprometido por *overfitting* e viés amostral. Nesse cenário, o aumento artificial de dados por meio de modelos generativos, como GANs e difusão, torna-se uma estratégia relevante para expandir a variabilidade do conjunto de treinamento e melhorar a robustez dos classificadores.

2.3.1 Convolutional Neural Networks (CNNs)

As CNNs, inspiradas no funcionamento do córtex visual, exploram a correlação local entre *pixels* por meio de filtros convolucionais, capturando texturas, bordas e estruturas hierárquicas da imagem. Essa propriedade de invariância espacial torna as CNNs especialmente adequadas para domínios como a histopatologia, em que padrões morfológicos e estruturais, como disposição celular, densidade nuclear e organização tecidual, são determinantes no diagnóstico. Entre algumas arquiteturas representativas, pode-se citar a ResNet e a DenseNet.

2.3.1.1 DenseNet

A *Densely Connected Convolutional Network* (DenseNet) (Huang et al., 2017) conecta cada camada a todas as camadas subsequentes em um mesmo bloco denso, promovendo a reutilização de características e o fluxo eficiente de gradientes. As saídas das camadas anteriores

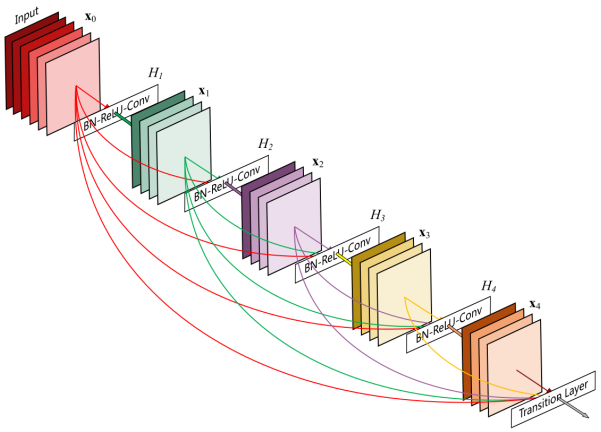
são concatenadas como entrada da próxima:

$$x_\ell = H_\ell([x_0, x_1, \ldots, x_{\ell-1}]), \tag{1}$$

em que $[x_0, x_1, \ldots, x_{\ell-1}]$ representa a concatenação de todos os mapas de características anteriores, e $H_\ell(\cdot)$ é uma composição de operações de normalização em lote, ReLU e convolução 3×3.

Cada bloco denso (Figura 2) gera um número fixo k de novos mapas de características, denominado taxa de crescimento. Assim, a dimensão de entrada da camada ℓ é $k_0 + k(\ell - 1)$, em que k_0 é o número inicial de canais. Entre blocos densos, são inseridas camadas de transição que reduzem a dimensionalidade espacial e o número de canais, compostas por uma convolução 1×1 seguida de *average pooling* 2×2. A saída do último bloco passa por *global average pooling* e uma camada totalmente conectada seguida de softmax para a predição das probabilidades de classe $y = \text{softmax}(W_c z_L)$.

Figura 2 – Esquema de um bloco denso pertencente à DenseNet.



Fonte: Huang et al. (2017).

2.3.1.2 ResNet

A Rede Residual (ResNet), proposta por He et al. (2016), introduziu a aprendizagem residual como solução para o problema de degradação em redes neurais profundas. Em vez de aprender diretamente uma transformação desejada $H(x)$, cada bloco residual (Figura 3) aprende uma função residual $F(x) = H(x) - x$, resultando na seguinte formulação:

$$y = F(x, W_i) + x, \tag{2}$$

em que x e y são, respectivamente, as entradas e saídas do bloco, e $F(x, W_i)$ representa a composição de camadas convolucionais, normalização e ativações não lineares parametrizadas pelos pesos W_i. Quando há mudança de dimensão entre x e $F(x)$, uma projeção linear W_s é

usada:

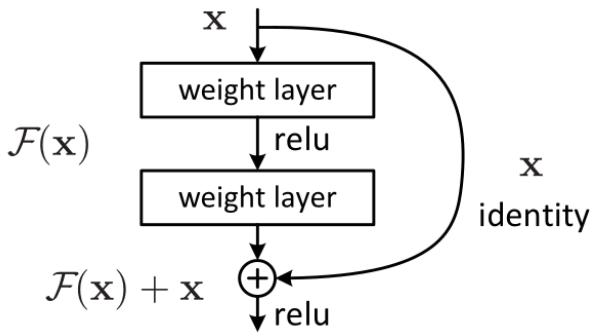
$$y = F(x, W_i) + W_s x. \tag{3}$$

O modelo completo é construído pela composição hierárquica de blocos residuais, formando uma função $f_\theta(x)$ com profundidade L:

$$f_\theta(x) = f_L(\cdots f_2(f_1(x))\cdots), \tag{4}$$

em que cada f_l é um bloco residual. Essa estrutura facilita o treinamento de redes muito profundas, pois as conexões de atalho preservam o fluxo de gradientes. Após o último bloco, aplica-se um pooling global e uma camada totalmente conectada seguida de softmax para obter as probabilidades de classe $y = \mathrm{softmax}(W_c z_L)$.

Figura 3 – Bloco de aprendizado residual.



Fonte: He et al. (2016).

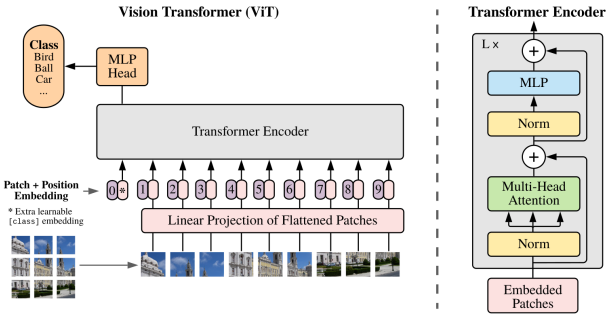
2.3.2 *Vision Transformer* (ViT)

Mais recentemente, o *Vision Transformer* (ViT), proposto por Dosovitskiy et al. (2020), adaptou o *Transformer* (originalmente desenvolvido para processamento de linguagem natural) para tarefas de classificação de imagens. Ao invés de utilizar camadas convolucionais para extrair características espaciais, o ViT divide a imagem em pequenos patches, que são então linearmente projetados em *embeddings* e processados por um *encoder Transformer* padrão. Essa abordagem permite que o modelo capture relações globais entre diferentes partes da imagem, aproveitando a capacidade do *Transformer* de modelar dependências de longo alcance, uma vantagem em relação às CNNs, que capturam apenas janelas locais. Contudo, os *Transformers* exigem grandes volumes de dados rotulados e alto custo computacional, limitando sua eficácia em, por exemplo, *datasets* limitados em domínios médicos. Uma representação desta arquitetura está ilustrada na Figura 4.

Formalmente, a imagem de entrada (W de largura, H de altura e C canais), representada por $x \in \mathbb{R}^{H\times W\times C}$, é dividida em uma grade de $N = \frac{HW}{P^2}$ patches não sobrepostos $x_p^1, x_p^2, \ldots, x_p^N$, em que cada patch $x_p^i \in \mathbb{R}^{P\times P\times C}$. Cada patch é então achatado em um vetor $x_p^i \in \mathbb{R}^{P^2\cdot C}$ e projetado linearmente em um *embedding* $z_0^i = Ex_p^i$ de tamanho D, usando uma matriz de

projeção $E \in \mathbb{R}^{(P^2C)\times D}$. Esses *embeddings* de patches são então somados a *embeddings* de posição $p^i \in \mathbb{R}^D$ para incorporar informações espaciais: $z_0^i \leftarrow z_0^i + p^i$. A sequência resultante $Z_0 = [z_0^1, z_0^2, \ldots, z_0^N]$ é então alimentada em um encoder *Transformer*, que consiste em múltiplas camadas (L) de atenção multi-cabeça (*Multi-Head Self-Attention*, MHSA) e blocos MLP, com normalização de camada (LayerNorm, LN) aplicada antes de cada bloco e conexões residuais ($\oplus$) após. A saída do token [class] z_L^0 do último layer do *Transformer* é usada como a representação da imagem para a tarefa de classificação, passando por uma cabeça de classificação (camada linear + Softmax) para produzir as probabilidades das classes $y = \text{softmax}(W_c z_L^0)$.

Figura 4 – Esquema representativo do modelo ViT e do *encoder Transformer* padrão.

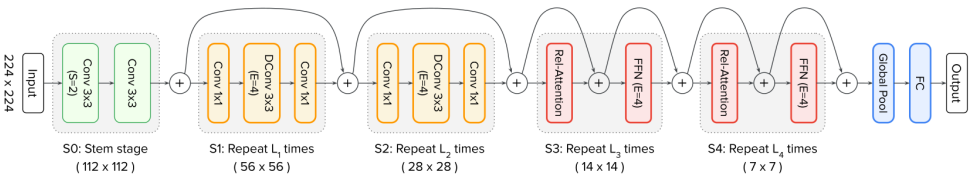


Fonte: Dosovitskiy et al. (2020).

2.3.3 CoAtNet

Visando unir as vantagens de ambos os paradigmas, surgiram modelos híbridos, como o CoAtNet (Dai et al., 2024), que integra convoluções e mecanismos de atenção. Nesses modelos, as convoluções atuam como extratoras de características locais e os blocos de atenção refinam relações não locais, resultando em melhor generalização com menor dependência de dados. O algoritmo CoAtNet combina convoluções e autoatenção em uma arquitetura híbrida para classificação de imagens, unindo a forte capacidade de generalização das convoluções com a alta capacidade de modelagem global dos *Transformers*. A arquitetura (Figura 5) é construída a partir de dois princípios fundamentais: convoluções e autoatenção podem ser unificadas via atenção relativa simples; e empilhar camadas convolucionais e camadas de atenção de forma hierárquica melhora simultaneamente a capacidade e a generalização do modelo.

Figura 5 – Esquema da arquitetura CoAtNet.



Fonte: Dai et al. (2024).

Essa rede processa uma imagem de entrada $x \in \mathbb{R}^{H\times W\times C}$ por meio de cinco estágios $(S_0, S_1, S_2, S_3, S_4)$, nos quais a resolução espacial é progressivamente reduzida e o número de

canais é aumentado. Os primeiros estágios (S_0, S_1) utilizam blocos convolucionais do tipo MBConv para capturar padrões locais, definidos por:

$$y_i = \sum_{j \in L(i)} w_{i-j} \odot x_j, \tag{5}$$

em que $L(i)$ denota uma vizinhança local (por exemplo, 3×3). Nos estágios superiores (S_2, S_3, S_4), o modelo substitui gradualmente convoluções por blocos *Transformer* com atenção relativa:

$$y_i = \sum_{j \in G} \mathrm{softmax}\left(x_i^T x_j + w_{i-j}\right) x_j, \tag{6}$$

em que G representa o conjunto global de posições espaciais e w_{i-j} introduz a equivariância translacional da convolução na atenção. Essa formulação permite ao CoAtNet capturar dependências globais sem perder as propriedades de invariância local das convoluções. A saída do último estágio é agregada via pooling global e passada por uma camada linear seguida de uma softmax, produzindo as probabilidades de classe $y = \mathrm{softmax}(W_c z)$.

2.4 MÉTRICAS DE CLASSIFICAÇÃO

Há diversas maneiras de avaliar o desempenho de modelos de classificação em *datasets* supervisionados. Para isso, podemos utilizar algumas métricas comuns, como acurácia, precisão, recall, F1-score, entre outras, a depender do contexto e dos objetivos do projeto. Por exemplo, o F1-score é uma métrica que combina a precisão e o recall em uma única medida, sendo especialmente útil em casos de classes desbalanceadas; a precisão é a proporção de verdadeiros positivos em relação ao total de positivos previstos, o que pode ser crítico em aplicações em que falsos positivos são indesejáveis; o recall, por sua vez, é a proporção de verdadeiros positivos em relação ao total de positivos reais, sendo importante em situações em que falsos negativos são problemáticos. Porém, no contexto em que se insere este trabalho, considerando a revisão de trabalhos relacionados, optou-se por utilizar a acurácia como métrica principal para avaliar o desempenho dos modelos de classificação desenvolvidos, contribuindo para efeitos de comparação com futuras publicações.

2.5 MODELOS GENERATIVOS

Modelos generativos consistem em uma classe de algoritmos de aprendizado profundo capazes de aproximar a distribuição de probabilidade de um conjunto de dados e amostrar novas instâncias sintéticas que preservem suas propriedades estatísticas e estruturais. Diferentemente dos modelos discriminativos, cujo objetivo é estimar a probabilidade condicional $p(y|x)$ (ou seja, classificar ou rotular uma amostra x), os modelos generativos buscam modelar a distribuição conjunta $p(x, y)$ ou, mais frequentemente, a distribuição marginal $p(x)$. Em termos práticos,

isso significa aprender uma função de geração $G(z)$ que mapeia vetores de ruído $z \sim p_z(z)$, amostrados de uma distribuição simples (como gaussiana), para o espaço de dados observáveis, de modo que as amostras artificiais se tornem indistinguíveis das reais.

Historicamente, os primeiros avanços relevantes nessa área surgiram com os *Variational Autoencoders* (VAEs) (Rezende; Mohamed; Wierstra, 2014), que introduziram o conceito de aprendizado de representações latentes contínuas. Os VAEs treinam uma rede codificadora e uma decodificadora conjuntamente, maximizando a evidência de probabilidade dos dados através de uma aproximação variacional do espaço latente. Entretanto, as imagens geradas por VAEs tendem a ser excessivamente suaves. As Redes Adversariais Generativas (GANs) (Goodfellow et al., 2014) substituíram a maximização explícita de probabilidade por um jogo adversarial entre dois modelos: um gerador, que produz amostras sintéticas, e um discriminador, que tenta distingui-las das reais. Essa abordagem eliminou a necessidade de definir uma função de verossimilhança explícita, permitindo gerar imagens de alta fidelidade visual. Desde então, as GANs se tornaram o paradigma dominante em síntese de imagens, com múltiplas variações projetadas para melhorar estabilidade, diversidade e qualidade (Iglesias; Talavera; Díaz-Álvarez, 2023). Mais recentemente, emergiram os modelos de difusão (Ho; Jain; Abbeel, 2020; Sohl-Dickstein et al., 2015), que reinterpretam o problema da geração como um processo iterativo de remoção de ruído, uma formulação inspirada em física estatística. Esses modelos demonstraram desempenho notável em qualidade e realismo das amostras, superando GANs em diversas tarefas de geração e restauração de imagens (Cao et al., 2024).

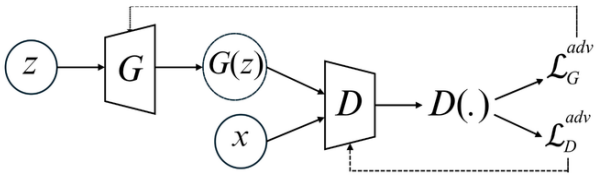
Nos contextos de imagens histológicas, essas abordagens têm papel fundamental para aumentar a diversidade e a robustez de *datasets* limitados, além de mitigar o *overfitting* em classificadores e simular tecidos realistas, mantendo padrões diagnósticos consistentes. Nas Subseções 2.5.1 e 2.5.2, apresentam-se os principais fundamentos das arquiteturas generativas empregadas neste trabalho.

2.5.1 Modelos Adversariais Generativos

Generative Adversarial Networks (GANs) são um tipo específico de arquitetura de Redes Neurais Artificiais, introduzidas pela primeira vez em Goodfellow et al. (2014). A arquitetura de uma GAN é baseada em teoria dos jogos, especificamente um jogo de minmax, consistindo em uma rede geradora e em uma rede discriminadora, cujas funções objetivo são opostas: a rede geradora G deve gerar amostras artificiais cada vez mais próximas da distribuição objetivo; e a rede discriminadora D deve aprender a diferenciar as amostras reais das amostras artificiais. A Figura 6 contém o funcionamento padrão de uma GAN.

Em teoria (Goodfellow et al., 2014), as GANs possuem uma solução única chamada de Equilíbrio de Nash, em que ambas geradora e discriminadora não são capazes de melhorar suas funções de perda. Entretanto, geralmente o treinamento converge para pontos de equilíbrio subótimos em vez de atingir o equilíbrio teórico ideal (Heusel et al., 2017). Ademais, as GANs enfrentam uma gama de problemas: colapso de modo, desaparecimento de gradiente,

Figura 6 – Esquema representativo do funcionamento de uma GAN.



Fonte: Elaboração própria.

instabilidade e problema de parada. Colapso de modo descreve a situação em que uma grande variedade de vetores latentes z geram um conjunto limitado de saídas; o que não é ideal, pois se espera que estes modelos consigam generalizar e gerem amostras com variabilidade. O desaparecimento de gradiente ocorre quando $D(x) = 1$ e $D(G(z)) = 0$, o que gera gradientes próximos de 0 que pouco contribuem para atualizações em G. A instabilidade, por conta da natureza dessas arquiteturas, descreve o impacto do aprendizado por mudanças grandes no gradiente de uma das redes. Por outro lado, o problema da parada acontece, pois, nessas arquiteturas, a função de perda não se traduz na otimização dessas redes, e muito menos no estado de qualidade das imagens geradas, o que impossibilita utilizar técnicas como *early stopping*. Esses desafios motivaram pesquisas sobre melhorias na arquitetura das GANs, que são melhor exploradas nas Subseções seguintes.

2.5.1.1 *Deep Convolutional GAN*

A *Deep Convolutional GAN* (DCGAN) (Radford; Metz; Chintala, 2015) surgiu apenas um ano após a primeira GAN, com a proposta de estabilizar e aprimorar o treinamento das GANs no escopo de geração de imagens. A função de perda seguiu sendo a mesma; as melhorias se concentraram na arquitetura. Primeiramente, as camadas totalmente conectadas (*Multi-Layer Perceptron*) foram substituídas por camadas convolucionais, por considerarem características locais. Utilizou-se normalização de *batch*, com exceção das camadas de entrada do gerador e de saída do discriminador, para estabilizar os gradientes. As funções de ativação também sofreram alterações: do discriminador é a LeakyReLU; e do gerador é a ReLU, exceto na camada de saída, que é tangente hiperbólica (Tanh). Além disso, as camadas de *max pooling* no gerador são substituídas por fractional-strided convolutions, e no discriminador por strided convolutions. Essas mudanças se provaram eficazes, sendo reaproveitadas na maioria das GANs. O algoritmo do gerador e do discriminador se encontra na Figura 7.

2.5.1.2 *Wasserstein GAN with Gradient Penalty*

A *Wasserstein GAN with Gradient Penalty* (WGAN-GP) (Gulrajani et al., 2017) é uma modificação da *Wasserstein GAN* (WGAN) (Arjovsky; Chintala; Bottou, 2017), que utiliza a distância de Wasserstein-1 para o cálculo da função de perda, com penalidade de gradiente. Nesses modelos, o discriminador agora é chamado de *critic*, pois sua função objetivo é construir

Figura 7 – Algoritmo de treinamento da DCGAN.

Algoritmo: Treinamento DCGAN
Requisitos: Conjunto de dados real $X = \{x^{(1)}, \ldots, x^{(m)}\}$, número de iterações N, taxa de aprendizado α, hiperparâmetros Adam (β_1, β_2)
Inicialização: Pesos do Gerador G e Discriminador D com $\mathcal{N}(0, 0.02)$
Para $iter = 1$ até N: Amostrar mini-batch de ruído $\{z^{(1)}, \ldots, z^{(m)}\} \sim p_z(z)$ Para $i = 1$ até m: $\tilde{x}^{(i)} \leftarrow G(z^{(i)})$ Fim Para $i = 1$ até m: $g_d^{(i)} \leftarrow \nabla_{\theta_d}[\log D(x^{(i)}) + \log(1 - D(\tilde{x}^{(i)}))]$ Fim $\theta_d \leftarrow \theta_d + \alpha \cdot \text{Adam}\left(\frac{1}{m}\sum_{i=1}^{m} g_d^{(i)}\right)$ Para $i = 1$ até m: $g_g^{(i)} \leftarrow \nabla_{\theta_g} \log D(G(z^{(i)}))$ Fim $\theta_g \leftarrow \theta_g - \alpha \cdot \text{Adam}\left(\frac{1}{m}\sum_{i=1}^{m} g_g^{(i)}\right)$ Fim
Retorno: Gerador G e Discriminador D treinados.

Fonte: Elaboração própria.

uma função que aproxima a distância entre distribuições, ao atribuir valores mais altos a amostras reais e mais baixos a amostras geradas, em vez de simplesmente classificá-las como reais ou artificiais. O objetivo dessa mudança é lidar com a instabilidade do treinamento e com o colapso de modo. A função de perda do discriminador (*critic*) é:

$$L_D = \underbrace{\mathbb{E}_{\tilde{x}\sim\mathbb{P}_g}[D(\tilde{x})] - \mathbb{E}_{x\sim\mathbb{P}_{\text{real}}}[D(x)]}_{\textit{critic loss}} + \underbrace{\lambda\mathbb{E}_{\hat{x}\sim\mathbb{P}_{\hat{x}}}\left[(|\nabla_{\hat{x}}D(\hat{x})|_2 - 1)^2\right]}_{\textit{gradient penalty}} \tag{7}$$

em que $\tilde{x}$ é equivalente a $G(z)$ em outras arquiteturas, representando a amostra artificial gerada; $\mathbb{E}_{\tilde{x}\sim\mathbb{P}_g}[D(\tilde{x})]$ é a média dos valores escalares de saída do discriminador para as amostras artificiais, e $\mathbb{E}_{x\sim\mathbb{P}_{\text{real}}}[D(x)]$ para as amostras reais.

A diferença crítica é que, ao invés de calcular a média dos logs como na GAN padrão, calcula-se diretamente a diferença entre as pontuações atribuídas às amostras falsas e reais, o que corresponde a uma estimativa da distância de Wasserstein entre as distribuições. Além disso, adiciona-se um termo de penalidade de gradiente para impor a restrição de 1-Lipschitz no *critic*, garantindo maior estabilidade no treinamento. A restrição de 1-Lipschitz (Equação 8) exige

que a função aprendida pelo *critic* $D(x)$ não varie mais do que 1 unidade para cada 1 unidade de mudança na entrada. Para funções diferenciáveis, isso é equivalente a impor que a norma do gradiente em relação à entrada seja no máximo 1 em todos os pontos (Equação 9). Essa condição é essencial porque a dualidade de Kantorovich–Rubinstein, que define a distância de Wasserstein-1, só é válida quando se maximiza sobre funções 1-Lipschitz.

$$|D(x_1) - D(x_2)| \leq |x_1 - x_2|_2 \quad \forall x_1, x_2 \tag{8}$$

$$|\nabla_x D(x)|_2 \leq 1 \quad \forall x \tag{9}$$

A WGAN original impunha essa restrição por meio de *weight clipping*, limitando os pesos da rede a um intervalo fixo, o que indiretamente força os gradientes a não crescerem demais. No entanto, isso degrada a capacidade expressiva do modelo e causa instabilidade. A WGAN-GP propôs uma alternativa mais estável: em vez de forçar $|\nabla_x D(x)|_2 \leq 1$, penaliza-se qualquer desvio da norma do gradiente em relação a 1, utilizando:

$$(|\nabla_x D(x)|_2 - 1)^2. \tag{10}$$

Na prática, calcula-se a norma do gradiente em pontos $\hat{x}$ interpolados entre amostras reais e geradas, tira-se a média desse termo ao longo do *batch*, e multiplica-se pelo hiperparâmetro λ, que controla o peso da penalidade. Essa penalização força o *critic* a satisfazer aproximadamente a condição de 1-Lipschitz, permitindo o uso correto da distância de Wasserstein e resultando em treinamento mais estável.

A função de perda do gerador continua sendo a mesma da WGAN:

$$L_G = -\mathbb{E}_{\tilde{x} \sim \mathbb{P}_g}[D(\tilde{x})], \tag{11}$$

em que o gerador produz amostras $\tilde{x} = G(z)$ a partir de vetores de ruído z, e essas amostras seguem a distribuição $\mathbb{P}_g$. O *critic* avalia cada amostra gerada por meio de $D(\tilde{x})$, atribuindo um valor escalar que indica o quão semelhante ela é a uma amostra real.

Como a função de perda tem um sinal negativo, o objetivo do gerador é maximizar a avaliação do *critic*, ou seja, produzir amostras que o *critic* considere próximas quanto possível das reais. O procedimento de treinamento do gerador e do discriminador está descrito na Figura 8.

2.5.1.3 *Relativistic average GAN*

A *Relativistic average GAN* (RaGAN) (Jolicoeur-Martineau, 2019) é uma variação da RGAN (Jolicoeur-Martineau, 2019) que modifica o discriminador para estimar a probabilidade de um conjunto de amostras reais ser, em média, mais realista que um conjunto de amostras artificiais. Essa abordagem incorpora na arquitetura o conhecimento de que o *batch* possui amostras reais e amostras artificiais, fazendo com que a função objetivo do gerador dependa de ambas as amostras.

Figura 8 – Algoritmo de treinamento da WGAN-GP.

Algoritmo: Treinamento WGAN-GP

Requisitos: Número de iterações N, passos do crítico n_{critic}, tamanho do batch m, taxa de aprendizado α, penalidade de gradiente λ

Enquanto não convergiu:
  Para $t = 1, \ldots, n_{\text{critic}}$:
    Amostrar $\{x^{(i)}\}_{i=1}^{m} \sim p_{\text{data}}$
    Amostrar $\{z^{(i)}\}_{i=1}^{m} \sim p_z$
    $\tilde{x}^{(i)} \leftarrow G(z^{(i)})$
    Amostrar $\{\epsilon^{(i)}\}_{i=1}^{m} \sim \mathcal{U}(0,1)$
    $\hat{x}^{(i)} \leftarrow \epsilon^{(i)} x^{(i)} + (1-\epsilon^{(i)})\tilde{x}^{(i)}$
    $\mathcal{L}_D \leftarrow \frac{1}{m}\sum_{i=1}^{m}\left[D(\tilde{x}^{(i)}) - D(x^{(i)}) + \lambda(\|\nabla_{\hat{x}^{(i)}} D(\hat{x}^{(i)})\|_2 - 1)^2\right]$
    Atualizar θ_d usando Adam com taxa α, $\beta_1 = 0$, $\beta_2 = 0{,}9$
  Fim
  Amostrar $\{z^{(i)}\}_{i=1}^{m} \sim p_z$
  $\mathcal{L}_G \leftarrow -\frac{1}{m}\sum_{i=1}^{m} D(G(z^{(i)}))$
  Atualizar θ_g usando Adam com taxa α, $\beta_1 = 0$, $\beta_2 = 0{,}9$

Retorno: Parâmetros treinados do gerador e do crítico.

Fonte: Adaptado de Gulrajani et al. (2017).

Para tal efeito, a saída do discriminador é definida em função das variáveis reais (Equação 12) e artificiais (Equação 13).

$$\tilde{D}(x) = \text{sigmoid}\big(C(x) - \mathbb{E}_{\tilde{x}\sim\mathbb{Q}}[C(\tilde{x})]\big) \tag{12}$$

$$\tilde{D}(\tilde{x}) = \text{sigmoid}\big(C(\tilde{x}) - \mathbb{E}_{x\sim\mathbb{P}}[C(x)]\big) \tag{13}$$

$C(x)$ representa o valor crítico (*critic score*) do discriminador para a amostra x, refletindo o quão realista o discriminador considera a amostra antes de aplicar a função sigmoide. Esse valor é utilizado para comparar cada amostra com a média das amostras da classe oposta dentro do *mini-batch*, garantindo uma abordagem relativística mais estável.

As funções de perda do gerador e do discriminador na RaGAN podem ser escritas como:

$$L_G^{\text{RaGAN}} = \mathbb{E}_{x\sim\mathbb{P}}[f_2(C(x) - \mathbb{E}_{\tilde{x}\sim\mathbb{Q}}C(\tilde{x}))] + \mathbb{E}_{\tilde{x}\sim\mathbb{Q}}[f_1(C(\tilde{x}) - \mathbb{E}_{x\sim\mathbb{P}}C(x))], \tag{14}$$

$$L_D^{\text{RaGAN}} = \mathbb{E}_{x\sim\mathbb{P}}[f_1(C(x) - \mathbb{E}_{\tilde{x}\sim\mathbb{Q}}C(\tilde{x}))] + \mathbb{E}_{\tilde{x}\sim\mathbb{Q}}[f_2(C(\tilde{x}) - \mathbb{E}_{x\sim\mathbb{P}}C(x))]. \tag{15}$$

Nesta formulação, as funções f_1 e f_2 definem o tipo de perda aplicada. Elas correspondem às funções de entropia cruzada binária, isto é, $f_1(y) = -\log(\text{sigmoid}(y))$ e $f_2(y) = -\log(1 - \text{sigmoid}(y))$. Na prática, as expectativas $\mathbb{E}_{\tilde{x}\sim\mathbb{Q}}C(\tilde{x})$ e $\mathbb{E}_{x\sim\mathbb{P}}C(x)$ são aproximadas pelas médias do *mini-batch* atual, permitindo uma implementação eficiente e estável do treinamento. O

treinamento da RaGAN pode ser observado na Figura 9.

Figura 9 – Algoritmo de treinamento da RaGAN.

Algoritmo: Treinamento da RaGAN

Requisitos: Número de iterações do discriminador n_D ($n_D = 1$ salvo se desejar treinar D até a otimalidade), tamanho do batch m, e funções f_1 e f_2 que definem a função objetivo do discriminador.

Enquanto θ não convergiu:

Para $t = 1, \ldots, n_D$:

Amostrar $\{x^{(i)}\}_{i=1}^{m} \sim \mathbb{P}$

Amostrar $\{z^{(i)}\}_{i=1}^{m} \sim \mathbb{P}_z$

Calcular $\overline{C}_w(x_r) = \dfrac{1}{m}\sum_{i=1}^{m} C_w(x^{(i)})$

Calcular $\overline{C}_w(x_f) = \dfrac{1}{m}\sum_{i=1}^{m} C_w(G_\theta(z^{(i)}))$

Atualizar w utilizando SGD ascendendo com

$$\nabla_w \frac{1}{m}\sum_{i=1}^{m}\left[f_1(C_w(x^{(i)}) - \overline{C}_w(x_f)) + f_2(C_w(G_\theta(z^{(i)})) - \overline{C}_w(x_r))\right]$$

Fim

Amostrar $\{x^{(i)}\}_{i=1}^{m} \sim \mathbb{P}$ e $\{z^{(i)}\}_{i=1}^{m} \sim \mathbb{P}_z$

Calcular $\overline{C}_w(x_r) = \dfrac{1}{m}\sum_{i=1}^{m} C_w(x^{(i)})$

Calcular $\overline{C}_w(x_f) = \dfrac{1}{m}\sum_{i=1}^{m} C_w(G_\theta(z^{(i)}))$

Atualizar θ utilizando gradientes ascendendo com

$$\nabla_\theta \frac{1}{m}\sum_{i=1}^{m}\left[f_1(C_w(G_\theta(z^{(i)})) - \overline{C}_w(x_r)) + f_2(C_w(x^{(i)}) - \overline{C}_w(x_f))\right]$$

Retorno: Parâmetros treinados de G_θ e do discriminador C_w.

Fonte: Adaptado de Jolicoeur-Martineau (2019).

2.5.1.4 XGANs

As XGANs integram mecanismos de explicabilidade (XAI) ao processo adversarial, extraindo do discriminador ou crítico mapas de importância por pixel ($E \in \mathbb{R}^{H\times W\times C}$) que indicam as regiões mais influentes na decisão. Esses mapas modulam o gradiente do gerador, enfatizando áreas relevantes para D e equilibrando o jogo adversarial, o que empiricamente resulta em maior estabilidade, qualidade e diversidade das amostras geradas.

Dadas amostras $\tilde{x} = G(z)$, os mapas E podem ser obtidos por métodos como *Saliency*, *Gradient⊙Input* e *DeepLIFT*, seguidos de normalização:

$$\hat{E} = \frac{\mathrm{ReLU}(E)}{\varepsilon + \frac{1}{HWC}\sum_{u=1}^{HWC}\mathrm{ReLU}(E)_u}, \tag{16}$$

com $\varepsilon > 0$ pequeno, garantindo $\hat{E} \geq 0$ e média unitária.

O gradiente da perda adversária padrão do gerador, $U = \partial\mathcal{L}_G^{\text{adv}}/\partial\tilde{x}$, é então modulado por $\hat{E}$

segundo:

$$\widetilde{U} = (1 - \lambda_{\text{ed}})U + \lambda_{\text{ed}}(\hat{E} \odot U), \tag{17}$$

em que $\lambda_{\text{ed}} \in [0, 1]$ controla a intensidade educacional. Essa modulação ajusta o gradiente de forma espacial sem alterar a complexidade do treinamento, resultando em uma atualização:

$$\nabla_{\theta_g}^{\text{XGAN}} = J_G^{\top} \text{vec}(\widetilde{U}), \tag{18}$$

na qual J_G é o Jacobiano do gerador em relação a seus parâmetros.

A modulação por pixel faz com que o gerador concentre sua capacidade nas regiões mais relevantes para o discriminador, reduzindo atualizações aleatórias e mitigando o colapso de modo. O custo computacional é equivalente ao de uma atualização padrão, permitindo integração direta a outras arquiteturas.

2.5.1.5 StyleGAN3

A *StyleGAN3* representa a terceira iteração da família de modelos generativos propostos por Karras et al. (2021b), e tem como principal contribuição a eliminação de *aliasing* na síntese de imagens, alcançando uma geração mais estável e consistente sob transformações espaciais. As camadas de *upsampling* e *downsampling* convencionais (empregadas em arquiteturas convolucionais como as das GANs) introduzem *aliasing* devido ao uso de operações não lineares entre convoluções e interpolações discretas. Esse artefato causa instabilidade e dependência das amostras geradas em relação ao alinhamento dos *pixels*, o que se manifesta como movimentos incoerentes ou distorções quando as imagens são submetidas a rotações ou translações.

Para mitigar esse problema, a StyleGAN3 substitui as operações discretas de reamostragem por um pipeline contínuo, baseado em filtros anti-*aliasing* do tipo FIR (*Finite Impulse Response*) aplicados antes e após cada convolução. Essa abordagem garante a chamada *equivariância translacional*, isto é, pequenas translações no espaço de entrada resultam em deslocamentos correspondentes na saída, sem alterar a estrutura da imagem. A convolução é implementada no domínio contínuo, e as funções de ativação são cuidadosamente tratadas para preservar a suavidade do sinal, evitando componentes de alta frequência.

O gerador G transformou o vetor latente z em um espaço intermediário w, de onde são obtidos os parâmetros de estilo que modulam cada bloco convolucional, por uma rede de mapeamento f_θ. O processo de geração segue:

$$w = f_\theta(z), \qquad x = G(w). \tag{19}$$

Cada camada de G utilizou convoluções moduladas (*modulated convolutions*) e normalização demoduladora, mas com o novo esquema contínuo de filtragem FIR aplicado ao *upsampling*, evitando artefatos de *aliasing*. O discriminador D segue uma estrutura espelhada, também com

filtros FIR em todas as operações de *downsampling*, assegurando consistência entre os domínios do gerador e do discriminador.

O gerador foi atualizado a cada passo de gradiente do discriminador ($n_D = 1$), e a perda total incorporou o termo de regularização R_1 aplicado sobre as amostras reais:

$$\mathcal{L}_D^{R_1} = \mathbb{E}_{x \sim p_{\text{real}}} \left[|\nabla_x D(x)|_2^2 \right], \tag{20}$$

que penaliza gradientes excessivos e melhora a estabilidade do critic. O termo adversarial seguiu a formulação não saturante:

$$\mathcal{L}_G = -\mathbb{E}_{z \sim p_z}[\log D(G(z))]. \tag{21}$$

2.5.2 Modelos de Difusão

O *Denoising Diffusion Probabilistic Model* (DDPM) (Ho; Jain; Abbeel, 2020) foi o primeiro modelo de difusão desenvolvido com base em Sohl-Dickstein et al. (2015). Seu funcionamento consiste na remoção probabilística de ruído dos dados mediante uma cadeia de Markov, em que se utiliza ruído gaussiano para simplificar o problema de amostragem de uma distribuição complexa. Isso é feito transformando a tarefa de inferência e aprendizado em uma sequência de problemas de amostragem mais simples, nos quais, a cada passo, remove-se ruído dos dados. Demonstrou-se que a procedência do ruído não necessariamente precisa ser gaussiana (Bansal et al., 2023), porém essa abordagem foi mais reproduzida em diferentes abordagens de difusão, portanto, foi a única utilizada neste trabalho. O DDPM estabeleceu precedentes para outros modelos de difusão, possuindo sempre um processo direto (*forward process*) e um processo reverso (*reverse process*).

Há outras abordagens de modelos de difusão além do DDPM(Cao et al., 2024; Yang et al., 2023): modelos generativos baseados em score (ou score-based generative models) (SGMs) e modelos de difusão com equações diferenciais estocásticas (Score SDEs). Um dos primeiros SGMs foi o Noise Conditional Score Network (NCSN), que se utiliza de equações de Langevin Monte Carlo para a amostragem. Nesse caso, busca-se calcular o score de Stein, em função dos dados de entrada, para calcular o gradiente da função de densidade de probabilidade das amostras. O treinamento e a amostragem são etapas independentes, permitindo que o score function seja calculado no treinamento, mas na amostragem a reversão de ruído utiliza-se do score de diversas maneiras, como: Langevin Monte Carlo; equações diferenciais ordinárias (ODEs) e suas combinações (Karras et al., 2022; Lu et al., 2022; Song et al., 2021; Song et al., 2021; Zhang; Chen, 2022); além das equações diferenciais estocásticas (SDEs), com a abordagem chamada Score SDE, que pode ser considerada uma categoria generalizada de modelos de difusão (Yang et al., 2023). Os exemplos mais representativos dessa classe são o NCSN++ e o DDPM++ (Song et al., 2021), nos quais, em vez de discretizar diretamente as equações de amostragem, as integram-nas numericamente em tempo contínuo, resolvendo as SDEs ou ODEs associadas para gerar amostras de forma mais estável e eficiente. O algoritmo de

NCSN++, por representar uma melhoria em relação ao NCSN e uma alternativa à técnica do DDIM, também foi implementado para experimentos neste trabalho.

2.5.2.1 *Denoising Diffusion Probabilistic Model* (DDPM)

A base dos modelos de difusão é: a distribuição em estudo, neste caso, o espaço amostral das imagens do *dataset*, é chamada de $p*$, e, sendo x_0 uma variável aleatória independente amostra de $p*$, constrói-se uma sequência de distribuições entre $p*$ e uma distribuição base q, que deve ser fácil de amostrar (nesse caso, gaussiana). Essas distribuições são criadas através do processo direto, e são marginalmente "próximas" em algum sentido apropriado.

Com essas condições, é possível treinar um modelo amostrador reverso (*reverse sampler*) para o passo t em que uma função potencialmente estocástica F_t, tal que, se $x_t \sim p_t$, então a distribuição marginal de $F_t(x_t)$ é exatamente p_{t-1}:

$$F_t(z) : z \sim p_t \equiv p_{t-1}. \tag{22}$$

Dizer que p_t e p_{t-1} são próximas significa que pensamos na sequência de distribuições como a discretização de uma função p(x, t), que evolui no tempo t = 0 até t = 1, de p_0 até p_T:

$$p(x, k\Delta t) = p_k(x), \tag{23}$$

em que $\Delta t = \frac{1}{T}$. Ou seja, tomamos o limite em tempo contínuo, agora T controla a proximidade das distribuições adjacentes, e, para garantir que p_T seja independente dos números de passos, a variância da gaussiana de cada incremento é definida como:

$$x_k = x_{k-1} + \mathcal{N}(0, \sigma^2), \tag{24}$$

então, $x_T \sim \mathcal{N}(x_0, T\sigma^2)$. Dessa forma, a variância de cada incremento deve ser escalada por $\Delta t = \frac{1}{T}$, escolhendo $\sigma = \sigma_q\sqrt{\Delta t}$, com σ_q^2 sendo a variância terminal desejada. A variância de p_t será sempre σ_q^2 independentemente de T. Ou seja, a partir de agora, t é um valor contínuo entre [0,1], fazendo o ruído adicionado ter uma variância entre 0 e $T\Delta t$ (ou seja, 1). x_t agora denota x no tempo discretizado t, então uma amostra da distribuição p_t será:

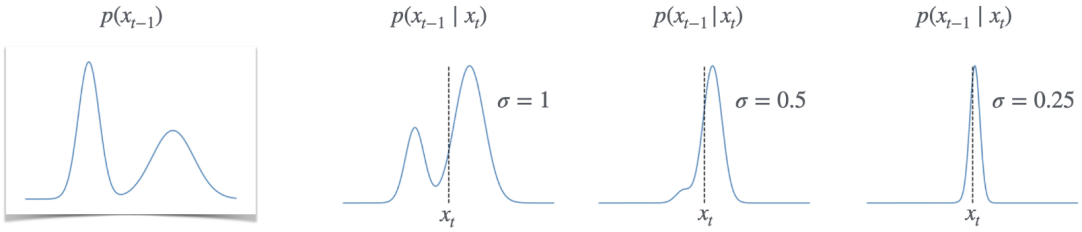
$$x_{t+\delta t} := x_t + \eta_t, \tag{25}$$

em que $\eta_t \sim \mathcal{N}(0, \sigma_q^2\Delta t)$, o que implica que $x_t \sim \mathcal{N}(x_0, (\sigma_q\sqrt{\Delta t})^2)$.

Estas são as bases teóricas do processo de difusão que regram todos os modelos desta categoria.

O modelo DDPM (*Denoising Diffusion Probabilistic Model*), por exemplo, utiliza um processo reverso estocástico para reverter a difusão gaussiana. Dado o processo direto descrito por:

Figura 10 – Grafos representativos da influência da intensidade do ruído nas distribuições gaussianas.



Fonte: Nakkiran et al. (2024)

$$x_{t+\Delta t} = x_t + \eta_t, \qquad \eta_t \sim \mathcal{N}\big(0, \sigma_q^2 \Delta t, I\big), \tag{26}$$

em que a escala $\sigma = \sigma_q \sqrt{\Delta t}$ garante que a variância terminal seja independente de T; o passo reverso ideal é amostrar de $p(x_{t-\Delta t} \mid x_t)$, que é aproximadamente gaussiano:

$$p(x_{t-\Delta t} \mid x_t = z) \approx \mathcal{N}\big(\mu_z; \sigma_q^2 \Delta t I\big), \qquad \mu_z = \mathbb{E}[x_{t-\Delta t} \mid x_t = z]. \tag{27}$$

A distribuição é inteiramente caracterizada pela média μ_z, pois a variância $\sigma_q^2 \Delta t$ é conhecida. Assim, basta aprender a função $\mu_t : \mathbb{R}^d \to \mathbb{R}^d$ que aproxima a média condicional:

$$\mu_t(z) \approx \mathbb{E}[x_t \mid x_{t+\Delta t} = z]. \tag{28}$$

Durante o treinamento, estima-se essa função por regressão, minimizando o erro quadrático médio entre a predição do modelo e o valor verdadeiro:

$$\mu_t = \arg \min_{f:\mathbb{R}^d \to \mathbb{R}^d} \mathbb{E}_{x_t, x_{t+\Delta t}} \left[|f(x_{t+\Delta t}) - x_t|_2^2 \right]. \tag{29}$$

Em prática, uma rede neural $f * \theta(x_t, t)$ é treinada sobre diferentes passos de ruído $t \sim \mathrm{Unif}[0, 1]$, compartilhando parâmetros para todos os tempos.

Durante a inferência, amostras são geradas de trás para frente, partindo de $x_1 \sim \mathcal{N}(0, \sigma_q^2 I)$ e aplicando iterativamente o amostrador reverso:

$$x_{t-\Delta t} \leftarrow \mu_{t-\Delta t}(x_t) + \mathcal{N}(0, \sigma_q^2 \Delta t, I). \tag{30}$$

Cada passo remove uma fração do ruído, aproximando gradualmente as amostras de p_0.

Uma variação importante deste algoritmo é o treinamento para prever x_0 em vez de $x_{t-\Delta t}$. Nesse caso, a relação entre as duas expectativas é:

$$\mathbb{E}[x_{t-\Delta t} \mid x_t] = \frac{\Delta t}{t} \mathbb{E}[x_0 \mid x_t] + \left(1 - \frac{\Delta t}{t}\right) x_t, \tag{31}$$

o que reduz a variância do estimador durante o aprendizado. Isso é amplamente utilizado em implementações modernas de difusão, e representa uma simplificação do funcionamento do amostrador reverso.

2.5.2.1.1 Denoising Diffusion Implicit Model (DDIM)

O DDIM (*Denoising Diffusion Implicit Model*) (Song; Meng; Ermon, 2020) é similar ao DDPM estocástico, com a diferença de que o processo reverso é *determinístico*. Considerando a distribuição conjunta $(x_0, x_{\Delta t}, \cdots, x_1)$, a média condicional $\mu_t(z) \approx \mathbb{E}[x_t \mid x_{t+\Delta t} = z]$ atua como um mapa determinístico que transporta p_t para $p_{t-\Delta t}$. Diferentemente do DDPM, não há injeção de ruído em cada passo, e a atualização segue o mapeamento mostrado na Figura 11.

O passo reverso é definido como:

$$x_{t-\Delta t} \leftarrow x_t + \lambda\big(\mu_{t-\Delta t}(x_t) - x_t\big), \tag{32}$$

em que $\lambda := \dfrac{\sigma_t}{\sigma_t + \sigma_{t-\Delta t}}$ e $\sigma_t \equiv \sigma_q\sqrt{t}$.

Figura 11 – Algoritmo de amostragem do DDIM.

Algoritmo: Amostragem DDIM
Requisitos: Modelo treinado f_θ
Dados: variância terminal σ_q, passo Δt $x_1 \leftarrow \mathcal{N}(0, \sigma_q^2 I)$ Para $t = 1, (1-\Delta t), (1-2\Delta t), \ldots, 0$: $\quad \lambda \leftarrow \dfrac{\sqrt{t}}{\sqrt{t-\Delta t} + \sqrt{t}}$ $\quad x_{t-\Delta t} \leftarrow x_t + \lambda\big(f_\theta(x_t, t) - x_t\big)$ Fim
Retorno: x_0

Fonte: Adaptado de Nakkiran et al. (2024).

2.5.2.2 *Score-Based Generative Models*

Os *Score-Based Generative Models* (SGMs) reformularam a geração de dados por difusão em termos de funções de score (*score functions*), as quais são os gradientes do logaritmo da densidade de probabilidade dos dados. Dada uma distribuição $p(x)$, o score é definido como:

$$\nabla_x \log p(x). \tag{33}$$

O gradiente aponta na direção de maior crescimento da densidade, indicando como se mover no espaço de dados para aumentar a probabilidade. Assim, aprender o score é equivalente a aprender o campo vetorial que descreve a geometria da distribuição de interesse. Para aproximar esse campo vetorial, perturba-se uma amostra $x_0 \sim p^*(x)$ com ruído gaussiano de variância

crescente σ_t^2, obtendo-se uma família de distribuições intermediárias $q(x_t)$ para $t \in [1, T]$ (similarmente ao DDPM), com:

$$q(x_t \mid x_0) = \mathcal{N}(x_t | x_0, \sigma_t^2 I). \tag{34}$$

em que x_t é uma versão ruidosa de x_0, deslocada pelo ruído gaussiano $\epsilon \sim \mathcal{N}(0, I)$:

$$x_t = x_0 + \sigma_t \epsilon. \tag{35}$$

Quando t cresce, o ruído domina e $q(x_T)$ tende a uma gaussiana isotrópica.

O treinamento de um SGM consiste em ajustar uma rede $s_\theta(x_t, t)$ para aproximar o score verdadeiro $\nabla_{x_t} \log q(x_t)$. Como não conhecemos $q(x_t)$ em forma fechada, utiliza-se o *denoising score matching*, pois a diferença entre o score verdadeiro e o estimado pode ser reescrita em termos do ruído ϵ. O objetivo de treinamento é:

$$\mathcal{L}(\theta) = \mathbb{E}t, x_0, \epsilon \Big[\lambda(t) \parallel \epsilon + \sigma_t \, s_\theta(x_t, t) \parallel_2^2 \Big], \tag{36}$$

com $x_t = x_0 + \sigma_t \epsilon$. Aqui, ϵ é o ruído gaussiano adicionado; σ_t controla a intensidade do ruído em cada escala; $s_\theta(x_t, t)$ é a predição do score pela rede neural; $\lambda(t)$ é uma função de ponderação que pode ser escolhida para estabilizar o treinamento.

Se $s_\theta(x_t, t) \approx -\epsilon/\sigma_t$, a rede aprendeu o score correto, pois

$$\nabla_{x_t} \log q(x_t \mid x_0) = -\frac{x_t - x_0}{\sigma_t^2} = -\frac{\epsilon}{\sigma_t}. \tag{37}$$

Uma vez treinado o campo de scores, é possível gerar novas amostras com uma variedade de técnicas: dinâmica de Langevin (Song; Ermon, 2019; Song et al., 2021), equações diferenciais estocásticas (Song et al., 2021), equações diferenciais ordinárias (Karras et al., 2022; Lu et al., 2022), e combinações destas. As equações diferenciais estocásticas (SDEs) reversas que removem gradualmente o ruído consistem numa abordagem chamada *Score SDE*. A forma geral de uma SDE que descreve o processo direto (ruidoso) é:

$$dx = f(x, t), dt + g(t), dw, \tag{38}$$

em que $f(x, t)$ é o termo de drift, $g(t)$ controla a intensidade do ruído, e dw é o incremento de um processo de Wiener (movimento browniano).

A teoria mostra que esse processo pode ser revertido se conhecermos os scores. O processo reverso é dado por:

$$dx = \Big[f(x, t) - g(t)^2 \nabla_x \log q_t(x) \Big] dt + g(t) d\bar{w}, \tag{39}$$

com $d\bar{w}$ um movimento browniano no tempo invertido. A substituição de $\nabla_x \log q_t(x)$ pelo

estimador $s_\theta(x, t)$ aprendido fornece um método prático de geração.

2.5.2.2.1 Noise Conditional Score Network++ (NCSN++)

Neste trabalho, foi utilizado o algoritmo *Noise Conditional Score Network* ++ (NCSN++) (Song et al., 2021), representativo dos modelos baseados em score com equações estocásticas (Score SDE), em que a amostragem funciona com um preditor-corretor. Trata-se de um SGM em que os conceitos de SGM com Score SDE servem de preditor, e uma cadeia de Markov Monte Carlo (neste caso, dinâmica de Langevin) serve de corretor; essas técnicas são combinadas para corrigir a distribuição das amostras e melhorar a qualidade das imagens geradas.

A principal diferença do NCSN++ para outros SGM é a amostragem (Figura 12). Considerando o tempo variando de t a $t - \Delta t$ (discretos), o preditor utiliza a SDE reversa e o score estimado pela rede neural $s\theta(x_t, t)$ para produzir uma estimativa inicial da amostra no instante de tempo anterior $(x_t - \Delta t)$, movendo a amostra na direção correta, mas acumulando erros de discretização ao longo do tempo. O corretor (Figura 13) refina a estimativa, corrigindo a distribuição marginal da amostra e mitigando erros do preditor (Equação 40).

$$x_i \leftarrow x_i + \epsilon_i s_{\theta^*}(x_i, \sigma_i) + \sqrt{2\epsilon_i} z \tag{40}$$

Figura 12 – Algoritmo de amostragem do Preditor–Corretor.

Algoritmo: Amostragem Preditor–Corretor
Requisitos: Modelo de score treinado s_{θ^*}, número de passos N, número de passos do corretor M
$x_N \sim \mathcal{N}(0, \sigma_{\text{max}}^2 I)$ Para $i = N - 1$ até 0: **Preditor:** $x'_i \leftarrow x_{i+1} + (\sigma_{i+1}^2 - \sigma_i^2) s_{\theta^*}(x_{i+1}, \sigma_{i+1})$ Amostrar $z \sim \mathcal{N}(0, I)$ $x_i \leftarrow x'_i + \sqrt{\sigma_{i+1}^2 - \sigma_i^2}\, z$ **Corretor:** Para $j = 1$ até M: Amostrar $z \sim \mathcal{N}(0, I)$ $x_i \leftarrow x_i + \epsilon_i s_{\theta^*}(x_i, \sigma_i) + \sqrt{2\epsilon_i}\, z$ Fim Fim
Retorno: x_0

Fonte: Adaptado de Nakkiran et al. (2024).

2.6 MÉTRICAS DE QUALIDADE

Métricas de avaliação confiáveis são um requisito importante para o contínuo desenvolvimento de melhores algoritmos de aprendizado de máquina. Em tarefas não supervisionadas (nas

Figura 13 – Algoritmo do Corretor.

Algoritmo: Corretor

Requisitos: Níveis de ruído $\{\sigma_i\}_{i=1}^{N}$, relação sinal–ruído r, passos N, passos do corretor M

$x_N^0 \sim \mathcal{N}(0, \sigma_{\text{max}}^2 I)$
Para $i = N$ até 1:
  Para $j = 1$ até M:
    Amostrar $z \sim \mathcal{N}(0, I)$
    $g \leftarrow s_{\theta^*}(x_i^{j-1}, \sigma_i)$
    $\epsilon \leftarrow 2\left(r \cdot \frac{\|z\|_2}{\|g\|_2^2}\right)$
    $x_i^j \leftarrow x_i^{j-1} + \epsilon g + \sqrt{2\epsilon} z$
  Fim
  $x_{i-1}^0 \leftarrow x_i^M$
Fim

Retorno: x_0^0

Fonte: Adaptado de Nakkiran et al. (2024).

quais não há rótulos que sirvam como referência de "correto" ou "incorreto") a avaliação da qualidade das amostras geradas depende, em grande parte, de julgamentos subjetivos. Por isso, busca-se substituir a percepção humana por critérios objetivos, baseados em distâncias matemáticas e semelhanças estatísticas entre as distribuições de imagens reais e sintéticas. Algumas métricas propostas tentam validar os modelos generativos quantitativamente, outras possuem abordagem qualitativa ao utilizar-se de estudos de usuário ou analisar o funcionamento interno dos modelos. Porém, não há concordância e trata-se de uma tarefa difícil de avaliar. As métricas para imagens sintéticas mais amplamente utilizadas nesse contexto são o *Inception Score* (IS) (Salimans et al., 2016), *Fréchet Inception Distance* (FID) (Heusel et al., 2017) e *Kernel Inception Distance* (KID) (Binkowski et al., 2018). Elas fornecem avaliações automáticas e reprodutíveis, mas também apresentam limitações conceituais.

O *Inception Score* (IS), proposto por Salimans et al. (2016), quantifica simultaneamente a qualidade e diversidade de imagens artificiais. Utiliza-se uma rede InceptionNet (Szegedy et al., 2016) pré-treinada, geralmente no *dataset* ImageNet (Deng et al., 2009), para identificar se as imagens são diversas. Sua principal limitação é não comparar imagens geradas com imagens reais, mas apenas avaliar a coerência interna do conjunto sintético. Ou seja, é possível que o IS seja alto mesmo com imagens não semelhantes às reais. Além disso, é uma métrica altamente dependente da InceptionNet: como a rede é normalmente treinada no ImageNet, seus resultados podem ser pouco representativos para outros domínios, como imagens médicas e histológicas (Borji, 2019). É sensível à resolução da imagem e é incapaz de detectar colapso de modo.

O Fréchet Inception Distance (Heusel et al., 2017) surgiu como uma evolução do IS, superando parte de suas limitações. Essa métrica compara diretamente as distribuições estatísticas de imagens reais e sintéticas. As *features* após o processamento pela InceptionNet são extraídas, e interpretadas como distribuições de características. A seguir, mede-se a distância entre essas

distribuições; ou seja, quanto menor a distância, mais próximas às distribuições real e artificial são. Essa métrica é capaz de detectar colapso de modo, mas possui outras limitações: parte-se da premissa de que as *features* seguem uma distribuição gaussiana; e é sensível ao tamanho da amostra, sendo melhor em conjuntos maiores. Apesar disso, ele é eficiente e é uma das métricas mais aceitas para avaliação.

Para mitigar os problemas do FID, o Kernel Inception Distance (Binkowski et al., 2018) foi introduzido. Ele é baseado no conceito de Maximum Mean Discrepancy (MMD) e no método de comparar distribuições extraídas de uma InceptionNet; porém, não se assume que as distribuições sejam gaussianas, tornando-o mais flexível e confiável em conjuntos de dados menores. Sua limitação é o desempenho ser dependente do *kernel* utilizado para medir a discrepância entre as distribuições. Porém, é considerado uma métrica sólida e reprodutível.

2.7 TRABALHOS RELACIONADOS

O avanço da inteligência artificial (IA) aplicada à histopatologia tem proporcionado transformações significativas na análise de imagens coradas com hematoxilina-eosina (H&E), técnica fundamental para o diagnóstico e a classificação de neoplasias em tecidos humanos e animais. A patologia digital consolidou-se como uma vertente capaz de integrar processamento de imagem e aprendizado de máquina para automatizar tarefas tradicionalmente dependentes da observação microscópica manual. Esse cenário é impulsionado pela crescente disponibilidade de bases de dados histológicas, entre as quais se destacam os conjuntos de imagens de câncer colorretal (CR), câncer de mama (UCSB) e tecido hepático de camundongos (LG), empregados no presente estudo. Esses *datasets*, por apresentarem características complementares (diferenças estruturais entre tecidos epiteliais e glandulares e variações morfológicas relacionadas ao sexo em amostras hepáticas), tornaram-se referências para o desenvolvimento e a validação de modelos de classificação e geração de imagens H&E.

Na literatura recente, diversos estudos exploraram arquiteturas de redes neurais convolucionais (CNNs) em tarefas de classificação de imagens histológicas. O trabalho de Oliveira et al. (2024) propôs um modelo híbrido baseado na combinação de deep features obtidas por transfer learning em arquiteturas AlexNet e ResNet-50, seguido por seleção de atributos com o algoritmo ReliefF e classificação por meio de um *ensemble* de cinco métodos supervisionados. Os autores aplicaram o esquema a imagens de mama, cólon e fígado e alcançaram acurácia de 99,32% no *dataset* UCSB, 98% no conjunto colorretal e 98% no conjunto LG, com número reduzido de descritores e uso de validação k-fold. O estudo indicou que camadas intermediárias da ResNet-50 são capazes de reter informações relevantes de textura e morfologia em amostras H&E, e que a combinação de deep features com classificadores de natureza distinta pode aumentar a precisão sem ampliação da arquitetura.

Pereira et al. (2026) também exploraram CNNs na análise de imagens histológicas, avaliando diferentes abordagens de extração de atributos, incluindo descritores fractais, deep features e

representações explicáveis derivadas de Grad-CAM e LIME. A classificação foi conduzida por um modelo polinomial de Hermite regularizado por LASSO, alcançando acurácia próxima de 99% em diferentes conjuntos de imagens, entre eles amostras de mama, cólon e fígado. O estudo reforçou a eficácia de métodos que associam descritores geométricos e aprendizado profundo, evidenciando que a combinação entre abordagens de natureza distinta melhora a identificação de padrões teciduais em H&E e reduz a dependência de bases extensas. Na mesma linha, Roberto et al. (2024) apresentaram um modelo de ensemble de classificadores que associa descritores fractais de textura e deep features extraídas pela ResNet-50. O método foi avaliado nos mesmos conjuntos de dados e obteve acurácia entre 88,45% e 99,77%, mesmo em bases com classes desbalanceadas e diferentes resoluções espaciais. Esses resultados indicam a importância de se utilizar múltiplas representações de imagem para capturar a variabilidade inerente a tecidos histológicos, especialmente em contextos com amostras limitadas, como no caso de bancos de imagens médicas.

Além das abordagens de classificação, a geração sintética de imagens vem se consolidando como ferramenta auxiliar no treinamento de modelos discriminativos. Os modelos generativos adversariais (GANs), introduzidos por Goodfellow et al. (2014), foram amplamente empregados para aumentar a diversidade de dados e reduzir problemas de sobreajuste. Ainda assim, limitações inerentes às GANs (como colapso de modo e sensibilidade a hiperparâmetros) motivaram a adoção de abordagens alternativas baseadas em modelos de difusão, que se mostraram mais estáveis e com capacidade superior de modelar distribuições de dados complexas. Esses modelos, introduzidos por Sohl-Dickstein et al. (2015) e consolidados por Ho, Jain & Abbeel (2020), empregam processos estocásticos reversíveis para aprender a reconstruir amostras a partir de ruído. Em aplicações médicas, estudos recentes (Cao et al., 2024) demonstraram que os Diffusion Models produzem amostras mais diversas e visualmente coerentes em comparação às GANs, além de melhor desempenho em métricas objetivas de qualidade e diversidade. Na área de deep learning aliada à histopatologia, a comparação entre modelos de difusão e GANs é pouco explorada.

A padronização de coloração e a reprodutibilidade dos modelos permanecem desafios centrais na área. Khan et al. (2025) demonstraram, em um estudo multicêntrico envolvendo 66 laboratórios e três tipos de tecidos (cólon, rim e pele), que variações de protocolo e de escaneamento produzem diferenças significativas na coloração H&E, afetando o desempenho de redes neurais. Foram comparados métodos tradicionais de normalização (Macenko, Reinhard, Vahadane e histogram matching) e técnicas baseadas em aprendizado profundo, incluindo CycleGAN e Pix2Pix. Os métodos de IA apresentaram maior consistência entre centros e contribuíram para melhorar a generalização de modelos de segmentação e classificação em tarefas histológicas. Essa constatação reforça a necessidade de abordagens que reduzam a dependência de variabilidade entre fontes de dados e assegurem a qualidade cromática das imagens.

No contexto específico do câncer de mama, a literatura recente aponta que a escassez de bases públicas amplas e bem documentadas limita a criação de modelos generalizáveis. Um

levantamento conduzido por Tafavvoghi et al. (2024) identificou apenas 17 conjuntos de lâminas inteiras (WSI) disponíveis publicamente, totalizando cerca de 10.385 imagens. A maioria dos estudos revisados utilizava um único *dataset* para treinamento e validação, sem avaliação externa, o que contribui para resultados potencialmente superestimados e pouca replicabilidade. O estudo também destacou a ausência de metadados padronizados, fator que compromete a interoperabilidade e o reuso dos dados.

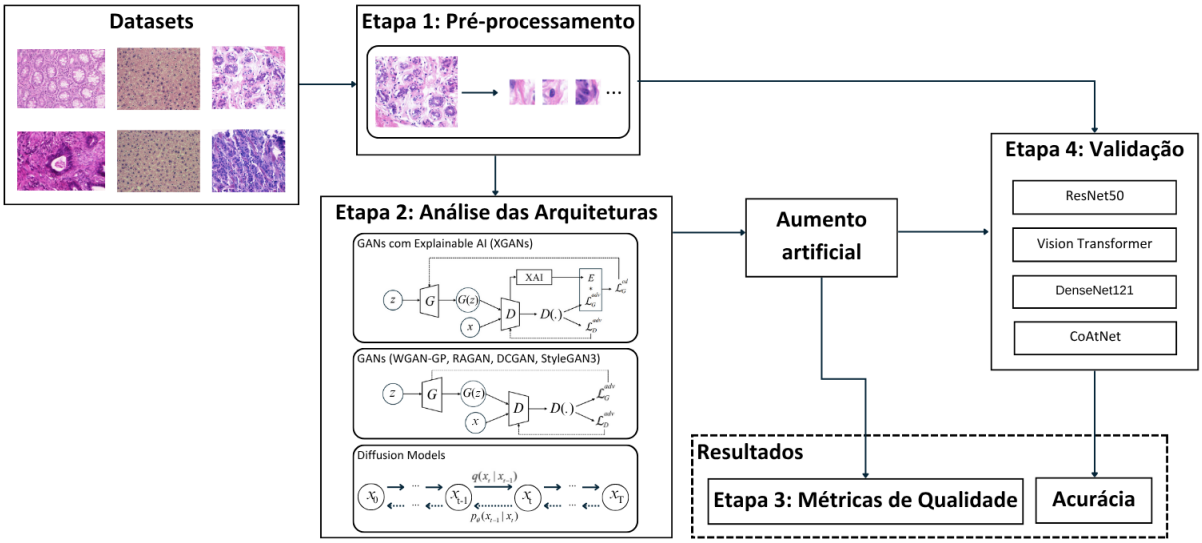
Essas lacunas motivam a busca por soluções que unam a geração sintética de dados à ampliação controlada de bases reais. O uso combinado de imagens de câncer colorretal e de mama com tecidos hepáticos de roedores, como proposto neste trabalho, permite avaliar a capacidade de generalização dos modelos sob condições morfológicas distintas. A partir disso, a integração de modelos generativos (GANs e *diffusion models*) à tarefa de aumento de dados para imagens H&E representa uma contribuição inédita em estudos que utilizam simultaneamente domínios histológicos humanos e animais. Essa integração permite não apenas aumentar o volume de dados de treinamento, mas também investigar de forma controlada o impacto da diversidade sintética na acurácia de classificadores supervisionados: CNNs, *Transformers* e arquiteturas híbridas.

Assim, a literatura indica uma tendência entre diferentes abordagens: a necessidade de aprimorar a representatividade dos dados histológicos, reduzir a dependência de amostras reais e integrar modelos generativos e discriminativos em um mesmo fluxo experimental. Nesse contexto, a presente pesquisa se insere como uma continuação lógica e necessária, ao comparar empiricamente o desempenho de GANs e Diffusion Models na geração de imagens histológicas H&E de câncer de mama, câncer colorretal e fígado de camundongos, avaliando a qualidade das amostras geradas e sua influência direta em tarefas de classificação, alinhadamente às tendências atuais da patologia computacional.

3 MATERIAIS E MÉTODOS

A abordagem computacional foi estruturada em quatro etapas distintas. Na primeira etapa, realizou-se o pré-processamento dos conjuntos de dados, padronizando os tamanhos das imagens segundo uma estratégia semelhante à de Rozendo et al. (2024), seguido de normalização e equalização de histograma. Na segunda etapa, modelos generativos representativos de diferentes arquiteturas (GANs e modelos de difusão) foram treinados para produzir aumentos artificiais dos conjuntos de dados. Na terceira etapa, aplicou-se aumento de dados à taxa de 100% do conjunto de treino para cada contexto de imagem histológica, tendo sido coletadas métricas que avaliam a qualidade e a variabilidade das imagens sintéticas geradas. Por fim, na quarta etapa, diversas arquiteturas convolucionais (He et al., 2016; Huang et al., 2017) e modelos baseados em *Transformers* (Dosovitskiy et al., 2020; Dai et al., 2024) foram empregados para classificar cada conjunto histológico, considerando as diferentes estratégias de aumento de dados. Um diagrama esquemático do método encontra-se na Figura 14, oferecendo uma visão geral das etapas e processos envolvidos. Descrições detalhadas de cada etapa são fornecidas nas seções subsequentes.

Figura 14 – Diagrama esquemático do método proposto, destacando as quatro etapas principais: pré-processamento, treinamento dos modelos generativos, aumento de dados e classificação de imagens histológicas.



Fonte: Elaboração própria.

3.1 ETAPA 1: PRÉ-PROCESSAMENTO DE IMAGENS

Os *datasets* utilizados neste trabalho são compostos por imagens histológicas coloridas com hematoxilina-eosina (H&E), obtidas de três fontes distintas: tecido colorretal (CR), tumor mamário (UCSB) e tecido hepático de camundongos (LG).

O primeiro conjunto de dados refere-se ao câncer colorretal (CR), sendo composto por imagens histológicas provenientes de 16 seções coradas com hematoxilina-eosina (H&E), correspondentes aos estágios T3 ou T4 da doença, conforme descrito por (Sirinukunwattana et al., 2017). As seções histológicas foram digitalizadas utilizando o equipamento *Zeiss MIRAX MIDI*, que oferece uma resolução de *pixel* de 0,465 μm, permitindo a obtenção de imagens de alta qualidade e detalhamento. As imagens resultantes foram classificadas em dois grupos distintos: benigno e maligno. Para a realização deste trabalho, foram selecionadas 151 imagens com dimensões de 775×522 *pixels*, sendo 67 referentes a casos benignos e 84 a casos malignos, o que possibilita uma análise comparativa entre as duas condições clínicas.

No contexto do câncer de mama, o banco de dados utilizado é constituído por 58 imagens histológicas obtidas a partir de biópsias coradas com H&E, disponibilizadas pela University of California, Santa Barbara (UCSB) (Gelasca et al., 2008). As imagens apresentam dimensões de 768×896 *pixels*, estão no padrão de cor RGB e possuem uma taxa de quantização de 24 *bits*, características que garantem riqueza de detalhes e fidelidade cromática. Este conjunto de dados é amplamente utilizado em pesquisas de patologia digital, sendo reconhecido pela qualidade das imagens e pela representatividade dos casos incluídos.

Adicionalmente, foi empregada a base liver gender (LG), fornecida pelo Atlas of Gene Expression in Mouse Ageing Project (AGEMAP) (AGEMAP, 2020). Esta base é composta por imagens representativas de tecido hepático de camundongos, separadas de acordo com o gênero dos animais, ou seja, macho e fêmea. As imagens possuem dimensões de 417×312 *pixels* e totalizam 265 exemplos, sendo 150 referentes ao gênero masculino e 115 ao gênero feminino. A distinção entre as classes permite a investigação de possíveis diferenças morfológicas associadas ao gênero, além de ampliar a diversidade dos tipos de dados utilizados para o treinamento e avaliação dos modelos generativos.

Tabela 1 – Resumo dos *datasets* utilizados no estudo.

Dataset	**Qtd. de Imagens**	**Classes**
CR (Sirinukunwattana et al., 2017)	151	Benigno, Maligno
UCSB (Gelasca et al., 2008)	58	Benigno, Maligno
LG (AGEMAP, 2020)	265	Macho, Fêmea

Fonte: Elaboração própria.

Aplicou-se normalização de intensidade às imagens para mitigar variações de iluminação, saturação e contraste decorrentes de diferenças nos protocolos de coloração H&E, que podem comprometer a estabilidade estatística e a reprodutibilidade dos modelos generativos (Macenko et al., 2009). No contexto de GANs e modelos de difusão, essa padronização inicial é essencial, por estabilizar o treinamento do discriminador e permite que o gerador aprenda representações morfologicamente coerentes. A normalização foi aplicada como:

$$I_n(x, y) = \frac{I'(x, y) - \mu_{I'}}{\sigma_{I'}}, \tag{1}$$

em que $\mu_{I'}$ e $\sigma_{I'}$ correspondem, respectivamente, à média e ao desvio padrão das intensidades normalizadas de cada canal.

Essa operação assegurou que as distribuições de cor e textura permanecessem em uma faixa estatística consistente. Em seguida, aplicou-se a equalização de histograma às amostras, expressa como:

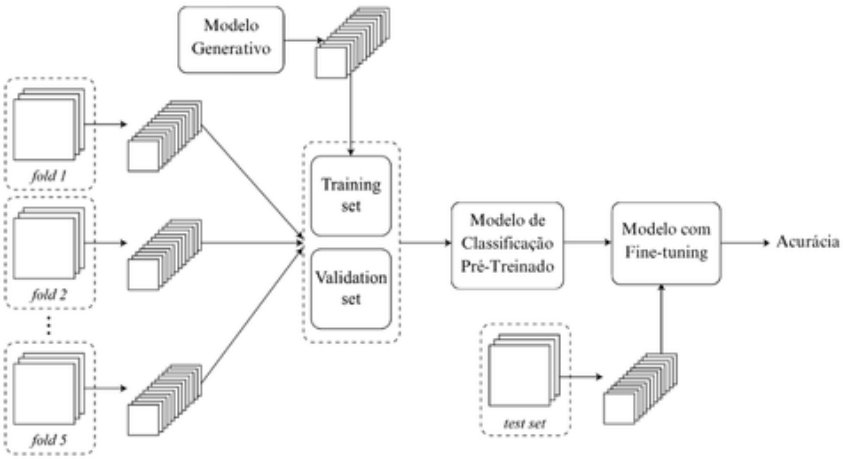
$$I_e(x,y) = \left\lfloor (L-1)\,\frac{\sum_{k=0}^{I'(x,y)} h(k)}{N} \right\rfloor \tag{2}$$

em que $h(k)$ denota o histograma do canal, $\sum_{k=0}^{I'(x,y)} h(k)$ representa a distribuição acumulada de intensidades, N é o número total de *pixels*, e L é o número de níveis de intensidade.

Essa técnica ajusta os níveis de intensidade rumo a uma distribuição mais uniforme, evidenciando detalhes que poderiam estar ocultos devido ao baixo contraste.

As imagens foram organizadas em grupos de treino, validação e teste nas proporções de, respectivamente, 80%, 10% e 10%, garantindo que os modelos foram treinados em conjuntos representativos e avaliados em dados de teste independentes. As imagens foram divididas em imagens menores de dimensão 64×64 *pixels*, sem sobreposição e evitando duplicidade de pedaços oriundos da mesma imagem original em conjuntos diferentes. Além disso, foi empregada a técnica de *k-fold repetition*, na qual os conjuntos foram subdivididos em 5 *folds* (Figura 15). Em cada experimento, um *fold* foi utilizado para validação enquanto os restantes 4 foram utilizados para treinamento, assegurando que cada amostra fosse testada sem contaminação cruzada.

Figura 15 – Divisão dos *folds* e conjuntos de treino, validação e teste para os experimentos.



Fonte: Adaptado de Rozendo et al. (2024).

Na Tabela 2, há a distribuição final de imagens de 64×64 *pixels* para cada classe nos três *datasets*. Esta padronização permitiu comparações consistentes entre os diferentes conjuntos de

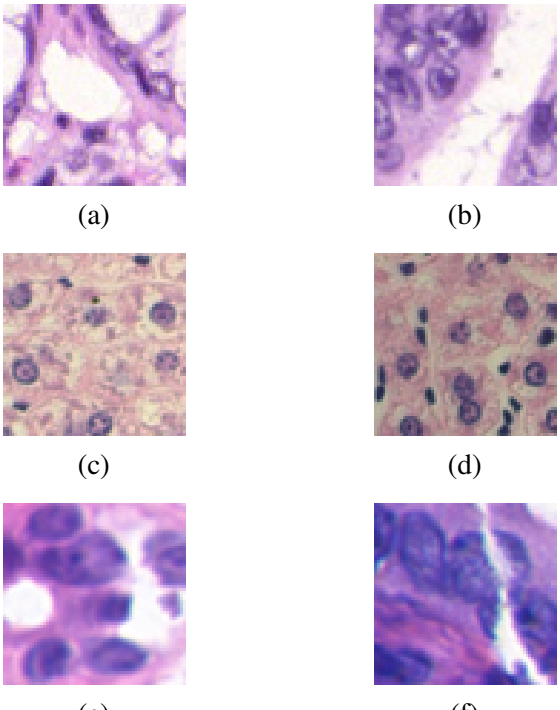
dados, garantindo que as análises quantitativas e qualitativas subsequentes refletissem diferenças metodológicas reais entre os modelos. Exemplos dessas imagens encontram-se na Figura 16.

Tabela 2 – Distribuição de imagens de dimensão 64×64 *pixels* em cada classe dos *datasets*, em que Classe 1 representa benigna (CR e UCSB) ou macho (LG) e Classe 2 representa maligna (CR e UCSB) ou fêmea (LG).

Dataset	**Classe 1**	**Classe 2**
CR	4664	5736
UCSB	4608	3744
LG	2400	1840

Fonte: Elaboração própria.

Figura 16 – Exemplos de imagens histológicas H&E, com corte de 64×64; (a) e (b) de tecido colorretal, (c) e (d) de tecido hepático e (e) e (f) de tecido mamário.



(a) (b)

(c) (d)

(e) (f)

Fonte: Elaboração própria, com imagens de Gelasca et al. (2008), AGEMAP (2020), Sirinukunwattana et al. (2017).

3.2 ETAPA 2: ESPECIFICAÇÃO E CONFIGURAÇÃO DAS ARQUITETURAS

As arquiteturas exploradas neste estudo foram selecionadas com base em trabalhos prévios que reportam resultados relevantes tanto para modelos generativos quanto para classificadores (Rozendo et al., 2024). A análise abrangeu paradigmas distintos de aprendizagem profunda, incluindo modelos guiados por ruído, formulações adversariais e arquiteturas híbridas que

combinam operações convolucionais com mecanismos de atenção. Essa diversidade arquitetural possibilitou um exame abrangente do potencial de cada modelo em termos de capacidade de generalização e interpretabilidade de representações internas, aspectos cruciais para aplicações em imagens histológicas que demandam coerência morfológica e explicabilidade visual.

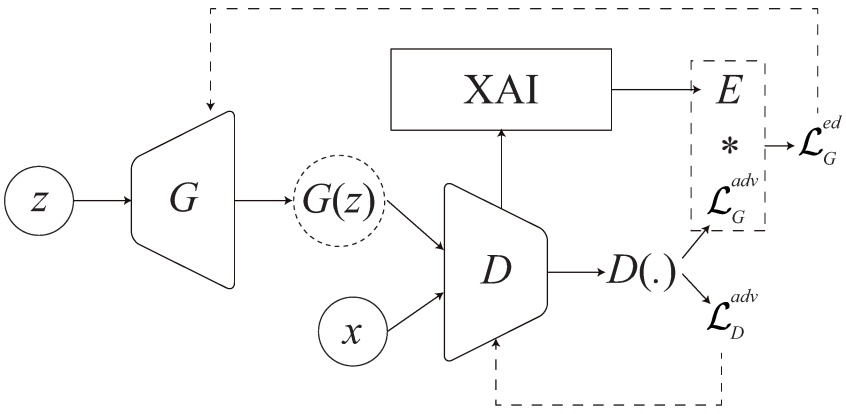
3.2.1 Modelos Adversariais Generativos

Modelos Adversariais Generativos constituem uma classe de modelos generativos treinados via estratégia adversarial, concebida como um jogo de soma zero entre duas redes complementares: o gerador G e o discriminador D. O gerador G recebe como entrada um vetor de ruído $z \sim p_z(z)$, em que p_z tipicamente corresponde a uma distribuição simples, como Gaussiana ou uniforme, e produz uma amostra sintética $G(z)$. O discriminador D, por sua vez, avalia uma amostra, real ou gerada, e retorna um valor escalar $D(x)$, interpretado como a probabilidade de que x pertença à distribuição real p^*. A função objetivo clássica que orienta o treinamento é:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p^*}[\log D(x) + \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))]. \tag{3}$$

Apesar do uso disseminado em visão computacional, GANs são notoriamente instáveis durante o treinamento, em parte porque o gerador recebe do discriminador apenas um valor escalar como gradiente. Para fornecer uma orientação mais informativa ao gerador, o modelo desenvolvido aqui baseou-se na abordagem proposta em Rozendo et al. (2024), que integrou técnicas de Inteligência Artificial Explicável (XAI) ao processo adversarial (Figura 17). Três algoritmos modificados dessa forma são destacados: XRaGAN, XWGAN-GP e XDCGAN, correspondentes a RaGAN (Jolicoeur-Martineau, 2019), WGAN-GP (Gulrajani et al., 2017) e DCGAN (Radford; Metz; Chintala, 2015), respectivamente, acrescidos de XAI.

Figura 17 – Esquema do modelo XGAN.



Fonte: Adaptado de Rozendo et al. (2024).

No fluxo de trabalho XGAN, um vetor de ruído z é mapeado pelo gerador G para uma imagem sintética $G(z)$, pontuada pelo discriminador D juntamente com imagens reais x. A decisão do

discriminador é então processada por métodos de XAI para produzir uma explicação ao nível de *pixel* E que modula o objetivo do gerador, fechando o ciclo $z \to G(z) \to D(\cdot) \to E \to L_G^{ed}$.

Em contraste com as GANs padrão, nas quais o gerador depende exclusivamente da saída escalar de $D(\cdot)$, XGANs exploram E para enriquecer o sinal de gradiente. A função de perda do gerador, denominada treinamento educacional, é:

$$L_G^{ed} = L_G^{adv} \cdot E, \tag{4}$$

em que L_G^{adv} corresponde à perda adversarial tradicional do gerador, e E representa as explicações de XAI extraídas do discriminador, obtidas por métodos como *Saliency* (Simonyan; Vedaldi; Zisserman, 2014), *Gradient⊙Input* ou *DeepLIFT* (Shrikumar; Greenside; Kundaje, 2017).

A formulação de L_G^{ed} pode ser aplicada a diferentes arquiteturas adversariais. No XDCGAN, a perda adversarial de referência L_G^{adv} baseou-se na entropia cruzada binária. No XWGAN-GP, L_G^{adv} foi definida pela distância de Wasserstein, acrescida de uma penalização de gradiente para impor a restrição 1-Lipschitz. No XRaGAN, preserva-se a natureza relativística da perda adversarial, mas ela também foi multiplicada pelo fator E. Assim, XGANs mantêm as propriedades de suas arquiteturas originais enquanto incorporam o termo multiplicativo E para guiar o treinamento com informação adicional proveniente do discriminador.

O termo de explicação E é obtido computando-se a relevância de cada *pixel* i na saída do discriminador $D(x)$ em relação à imagem de entrada x. Três métodos de atribuição foram empregados. No método *Saliency*, a relevância corresponde ao valor absoluto do gradiente da saída do discriminador em relação à entrada, definido como

$$E_i^{\text{Sal}} = \left| \frac{\partial D(x)}{\partial x_i} \right|, \tag{5}$$

que mede a sensibilidade de $D(x)$ a pequenas perturbações no *pixel* x_i.

O método *Gradient⊙Input* multiplica a ativação de entrada pelo seu gradiente correspondente, enfatizando regiões em que tanto o valor do *pixel* quanto sua influência na saída são elevados:

$$E_i^{\text{Grad}\odot\text{Inp}} = x_i \cdot \frac{\partial D(x)}{\partial x_i}. \tag{6}$$

Por fim, o *DeepLIFT* atribui relevância comparando ativações neuronais a uma referência $\bar{x}$, computando a contribuição de cada *pixel* como:

$$E_i^{\text{DL}} = (x_i - \bar{x}_i) \cdot \frac{\Delta D(x)}{\Delta x_i}, \tag{7}$$

em que $\Delta D(x)$ e Δx_i representam as diferenças entre as ativações atuais e as da entrada de referência. Esses mapas de explicação são posteriormente normalizados e usados como fator de realimentação multiplicativo E na perda educacional L_G^{ed}, guiando o gerador para regiões que o

discriminador considera mais relevantes.

Outra arquitetura de GAN considerada neste estudo pertence à família *StyleGAN* (Karras et al., 2021b). Um desafio central abordado no *StyleGAN3* é o problema de *aliasing*, que decorre do uso de filtros de *upsampling* não ideais e da aplicação local de não linearidades (por exemplo, ReLU), potencialmente introduzindo frequências que não podem ser representadas na grade discreta e resultando em artefatos visuais. Para mitigar esse problema, o *StyleGAN3* trata todos os sinais como funções contínuas limitadas em banda, amostradas em grades discretas, garantindo equivariância por translação e mitigando artefatos de *aliasing*.

3.2.2 Modelos de Difusão

O *Denoising Diffusion Probabilistic Model* (DDPM) (Sohl-Dickstein et al., 2015; Ho; Jain; Abbeel, 2020) é um modelo generativo composto por dois processos: um processo direto fixo que adiciona ruído Gaussiano incrementalmente aos dados, e um processo reverso aprendido que reconstrói os dados a partir de ruído puro de volta à distribuição dos dados. Cada processo evolui ao longo de um número finito de passos discretos de tempo $t \in \{1, \ldots, T\}$, que podem ser vistos como a discretização de uma variável temporal contínua em formulações subsequentes. O processo direto é definido como

$$q(x_t|x_{t-1}) = \mathcal{N}\Big(\sqrt{1-\beta_t}\, x_{t-1},\, \beta_t I\Big)\,, \tag{8}$$

em que β_t é um agendamento de variância fixo. A distribuição marginal correspondente permite amostragem direta em qualquer passo:

$$q(x_t|x_0) = \mathcal{N}\big(\sqrt{\bar{\alpha}_t}\, x_0,\, (1-\bar{\alpha}_t)I\big)\,, \tag{9}$$

com $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$. O processo reverso $p_\theta(x_{t-1}|x_t)$ é modelado como uma Gaussiana cuja média $\mu_\theta(x_t, t)$ é parametrizada para predizer o componente de ruído $\epsilon_\theta(x_t, t)$ em vez da própria média, simplificando o treinamento a um objetivo de remoção de ruído:

$$L_{\text{simple}}(\theta) = \mathbb{E}_{t,x_0,\epsilon}\left[\left\|\epsilon - \epsilon_\theta\big(\sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, t\big)\right\|^2\right]. \tag{10}$$

A partir dessa formulação discreta, o *Denoising Diffusion Implicit Model* (DDIM) (Song; Meng; Ermon, 2020) generaliza o processo reverso ao introduzir uma transformação determinística ou parcialmente estocástica que conecta x_t e x_{t-1} por meio de uma formulação análoga a uma equação diferencial ordinária. Essa modificação acelera a amostragem e melhora o controle sobre a variabilidade das imagens. A regra de atualização do DDIM é dada por

$$x_{t-1} = \sqrt{\alpha_{t-1}}\left(\frac{x_t - \sqrt{1-\alpha_t}\,\epsilon_\theta(x_t, t)}{\sqrt{\alpha_t}}\right) + \sqrt{1-\alpha_{t-1}}\,\epsilon_\theta(x_t, t), \tag{11}$$

em que $\epsilon_\theta(x_t, t)$ representa o ruído previsto e $\{\alpha_t\}$ são coeficientes do agendamento de ruído que controlam a magnitude da perturbação em cada passo.

Complementarmente, também foi explorado o *Noise Conditional Score Network++* (NCSN++) (Song et al., 2021), um modelo generativo de ponta baseado em *scores*, formulado no arcabouço unificado de equações diferenciais estocásticas (SDE) dos modelos de difusão. Na configuração *variance exploding* (VE), a geração de amostras foi realizada por amostradores preditor–corretor (PC), em que a etapa corretiva, baseada em *Langevin Monte Carlo* (MCMC), refina iterativamente cada amostra como:

$$x_i \leftarrow x_i + \epsilon_i \, s_\theta(x_i, \sigma_i) + \sqrt{2\epsilon_i}\, z, \quad z \sim \mathcal{N}(0, I), \tag{12}$$

em que s_θ denota a função de *score* aprendida e σ_i é a escala de ruído.

Em ambas as formulações baseadas em difusão, a rede neural parametrizada foi implementada utilizando uma arquitetura U-Net (Ronneberger; Fischer; Brox, 2015), amplamente adotada nestes modelos devido à sua eficácia em preservar a informação espacial em alta resolução. Estes modelos foram então treinados em *patches* H&E para sintetizar amostras histopatológicas realistas com fins de aumento de dados.

3.3 ETAPA 3: MÉTRICAS DE QUALIDADE PARA AVALIAÇÃO DE MODELOS GENERATIVOS

A avaliação quantitativa de modelos generativos requer métricas capazes de capturar aspectos distintos das amostras geradas, incluindo qualidade perceptual, diversidade e similaridade estatística em relação aos dados reais. Nesta investigação, três métricas amplamente adotadas foram consideradas: o Inception Score (IS) (Salimans et al., 2016), o Fréchet Inception Distance (FID) (Heusel et al., 2017) e o Kernel Inception Distance (KID) (Binkowski et al., 2018).

O *Inception Score* (IS) foi originalmente proposto para avaliar simultaneamente a qualidade e a diversidade das amostras geradas sem exigir acesso direto aos dados reais. A ideia subjacente é que imagens de alta qualidade produzem distribuições condicionais $p(y \mid x)$ de baixa entropia, refletindo previsões confiantes, ao passo que um conjunto diverso de imagens produz uma distribuição marginal $p(y)$ com alta entropia. A métrica foi formalmente definida como:

$$IS = \exp\left(\mathbb{E}_{x \sim p_g}\Big[D_{\mathrm{KL}}\big(p(y \mid x) \,\|\, p(y)\big) \Big] \right), \tag{13}$$

em que $p(y)$ denota a distribuição marginal sobre todos os rótulos previstos. Valores mais altos de IS indicam melhor desempenho generativo, capturando conjuntamente a fidelidade das amostras individuais e a diversidade global do conjunto gerado.

Outra métrica avaliada foi o Fréchet Inception Distance (FID), que quantifica a diferença entre as distribuições de características de imagens reais e geradas modelando ambas como

gaussianas multivariadas. Sejam (μ_r, Σ_r) e (μ_g, Σ_g) as médias e covariâncias das características extraídas de imagens reais e sintéticas, respectivamente. O FID foi definido como:

$$FID = \|\mu_r - \mu_g\|_2^2 + \mathrm{Tr}\Big(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}\Big), \tag{14}$$

em que $\|\cdot\|_2$ denota a norma euclidiana e $\mathrm{Tr}(\cdot)$ o traço de matriz. Valores menores de FID indicam maior similaridade estatística entre os conjuntos de características, refletindo tanto a qualidade quanto a diversidade das amostras geradas.

Por fim, o Kernel Inception Distance (KID) também foi empregado como abordagem não paramétrica. O KID baseia-se na *Maximum Mean Discrepancy* (MMD) com kernel polinomial. Seja $\phi(x)$ o vetor de características extraído pela rede Inception v3 para uma imagem x. O KID é definido como:

$$\begin{aligned} KID &= \mathrm{MMD}^2(X, Y), \\ \mathrm{MMD}^2(X, Y) &= \frac{1}{m(m-1)} \sum_{i \neq j} k(x_i, x_j) + \frac{1}{n(n-1)} \sum_{i \neq j} k(y_i, y_j) - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k(x_i, y_j), \end{aligned} \tag{15}$$

em que $X = \{\phi(x_r)\}_{r=1}^{m}$ representa o conjunto de amostras reais, $Y = \{\phi(x_g)\}_{g=1}^{n}$ o conjunto de amostras geradas, e $k(\cdot, \cdot)$ o kernel polinomial empregado.

O KID é capaz de fornecer uma estimativa não enviesada da distância entre distribuições mesmo para conjuntos de dados pequenos e elimina a necessidade de calcular a raiz quadrada de matriz exigida no FID, reduzindo, assim, a instabilidade numérica. Em conjunto, IS, FID e KID fornecem uma avaliação quantitativa abrangente da qualidade generativa, fidelidade e diversidade de imagens histopatológicas sintéticas.

3.4 ETAPA 4: TREINAMENTO E AVALIAÇÃO DOS MODELOS

Para a classificação de imagens, consideraram-se arquiteturas tanto convolucionais quanto baseadas em *Transformers*, a fim de capturar representações locais e globais em imagens histológicas. Nesse procedimento, pesos pré-treinados foram empregados nas camadas iniciais, enquanto as camadas finais foram ajustadas finamente para aprender características específicas do domínio. Os pesos pré-treinados foram obtidos das bibliotecas *TorchVision* (TorchVision maintainers and contributors, 2016) e *Timm* (Wightman, 2019), que disponibilizam modelos previamente treinados em conjuntos de grande escala, como o ImageNet, oferecendo uma representação inicial robusta de características visuais.

Os modelos baseados em CNN selecionados foram a ResNet50 (He et al., 2016) e a DenseNet121 (Huang et al., 2017). A ResNet50 introduziu conexões residuais de atalho para facilitar a propagação do gradiente e estabilizar o processo de treinamento. A DenseNet121 utilizou blocos densos, permitindo que cada camada recebesse entradas de todas as camadas preceden-

tes, ampliando a reutilização de características e melhorando a generalização com um número reduzido de parâmetros. Modelos baseados em *Transformers* foram incluídos para investigar mecanismos de atenção capazes de modelar dependências de longo alcance. O *Vision Transformer* (ViT) (Dosovitskiy et al., 2020) dividiu as imagens em *patches* processados por blocos de autoatenção, capturando de forma eficaz relações contextuais ao longo das regiões espaciais. O CoAtNet (Dai et al., 2024) integrou módulos convolucionais e de atenção em uma estrutura híbrida, explorando convoluções para extrair padrões locais e blocos de atenção para modelar dependências globais, viabilizando representações hierárquicas ricas e adequadas à classificação de imagens histopatológicas.

3.4.1 Métrica de Avaliação e Classificação

Para quantificar o desempenho dos modelos, considerou-se a acurácia média obtida nas fases de teste de cada *fold* da validação cruzada. A acurácia foi definida como:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{16}$$

em que TP, TN, FP e FN denotam verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos, respectivamente. Reportamos a média ao longo dos 5 *folds* estratificados para refletir a variabilidade decorrente da partição em *folds*.

3.5 AMBIENTE DE DESENVOLVIMENTO E *SOFTWARE*

Todos os experimentos foram implementados em Python 3.10 utilizando o *PyTorch* [1] 2.1 com *TorchVision* 0.17 (TorchVision maintainers and contributors, 2016) e *Timm* (Wightman, 2019). O ambiente de execução foi configurado em sistema Ubuntu 22.04 LTS, com acelerador gráfico NVIDIA RTX 4060 e suporte a CUDA 12.6.

O treinamento dos modelos foi realizado com *mixed-precision optimization* (*float16*), a fim de reduzir o consumo de memória e acelerar o processo de aprendizado sem comprometer a estabilidade numérica. Todo o código foi modularizado para permitir a replicação dos experimentos, padronizando os mesmos hiperparâmetros de otimização e inicialização entre as diferentes arquiteturas avaliadas.

[1] PyTorch. Disponível em: https://pytorch.org/.

4 RESULTADOS E DISCUSSÃO

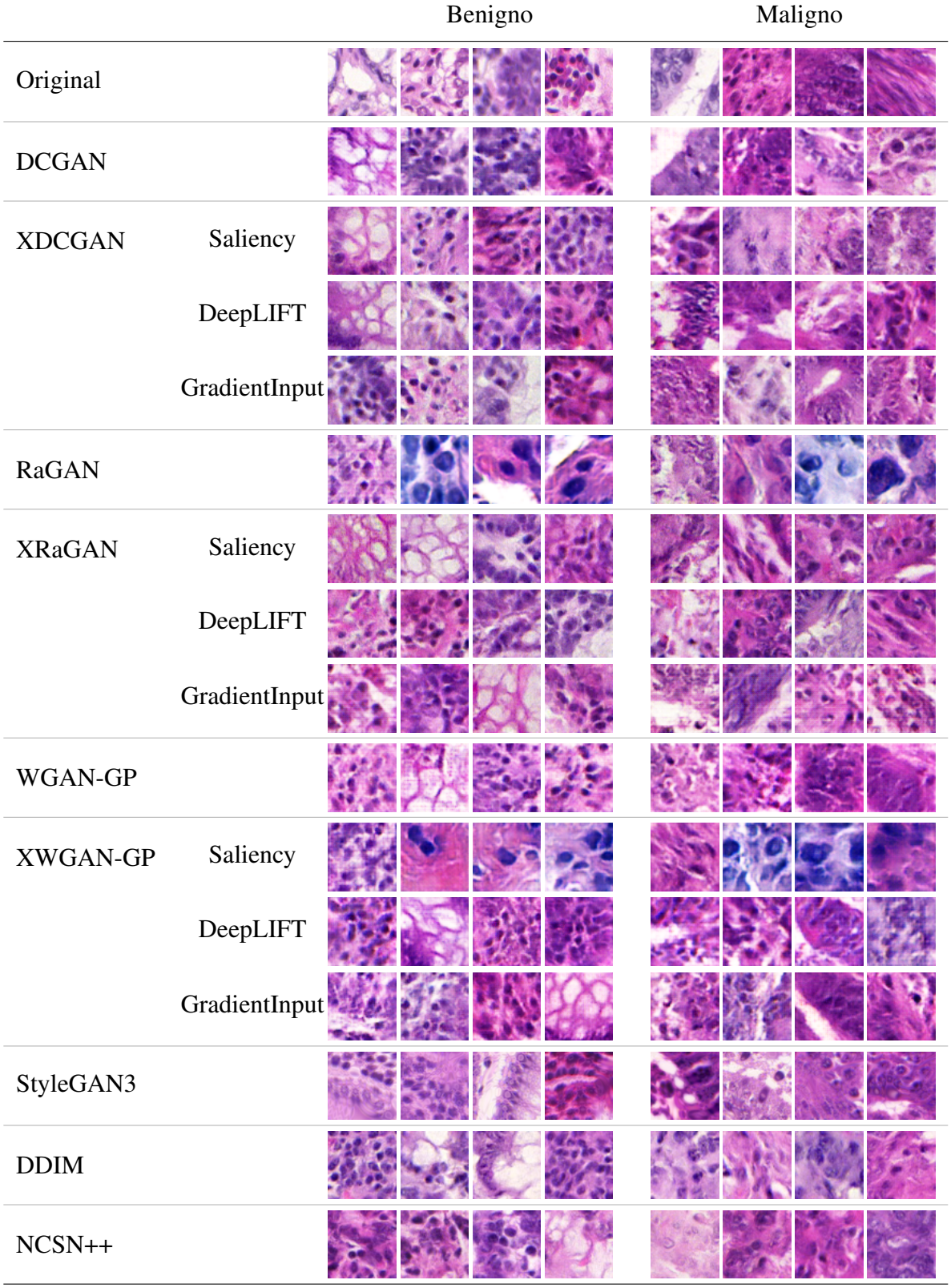
Os resultados dos testes realizados são apresentados nesta seção. Primeiramente, a qualidade das imagens artificiais geradas foi avaliada nas métricas de FID, KID e IS, para as variantes de GANs com treinamento educacional a partir de métodos XAI (*Saliency*, *DeepLIFT* e *Gradient⊙Input*), contrastando-os com os modelos básicos, o StyleGAN3, e os modelos de difusão *Denoising Diffusion Implicit Models* (DDIM) e Noise Conditional Score Network ++ (NCSN++). Em seguida, a comparação do desempenho para os algoritmos classificadores DenseNet, ResNet, CoAtNet e ViT, combinados com técnicas de aumento, e os ganhos representativos em cima do desempenho *baseline* desses algoritmos. Os experimentos foram conduzidos nos três *datasets* histológicos (UCSB, LG e CR). Exemplos das imagens originais e das artificiais geradas por cada conjunto de arquitetura utilizado no aumento artificial podem ser observados nas Figura 18, Figura 19, Figura 20.

4.1 QUALIDADE DAS IMAGENS: FID, KID E IS

A fim de avaliar a qualidade das imagens geradas via diferentes técnicas de aumento de dados exploradas neste trabalho, na Tabela 3 são apresentados os resultados de FID (quanto menor, melhor), KID (quanto menor, melhor) e IS (quanto maior, melhor) para cada *dataset* e gerador. A partir dos resultados obtidos, observa-se que, entre todos os algoritmos investigados, o DDIM obteve consistentemente os melhores valores absolutos de FID e KID em todos os cenários. Esses resultados reforçam a superioridade dos modelos de difusão na representação de distribuições complexas, superando as GANs, mesmo aquelas aprimoradas com técnicas de XAI.

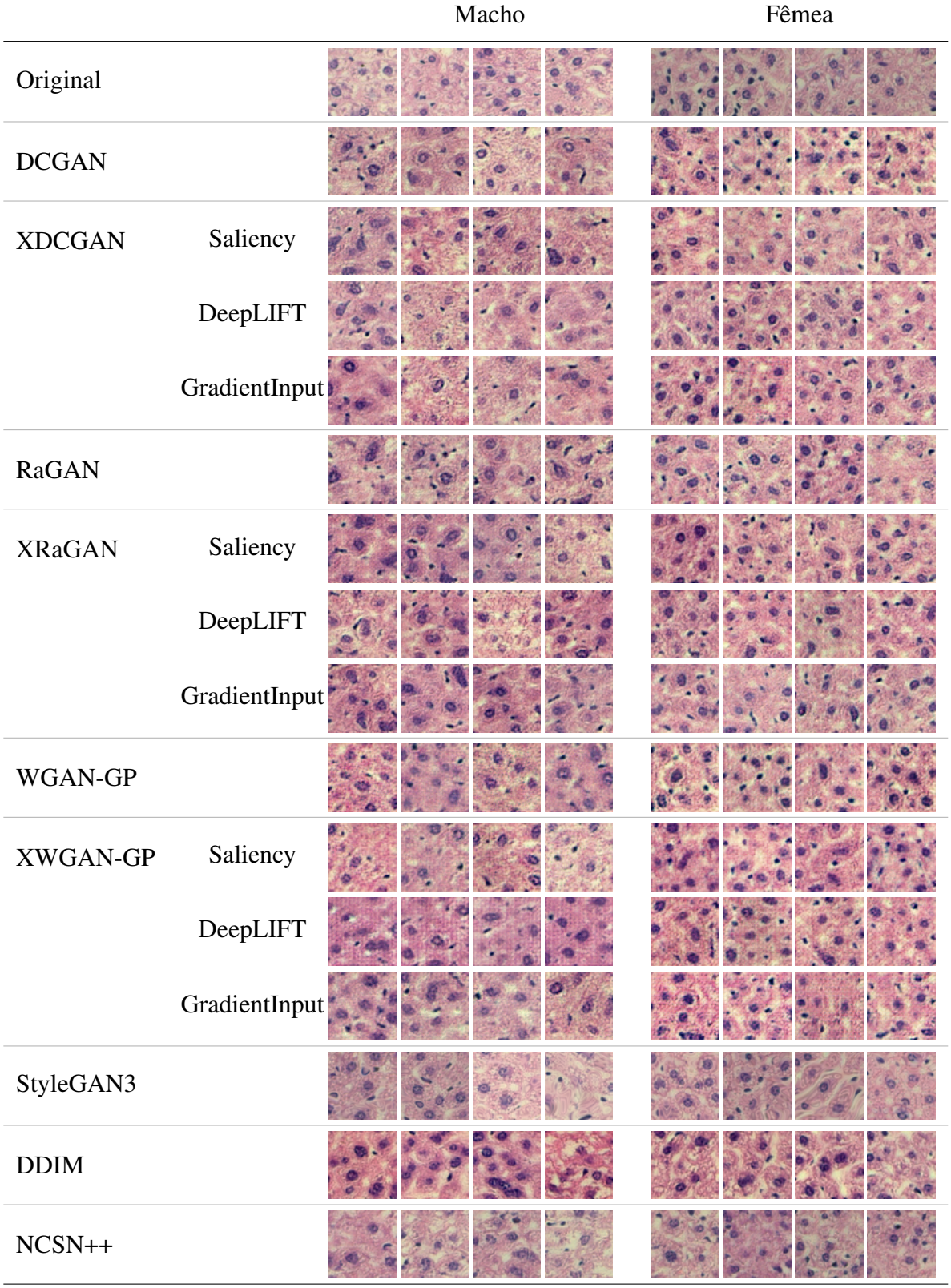
Um destaque importante é a consistência observada entre as métricas avaliadas. A inclusão do KID, calculado a partir das mesmas imagens artificiais utilizadas em Rozendo et al. (Rozendo et al., 2024), confirmou os melhores resultados apontados por FID e IS. Em todas as combinações analisadas, os valores de KID coincidiram com os melhores de FID e/ou IS entre as XGANs, reforçando a validade das métricas adotadas e indicando que os diferentes métodos capturam aspectos complementares da qualidade das imagens geradas. Além disso, ao analisar quantitativamente os resultados, observa-se que, para o *dataset* CR, o DDIM alcançou FID de 33,36, KID de 0,0237 e IS de 2,30, valores significativamente inferiores aos das GANs (por exemplo, DCGAN com Saliency: FID de 57,01, KID de 0,0573, IS de 2,43), evidenciando melhor aproximação da distribuição real. No *dataset* LG, o DDIM apresentou FID de 34,22 e KID de 0,0439, superando consistentemente as XGANs, que variaram entre FID de 88,30 a 114,01 e KID de 0,1150 a 0,1450, reforçando a maior cobertura do espaço latente; porém, o destaque neste *dataset* foi a StyleGAN3, com os melhores valores de IS (2,03) e de KID (0,0419). Para UCSB, o NCSN++ obteve o desempenho superior com FID de 40,86 e KID de 0,0315, destacando-se até mesmo do outro modelo de difusão. Neste cenário, as imagens sintéticas de

Figura 18 – Exemplos de imagens geradas artificialmente para o *dataset* CR com 4 amostras por classe (Benigno e Maligno).



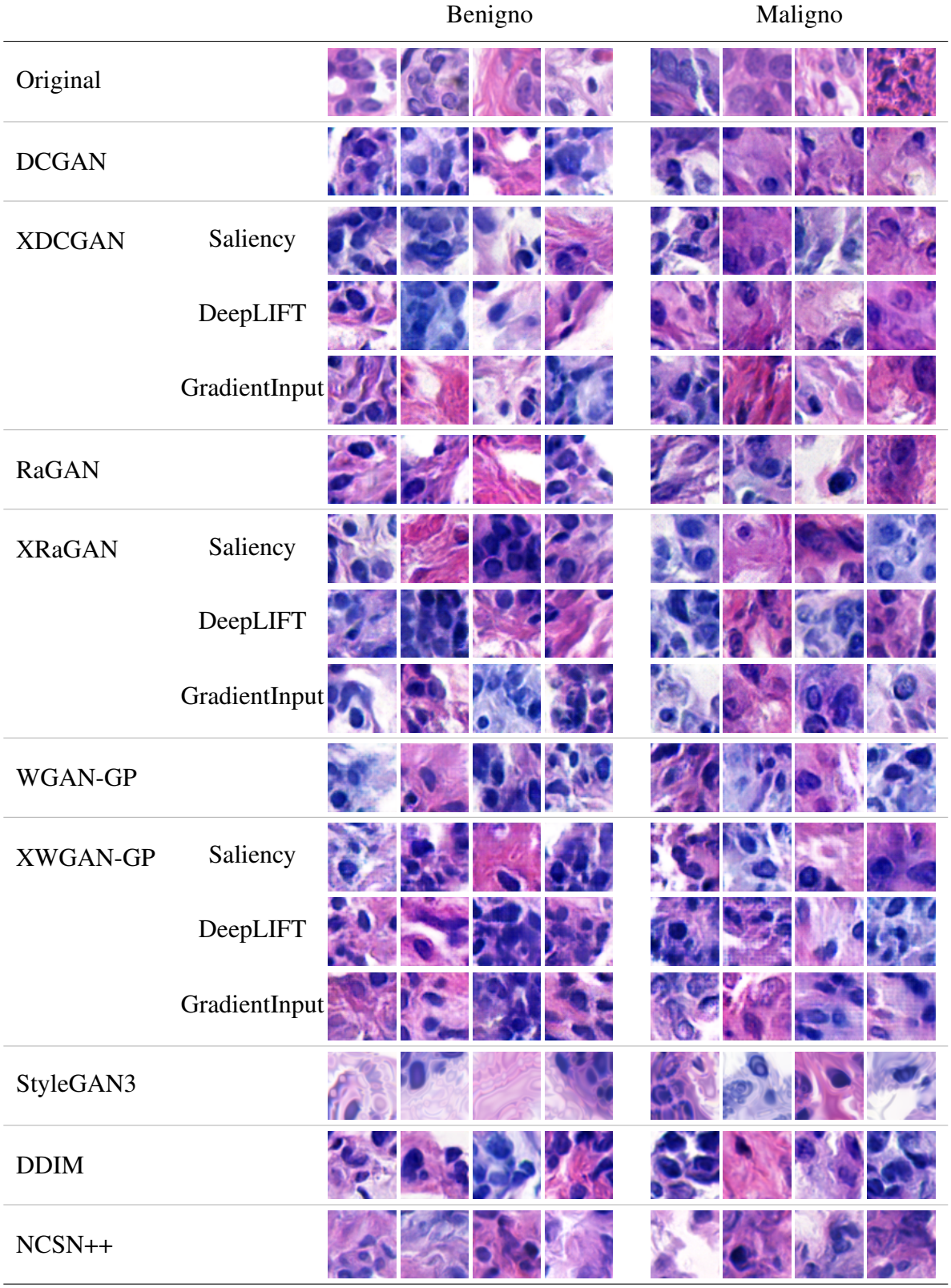
Fonte: Elaboração própria.

Figura 19 – Exemplos de imagens geradas artificialmente para o *dataset* LG com 4 amostras por classe (Macho e Fêmea).



Fonte: Elaboração própria.

Figura 20 – Exemplos de imagens geradas artificialmente para o *dataset* UCSB com 4 amostras por classe (Benigno e Maligno).



Fonte: Elaboração própria.

Tabela 3 – Resultados de FID ↓, KID ↓ e IS ↑ para os *datasets* CR, LG e UCSB.

		CR			LG			UCSB		
		FID	IS	KID	FID	IS	KID	FID	IS	KID
DCGAN	Sem aumento	59,13	2,41	0,0661	108,18	1,40	0,1450	72,77	2,78	0,0680
	Saliency	57,01	2,43	0,0573	114,01	1,38	0,1368	70,72	2,72	0,0701
	DeepLIFT	62,45	2,38	0,0712	98,24	1,42	0,1394	62,34	2,83	0,0598
	Grad ⊙ Input	60,32	2,43	0,0653	93,88	1,39	0,1543	69,90	2,72	0,0663
RaGAN	Sem aumento	68,39	2,40	0,0524	95,91	1,45	0,1337	68,42	2,62	0,0428
	Saliency	50,39	2,45	0,0395	88,30	1,42	0,1150	63,21	2,79	0,0421
	DeepLIFT	46,02	2,50	0,0337	101,78	1,40	0,1428	69,73	2,62	0,0463
	Grad ⊙ Input	58,57	2,56	0,0425	95,12	1,40	0,1287	66,68	2,65	0,0507
WGAN-GP	Sem aumento	82,81	2,21	0,0777	137,46	1,45	0,1850	81,07	2,44	0,0587
	Saliency	72,01	2,28	0,0555	128,26	1,43	0,1768	92,27	2,50	0,0660
	DeepLIFT	76,86	2,29	0,0659	153,10	1,57	0,1870	89,34	2,55	0,0634
	Grad ⊙ Input	74,50	2,17	0,0643	143,97	1,50	0,1973	87,65	2,74	0,0561
StyleGAN3		50,76	**2,95**	0,0386	38,84	**2,03**	**0,0419**	115,96	**2,86**	0,1018
DDIM		**33,36**	2,30	**0,0237**	**34,22**	1,64	0,0439	61,40	2,45	0,0493
NCSN++		43,97	2,00	0,0376	41,28	1,36	0,0530	**40,86**	2,50	**0,0315**

Fonte: Elaboração própria.

maior qualidade não apenas apresentam características visuais convincentes (capturadas pelo IS), mas também preservam a estrutura estatística das distribuições reais (avaliada por FID e KID).

Em complemento, a concordância entre FID e KID é particularmente relevante, pois o KID é uma métrica não paramétrica baseada no *Maximum Mean Discrepancy*, é menos sensível a pressupostos de gaussianidade que podem limitar a interpretação do FID. Dessa forma, o KID valida, adicionalmente, a fidelidade estatística das amostras. Os resultados quantitativos indicaram ainda que modelos com melhor FID e KID exibem menor tendência ao colapso de modo. Por exemplo, as DCGANs com XAI apresentaram redução relativa de KID entre 10% e 15% em comparação com versões sem XAI, sugerindo que as técnicas de explicabilidade também contribuem para uma cobertura mais ampla da distribuição real. Entretanto, mesmo com melhorias locais, os valores absolutos de FID e KID das GANs permanecem acima dos observados no DDIM, evidenciando a superioridade intrínseca dos modelos de difusão na modelagem de distribuições complexas de alta dimensão, bem como as XGANs não ultrapassam a StyleGAN3 e os modelos de difusão, indicando a evolução no processo de geração com estes modelos, que surgiram após as GANs *baseline*.

4.2 DISCUSSÃO DOS RESULTADOS DE CLASSIFICAÇÃO

A análise dos resultados apresentados nas Tabelas 4, 5 e 6 evidencia o impacto do aumento artificial de dados sobre diferentes arquiteturas de classificação e *datasets*. Em termos gerais, a utilização de imagens geradas artificialmente promoveu ganhos modestos em relação ao *baseline*, mas com variações significativas entre arquiteturas e tipos de geradores.

Tabela 4 – Média das acurácias (%) com 5-folds para o *dataset* CR.

Modelos	DenseNet121	ResNet50	ViT	CoAtNet
Sem aumento artificial	89,74%	88,34%	81,29%	93,19%
DCGAN	91,93%	89,53%	82,67%	92,30%
XDCGAN + Saliency	90,68%	90,17%	81,07%	93,44%
XDCGAN + DeepLIFT	91,61%	89,93%	79,97%	93,15%
XDCGAN + Grad $\odot$ Input	90,95%	90,40%	81,98%	93,44%
RaGAN	91,80%	89,10%	83,53%	92,73%
XRaGAN + Saliency	91,64%	90,41%	81,87%	**93,88%**
XRaGAN + DeepLIFT	**92,22%**	90,89%	84,39%	93,06%
XRaGAN + Grad $\odot$ Input	91,02%	90,23%	80,56%	92,64%
WGAN-GP	91,94%	90,86%	81,60%	91,95%
XWGAN-GP + Saliency	90,24%	88,76%	84,41%	93,64%
XWGAN-GP + DeepLIFT	91,28%	88,88%	78,25%	93,36%
XWGAN-GP + Grad $\odot$ Input	91,28%	89,67%	83,34%	93,13%
StyleGAN3	91,97%	**91,26%**	**88,40%**	93,26%
DDIM	91,77%	91,16%	84,17%	93,03%
NCSN++	90,56%	90,54%	84,54%	92,39%

Fonte: Elaboração própria.

Quando o conjunto CR é considerado, os efeitos do aumento artificial de dados tornam-se particularmente evidentes em modelos baseados em mecanismos de atenção. O ViT apresentou um ganho expressivo de 7,11 pontos percentuais, passando de 81,29% (*baseline*) para 88,40% com a StyleGAN3, o que evidencia que arquiteturas do tipo *Transformer* exploram de maneira mais eficiente a diversidade de padrões sintéticos e capturam dependências de longo alcance introduzidas pelas imagens geradas. Entre as variantes de XGAN, a XRaGAN com DeepLIFT atingiu 92,22% na DenseNet121, superando o *baseline* de 89,74% em 2,48 pontos percentuais, enquanto a XRaGAN com Saliency elevou a acurácia da CoAtNet para 93,88%, ultrapassando os 93,19% sem aumento. Esses resultados indicam que técnicas XAI incorporadas ao processo de geração não apenas promovem maior diversidade semântica nas amostras, mas também podem

Tabela 5 – Média das acurácias (%) com 5-folds para o *dataset* UCSB.

Modelos	DenseNet121	ResNet50	ViT	CoAtNet
Sem aumento artificial	77,45%	75,95%	75,00%	76,17%
DCGAN	75,88%	74,98%	72,21%	75,01%
XDCGAN + Saliency	78,51%	**77,59%**	72,36%	74,49%
XDCGAN + DeepLIFT	78,00%	77,21%	72,04%	74,76%
XDCGAN + Grad $\odot$ Input	77,56%	77,08%	71,44%	74,03%
RaGAN	76,30%	75,53%	71,30%	72,77%
XRaGAN + Saliency	76,88%	75,88%	72,91%	72,89%
XRaGAN + DeepLIFT	76,97%	74,48%	71,75%	74,85%
XRaGAN + Grad $\odot$ Input	79,40%	76,66%	71,61%	75,44%
WGAN-GP	78,22%	76,63%	72,86%	73,54%
XWGAN-GP + Saliency	76,28%	76,61%	70,91%	72,94%
XWGAN-GP + DeepLIFT	76,06%	76,42%	73,51%	73,96%
XWGAN-GP + Grad $\odot$ Input	77,38%	75,75%	70,39%	76,59%
StyleGAN3	77,49%	74,00%	**76,26%**	75,22%
DDIM	78,34%	76,19%	75,61%	**79,03%**
NCSN++	**79,46**%	76,44%	74,36%	76,11%

Fonte: Elaboração própria.

ampliar a eficácia do treinamento supervisionado ao fornecer áreas mais informativas e relevantes na etapa de geração.

No conjunto UCSB, por exemplo, caracterizado por poucas amostras e significativa presença de regiões com baixa informação diagnóstica (frequentemente quase totalmente em branco), os efeitos do aumento artificial foram mais heterogêneos, embora positivos. O DDIM destacou-se no CoAtNet, alcançando 79,03%, o que representa um acréscimo de 2,86 pontos percentuais em relação ao *baseline* (76,17%). A StyleGAN3 elevou a acurácia do ViT para 76,26% (+1,26 pontos), enquanto o NCSN++ atingiu 79,46% na DenseNet121 (+2,01 pontos). Outras variantes, como XDCGAN+Saliency (77,59% na ResNet50) e XWGAN-GP+DeepLIFT (73,96% na CoAtNet), exibiram resultados mais variados. Esses achados sugerem que, em cenários com baixa representatividade e alta esparsidade, a eficácia do aumento depende fortemente da sensibilidade do classificador à variabilidade sintética e da capacidade do gerador de sintetizar amostras com conteúdo informativo relevante.

No *dataset* LG, que apresenta elevada separabilidade entre classes (*baseline* >95%), os ganhos com aumento foram naturalmente mais sutis, refletindo um teto de desempenho inerente ao problema. Ainda assim, os resultados foram notavelmente consistentes. O CoAtNet

Tabela 6 – Média das acurácias (%) com 5-folds para o *dataset* LG.

Modelos	DenseNet121	ResNet50	ViT	CoAtNet
Sem aumento artificial	95,87%	95,17%	97,05%	99,50%
DCGAN	97,83%	97,59%	96,79%	**99,98%**
XDCGAN + Saliency	97,57%	**98,44%**	95,64%	99,86%
XDCGAN + DeepLIFT	97,52%	97,62%	95,47%	99,34%
XDCGAN + Grad $\odot$ Input	97,45%	98,11%	96,06%	99,95%
RaGAN	98,04%	97,57%	95,24%	99,74%
XRaGAN + Saliency	96,96%	98,14%	95,50%	99,69%
XRaGAN + DeepLIFT	97,45%	97,83%	95,33%	98,02%
XRaGAN + Grad $\odot$ Input	97,67%	97,45%	95,85%	99,83%
WGAN-GP	97,52%	98,14%	91,11%	99,60%
XWGAN-GP + Saliency	97,57%	97,83%	96,32%	99,81%
XWGAN-GP + DeepLIFT	98,09%	97,43%	96,79%	99,17%
XWGAN-GP + Grad $\odot$ Input	98,11%	97,88%	95,19%	99,79%
StyleGAN3	97,92%	97,83%	**97,62%**	99,95%
DDIM	**98,14%**	98,04%	97,03%	99,86%
NCSN++	97,83%	97,95%	96,98%	99,76%

Fonte: Elaboração própria.

manteve acurácias próximas de 100% com quase todos os geradores (DCGAN: 99,98%; StyleGAN3: 99,95%; DDIM: 99,86%), demonstrando o desempenho das arquiteturas híbridas CNN-*Transformer* na extração de características discriminativas, mesmo diante de variações na qualidade das imagens artificiais. A DenseNet121 obteve ganhos marginais com DDIM (98,14%) e XWGAN-GP+DeepLIFT (98,09%), enquanto o ViT atingiu seu melhor desempenho com StyleGAN3 (97,62%). Embora a diversidade introduzida pelas imagens artificiais tenha produzido ganhos sutis, a escolha do gerador e do classificador ainda influenciou diferenças de até 1,5 pontos percentuais, o que pode ser relevante em cenários que exigem alta precisão diagnóstica.

Cabe ainda destacar que, em todos os conjuntos avaliados, os geradores tradicionais apresentaram padrões consistentes: o DCGAN puro proporcionou ganhos discretos (por exemplo, 91,93% na DenseNet121 e 92,30% na CoAtNet no conjunto CR), e o WGAN-GP manteve desempenho estável, embora inferior ao das variantes XGAN. Apesar de o DDIM ter apresentado os melhores indicadores de qualidade de imagem (FID de 33,36; KID de 0,0237; IS de 2,30), sua contribuição para a acurácia permaneceu limitada, sugerindo que a fidelidade visual e estatística, embora necessárias, não garantiram ganhos proporcionais de desempenho em cenários com

alta heterogeneidade estrutural, nos quais a relevância semântica e a diversidade contextual são determinantes.

De modo geral, a análise comparativa dos três conjuntos revela padrões claros e diretrizes úteis para aplicações futuras: modelos baseados em atenção (ViT) mostraram-se altamente responsivos à diversidade sintética em *datasets* complexos como CR, com aumentos de até 7,11 pontos percentuais, evidenciando sua capacidade de explorar padrões globais e dependências contextuais em dados aumentados; redes convolucionais profundas (DenseNet121, ResNet50) apresentaram ganhos mais modestos (1–3 pontos), mas alcançaram melhorias expressivas quando associadas a variantes de XGAN; e, por fim, arquiteturas híbridas (CoAtNet) demonstraram desempenho consistentemente elevado (>93% no CR, >99% no LG), com baixa sensibilidade à variação do gerador, tornando-se ótimas opções em diferentes cenários de aplicação, e a fidelidade visual, embora essencial, não se correlacionou linearmente com o desempenho preditivo, por exemplo, o DDIM obteve FID de 33,36 (CR) e 34,22 (LG), mas os ganhos de acurácia foram limitados.

Assim, os resultados indicam que a eficácia do aumento artificial depende fortemente de três fatores principais: a complexidade e heterogeneidade do conjunto de dados, a arquitetura do classificador e a capacidade do gerador de sintetizar amostras representativas. Modelos *Transformer* e híbridos mostraram maior sensibilidade à diversidade introduzida por imagens artificiais, enquanto redes convolucionais profundas responderam de forma mais contida, exceto quando associadas a variantes de XGAN. Por fim, a combinação entre métricas de qualidade (FID, KID e IS) e avaliação de desempenho preditivo mostrou-se indispensável para a seleção de abordagens eficazes em geração e aumento de dados histológicos.

5 CONCLUSÃO

Neste trabalho, foi proposta e avaliada uma abordagem computacional baseada em variantes de *Generative Adversarial Networks* aprimoradas por explicabilidade (XGANs), na arquitetura *StyleGAN3* e nos modelos de difusão DDIM e NCSN++, com o objetivo de investigar a geração de imagens histológicas sintéticas e seu impacto em tarefas de classificação. Os resultados indicaram que o modelo DDIM apresentou desempenho superior em quase todos os conjuntos de dados em termos das métricas KID e FID, corroborando evidências recentes que consolidam os modelos de difusão como o atual *state-of-the-art* em síntese de imagens. Entre as GANs avaliadas, a StyleGAN3 destacou-se como a alternativa mais consistente, enquanto as variantes XGAN obtiveram ganhos competitivos sob diversas configurações experimentais, reforçando a viabilidade de integrar mecanismos de explicabilidade sem perdas substanciais de qualidade generativa.

Além disso, o aumento artificial de dados resultou em ganhos claros de desempenho em relação ao *baseline*, particularmente em classificadores baseados em atenção. No conjunto CR, os modelos ViT e CoAtNet apresentaram ganhos expressivos de acurácia, enquanto no conjunto LG, caracterizado por maior separabilidade intrínseca entre classes, as melhorias foram mais modestas, porém consistentes. Esses resultados indicam que, embora os modelos de difusão tenham apresentado a maior qualidade de síntese, as arquiteturas StyleGAN3 e XGANs demonstraram maior efetividade no aumento artificial de dados, especialmente em classificadores baseados em atenção, ao equilibrar diversidade e estabilidade.

Os achados deste estudo ressaltam que a abordagem desenvolvida, que integra modelos generativos a classificadores baseados em atenção, representa uma estratégia promissora para aprimorar a análise automatizada de imagens histológicas. Como direção futura, propõe-se o desenvolvimento de abordagens híbridas que combinem a fidelidade generativa da StyleGAN3 ou dos modelos de difusão com mecanismos avançados de interpretabilidade, buscando alcançar um equilíbrio entre alta fidelidade visual e desempenho aprimorado em classificação.

REFERÊNCIAS

AGEMAP, N. I. o. A. **The Atlas of Gene Expression in Mouse Aging Project (AGEMAP)**. 2020. Disponível em: https://ome.grc.nia.nih.gov/iicbu2008/agemap/index.html.

ARJOVSKY, M.; CHINTALA, S.; BOTTOU, L. Wasserstein generative adversarial networks. In: PMLR. **International conference on machine learning**. [S.l.], 2017. p. 214–223.

BANSAL, A. et al. Cold diffusion: Inverting arbitrary image transforms without noise. In: OH, A. et al. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2023. v. 36, p. 41259–41282. Disponível em: https://proceedings.neurips.cc/paper_files/paper/2023/file/80fe51a7d8d0c73ff7439c2a2554ed53-Paper-Conference.pdf.

BINKOWSKI, M. et al. Demystifying mmd gans. In: **International Conference on Learning Representations (ICLR)**. [S.l.: s.n.], 2018.

BORJI, A. Pros and cons of gan evaluation measures. **Computer Vision and Image Understanding**, v. 179, p. 41–65, 2019. ISSN 1077-3142. Disponível em: https://www.sciencedirect.com/science/article/pii/S1077314218304272.

BRAY, F. et al. Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. **CA: A Cancer Journal for Clinicians**, Wiley Periodicals LLC on behalf of the American Cancer Society, United States, v. 74, n. 3, p. 229–263, May–Jun 2024. ISSN 0007-9235. Disponível em: https://doi.org/10.3322/caac.21834.

CAO, H. et al. A survey on generative diffusion models. **IEEE Transactions on Knowledge and Data Engineering**, v. 36, n. 7, p. 2814–2830, 2024.

CHEN, P.-H. C.; LIU, Y.; PENG, L. How to develop machine learning models for healthcare. **Nature Materials**, v. 18, n. 5, p. 410–414, 2019. ISSN 1476-4660. Disponível em: https://doi.org/10.1038/s41563-019-0345-0.

DAI, Z. et al. Coatnet: marrying convolution and attention for all data sizes. In: **Proceedings of the 35th International Conference on Neural Information Processing Systems**. Red Hook, NY, USA: Curran Associates Inc., 2024. (NIPS '21). ISBN 9781713845393.

DENG, J. et al. Imagenet: A large-scale hierarchical image database. In: **2009 IEEE Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2009. p. 248–255.

DHARIWAL, P.; NICHOL, A. Diffusion models beat gans on image synthesis. In: **Advances in Neural Information Processing Systems**. [s.n.], 2021. v. 34, p. 8780–8794. Disponível em: https://proceedings.neurips.cc/paper_files/paper/2021/hash/49ad23d1ec9fa4bd8d77d02681df5cfa-Abstract.html.

DOSOVITSKIY, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. **arXiv preprint arXiv:2010.11929**, 2020.

GELASCA, E. D. et al. Evaluation and benchmark for biological image segmentation. In: **IEEE International Conference on Image Processing**. [S.l.: s.n.], 2008. Place: San Diego, CA.

GOODFELLOW, I. J. et al. Generative adversarial nets. **Advances in neural information processing systems**, v. 27, 2014.

GULRAJANI, I. et al. Improved training of wasserstein gans. In: **Proceedings of the 31st International Conference on Neural Information Processing Systems**. Red Hook, NY, USA: Curran Associates Inc., 2017. (NIPS'17), p. 5769–5779. ISBN 9781510860964.

HARB, R.; POCK, T.; MüLLER, H. Diffusion-based generation of histopathological whole slide images at a gigapixel scale. **arXiv preprint**, 2024. Disponível em: https://arxiv.org/abs/2403.02498.

HE, K. et al. Deep residual learning for image recognition. In: **2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2016. p. 770–778.

HEUSEL, M. et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: **Advances in Neural Information Processing Systems (NeurIPS)**. [S.l.: s.n.], 2017.

HO, J.; JAIN, A.; ABBEEL, P. Denoising diffusion probabilistic models. In: **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2020. v. 33, p. 6840–6851. Disponível em: https://proceedings.neurips.cc/paper_files/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

HUANG, G. et al. Densely connected convolutional networks. In: **2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2017. p. 2261–2269.

IGLESIAS, G.; TALAVERA, E.; DíAZ-ÁLVAREZ, A. A survey on gans for computer vision: Recent research, analysis and taxonomy. **Computer Science Review**, v. 48, p. 100553, 2023. ISSN 1574-0137. Disponível em: https://www.sciencedirect.com/science/article/pii/S1574013723000205.

Instituto Nacional de Câncer (INCA). **Controle do câncer de mama no Brasil: dados e números 2024**. Rio de Janeiro, RJ: Ministério da Saúde, Instituto Nacional de Câncer (INCA), 2024. Disponível na Biblioteca Virtual em Saúde Prevenção e Controle de Câncer: http://controlecancer.bvs.br/. Organizadoras: Caroline Madalena Ribeiro, Danielle Nogueira Ramos, Itamar Bento Claro, Maria Beatriz Kneipp Dias, Mônica de Assis. Disponível em: http://www.inca.gov.br.

International Agency for Research on Cancer (IARC). **Breast Cancer Screening**. Lyon, France: International Agency for Research on Cancer, 2016. v. 15. 7–8 p. (IARC Handbooks of Cancer Prevention, v. 15). ISBN 978-92-832-3015-1, 978-92-832-3017-5.

JOLICOEUR-MARTINEAU, A. The relativistic discriminator: A key element missing from standard gan. In: **Proceedings of the International Conference on Learning Representations**. Nice, France: [s.n.], 2019.

KARRAS, T. et al. Elucidating the design space of diffusion-based generative models. In: KOYEJO, S. et al. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2022. v. 35, p. 26565–26577. Disponível em: https://proceedings.neurips.cc/paper_files/paper/2022/file/a98846e9d9cc01cfb87eb694d946ce6b-Paper-Conference.pdf.

KARRAS, T. et al. Alias-free generative adversarial networks. In: **Proc. NeurIPS**. [S.l.: s.n.], 2021.

KARRAS, T. et al. Alias-free generative adversarial networks. In: **Advances in Neural Information Processing Systems (NeurIPS)**. [S.l.: s.n.], 2021. p. 852–863.

Khan, U. et al. Staining normalization in histopathology: Method benchmarking using multicenter dataset. **arXiv e-prints**, p. arXiv:2506.19106, jun. 2025.

KIM, J. et al. Global patterns and trends in breast cancer incidence and mortality across 185 countries. **Nature Medicine**, Springer Nature America, Inc., United States, v. 31, n. 4, p. 1154–1162, April 2025. ISSN 1546-170X. Disponível em: https://doi.org/10.1038/s41591-025-03502-3.

LU, C. et al. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. **Advances in neural information processing systems**, v. 35, p. 5775–5787, 2022.

MACENKO, M. et al. A method for normalizing histology slides for quantitative analysis. In: **Proceedings of the IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI)**. [S.l.: s.n.], 2009. p. 1107–1110.

NAKKIRAN, P. et al. Step-by-step diffusion: An elementary tutorial. **arXiv preprint arXiv:2406.08929**, 2024.

NIEHUES, J. M. et al. Using histopathology latent diffusion models as privacy-preserving dataset augmenters improves downstream classification performance. **Computers in Biology and Medicine**, Elsevier Ltd., v. 175, p. 108410, 2024. Disponível em: https://doi.org/10.1016/j.compbiomed.2024.108410.

OLIVEIRA, C. I. de et al. Hybrid models for classifying histological images: An association of deep features by transfer learning with ensemble classifier. **Multimedia Tools and Applications**, v. 83, n. 8, p. 21929–21952, 2024. ISSN 1573-7721. Disponível em: https://doi.org/10.1007/s11042-023-16351-4.

PEREIRA, D. C. et al. Evaluation of fractal descriptors, deep features and xai representations with lasso-regularized hermite polynomial classifier for h&e histological image classification. **Biomedical Signal Processing and Control**, v. 112, p. 108393, 2026. ISSN 1746-8094. Disponível em: https://www.sciencedirect.com/science/article/pii/S1746809425009048.

RADFORD, A.; METZ, L.; CHINTALA, S. Unsupervised representation learning with deep convolutional generative adversarial networks. **CoRR**, abs/1511.06434, 2015. Disponível em: https://api.semanticscholar.org/CorpusID:11758569.

REZENDE, D. J.; MOHAMED, S.; WIERSTRA, D. Stochastic backpropagation and approximate inference in deep generative models. In: PMLR. **International conference on machine learning**. [S.l.], 2014. p. 1278–1286.

ROBERTO, G. F. et al. An ensemble of learned features and reshaping of fractal geometry-based descriptors for classification of histological images. **Pattern Analysis and Applications**, v. 27, n. 1, p. 8, 2024. ISSN 1433-755X. Disponível em: https://doi.org/10.1007/s10044-024-01223-w.

RONNEBERGER, O.; FISCHER, P.; BROX, T. U-net: Convolutional networks for biomedical image segmentation. In: SPRINGER. **International Conference on Medical image computing and computer-assisted intervention**. [S.l.], 2015. p. 234–241.

ROSHANDEL, G.; GHASEMI-KEBRIA, F.; MALEKZADEH, R. Colorectal cancer: Epidemiology, risk factors, and prevention. **Cancers**, v. 16, n. 8, 2024. ISSN 2072-6694. Disponível em: https://www.mdpi.com/2072-6694/16/8/1530.

ROZENDO, G. B. et al. Data augmentation in histopathological classification: An analysis exploring GANs with XAI and vision transformers. v. 14, n. 18, p. 8125, 2024. ISSN 2076-3417. Disponível em: https://www.mdpi.com/2076-3417/14/18/8125.

SALIMANS, T. et al. Improved techniques for training gans. In: **Advances in Neural Information Processing Systems (NeurIPS)**. [S.l.: s.n.], 2016.

SHORTEN, C.; KHOSHGOFTAAR, T. M. A survey on image data augmentation for deep learning. **Journal of Big Data**, v. 6, n. 1, p. 60, 2019. ISSN 2196-1115. Disponível em: https://doi.org/10.1186/s40537-019-0197-0.

SHRIKUMAR, A.; GREENSIDE, P.; KUNDAJE, A. Learning important features through propagating activation differences. In: PMLR. **Proceedings of the 34th International Conference on Machine Learning (ICML)**. [S.l.], 2017. v. 70, p. 3145–3153.

SIMONYAN, K.; VEDALDI, A.; ZISSERMAN, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. In: **2nd International Conference on Learning Representations (ICLR) Workshop**. [S.l.: s.n.], 2014.

SIRINUKUNWATTANA, K. et al. Gland segmentation in colon histology images: The glas challenge contest. v. 35, p. 489–502, 2017. Publisher: Elsevier.

SOHL-DICKSTEIN, J. et al. Deep unsupervised learning using nonequilibrium thermodynamics. In: **Proceedings of the 32nd International Conference on Machine Learning**. Lille, France: PMLR, 2015. v. 37, p. 2256–2265. Disponível em: http://proceedings.mlr.press/v37/sohl-dickstein15.pdf.

SONG, J.; MENG, C.; ERMON, S. Denoising diffusion implicit models. **ArXiv**, abs/2010.02502, 2020. Disponível em: https://api.semanticscholar.org/CorpusID:222140788.

SONG, Y. et al. Maximum likelihood training of score-based diffusion models. **Advances in neural information processing systems**, v. 34, p. 1415–1428, 2021.

SONG, Y.; ERMON, S. Generative modeling by estimating gradients of the data distribution. **Advances in neural information processing systems**, v. 32, 2019.

SONG, Y. et al. Score-based generative modeling through stochastic differential equations. In: **International Conference on Learning Representations**. [s.n.], 2021. Disponível em: https://openreview.net/forum?id=PxTIG12RRHS.

SZEGEDY, C. et al. Rethinking the inception architecture for computer vision. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2016. p. 2818–2826.

TAFAVVOGHI, M. et al. Publicly available datasets of breast histopathology h&e whole-slide images: A scoping review. **Journal of Pathology Informatics**, v. 15, p. 100363, 2024. ISSN 2153-3539. Disponível em: https://www.sciencedirect.com/science/article/pii/S2153353924000026.

THAKKAR, D. et al. **Comparative Analysis of Diffusion Generative Models in Computational Pathology**. 2024. Preprint submitted to Under Review.

TorchVision maintainers and contributors. **TorchVision: PyTorch's Computer Vision library**. GitHub, 2016. Disponível em: https://github.com/pytorch/vision.

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

XU, C. et al. Stain normalization of histopathological images based on deep learning: A review. **Diagnostics**, v. 15, n. 8, 2025. ISSN 2075-4418. Disponível em: https://www.mdpi.com/2075-4418/15/8/1032.

YANG, L. et al. Diffusion models: A comprehensive survey of methods and applications. **ACM Comput. Surv.**, Association for Computing Machinery, New York, NY, USA, v. 56, n. 4, nov. 2023. ISSN 0360-0300. Disponível em: https://doi.org/10.1145/3626235.

ZHANG, Q.; CHEN, Y. Fast sampling of diffusion models with exponential integrator. In: **NeurIPS 2022 Workshop on Score-Based Methods**. [s.n.], 2022. Disponível em: https://openreview.net/forum?id=hiZ98L9tX1k.