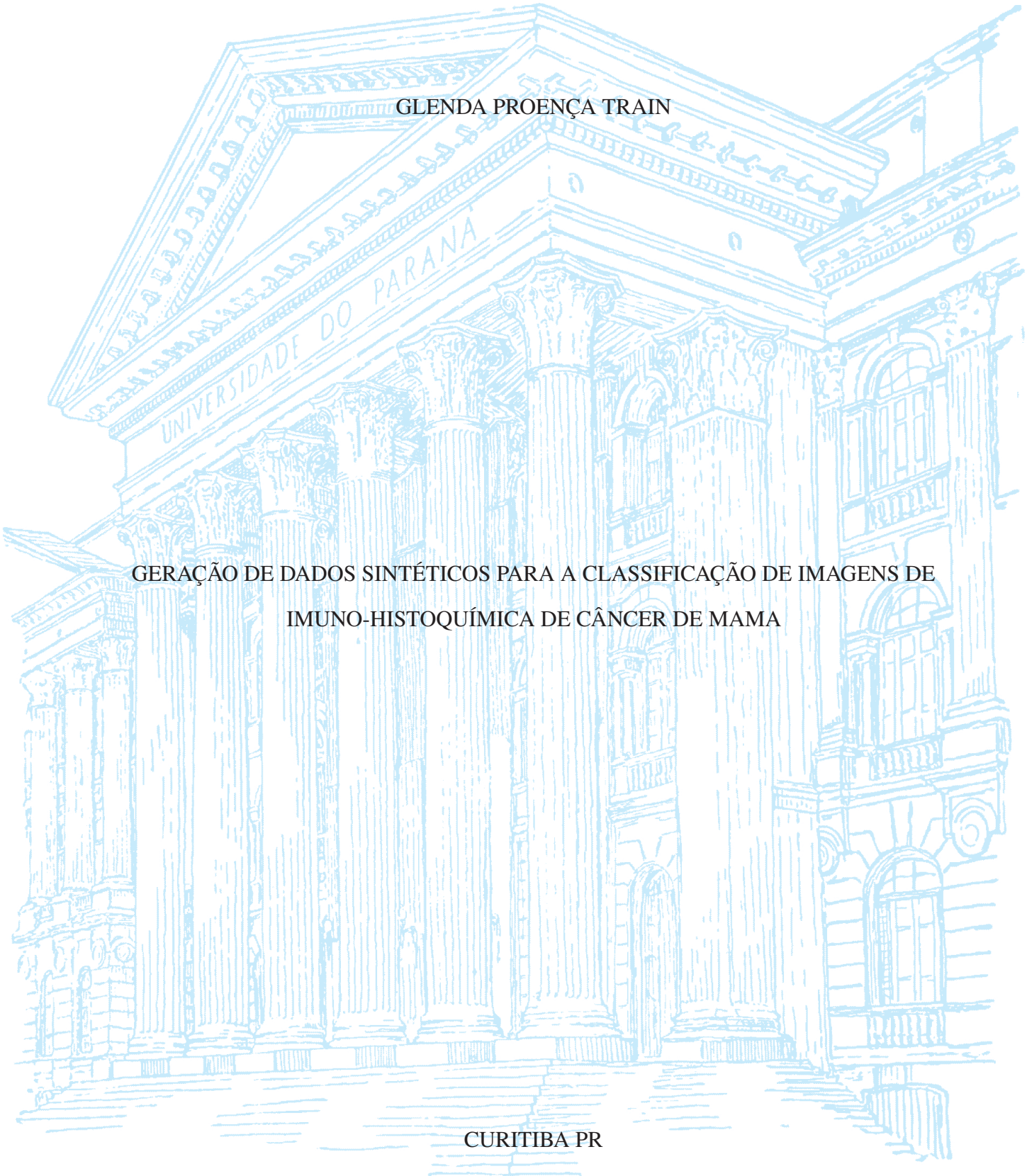
UNIVERSIDADE FEDERAL DO PARANÁ

GLENDA PROENÇA TRAIN



GERAÇÃO DE DADOS SINTÉTICOS PARA A CLASSIFICAÇÃO DE IMAGENS DE IMUNO-HISTOQUÍMICA DE CÂNCER DE MAMA

CURITIBA PR

2024

GLENDA PROENÇA TRAIN

GERAÇÃO DE DADOS SINTÉTICOS PARA A CLASSIFICAÇÃO DE IMAGENS DE IMUNO-HISTOQUÍMICA DE CÂNCER DE MAMA

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Informática no Programa de Pós-Graduação em Informática, Setor de Ciências Exatas, da Universidade Federal do Paraná.

Área de concentração: *Ciência da Computação*.

Orientador: Prof. Dr. Lucas Ferrari de Oliveira.

Coorientador: Prof. Dr. Sergio Ossamu Ioshii.

CURITIBA PR

2024

DADOS INTERNACIONAIS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)
UNIVERSIDADE FEDERAL DO PARANÁ
SISTEMA DE BIBLIOTECAS – BIBLIOTECA DE CIÊNCIA E TECNOLOGIA

Train, Glenda Proença
Geração de dados sintéticos para a classificação de imagens de imuno-histoquímica de câncer de mama / Glenda Proença Train. – Curitiba, 2024.
1 recurso on-line : PDF.

Dissertação (Mestrado) - Universidade Federal do Paraná, Setor de Ciências Exatas, Programa de Pós-Graduação em Informática.

Orientador: Lucas Ferrari de Oliveira
Coorientador: Sergio Ossamu Ioshii

1. Neoplasias da Mama. 2. Receptores de Estrogênio. 3. Receptores de Progesterona. 4. Inteligência artificial. I. Universidade Federal do Paraná. II. Programa de Pós-Graduação em Informática. III. Oliveira, Lucas Ferrari de. IV. Ioshii, Sergio Ossamu. V. Título.

Bibliotecário: Leticia Priscila Azevedo de Sousa CRB-9/2029

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação INFORMÁTICA da Universidade Federal do Paraná foram convocados para realizar a arguição da dissertação de Mestrado de **GLENDA PROENÇA TRAIN** intitulada: **GERAÇÃO DE DADOS SINTÉTICOS PARA A CLASSIFICAÇÃO DE IMAGENS DE IMUNO-HISTOQUÍMICA DE CÂNCER DE MAMA**, sob orientação do Prof. Dr. LUCAS FERRARI DE OLIVEIRA, que após terem inquirido a aluna e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de mestra está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

CURITIBA, 05 de Setembro de 2024.

Assinatura Eletrônica
06/09/2024 09:36:56.0
LUCAS FERRARI DE OLIVEIRA
Presidente da Banca Examinadora

Assinatura Eletrônica
08/09/2024 20:03:43.0
EDUARDO TODT
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica
06/09/2024 12:28:43.0
FERNANDO ROBERTO PEREIRA
Avaliador Externo (INSTITUTO FEDERAL DE SANTA CATARINA)

Assinatura Eletrônica
09/09/2024 17:32:38.0
SERGIO OSSAMU IOSHII
Coorientador(a) (DEPARTAMENTO DE PATOLOGIA BáSICA)

”Não se deve temer nada na vida,
apenas entender.” – Marie Curie

AGRADECIMENTOS

Aos meus pais, Eliane e Moacir, e à minha irmã, Aline, por sempre confiarem em mim, além de me apoiarem e incentivarem em todas as escolhas que faço.

Ao Rodrigo, por me motivar e ajudar nas inúmeras vezes que pensei em desistir, sempre tornando os meus dias mais leves, divertidos e significativos.

Aos demais membros da minha família, por sempre torcerem por mim e celebrarem as minhas conquistas.

Ao meu orientador, Lucas, por me ouvir e me guiar no desenvolvimento de novas ideias, além de ser compreensivo nos momentos difíceis.

À Johanna, por me permitir continuar o seu trabalho, estando sempre presente para responder as minhas dúvidas e revisar as publicações.

À UFPR, por proporcionar diversas oportunidades que me permitiram evoluir e alcançar os meus objetivos.

Ao Hospital Erasto Gaertner e aos médicos Sergio, Milton e Milena, por contribuírem para o desenvolvimento da pesquisa, além de me permitirem experienciar a rotina em um laboratório de patologia.

À todas as mulheres incríveis que abriram os caminhos na ciência para que eu pudesse estar aqui hoje.

À todas as mulheres determinadas, fortes e corajosas que estão na batalha contra o câncer de mama.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001 e do Conselho Nacional de Pesquisa (CNPq: processo 441782/2018-3).

RESUMO

Devido ao rápido desenvolvimento da tecnologia na última década, a área de histopatologia começou sua era digital com a difusão das *Whole Slide Images* (WSIs). Com este avanço, muitos estudos foram propostos para automatizar o diagnóstico de diversas patologias em imagens médicas de imuno-histoquímica, visando reduzir o tempo e esforço dos médicos em tarefas exaustivas e demoradas. Neste contexto, um progresso recente, mas expressivo, vem sendo alcançado na automatização da classificação dos níveis do câncer de mama em imagens advindas de exames com receptores de hormônio. Este progresso tem sido alcançado com técnicas de Aprendizagem de Máquina e *Deep Learning*, com Redes Neurais Profundas conhecidas pela sua grande capacidade de generalização, mas também pela necessidade de um conjunto volumoso de dados para aprender os padrões do problema. A desvantagem da aplicação destes métodos no ambiente de imagens médicas se concentra na escassez e desbalanceamento dos dados anotados por especialistas, além da falta de disponibilização pública destes dados. Assim, para mitigar estes problemas, propomos uma metodologia para investigar o impacto da adição de imagens imuno-histoquímicas sintéticas na classificação dos níveis de intensidade da coloração de células cancerosas presentes nas WSIs dos biomarcadores de receptores de hormônio. Inicialmente, treinamos uma rede generativa pertencente ao estado da arte chamada StyleGAN2ADA, e avaliamos a qualidade quantitativa e qualitativa das imagens produzidas pela rede. Na análise quantitativa, alcançamos um DreamSIM de 0,091, indicando uma excelente similaridade entre as imagens sintéticas e reais. Em relação à avaliação qualitativa, implementamos o método de Categorização Rápida de Cenas (CRC) envolvendo a participação de três patologistas. Nessa validação, os médicos não detectaram as imagens sintéticas, demonstrando o alto padrão de qualidade das imagens geradas pela rede. Para melhorar a performance da classificação dos níveis de câncer, nós conduzimos cinco experimentos com os modelos SVM, CNN, DenseNet e ViT, incorporando imagens geradas via StyleGAN2ADA e AutoAugment. Os experimentos envolveram o balanceamento de classes e a adição de uma pequena quantidade de imagens sintéticas ao processo de treinamento. Como resultado, alcançamos um *f1-score* de 87,46% no conjunto de imagens de Receptores de Estrogênio (ER) e de 92,22% nas imagens de Receptores de Progesterona (PR), atingindo uma melhora de até 14,15 pontos percentuais com a StyleGAN2DA em relação à classificação com os dados originais. Por fim, visando trazer uma maior robustez e assertividade à pesquisa, estudamos algumas possíveis melhorias futuras à metodologia proposta.

Palavras-chave: Câncer de Mama. Receptor de Estrogênio. Receptor de Progesterona. Aumento de Dados. Inteligência Artificial. Deep Learning. Modelos Generativos. Classificação. Processamento de Imagens.

ABSTRACT

Due to the rapid development of technology in the last decade, the field of histopathology began its digital era with the dissemination of Whole Slide Images (WSIs). With this advancement, many studies have been proposed to automate the diagnosis of various pathologies in immunohistochemical medical images, aiming to reduce the time and effort of doctors in exhaustive and time-consuming tasks. In this context, recent but significant progress has been made in automating the classification of breast cancer levels in images from exams with hormone receptors. This progress has been achieved with Machine Learning and Deep Learning techniques, specifically Deep Neural Networks which are known for their great generalization capacity and need for a voluminous set of data to learn the problem patterns. The disadvantage of applying these methods in the medical imaging environment focuses on the scarcity and imbalance of data annotated by specialists, in addition to the lack of public availability of these data. Therefore, to mitigate these problems, we propose a methodology to investigate the impact of adding synthetic immunohistochemical images on classifying the color intensity levels of cancer cells present in the WSIs of hormone receptor biomarkers. Initially, we trained a generative network belonging to the state-of-the-art called StyleGAN2ADA, and we evaluated the quantitative and qualitative quality of the images produced by the network. In a quantitative analysis, we achieved a DreamSIM of 0.091, indicating an excellent similarity between the synthetic and real images. About qualitative assessment, we implemented the Rapid Scene Categorization (RSC) method involving the participation of three pathologists. In this validation, doctors did not detect the synthetic images, demonstrating the high-quality standard of images generated by the network. To improve the performance of cancer classification, we conducted five experiments using SVM, CNN, DenseNet, and ViT models, incorporating images generated via StyleGAN2ADA and AutoAugment. The experiments involved class balancing and adding a small amount of synthetic images to the training process. As a result, we achieved an f1-score of 87.46% in the Estrogen Receptors (ER) dataset and 92.22% in the images of Progesterone Receptors (PR), reaching an improvement of up to 14.15 percentage points with StyleGAN2DA compared to the classification with the original data. Finally, aiming to bring greater robustness and assertiveness to the research, we studied some possible future improvements to the proposed methodology.

Keywords: Breast Cancer. Estrogen Receptor. Progesterone Receptor. Data Augmentation. Artificial Intelligence. Deep Learning. Generative Models. Classification. Image Processing.

LISTA DE FIGURAS

LISTA DE TABELAS

Lista de Acrônimos

AC-GAN	*Auxiliary Classifier Generative Adversarial Network*
ADA	*Adaptive Discriminator Augmentation*
AdaIN	*Adaptive Instance Normalization*
ADNI	*Alzheimer 's Disease Neuroimaging Initiative*
AGGrGAN	*Aggregation of GAN*
AUPRC	*Area Under the Precision-Recall Curve*
AUROC	*Area Under the Receiver Operator Characteristic*
CLAHE	*Contrast Limited Adaptive Histogram Equalization*
CLBP	*Completed Local Binary Pattern*
CNN	*Convolutional Neural Network*
CRC	*Categorização Rápida de Cenas*
DAB	*Diaminobenzidine*
DCGAN	*Deep Convolutional Adversarial Network*
DCIS	*Ductal Carcinoma In situ*
DenseNet	*Densely Connected Convolutional Network*
DRC	*Doença Renal Crônica*
eGFR	*Estimated Glomerular Filtration Rate*
ER	*Estrogen Receptor*
FID	*Fréchet Inception Distance*
GAN	*Generative Adversarial Network*
GLCM	*Grey Level Co-occurrence Matrices*
Guided Grad-CAM	*Guided Gradient-weighted Class Activation Mapping*
H	*Hematoxilina*
H&E	*Hematoxilina e Eosina*
HER-2	*Human Epidermal growth factor Receptor-type 2*
HRP	*Horseradish Peroxidase*
HSV	*Hue-Saturation-Value*
IA	*Inteligência Artificial*
IARC	*International Agency for Research on Cancer*
IDC	*Invasive Ductal Carcinoma*
IHQ	*Imuno-histoquímica*
ILC	*Invasive Lobular Carcinoma*
IoU	*Intersection over Union*
IS	*Intensity Score*
KID	*Kernel Inception Distance*
KNN	*K-Nearest Neighbor*
LBP	*Local Binary Pattern*
LoRA	*Low-Rank Adaptation*
MLP	*Multilayer Perceptron*
MMD	*Maximum Mean Discrepancy*
MSE	*Mean Squared Error*
NET	*Neuroendocrine Tumor*
PET	*Positron Emission Tomography*

PFTAS	*Parameter-free Threshold Adjacency Statistic*
PPL	*Perceptual Path Length*
PR	*Progesterone Receptor*
ProGAN	*Progressive Growing Generative Adversarial Network*
PS	*Proportion Score*
PSNR	*Peak Signal to Noise Ratio*
RGB	*Red-Green-Blue*
RSC	*Rapid Scene Categorization*
SSIM	*Structural Similarity Index*
StyleGAN	*Style Generative Adversarial Network*
SVM	*Support Vector Machine*
TC	*Tomografia Computadorizada*
VGG	*Visual Geometry Group*
ViT	*Vision Transformer*
WGAN	*Wasserstein GAN*
WHO	*World Health Organization*
WSI	*Whole-Slide Imaging*

SUMÁRIO

1 INTRODUÇÃO

Câncer é um termo genérico para definir um grande grupo de doenças que podem afetar qualquer parte do corpo. Uma de suas características é a rápida criação de células anormais que crescem além de seus limites habituais e podem se espalhar para outras regiões do corpo. O espalhamento generalizado dessas células anormais é o principal fator da causa de morte por câncer (WHO, 2022). De acordo com a *World Health Organization* (WHO), o câncer é uma das principais causas de morte em todo o mundo, alcançando quase 10 milhões de óbitos em 2020 (WHO, 2022). No mesmo ano, a incidência global de câncer ultrapassou 19 milhões. No Brasil, são esperados 704 mil novos casos de câncer para o triênio de 2023-2025, sendo o câncer de mama o mais incidente (Santos et al., 2023).

De acordo com os dados da *International Agency for Research on Cancer* (IARC) em 2020, o câncer de mama ocupou a primeira posição ao considerar a incidência de todas as categorias de câncer (IARC, 2023). Cerca de 2,3 milhões de novos casos da doença foram estimados para o ano de 2020 em todo o mundo, representando cerca de 24,5% de todas as neoplasias diagnosticadas em mulheres (INCA, 2022). No Brasil, foram estimados 66.280 novos casos de câncer de mama em 2021 (INCA, 2022), sendo a doença com maior letalidade dentre todos os tipos de câncer, atingindo 18.032 óbitos em 2020 (INCA, 2020). Em um panorama mundial, a taxa de mortalidade da doença ocupa o segundo lugar, antecedida apenas pelo câncer de pulmão (IARC, 2023).

O câncer de mama surge principalmente nas células de revestimento dos ductos ou lóbulos do tecido glandular da mama. Com o tempo, a doença pode progredir e invadir o tecido mamário circundante, além de poder se espalhar para os gânglios linfáticos próximos ou para outros órgãos do corpo. O tratamento do câncer de mama pode ser altamente eficaz, principalmente quando a doença é identificada precocemente (WHO, 2021). Como forma de prevenção, muitas campanhas estão sendo feitas, como o Outubro Rosa, que é um movimento de caráter internacional com o objetivo de aumentar a conscientização sobre o câncer de mama, incentivando o auto-exame e dando acesso para os serviços de diagnóstico e tratamento (INCA, 2022).

Para chegar a um diagnóstico, os pacientes são submetidos a exames de imagens e biópsias, com o objetivo de categorizar o tipo da doença e determinar se o tecido é canceroso ou benigno (WHO, 2021). A amostra recolhida pela biópsia é analisada em diversas etapas no processo chamado de *Imuno-histoquímica* (IHQ), que é um método de coloração para detectar e localizar um antígeno a partir da interação dele com seu anticorpo (Rogalsky, 2021). Na etapa final, o patologista chega a uma conclusão, preenchendo o laudo do paciente (Kim et al., 2016).

Uma das formas de realizar a biópsia para detectar o câncer de mama com IHQ é por meio dos biomarcadores *Estrogen Receptor* (ER) e *Progesterone Receptor* (PR). Esses receptores são proteínas que estão dentro ou sobre as células que podem se ligar a certas substâncias no sangue (American Cancer Society, 2021c). Quando as células são cancerígenas, elas apresentam uma superexpressão desses receptores, recebendo altas quantidades dos hormônios de progesterona e estrogênio, o que contribui para o crescimento celular descontrolado (Yip e Rhodes, 2014). Após a lâmina de IHQ ser corada, uma das formas de pontuar a expressão desses receptores é por meio do método *Allred*. Esse método é uma soma das pontuações *Proportion Score* (PS) (indicando a proporção relativa de células cancerosas na imagem) e *Intensity Score* (IS) (avaliando a intensidade das células) (Mouelhi et al., 2018).

A pontuação PS pode variar de 0 a 5, enquanto a pontuação IS apresenta os valores 0, 1+, 2+ e 3+, dependendo do quanto os receptores se destacam no tecido analisado (Rogalsky, 2021). Cada um desses valores pode ser visto como uma classe associada manualmente por um patologista a cada amostra de tecido. Dessa forma, a pontuação depende da experiência e formação profissional dos patologistas ou histopatologistas. Como consequência, essa tarefa pode ser tediosa e consumir muito tempo, o que pode causar cansaço nos profissionais, levando à erros e diagnósticos incorretos (Han et al., 2017).

Com o rápido desenvolvimento da tecnologia na última década, a área de patologia passou por diversas mudanças ao começar sua era digital, proporcionada pela geração das *Whole-Slide Images* (WSIs). Assim, houve um esforço recente de pesquisadores da área de computação para automatizar o diagnóstico, visando auxiliar na geração do laudo das imagens de IHQ (Laurinavicius et al., 2016). Neste contexto, surgiram muitas pesquisas na área de classificação automática, para categorizar de forma mais específica as características da doença, e na área de segmentação, para determinar a localização das células cancerígenas nas imagens de IHQ (Mouelhi et al., 2018; Cordeiro et al., 2018; Rogalsky et al., 2021; Mridha et al., 2022; Rmili et al., 2022).

Uma etapa essencial anterior à aplicação destes métodos é a implementação de técnicas de processamento de imagens e de visão computacional. A aquisição de imagens médicas pode ser afetada por problemas como ruído, desfoque, pouco contraste e nitidez, o que pode complicar as etapas seguintes de diagnóstico. Assim, para cada problema, existe uma sequência de pré-processamentos que podem contribuir para melhorar a imagem, como as técnicas de realce de contraste, aumento de nitidez e borramento, por exemplo (Vasuki et al., 2017). Além de corrigir esses problemas, as técnicas podem reduzir a dimensionalidade dos dados ao extrair características das imagens. Essas características podem ser o formato, textura ou a contagem dos elementos de interesse, sintetizando os componentes mais relevantes encontrados na imagem (Vasuki et al., 2017).

Com imagens de boa qualidade se torna possível aplicar os métodos de segmentação e classificação. Para ambos métodos, se tornou comum nas pesquisas recentes o uso de modelos advindos das áreas de Aprendizado de Máquina e *Deep Learning* (Cordeiro, 2019; Rogalsky, 2021; Mridha et al., 2022). O Aprendizado de Máquina é uma subárea de Inteligência Artificial em que um modelo pode alcançar melhores resultados ao realizar um treinamento. Uma das formas de realizar esse treinamento é por meio da Aprendizagem Supervisionada, que é o caso da classificação, em que cada imagem será associada à uma classe (manualmente) e, a partir das características extraídas da imagem e das saídas esperadas, o algoritmo aprenderá a dizer a qual classe uma imagem pertence (Cunningham et al., 2008). Já o *Deep Learning* é um ramo do Aprendizado de Máquina que se baseia na definição das Redes Neurais Artificiais com muitas das chamadas “camadas ocultas”. No caso da classificação, essa técnica é capaz de encontrar as características e realizar a pontuação das classes de forma automática (Cunningham et al., 2008).

Na área de classificação, alguns exemplos de pesquisas são os trabalhos de Cordeiro (2019) e Rogalsky (2021). Em Cordeiro (2019), foram utilizados conjuntos de imagens de IHQ do tipo *Human Epidermal growth factor Receptor-type 2* (HER-2) da mama, visando classificar os diferentes níveis deste câncer. Essa classificação foi realizada com os modelos mais recentes de Aprendizado de Máquina e *Deep Learning*. Já em Rogalsky (2021), as imagens de IHQ foram extraídas de pacientes avaliados por dois biomarcadores de receptores de hormônio, compartilhando o objetivo de classificar os diferentes estágios da doença. Neste trabalho, houve um enfoque em técnicas de pré-processamento para que características fossem extraídas das imagens e entregues a um modelo de Aprendizado de Máquina.

Considerando a área de segmentação, as pesquisas se dividem em trabalhos com o foco em técnicas de processamento de imagens clássicas e os novos modelos de *Deep Learning*. Em Mouelhi et al. (2018), os autores utilizaram imagens de biomarcadores de receptores de hormônio com o objetivo de segmentar as células cancerígenas com técnicas clássicas de processamento de imagens. Já em Signaevsky et al. (2019), os autores construíram um conjunto de dados com imagens de IHQ de autópsia cerebral e segmentaram as regiões de interesse utilizando um modelo de *Deep Learning*.

Com o atual interesse dos pesquisadores na área de análise de imagens médicas de IHQ, um progresso expressivo foi alcançado, mas ainda existem desafios. Conjuntos de imagens médicas costumam ser limitados, com pouca variabilidade e com um alto desbalanceamento entre classes (Mukherkjee et al., 2022). Recentemente, essas limitações estão sendo superadas com técnicas de Aumento de Dados, tanto com métodos advindos da área de processamento de imagens quanto com a geração de imagens sintéticas proporcionadas pelas *Generative Adversarial Network*s (GANs) (Osuala et al., 2023). As GANs são arquiteturas de redes neurais profundas que aprendem a gerar novas imagens a partir de um treinamento realizado em uma amostra do conjunto de dados (Osuala et al., 2023).

Ao longo dos anos, a GAN vem recebendo melhorias em sua arquitetura, resultando em novas nomenclaturas, como *Deep Convolutional Adversarial Network* (DCGAN), *Wasserstein GAN* (WGAN), *Style Generative Adversarial Network* (StyleGAN) e StyleGAN2*ADA* (Radford et al., 2016; Arjovsky et al., 2017; Karras et al., 2019, 2020a). Em Islam e Zhang (2020), foi proposto um método de geração de imagens médicas sintéticas utilizando uma DCGAN. Os autores treinaram uma DCGAN com imagens *Positron Emission Tomography* (PET) de cérebro e geraram imagens de cada uma das classes referentes ao nível de comprometimento cognitivo. Ao realizar a classificação com a ajuda dessas imagens sintéticas, os autores conseguiram melhorar os resultados em 10 pontos percentuais em relação à classificação sem nenhum tipo de Aumento de Dados.

Tendo em vista este cenário, trabalhamos com imagens de imuno-histoquímica de câncer de mama advindas de dois exames distintos e definimos dois objetivos de pesquisa principais. Como primeiro objetivo, buscamos descobrir a capacidade de um modelo generativo do estado da arte, chamado StyleGAN2ADA, de gerar imagens celulares sintéticas de qualidade. Nesta etapa, utilizamos métricas de distância para uma avaliação quantitativa, e também convidamos três patologistas para avaliarem qualitativamente se os dados sintéticos são similares à realidade. No segundo objetivo, propomos quatro formas de classificar automaticamente os níveis de câncer das imagens e investigamos o impacto da adição de dados sintéticos ao conjunto de treinamento dos classificadores. Para atingir esses objetivos, exploramos os seguintes temas:

1. A avaliação quantitativa das imagens imuno-histoquímicas de Receptores de Estrogênio (ER) geradas pela rede StyleGAN2ADA, aplicando uma métrica de similaridade pertencente ao estado da arte (DreamSIM);

2. A avaliação qualitativa das imagens sintéticas produzidas pela StyleGAN2ADA a partir do conjunto de imagens de Receptores de Estrogênio (ER), contando com a participação de três patologistas em um experimento de Categorização Rápida de Cenas (CRC);

3. As vantagens e desvantagens entre um modelo de Aumento de Dados baseado em técnicas simples de processamento de imagens (AutoAugment) e um modelo derivado da área de redes neurais generativas (StyleGAN2ADA);

4. A implementação de diferentes classificadores para descobrir a pontuação de intensidade de coloração (IS) das imagens de Receptores de Estrogênio (ER) e Progesterona (PR),

como o *Support Vector Machine* (SVM), a *Convolutional Neural Network* (CNN), a *Densely Connected Convolutional Network* (DenseNet) e o *Vision Transformer* (ViT);

5. A avaliação com métricas gerais e métricas específicas por classe de quatro classificadores, um pertencente a área de Aprendizado de Máquina (SVM), dois à área de redes neurais convolucionais (CNN e DenseNet) e o último baseado na ideia de *Transformers* (ViT);

6. A análise do impacto do Aumento de Dados na classificação das imagens dos biomarcadores ER e PR;

7. O estudo das possíveis melhorias futuras para garantir uma maior robustez e confiabilidade da metodologia proposta.

2 FUNDAMENTAÇÃO TEÓRICA

Nesta seção serão dissertados os conceitos básicos para a compreensão de todas as etapas da pesquisa. Inicialmente, serão apresentadas as definições e os métodos de detecção da patologia estudada (seções 2.1 e 2.2), além do sistema de pontuação indicativo do estágio da doença (seção 2.3); em seguida serão brevemente descritos os pré-processamentos, com foco em imagens de histopatologia (seção 2.4); então serão explicadas as técnicas de extração de características (seção 2.5) e os modelos de classificação (seção 2.6); logo após serão apresentados os principais métodos de aumento de dados, com técnicas simples de processamento de imagens (seção 2.7) e com as GANs (seção 2.8) e suas variações (seções 2.9, 2.10 e 2.11) e, por fim, serão descritas as métricas de avaliação, tanto de classificação (seção 2.12) quanto de geração de imagens (seção 2.13).

2.1 CÂNCER DE MAMA

O câncer é uma doença que pode começar em quase qualquer órgão ou tecido do corpo, se caracterizando pelo crescimento descontrolado de células anormais que podem se espalhar para outras partes do corpo (WHO, 2023). Normalmente, as células humanas crescem e se multiplicam (processo chamado de divisão celular) para formar novas células quando o corpo necessita. Conforme essas células envelhecem ou se danificam, elas são substituídas pelas células novas. Às vezes, esse processo ordenado falha e células anormais ou danificadas crescem e se multiplicam quando não deveriam. Essa falha pode resultar na formação de tumores (Figura 2.1), que são amontoados de tecido que podem ser cancerígenos (malignos) ou não (benignos). Tumores cancerígenos se espalham para tecidos próximos e podem viajar através do sangue ou do sistema linfático para locais distantes no corpo e formar novos tumores (processo chamado de metástase), como mostrado na Figura 2.2 (National Cancer Institute, 2021).

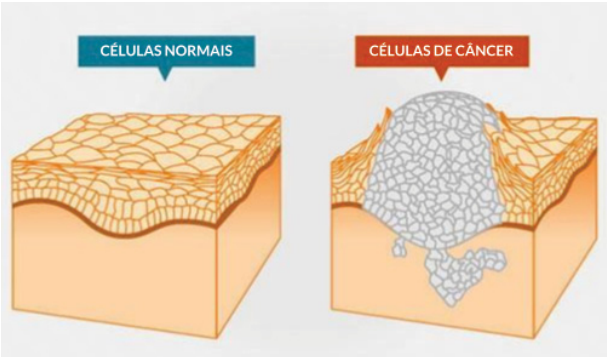


Figura 2.1: Comparação entre as células saudáveis (normais) e as células de câncer. As células normais possuem um crescimento controlado e permanecem nas regiões que deveriam para cumprir suas funções. Já as células cancerígenas apresentam um crescimento descontrolado e se espalham para os tecidos nas proximidades. Fonte: adaptado de National Cancer Institute (2021).

Dentre os tipos de câncer, o câncer de mama é o mais incidente atualmente e ele pode surgir em diferentes partes da mama. Os principais componentes da mama são: os lóbulos, que são as glândulas que produzem o leite materno; os dutos, que são pequenos canais que saem dos lóbulos e levam o leite até o mamilo; o mamilo, que é a abertura na pele onde os dutos se unem para que o leite possa sair da mama; o estroma, ou tecido conjuntivo, que envolve os dutos e

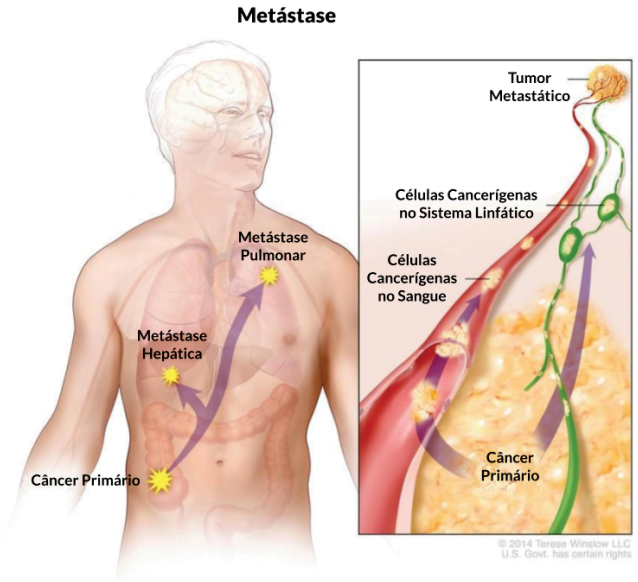


Figura 2.2: Exemplo do processo de metástase. Inicialmente, o câncer se originou no intestino, mas as células cancerígenas se espalharam para outros locais do corpo, como o fígado e o pulmão. O trajeto para esses outros órgãos pode ter ocorrido pelos vasos sanguíneos ou pelo sistema linfático. Fonte: adaptado de National Cancer Institute (2021).

lóbulos, ajudando a mantê-los no lugar e os vasos sanguíneos e linfáticos (Figura 2.3) (American Cancer Society, 2021c). Os cânceres de mama mais comuns ocorrem nos dutos ou nos lóbulos, sendo chamados de adenocarcinomas ("adeno" que significa glândula e "carcinoma" que indica um câncer epitelial) (American Cancer Society, 2021c).

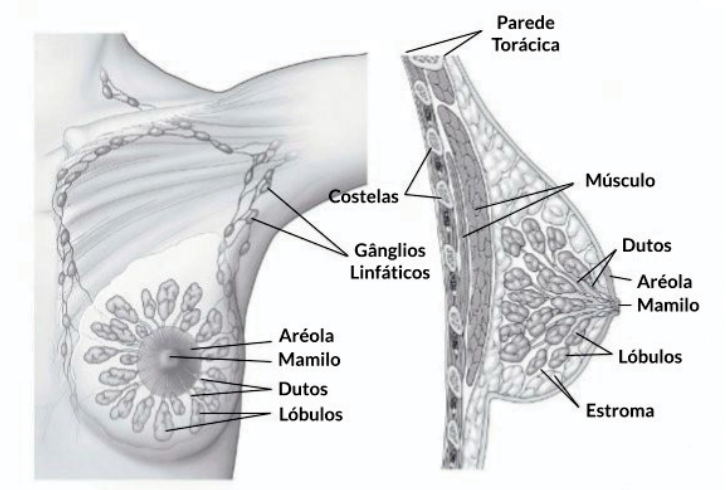


Figura 2.3: Componentes da mama em uma pessoa saudável. Fonte: adaptado de American Cancer Society (2021c).

Os tipos de câncer de mama são determinados por células específicas na mama que se tornam câncer. Muitos cânceres são carcinomas, que são tumores que começam nas células epiteliais que revestem os órgãos e tecidos do corpo. Quando os carcinomas se formam na mama, eles geralmente iniciam nas células dos dutos (carcinoma ductal) ou nos lóbulos (carcinoma lobular). O tipo de câncer de mama pode também se referir ao fato de ele se espalhar ou não. O Carcinoma Ductal *In situ* ou *Ductal Carcinoma In situ* (DCIS) é um pré-câncer que começa em um duto de leite e não cresce no resto do tecido mamário. Já o termo câncer de mama invasivo é usado para descrever qualquer tipo de câncer de mama que se espalhou para os tecidos mamários

próximos. Os tipos mais comuns são o Carcinoma Ductal Invasivo ou *Invasive Ductal Carcinoma* (IDC) e o Carcinoma Lobular Invasivo ou *Invasive Lobular Carcinoma* (ILC), sendo o IDC representante de cerca de 70-80% de todos os cânceres de mama (American Cancer Society, 2021b).

Quando uma suspeita de câncer é detectada por exame de imagem ou auto-exame, é necessária uma análise do tecido mamário para determinar a extensão da disseminação e caracterizar o tipo da doença. Para isso, o tecido pode ser obtido a partir de uma biópsia por agulha ou incisão cirúrgica (Rogalsky, 2021). Uma das formas de análise é por meio do teste de imuno-histoquímica para descobrir se as células possuem receptores de estrogênio (*Estrogen Receptor* (ER)) e progesterona (*Progesterone Receptor* (PR)). Os receptores são proteínas dentro ou sobre as células que podem se ligar a certas substâncias no sangue. Células mamárias normais têm receptores que se ligam aos hormônios de estrogênio e progesterona e precisam desses hormônios para que possam crescer (American Cancer Society, 2021a). A diferença para as células cancerígenas é que elas apresentam uma superexpressão desses receptores, contribuindo para o seu crescimento descontrolado (Figura 2.4) (Yip e Rhodes, 2014).

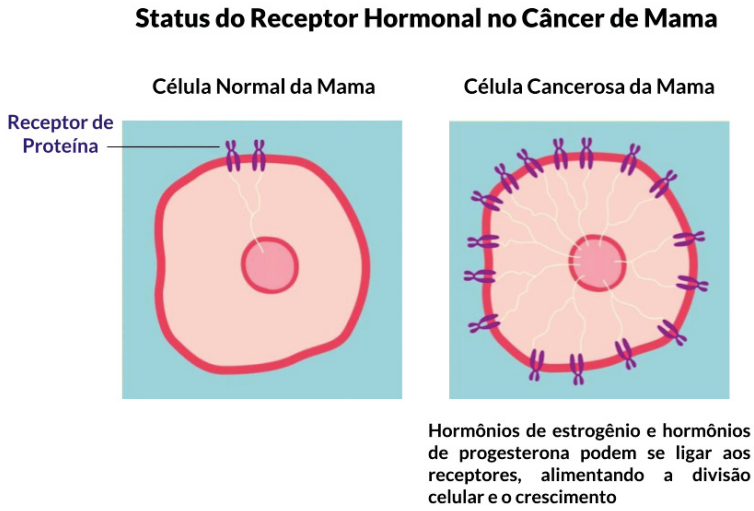


Figura 2.4: Comparação dos receptores de hormônio em uma célula saudável (esquerda) e uma célula cancerosa (direita) da mama. Na célula de câncer há uma superexpressão dos receptores de hormônio, o que contribui para o seu crescimento descontrolado. Fonte: adaptado de Stephan e Nelson (2021).

O teste de *Imuno-histoquímica* (IHQ) dirá que um tumor é positivo para os receptores hormonais se pelo menos 1% das células testadas apresentarem uma superexpressão de receptores de estrogênio e/ou progesterona. Caso contrário, o teste dirá que o tumor é negativo para receptores hormonais. Conhecer o *status* do receptor hormonal ajuda os médicos a decidir como tratar o câncer (American Cancer Society, 2021a). Vários tipos de tratamento são possíveis, como cirurgia, radioterapia, quimioterapia, terapia hormonal, imunoterapia e cuidados paliativos de suporte (Rogalsky, 2021). Se o câncer tem um ou ambos receptores hormonais em excesso, os medicamentos de terapia hormonal podem ser utilizados para diminuir os níveis do hormônio ou impedir que ele atue nas células do câncer de mama (American Cancer Society, 2021a). As terapias hormonais têm uma associação bem estabelecida com a expressão de IHQ para o *status* dos receptores de hormônio, sendo a quantificação desses receptores uma grande responsabilidade para os patologistas (Barnes et al., 2017).

2.2 IMUNO-HISTOQUÍMICA

Como o nome indica, a imuno-histoquímica é a combinação da área de histologia (estudo dos tecidos) com imunologia (estudo do sistema imunológico). A IHQ é um método de coloração que detecta e localiza um antígeno (substância estranha ao organismo), em uma amostra de tecido, através da interação com seu anticorpo (moléculas que atuam na defesa do organismo) (Weidner et al., 2009). Essa reação antígeno-anticorpo tornou a IHQ um importante método auxiliar para os patologistas, pois permitiu visualizar especificamente a distribuição e quantidade de determinadas moléculas dentro de uma amostra (Kim et al., 2016).

A Figura 2.5 apresenta uma visão geral e simplificada de um exemplo do processo de diagnóstico se baseando em métodos de imuno-histoquímica. No item 1, chamado de recepção, as amostras (um baço, no exemplo da imagem) vão ser identificadas e os dados clínicos avaliados. No item 2, com base na consistência, forma, coloração e tamanho é realizado um pré-diagnóstico (baço aumentado ou Esplenomegalia, no exemplo) e também são selecionadas as fatias de tecido mais significativas para o diagnóstico. Em 3, se inicia o processamento dos tecidos, geralmente começando com uma desidratação com o uso de álcool para remover a água da amostra (Bio-Techne, 2022). Em seguida, é feita uma diafanização ou clarificação por meio do xilol para remover o álcool dos tecidos.

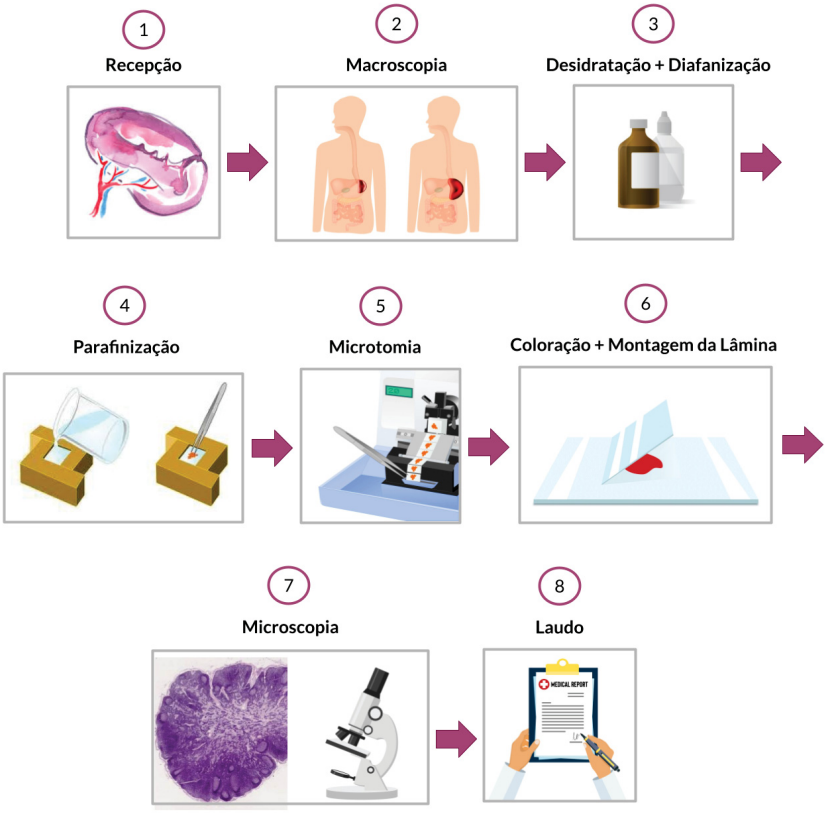


Figura 2.5: Etapas do processo geral de dignóstico com base em técnicas de imuno-histoquímica. Fonte: elaborado pela autora.

No item 4, é a fase da parafinização ou impregnação, que consiste na confecção de blocos de parafina, para manter os detalhes morfológicos e permitir uma preservação da amostra a longo prazo. Após o resfriamento da parafina, o tecido é cortado com uma lâmina afiada em fatias ultra finas utilizando um micrótomo (item 5). Cada fatia pode então passar pelo processo de coloração histoquímica e ser inserida em uma lâmina (item 6). Na fase do diagnóstico, as

lâminas são avaliadas no microscópio por um patologista (etapa chamada de microscopia, no item 7) e, a partir dessa observação, é gerado o laudo da amostra (item 8) (Bio-Techne, 2022).

Um exemplo de aplicação do método de IHQ seria utilizar blocos fixos de formalina incorporados em parafina com um sistema baseado em polímeros (Rogalsky, 2021). Com esse sistema existe a possibilidade de conjugar grandes quantidades de anticorpos, exercendo uma capacidade superior de amplificação da interação antígeno-anticorpo, pois é possível concentrar muitas moléculas propiciadoras de visualização por moléculas de antígeno (Weidner et al., 2009). Assim, a visualização amplificada acontece através das conexões de enzimas (*Horseradish Peroxidase* (HRP)), utilizadas como um segundo anticorpo, com a *Diaminobenzidine* (DAB), que é um cromogênio (substância capaz de reagir com outro composto para se converter em um pigmento ou corante). Esse processo pode ser observado na imagem da esquerda da Figura 2.6, onde a reação produziu uma coloração marrom-dourada (Weidner et al., 2009).

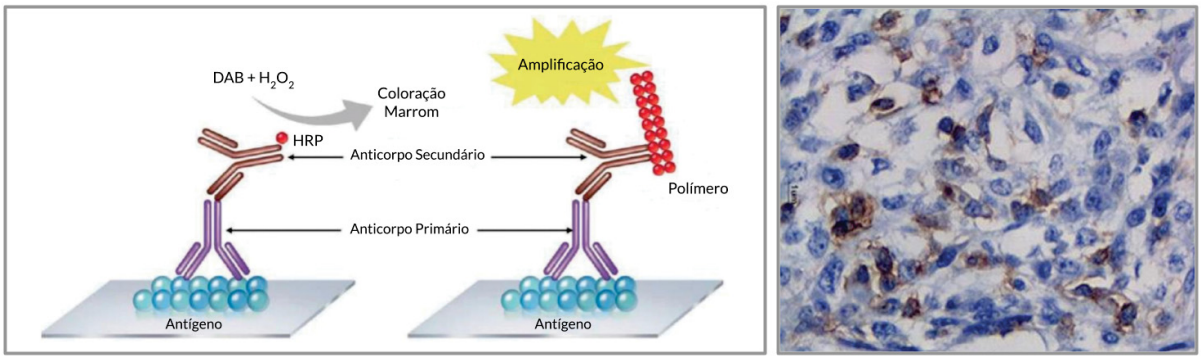


Figura 2.6: Ilustração da amplificação obtida pelo sistema de polímeros (esquerda) e exemplo de uma imagem de IHQ obtida pelo sistema ilustrado (direita). Fonte: adaptado de Weidner et al. (2009) e Magalhães et al. (2014).

Para que os patologistas possam identificar os tipos de células além da localização das células imuno-positivas destacadas pela coloração, é necessário um método de contraste. O método de contraste mais popular é a *Hematoxilina* (H), a qual dá aos núcleos das células uma coloração azul contrastando com o cromogênio (direita da Figura 2.6) (Rogalsky, 2021). Com o resultado do processo de IHQ, os patologistas avaliam as imagens e determinam o diagnóstico de acordo com pontuações quantitativas e qualitativas, como a proporção de células imuno-positivas e a avaliação de intensidade da coloração, respectivamente (Kim et al., 2016).

A fim de possibilitar o registro desses dados em um sistema de informações clínicas ou laboratoriais, as lâminas são transformadas em *Whole-Slide Images* (WSIs). As WSIs são imagens de alta resolução criadas a partir das lâminas de vidro microscópicas digitalizadas por um *scanner* próprio ou até mesmo pelo equipamento da fase de microscopia. A geração dessas imagens pode ser feita em diferentes ampliações, sendo as de 10x, 20x e 40x as mais comuns, apresentadas na Figura 2.7. Por exemplo, com uma ampliação de 40x, a resolução pode corresponder a 0,25 mícrons por pixel, o que significa que uma região de tamanho 15mm x 15mm pode corresponder a 60.000 x 60.000 pixels (Mukundan, 2017).

2.3 PONTUAÇÃO ALLRED

Uma das formas de pontuar as lâminas de imuno-histoquímica com receptores hormonais de câncer de mama é através do método *Allred*. Esse método é uma pontuação combinativa variável de 0 a 8, composto por uma proporção de células positivas coradas somada a uma intensidade média das células positivas. A proporção é uma técnica quantitativa de porcentagem relativa (chamada de *Proportion Score* (PS)) e a intensidade média é uma avaliação qualitativa da

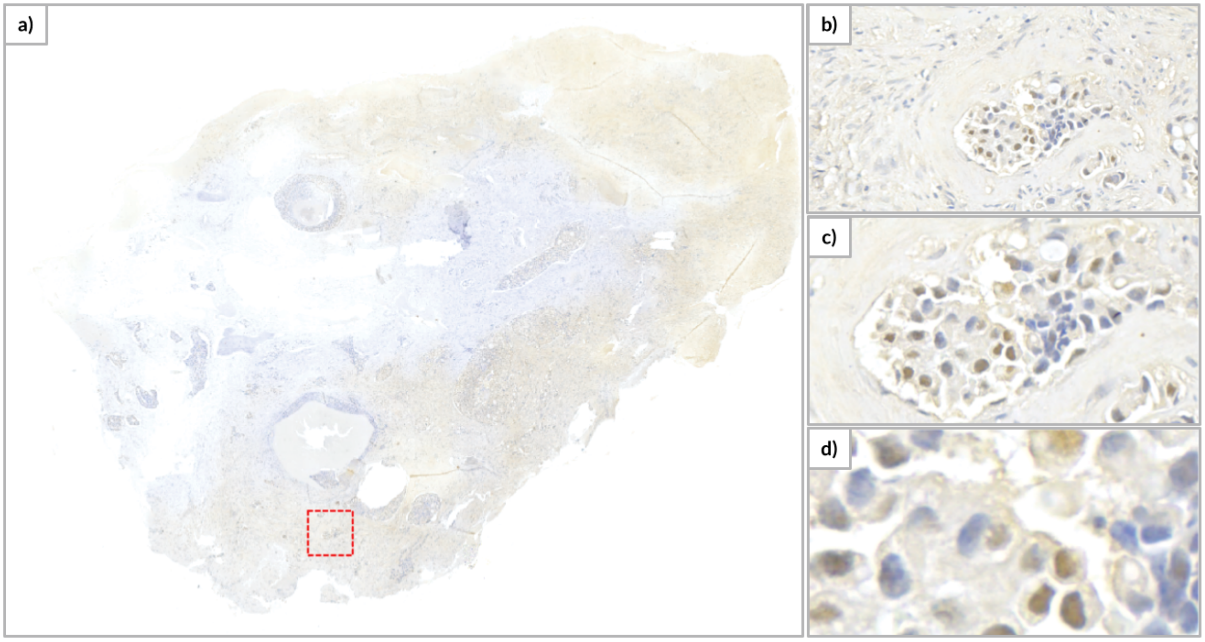


Figura 2.7: Exemplo das diferentes ampliações em uma WSI de mama corada para o Receptor de Estrogênio (ER). O retângulo vermelho indica a área em que a imagem foi aumentada. Em *a)* é apresentada a imagem original; em *b)* a imagem com a ampliação de 10x; em *c)* com ampliação de 20x e em *d)* de 40x. Fonte: elaborado pela autora.

intensidade (chamada de *Intensity Score* (IS)), indicando o quão bem os receptores se destacam após a coloração (Harvey et al., 1999; Mouelhi et al., 2018). A Tabela 2.1 apresenta o sistema de pontuação *Allred*, com as definições do PS à esquerda, do IS à direita e a soma dos dois ao final. Assim, se a lâmina apresentar uma proporção de células positivas (PS) de 5 e uma pontuação de intensidade (IS) de 2, o valor da pontuação Allred será de 7 unidades. Quanto maior a pontuação, mais receptores foram encontrados na amostra e também mais visíveis eles são (Harvey et al., 1999).

Tabela 2.1: Sistema de pontuação *Allred*. Fonte: adaptado de Rogalsky (2021).

Proporção de Células Positivas		Avaliação de Intensidade das Células	
PS	Intervalo (%)	**IS**	Tipo
0	0	0	Negativo
1	< 1	1+	Fracamente Positivo
2	1 - 10	2+	Moderadamente Positivo
3	11 - 33	3+	Fortemente Positivo
4	34 - 66		
5	> 67	Pontuação *Allred* = PS + IS	

Ainda não foram definidos guias ou escalas de intensidade padrões para definir qual pontuação de intensidade IS cada célula ou região celular deve receber. A avaliação e classificação da amostra é feita com base no visual e na gradação de cores em comparação com o tecido de controle, seguindo o histórico e experiência profissional dos especialistas. Como o PS é uma técnica quantitativa, o cálculo depende apenas da porcentagem de células cancerígenas em relação a todas as células (Rogalsky, 2021).

2.4 MÉTODOS DE PRÉ-PROCESSAMENTO

A fase de pré-processamento de imagens é responsável por preparar as imagens para os processamentos posteriores, como a extração de características e a classificação. O objetivo desta etapa se resume a melhorar as informações de uma imagem, suprimindo eventuais distorções indesejadas ou aprimorando algumas características importantes (Sonka et al., 1993). Nesta seção, serão apresentados os conceitos fundamentais desta etapa, assim como as descrições das principais técnicas utilizadas nesta pesquisa.

2.4.1 Espaço de Cores

A cor é a forma como o sistema visual humano mede uma parte do espectro eletromagnético, indo de 430 nm (cor violeta) a 790 nm (cor vermelha). Um espaço de cores é uma notação que especifica as cores na percepção humana desse espectro, sendo o espaço formado por todas as combinações dessas cores. Cada cor corresponde a um ponto dentro do espaço de cores, sendo os mais comuns na área de processamento de imagens o *Red-Green-Blue* (RGB), *Hue-Saturation-Value* (HSV) e LAB (Cordeiro, 2019; Rogalsky, 2021).

- **RGB:** é um sistema de cores aditivo no qual uma cor é igual a composição de três cores primárias, que são o vermelho, verde e azul. O espaço de cores RGB pode ser representado por um cubo em um sistema de coordenadas cartesianas, utilizando valores dentro de um intervalo de 0 a 1. Neste sistema, a cor preta é representada na origem (0, 0, 0) e o vermelho, verde e azul estão ao longo dos eixos x, y, z, como é apresentado na Figura 2.8 a). Uma cor é representada por um ponto no espaço, com coordenadas (R, G, B), e todas as cores realizáveis devem corresponder a pontos que estão dentro do cubo. Grande parte dos sistemas eletrônicos descrevem as cores no modelo RGB, mas as percepções humanas são melhores descritas por outros sistemas de cores que consideram mais diretamente a sensação subjetiva da cor (Ledley et al., 1990).

- **HSV:** é o modelo de cores mais semelhante à forma humana de percepção das cores, descrevendo-as pela tonalidade, saturação e brilho. O espaço de cores HSV é formado pelos canais de matiz (H ou *hue*), saturação (S) e valor (V). Neste modelo, a matiz é um atributo que descreve a pureza da cor; a saturação é a medida do grau em que uma cor pura é diluída pela luz branca e o valor é o brilho, indicando o quão clara ou escura é a cor. Esse espaço de cores é derivado do cubo RGB centrado em seu vértice branco. Assim, os seis vértices ao redor do centro branco formam um hexágono e especificam as cores primárias e secundárias. Geralmente, o espaço é representado por um cone, associando a matiz à um ângulo que varia de 0º a 360º, a saturação à uma distância radial e o valor ao eixo vertical, ambos variando de 0 a 1, como apresentado na Figura 2.8 b) (Gonzalez e Woods, 2008).

- **LAB:** este espaço de cores é composto por um canal para luminância (luminosidade) e outros dois canais de cores que são o a* e o b*, conhecidos como camadas de cromaticidade. Na Figura 2.8 c), o eixo vertical representa a luminância (L*), que pode assumir valores de 0 (preto) a 100 (branco), e os outros eixos variam de positivo a negativo. Valores a* positivos indicam quantidades de vermelho, enquanto valores negativos representam as quantidades de verde. De forma similar, valores b* positivos indicam quantidades de amarelo e valores negativos informam quantidades de azul. Por ser capaz de separar a luminância como um canal, este espaço de cores é muito utilizado

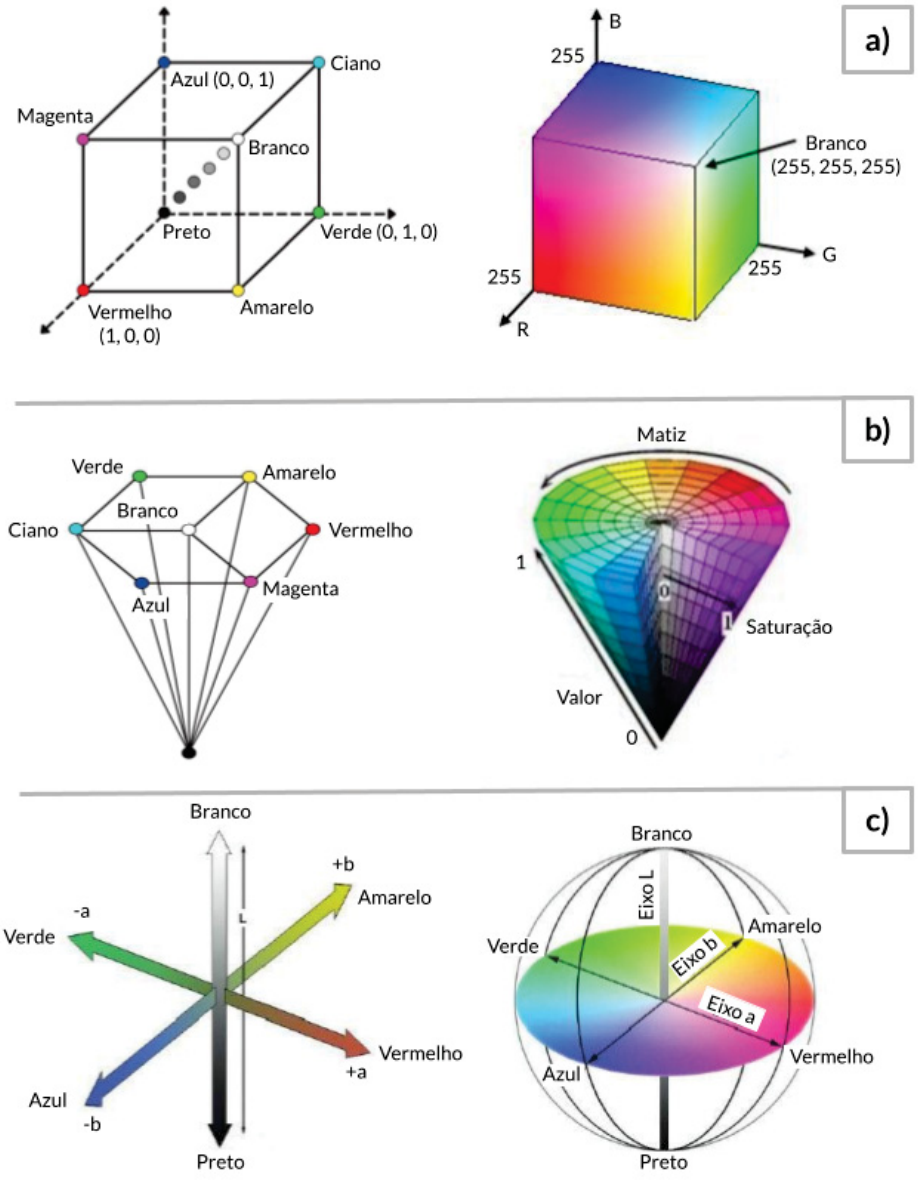


Figura 2.8: Representações dos espaços de cores RGB em *a)*; HSV em *b)* e LAB em *c)*. À esquerda são apresentadas as versões simplificadas dos espaços e na direita são apresentadas ilustrações que representam o espalhamento das cores conforme alterações são feitas em cada eixo. Fonte: adaptado de M.Ghandour e Turab (2016); Rogalsky (2021).

para técnicas de realce de contraste, ajudando a equilibrar regiões com sombras ou muita claridade nas imagens (Bora et al., 2015).

2.4.2 Realce de Contraste

O realce de contraste é um método capaz de melhorar a qualidade visual das imagens e evidenciar detalhes que podem ser importantes para as etapas posteriores de processamento. Essa melhora visual é decorrente do ajuste da diferença de brilho entre os objetos que se deseja realçar e o fundo. Nessa área, existem muitas técnicas relacionadas, sendo a equalização de histograma uma das mais utilizadas devido à sua simplicidade de implementação e resultados satisfatórios (Peng et al., 2022). Esta equalização será descrita abaixo, assim como uma de suas extensões, o método chamado *Contrast Limited Adaptive Histogram Equalization* (CLAHE).

- **Equalização de Histograma:** é uma técnica que calcula o histograma de intensidade de uma imagem e ajusta as frequências para serem bem distribuídas ao longo das classes. É um método indicado para equilibrar o contraste em imagens muito escuras ou muito claras. Por ser um método global, ou seja, aplicado à imagem como um todo, o uso da equalização de histograma pode aumentar o ruído na imagem e criar artefatos indesejados, como é possível observar na Figura 2.9 b) (Peng et al., 2022).

- **CLAHE:** é uma extensão da equalização de histograma com o objetivo de solucionar os problemas de aumento de ruído e geração de artefatos. A técnica aplica a equalização em pequenas regiões da imagem, chamadas de blocos, ao invés da imagem inteira, combinando os blocos vizinhos com interpolação bilinear. Em imagens coloridas, o CLAHE é geralmente aplicado no canal de luminância do espaço de cores LAB, conservando os canais de matiz e saturação da imagem (Yadav et al., 2014). Um exemplo é apresentado na Figura 2.9 c).

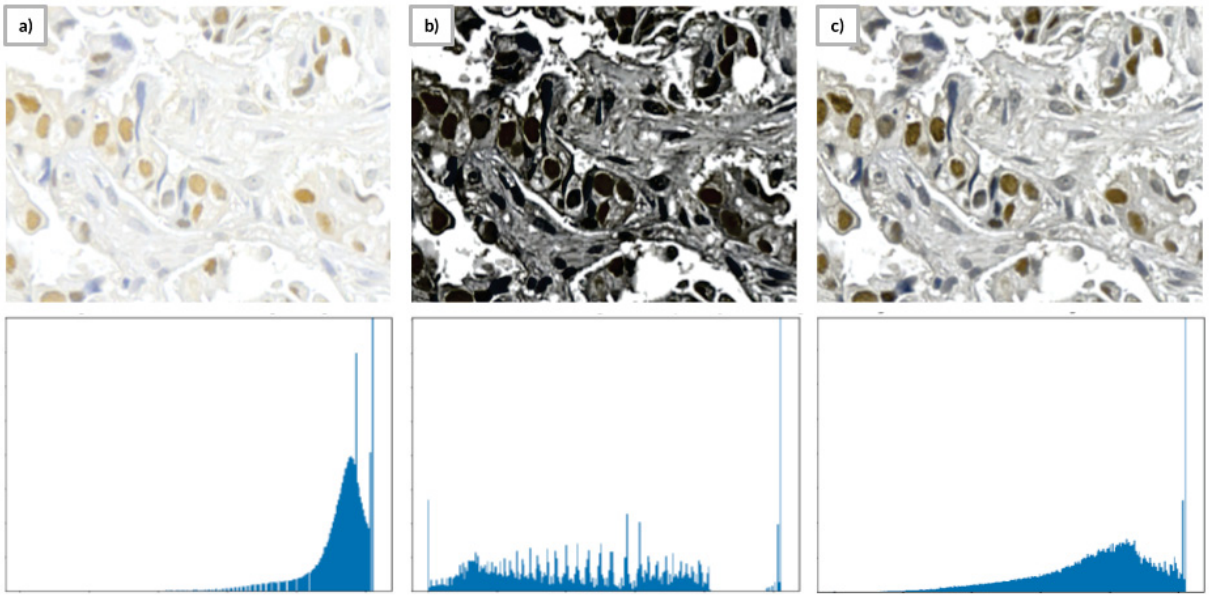


Figura 2.9: Exemplo das mudanças no histograma de intensidade de uma imagem da mama em nível celular. Em *a)* é apresentada a imagem original com seu histograma correspondente abaixo. Ao lado, em *b)*, é mostrada a imagem com realce de contraste por equalização do histograma e, em *c)*, pela técnica CLAHE. Fonte: elaborado pela autora.

2.4.3 Segmentação

A segmentação é um processo que divide uma imagem em suas partes constituintes ou objetos, chamados de segmentos. Com essa divisão, a complexidade da imagem é reduzida, permitindo o processamento ou análise adicional de cada um dos segmentos encontrados (Zaitoun e Aqel, 2015). De forma simples, é uma técnica para identificar objetos, pedestres, animais ou outros elementos importantes na imagem. Na área de imagens médicas, as técnicas de limiarização e separação por cores são essenciais para a segmentação.

- **Limiarização:** é uma técnica utilizada para binarizar uma imagem com base nas intensidades de pixel. Se a intensidade de pixel na imagem de entrada for maior do que um limite pré-definido, o pixel de saída correspondente será marcado com um valor que representa o objeto que se deseja destacar (primeiro plano). Se a intensidade do pixel de entrada for menor ou igual ao limite, a localização do pixel de saída será marcada com o valor associado ao fundo (Rogalsky, 2021). Existem muitos métodos de limiarização, sendo o *Otsu* um dos mais populares devido a sua capacidade de encontrar o valor de limite ideal de forma automática (Otsu, 1979). Um exemplo utilizando a técnica *Otsu* é apresentado na Figura 2.10.

- **Separação por Cores:** é um método que permite a segmentação de objetos dentro das imagens com base nas cores, utilizando os valores do canal de matiz do espaço de cores HSV. A matiz é um componente que está em uma escala angular, o que permite extrair a cor pura de uma imagem através da definição de um intervalo de graus (Rogalsky, 2021). Por exemplo, considerando a representação horizontal da matiz na Figura 2.11, o vermelho corresponde ao grau 0, o verde ao grau 120 e o azul ao 240. Tons de verde podem ser extraídos pelo intervalo do grau 110 ao 130. A definição de uma faixa de valores para os ângulos pode significar uma melhor separação de cores do que o uso de apenas um grau (Rogalsky, 2021).

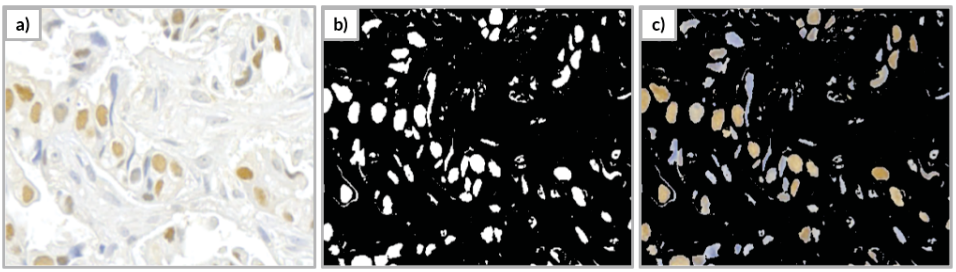


Figura 2.10: Exemplo do uso da técnica de limiarização *Otsu* em uma imagem de mama a nível celular. Em *a)* é apresentada a imagem original; em *b)* é exibida a máscara obtida pela técnica e em *c)* a segmentação alcançada com a máscara encontrada. Fonte: elaborado pela autora.

Figura 2.11: Representação horizontal do canal de matiz, com saturação e valor iguais a 1. Fonte: adaptado de Rogalsky (2021).

2.5 EXTRAÇÃO DE CARACTERÍSTICAS

A extração de características se refere ao processo de transformação dos dados brutos em características simbólicas ou numéricas que podem ser processadas, preservando as informações essenciais do conjunto de dados original (Hira e Gillies, 2015). No cenário de imagens, geralmente são definidos vetores de características d-dimensionais para representar um imagem M x N, sendo d menor do que M x N (Cordeiro, 2019). Essas características podem ser entendidas como atributos ou descrições distintivas utilizadas em problemas de reconhecimento de padrões para diferenciar e rotular cada dado. Os métodos mais comuns de extração se baseiam em características geométricas, estatísticas, de textura ou de cor (Mutlag et al., 2020).

Um exemplo de características estatísticas são os histogramas HSV, que são baseados na contagem dos valores de intensidade de cada canal (H, S e V). Por definição, um histograma é uma função de distribuição que mostra a frequência na qual as intensidades aparecem em uma imagem de nível L (equivalente ao tamanho do histograma). O histograma unidimensional é denotado pela equação 2.1 abaixo, onde r_k denota a intensidade com k indo de 0 a L-1, n_k representa o número de pixels da imagem *f(x, y)* com intensidade r_k, com as classes (ou *bins*) subdividindo a escala de intensidade. O histograma unidimensional pode ser utilizado para contabilizar as frequências de intensidades de apenas um canal de uma imagem, sendo geralmente armazenado em um vetor (Rogalsky et al., 2021).

$$h(r_k) = n_k \tag{2.1}$$

Como o cálculo é feito passando por cada pixel, é possível gerar histogramas de dois ou mais canais, armazenando os valores em matrizes ou concatenando o histograma de cada canal em um vetor. Dessa forma, quando são utilizados dois canais, o histograma bidimensional pode ser definido pela equação 2.2, onde r_k e s_l correspondem às intensidades dos dois canais, com k e l indo de 0 a L1 - 1 e L2 - 1, respectivamente. O termo $n_{k,l}$ denota o número de pixels na imagem *f(x, y)* que possuem intensidades r_k e s_l como índices da matriz na qual o histograma está armazenado (Rogalsky et al., 2021). A mesma lógica pode ser aplicada para obter histogramas com mais de 2 dimensões.

$$h(r_k, s_l) = n_{k,l} \tag{2.2}$$

2.6 CLASSIFICAÇÃO

O aprendizado supervisionado é a técnica mais comum do Aprendizado de Máquina atualmente. O nome deriva da ideia de supervisionar o processo de treinamento o tempo todo. Para realizar o treinamento, inicialmente é necessário criar o conjunto de dados, que consiste no emparelhamento das entradas com as saídas corretas. Durante o treino, o algoritmo buscará padrões nas entradas que se correlacionam com as saídas desejadas. Após o processo de aprendizagem, o algoritmo receberá novas entradas não vistas (chamado de conjunto de teste) e determinará qual a saída mais adequada para aquela entrada (Cunningham et al., 2008). No contexto de aprendizado supervisionado existem diversos métodos, sendo um deles a classificação. Esse método consiste em um classificador que recebe um exemplo de entrada (como uma imagem, vetor ou série temporal) e tem como objetivo atribuir a classe correspondente a esse exemplo. Durante o processo de treinamento, o classificador aprende a associar corretamente as classes aos exemplos fornecidos (Cunningham et al., 2008). Nesta seção, será apresentada uma técnica de organização do processo de classificação e serão descritos os quatro classificadores mais relevantes para esta pesquisa.

2.6.1 Organização do Processo de Classificação

A aplicação da técnica de *Holdout* é uma das formas mais comuns de organizar o processo de classificação. A ideia da técnica se resume a particionar os dados em 3 conjuntos, sendo a primeira partição o conjunto de treinamento, a segunda o conjunto de validação e a terceira o conjunto de teste. Assim, o classificador aprende com o conjunto de treinamento e é avaliado com o conjunto de validação durante o treinamento. Após o modelo aprender os padrões dos dados, o conjunto de teste é utilizado para descobrir o desempenho do modelo treinado (Maleki et al., 2020). Esse método é recomendado para conjunto de dados de larga escala, pois cada um dos conjuntos particionados será grande o suficiente para representar bem o problema. No caso de conjuntos de dados de tamanho reduzido, um pequeno conjunto de teste pode não fornecer um estimativa confiável do modelo (Maleki et al., 2020). Além disso, o modelo treinado pode ser sensível à escolha do conjunto de validação, que por ser pequeno, pode resultar em modelos com baixa capacidade de generalização (Maleki et al., 2020).

Para aumentar a confiabilidade dos resultados e avaliar a capacidade de generalização do modelo, a Validação Cruzada *k-fold* é a técnica mais recomendada. Neste método, o conjunto de dados é inicialmente dividido em conjunto de treino (90%, por exemplo) e conjunto de teste (10%, por exemplo). Então, o conjunto de treino é dividido em k subconjuntos mutuamente exclusivos de mesmo tamanho (chamados de pastas ou *folds*). Um desses conjuntos é utilizado como conjunto de validação e os outros compõem o conjunto de treinamento (Maleki et al., 2020). Esse processo é realizado k vezes, alternando qual *fold* será o conjunto de validação. Por fim, uma última avaliação de desempenho é realizada no conjunto de teste, como mostrado na Figura 2.12. Comparada com a técnica *Holdout*, a Validação Cruzada tende a fornecer uma estimativa mais precisa do erro de generalização alcançado pelo modelo (Maleki et al., 2020).

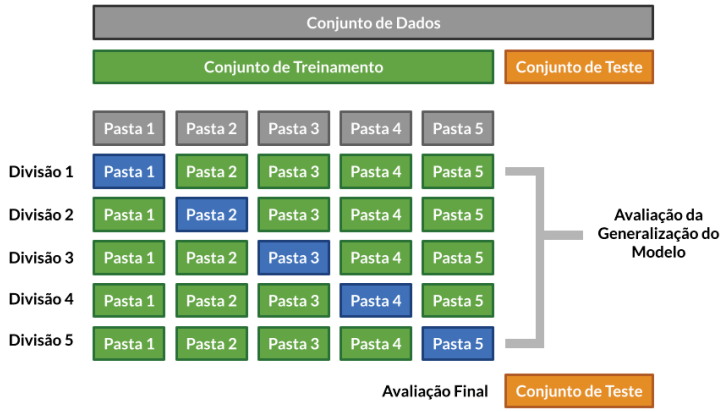


Figura 2.12: Funcionamento da Validação Cruzada de 5 pastas. As regiões em verde indicam o conjunto de treinamento, as regiões em laranja representam o conjunto de teste e as regiões em azul simbolizam o conjunto de validação. Fonte: adaptado de Maleki et al. (2020).

2.6.2 SVM

O *Support Vector Machine* (SVM) é uma estratégia de Aprendizado de Máquina inicialmente proposta para realizar a tarefa de classificação binária. A ideia principal se resume ao uso de hiperplanos, que podem ser entendidos como uma divisão do espaço euclidiano em dois, de forma que cada lado contenha dados de uma única classe (Cortes e Vapnik, 1995). O objetivo do modelo é encontrar o melhor hiperplano em um espaço n-dimensional que separe os dados para as suas classes potenciais. Para isso, o hiperplano deve ser posicionado com a maior distância possível dos dados. Além dos hiperplanos, também são definidos os vetores de suporte,

que são os dados com distância mínima ao hiperplano. Devido à sua localização próxima, a influência desses dados na posição exata do hiperplano é maior do que em outros pontos (Tan et al., 2005). Um exemplo com estes conceitos é apresentado na Figura 2.13 a).

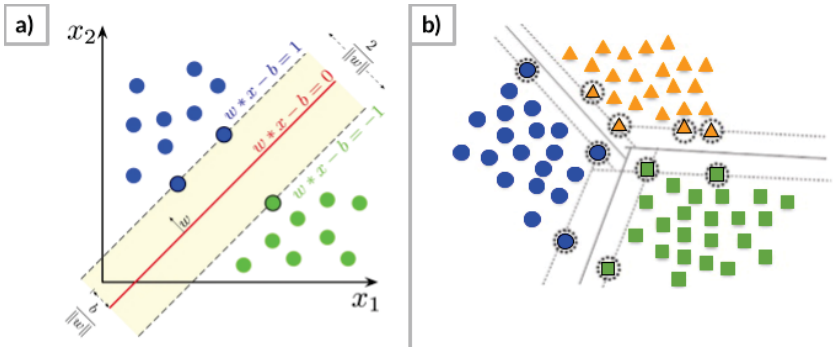


Figura 2.13: Exemplos de classificações realizadas pelo modelo SVM. Em *a)*, é apresentado um problema binário, sendo a reta central o hiperplano e as retas pontilhadas os vetores de suporte. O mesmo ocorre em *b)*, com a diferença de que são consideradas 3 classes (problema multiclasse). Fonte: adaptado de Herrero-Lopez (2011) e Pecha e Horák (2020).

Ademais, são definidos os *kernels*, que são responsáveis por converter o espaço de dados de entrada em um espaço de dimensão superior, aplicando uma função de mapeamento entre dois pontos. O uso do truque de *kernel* ajuda a transformar os dados em um mapeamento de maior dimensão, mantendo as operações no conjunto de características original, sem gastar tempo e esforço computacional em um espaço com mais dimensões. Em resumo, o truque de *kernel* oferece uma forma mais eficiente e mais econômica de transformar os dados em dimensões maiores (Bhattacharya et al., 2019). Uma ilustração desta transformação pode ser observada na Figura 2.14.

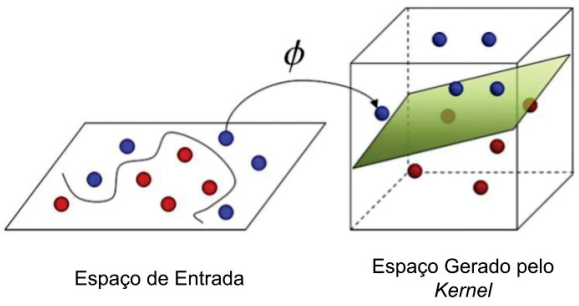


Figura 2.14: Ilustração da aplicação do *kernel* para transformar o espaço bidimensional de entrada em um espaço de 3 dimensões. Inicialmente, não há nenhuma maneira linear capaz de separar os dados. Após a aplicação do *kernel* se torna possível dividir as classes com um hiperplano. Fonte: adaptado de Bhattacharya et al. (2019).

Na sua forma mais simples, os SVMs são aplicados na classificação binária, separando os dados nas classes 0 e 1. No entanto, o princípio do modelo pode ser estendido para considerar mais de 2 classes, dividindo o problema em vários casos de classificação binária (*one-vs-one*). Nessa abordagem, cada classificador separa os dados de duas classes diferentes e, compreendendo todos os modelos, é possível obter um classificador multiclasse (Tan et al., 2005). O número de classificadores binários necessários para realizar a divisão em n classes é apresentado na Equação 2.3 (Herrero-Lopez, 2011) e um exemplo de classificação multiclasse é apresentado na Figura 2.13 b).

$$n\frac{(n-1)}{2} \tag{2.3}$$

2.6.3 CNN

A *Convolutional Neural Network* (CNN), também chamada de *ConvNet*, é um dos modelos mais conhecidos de Aprendizado Profundo, geralmente utilizada para resolver tarefas de visão computacional, principalmente classificação de imagens. Em resumo, uma CNN é um algoritmo que recebe uma entrada (considerada como imagem nesta pesquisa), atribui uma importância (com pesos e vieses que são aprendidos durante o treino) à vários aspectos do dado recebido e é capaz de diferenciar uma entrada da outra. Um dos aspectos principais deste modelo é a capacidade de aprender a encontrar uma representação a partir dos dados de forma automática (LeCun et al., 2010). Existem 3 tipos de componentes principais em uma CNN, que são as camadas de convolução, camadas de *pooling* e camadas totalmente conectadas (Spanhol et al., 2016a). Uma visão geral da arquitetura da rede é apresentada na Figura 2.15.

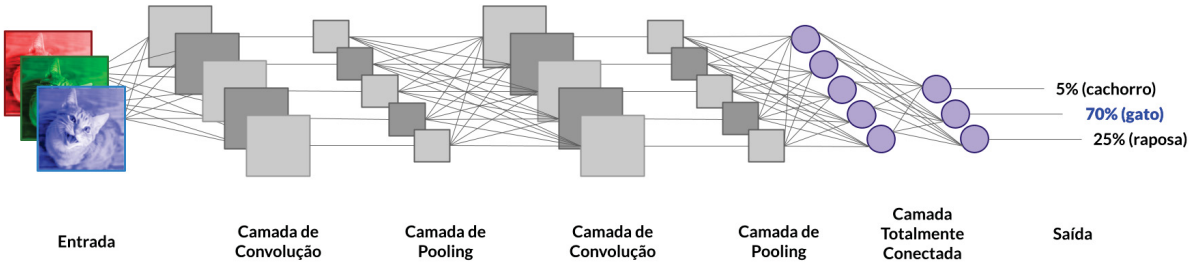


Figura 2.15: Exemplo de uma CNN para classificação de imagens de animais. Inicialmente, a rede recebe uma imagem de gato e passa por uma sequência de camadas de convolução seguidas por camadas de *pooling*. Ao final dessas camadas, a rede obtém as características principais da imagem e então realiza a classificação por meio da camada totalmente conectada. Fonte: elaborado pela autora.

- **Camada de Convolução**: um dos papéis da CNN é extrair informações das imagens de forma que seja mais fácil processá-las, sem perder características que são críticas para alcançar bons resultados. Para isso, um dos componentes responsáveis é a camada de convolução, que multiplica a imagem de entrada por diferentes filtros (pequenas matrizes, também chamados de *kernels*), delizando-os por toda a imagem para obter mapas de ativação (LeCun et al., 2010). O mapa de ativação armazena todas as características distintivas da imagem e, ao mesmo tempo, reduz a quantidade de dados a serem processados (Ajit et al., 2020). Um exemplo da operação de convolução é apresentado na Figura 2.16. As *ConvNets* podem apresentar mais de uma camada convolucional, sendo as primeiras camadas responsáveis por capturar características de baixo nível, como bordas e cores, e as camadas adicionais buscam por características de alto nível, como formas geométricas (Ajit et al., 2020).

- **Camada de *Pooling***: frequentemente posicionada após a camada convolucional, a camada de *pooling* é um passo importante para reduzir ainda mais as dimensões do mapa de ativação, mantendo apenas as características importantes (Spanhol et al., 2016a). Existem diversas variações desta técnica, como *Pooling* Máximo, *Pooling* Médio e *Pooling* Estocástico. Dentre eles, o mais popular é o *Pooling* Máximo, que seleciona o valor mais alto de cada submatriz do mapa de ativação e forma uma matriz separada a

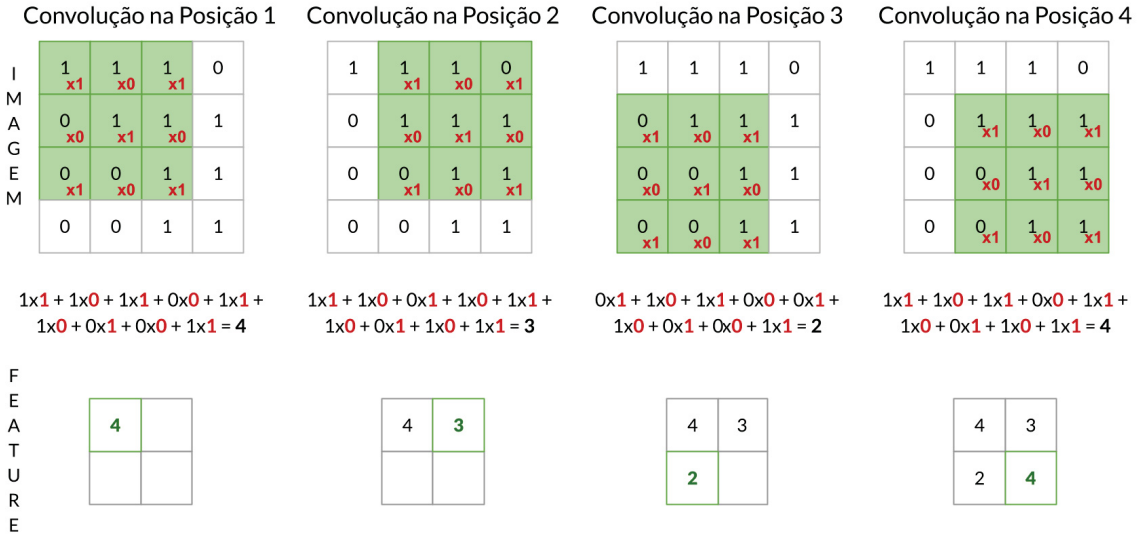


Figura 2.16: Exemplo da operação de convolução. A matriz 4x4 representa a imagem; a submatriz 3x3 (em verde) o filtro e as matrizes 2x2 as características resultantes do processo de convolução. Inicialmente, é realizado um produto escalar do filtro com as 3 primeiras linhas e 3 primeiras colunas da imagem na posição 1. O filtro vai se deslocando de 1 em 1 posição (*stride* = 1) até chegar nas 3 últimas linhas e colunas, na posição 4. Cada posicionamento do filtro produz um valor, que pode ser visto na área das *features*. Então, a imagem resultante corresponde à junção dos resultados da passagem do filtro em todas as posições. Fonte: elaborado pela autora.

partir dele. Isso garante que as características que podem ser aprendidas permaneçam limitadas em número, preservando também as principais características de qualquer imagem (Ajit et al., 2020). Um exemplo da técnica aplicada na camada de *pooling* é apresentado na Figura 2.17.

- **Camada Totalmente Conectada**: com a redução de dimensões da imagem proporcionada pelas camadas anteriores, a matriz resultante pode ser achatada em um vetor unidimensional e dada de entrada para uma Rede Neural *Multilayer Perceptron* (MLP). Essa rede consiste em camadas de neurônios conectados entre si por pesos e sinais de saída. A camada de entrada geralmente consiste em um vetor de características e a camada de saída faz alguma previsão sobre a entrada. Entre a camada de entrada e a camada de saída podem existir uma ou mais camadas de neurônios chamadas de camadas ocultas (Ponti et al., 2017). Essas camadas permitem que a rede aprenda representações mais complexas e não lineares dos dados. Na CNN, essa camada é capaz de distinguir entre características dominantes e certas características de baixo nível nas imagens e classificá-las por meio de um treinamento de várias épocas com as técnicas de *Feedforward* e *Backpropagation*. Durante essas iterações, os pesos e os vieses (*bias*) dos filtros da camada de convolução também são aprendidos com essas técnicas (Ponti et al., 2017). Um exemplo desta camada é mostrada na Figura 2.18.

2.6.4 DenseNet

Com a difusão das CNNs para solucionar uma grande variedade de problemas, muitas arquiteturas de redes foram propostas, visando aumentar a complexidade e eficiência do processo de treinamento (Szegedy et al., 2015; Srivastava et al., 2015; He et al., 2016). Simultaneamente ao desenvolvimento das melhorias, surgiram novos desafios como o *vanishing gradient*, o

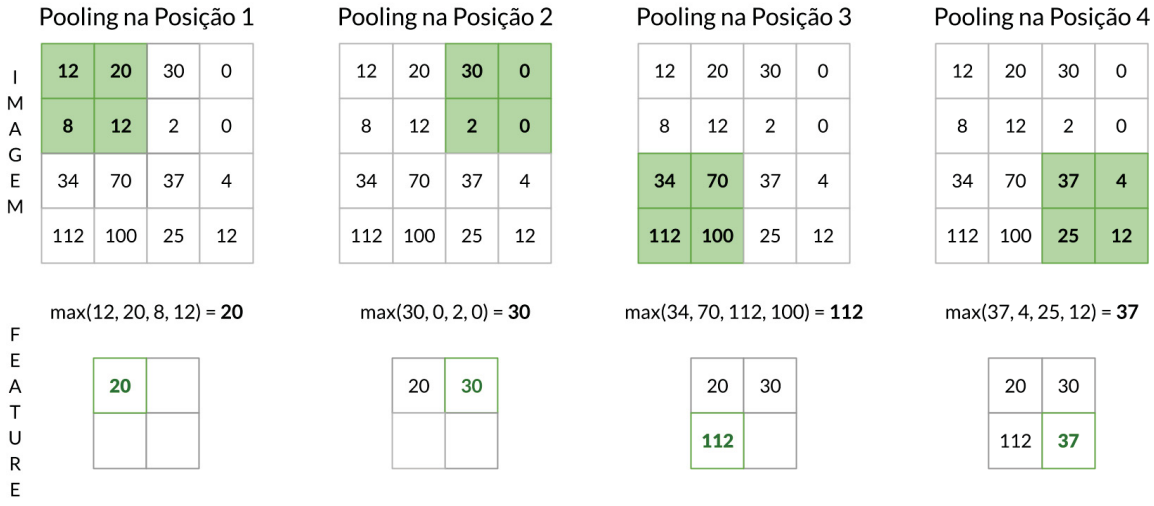


Figura 2.17: Exemplo da operação de *Pooling* Máximo. A matriz 4x4 representa a imagem; a submatriz 2x2 (em verde) o filtro e as matrizes 2x2 o resultado do processo de *pooling*. Inicialmente, o filtro é posicionado no ínicio da matriz e o valor máximo encontrado na região do filtro irá preencher a imagem resultante. O filtro vai se deslocando pela imagem, ultrapassando 2 linhas e 2 colunas por vez (*stride* = 2), até chegar ao final da imagem, na posição 4. Então, a imagem resultante corresponde à junção dos resultados da passagem do filtro em todas as posições. Fonte: adaptado de Ajit et al. (2020).

enfraquecimento da propagação de características e aumento do número de parâmetros das redes (Huang et al., 2017).

O problema do *vanishing gradient* acontece quando os gradientes utilizados para ajustar os pesos das camadas durante o *backpropagation* se tornam muito pequenos à medida que são propagados para trás pela rede. Como resultado, as camadas iniciais da rede não recebem informações suficientes para aprender, tornando o treinamento ineficaz ou muito lento. O *vanishing gradient* pode enfraquecer a transmissão das informações pelas camadas rede, resultando na baixa preservação de características importantes para a solução do problema (Huang et al., 2017).

Para lidar com estas dificuldades, os autores de Huang et al. (2017) propuseram a *Densely Connected Convolutional Network* (DenseNet). Uma das muitas inovações da DenseNet foi o uso de conexões densas entre as camadas, chamadas de *Dense Blocks*, onde todas as camadas são conectadas entre si. Dessa forma, cada camada recebe entradas adicionais de todas as camadas anteriores e passa seus mapas de características para as próximas camadas. Este processo permitiu um aprendizado mais eficiente, reduziu o número de parâmetros da rede e garantiu o reuso de características durante o treino.

A Figura 2.19 apresenta a arquitetura simplificada da DenseNet. As camadas densas são intercaladas pelas chamadas camadas de transição, responsáveis por mudar as dimensões dos mapas de características aplicando as operações de convolução e *pooling*. Por fim, os autores definiram várias arquiteturas diferentes para a DenseNet, variando a profundidade da rede, como a DenseNet-121, a DenseNet-169, a DenseNet-201 e a DenseNet-264, com 121, 169, 201 e 264 camadas, respectivamente.

2.6.5 ViT

Dentre as arquiteturas mais recentes de Aprendizado de Máquina, surgiram os modelos chamados de Transformers. Eles foram introduzidos no artigo *Attention is All You Need* (Vaswani et al., 2017) e podem ser utilizados para diversas tarefas, incluindo geração de texto, tradução

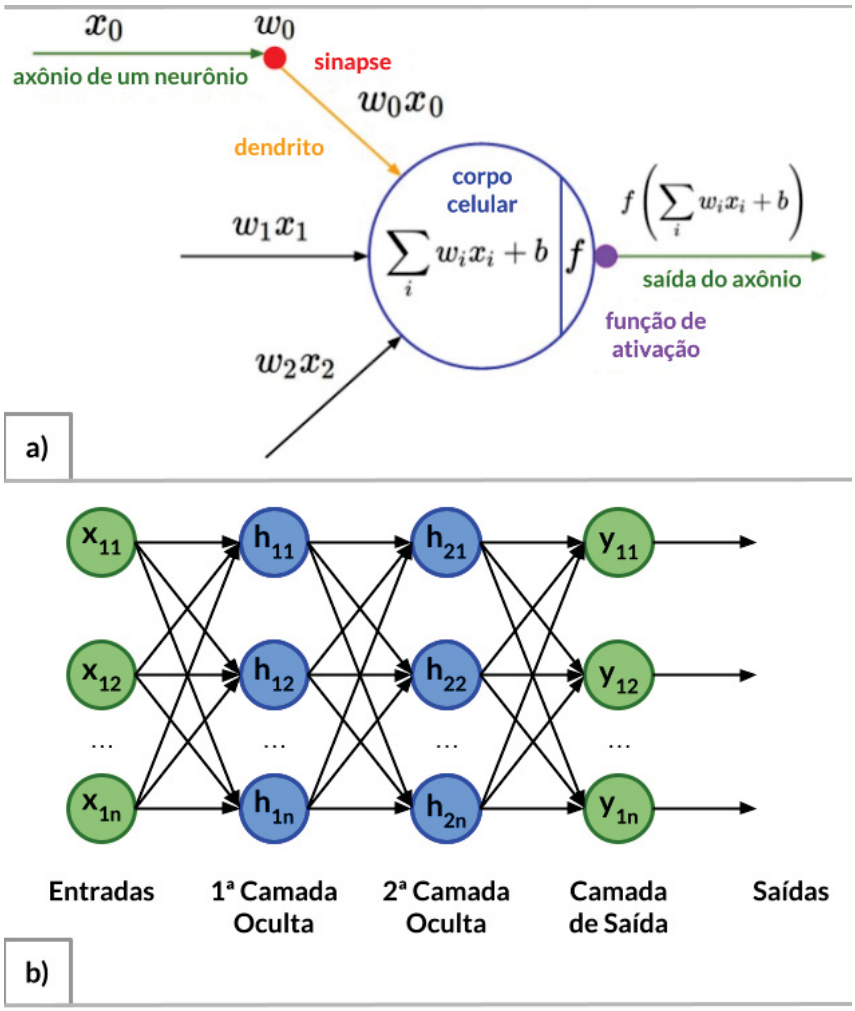


Figura 2.18: Exemplo de um neurônio e de uma arquitetura de Rede Neural com camadas totalmente conectadas. Em *a)* são apresentadas as entradas (x_n), os pesos (w_n), o *viés* ou *bias* (b) e a função de ativação (f), além da combinação linear que ocorre dentro do neurônio. Na parte *b)* da figura, são mostradas as camadas de uma Rede Neural totalmente conectada, sendo os primeiros neurônios em verde a camada de entrada, os neurônios em azul as camadas ocultas e os neurônios verdes do fim a camada de saída. Fonte: elaborado pela autora.

entre linguagens e classificação. Como uma forma de trazer uma alternativa às CNNs, visando utilizar menos recursos computacionais e acelerar o processo de treinamento, os autores de Dosovitskiy et al. (2021) propuseram uma arquitetura adaptada dos Transformers para classificar imagens ao invés de manipular textos, chamada de *Vision Transformer* (ViT).

A ideia do ViT é dividir a imagem de entrada em *patches* de tamanho constante e utilizar mecanismos de atenção para capturar características em diferentes níveis. Dessa forma, o modelo consegue capturar não só o contexto global da imagem, mas também descobrir informações sobre as relações entre os *patches*. A etapa de classificação é implementada por uma MLP com uma camada oculta durante o treinamento e por uma única camada linear durante o *fine-tuning* (Dosovitskiy et al., 2021). Abaixo são apresentadas as descrições dos componentes mais relevantes do ViT.

- **Patches e Positional Encoding:** inicialmente, a imagem de entrada é dividida em *patches* de tamanho fixo, sendo o tamanho um hiperparâmetro do modelo. Cada *patch* é transformado em um vetor que será associado a um padrão periódico com o objetivo de informar a posição do *patch* em relação a imagem de entrada. Esta fusão é chamada de *Positional Encoding* e os padrões periódicos mais utilizados são as funções de seno e

Figura 2.19: Arquitetura da rede DenseNet. Fonte: adaptado de Huang et al. (2017).

cosseno. A função desta etapa é manter a estrutura espacial da imagem e permitir que o modelo capture relações entre os *patches* em diferentes posições (Dosovitskiy et al., 2021).

- **Camadas de Self-Attention:** as camadas de atenção permitem que cada *patch* "preste atenção" em todos os outros *patches* na mesma camada, capturando dependências e relações contextuais. Essa etapa envolve três matrizes, sendo a Q (*Query*) a responsável pela busca por informações relevantes entre um determinado vetor e os demais; a K (*Key*) que é a chave que permite que Q encontre as informações relevantes e a V (*Value*) que é a resposta à Q levando em consideração o alinhamento entre Q e K (Vaswani et al., 2017).

  – Para exemplificar, no cenário de processamento de texto, se a frase "gato pequeno e amarelo" estivesse sendo analisada, Q tentaria descobrir se existem outras palavras (*patches*) que estão relacionadas com a palavra "gato"; K informaria que existem as palavras "pequeno" e "amarelo" e V teria as informações para ajustar o conhecimento inicial de "gato" para o espaço referente a um gato pequeno e amarelo.

- **Multi-head Attention:** é a extensão do conceito de *self-attention*, que permite ao modelo aprender diferentes representações contextuais em paralelo. A ideia básica é utilizar várias camadas de atenção (*head*), cada uma com seus próprios conjuntos de pesos nas matrizes Q, K e V. Nas camadas finais do modelo, os resultados de todas as camadas de atenção (*heads*) são combinados para produzir uma pontuação final de atenção. Dessa forma, o modelo consegue codificar múltiplas relações e nuances para cada *patch* (Vaswani et al., 2017).

- **Camadas de Feedforward:** as camadas de *feedforward* fornecem um processamento adicional após a camada de atenção para que o modelo aprenda padrões e relações não lineares. Inicialmente, essa estrutura consiste de uma camada linear com o objetivo de projetar os dados em um espaço de maior dimensão. Em seguida, é aplicada uma função de ativação, geralmente a ReLU, para introduzir o conceito de não-linearidade no modelo. Por fim, a estrutura conta com mais uma camada linear, visando ajustar as dimensões da saída para ser utilizada nas camadas subsequentes (Vaswani et al., 2017).

- **MLP-Head:** com o objetivo de transformar a saída das camadas anteriores (*encoder*) em uma previsão final para realizar a tarefa de classificação, o ViT conta com uma camada final chamada MLP-Head. Ela consiste de uma ou mais camadas densas (totalmente conectadas), uma função de ativação e uma projeção linear seguida da aplicação da Softmax para gerar uma distribuição de probabilidades sobre as classes (Dosovitskiy et al., 2021).

A Figura 2.20 apresenta a visão geral do ViT. A quantidade de camadas no *encoder* do ViT pode variar com base no tamanho do modelo e nos requisitos da tarefa a ser resolvida. Os

modelos ViT mais comuns são o ViT-Base (com 12 camadas de atenção e 12 de *feedforward*, com 12 *heads* em cada), o ViT-Large (com 24 camadas e 16 *heads* em cada) e o ViT-Huge (com 32 camadas e 16 *heads* em cada).

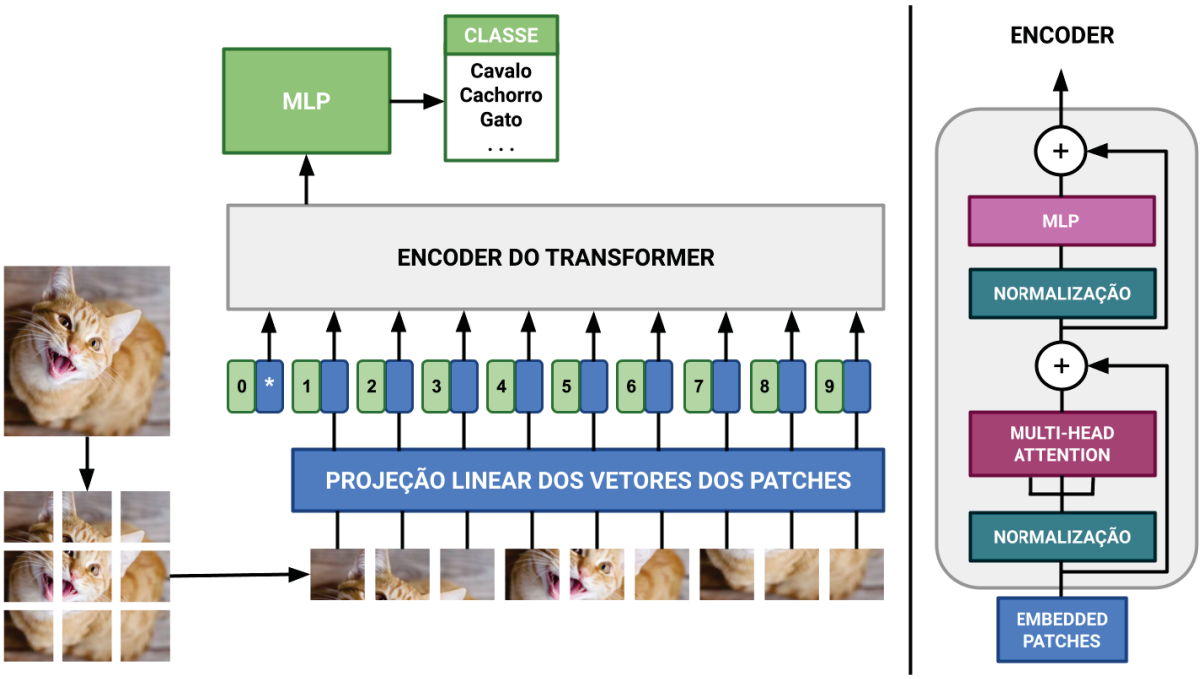


Figura 2.20: Arquitetura do modelo ViT. À esquerda, a imagem de entrada é dividida em *patches* de tamanho fixo, cada *patch* é representado de forma linear e adicionado ao *positional encoding* correpondente. A posição 0 carrega informações sobre a classe da imagem para auxiliar no processo de classificação. O resultado desta etapa é entregue ao *encoder* e a classificação é realizada por uma MLP. À direita é apresentada a arquitetura do *encoder* do Transformer, com as camadas de normalização, atenção e a MLP. Fonte: adaptado de Dosovitskiy et al. (2021).

2.7 TÉCNICAS DE AUMENTO DE DADOS

A ideia central das técnicas de Aumento de Dados é melhorar a suficiência e a diversidade dos dados de treinamento gerando conjuntos de dados sintéticos. Os dados aumentados podem ser considerados como sendo extraídos de uma distribuição muito próxima da distribuição dos dados reais. Dessa forma, o conjunto de dados aumentado pode apresentar características mais abrangentes (Yang et al., 2022). Na área de imagens, esses dados sintéticos podem ser gerados com a aplicação de diferentes transformações que adicionam algum tipo de informação nas imagens (Ruiz et al., 2019, 2020a,b).

Uma das formas mais básicas de geração de imagens sintéticas é através de manipulações simples de imagens, como as transformações apresentadas na Tabela 2.2 (Yang et al., 2022). Com tantas opções disponíveis, os autores de Cubuk et al. (2019) propuseram uma forma automática de definir qual técnica de aumento de dados utilizar, chamada de AutoAugment. O AutoAugment formula o problema de encontrar a melhor política de aumento de dados como um caso de busca discreta. Inicialmente, o espaço de busca é definido como uma amostragem das técnicas de aumento de dados, que estabelecem informações sobre qual operação utilizar, a probabilidade de aplicar esta técnica e a magnitude da operação. Então, o algoritmo de busca aplica *Reinforcement Learning* para descobrir as melhores técnicas de aumento de dados para o problema. Um exemplo das melhores transformações encontradas com a técnica no conjunto de dados ImageNet pode ser observada na Figura 2.21

Tabela 2.2: Exemplos e descrições das manipulações simples de imagens para a geração de dados sintéticos. Fonte: adaptado de Yang et al. (2022).

Métodos	**Descrição**
Inversão	Inverte a imagem horizontalmente, verticalmente ou ambos.
Rotação	Rotaciona a imagem em um determinado ângulo.
Proporção de Escala	Aumenta ou reduz o tamanho da imagem.
Injeção de Ruído	Adiciona ruído na imagem.
Espaço de Cores	Muda os canais de cores da imagem.
Equalização	Altera o contraste da imagem.
Nitidez	Modifica a nitidez da imagem.
Translação	Move a imagem horizontalmente, verticalmente ou ambos.
Corte	Corta uma sub-região da imagem.
Posterização	Reduz o número de bits para cada canal de cor.

Com técnicas simples de processamento de imagens é possível aumentar a diversidade dos dados, mas existem desvantagens. Somente se os dados gerados pertencerem a uma distribuição próxima dos dados reais que o uso destas manipulações básicas de imagens será significativo (Yang et al., 2022). Por exemplo, técnicas que rotacionam imagens de tomografia cerebral podem não contribuir para melhorar o conjunto de dados, pois essas imagens são sempre extraídas no mesmo posicionamento.

Além disso, alguns métodos básicos como translação e rotação sofrem com o efeito de preenchimento, em que algumas áreas da imagem são deslocadas para fora do limite e perdidas (Yang et al., 2022). Portanto, alguns métodos de interpolação serão utilizados para compensar a informação perdida. Esse efeito pode ser observado na primeira linha da política 1 da Figura 2.21. Por fim, durante o uso destas técnicas também pode ocorrer a movimentação do objeto de interesse para fora do limite da imagem, o que ocasionaria uma perda de informação prejudicial ao conjunto de dados resultante (Yang et al., 2022).

2.8 GAN

As Redes Adversárias Generativas (*Generative Adversarial Network*s (GANs)) são arquiteturas de redes neurais profundas compostas por duas redes que competem entre si (Data Science Academy, 2022). Essas redes podem ser vistas como participantes de um jogo *MinMax*, onde a rede geradora (G) enfrenta a outra rede chamada discriminadora (D). A rede geradora é responsável por sintetizar novas amostras e a discriminadora por predizer essas amostras, informando se elas pertecem ao conjunto de treinamento ou se foram criadas pela rede geradora (Osuala et al., 2023).

A rede discriminadora gera um valor (p) indicando a probabilidade da amostra recebida ser um exemplo de treinamento real em vez das amostras falsas da rede geradora, o que pode ser visto como uma tarefa de classificação binária (Osuala et al., 2023). O objetivo de treino de D é minimizar a função de custo da rede, enquanto o objetivo de G é o oposto, maximizar a função

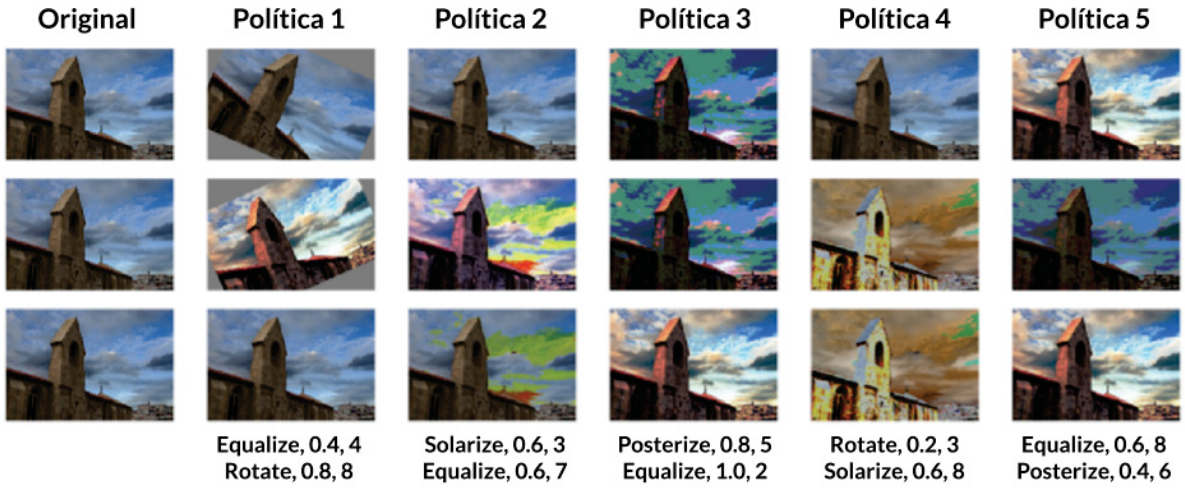


Figura 2.21: Exemplo da aplicação da técnica AutoAugment em uma imagem do conjunto de dados ImageNet. Cada coluna apresenta uma política encontrada pelo modelo e cada linha apresenta exemplos da aplicação de cada política. Cada imagem está associada a uma ou mais operações, e cada operação pode ser aplicada de acordo com uma probabilidade e magnitude encontradas durante o treinamento. Fonte: adaptado de Cubuk et al. (2019).

de custo para confundir D. O resultado é uma função de soma zero para dois jogadores (D e G) e as formulações matemáticas podem ser encontradas em Osuala et al. (2023).

A Figura 2.22 apresenta um exemplo genérico da arquitetura da GAN. Inicialmente, a entrada do gerador é um vetor aleatório (ruído) e sua saída inicial também é ruído. Com o tempo, G vai recebendo as avaliações de D e aprende a sintetizar imagens mais realistas (Data Science Academy, 2022). A rede discriminadora também melhora com o tempo, comparando as amostras geradas com as reais, tornando mais difícil para o gerador enganá-la. Em teoria, na convergência, as amostras do gerador se tornam indistinguíveis dos dados de treinamento reais e as probabilidades do discriminador se tornam 50% para qualquer amostra dada (Osuala et al., 2023).

2.9 STYLEGAN

A *Style Generative Adversarial Network* (StyleGAN) é uma extensão da arquitetura da GAN criada com o objetivo de controlar as propriedades de estilo das imagens geradas (Karras et al., 2019). Mais especificamente, os autores se basearam na arquitetura da *Progressive Growing Generative Adversarial Network* (ProGAN), que introduziu a ideia de treinamento progressivo, no qual o gerador é treinado com imagens de baixa resolução (4x4) e são adicionadas camadas de resolução maiores (*Upsampling*) com o tempo. De forma semelhante, o discriminador recebe imagens de alta resolução e reduz as dimensões (*Downsampling*) até voltar ao tamanho 4x4 (Karras et al., 2018). Assim, a ProGAN aprende inicialmente as características mais básicas das imagens de baixa resolução e incorpora mais detalhes assim que a resolução aumenta. Essa técnica acelera o treino da rede e consegue gerar imagens de alta qualidade, mas sua habilidade de controlar características específicas das imagens é bem limitada (Karras et al., 2019).

Dessa forma, a StyleGAN foi projetada para lidar com o emaranhamento de características, que é um fenômeno em que diferentes características de uma imagem estão interconectadas de uma maneira complexa e não linear, dificultando o entendimento ou controle de cada característica separadamente. Por exemplo, em imagens de faces, características como a forma do rosto, a cor dos olhos, a textura da pele e a expressão facial podem estar interconectadas de maneira

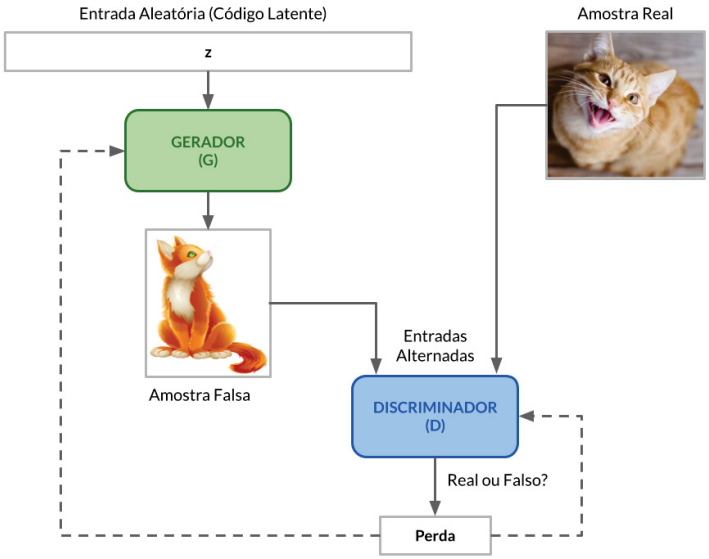


Figura 2.22: Arquitetura genérica da GAN. Inicialmente, o gerador recebe uma entrada aleatória (ruído) e gera uma amostra. Essa amostra vai ser entregue ao discriminador, que irá predizer se essa amostra é falsa ou é pertencente aos dados do conjunto de treino (imagens reais). A função de custo será calculada e a perda será utilizada para atualizar os dois modelos. Fonte: elaborado pela autora.

complexa. A alteração em uma dessas características pode ter efeitos imprevisíveis em outras características (Karras et al., 2019). A arquitetura da StyleGAN pode ser conferida na Figura 2.23, sendo seus componentes descritos abaixo.

- **Rede de Mapeamento**: o objetivo da rede de mapeamento é codificar o vetor de entrada (código latente) em um vetor intermediário capaz de controlar diferentes características visuais. Para reduzir as desvantagens do emaranhado de características, essa rede consiste em 8 camadas totalmente conectadas e sua saída (w) apresenta o mesmo tamanho da entrada (512x1). A adição de uma rede neural para preparar (mapear) a entrada possibilita a geração de um vetor que não precise seguir a distribuição do conjunto de treinamento e também a reduzir a correlação entre características (Karras et al., 2019). A ilustração da rede de mapeamento é apresentada à esquerda na Figura 2.24.

- **Módulo de Estilo *AdaIN***: o módulo *Adaptive Instance Normalization* (AdaIN) transfere a informação codificada, criada pela rede de mapeamento, para a imagem gerada. O módulo é adicionado em cada nível de resolução da Rede de Síntese (gerador) e define a expressão visual das características daquele nível (Karras et al., 2019). A Figura 2.24 apresenta a aplicação do módulo, onde o vetor intermediário (w) é transformado pela camada totalmente conectada (A) em uma escala e viés. Então, cada canal de saída da camada de convolução é normalizado pela média e variância e, em seguida, são transformados pela escala e viés encontrados anteriormente. Esse processo define a importância de cada filtro na convolução e essa sintonia traduz a informação de w para uma representação visual.

- **Variação Estocástica**: existem muitos aspectos nas imagens que são pequenos e podem ser vistos como estocásticos. No caso de imagens faciais, alguns exemplos são sardas, posicionamento dos fios de cabelo e rugas. Essas características tornam as imagens mais realistas e aumentam a variabilidade dos dados. A forma comum de inserir essas

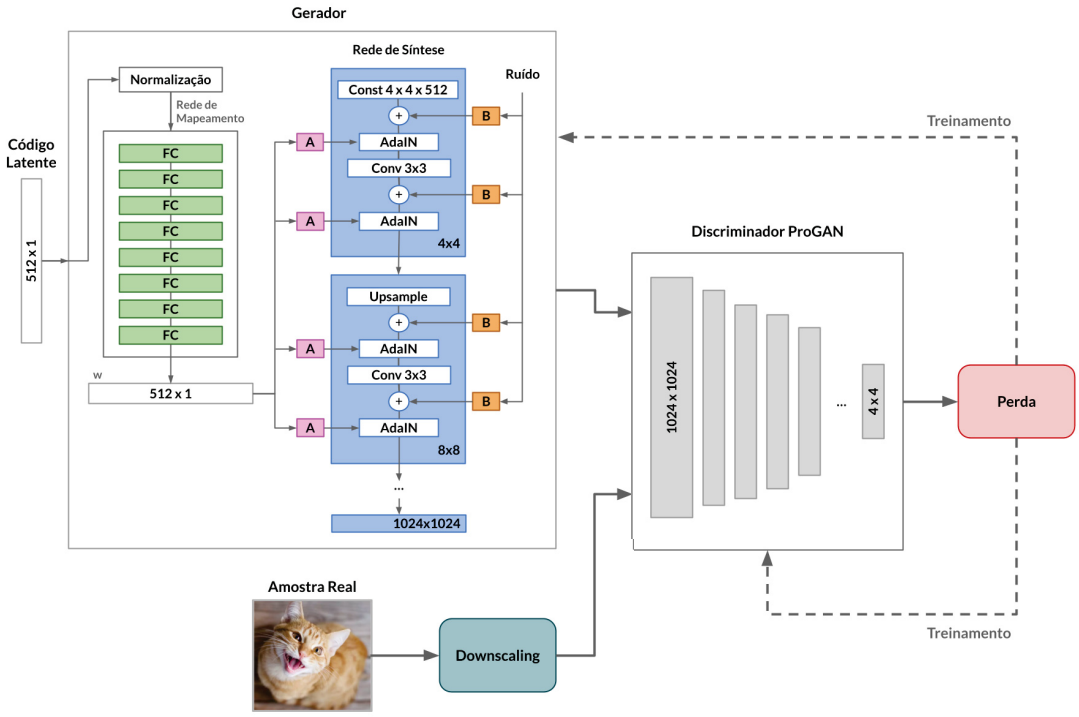


Figura 2.23: Arquitetura da StyleGAN. À esquerda é apresentado o gerador, com a Rede de Mapeamento (em verde), logo em seguida a Rede de Síntese (em azul) com a adição de estilo (em rosa) e a variação estocástica (em alaranjado). À direita é ilustrado o discriminador, que manteve a arquitetura da ProGAN. Fonte: adaptado de Horev (2018).

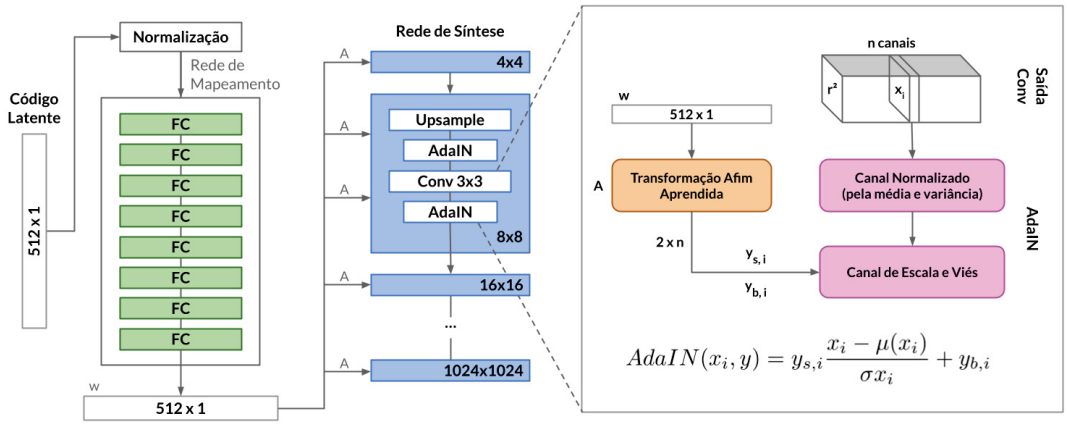


Figura 2.24: Arquitetura parcial da StyleGAN com foco no módulo da AdaIN. À esquerda é apresentada a Rede de Mapeamento, que produz o vetor intermediário w para ser utilizado em várias camadas da rede de síntese (em azul). O módulo da AdaIN é expandido para detalhar seus componentes de normalização. Fonte: adaptado de Horev (2018).

pequenas características na geração de imagens seria adicionar ruído no vetor de entrada. Entretanto, devido ao emaranhado de características, seria complexo controlar o efeito do ruído. Assim, a StyleGAN adiciona essas variações estocásticas de forma similar à AdaIN, escalando o ruído e adicionando ele em cada canal antes do módulo da AdaIN (Karras et al., 2019). A Figura 2.23 ilustra a inserção do ruído em algumas camadas da Rede de Síntese.

- **Truque de Truncamento**: a ideia do truncamento de vetor se resume a reduzir a influência das características de alta frequência e a aleatoriedade. Com isso, a rede pode se concentrar em gerar imagens mais controladas e coerentes, enquanto ainda mantém a diversidade e a variação nas imagens geradas. O truncamento ocorre nas camadas intermediárias da Rede de Síntese, que ao invés de receberem o vetor resultante da rede de mapeamento (w), esse vetor é truncado de acordo com um valor, sendo forçado a se aproximar do vetor médio. Se esse valor for baixo, as imagens geradas serão mais semelhantes às imagens médias do conjunto de treino e, se for alto, as imagens geradas serão mais variadas e distintas das imagens médias. Utilizando diferentes valores para cada bloco, o modelo pode controlar o quão longe da média cada conjunto de características vai estar (Karras et al., 2019).

- **Mistura de Estilo**: durante o treinamento da StyleGAN, o modelo pode ficar preso em uma solução subótima, gerando imagens que são muito semelhantes entre si. Para mitigar este problema, os autores introduziram a técnica de Mistura de Estilo (*Style Mixing*). Na etapa de treino, o modelo é ajustado para gerar duas imagens diferentes a partir de dois vetores de entrada aleatórios, sendo a escolha do par de imagens feita de forma aleatória. No entanto, ao invés de alimentar diretamente os vetores no gerador, uma proporção aleatória de camadas intermediárias da Rede de Síntese de uma imagem é misturada com as camadas intermediárias da rede de outra imagem. Assim, o modelo gera uma imagem misturada a partir das camadas intermediárias misturadas. Esse processo ajuda o modelo a aprender a separar as características das imagens de treinamento e gerar imagens mais diversificadas e com melhor qualidade, além de previnir que a rede assuma que estilos adjacentes (em níveis diferentes) são correlacionados (Karras et al., 2019).

Além de todas essas alterações, uma melhora adicional da StyleGAN foi a atualização de diversos hiperparâmetros da rede, como a duração do treinamento, a função de perda e a substituição do *up/downsampling* simples pela interpolação bilinear (Karras et al., 2019). Para resumir os aspectos principais da StyleGAN será utilizado o exemplo de um conjunto de dados de tipos de flores. Se o objetivo for controlar as características de estilo das imagens geradas, como cores e padrões de pétalas, o módulo da AdaIN seria responsável por ajustar o vetor para cada imagem, podendo incluir informações sobre as cores desejadas, a forma das pétalas e outros detalhes.

Mas se o problema em questão for a alta correlação entre as imagens geradas, a técnica de *Style Mixing* seria uma solução, permitindo o aprendizado de separação de características, como forma, cores e textura. Além disso, essa técnica possibilitaria a mistura desses atributos de maneiras diferentes para gerar imagens mais diversas e realistas, incluindo flores que não foram vistas no conjunto de treinamento original. Por fim, para controlar a variação das imagens geradas é utilizado o truncamento de vetor. Se o valor de truncamento for baixo, as imagens geradas terão características mais médias e típicas das flores do conjunto de treinamento. Se a maioria das flores do conjunto de treino tiverem pétalas rosas, as flores geradas terão pétalas

predominantemente rosas. No entanto, se o objetivo for gerar imagens de flores mais criativas e variadas, é necessário aumentar o valor de truncamento de vetor.

2.10 STYLEGAN2

A StyleGAN2 é uma evolução da StyleGAN que surgiu com o objetivo de reparar alguns problemas encontrados em sua predecessora. Um desses problemas foi a formação de artefatos semelhantes a bolhas de água nas imagens geradas (Figura 2.25 *a)*). Os autores atribuíram esse problema à camada de normalização *AdaIN*, em que o gerador cria esses artefatos para contornar uma falha de projeto na sua arquitetura. A solução encontrada foi a adição da demodulação dos pesos, que é uma técnica que modula a magnitude dos pesos (ou filtros) nas camadas de convolução do gerador com base nas características do vetor de entrada. Essa técnica não só resolve a aparição dos artefatos como ajuda a regular a magnitude dos pesos e melhorar a qualidade e estabilidade das imagens geradas (Karras et al., 2020b).

A segunda principal melhoria foi a alteração do crescimento progressivo para um gerador baseado em *Skip Connections* (Karnewar e Wang, 2020) e um discriminador com arquitetura das *Residual Networks* (He et al., 2015). Essa alteração foi necessária porque o crescimento progressivo tem uma forte preferência pela localização de algumas características, como dentes e olhos. A Figura 2.25 *b)* mostra o alinhamento dos dentes ficando preso em uma localização mesmo com a mudança de posicionamento da cabeça (Karras et al., 2020b).

Por fim, a última contribuição principal foi a apresentação de um novo método para encontrar o código latente que reproduz uma determinada imagem. As alterações foram realizadas na função de perda com a adição da métrica *Perceptual Path Length* (PPL). A Figura 2.25 *c)* demonstra os efeitos dessa melhoria, sendo a primeira linha composta pelas imagens originais e a segunda linha pelas imagens reproduzidas. A imagens da esquerda foram reproduzidas a partir de imagens falsas (geradas) enquanto as imagens da direita foram projetadas a partir de imagens reais. Os autores concluíram que imagens falsas podem ser reproduzidas quase perfeitamente, sendo que as imagens reais apresentam diferenças claras para as imagens geradas (Karras et al., 2020b).

2.11 STYLEGAN2ADA

Um dos problemas encontrados ao treinar as arquiteturas da GAN com conjuntos de dados de tamanho limitado é o efeito de *overfitting* do discriminador. Esse efeito faz com que o discriminador produza respostas sem sentido resultando na divergência do treinamento. Para solucionar o *overfitting*, o Aumento de Dados é a técnica adotada por quase todas as áreas de *Deep Learning* (Karras et al., 2020a). Assim, foram nessas questões que os autores de Karras et al. (2020a) se inspiraram para propor a StyleGAN2*ADA*.

A StyleGAN2*ADA* é uma versão aprimorada da StyleGAN2, que tem o objetivo de adicionar técnicas de Aumento de Dados no processo de treinamento para diversificar e aumentar a quantidade de exemplos do conjunto de dados. Nessa rede, são aplicadas uma série de transformações aleatórias nas imagens de treino, como rotações, escalas, recortes e espelhamentos, para gerar novas imagens com variações nas características das imagens originais. Ao mesmo tempo, o discriminador é treinado para diferenciar as imagens criadas pelo gerador e as imagens reais do conjunto de treinamento, independentemente de qualquer transformação aplicada sobre elas (Karras et al., 2020a).

Durante o treinamento da StyleGAN2*ADA*, o Aumento de Dados é aplicado em um pequeno conjunto de amostras (chamado de *mini-batch*) das imagens de treino. Cada *mini-batch*

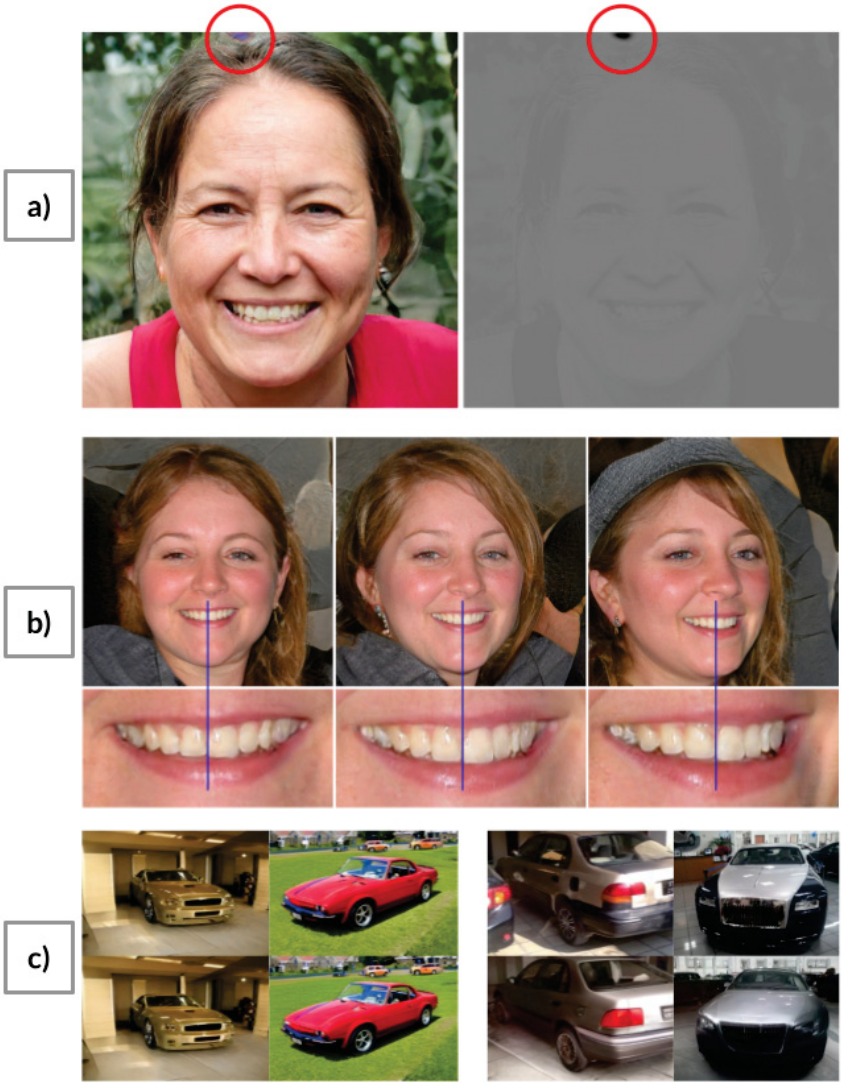


Figura 2.25: Exemplos dos problemas encontrados na StyleGAN e das melhorias proporcionadas pela StyleGAN2. Em *a)* é apresentada uma imagem com o problema das bolhas de água (presente no círculo vermelho); em *b)* é mostrado um exemplo da fixação da localização dos dentes mesmo com a mudança de posicionamento da cabeça na imagem gerada e, por fim, em *c)* é demonstrada a capacidade de reprodução de imagens da StyleGAN2, onde as imagens da esquerda são imagens falsas (acima) com suas reproduções (embaixo) e na direita são imagens reais (acima) com suas reproduções (embaixo). Fonte: adaptado de Karras et al. (2020b).

é pré-processado aleatoriamente com transformações diferentes, de modo que cada iteração do treinamento utilize um conjunto de imagens diferente do conjunto de treinamento original (Karras et al., 2020a). A probabilidade de aplicar o aumento em uma imagem no *mini-batch* é controlada pelo valor p, sendo o mesmo valor de p usado para todas as transformações. O valor inicial de p é 0 e seu valor é ajustado ao longo do tempo de acordo com uma heurística que avalia o *overfitting*. Se a heurística indicar muito *overfitting*, então p é incrementado e, da mesma forma, se houver pouco *overfitting*, p é decrementado. Esse método de ajuste automático de p foi chamado de *Adaptive Discriminator Augmentation* (ADA) (Karras et al., 2020a).

A Figura 2.26 apresenta o fluxo do modelo em *a)* e alguns exemplos de imagens geradas com diferentes valores de p, em *b)*. Por fim, os autores de Karras et al. (2020a) garantem que o método encontra a distribuição correta e não a distribuição aumentada (o que seria chamado de vazamento). Essa garantia se resume ao uso de transformações invertíveis e um p abaixo de 85%. As transformações escolhidas para comporem o Aumento de Dados consideraram 5 categorias, que foram: Pixel *Blitting*, Transformações Geométricas, Transformações de Cores, Filtragem do Espaço de Imagem e Corrupções do Espaço de Imagem (Karras et al., 2020a). Alguns exemplos dessas transformações podem ser observados na Figura 2.27.

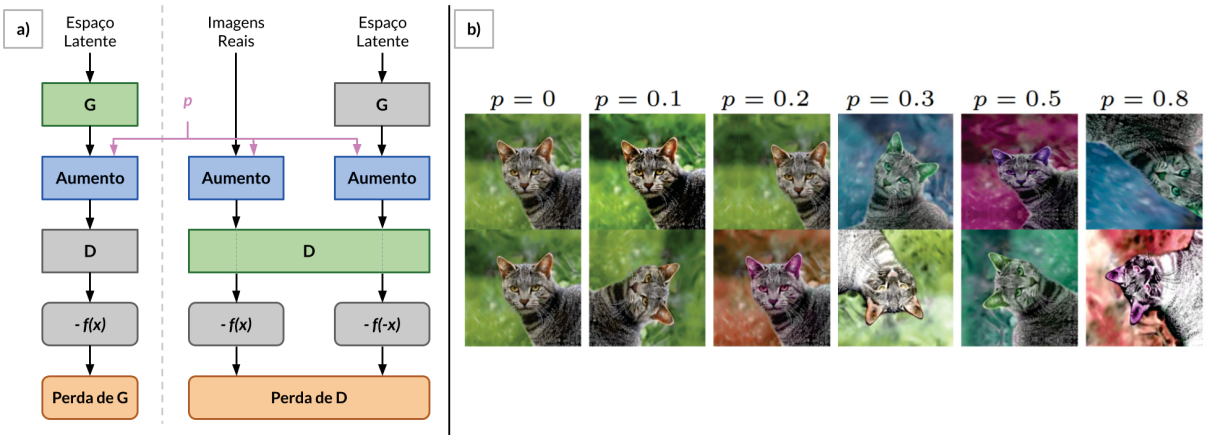


Figura 2.26: Em *a)* é apresentado o fluxo da StyleGAN2*ADA*, tanto do Gerador (esquerda da linha trecejada) quanto do Discriminador (direita da linha tracejada). Em *b* são mostrados exemplos de imagens geradas com diferentes valores de p. Fonte: adaptado de Karras et al. (2020a).

Figura 2.27: Exemplos das transformações consideradas no processo de Aumento de Dados da StyleGAN2*ADA*. Fonte: adaptado de Karras et al. (2020a).

2.12 MÉTRICAS DE AVALIAÇÃO DA CLASSIFICAÇÃO

Na área de classificação, diferentes métricas podem ser utilizadas para avaliarem os resultados. Dependendo do comportamento dos dados, algumas métricas podem ser mais adequadas do que outras. Além disso, a forma de avaliação dos resultados pode ajudar muito na compreensão dos resultados. Em problemas com desbalanceamento de classes, por exemplo, métricas calculadas individualmente por classe apontam mais diretamente as dificuldades dos modelos do que as métricas gerais (Sokolova e Lapalme, 2009). Nessa área, as métricas mais comuns são a Acurácia, Precisão, *Recall* e *F1-Score*, que serão descritas abaixo. Todas as métricas apresentadas se baseiam nas definições da Tabela 2.3.

Tabela 2.3: Matriz de confusão para classificação binária. Fonte: adaptado de Vujovic (2021).

		Rótulos	
		Positivo	**Negativo**
Predições	**Positivo**	Verdadeiro Positivo (VP)	Falso Positivo (FP)
	Negativo	Falso Negativo (FN)	Verdadeiro Negativo (VN)

- **Acurácia**: métrica que indica o quão bom foi o desempenho do classificador em relação aos rótulos pré-definidos por um especialista. Pode ser calculada pelo número de previsões corretas dividido pelo total de previsões, como é mostrado pela Equação 2.4 (Cordeiro, 2019).

$$Acurácia = \frac{VP + VN}{VP + FP + VN + FN} \tag{2.4}$$

- **Precisão**: métrica que avalia os acertos da classe positiva em relação a todos os exemplos preditos como positivos, ou seja, os VPs comparados com VPs + FPs. O cálculo dessa métrica é apresentado na Equação 2.5 (Vujovic, 2021).

$$Precisão = \frac{VP}{VP + FP} \tag{2.5}$$

- ***Recall***: métrica que contabiliza os acertos da classe positiva em relação a todos os exemplos que realmente são positivos, ou seja, os VPs comparados com VPs + FNs. A Equação 2.6 contém a definição desta métrica (Vujovic, 2021).

$$Recall = \frac{VP}{VP + FN} \tag{2.6}$$

- ***F1-Score***: a pontuação *F1-Score* tem o objetivo de considerar o equilíbrio entre a Precisão e o *Recall* do classificador nos seus cálculos. Se a Precisão for baixa, o *F1-Score* é baixo, e se o *Recall* for baixo, o *F1-Score* também vai apresentar valores baixos. A Equação 2.7 apresenta a fórmula dessa métrica (Vujovic, 2021).

$$F1 - Score = \frac{2 * Precisão * Recall}{Precisão + Recall} \tag{2.7}$$

2.13 MÉTRICAS DE AVALIAÇÃO DA GERAÇÃO DE IMAGENS

No cenário das GANs, tanto o gerador quanto o discriminador são treinados em conjunto para manter um equilíbrio. Dessa forma, não há nenhuma função de perda objetiva utilizada para treinar o modelo do gerador, o que significa que o modelo deve ser avaliado analisando a qualidade das imagens sintéticas geradas (Salimans et al., 2016). Nesta seção, serão brevemente descritos os métodos qualitativos e quantitativos para avaliar modelos de geração de imagens sintéticas.

2.13.1 Métodos Qualitativos

Métodos qualitativos são aquelas medidas que não são somente numéricas e muitas vezes envolvem a avaliação humana ou avaliação por comparação. Esses métodos são geralmente utilizados como um ponto de partida para descobrir se as imagens produzidas por um modelo generativo possuem uma boa qualidade ou se enfretam os problemas de *underfitting* ou *overfitting* (Borji, 2018). Abaixo, serão enumeradas as métricas qualitativas (extraídas de Borji (2018)) mais utilizadas para avaliar imagens sintéticas.

- **Categorização Rápida de Cenas:** nesse experimento, os participantes são questionados para distinguir as amostras geradas das imagens reais em um curto período de apresentação (uma fração de segundo).
- **Classificação e Julgamento por Preferência:** nessa avaliação, os participantes classificam os modelos em termos de fidelidade de suas imagens geradas. Também pode ser aplicada para comparar duas configurações de rede diferentes.
- **Investigação e Visualização do Interior das Redes:** técnica que explora e ilustra as representações internas e dinâmicas dos modelos, assim como as características aprendidas.

2.13.2 Métodos Quantitativos

A avaliação quantitativa do gerador se refere ao cálculo de pontuações numéricas específicas utilizadas para resumir a qualidade das imagens geradas. Essas métricas podem ser calculadas para comparar um par de imagens ou para verificar o quão bem as imagens geradas se encaixam na distribuição das imagens reais (Borji, 2018). A seguir, serão listados os métodos quantitativos mais utilizados para avaliar a qualidade e similaridade entre imagens.

- **SSIM:** o *Structural Similarity Index* (SSIM) é uma métrica de comparação entre duas imagens baseada no sistema de percepção humano, sendo capaz de identificar informações estruturais em uma cena. O SSIM extrai 3 componentes principais de uma imagem, que são a luminância, o contraste e a estrutura. A luminância é medida pela média (μ) de todos os valores de pixel, o contraste é calculado pelo desvio padrão (σ) e a estrutura é definida como a divisão da imagem de entrada menos a média pelo desvio padrão. A equação 2.8 apresenta a fórmula da métrica, sendo C_1 e C_2 constantes baseadas no intervalo dinâmico de pixels e x e y as duas imagens sendo comparadas.

Cada par de componentes tem suas funções próprias de comparação, sendo o resultado dessas comparações o valor de similaridade. Um valor de +1 indica que as duas imagens são muito semelhantes, enquanto um valor de -1 informa que as imagens fornecidas são muito diferentes (Wang et al., 2004).

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{2.8}$$

- **PSNR:** o *Peak Signal to Noise Ratio* (PSNR) é uma métrica utilizada para avaliar a qualidade de imagens e vídeos. Assim, dada uma imagem original e uma imagem corrompida ou alterada, a métrica mede a similaridade entre essas imagens. O cálculo do PSNR pode ser feito a partir do *Mean Squared Error* (MSE), que é a média dos quadrados das diferenças entre os valores de pixel correspondentes das duas imagens. O cálculo do PSNR pode ser conferido na Equação 2.9, onde *MAX* é o valor máximo possível para um pixel (por exemplo, 255 para uma imagem de 8 bits). Quanto mais alto o resultado, mais semelhantes são as imagens e, quanto mais baixo, mais diferentes (Hemanth et al., 2019).

$$PSNR = 10 \ . \ \log_{10}\left(\frac{MAX^2}{MSE}\right) \tag{2.9}$$

- **DreamSIM:** pertencente ao estado da arte das métricas de similaridade entre imagens, a DreamSIM surgiu da expansão do conceito de similaridade de pixels e *patches* para uma forma de análise mais holística, considerando julgamentos humanos em sua definição. Para isso, os autores de Fu et al. (2023) construíram um conjunto de dados que leva em consideração uma imagem de referência e duas outras similares (imagem *A* e imagem *B*) de acordo com anotações humanas. Então, na fase de modelagem, a extração de características é realizada com um *ensemble* de métricas de similaridade existentes combinadas com os modelos de MLP e Transformers. A perda é baseada na *Hinge Loss* entre as distâncias do cosseno da imagem de referência em relação às imagens similares, sendo os pesos do modelo atualizados pela técnica *Low-Rank Adaptation* (LoRA). Quanto mais próximo a zero o valor da DreamSIM (y), mais similares as imagens são. A visão geral do modelo pode ser observada na Figura 2.28.

- **FID:** a pontuação *Fréchet Inception Distance* (FID) é uma métrica que calcula a distância entre dois vetores de características calculados para imagens reais e geradas. Uma rede *Inception-v3* é utilizada para extrair as características (vetor de codificação) de cada imagem e então é definida uma distância entre esse vetores com base nos cálculos de covariância e vetor médio. A equação 2.10 apresenta a fórmula do FID, onde FID^2 é a pontuação FID calculada em unidades quadradas, m e m_w se referem à média das características das imagens reais e geradas e C e C_w são as matrizes de covariância para os vetores de características reais e gerados. O primeiro módulo calcula a diferença da soma ao quadrado entre os dois vetores médios (norma L2) e o *Tr* se refere a operação de álgebra linear de traço (soma de elementos ao longo da diagonal principal da matriz quadrada) (Heusel et al., 2018).

$$FID^2((m, C), (m_w, C_w)) = ||m - m_w||_2^2 + Tr(C + C_w - 2(CC_w)^{1/2}) \tag{2.10}$$

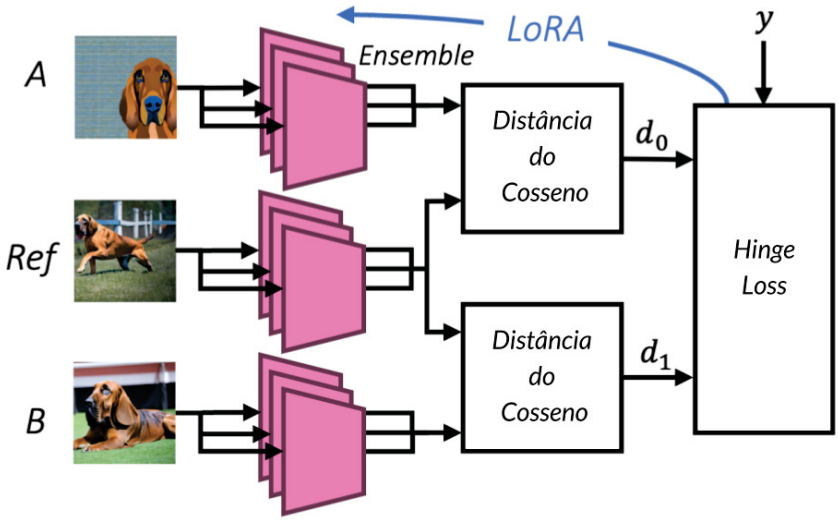


Figura 2.28: Visão geral da arquitetura da técnica DreamSIM. Fonte: adaptado de Fu et al. (2023).

- **KID:** o *Kernel Inception Distance* (KID) é uma métrica similar ao FID que calcula a discrepância média máxima quadrada (*Maximum Mean Discrepancy* (MMD)) entre as características de imagens reais e geradas. Essa métrica foi proposta com o objetivo de ser menos sensível a pequenas variações nas características das imagens. A equação do KID em unidades quadradas é apresentada abaixo (Equação 2.11), onde x_i é uma imagem real pertencente a X (conjunto de imagens reais com tamanho n), y_i é uma imagem gerada pertencente a Y (conjunto de imagens geradas com tamanho m), ϕ é a função de mapeamento de características obtida a partir de uma *Inception-v3* e K é o *kernel* utilizado pelo MMD para estimar as diferenças entre pares de pontos nas duas distribuições. A ideia dos dois primeiros termos é medir a similaridade entre as distribuições de características das imagens dos conjuntos X e Y. Essa similaridade é calculada como a média da função de *kernel* aplicada a todas as combinações possíveis das imagens do conjunto. Já o último termo mede a diferença entre as distribuições de características de X e Y (Horak et al., 2020).

$$\begin{aligned} KID^2(x_{1:n}, y_{1:m}) = & \frac{1}{n^2}\sum_{i,j=1}^{n} K(\phi(x_i), \phi(x_j)) + \\ & \frac{1}{m^2}\sum_{i,j=1}^{m} K(\phi(y_i), \phi(y_j)) - \\ & \frac{2}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m} K(\phi(x_i), \phi(y_j)) \end{aligned} \tag{2.11}$$

3 REVISÃO DA LITERATURA

Nesta seção, serão apresentadas breves descrições dos trabalhos realizados na área de imagens médicas nos últimos anos. A seleção dos trabalhos foi elaborada considerando quatro temas: classificação e segmentação de imagens de imuno-histoquímica, aumento de dados com *Generative Adversarial Network*s (GANs) em imagens médicas e técnicas de interpretabilidade dos modelos de *Deep Learning* na área de saúde. Além dessas breves descrições, serão apresentados com detalhes os trabalhos relacionados com o tema da pesquisa, o resumo dos artigos da área e as comparações com os resultados da nossa pesquisa (Tabela 3.1).

3.1 CLASSIFICAÇÃO DE IMAGENS DE IMUNO-HISTOQUÍMICA

Com o rápido desenvolvimento da tecnologia na última década, a área de patologia passou por diversas mudanças ao começar sua era digital, proporcionada pela geração das *Whole-Slide Images* (WSIs) (Laurinavicius et al., 2016). Assim, houve um esforço recente de pesquisadores para realizar a classificação automática de diversos tipos de imagens de imuno-histoquímica. Abaixo serão listados e brevemente apresentados os trabalhos selecionados deste tema.

- Cordeiro et al. (2018) propuseram um algoritmo para a pontuação de imagens imuno-histoquímicas de *Human Epidermal growth factor Receptor-type 2* (HER-2) para detectar o câncer de mama. Inicialmente, os autores dividiram a WSI da lâmina em pequenas porções de mesmo tamanho, chamadas de *patches*. Então, foram extraídas características de cada *patch* se baseando na coloração da imagem. Por fim, vários modelos de *Machine Learning* foram treinados para classificar os *patches* em 4 classes diferentes. Os autores alcançaram acurácias próximas à 90% na classificação dos *patches*, e de 94,12% ao combinar todos os *patches* de uma lâmina e classificar a WSI completa nas 4 categorias de câncer.

- Tang et al. (2019) focaram em imagens de imuno-histoquímica cerebrais que retrataram o acúmulo de deposição extracelular de placas beta amilóides, visando diagnosticar a Doença de Alzheimer. A primeira etapa do método foi a extração das regiões de interesse das WSIs, que foi obtida com técnicas de normalização, morfologia matemática e segmentação. Essas regiões resultantes foram anotadas por especialistas considerando 3 classes que definem os diferentes tipos de placas. Por fim, a classificação foi realizada com uma *Convolutional Neural Network* (CNN) e ocorreu a adição de técnicas de Aumento de Dados para mitigar o desbalanceamento das classes. Os autores alcançaram 99% de *Area Under the Receiver Operator Characteristic* (AUROC) e 74% de *Area Under the Precision-Recall Curve* (AUPRC) e também investigaram o funcionamento do modelo com técnicas de interpretabilidade.

- Rogalsky et al. (2021) criaram uma base de dados com WSIs de imuno-histoquímica da mama de pacientes avaliados pelos biomarcadores *Estrogen Receptor* (ER) e *Progesterone Receptor* (PR). Essas WSIs foram decompostas em imagens menores de mesmo tamanho (*patches*) e associadas às classes presentes nos prontuários dos pacientes. Esses *patches* passaram pelo processo de extração de características, que consistiu na criação de histogramas com dimensões variadas no espaço de cores *Hue-Saturation-Value*

(HSV). Um *Support Vector Machine* (SVM) foi treinado com essas características, alcançando *f1-scores* próximos à 65% nas imagens do biomarcador ER e de 58% no PR.

- Mridha et al. (2022) trabalharam com imagens do tipo HER-2, utilizando um conjunto de dados público contendo 4 classes responsáveis por indicar o estágio do câncer de mama. Essas imagens poderiam pertencer a duas categorias, sendo uma delas advinda de uma aplicação mais direta com *Hematoxilina e Eosina* (H&E) e outra decorrente do processo de *Imuno-histoquímica* (IHQ). Uma nova CNN foi proposta, se baseando na arquitetura da Inception-V3, com adições de camadas e ajuste de parâmetros. Essa CNN foi treinada e obteve acurácia de 85% nas imagens H&E e de 88% em IHQ, superando o estado da arte.

- Che et al. (2023) definiram um método automático de pontuação de imagens HER-2 para a detecção do câncer de mama. Para isso, os autores criaram uma base de imagens que foi rotulada por dois patologistas. Cada WSI do conjunto de dados recebeu um valor de 0 a 3 indicando a intensidade das manchas na imagem. Inicialmente, foram extraídas as regiões com lesões significativas das imagens (*patches*) e divididas em conjuntos de treino e validação. Para realizar a classificação, o modelo envolveu um classificador binário para detectar a presença de tumores combinado com uma segmentação. Essa combinação gerou um mapa de probabilidades que permitiu a extração de 3 características responsáveis pela classificação final com base em regras simples. Com essa metodologia de 3 fases, os autores atingiram acurácias acima de 95% em todas as classes.

Dentre as alternativas de classificação de imagens de imuno-histoquímica apresentadas, um progresso expressivo foi alcançado, mas ainda existem algumas limitações. Uma delas foi a omissão do desbalanceamento dos dados, que poderia indicar quantos exemplos de cada classe foram utilizados para treinar e testar o modelo. Manter os dados balanceados garante a validade de métricas como a acurácia e a qualidade do treinamento dos modelos baseados em redes neurais. Além disso, poucas métricas foram calculadas e não houve uma padronização sobre elas, o que dificultou a comparação entre trabalhos. Por fim, somente um trabalho apresentou métricas por classe, sendo frequente o cálculo de apenas métricas gerais, o que limitou a compreensão do desempenho dos modelos.

3.2 SEGMENTAÇÃO DE IMAGENS DE IMUNO-HISTOQUÍMICA

Além da classificação, a segmentação de imagens médicas é de extrema importância para garantir a veracidade de alguns diagnósticos. No cenário de imagens de imuno-histoquímica nucleares, a segmentação pode contribuir para calcular a proporção de células positivas em uma determinada porção da imagem (Rogalsky, 2021). Assim, serão apresentados abaixo algumas das técnicas mais recentes de segmentação em imagens médicas de imuno-histoquímica.

- Signaevsky et al. (2019) construíram um conjunto de dados com imagens de imuno-histoquímica de autópsia cerebral de indivíduos com 4 diferentes tipos de tauopatias (classe de doenças neurodegenerativas). Os autores optaram por um versão modificada da CNN *SegNet*, com o objetivo de segmentar as regiões de interesse das imagens. Também foram aplicadas técnicas de Aumento de Dados para contribuir com o treinamento, como espelhamentos e rotações. Como conclusão, os autores enfatizaram que além da pesquisa apresentar um conjunto de métodos reprodutíveis, rápidos e imparciais, foi

possível alcançar bons resultados, com um *f1-score* de 81% ao comparar os resultados com as anotações do especialista.

- Tamaki et al. (2021) utilizaram os *patches* extraídos em Rogalsky (2021) para propôr um método de delimitação das células receptoras de progesterona e estrogênio. Os pré-processamentos considerados foram: aumento de contraste, agrupamento médio para cada pixel, limiarização e filtragem Gaussiana. Por fim, a segmentação foi composta de 3 etapas: aplicação do filtro Laplaciano modificado, filtros morfológicos e o algoritmo *Watershed*. Os resultados mostraram que o conjunto de métodos obteve sucesso em delimitar as regiões celulares, mas não realizou a delimitação dos contornos individuais de cada célula.

- Rmili et al. (2022) propuseram um conjunto de métodos para realizar a segmentação de imagens imuno-histoquímicas de *Neuroendocrine Tumor*s (NETs). Para mitigar os problemas de variações de coloração e iluminação, os autores buscaram descobrir qual espaço de cor se adequava melhor ao problema. Além disso, foi criado um procedimento para selecionar apenas o canal mais relevante do espaço de cores, visando reduzir a quantidade de dados analisados. Por fim, a segmentação contou com métodos como limiariazação adaptativa, filtragem Laplaciana, morfologia matemática e o algoritmo *Watershed*. Os resultados se aproximaram de 98% para as métricas *recall*, precisão, *f1-score* e coeficiente Dice, demonstrando uma melhora de 10% do estado da arte analisado.

A segmentação pode auxiliar de diversas formas na definição do diagnóstico, mas também apresenta limitações. Para que seja possível avaliar os métodos de segmentação e também para treinar modelos de *Deep Learning*, são necessários dados anotados que consistem na delimitação do posicionamento dos objetos de interesse em milhares de imagens, o que consome bastante tempo e esforço dos especialistas. Além disso, não há uma padronização da forma de avaliação dos resultados, dificultando a comparação de métodos advindos de trabalhos diferentes.

3.3 GERAÇÃO DE IMAGENS MÉDICAS COM GANS

Alguns problemas comuns no cenário de análise de imagens médicas são: a quantidade limitada de imagens rotuladas, a falta de variabilidade nos dados e o desbalanceamento das classes presentes nesses rótulos (Mukherkjee et al., 2022). A razão desses problemas está associada à necessidade de especialistas com um bom conhecimento dos dados específicos para realizarem as anotações, resultando em um processo caro em termos de tempo, dinheiro e esforço (Islam e Zhang, 2020). Considerando esse panorama, serão apresentadas abaixo algumas soluções com o uso de geração sintética de dados utilizando GANs (*Generative Adversarial Networks*).

- Hong et al. (2021) visaram permitir a síntese de imagens médicas 3D de alta qualidade propondo uma extensão da rede *StyleGAN2*, chamada de *3D-StyleGAN*. Os experimentos foram realizados com imagens de ressonância magnética do cérebro, abrangendo vários conjuntos de dados com disponibilidade pública. Os autores apresentaram uma análise visual das imagens geradas e destacaram que as estruturas anatômicas foram geradas corretamente e posicionadas coerentemente. Além disso, foram realizados experimentos de reconstrução das imagens, utilizando o espaço latente, e com a combinação de estilos característica da *StyleGAN*. Como conclusão, foi pontuado que com a alta qualidade das

imagens geradas e com o controle da variabilidade anatômica pelos vetores de estilo, a proposta pode contribuir para enfrentar diversos problemas clínicos importantes.

- Liao et al. (2021) classificaram a taxa de *Estimated Glomerular Filtration Rate* (eGFR) das imagens ultrassonográficas renais, com o objetivo de detectar os 4 estágios da *Doença Renal Crônica* (DRC). Eles tinham como objetivo resolver problemas de desequilíbrio de classes e dados insuficientes durante o processo de treinamento de um modelo de *Deep Learning*. Dessa forma, os autores implementaram uma ACWGAN-GP (extensão da *Wasserstein GAN* (WGAN) e *Auxiliary Classifier Generative Adversarial Network* (AC-GAN)) para gerar novas imagens de classes com poucos exemplos. Na etapa de classificação eles optaram pela MobileNetV2, ResNet50 e pelo classificador pertencente à rede geradora. Como resultado, eles alcançaram um *f1-score* médio de 81% com o modelo MobileNetv2 ao adicionar os dados gerados ao conjunto de treinamento.

- Gulakala et al. (2022) buscaram resolver o problema de classificação das imagens de radiografia de tórax considerando três classes (Covid-19, saudável e pneumonia). Devido à escassez de imagens de raios X com Covid-19, os autores desenvolveram uma GAN U-Net residual capaz de lidar com a aleatoriedade da iluminação utilizando a função *Swish*, normalizações e *skip connections*. A proposta para o modelo de classificação foi uma nova arquitetura CNN capaz de superar as redes convolucionais presentes no estado da arte mesmo sendo 40% mais leve. Finalmente, os autores adicionaram 2.000 imagens sintéticas ao conjunto de treinamento e alcançaram uma precisão de 99,2% no conjunto de teste.

- Mukherkjee et al. (2022) discutiram sobre os obstáculos dos sistemas de auxílio ao diagnóstico, como o alto desbalaceamento de classes e a variabilidade insuficiente dos dados. Enfatizaram que o uso de técnicas de Aumento de Dados com operações básicas de processamento de imagens podem adicionar informações irrelevantes ou mudar o padrão de diagnóstico. Com o objetivo de solucionar esses problemas, os autores propuseram um novo método de geração de imagens chamado de *Aggregation of GAN* (AGGrGAN), que combinou imagens sintéticas de 3 GANs diferentes: 2 variantes da *Deep Convolutional Adversarial Network* (DCGAN) e a WGAN. O método foi avaliado em imagens de ressonância magnética contendo várias categorias diferentes de tumores no cérebro. Ao adicionar 300 imagens sintéticas no treinamento das CNNs mais recentes, os autores aumentaram em mais de 3 pontos percentuais a acurácia dos modelos.

- Osuala et al. (2023) analisaram 164 publicações com aplicações de técnicas de treinamento adversarial para a geração de dados sintéticos no contexto de imagens médicas de câncer. O objetivo dos autores foi fazer uma revisão da literatura destacando soluções não exploradas e com potencial de pesquisa. Além disso, foi proposto o *SynTrust*, um *framework* de meta-análise que consistiu em um conjunto de métricas rigorosas para avaliar a confiabilidade e a validade dos estudos na área. De 16 trabalhos analisados, 11 preencheram todos os critérios essenciais definidos pelo *SynTrust*. Como conclusão, os autores enfatizaram o alto nível de confiança dos estudos nessa área, ressaltando a versatilidade e os resultados altamente aplicáveis de algumas GANs.

- Krinski et al. (2023) propuseram uma análise de como as técnicas de Aumento de Dados melhoram o treinamento das redes de segmentação semântica em tomografias computadorizadas de pulmões com lesões da *Covid-19*. Os autores avaliaram 20

técnicas de Aumento de Dados baseadas em métodos de processamento de imagens e 4 baseadas em GANs (variações com *StarGAN* e StyleGAN2*ADA*). Os experimentos foram aplicados em 5 conjuntos de dados diferentes (individualmente e unificados), sendo a organização do processo de classificação realizada com Validação Cruzada de 5 pastas. As classes consideradas foram lesão e fundo, e as métricas de avaliação selecionadas contaram com o *F-Score* e *Intersection over Union* (IoU). Os melhores resultados foram alcançados com as técnicas de aumento que alteram a forma das lesões nas imagens e não as cores.

O uso de GANs para geração de imagens sintéticas na área da saúde ainda é um tema recente. Até o momento, não houve uma padronização dos métodos qualitativos e quantitativos para avaliar a qualidade das imagens geradas. Além disso, pelas variações das GANs serem baseadas em rede neurais, esses modelos são considerados caixas pretas por não esclarecerem quais características são consideradas para gerar uma nova imagem.

3.4 INTERPRETABILIDADE DE MODELOS DE DEEP LEANING

Os modelos de análise de imagens médicas baseados em *Deep Learning* ainda são vistos com muita incerteza nos ambientes clínicos. O objetivo principal das ferramentas de *Inteligência Artificial* (IA) para imagens médicas é auxiliar os médicos em suas tomadas de decisão, combinando vários fatores em um modelo que retorne uma saída acionável. Sem qualquer explicação dessa saída, a utilidade do modelo é limitada, pois não revela o processo de raciocínio, limitações e vieses (Salahuddin et al., 2022). Abaixo serão brevemente descritos artigos sobre interpretabilidade de modelos de *Deep Learning*, visando compreender como é possível tornar as IAs mais transparentes, robustas, justas e responsáveis na área de imagens médicas.

- Schutte et al. (2021) definiram um novo método de interpretabilidade para compreender as previsões dos modelos de caixa preta de análise de imagens. Para isso, foi treinada uma StyleGAN com imagens de raio X de joelho e também com imagens histológicas de linfonodos metastáticos indicativos do câncer de mama. Com a StyleGAN treinada, foi descoberta a modificação mínima no espaço latente capaz de mudar a previsão do modelo. Na base de dados de raio X, as imagens foram gradualmente modificadas durante a geração a ponto de mostrarem a progressão da osteoartrite, evidenciando o afinamento dos espaços das juntas. Os autores pontuaram que o método proposto é intuitivo para localizar as características preditivas na imagem e enteder como elas impactam na previsão, mostrando versões modificadas da imagem de entrada que podem produzir diferentes previsões.

- Salahuddin et al. (2022) discutiram a dificuldade de incorporação de redes neurais profundas no fluxo de trabalho clínico devido ao vago entendimento sobre o processo de tomada de decisão destes modelos. Dada a importância de solucionar este problema, os autores resumiram os detalhes técnicos, limitações e aplicações de métodos de interpretabilidade disponíveis para permitir um maior grau de transparência dessas redes na área de saúde. Dentre os métodos apresentados, o *Case-Based Models* e o *Concept Learning* alcançaram os melhores resultados.

A capacidade de explicar o processo de tomada de decisão de uma IA pode contribuir para uma maior adoção de modelos de *Deep Learning* no ambiente médico, mas ainda existem

muitos obstáculos. A complexidade dos métodos de interpretabilidade e a necessidade do auxílio de um especialista na maioria desses métodos são alguns exemplos de limitações enfrentadas no desenvolvimento de IAs interpretáveis na área médica.

3.5 TRABALHOS RELACIONADOS

Nesta subseção, serão apresentadas as descrições detalhadas dos cinco trabalhos com maior semelhança com o tema escolhido. A seleção dos trabalhos levou em consideração o tipo de imagem analisada, os métodos de pré-processamento implementados, os modelos de classificação escolhidos e as aplicações correlatas de geração de imagens médicas com GANs.

3.5.1 Abordagem Automática para Pontuação de Imagens HER-2 de Câncer de Mama

Nesta dissertação (Cordeiro, 2019), a autora trabalhou com conjuntos de dados de imuno-histoquímica HER-2, compostos de WSIs de tecidos mamários corados para testar a reação da proteína HER-2, capaz de influenciar em como o câncer de mama se desenvolve. Dependendo da coloração apresentada, uma imagem pode ser fortemente positiva (classe 3+), moderadamente positiva ou incerta (classe 2+) ou negativa (classes 0/1+).

A partir das WSIs com ampliação de 40x, foram extraídos *patches* com dimensões de 250 x 250 pixels, selecionando automaticamente partes relevantes das imagens e desconsiderando as porções com predomínio de fundo. No total, a autora utilizou 1867 *patches* extraídos de dois conjuntos de dados, o primeiro construído por ela (HistoBC-HER2) e o segundo da Universidade de Warwick do Reino Unido (*Warwick's Dataset*) (Qaiser et al., 2017).

Para a extração de características, foram experimentados os descritores de textura *Local Binary Pattern* (LBP) e *Parameter-free Threshold Adjacency Statistic* (PFTAS), aplicados na imagem colorida e na imagem *Diaminobenzidine* (DAB). Também foram testadas abordagens baseadas nas características de cor como os histogramas HSV e *Red-Green-Blue* (RGB) e, por fim, foram aplicadas algumas CNNs pré-treinadas para obter as características, como VGG16 e ResNet50. Após esta etapa, foram implementados os modelos SVM, *K-Nearest Neighbor* (KNN), *Multilayer Perceptron* (MLP) e *Decision Tree* para realizar a classificação.

Os experimentos foram divididos em duas etapas, sendo a primeira a classificação a nível de *patches*, que extraiu *patches* significativos de cada WSI para construir o conjunto de treinamento e todos os *patches* da WSI receberam uma classificação. Em seguida, o resultado da classificação inicial foi utilizado para gerar um histograma de ocorrência de classe para cada WSI. Então, esse histograma serviu como entrada para um classificador determinar a pontuação HER-2 da WSI.

Nos experimentos a nível de *patches*, os melhores resultados foram obtidos com o extrator de características ResNet50 em conjunto com o classificador SVM, chegando a 90,84% de acurácia. Na etapa seguinte, chamada de nível de paciente, os autores alcançaram 90,20% de acurácia no *Warwick's Dataset* utilizando também a ResNet50, mas com o classificador MLP. Já no conjunto de dados construído durante o trabalho, a maior acurácia obtida foi de 79,26% com a ResNet50 e com o classificador KNN.

3.5.2 Pontuação Semi-Automática dos Biomarcadores ER e PR em Imagens de Câncer de Mama

Na dissertação de Rogalsky (2021), a autora criou um conjunto de dados com 135 pacientes avaliados pelos biomarcadores ER e PR, totalizando 270 imagens digitalizadas de imuno-histoquímica DAB. As WSIs de tecido mamário foram aumentadas em 40x por uma lente de ampliação objetiva e, a partir dessas imagens aumentadas, foram extraídos manualmente

segmentos (chamados de *patches*) com dimensões de 400 x 300 pixels, totalizando 4680 *patches* referentes à 78 pacientes.

O processo de anotação das imagens foi separado em duas categorias, uma a nível de paciente, na qual cada WSI recebeu a classe pertencente ao prontuário do paciente, e a outra a nível de *patches*, em que as porções de imagem extraídas tiveram suas classes reavaliadas por um especialista ao invés de serem retiradas do prontuário. Devido à questões de tempo, nem todos os *patches* extraídos foram reavaliados. As classes consideradas foram divididas de acordo com a pontuação *Intensity Score* (IS): 0 (células saudáveis); 1+ (células fracamente positivas); 2+ (células moderadamente positivas) e 3+ (células fortemente positivas).

O primeiro pré-processamento aplicado foi a *Contrast Limited Adaptive Histogram Equalization* (CLAHE), uma equalização de histograma local com limite de contraste. As imagens foram convertidas para o espaço de cores LAB e o canal de luminância foi dado como entrada para a técnica de equalização. O segundo passo foi a limiarização, que iniciou com a aplicação de um filtro de deslocamento seguido de um filtro de borramento Gaussiano para nivelar gradientes e evitar a amplificação de ruído. Então, a imagem foi convertida para a escala de cinza e passou pela técnica Otsu, uma abordagem que encontrou um valor ótimo para realizar a limiarização.

Para extrair as células positivas (em marrom) e negativas (em azul) foi realizada uma separação de cores com o canal de matiz do espaço de cores HSV. Para cada uma das duas cores foi definido um intervalo representativo e com ele foi possível segmentar as células. A partir do resultado, foram extraídas as características normalizadas utilizando 15 combinações de histogramas do espaço de cores HSV, histogramas *Completed Local Binary Pattern* (CLBP) e estatísticas *Grey Level Co-occurrence Matrices* (GLCM). Essas características foram passadas para o classificador SVM seguindo a estratégia *Leave-One-Patient-Out* para treinamento e teste acompanhada pela função de decisão *One-Vs-Rest*.

Nos experimentos a nível de paciente, foram selecionados todos os *patches* da região de interesse de cada WSI e, a partir deles, foram calculados os histogramas normalizados acumulados. Assim, cada WSI detinha um único vetor de características e uma classe advinda do prontuário do paciente. Além disso, foram avaliadas duas formas de seleção desses *patches*, uma delas considerando somente os *patches* reavaliados e outra abrangendo todos os *patches* extraídos. Já na abordagem a nível de *patches*, cada imagem foi classificada diretamente, pois todas as imagens já haviam sido associadas à uma classe.

Os resultados de nível de paciente ficaram entre 42% e 63% de *f1-score*, considerando o uso de todos os *patches* em ambos biomarcadores. Já no caso dos *patches* reavaliados, esses valores aumentaram, oscilando entre 55% e 76%, também levando em conta os dois biomarcadores. Por fim, nos resultados da abordagem por *patches*, os autores alcançaram um *f1-score* de 100% para ambos biomarcadores com o uso de histogramas HSV. Assim, a autora chegou à conclusão de que fragmentar as WSIs pode ser útil para melhorar a classificação, e observou que o uso de histogramas HSV como extratores de características e do SVM como classificador pode ser eficiente para resolver este problema.

3.5.3 Segmentação e Classificação Não-Supervisionadas para Pontuação de Imagens de Câncer de Mama

Os autores de Mouelhi et al. (2018) construíram uma base de dados com a digitalização de tecidos de biópsias pertencentes a 28 pacientes com Carcinoma Ductal Invasivo da mama, visando evidenciar os biomarcadores ER e PR com o uso de corantes. A partir dessa digitalização, foram extraídas 84 imagens e elas foram ampliadas por fatores de 20x, 40x e 80x. Para cada uma dessas imagens foram associadas duas classes, uma pertencente ao sistema de pontuação IS (0,

1+, 2+ e 3+) e outra sendo o *Proportion Score* (PS) (0 a 5). O objetivo final do trabalho foi a pontuação *Allred*, que foi calculada como a soma do IS com o PS, podendo variar de 0 a 8.

Na etapa de segmentação, a imagem de imuno-histoquímica recebeu um aumento de contraste pela técnica de equalização do histograma de intensidade. Então, essa imagem foi ajustada por um alongamento de contraste linear para evidenciar os núcleos das células. Em seguida, foi utilizada uma limiarização adaptativa, uma filtragem morfológica usando o operador Laplaciano modificado e alguns filtros morfológicos. Para finalizar a segmentação, os autores optaram pela técnica de *Watershed* modificada, que além de retornar os contornos dos núcleos, solucionou o problema de um contorno contendo dois ou mais núcleos sobrepostos.

Para a classificação, inicialmente algumas das células benignas foram filtradas se baseando em informações dos especialistas em relação aos tamanhos e características morfológicas. Então, foram aplicadas quatro técnicas de separação por cor: extração do componente azul; extração do componente marrom; diferença dos canais vermelho e azul e *Color Deconvolution*. Essas técnicas foram combinadas por quatro métodos de agregação diferentes (mínimo, máximo, média e mediana) e o resultado foi comparado com três limiares dinâmicos para descobrir qual classe IS deveria ser associada à imagem.

Além do IS, a segmentação também permitiu o cálculo do PS com base na razão entre o número de núcleos cancerosos positivos e o número total de núcleos negativos e positivos em toda a imagem. Com o IS e o PS obtidos, a pontuação *Allred* foi calculada sendo avaliada pelas métricas de acurácia, sensibilidade e especificidade. Os melhores resultados foram alcançados com a combinação por mediana, com 98,6% de sensibilidade, 99,3% de especificidade e 98,9% de acurácia.

3.5.4 Geração de Imagens Sintéticas de Cérebro Utilizando DCGAN

Com o grande sucesso dos modelos de *Deep Learning* em tarefas de visão computacional, também surgiram os desafios em relação ao acesso aos dados. Um dos problemas observados pelos autores (Islam e Zhang, 2020) ao se tratar de conjuntos de imagens médicas foi que geralmente esses conjuntos são pequenos e limitados. A razão disso se dá pela necessidade de especialistas com um bom conhecimento dos dados específicos para anotarem essas imagens, resultando em um processo caro em termos de tempo, dinheiro e esforço.

Dessa forma, a implementação de um modelo automático de diagnóstico com dados restritos e desbalanceados se torna desafiadora. Para mitigar esses problemas, os autores de Islam e Zhang (2020) propuseram uma forma de gerar imagens médicas sintéticas utilizando uma DCGAN. As imagens anotadas foram obtidas do *Alzheimer 's Disease Neuroimaging Initiative* (ADNI) e foram selecionadas 411 imagens do tipo *Positron Emission Tomography* (PET). Dessas imagens, 98 pertenciam à classe da Doença de Alzheimer (AD), 105 ao Controle Normal (NC) e 208 ao Comprometimento Cognitivo Leve (MCI).

Para cada uma das três classes, foi treinada uma DCGAN e cada modelo gerou 64 imagens PET. Essas imagens foram avaliadas pelas métricas *Peak Signal to Noise Ratio* (PSNR) e *Structural Similarity Index* (SSIM), chegando a 32,83 e 77,48 respectivamente. Os autores também apresentaram amostras visuais dos resultados, além de desenvolverem um modelo de CNN 2D de classificação para avaliar o impacto das imagens geradas. Assim, foram considerados dois experimentos, um deles com o conjunto de treinamento contendo apenas as imagens originais, e um segundo experimento complementando os dados de treino com essas 64 imagens de cada classe geradas pelas DCGANs.

Como resultado, o modelo de classificação atingiu 71,45% de acurácia nas classes CN e AD, aumentando 10 pontos percentuais em comparação ao modelo treinado sem as imagens geradas pela DCGAN. Em conclusão, foi observado que o modelo pode ser generalizado para o

diagnóstico de outras doenças com tipos de imagens diferentes, contribuindo para complementar o conjunto de treinamento entregue aos modelos. Além disso, os autores enfatizaram que a geração de imagens médicas sintéticas é uma área de pesquisa promissora e econômica para desenvolver tecnologias de diagnóstico automático.

3.5.5 Geração de Imagens Sintéticas de Células Cervicais

Motivados pela prevenção do câncer cervical, os autores de Yu et al. (2021) propuseram um estudo sobre o uso de dados sintéticos na classificação de imagens celulares. Neste contexto, a proporção de células anormais é reduzida em comparação com a quantidade de células normais (benignas). Dessa forma, os autores enfrentaram o desafio do desbalanceamento entre classes, um fator que pode impactar significativamente diferentes modelos de classificação, especialmente redes neurais.

Inicialmente, os autores criaram o conjunto de dados a partir de imagens microscópicas do teste citológico e particionaram as células únicas nas categorias “normal” e “anormal”. O conjunto de dados contou com imagens de 227 x 227 pixels, sendo verificado duas vezes pelos patologistas. Os dados foram divididos aleatoriamente, resultando em 961 imagens anormais e 16.961 normais e 241 imagens anormais e 3.961 normais nos conjuntos de treino e teste, respectivamente.

Para realizar os experimentos os autores implementaram uma AlexNet para classificação e uma GAN simples de 5 camadas para gerar imagens sintéticas de células anormais. Os experimentos avaliaram o *undersampling* dos dados reais classificados com a AlexNet (chamado de *task 1*); o uso da AlexNet pré-treinada no conjunto de dados ImageNet com adição de *fine-tuning* com os dados do *undersampling* (*task 2*); a geração de 16.000 imagens de células anormais com a GAN, seguida de uma classificação final com a AlexNet contando com 16.961 imagens de cada classe (*task 3*) e o pré-treino da AlexNet realizado nas 16.000 imagens sintéticas de classe anormal e 16.000 imagens reais de classe normal, executando posteriormente o *fine-tuning* com 961 imagens reais de cada classe.

Os autores avaliaram os resultados com as métricas gerais de precisão, sensibilidade, especificidade, acurácia, *f1-score* e AUROC. O terceiro experimento (*task 3*) obteve os melhores resultados em relação à acurácia (95,1%) e o *f1-score* (68,6%), enquanto o quarto experimento (*task 4*) obteve o maior valor de AUROC (98,4%). Assim, os autores concluíram que a geração de amostras sintéticas anormais é uma forma efetiva de resolver o problema de desbalanceamento de classes no contexto de imagens celulares cervicais.

Tabela 3.1: Resumo dos artigos relacionados com a pesquisa. As linhas horizontais duplas separam os trabalhos por categoria, sendo a primeira seção os artigos que lidam com classificação; a segunda, os artigos de segmentação; a terceira, os artigos sobre geração de imagens sintéticas e a última apresenta os dados da nossa pesquisa. Fonte: elaborado pela autora.

Autores	**Tipo de Imagem**	**Modelos**	**Resultados**
Cordeiro et al. (2018)	IHQ de HER-2	KNN, SVM e MLP	Acurácia de 90,84%
Tang et al. (2019)	IHQ Cerebral	CNN	AUROC de 99% e AUPRC de 74%
Rogalsky et al. (2021)	IHQ de ER e PR	SVM	F1-Score de 65% (ER) e de 58% (PR)
Mridha et al. (2022)	H&E e IHQ de HER-2	CNN baseada na Inception-V3	Acurácia de 85% (H&E) e de 88% (IHQ)
Che et al. (2023)	IHQ de HER-2	Classificador de 3 Fases	Acurácia de 97,9%
Mouelhi et al. (2018)	IHQ de ER e PR	Watershed e Métodos de Agregação	Acurácia de 98,9%
Signaevsky et al. (2019)	IHQ Cerebral	CNN SegNet	Acurácia de 81%
Tamaki et al. (2021)	IHQ de ER e PR	Watershed	Análises Visuais
Rmili et al. (2022)	IHQ de NETs	Watershed	F1-Score de 98,6% e Coeficiente Dice de 97,9%
Islam e Zhang (2020)	PET Cerebral	DCGAN e CNN	Acurácia de 71,45%
Yu et al. (2021)	Imagens Celulares Cervicais	GAN e AlexNet	Acurácia de 95,1% e F1-Score de 68,6%
Hong et al. (2021)	Ressonância Magnética Cerebral	3D-StyleGAN	MMD de 4470 e SSIM de 0,93
Gulakala et al. (2022)	Raio-X de Tórax	GAN U-Net Residual e CNN baseada na DenseNet	Acurácia de 99,2%
Mukherkjee et al. (2022)	Ressonância Magnética Cerebral	Agregação de GANs (DCGAN e WGAN) e VGG-19	Acurácia de 93%
Liao et al. (2021)	Ultrassom Renal	ACWGAN-GP e MobileNetV2	Acurácia de 81,9%
Krinski et al. (2023)	TC de Pulmões com Lesões da Covid-19	StarGAN, StyleGAN2ADA, RegNetx-002 e U-net++	F-Score de 87,37% e IoU de 81,11%
Trabalho Proposto	IHQ de ER e PR	StyleGAN2ADA, SVM, CNN, DenseNet, ViT, DreamSIM e CRC	Acurácia e F1-Score de 95,45% e 87,46% (ER) e 97,27% e 92,22% (PR)

4 MATERIAIS E MÉTODOS

Nas subseções abaixo iremos descrever os materiais e métodos necessários para realizar os experimentos que compõem a nossa metodologia. Iniciamos descrevendo os conjuntos de dados escolhidos (subseção 4.1) e, em seguida, apresentamos duas formas de realizar o aumento de dados, com uma técnica simples (subseção 4.2) e com uma rede generativa (subseção 4.3). Para compreendermos a qualidade das imagens sintéticas da rede generativa, definimos duas avaliações visando considerar aspectos quantitativos (subseção 4.4) e qualitativos (subseção 4.5) das imagens. Por fim, apresentamos os modelos de classificação implementados para categorizar os níveis de câncer de mama (subseções 4.6, 4.7, 4.8 e 4.9).

4.1 VISÃO GERAL DOS CONJUNTOS DE DADOS

Visando analisar a viabilidade dos métodos estudados (capítulo 2), realizamos os experimentos com os conjuntos de dados que compõem o *HistoBC-HR* de Rogalsky (2021). Para isso, selecionamos os dados dos *patches* de Receptores de Estrogênio (ER) e dos Receptores de Progesterona (PR), além de escolhermos a pontuação IS como alvo. Podemos justificar as nossas decisões pelo fato da pontuação IS apresentar um alto desbalanceamento de classes nesses dados (Tabela 4.1), nos permitindo a possibilidade de estudar formas de mitigar este problema. Além disso, como esses conjuntos de dados foram criados em parceria com os patologistas do Hospital Erasto Gaertner, os especialistas detinham um conhecimento aprofundado sobre as patologias das imagens, viabilizando a realização do nosso experimento qualitativo (seção 4.5).

Tabela 4.1: Distribuição das classes IS dos *patches* das imagens de Receptores de Estrogênio (ER) e Progesterona (PR) que compõem o *HistoBC-HR* de Rogalsky (2021). Fonte: elaborado pela autora.

	Número de Amostras de Cada Classe IS				
Tipo de Exame	0	1+	2+	3+	Total
ER	414	149	293	945	1801
PR	515	171	226	713	1625

Esses dados contam com imagens de 400 x 300 pixels que foram redimensionadas para o tamanho de 256 x 256 pixels (*StyleGAN2ADA* aceita somente entradas quadradas pertencentes às potências de 2). Essas imagens estão associadas a 4 classes diferentes da pontuação de intensidade (IS) que podem ser conferidas na Figura 4.1, sendo as distribuições das classes apresentadas na Tabela 4.1. Mais detalhes sobre o conjunto de dados podem ser encontrados na seção 3.5.2.

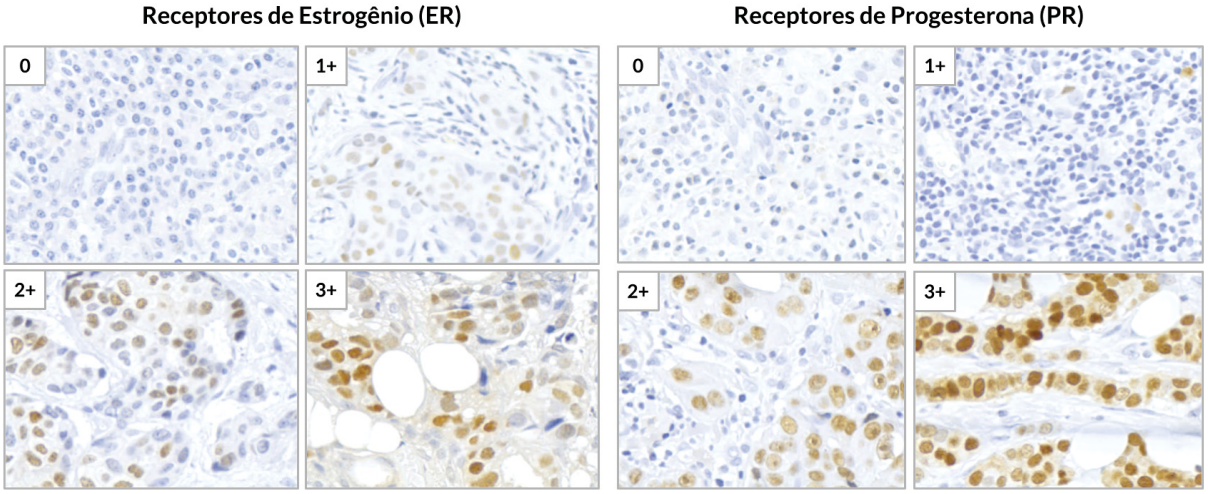


Figura 4.1: Exemplos de cada classe IS das imagens dos biomarcadores de ER e PR do *HistoBC-HR*. Cada imagem está associada à um valor da pontuação IS, sendo a classe 0 considerada uma amostra negativa; a classe 1+ como fracamente positiva; a classe 2+ como moderadamente positiva e a classe 3+ como fortemente positiva. Fonte: elaborado pela autora.

4.2 GERAÇÃO DE IMAGENS SINTÉTICAS COM AUTOAUGMENT

Considerando o cenário de análise de imagens médicas, a quantidade limitada de imagens rotuladas, a falta de variabilidade nos dados e o desbalanceamento das classes presentes nesses rótulos são problemas frequentes. Observando o conjunto de dados *HistoBC-HR*, todos esses problemas foram encontrados. Por exemplo, a classe 1+ possui apenas 149 exemplos no biomarcador ER, enquanto a classe 3+ tem 945 imagens, demonstrando pouca variabilidade e limitação do número de exemplos da classe 1+ e um desbalanceamento entre ela e a classe 3+.

Questões como escassez de dados anotados e desbalanceamento de classes têm sido resolvidos com técnicas simples de aumento de dados aplicadas às imagens de treinamento, como rotações, translações, espelhamentos, transformações geométricas e alterações de cores. Com tantas opções disponíveis, os autores de Cubuk et al. (2019) propuseram uma forma automática para definir quais técnicas utilizar, chamada de AutoAugment, descrita na seção 2.7.

Assim, visando avaliar uma técnica simples para realizar o aumento das imagens de câncer de mama, foi utilizado o AutoAugment fornecido pela biblioteca *Pytorch*. A versão escolhida considerou pesos pré-treinados no conjunto de dados *CIFAR-10*, que conta com transformações como inversões, contraste, rotações, translações, ajustes de brilho, alterações de cores e equalizações com diferentes probabilidades de uso e magnitude das operações. Mais detalhes sobre as transformações podem ser conferidos em Cubuk et al. (2019).

4.3 GERAÇÃO DE IMAGENS SINTÉTICAS COM GANS

Além dos métodos simples de aumento de dados, as GANs estão entre as técnicas mais recentes capazes de gerar imagens sintéticas. Dentre as arquiteturas atuais, a *Style Generative Adversarial Network* (StyleGAN) se destacou na geração de imagens faciais de alta qualidade, além de permitir combinações de estilo para gerar novas imagens. Devido à quantidade limitada de dados, treinar um modelo para gerar imagens de alta qualidade se tornou um novo desafio. Como solução, realizamos um experimento com o uso da StyleGAN2*ADA*, projetada para lidar com poucas imagens de entrada.

Neste experimento, realizamos a separação das imagens de mesma classe, formando quatro conjuntos de treinamento (0, 1+, 2+ e 3+) para cada tipo de exame (ER e PR). Então, treinamos uma StyleGAN2*ADA* específica para cada classe e geramos novas imagens para adicionarmos ao conjunto de treino original (mais informações na seção 5.3). Por exemplo, se o conjunto de treino inicial contasse com um total de 800 imagens e fossem geradas 100 imagens em cada uma das quatro StyleGAN2*ADAs*, o conjunto de dados final conteria 1200 imagens. O fluxo geral de treinamento e geração de imagens sintéticas pode ser observado na Figura 4.2.

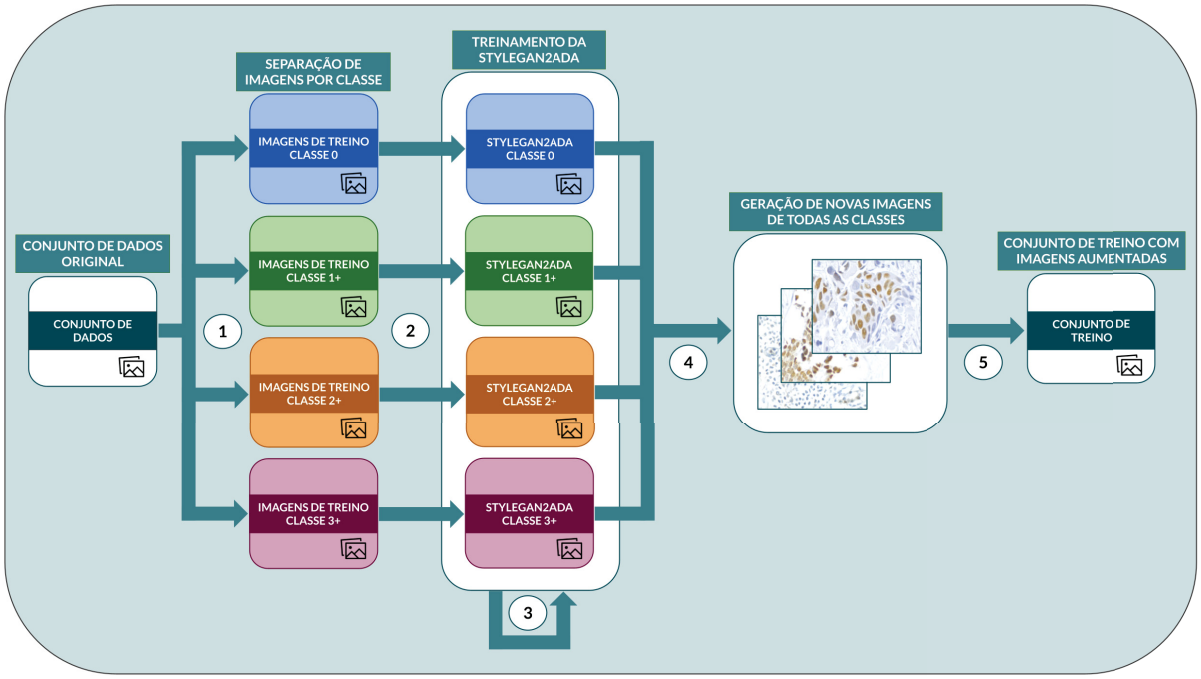


Figura 4.2: Fluxo geral do uso da *StyleGAN2ADA* na geração de imagens sintéticas. Em 1, o conjunto de dados passou por uma divisão visando separar as imagens de mesma classe, formando quatro novos conjuntos de imagens; em 2, os conjuntos de imagens foram entregues às quatro *StyleGAN2ADAs* para realizar a etapa de treino; em 3, os modelos foram treinados para cada conjunto de imagens; em 4, foram geradas imagens de cada uma das quatro *StyleGAN2ADAs* que foram inseridas em um conjunto de treino, em 5, que será entregue a um classificador posteriormente. Fonte: elaborado pela autora.

4.4 ANÁLISE QUANTITATIVA DAS IMAGENS SINTÉTICAS

Para avaliar a qualidade das imagens geradas pela rede StyleGAN2ADA, aplicamos uma análise quantitativa utilizando a métrica DreamSIM. A escolha da métrica levou em consideração o experimento nos dados de ER apresentado na Tabela 4.2. Neste experimento, selecionamos 15 imagens de cada classe do conjunto de validação, totalizando 60 imagens reais. Cada uma dessas 60 imagens foram comparadas com todas as imagens reais do conjunto de treino, usando as métricas tradicionais de similaridade (PSNR e SSIM) e a métrica DreamSIM. Para cada imagem de validação, nós armazenamos a classe da imagem mais similar pertencente ao conjunto de treino segundo as métricas (os menores valores para DreamSIM e os maiores valores para PSNR e SSIM). Depois da execução, nós contamos as vezes em que a classe da imagem de validação correspondia à classe da imagem de treino considerada mais similar (acertos) e também as vezes em que as classes diferiam (erros).

Como as métricas SSIM e PSNR são baseadas na localização dos pixels, elas apresentaram dificuldades em encontrar imagens da mesma classe no conjunto de treinamento. Essas

dificuldades podem estar relacionadas com os desafios das imagens de IHQ, que apresentam diferentes rotações, cores, escalas e formatos de agrupamentos de células. Portanto, métricas que funcionam bem em outros tipos de imagens que mantêm a mesma orientação, como raio-x e tomografia, não alcançam bons resultados no cenário de IHQ. Já a DreamSIM, que treinou levando em consideração imagens similares a uma imagem de referência, apresentou um desempenho superior, com 55 acertos e apenas 5 erros nas 60 imagens.

Após realizar a escolha da métrica, nós definimos uma sequência de passos para avaliar as imagens sintéticas (Figura 4.3). Inicialmente, utilizamos a StyleGAN2ADA treinada para gerar 100 imagens de cada classe da pontuação IS (passo 1 da Figura 4.3), totalizando 400 imagens sintéticas. Então, comparamos cada imagem sintética com todas as imagens do conjunto de treinamento original com a métrica DreamSIM (passo 2 da Figura 4.3). Como as características de imagens da mesma classe apresentam grande variabilidade, nós escolhemos a imagem real do conjunto de treino com o menor valor da métrica DreamSIM (mais similar) em relação à imagem sintética avaliada (passo 4 da Figura 4.3). Por exemplo, aplicamos a métrica DreamSIM para comparar a imagem sintética i (imagem_g_i) com todo o conjunto de treinamento. Como resultado, teremos a imagem real i (imagem_r_i), que apresentou a maior similaridade com a imagem sintética i. Como valor final, calculamos a média considerando todos os pares de imagens (imagem_g_i e imagem_r_i) encontrados.

Tabela 4.2: Descrição dos experimentos quantitativos realizados. Fonte: elaborado pela autora.

Métrica	Acertos	Erros	Acurácia	Descrição do Erro
DreamSim	55	5	91,66%	Confusões entre a classe 1+ com 0 e 2+ e dificuldades entre a classe 2+ com 3+.
SSIM	17	43	28,33%	Dos 60 exemplos, indicou que 53 são mais similares à classe 1+.
PSNR	22	38	36,66%	Dos 60 exemplos, indicou que 50 são mais similares à classe 0.

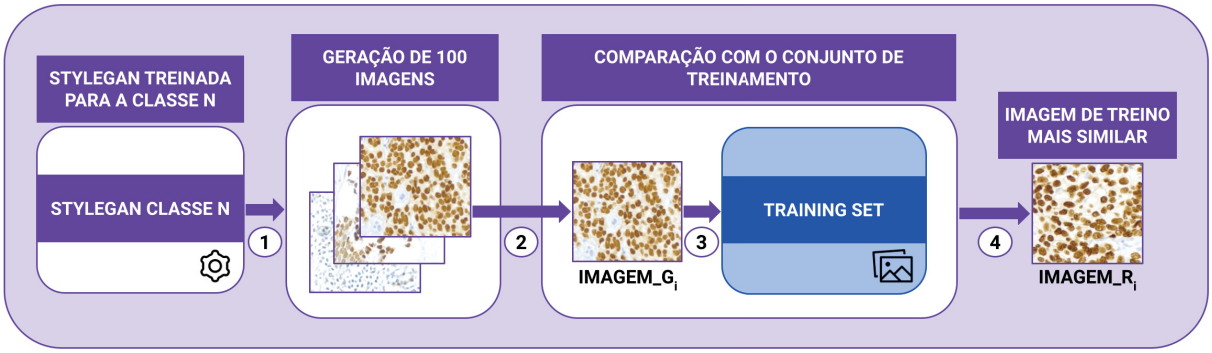


Figura 4.3: Fluxo das etapas da análise quantitativa utilizando a métrica DreamSIM. Fonte: elaborado pela autora.

4.5 ANÁLISE QUALITATIVA DAS IMAGENS SINTÉTICAS

Para realizar a análise qualitativa das imagens, utilizamos o método de Categorização Rápida de Cenas (CRC). O CRC se originou da ideia de que humanos podem reportar as principais

características de uma cena em um relance (Borji, 2018). Então, nos baseamos no método proposto por Denton et al. (2015) para construir uma interface (Figura 4.4) e convidamos três especialistas em patologia do Hospital Erasto Gaertner para participar desta análise.

Para construir a interface, nós consideramos duas configurações de experimentos. Inicialmente, selecionamos 18 imagens geradas de cada classe com o menor valor da métrica DreamSIM em relação ao conjunto de treino, totalizando 72 imagens sintéticas. Além disso, selecionamos aleatoriamente 7 imagens reais de cada classe, totalizando 28 imagens originais.

Nós embaralhamos aleatoriamente as 100 imagens e as apresentamos individualmente na interface. As imagens foram apresentadas na tela por 3 segundos, e o objetivo dos especialistas foi indicar quando, na opinião deles, a imagem era real ou gerada. Cada especialista realizou o teste individualmente nas mesmas condições de equipamento e iluminação.

Para definir o limite de 3 segundos, nos baseamos no limite de tempo proposto por Denton et al. (2015) e nos resultados de experimentos iniciais com os especialistas. Dessa forma, o teste visual com 3 segundos permitiu uma percepção geral de todos os objetos de interesse nas imagens e preveniu que os especialistas prestassem atenção aos detalhes.

Por fim, para avaliar a qualidade das imagens sintéticas de forma geral (não só das imagens mais similares ao conjunto de treino), nós selecionamos aleatoriamente 12 imagens sintéticas de cada classe (totalizando 48 imagens sintéticas) e 12 imagens reais aleatórias de cada classe (totalizando 48 imagens reais) para realizar o mesmo experimento com os mesmos especialistas. Visando garantir a confiabilidade do experimento, nós divulgamos os acertos e erros dos especialistas somente após a realização do último experimento.

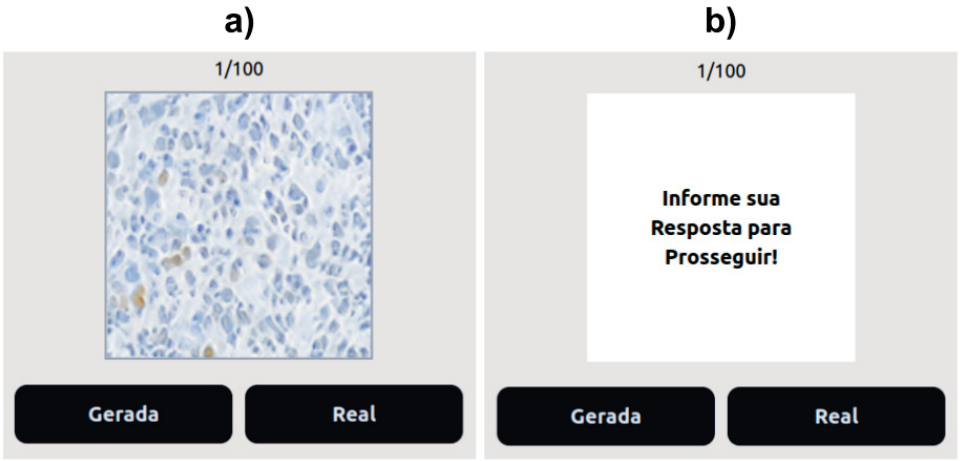


Figura 4.4: Exemplo da interface de análise qualitativa utilizada pelos três especialistas. A imagem é apresentada na tela em *a)*, e, depois de 3 segundos, é substituída por uma tela branca em *b)*, até que o participante indique sua resposta (gerada ou real). Fonte: elaborado pela autora.

4.6 METODOLOGIA ROGALSKY PARA CLASSIFICAÇÃO

Com os conjuntos de dados e os métodos de aumento definidos, implementamos o primeiro experimento com base no método de classificação proposto por Rogalsky (2021). A sequência de etapas que compõem esse método pode ser observada na Figura 4.5. Inicialmente, submetemos a imagem original a um realce de contraste com o método CLAHE, uma equalização do histograma local com limite de contraste. Como segundo passo implementamos a limiarização, que iniciou com a aplicação de um filtro de borramento gaussiano para nivelar gradientes e evitar amplificação de ruído. Então, convertemos a imagem para a escala de cinza e aplicamos a técnica

Otsu, uma abordagem que encontrou um valor ótimo para realizar a limiarização (componente de pré-processamento na etapa 1 da Figura 4.5).

O uso da técnica *Otsu* na etapa anterior permitiu a realização de uma segmentação inicial das células na imagem. Assim, utilizamos essa segmentação para transformar o espaço de cores em HSV, separando os canais H, S, e V, e extraímos as células positivas (em marrom) e negativas (em azul) com uma separação de cores (componentes de segmentação e separação de cores nas etapas 2 e 3 da Figura 4.5). A partir dos canais H e S e da separação de cores, calculamos os histogramas de intensidade, que passaram por uma normalização para manter os valores entre 0 e 100 (componente de extração de características na etapa 4 da Figura 4.5).

Assim, com os histogramas 2D de intensidade, obtemos 960 características para cada uma das imagens do conjunto de dados. Nós dividimos os dados em conjunto de treino, validação e teste de acordo com as especificações da seção 5.3. Então, entregamos os conjuntos de treino e validação ao modelo SVM compondo a etapa de treinamento (componente de classificação na etapa 5 da Figura 4.5). Durante esta fase, o conjunto de validação viabilizou a otimização de hiperparâmetros do modelo. Por fim, passamos o conjunto de teste ao SVM treinado (componente de classificação na etapa 6 da Figura 4.5), e calculamos as métricas gerais e por classe, como precisão, *recall* e *f1-score* (componente de métricas na etapa 7 da Figura 4.5). Em resumo, nosso objetivo com este modelo consistiu apenas na classificação dos *patches* considerando as 4 classes da pontuação IS.

Durante a implementação dos métodos propostos por Rogalsky (2021), encontramos algumas divergências. No código implementado pela autora, houve um pequeno equívoco e alguns rótulos foram inseridos nos vetores de características, resultando em métricas perfeitas. Dessa forma, ao ajustarmos os vetores de características, não alcançamos o *f1-score* de 100% nos experimentos com o mesmo conjunto de dados. Com essa descoberta, podemos dizer que a classificação das imagens de IHQ desse conjunto de dados ainda não é um problema bem resolvido, o que nos motivou a realizar novos experimentos.

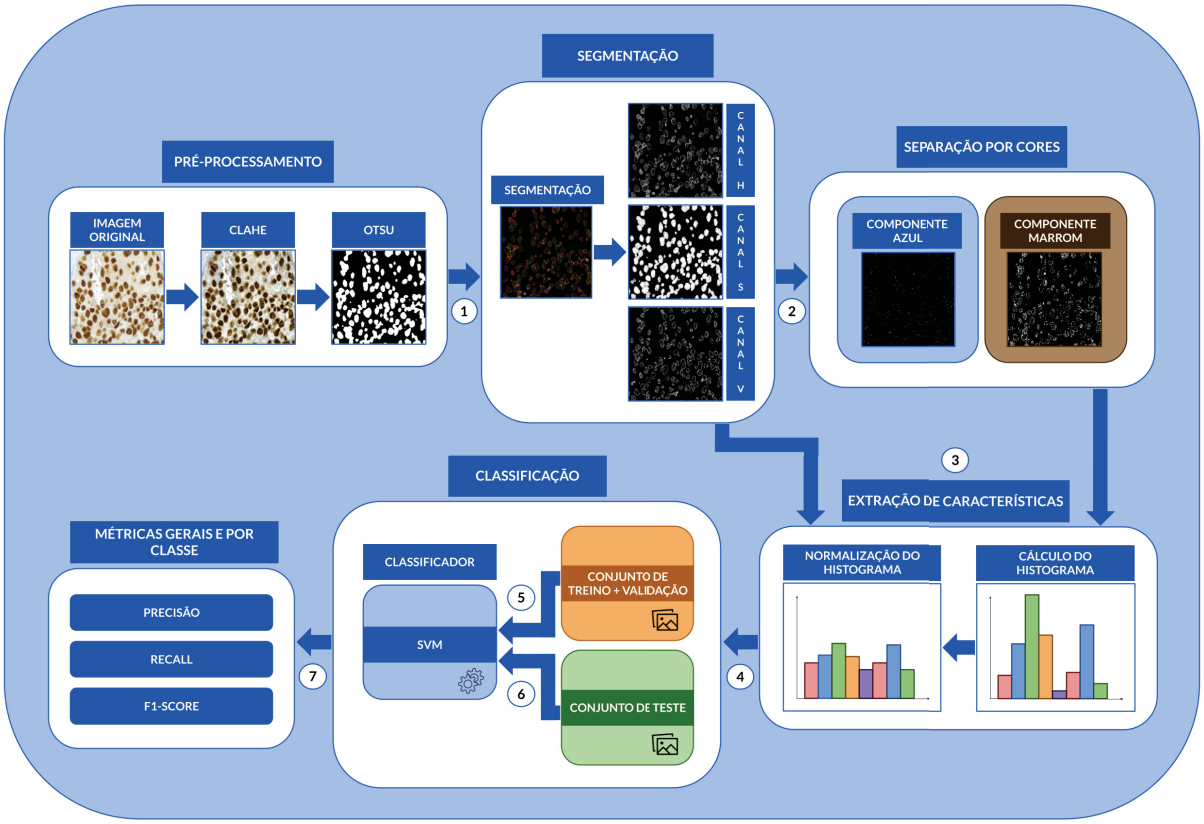


Figura 4.5: Fluxo geral do conjunto de métodos implementados na Metodologia Rogalsky. Fonte: elaborado pela autora.

4.7 MÉTODO COM CNN PARA CLASSIFICAÇÃO

Com a popularização das WSIs de imuno-histoquímica, os modelos de *Deep Learning* se sobressaíram, principalmente as CNNs (Cordeiro et al., 2018; Tang et al., 2019; Mridha et al., 2022). A propagação destes modelos se explicou pela capacidade dessas redes de extraírem automaticamente as características, alcançando resultados similares ou melhores do que um conjunto de métodos manualmente ajustados para a extração de características. Tendo em vista este cenário, implementamos uma forma alternativa de classificação utilizando uma CNN, que foi baseada na arquitetura apresentada em Tang et al. (2019).

Inicialmente, dividimos o conjunto de dados em conjunto de treino, conjunto de validação e conjunto de teste de acordo com as definições da seção 5.3. Então, utilizamos os conjuntos de treino e validação no treinamento de uma CNN com a arquitetura consistindo em 6 camadas de convolução e *Max Pooling*, seguidas por 2 camadas densas com 100 neurônios. Definimos o número de neurônios nas camadas ocultas sendo 64, 64, 128, 256, 256 e 512. Nós consideramos um *dropout* de 0,2, a resolução da imagem de 224 x 224 pixels, uma taxa de aprendizado de 0,00008, o otimizador *Adam* e a função de perda *Multi Label Soft Margin Loss*. A saída da rede indicou a probabilidade da imagem de entrada pertencer a cada uma das classes IS (0, 1+, 2+ e 3+). Avaliamos as previsões da rede com as mesmas métricas descritas na subseção anterior. A arquitetura e todas as etapas do fluxo de funcionamento da CNN são apresentadas na Figura 4.6.

4.8 MÉTODO COM DENSENET PARA CLASSIFICAÇÃO

Com o investimento no desenvolvimento de modelos de *Deep Learning*, muitas arquiteturas foram propostas para otimizar questões de desempenho, com diferentes tipos de normalizações, número e profundidade de camadas e uso das informações de contexto em uma imagem (Huang et al., 2017; Dosovitskiy et al., 2021). Assim, para alinhar o nosso classificador com as arquiteturas pertencentes ao estado da arte, implementamos uma variação da DenseNet com 121 camadas, chamada de DenseNet121.

Para construir a DenseNet121, nós utilizamos o *Pytorch*, que é uma biblioteca aberta de aprendizado de máquina. Nós aplicamos um *fine-tuning* por 30 épocas, considerando como base os pesos pré-treinados no conjunto de dados ImageNet1K. Os hiperparâmetros foram mantidos conforme as recomendações da biblioteca, com a função de perda sendo a Entropia Cruzada, o tamanho da imagem sendo 224 x 224 pixels e o otimizador Adam. Em relação à divisão dos dados e as métricas, a DenseNet seguiu os mesmos passos descritos na seção da CNN (subseção 4.7). A arquitetura simplificada e o fluxo de execução podem ser observados na Figura 4.6.

4.9 MÉTODO COM VIT PARA CLASSIFICAÇÃO

Contando com os mecanismos de atenção, os Transformers revolucionaram o campo de Processamento de Linguagem Natural. Recentemente, esses modelos extrapolaram os limites de área de aplicação, sendo utilizados para classificação de imagens, reconhecimento de fala, detecção de objetos e aprendizagem por reforço (Islam et al., 2024). Visando trazer uma alternativa recente para as CNNs, testamos um *Vision Transformer* (ViT) para realizar a classificação das imagens de IHQ de câncer de mama.

O ViT que escolhemos (*ViTB16*) também foi implementado com a biblioteca *Pytorch* e o processo de *fine-tuning* foi realizado com base nos pesos pré-treinados no conjunto de dados *ImageNet1K*. Nós definimos a normalização das imagens para o tamanho 224 x 224 pixels e mantivemos as configurações recomendadas pela biblioteca, como uma taxa de aprendizado de 0,001, a Entropia Cruzada como função de perda e o Adam como otimizador. Nós avaliamos o desempenho do modelo em experimentos iniciais e definimos o número de épocas para 60. Inicialmente, nós dividimos os dados de treinamento, validação e teste conforme os passos descritos na subseção 4.7, assim como a escolha das métricas. A arquitetura simplificada pode ser conferida na Figura 4.6.

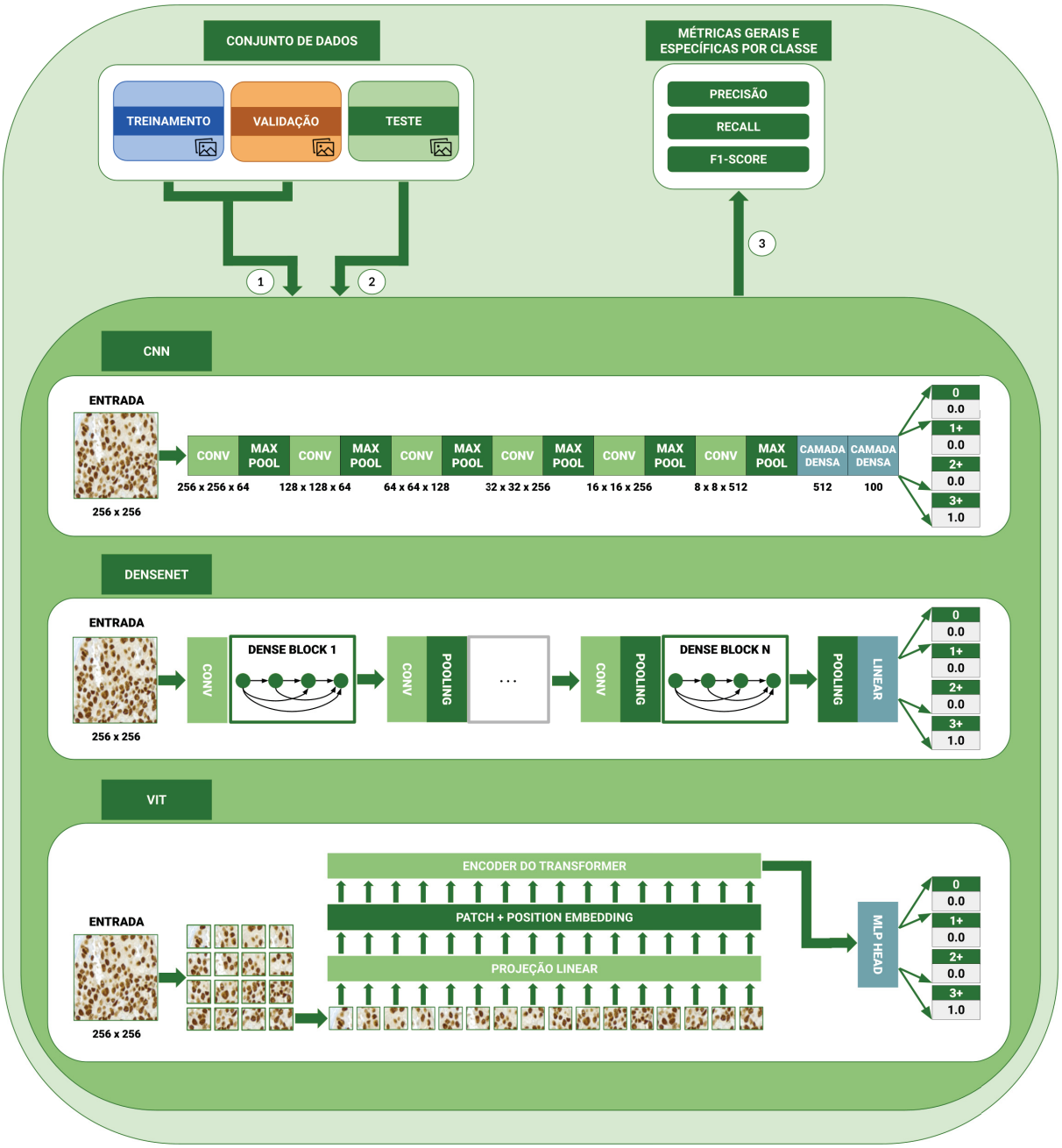


Figura 4.6: Arquitetura geral e fluxo dos classificadores CNN, DenseNet e ViT. Em 1, os modelos realizam o processo de treinamento com os conjuntos de treino e validação; em 2, o conjunto de teste é entregue aos modelos treinados; e em 3, os modelos são avaliados pelas métricas. Fonte: elaborado pela autora.

5 EXPERIMENTOS

Considerando a pequena quantidade de amostras e o desbalanceamento de classes dos conjuntos de dados de Receptores de Estrogênio (ER) e Progesterona (PR), iniciamos os nossos experimentos investindo na geração de imagens sintéticas. Para isso, adotamos os modelos AutoAugment e StyleGAN2ADA, descritos no capítulo anterior. Escolhemos a StyleGAN2ADA por pertencer ao estado da arte, além de ser projetada para lidar com conjuntos de dados de tamanho reduzido (Karras et al., 2020a). Por questões de comparação, selecionamos o AutoAugment como uma forma simples de aumento de dados, que aplica pesos pré-treinados em transformações básicas de processamento de imagens (Cubuk et al., 2019).

Após o treinamento e ajuste dos métodos de aumento de dados, realizamos avaliações quantitativas, qualitativas e análises visuais para compreender a qualidade e as características das imagens sintéticas. Além disso, organizamos cinco experimentos para classificar ambos os conjuntos de dados, utilizando as técnicas: Metodologia Rolgalsky, CNN, DenseNet e ViT, descritas e justificadas no capítulo 4. Por fim, analisamos cada um dos experimentos de acordo com métricas gerais e específicas por classe, incluindo precisão, *recall*, *f1-score* e acurácia. Nas próximas seções, organizamos e apresentamos os conceitos importantes para a definição e entendimento dos experimentos realizados nesta pesquisa.

5.1 AVALIAÇÃO DO TREINAMENTO DA STYLEGAN2ADA COM KID

Iniciando pelo treinamento da StyleGAN2ADA, optamos pelo uso da métrica *Kernel Inception Distance* (KID) ao invés da *Fréchet Inception Distance* (FID) para avaliar a qualidade das imagens produzidas pela rede ao longo das épocas. Devido às propriedades da FID, essa métrica pode ser mais sensível à variações na imagem de entrada, pois assume que as características extraídas pela rede *Inception* seguem uma distribuição Gaussiana (Heusel et al., 2018). Assim, pequenos ruídos podem causar grandes mudanças na forma desta distribuição, resultando em um aumento excessivo na distância *Fréchet* calculada.

Para mitigar essas questões, implementamos a métrica KID para monitorar o processo de treinamento da StyleGAN2ADA. A KID calcula a discrepância média máxima quadrada (MMD) entre as características das imagens reais e das imagens geradas, não dependendo diretamente das médias e matrizes de covariância das características extraídas. Dessa forma, a KID tende a apresentar distâncias no espaço das características que não mudam drasticamente com pequenas variações, resultando em uma métrica mais robusta e menos sensível aos ruídos (Horak et al., 2020).

5.2 TREINAMENTO DA STYLEGAN2ADA SEM PESOS PRÉ-TREINADOS

Visando uma maior qualidade nas imagens sintéticas produzidas pela StyleGAN2ADA, consideramos treiná-la a partir de um *finetuning* dos pesos pré-treinados no conjunto de dados BreCaHAD, que contém 1944 imagens celulares de 512 x 512 pixels (Aksac et al., 2019). Apresentamos alguns exemplos de imagens reais e sintéticas obtidas pelos autores após realizarem o treinamento com o conjunto de dados BreCaHAD (Figura 5.1). Como os pesos disponibilizados por Karras et al. (2020a) foram obtidos de uma arquitetura de 512 x 512 pixels, aplicamos uma interpolação bilinear para aumentar as imagens dos *patches* de 256 x 256 para o

tamanho necessário. No entanto, o resultado foi insatisfatório nos experimentos iniciais, com KIDs próximos à 30%.

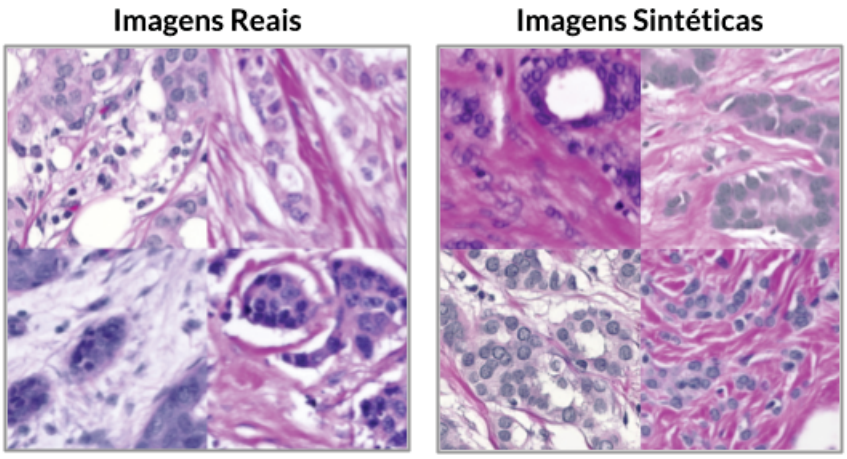


Figura 5.1: Exemplo das imagens do conjunto de dados BreCaHAD (esquerda) e das imagens sintéticas geradas pela StyleGAN2ADA treinada pelos autores de Karras et al. (2020a) (direita). Fonte: adaptado de Karras et al. (2020a).

Dessa forma, acreditamos que a pequena quantidade de amostras nos conjuntos de dados (ER e PR) e o aumento das dimensões para se ajustarem aos pesos inviabilizou a geração de imagens de qualidade em resoluções maiores. Assim, desconsideramos os pesos pré-treinados e mantivemos a StyleGAN2ADA com uma arquitetura para imagens de 256 x 256 pixels. Optamos por considerar 2000kimg para o treinamento de todas as StyleGAN2ADAs para cada uma das quatro classes (0, 1+, 2+ e 3+). Escolhemos esta quantidade porque foi o ponto no qual estabilizamos os valores de perda da rede durante o processo de treinamento.

5.3 ORGANIZAÇÃO DO PROCESSO DE TREINO COM VALIDAÇÃO CRUZADA

Para organizar o processo de treinamento dos classificadores, optamos pela técnica de validação cruzada de cinco pastas. Escolhemos esta técnica devido aos tamanhos reduzidos dos conjuntos de dados, que podem levar à interpretações errôneas dos resultados com o uso de outras técnicas, como a *Holdout* (Maleki et al., 2020). Com a validação cruzada, variamos os conjuntos de treinamento, validação e teste, e calculamos as médias de cada uma das cinco combinações, garantindo uma estimativa mais confiável do desempenho dos modelos de classificação. Definimos cinco pastas visando manter o equilíbrio entre a confiabilidade dos resultados e também o tempo de treinamento e execução dos modelos.

No início da nossa implementação, dividimos o conjunto de dados em cinco pastas e utilizamos quatro delas para treino. Diferentemente da validação cruzada usual, nós selecionamos a última pasta para compor o conjunto de teste ao invés da validação. Então, das quatro pastas de treinamento, escolhemos uma delas para ser utilizada como conjunto de validação. Apresentamos a ordem de seleção das pastas na Figura 5.2. Adaptamos a técnica com o objetivo de aumentar a variância dos dados de teste, apresentando uma estimativa de desempenho mais próxima à realidade.

5.4 AJUSTE DOS HIPERPARÂMETROS

Com os métodos de aumento de dados e de classificação definidos, focamos na otimização de alguns hiperparâmetros. Para ajustar o fator de truncamento da StyleGAN2ADA, o ideal seria

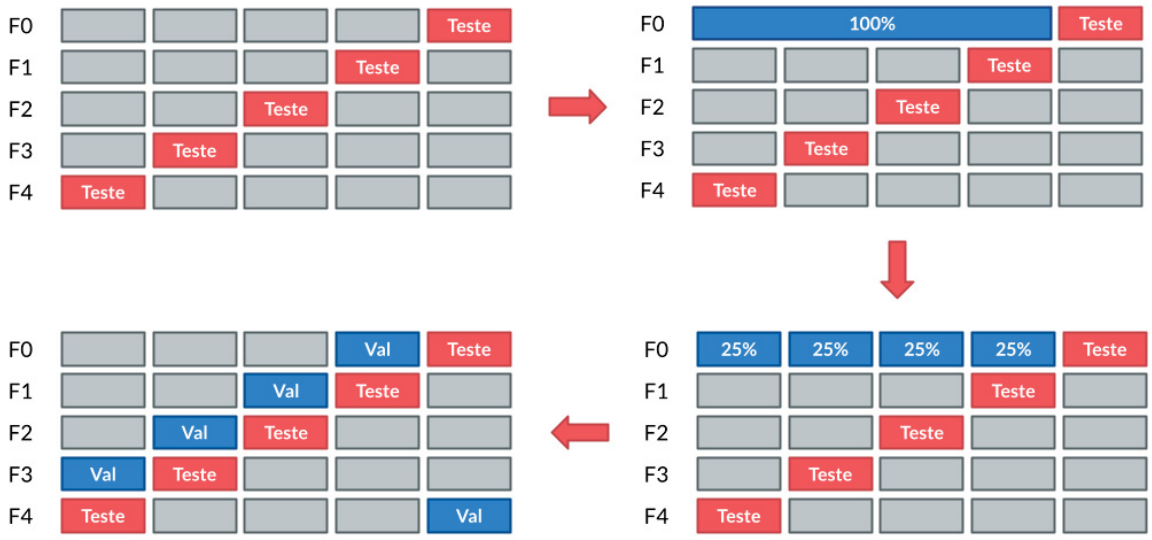


Figura 5.2: Exemplo do fluxo de construção das pastas da validação cruzada que utilizamos nos experimentos. Após definirmos as cinco pastas de teste em cada iteração (em rosa), selecionamos uma das pastas restantes para compor o conjunto de validação (em azul). Por fim, construímos o conjunto de treinamento com os dados restantes de cada iteração (em cinza). Fonte: elaborado pela autora.

avaliar quantitativamente e qualitativamente as imagens sintéticas com diferentes valores entre 0 e 1. Por questões de tempo, optamos por realizar os experimentos com um truncamento de 0,5. Esse valor considera o aumento da variância nas imagens sintéticas ao mesmo tempo que mantém as características das imagens originais (Karras et al., 2020a). Também baseamos a nossa escolha de acordo com experimentos iniciais de classificação, considerando valores de truncamento entre 0,3 até 0,7 no conjunto de dados de Receptores de Estrogênio (ER).

Em relação aos classificadores, consideramos o ajuste da taxa de aprendizado e também a adição da *early stopping* com o objetivo de melhorar a adaptação dos modelos para diferentes quantidades de dados. A *early stopping* interrompe o processo de treinamento caso ocorram sinais de *overfitting* (Bai et al., 2021). Na nossa implementação, o treinamento é interrompido se ocorrerem mais de três casos em que a perda da validação apresente uma piora de 10% em relação à época com menor perda. Já a taxa de aprendizado é reduzida pela metade em todos os três casos permitidos pela *early stopping*. Assim, o modelo se ajusta de forma a dar passos menores de aprendizado, refinando o processo de treinamento. Escolhemos os valores mencionados de acordo com experimentos menores e com a análise da curva de aprendizado.

5.5 EXPERIMENTOS DE ANÁLISE DAS IMAGENS SINTÉTICAS

Após treinarmos as StyleGAN2ADAs para todas as classes da pontuação IS considerando ambos conjuntos de dados, definimos três experimentos e uma análise visual para avaliar a qualidade das imagens sintéticas. Em relação à análise visual (VS), buscamos compreender as características das imagens geradas pelo AutoAugment e também pela StyleGAN2ADA, além de apresentar amostras das imagens obtidas pelas técnicas tanto no cenário de Receptores de Estrogênio (ER) quanto de Receptores de Progesterona (PR).

Sobre os experimentos, investimos em um teste quantitativo (QT) com o uso da métrica DreamSIM (seção 4.4) e duas avaliações qualitativas (QL1 e QL2) com o uso do método CRC (seção 4.5). Por questões de tempo e disponibilidade dos especialistas, realizamos

esses experimentos somente para as imagens dos Receptores de Estrogênio (ER) produzidas pela StyleGAN2ADA. Nós descrevemos cada experimento e a análise visual na Tabela 5.1 e representamos visualmente os fluxos de execução na Figura 5.3.

Tabela 5.1: Descrição dos experimentos de análise quantitativa, qualitativa e visual das imagens sintéticas. Fonte: elaborado pela autora.

Exp.	Descrição
VS	Análise visual das amostras aleatórias de imagens reais e sintéticas, tanto da StyleGAN2ADA quanto do AutoAugment.
QT	Análise quantitativa, com o cálculo da métrica DreamSIM para avaliar a similaridade perceptiva entre imagens sintéticas e reais.
QL1	Seleção de 28 imagens reais aleatórias e 72 imagens sintéticas mais similares com as imagens do conjunto de treino conforme a métrica DreamSIM. Realização do experimento de Categorização Rápida de Cenas (CRC) com limite de 3 segundos.
QL2	Seleção de 48 imagens reais e 48 imagens sintéticas de forma aleatória. Realização do experimento de Categorização Rápida de Cenas (CRC) com limite de 3 segundos.

5.6 DESCRIÇÃO DOS EXPERIMENTOS DE CLASSIFICAÇÃO

Quanto à pontuação automática dos níveis de câncer das imagens de Receptores de Estrogênio (ER) e Receptores de Progesterona (PR), nós definimos cinco experimentos para estudar o impacto da adição de imagens sintéticas no processo de treinamento dos modelos de classificação. Para isso, consideramos o AutoAugment e a StyleGAN2ADA como métodos de aumento de dados e selecionamos os modelos Metodologia Rogalsky, CNN, DenseNet e ViT para categorizar as imagens nos quatro valores da pontuação IS. Apresentamos as descrições destes experimentos na Tabela 5.2.

A Tabela 5.3 mostra o número de amostras de cada classe em cada um dos experimentos. Para os experimentos E2 e E3, adicionamos 100 imagens sintéticas ao conjunto de treinamento com o modelo AutoAugment e StyleGAN2ADA, respectivamente. Nesta etapa, nos concentramos em avaliar o impacto de uma pequena quantidade de aumento de dados, de forma que o número de imagens geradas não ultrapassasse o número de imagens reais em cada classe. Como a classe minoritária do conjunto de dados de ER possui 123 exemplos de treino e a PR conta com 102 exemplos, selecionamos o valor de 100 imagens sintéticas.

Os experimentos E4 e E5 focam no balanceamento de classes do conjunto de treinamento, adicionando imagens sintéticas até as classes minoritárias se igualarem em quantidade à classe majoritária. Para garantir que todas as classes recebessem imagens sintéticas, adicionamos mais 100 imagens geradas para cada classe após o balanceamento. Realizamos todos os experimentos com os quatro classificadores e comparamos os resultados com E1, que utilizou os dados originais do conjunto de dados, sem a adição de imagens sintéticas. Os experimentos E1, E2, E3, E4 e E5 são apresentados na Figura 5.4.

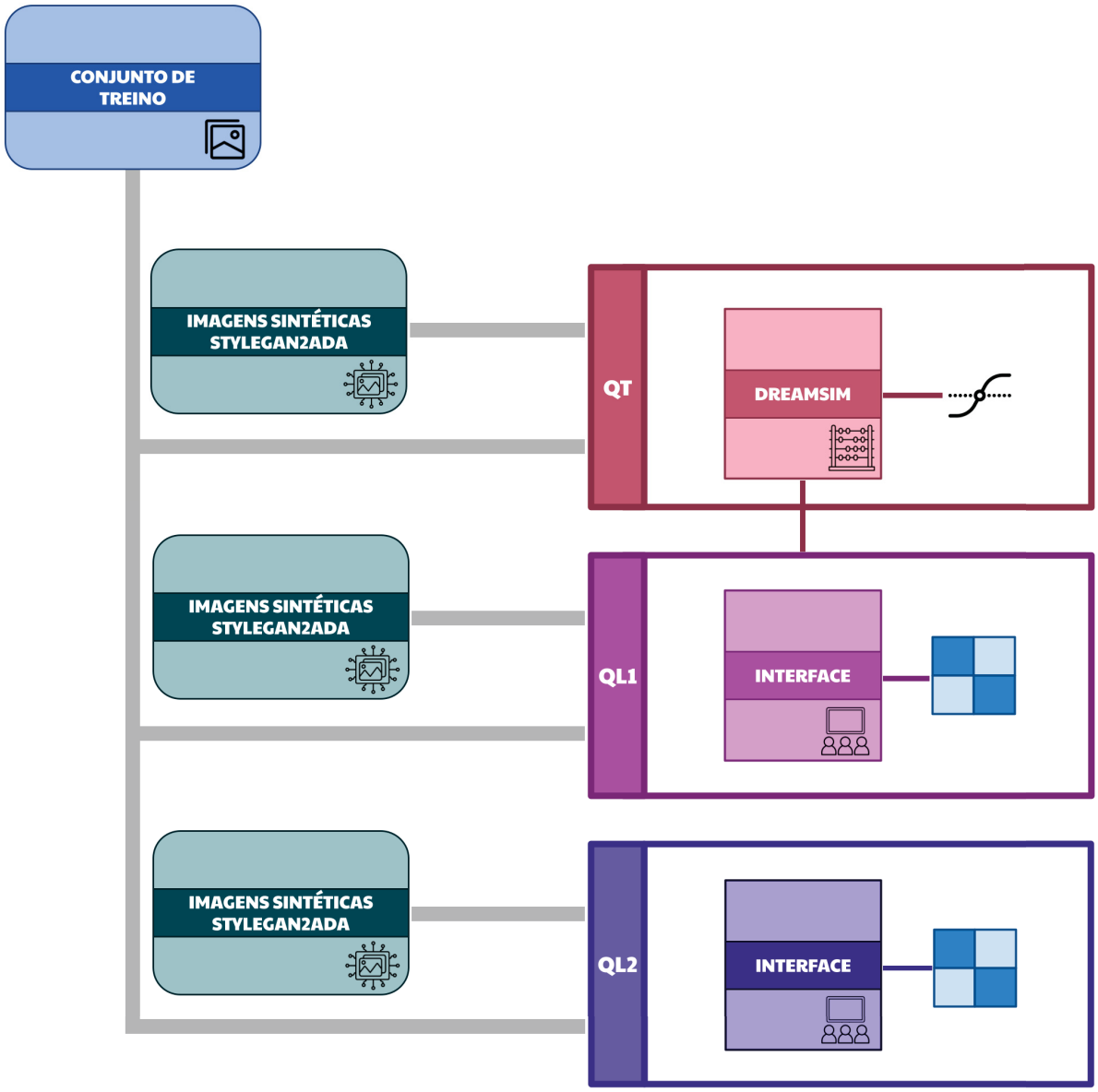


Figura 5.3: Resumo dos experimentos de análise da qualidade das imagens sintéticas produzidas pela StyleGAN2ADA. Inicialmente a análise quantitativa é realizada com a métrica DreamSIM, retornando o valor médio como resultado (QT). Por fim, a análise qualitativa pode ser realizada considerando imagens sintéticas com os menores valores de DreamSIM em relação ao conjunto de treino (QL1) ou imagens aleatórias (QL2). Os resultados desta etapa são os erros e acertos dos especialistas. Fonte: elaborado pela autora.

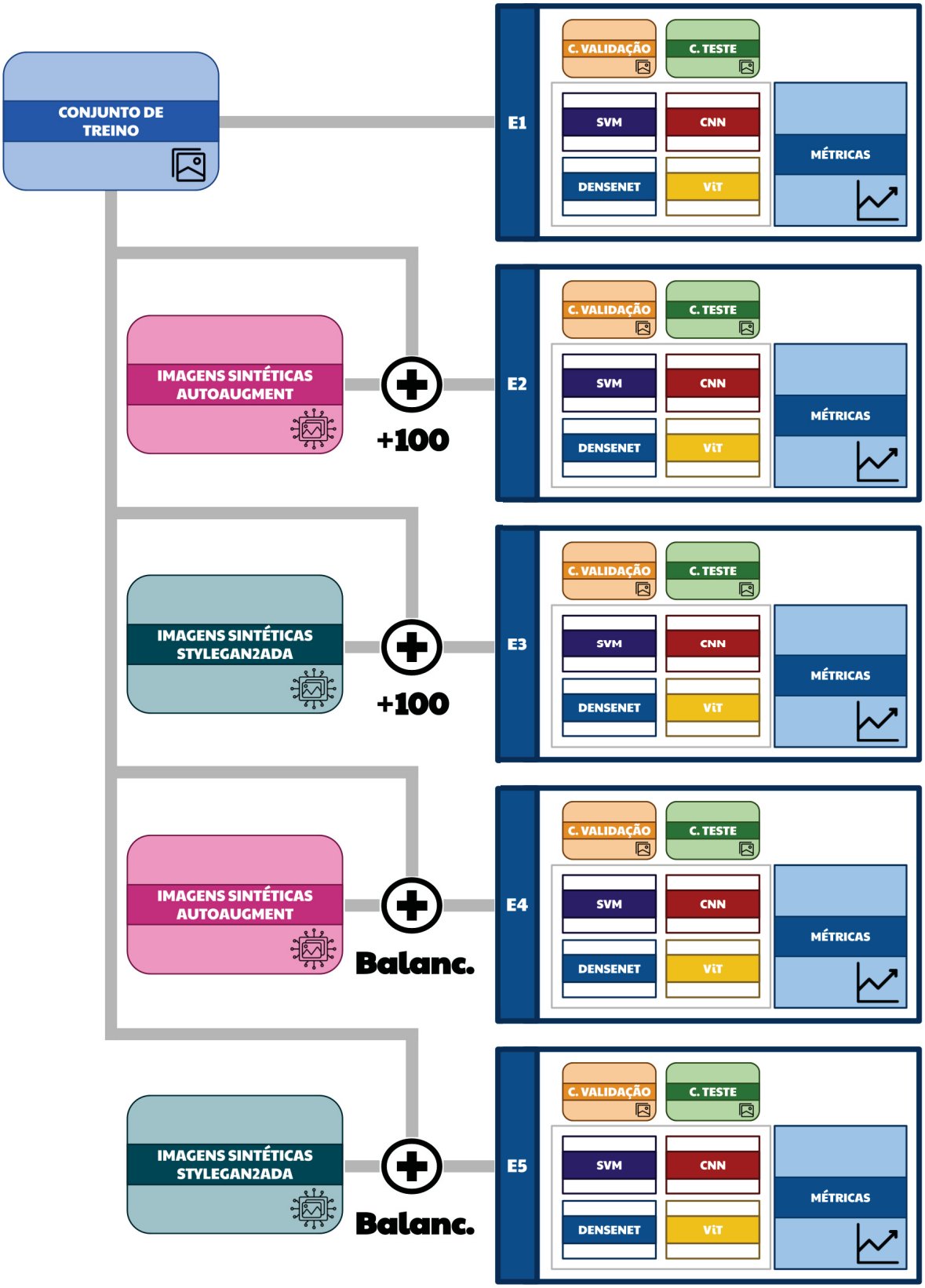


Figura 5.4: Resumo dos experimentos de classificação. Os experimentos podem ser realizados sem nenhum tipo de aumento de dados (E1); podem envolver a adição de imagens sintéticas ao conjunto de treinamento (E2 e E3) e o balanceamento das classes com imagens geradas (E4 e E5). Todos os experimentos utilizam os modelos propostos, que são a Metodologia Rogalsky (SVM), CNN, DenseNet e ViT. Os resultados são avaliados pelas métricas de precisão, *recall*, *f1-score* e acurácia. Fonte: elaborado pela autora.

Tabela 5.2: Experimentos da etapa de classificação das imagens de Receptores de Estrogênio (ER) e Receptores de Progesterona (PR). Todos os experimentos seguiram a mesma metodologia de treinamento, validação e teste descrita no experimento E1. Fonte: elaborado pela autora.

Exp.	**Descrição**
E1	Treinamento, validação e teste dos métodos de classificação propostos (Metodologia Rogalsky, CNN, DenseNet e ViT) com os dados originais. Os resultados obtidos no conjunto de teste foram avaliados com as métricas de precisão, *recall* e *f1-score*.
E2	Adição de 100 imagens sintéticas para cada classe IS ao conjunto de treinamento original, sendo geradas pelo modelo AutoAugment.
E3	Adição de 100 imagens sintéticas de cada classe IS ao conjunto de treinamento original, sendo geradas pelo modelo StyleGAN2ADA.
E4	Balanceamento das classes do conjunto de treinamento em relação à classe majoritária utilizando o modelo AutoAugment. Após o balanceamento, ocorreu a adição de mais 100 imagens sintéticas para cada classe IS.
E5	Balanceamento das classes do conjunto de treinamento em relação à classe majoritária utilizando o modelo StyleGAN2ADA. Após o balanceamento, ocorreu a adição de mais 100 imagens sintéticas para cada classe IS.

Tabela 5.3: Número de amostras de treinamento de cada classe para todos os experimentos com os conjuntos de dados ER e PR. Fonte: elaborado pela autora.

	Número de Amostras de Treinamento									
	ER					**PR**				
	0	**1+**	**2+**	**3+**	**Total**	**0**	**1+**	**2+**	**3+**	**Total**
E1	330	123	235	753	1441	319	102	137	417	975
E2	430	223	335	853	1841	419	202	237	517	1375
E3	430	223	335	853	1841	419	202	237	517	1375
E4	853	853	853	853	3412	517	517	517	517	2068
E5	853	853	853	853	3412	517	517	517	517	2068

6 RESULTADOS

Com os experimentos definidos, neste capítulo vamos apresentar o estudo e a análise dos resultados da pesquisa. Iniciamos com a avaliação do treinamento da StyleGAN2ADA (seção 6.1) e seguimos para os testes quantitativos (seção 6.2) e qualitativos (seção 6.3) das imagens produzidas pela rede. Então, observamos as amostras das imagens sintéticas geradas pelo AutoAugment e também pela StyleGAN2ADA, analisando visualmente a qualidade e as características mais relevantes para cada método (seção 6.4). Por fim, apresentamos os resultados dos cinco experimentos de classificação para ambos conjuntos de dados (seção 6.5), que são as imagens de IHQ dos Receptores de Estrogênio (ER) e Progesterona (PR).

6.1 RESULTADOS DO TREINAMENTO DA STYLEGAN2ADA

Inicialmente, treinamos as StyleGAN2ADAs seguindo as etapas apresentadas na seção 4.3, desconsiderando o uso de pesos pré-treinados. Ao final do treinamento, alcançamos KIDs abaixo de 4% para todas as classes em ambos conjuntos de dados, como mostramos na Tabela 6.1. Podemos observar que atingimos os maiores KIDs com as redes responsáveis pelas classes 1+ e 2+. Como mostramos no capítulo anterior, essas classes contêm os menores números de amostras, o que pode ter dificultado os aprendizados das redes.

Além disso, conforme os especialistas, essas duas classes são consideradas como *borderline* ou limítrofe, indicando que as células apresentam características intermediárias entre benígnas (normais) e malígnas (cancerosas). Essa particularidade pode levar à muita incerteza de classificação até mesmo para os patologistas (Shylasree et al., 2024). Dessa forma, concluímos que mesmo com poucos exemplos, as redes conseguiram aprender bem os fatores essenciais das imagens, pois as distâncias entre as distribuições de características das imagens geradas e das imagens originais não passaram de 4%.

Tabela 6.1: Resultados dos KIDs das StyleGAN2ADAs para cada uma das classes da pontuação IS para ambos conjuntos de dados (ER e PR). Fonte: elaborado pela autora.

	Classe			
Tipo de Exame	0	1+	2+	3+
ER	0,0199	0,0171	0,0235	0,0118
PR	0,0154	0,0396	0,0243	0,0153

6.2 RESULTADOS DA ANÁLISE QUANTITATIVA

Com as imagens sintéticas obtidas a partir da StyleGAN2ADA, aplicamos a sequência de passos da análise quantitativa apresentada na seção 4.4. Por questões de tempo, optamos por realizar esta análise apenas para os dados de Receptores de Estrogênio (ER), pois o cálculo da DreamSIM para cada imagem sintética em relação ao conjunto de treinamento necessita de muito tempo e recursos para ser executado.

Após medirmos todos os DreamSIMs e selecionarmos os menores valores da métrica entre cada imagem sintética e o conjunto de treino, nós calculamos a média de todos esses

valores e alcançamos um DreamSIM médio de 0,091. Dessa forma, concluímos que existe uma grande similaridade entre as imagens sintéticas e reais, revelando uma alta qualidade das imagens geradas pela StyleGAN2ADA. Nós podemos justificar essa afirmação pelo intervalo dos valores da métrica DreamSIM, que podem variar de 0 a 1, indicando que quanto mais perto de 0, mais similares as imagens são, e quanto mais próximos de 1, mais diferentes são as imagens (Fu et al., 2023).

Para exemplificar os resultados desta etapa, na primeira linha da Figura 6.1, apresentamos quatro imagens reais referentes a cada classe da pontuação IS. Nas linhas seguintes, mostramos as imagens geradas com os menores valores de DreamSIM em relação à imagem real, com cada coluna referenciando a uma classe. Ao analisarmos a figura, observamos que as imagens geradas mantiveram as características das imagens reais, mas variaram a iluminação e contraste, tanto no fundo quanto nas células. Dessa forma, assumimos que essas variações contribuem para diversificar visualmente os dados sem afetar as características mais relevantes de cada classe.

6.3 RESULTADOS DA ANÁLISE QUALITATIVA

Após a geração de imagens sintéticas com a StyleGAN2ADA, convidamos três patologistas do Hospital Erasto Gaertner para participarem dos experimentos qualitativos descritos na seção 4.5. Por questões de tempo e disponibilidade dos especialistas, optamos por selecionar apenas os dados dos Receptores de Estrogênio (ER) para esta etapa.

Iniciamos com o experimento QL1, que sorteou aleatoriamente 28 imagens reais e selecionou 72 imagens sintéticas de acordo com os menores valores da métrica DreamSIM para apresentar na interface da Categorização Rápida de Cenas (CRC). Para avaliarmos uma proporção equilibrada entre imagens sintéticas e reais, implementamos também o experimento QL2, que contou com 48 imagens reais e 48 sintéticas escolhidas aleatoriamente. Apresentamos as matrizes de confusão das respostas dos especialistas nas figuras 6.2 (QL1) e 6.3 (QL2).

Em relação ao experimento QL1, analisando somente as 28 imagens reais (segunda linha de cada matriz), notamos que os especialistas acertaram mais da metade, enfatizando o *especialista 2* que acertou todas as imagens. De acordo com o *especialista 1* e o *especialista 3*, 10 e 9 das 28 imagens reais foram marcadas como potencialmente geradas, possivelmente devido à existência de ruídos ou artefatos na imagem (ver Figura 6.2).

Considerando os resultados das imagens geradas (primeira linha de cada matriz), mais da metade das respostas de todos os especialistas indicaram que as imagem geradas eram reais. Das 72 imagens sintéticas, o *especialista 1* reportou que 46 delas eram reais, o *especialista 2* indicou 41 e o *especialista 3* informou 38. Assim, concluímos que essas baixas taxas de acerto indicam que a qualidade das imagens geradas são boas o suficiente para que os especialistas, neste experimento, não detectassem as imagens sintéticas. Dessa forma, demonstramos que as imagens sintéticas geradas pela StyleGAN2ADA se aproximam visualmente das imagens celulares reais.

Ao examinarmos os erros de todos os especialistas nas 72 imagens geradas, nós contamos 15 erros em comum. Os erros em comum cobriram todas as classes, enfatizando a classe 1+, com oito ocorrências. Acreditamos que esses resultados confirmem a similaridade entre as imagens geradas e as reais, pois os erros dos especialistas foram diferentemente distribuídos pelo conjunto de classes. Considerando os acertos em comum aos três especialistas, nós somente registramos quatro acertos das 72 imagens geradas. Os acertos em comum incluíram três ocorrências da classe 0 e uma da classe 1+. Assim, concluímos que a maioria das respostas corretas dos especialistas ocorreram em imagens diferentes, o que pode indicar o alto padrão de qualidade das imagens geradas, além dos indícios de preservação das principais características de cada classe.

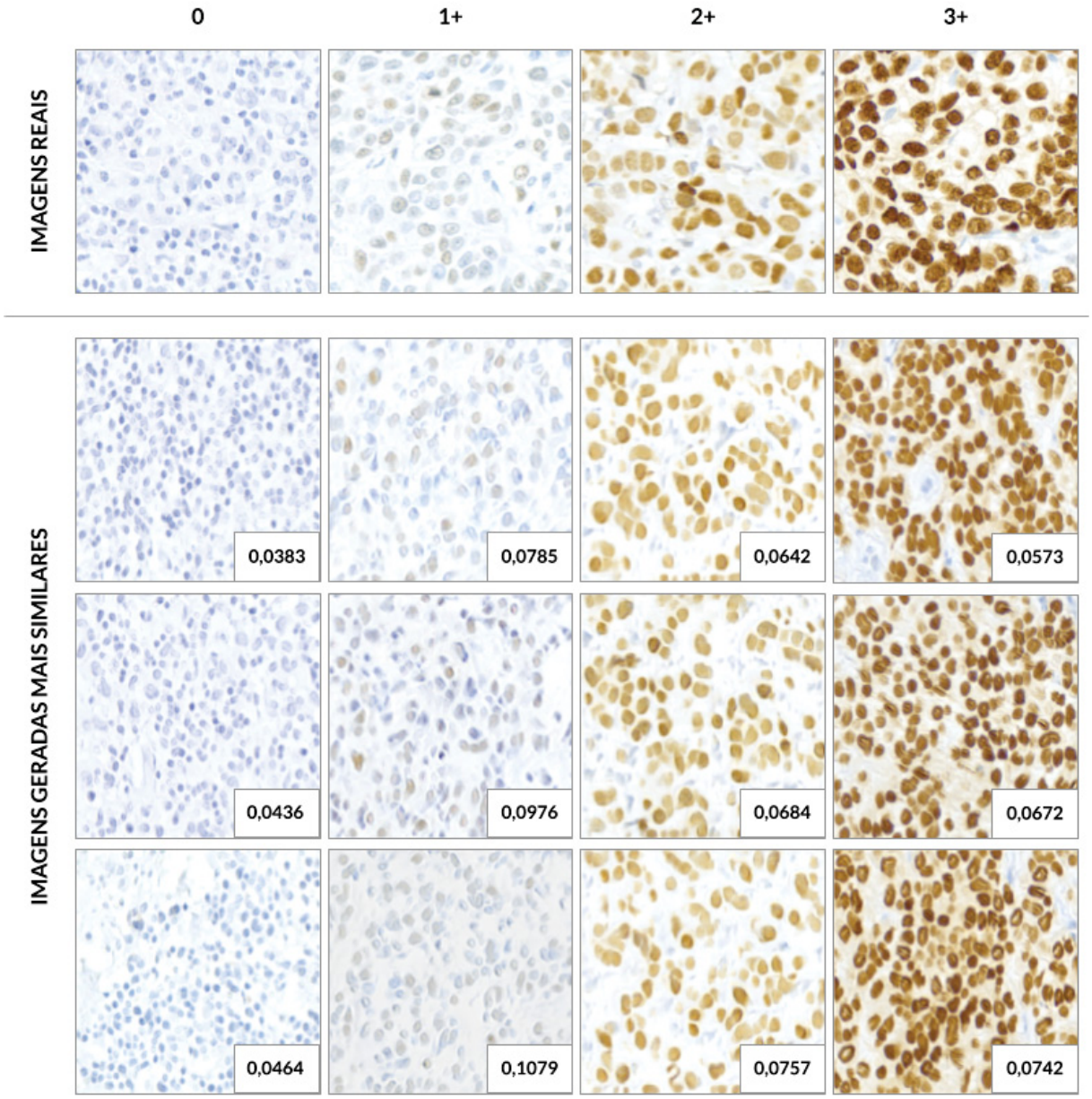


Figura 6.1: Exemplo dos resultados da análise quantitativa. Na primeira linha, são apresentados exemplos das imagens reais, com cada coluna referenciando uma classe da pontuação IS. A demais linhas mostram as imagens sintéticas mais similares à imagem real da coluna correspondente de acordo com a métrica DreamSIM. Fonte: elaborado pela autora.

Por fim, em relação ao experimento QL2 (Figura 6.3), das 48 imagens reais (segunda coluna de cada matriz), os especialistas também acertaram mais da metade, com ênfase para o *especialista 3*, que acertou 38 imagens. Observando os resultados nas imagens geradas (primeira linha de cada matriz), novamente mais da metade das respostas dos especialistas indicaram que as imagens sintéticas eram reais. Das 48 imagens geradas, o *especialista 1* reportou 30 delas como reais, o *especialista 2* indicou 34 e o *especialista 3* informou 25. Dessa forma, mostramos mais uma vez a semelhança visual das imagens geradas pela StyleGAN2ADA com a realidade.

Sobre os erros em comum aos três especialistas nas imagens geradas, contabilizamos oito erros em todas as classes, mostrando novamente que os erros foram distribuídos pelo conjunto de imagens. Os especialistas não apresentaram acertos em comum neste experimento,

Figura 6.2: Matrizes de confusão do experimento QL1, representando os erros e acertos dos três especialistas na interface de análise qualitativa. Fonte: elaborado pela autora.

nos levando à conclusão que não existem características específicas pertencentes às imagens geradas que levaram os especialistas a acreditarem que essas imagens são sintéticas. Assim, podemos concluir que as imagens sintéticas produzidas pela StyleGAN2ADA apresentam um alto padrão de forma geral, independente se a escolha das imagens foi feita de forma aleatória (QL2) ou se foram selecionadas as imagens sintéticas mais similares ao conjunto de treino (QL1).

Figura 6.3: Matrizes de confusão do experimento QL2, representando os erros e acertos dos três especialistas na interface de análise qualitativa. Fonte: elaborado pela autora.

6.4 ANÁLISE VISUAL DAS IMAGENS SINTÉTICAS

Nesta seção, focamos em mostrar as diferenças entre as imagens sintéticas obtidas com o AutoAugment e as geradas pela StyleGAN2ADA. Devido à questões de tempo e disponibilidade dos especialistas, optamos por esta análise visual simplificada para apresentar alguns exemplos de imagens geradas com ambos modelos e conjuntos de dados. Também nos concentramos em analisar essas imagens de uma perspectiva de classificação, visando supor quais características visuais poderiam impactar no desempenho dos classificadores.

Iniciando com uma análise visual das imagens do conjunto de dados ER, apresentamos amostras reais de cada classe acompanhadas pelas imagens sintéticas geradas pelo AutoAugment e pela StyleGAN2ADA na Figura 6.4. Observando as imagens do AutoAugment em relação às imagens reais, podemos notar que a técnica aumentou o contraste e modificou a coloração das células negativas na amostra da classe 0. Essas alterações escureceram as células saudáveis e

mudaram a tonalidade de azul para marrom. Acreditamos que modificações como essa possam confundir os modelos de classificação, pois essas características são evidentes em exemplos de outras classes, como a 2+ e 3+.

Em relação às classes 1+ e 2+, podemos ver que o AutoAugment aplicou técnicas de rotação e translação nas imagens originais. No contexto de imagens celulares, métodos como esses podem contribuir para trazer uma maior variação das imagens, mas eles também criam novas regiões sem informações. Essas regiões em preto podem trazer algum viés para o modelo de classificação, além de comporem imagens que se distanciam da realidade, dificultando o processo de interpretabilidade da metodologia.

Por fim, a amostra aumentada da classe 3+ amplificou o brilho da imagem original, deixando as células cancerosas com uma tonalidade mais clara. Na maioria dos cenários, alterações de brilho podem contribuir para diversificar o conjunto de dados, mas neste caso, essa modificação pode levar a confusões entre as classes 3+ e 2+, já que a coloração das células é de extrema importância para a classificação dos níveis de câncer (Rogalsky et al., 2021).

Sobre as imagens da StyleGAN2ADA, podemos notar que ocorreram transformações de brilho e contraste, mas essas alterações não interferiram significativamente na coloração das células, somente no fundo. Além disso, a rede trouxe variações de formas, posicionamentos e organizações das células, sempre preservando as características visuais principais de cada classe.

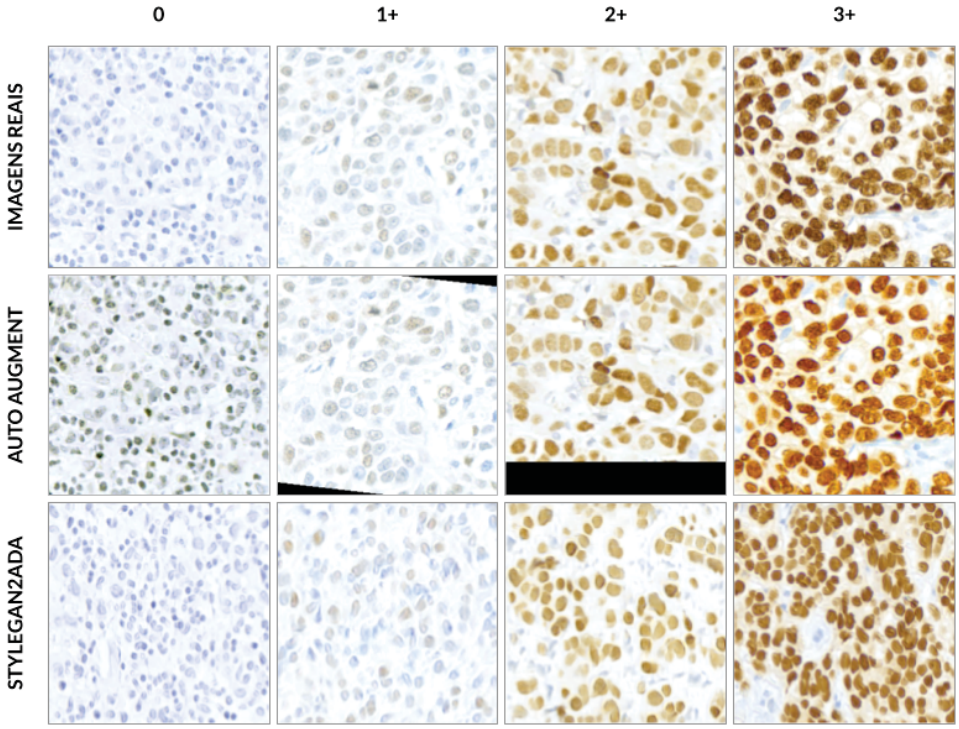


Figura 6.4: Exemplos das imagens reais, das imagens sintéticas obtidas pelo AutoAugment e das imagens geradas pela StyleGAN2ADA no conjunto de dados de Receptores de Estrogênio (ER). Fonte: elaborado pela autora.

Analisando as imagens do conjunto de dados PR na Figura 6.5, observamos que o AutoAugment também alterou o contraste na amostra da classe 0, mas não modificou a coloração das células saudáveis. Acreditamos que essa transformação possa contribuir para o modelo de classificação ampliar o seu conhecimento sobre as células negativas, aumentando a robustez em relação à diferenças de contrastes causadas na etapa de aquisição das imagens. A aplicação de borramento que ocorreu na classe 3+ também poderia ajudar a melhorar a robustez do classificador em relação aos ruídos de entrada desta etapa.

Podemos notar que na classe 1+ o modelo aumentou excessivamente o contraste da imagem, gerando uma amostra que difere de todas as classes pertencentes ao conjunto de dados. Esse exemplo nos mostra que, mesmo utilizando um modelo automático, equilibrar os valores de magnitude das transformações para o problema atual é um grande desafio. Por fim, para gerar a amostra da classe 2+, o modelo utilizou a rotação, que apresenta a vantagem de alterar o posicionamento das células na imagem, mas a desvantagem de criar novas regiões sem informações relevantes.

Em relação as imagens da StyleGAN2ADA, podemos observar que elas possuem os mesmos pontos positivos das imagens sintéticas do conjunto de dados ER, mas com algumas imperfeições. A amostra da classe 1+ apresenta uma resolução menor em relação às outras imagens. Acreditamos que essas imperfeições possam ter aparecido devido a quantidade reduzida de imagens de treino, com somente 171 exemplos. Contudo, as imagens geradas pela rede trouxeram diversidade para o conjunto de dados, mantendo as cores principais de cada classe e variando os formatos, tamanhos e agrupamentos das células.

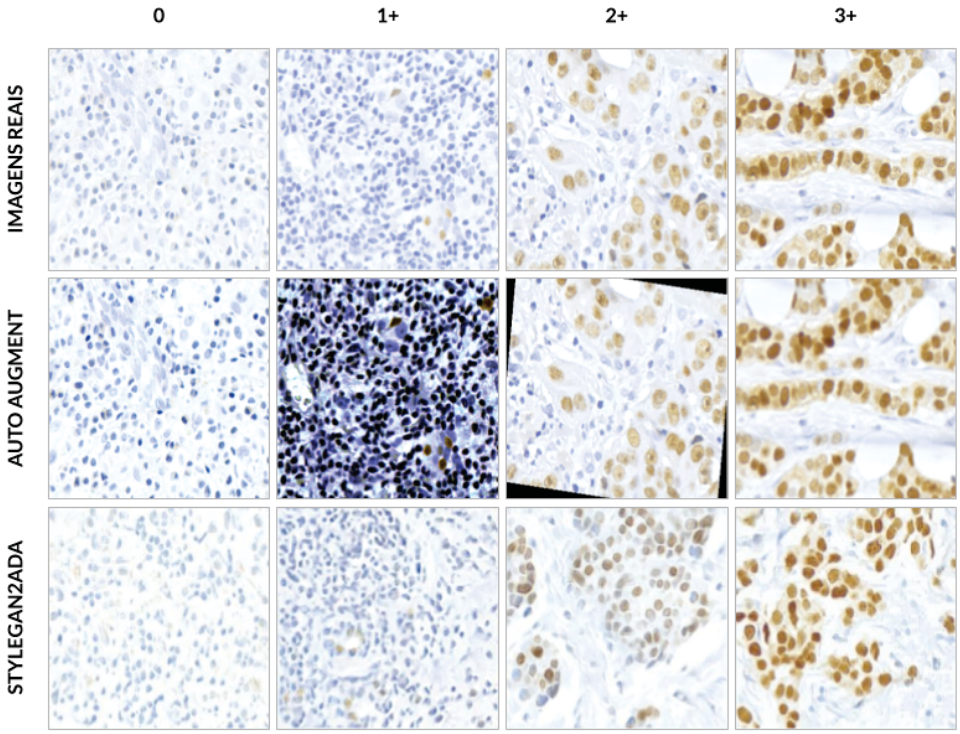


Figura 6.5: Exemplos das imagens reais, das imagens sintéticas obtidas pelo AutoAugment e das imagens geradas pela StyleGAN2ADA no conjunto de dados de Receptores de Progesterona (PR). Fonte: elaborado pela autora.

6.5 RESULTADOS DE CLASSIFICAÇÃO

Nesta seção, vamos analisar os resultados dos experimentos de classificação (E1, E2, E3, E4 e E5) aplicados no conjunto de teste, com as descrições presentes no capítulo 5. Dividimos esta seção em duas subseções, iniciando com os resultados nas imagens dos Receptores de Estrogênio (ER) e seguindo para as imagens dos Receptores de Progesterona (PR). Apresentamos as distribuições das classes das amostras de teste em ambos conjuntos de dados na Tabela 6.2.

Tabela 6.2: Número de amostras do conjunto de teste para cada classe das imagens dos Receptores de Estrogênio (ER) e Progesterona (PR). Fonte: elaborado pela autora.

	Número de Amostras de Teste				
	0	**1+**	**2+**	**3+**	**Total**
ER	82	30	59	189	360
PR	103	34	45	143	325

6.5.1 Conjunto de Dados ER

Os gráficos de barras resumem os resultados gerais dos classificadores em relação aos *f1-scores* obtidos no conjunto de dados ER. A Figura 6.6 apresenta os resultados médios de cada um dos métodos de classificação em relação a cada um dos experimentos (E1, E2, E3, E4 e E5). As mesmas informações são expostas na Figura 6.7, com a diferença de que os *f1-scores* são referentes às classes minoritárias (1+ e 2+), que apresentaram as maiores disparidades nos valores das métricas entre experimentos. Optamos por trazer esses gráficos no início da seção para exibir uma visão geral dos resultados, mas eles serão discutidos ao longo do texto juntamente com as tabelas de desempenho de cada método de classificação. Alteramos o nome da Metodologia Rogalsky para SVM por questões de simplicidade dos gráficos.

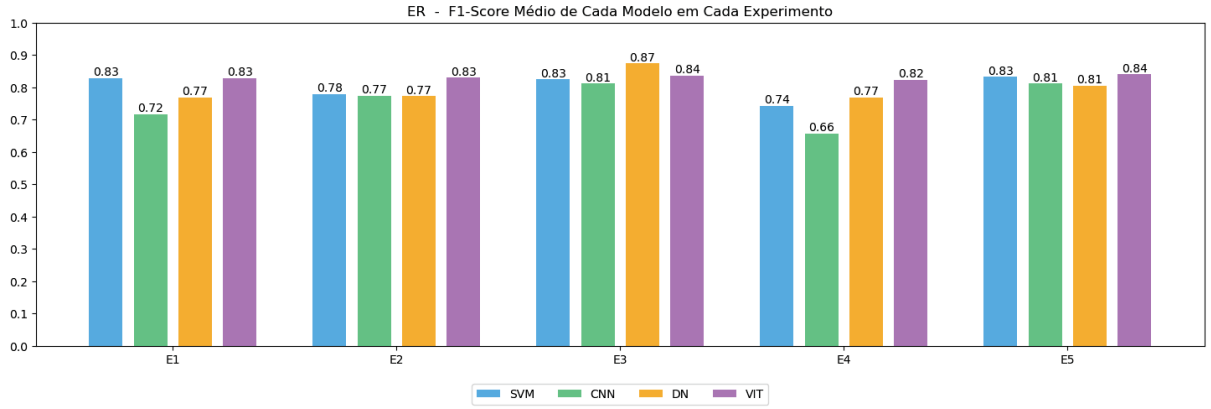


Figura 6.6: Gráfico de barras dos *f1-scores* de cada método de classificação em cada experimento, considerando a média dos resultados de todas as classes. A primeira barra de cada conjunto de quatro barras representa a Metodologia Rogalsky (SVM), a segunda a CNN, a terceira a DenseNet (DN) e a quarta o ViT. Os resultados são referentes às médias das cinco pastas do conjunto de teste de ER. Fonte: elaborado pela autora.

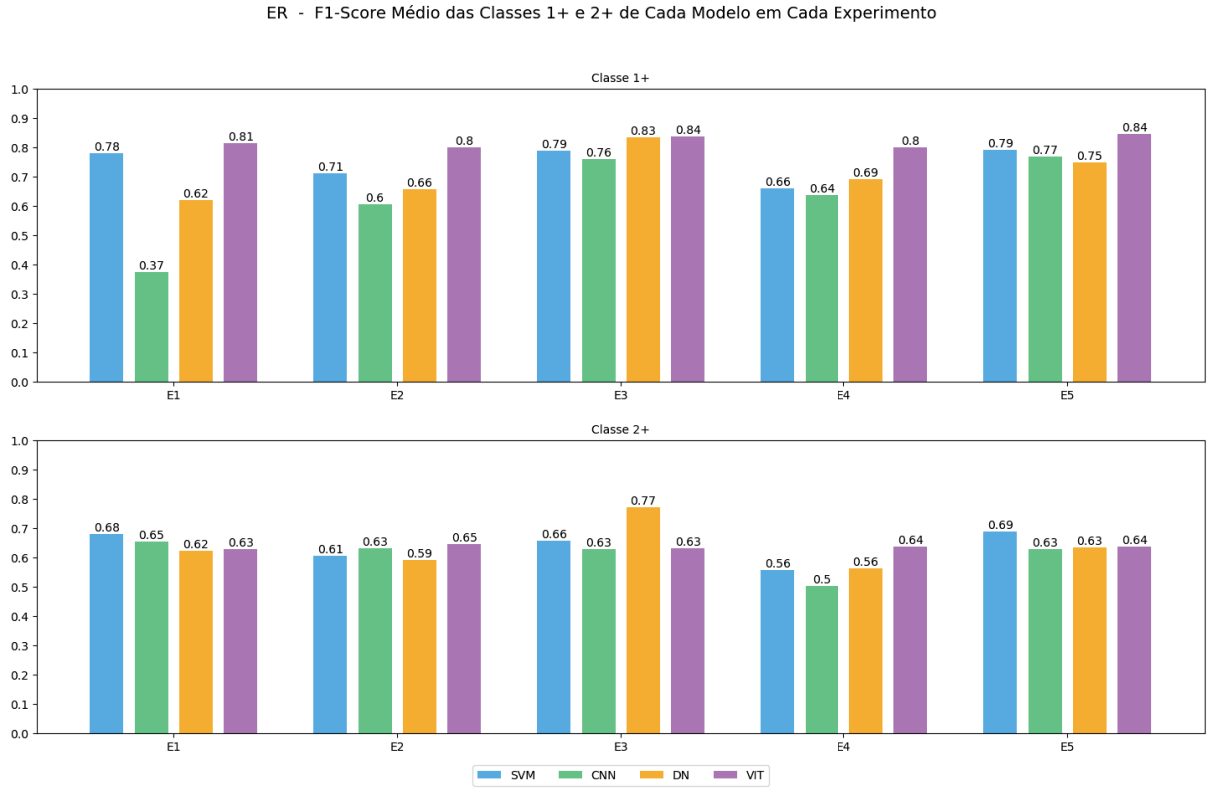


Figura 6.7: Gráfico de barras dos *f1-scores* de cada método de classificação em cada experimento, considerando as classes minoritárias (1+ e 2+). A primeira barra de cada conjunto de quatro barras representa a Metodologia Rogalsky (SVM), a segunda a CNN, a terceira a DenseNet (DN) e a quarta o ViT. Os resultados são referentes às médias das cinco pastas do conjunto de teste de ER. Fonte: elaborado pela autora.

Sobre os resultados da **Metodologia Rogalsky**, podemos observar na Figura 6.6 que os experimentos que aplicaram o AutoAugment (E2 e E4) pioraram os *f1-scores* médios em 5 e 9 pontos percentuais em comparação ao experimento sem aumento de dados (E1). Analisando a Figura 6.7, vemos que as classes 1+ e 2+ refletem esse comportamento. Exploramos os resultados com maior profundidade na Tabela 6.3, que contém as métricas de precisão, *recall* e *f1-score*.

Com esta tabela podemos perceber que os experimentos E2 e E4 prejudicaram a precisão da solução em até 18 pontos percentuais na classe 1+ em relação ao experimento E1. Já a classe 2+ apresentou uma queda maior no *recall*, reduzindo em até 9 pontos percentuais comparando com E1. Assim, podemos assumir que as imagens sintéticas da classes 1+ e 2+ produzidas pelo AutoAugment prejudicaram o aprendizado do modelo nos casos em que as amostras dessas classes eram positivas, possivelmente causando confusões entre elas.

Ademais, o experimento E3 piorou minimamamente (0,31 pontos percentuais) o *f1-score* geral médio e o experimento E5 alcançou uma melhora de 0,38 pontos percentuais em relação à E1. Como a Metodologia Rogalsky implementa o SVM, que é um modelo que não apresenta problemas com dados desbalanceados (Cortes e Vapnik, 1995), a adição de imagens com a StyleGAN2ADA (E3 e E5) não contribuiu para uma melhora significativa no desempenho do método de classificação, mas também não piorou os resultados de forma considerável. Dessa forma, mostramos que, neste método, as imagens sintéticas da rede superam o AutoAugment e não prejudicam o processo de classificação.

Tabela 6.3: Resultados da Metodologia Rogalsky nos experimentos E1, E2, E3, E4 e E5. A região de cima da tabela apresenta as precisões, *recalls* e *f1-scores* específicos por classe e a debaixo as médias, considerando todas as classes. Os valores destacados em azul mostram melhorias e os valores em vermelho indicam pioras em relação aos resultados do experimento E1. Os resultados são referentes às médias das cinco pastas do conjunto de teste dos Receptores de Estrogênio (ER). Fonte: elaborado pela autora.

	ER - Métricas Metodologia Rogalsky														
	Precisão (%)					Recall (%)					F1-Score (%)				
Classe	E1	E2	E3	E4	E5	E1	E2	E3	E4	E5	E1	E2	E3	E4	E5
0	89,67	85,20	91,29	83,68	92,61	97,21	95,51	96,04	93,72	93,70	93,26	90,05	93,59	88,39	93,12
1+	**82,02**	**68,28**	75,64	**64,00**	75,75	74,59	74,72	82,81	71,12	83,29	77,94	71,07	78,95	66,05	79,24
2+	65,08	58,05	66,32	51,81	65,63	**72,80**	**64,86**	66,77	**63,71**	73,18	67,93	60,60	65,73	55,82	68,87
3+	95,11	95,87	94,16	96,11	94,44	89,78	84,99	89,89	80,21	89,32	92,34	90,02	91,93	87,06	91,78
Média	82,97	76,85	81,85	73,90	82,11	83,60	80,02	83,88	77,19	84,87	**82,86**	**77,94**	**82,55**	**74,33**	**83,25**
DP	04,14	05,44	03,65	06,67	03,39	05,17	07,10	05,43	10,53	04,85	02,34	05,12	02,92	05,91	03,06

Notamos que houve uma melhora nos *f1-scores* médios dos experimentos E2, E3 e E5 com o modelo **CNN** em relação à E1 (Figura 6.6). Analisando os resultados das classes minoritárias na Figura 6.7, somente a classe 2+ seguiu esse padrão de comportamento, pois na categoria 1+ todos os experimentos melhoraram os resultados ao compararmos com E1. Investigamos essas questões na Tabela 6.4, que contém as métricas gerais e por classe dos experimentos realizados com a CNN.

Ao observarmos os resultados gerais de *f1-score*, notamos que os experimentos com a StyleGAN2ADA apresentaram os melhores valores, alcançando melhoras de até 9 pontos percentuais em relação à E1. Podemos justificar as melhoras pelas imagens sintéticas trazerem uma maior variabilidade ao conjunto de treinamento, ajudando na generalização do modelo (ao Huang et al., 2022). Neste caso, podemos afirmar que a adição de imagens sintéticas de E3 e o balanceamento de classes de E5 tiveram efeitos similares nos resultados finais (ambos com 81% de *f1-score*). Além disso, podemos perceber que estes experimentos contribuíram significativamente para a classe 1+, aumentando em até 39 pontos percentuais o *f1-score* desta categoria comparando com E1 (de 37% para 76%).

Nós conseguimos melhorar o *f1-score* geral em 5,6 pontos percentuais com a adição de 100 imagens sintéticas ao conjunto de treino com o AutoAugment (E2), mas alcançamos uma piora de 6 pontos percentuais ao balancearmos as classes com esta técnica (E4). Podemos justificar a melhora pelo fato da CNN ser um modelo mais simples que não possui pesos pré-treinados, podendo se beneficiar de uma maior quantidade de dados para estabilizar seu aprendizado (ao Huang et al., 2022). No caso de E4, as imagens sintéticas reduziram o desempenho do modelo em várias classes, incluindo as classes majoritárias 0 e 3+, com pioras de 15 e 19 pontos percentuais de *f1-score*, respectivamente. Acreditamos que a adição excessiva de imagens geradas sem uma supervisão da sua qualidade possa prejudicar o treinamento do modelo, pois o seu aprendizado se baseia majoritariamente nestas imagens sintéticas, trazendo muitas confusões entre as classes.

Tabela 6.4: Resultados da CNN nos experimentos E1, E2, E3, E4 e E5. A região de cima da tabela apresenta as precisões, *recalls* e *f1-scores* específicos por classe e a debaixo as médias, considerando todas as classes. Os valores destacados em azul mostram melhorias e os valores em vermelho indicam pioras em relação aos resultados do experimento E1. Os resultados são referentes às médias das cinco pastas do conjunto de teste dos Receptores de Estrogênio (ER). Fonte: elaborado pela autora.

	ER - Métricas CNN														
	Precisão (%)					Recall (%)					F1-Score (%)				
Classe	E1	E2	E3	E4	E5	E1	E2	E3	E4	E5	E1	E2	E3	E4	E5
0	85,81	89,66	91,69	74,59	93,06	97,50	97,03	95,62	77,54	94,25	**91,26**	93,19	93,61	**76,03**	93,64
1+	90,81	89,73	79,12	65,65	76,16	26,65	47,57	73,41	61,97	77,91	**37,45**	60,49	**76,12**	63,69	**76,74**
2+	63,25	64,50	64,33	49,54	61,63	68,48	62,12	62,10	51,86	64,76	**65,41**	63,25	62,84	**50,21**	62,85
3+	93,07	93,28	92,15	75,06	92,14	92,89	91,81	92,37	71,35	90,72	**92,95**	92,54	92,23	**73,09**	91,41
Média	83,24	84,29	81,82	66,21	80,75	71,38	74,63	80,87	65,68	81,91	**71,77**	**77,37**	**81,20**	**65,76**	**81,16**
DP	04,80	04,39	04,53	37,42	04,76	07,34	05,17	03,91	37,00	04,98	06,88	04,63	03,03	36,92	04,04

Considerando os *f1-scores* obtidos com a **DenseNet** na Figura 6.6, percebemos que todos os experimentos se igualaram ou superaram E1. Notamos também que esse comportamento se manteve na classe 1+, mas os experimentos E2 e E4 apresentaram quedas de desempenho na categoria 2+, como mostramos na Figura 6.7. Essas quedas podem ter acontecido devido aos realces indevidos e mudanças de coloração realizados pela técnica AutoAugment, como mencionamos na análise visual das imagens sintéticas.

Analisando a Tabela 6.5, reparamos que a DenseNet alcançou os melhores resultados em relação aos modelos anteriores, com 87,46% de *f1-score* geral no experimento E3. Além disso, notamos que os maiores ganhos do experimento E3 vieram na classe 1+, com um aumento de 21,5 pontos percentuais em relação à E1. Acreditamos que a maior profundidade das camadas da rede, as várias inovações trazidas pela arquitetura e o uso de pesos pré-treinados possam ter contribuído para um melhor desempenho.

No entanto, quando examinamos os resultados do experimento de balanceamento com a StyleGAN2ADA (E5), observamos que o valor do *f1-score* geral foi menor do que os resultados da CNN e da Metodologia Rogalsky. Dessa forma, enfatizamos que a otimização das quantidades de imagens sintéticas adicionadas ao conjunto de treinamento pode ser de extrema importância para alcançar melhores resultados no futuro, porque não existem garantias de que um modelo mais complexo sempre alcançará o melhor desempenho.

Por fim, atingimos melhoras de 0,44; 10,58; 0,05 e 3,72 pontos percentuais em *f1-scores* gerais ao comparamos os experimentos E2, E3, E4 e E5 com E1. Analisando esses resultados, concluímos que os experimentos com a StyleGAN2ADA foram os únicos capazes de trazer melhorias significativas em relação ao experimento sem aumento de dados (E1). Novamente, demonstramos a viabilidade da utilização de imagens sintéticas produzidas pela StyleGAN2ADA para melhorar a assertividade dos classificadores neste conjunto de dados.

Tabela 6.5: Resultados da DenseNet nos experimentos E1, E2, E3, E4 e E5. A região de cima da tabela apresenta as precisões, *recalls* e *f1-scores* específicos por classe e a debaixo as médias, considerando todas as classes. Os valores destacados em azul mostram melhorias e os valores em vermelho indicam pioras em relação aos resultados do experimento E1. Os resultados são referentes às médias das cinco pastas do conjunto de teste dos Receptores de Estrogênio (ER). Fonte: elaborado pela autora.

	ER - Métricas DenseNet														
	Precisão (%)					Recall (%)					F1-Score (%)				
Classe	E1	E2	E3	E4	E5	E1	E2	E3	E4	E5	E1	E2	E3	E4	E5
0	85,71	91,55	93,90	91,88	94,71	97,85	94,81	96,22	95,90	93,61	91,23	93,01	95,02	93,83	94,05
1+	84,72	82,26	82,46	71,29	75,23	50,38	56,63	84,41	68,04	75,60	**61,86**	65,80	**83,36**	69,03	74,92
2+	65,75	58,96	75,15	50,99	56,08	62,18	60,55	79,96	65,17	73,48	62,29	59,04	77,07	56,41	63,43
3+	91,90	90,49	96,16	93,88	94,44	92,63	92,58	92,76	83,96	86,08	92,12	91,44	94,40	88,43	90,01
Média	82,02	80,81	86,92	77,01	80,12	75,76	76,14	88,34	78,27	82,19	**76,88**	**77,32**	**87,46**	**76,93**	**80,60**
DP	05,82	06,46	04,05	05,59	04,52	08,99	08,59	05,68	06,52	06,82	05,47	04,95	03,98	03,88	04,08

Por fim, conquistamos os resultados mais consistentes com o **ViT**, que apresentou *f1-scores* acima de 82% em todos os experimentos (Figura 6.6). Percebemos que essa consistência de resultados similares entre experimentos também foi mantida nos cálculos das métricas para as classes 1+ e 2+, com alterações de 1 ou 2 pontos percentuais de um experimento para outro (Figura 6.7).

Observando a Tabela 6.6, podemos notar que o experimento E5 obteve os maiores ganhos em relação ao experimento sem aumento de dados (E1), melhorando em 1,18 pontos percentuais o *f1-score* geral. Não identificamos pioras ou melhoras significativas nos resultados dos experimentos com a técnica AutoAugment (E2 e E4). Assim, concluímos que o aumento do número de exemplos, o balanceamento das classes, a variabilidade e qualidade das imagens sintéticas da StyleGAN2ADA no experimento E5 podem ter contribuído para o aprendizado do modelo. Essas contribuições complementaram o conjunto de treinamento e podem ter ajudado o ViT a aprender as características locais das células no seu processo de divisão das imagens em *patches*, além de estabilizarem o treinamento da MLP presente em sua camada final.

Tabela 6.6: Resultados do ViT nos experimentos E1, E2, E3, E4 e E5. A região de cima da tabela apresenta as precisões, *recalls* e *f1-scores* específicos por classe e a debaixo as médias, considerando todas as classes. Os valores destacados em azul mostram melhorias e os valores em vermelho indicam piora em relação aos resultados do experimento E1. Os resultados são referentes às médias das cinco pastas do conjunto de teste dos Receptores de Estrogênio (ER). Fonte: elaborado pela autora.

	ER - Métricas ViT														
	Precisão (%)					Recall (%)					F1-Score (%)				
Classe	E1	E2	E3	E4	E5	E1	E2	E3	E4	E5	E1	E2	E3	E4	E5
0	95,20	94,14	95,23	94,76	56,08	94,38	95,78	97,05	94,70	96,25	94,77	94,94	96,12	94,69	95,96
1+	81,36	81,40	86,66	78,78	94,44	82,43	79,39	81,56	81,35	87,67	81,46	80,05	83,77	79,87	84,48
2+	65,46	63,77	61,84	61,84	80,12	61,85	66,67	65,39	67,00	62,64	62,98	64,55	63,04	63,68	63,74
3+	91,77	93,20	92,44	93,10	92,13	92,82	91,62	91,19	90,14	91,88	92,24	92,36	91,78	91,54	91,99
Média	83,45	83,13	84,04	82,12	83,92	82,87	83,36	83,80	83,30	84,61	**82,86**	**82,97**	**83,68**	**82,44**	**84,04**
DP	04,15	04,40	03,84	04,49	04,66	06,70	05,73	05,12	05,28	05,44	03,43	02,89	02,65	03,03	03,71

Na Tabela 6.7, agrupamos os *f1-scores* médios obtidos no conjunto de teste dos Receptores de Estrogênio (ER), considerando cada experimento com todos os métodos de classificação. Podemos notar que o experimento E4 apresentou os piores resultados, atingindo um desvio padrão de 36,6% com a CNN. Acreditamos que esses valores mostrem que técnicas simples de aumento de dados podem não ser adequadas para todos os tipos de imagens médicas, pois rotações, translações e mudanças de contraste podem interferir em características essenciais das patologias, prejudicando o aprendizado do classificador (Garcea et al., 2023).

Acerca dos melhores resultados, vemos que alcançamos maiores benefícios no experimento E3 com os dois modelos baseados em redes neurais convolucionais, enquanto atingimos um melhor desempenho com o experimento E5 para a Metodologia Rogalsky e o ViT. Diferentemente do que pensávamos, o balanceamento de classes não apresentou tanta melhoria quanto a adição de 100 imagens sintéticas ao conjunto de treino nos modelos CNN e DenseNet. Além disso, exceto na DenseNet, os resultados de E3 se aproximaram bastante de E5. Assim, reforçamos que otimizar o número de imagens geradas adicionadas ao conjunto de treino é uma etapa essencial para atingir melhores resultados com esta metodologia no futuro. Ademais, enfatizamos que os experimentos com a StyleGAN2ADA superaram os resultados de todos os outros experimentos, mostrando novamente a viabilidade desta técnica neste problema.

Tabela 6.7: Resultados dos métodos de classificação nos experimentos E1, E2, E3, E4 e E5, considerando a média dos *f1-scores* de todas as classes. Os valores destacados em azul informam os melhores resultados de cada método e os valores em vermelho indicam os piores desempenhos. Os resultados são referentes às médias das cinco pastas do conjunto de teste dos Receptores de Estrogênio (ER). Fonte: elaborado pela autora.

	ER - F1-Score Médio (%)				
Modelo	E1	E2	E3	E4	E5
Met. Rogalsky	82,86 ± 2,34	77,94 ± 5,12	82,55 ± 2,92	**74,33** ± 5,91	**83,25** ± 3,06
CNN	71,77 ± 6,88	77,37 ± 4,63	**81,20** ± 3,03	**65,76** ± 36,9	81,16 ± 4,04
DenseNet	**76,88** ± 5,47	77,32 ± 4,95	**87,46** ± 3,98	76,93 ± 3,88	80,60 ± 4,08
ViT	82,86 ± 3,43	82,97 ± 2,89	83,68 ± 2,65	**82,44** ± 3,03	**84,04** ± 3,71

Por fim, ainda no conjunto de dados de Receptores de Estrogênio (ER), apresentamos as acurácias médias de teste em cada experimento com todos os modelos de classificação na Tabela 6.8. Optamos por apresentar as acurácias para permitir comparações entre a nossa pesquisa e os trabalhos relacionados, mas destacamos que a acurácia não é uma métrica adequada para avaliar dados desbalanceados. Como a acurácia considera as quantidades de previsões corretas por classe proporcionalmente ao número de exemplos, ela desconsidera, na maioria das vezes, o desempenho da classe minoritária (Thölke et al., 2023). Podemos ver na tabela que mesmo o experimento E4 apresentando os menores valores da métrica, os desempenhos de todos os modelos e todos os experimentos são bem próximos. Dessa forma, enfatizamos que a escolha das métricas de avaliação da classificação são de extrema importância para resultados corretos e confiáveis em conjuntos de dados desbalanceados.

Tabela 6.8: Resultados dos métodos de classificação nos experimentos E1, E2, E3, E4 e E5, considerando a média das acurácias de todas as classes. Os valores destacados em azul informam os melhores resultados de cada método e os valores em vermelho indicam os piores desempenhos. Os resultados são referentes às médias das cinco pastas do conjunto de teste dos Receptores de Estrogênio (ER). Fonte: elaborado pela autora.

	ER - Acurácia Média (%)				
Modelo	E1	E2	E3	E4	E5
Met. Rogalsky	**93,79** ± 0,81	91,82 ± 1,96	93,65 ± 1,12	**90,11** ± 3,24	93,71 ± 1,05
CNN	92,54 ± 1,43	93,20 ± 1,36	**93,46** ± 1,41	**89,82** ± 8,57	93,15 ± 1,50
DenseNet	92,74 ± 1,40	92,52 ± 1,83	**95,45** ± 1,37	**91,35** ± 1,90	92,57 ± 1,80
ViT	93,90 ± 0,93	93,90 ± 0,74	93,96 ± 0,88	**93,51** ± 1,26	**94,08** ± 1,45

6.5.2 Conjunto de Dados PR

Com relação às imagens dos Receptores de Progesterona (PR), seguimos a mesma organização para realizar a análise dos resultados de classificação no conjunto de teste. Abaixo mostramos dois gráficos de barras para sumarizar os *f1-scores* gerais obtidos no conjunto de dados PR, considerando todos os classificadores. No primeiro (Figura 6.8) apresentamos os resultados médios de cada um dos métodos de classificação em relação a cada um dos experimentos (E1, E2, E3, E4 e E5). No segundo gráfico (Figura 6.9) expomos as mesmas informações com a diferença de que os *f1-scores* são referentes às classes minoritárias (1+ e 2+). Novamente, decidimos apresentar esses gráficos no ínicio da seção para exibir uma visão geral dos resultados, mas eles serão discutidos ao longo do texto em conjunto com as tabelas de desempenho de cada método de classificação.

Analisando as barras de resultado da **Metodologia Rogalsky** na Figura 6.8 (SVM), percebemos que melhoramos os *f1-scores* nos experimentos com a StyleGAN2ADA (E3 e E5) em relação ao experimento sem aumento de dados (E1). Já os experimentos com o AutoAugment (E2 e E4) pioraram o desempenho da classificação ao compararmos com E1. Também notamos esse comportamento nos gráficos das classes minoritárias (Figura 6.9), sendo que alcançamos um empate de melhor desempenho entre E3 e E5 na classe 1+ e constatamos que o experimento E3 apresentou um resultado superior na classe 2+.

Sobre as quedas de desempenho nos *f1-scores*, observamos uma perda de 13 pontos percentuais em E2 (classe 1+) e de 12 pontos percentuais (classe 2+) em E4 ao compararmos com E1. Esses valores são confirmados na Tabela 6.9, onde notamos reduções nos *f1-scores* em todas as classes nos experimentos E2 e E4 (representadas em vermelho na região de *f1-score* das classes), com ênfase para a perda de 13,52 pontos percentuais na classe 1+ em E2. Dessa forma, concluímos que algumas transformações simples de aumento de dados podem ser prejudiciais ao treinamento do classificador caso a técnica não esteja otimizada para o seu problema (Garcea et al., 2023).

Em relação aos melhores resultados, podemos observar que o experimento que realizou a adição de 100 imagens sintéticas com a StyleGAN2ADA (E3) aumentou a precisão da classe 1+ em 6,12 pontos percentuais. Além disso, o *recall* da classe 2+ subiu 6,88 pontos percentuais ao complementarmos o conjunto de treino com essas 100 imagens produzidas pela rede (valores em azul na região de precisão e *recall* das classes). Com isso, ganhamos 2,27 pontos percentuais no *f1-score* geral quando comparamos E3 com E1. Assim, mostramos que as imagens sintéticas desta rede podem contribuir para um melhor aprendizado mesmo em modelos que não dependem do uso de grandes quantidades de dados ou do balanceamento de classes (Cortes e Vapnik, 1995).

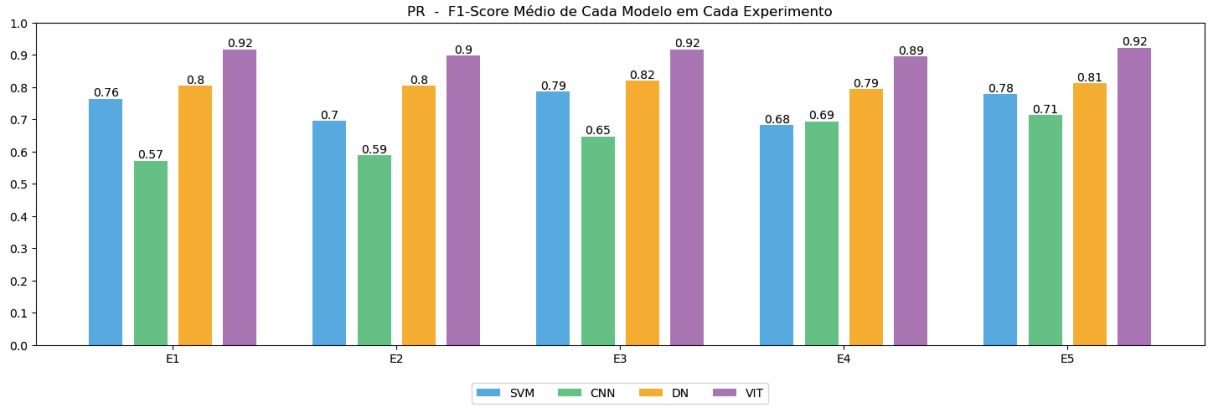


Figura 6.8: Gráfico de barras dos *f1-scores* de cada método de classificação em cada experimento, considerando a média dos resultados de todas as classes. A primeira barra de cada conjunto de quatro barras representa a Metodologia Rogalsky (SVM), a segunda a CNN, a terceira a DenseNet (DN) e a quarta o ViT. Os resultados são referentes às médias das cinco pastas do conjunto de teste de PR. Fonte: elaborado pela autora.

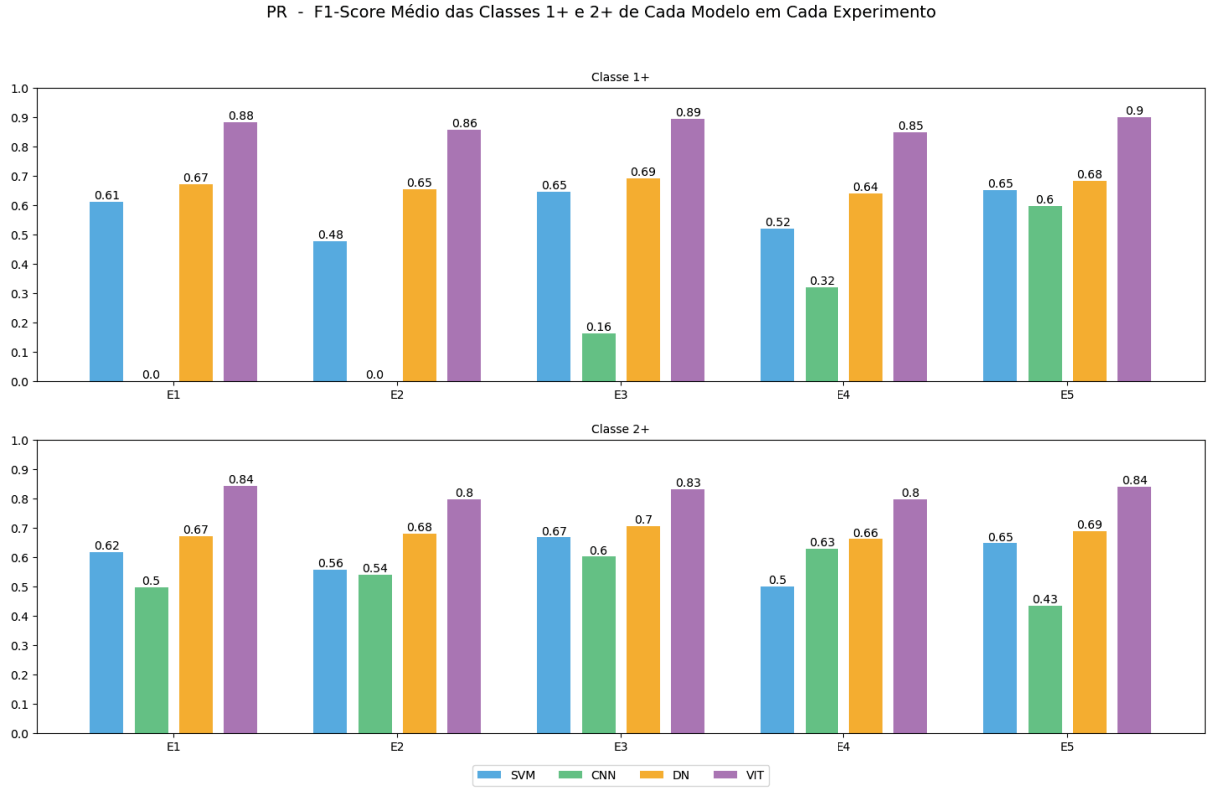


Figura 6.9: Gráfico de barras dos *f1-scores* de cada método de classificação em cada experimento, considerando as classes minoritárias (1+ e 2+). A primeira barra de cada conjunto de quatro barras representa a Metodologia Rogalsky (SVM), a segunda a CNN, a terceira a DenseNet (DN) e a quarta o ViT. Os resultados são referentes às médias das cinco pastas do conjunto de teste de PR. Fonte: elaborado pela autora.

Com a **CNN**, todos os experimentos com inserção de imagens sintéticas (E2, E3, E4 e E5) melhoraram os *f1-scores* gerais em relação ao experimento com os dados originais, como apresentamos na Figura 6.8. Analisando o gráfico de barras das classes minoritárias (Figura 6.9), notamos um comportamento incomum dos experimentos E1 e E2 na classe 1+. Como podemos ver, mesmo após o treinamento a CNN não acertou nenhum exemplo de teste nestes dois experimentos, zerando os *f1-scores*. A partir do experimento E3, a rede convolucional

passou a acertar, aumentando a métrica de 0% até 60% em E5. A classe 2+ também apresentou resultados atípicos, sendo o experimento E4 com o AutoAugment à alcançar os maiores *f1-scores* nesta classe.

Para aprofundar o nosso estudo sobre os resultados da CNN, trazemos a Tabela 6.10 com as precisões, *recalls* e *f1-scores* gerais e específicos para cada uma das classes da pontuação IS. Podemos notar aumentos na precisão e no *recall* da classe 1+, sendo o E5 superior aos outros experimentos (valores em azul na região de precisão e *recall* da classe 1+). Esses resultados são refletidos no *f1-score*, que aumentou de 0% para 59,63% com o balanceamento trazido pela StyleGAN2ADA no experimento E5 (valores em azul na região de *f1-score* da classe 1+).

No caso da classe 2+, vemos que o experimento E4 melhorou o *f1-score* em 13,06 pontos percentuais em relação à E1, sendo o melhor experimento para esta classe. No entanto, o experimento E5 superou E4 nos resultados médios, com uma diferença de 1,95 pontos percentuais. Assim, mostramos a importância e a relevância do balanceamento das classes do conjunto de treino com imagens sintéticas, principalmente para estabilizar o aprendizado de

Tabela 6.9: Resultados da Metodologia Rogalsky nos experimentos E1, E2, E3, E4 e E5. A região de cima da tabela apresenta as precisões, *recalls* e *f1-scores* específicos por classe e a debaixo as médias, considerando todas as classes. Os valores destacados em azul mostram melhorias e os valores em vermelho indicam piora em relação aos resultados do experimento E1. Os resultados são referentes às médias das cinco pastas do conjunto de teste dos Receptores de Progesterona (PR). Fonte: elaborado pela autora.

	PR - Métricas Metodologia Rogalsky														
	Precisão (%)					Recall (%)					F1-Score (%)				
Classe	E1	E2	E3	E4	E5	E1	E2	E3	E4	E5	E1	E2	E3	E4	E5
0	88,79	83,39	88,81	83,45	89,20	93,68	92,25	94,90	92,59	93,90	**91,02**	**87,36**	91,69	87,73	91,44
1+	**59,48**	55,24	**65,60**	54,86	63,86	64,25	46,06	65,00	56,20	67,50	**61,13**	**47,61**	64,57	52,02	65,04
2+	63,18	50,82	66,17	49,38	63,83	**60,86**	63,62	**67,74**	56,08	66,14	**61,59**	55,68	66,70	**50,11**	64,71
3+	94,60	94,44	94,46	92,87	93,37	89,01	81,97	88,87	76,57	87,38	**91,68**	87,58	91,56	**83,04**	90,23
Média	76,51	70,97	78,76	70,14	77,56	76,95	70,97	79,13	70,36	78,73	**76,36**	**69,56**	**78,63**	**68,23**	**77,86**
DP	07,40	08,08	05,50	08,80	04,13	05,63	12,08	05,74	13,72	05,55	04,91	06,81	04,13	06,83	03,01

Tabela 6.10: Resultados da CNN nos experimentos E1, E2, E3, E4 e E5. A região de cima da tabela apresenta as precisões, *recalls* e *f1-scores* específicos por classe e a debaixo as médias, considerando todas as classes. Os valores destacados em azul mostram melhorias e os valores em vermelho indicam piora em relação aos resultados do experimento E1. Os resultados são referentes às médias das cinco pastas do conjunto de teste dos Receptores de Progesterona (PR). Fonte: elaborado pela autora.

	PR - Métricas CNN														
	Precisão (%)					Recall (%)					F1-Score (%)				
Classe	E1	E2	E3	E4	E5	E1	E2	E3	E4	E5	E1	E2	E3	E4	E5
0	81,48	85,89	87,84	89,67	93,44	97,10	96,72	95,89	95,69	93,22	88,60	90,96	91,68	92,58	93,32
1+	**0,00**	**0,00**	**65,81**	**71,10**	**72,24**	**0,00**	**0,00**	**9,61**	**22,00**	**51,13**	**0,00**	**0,00**	**16,41**	**32,02**	**59,63**
2+	67,76	69,08	62,58	61,20	56,91	40,82	46,70	58,13	64,86	35,67	**49,68**	54,09	60,14	**62,74**	43,47
3+	85,69	87,21	87,30	91,25	83,10	95,61	93,70	94,04	89,03	95,44	90,36	90,30	90,54	90,11	88,83
Média	58,73	60,54	75,88	78,31	76,42	58,38	59,28	64,42	67,89	68,86	**57,16**	**58,84**	**64,69**	**69,36**	**71,31**
DP	03,87	04,03	13,35	07,94	06,41	03,55	04,25	03,94	06,01	03,99	02,80	03,05	05,57	06,49	03,95

modelos que não possuem pesos pré-treinados. Além disso, esses resultados nos motivam a pensar em experimentos que combinem os dois tipos de aumentos de dados, AutoAugment e StyleGAN2ADA, de modo que eles não sejam avaliados entre si, mas cooperem para aumentar ainda mais a assertividade dos métodos de classificação.

Sobre a **DenseNet**, observamos que o comportamento dos *f1-scores* não apresentaram grandes variações nos resultados gerais exibidos pela Figura 6.8. Notamos que esse comportamento também se manifestou nos resultados específicos por classe (Figura 6.9), sendo E3 o experimento superior aos demais. A Tabela 6.11 detalha os valores de precisões, *recalls* e *f1-scores* gerais e específicos para cada classe com este método de classificação.

Analisando os *f1-scores* de cada classe, os maiores ganhos do experimento E3 aconteceram nas classes 1+ e 2+, com melhorias de 2,17 e 3,4 pontos percentuais em relação ao experimento E1 (valores em azul na região do *f1-score* por classe). Essas melhorias impactaram no aumento de 1,53 pontos percentuais no resultado geral, permitindo o alcance de 82% no *f1-score*, sendo a DenseNet o primeiro classificador a passar de 80% nos dados de Receptores de Progesterona (PR). Dessa forma, concluímos que mesmo que as melhorias sejam pequenas de um experimento para o outro, elas são de extrema importância para diagnósticos mais assertivos no ambiente médico, além de mostrarem indícios de avanços para a área com o uso da StyleGAN2ADA.

Tabela 6.11: Resultados da DenseNet nos experimentos E1, E2, E3, E4 e E5. A região de cima da tabela apresenta as precisões, *recalls* e *f1-scores* específicos por classe e a debaixo as médias, considerando todas as classes. Os valores destacados em azul mostram melhorias e os valores em vermelho indicam piora em relação aos resultados do experimento E1. Os resultados são referentes às médias das cinco pastas do conjunto de teste dos Receptores de Progesterona (PR). Fonte: elaborado pela autora.

	PR - Métricas DenseNet														
	Precisão (%)					Recall (%)					F1-Score (%)				
Classe	E1	E2	E3	E4	E5	E1	E2	E3	E4	E5	E1	E2	E3	E4	E5
0	93,19	92,99	94,09	93,20	94,23	97,57	97,33	97,12	97,16	96,55	95,31	95,11	95,57	95,10	95,36
1+	72,73	71,50	75,56	68,46	73,69	63,76	60,42	64,33	60,98	63,77	**67,06**	65,34	**69,23**	64,04	68,27
2+	69,07	68,84	70,01	65,35	68,87	66,03	67,39	71,07	67,06	69,44	**67,04**	67,92	**70,44**	66,13	68,80
3+	92,01	92,81	92,46	92,75	92,25	93,02	93,82	93,02	91,50	93,02	92,49	93,30	92,72	92,10	92,57
Média	81,75	81,54	83,03	79,94	82,26	80,09	79,74	81,39	79,17	80,69	**80,47**	**80,42**	**82,00**	**79,34**	**81,25**
DP	05,79	03,50	03,40	04,84	05,08	06,53	03,64	03,75	04,34	07,53	04,48	02,77	02,82	02,98	03,37

Por fim, o **ViT** alcançou os melhores resultados da pesquisa nas imagens de PR, ultrapassando 90% de *f1-score* geral no gráfico de barras da Figura 6.8. Observando as barras de resultados das classes 1+ e 2+ na Figura 6.9, podemos ver uma diferença significativa entre os *f1-scores* do ViT em relação aos outros métodos de classificação, chegando a atingir 90% e 84% de acerto nas classes minoritárias 1+ e 2+, respectivamente. Como as diferenças entre os resultados gerais são pequenas, aprofundamos novamente o nosso estudo definindo a Tabela 6.12.

Focando somente nos resultados da classe 1+, podemos ver que o experimento E5 obteve maiores contribuições, com melhorias de 1,39 pontos percentuais em precisão e 2,18 em *recall*, decorrendo em um aumento de 1,88 em *f1-score*, comparando com E1. Analisando os resultados da classe 2+, observamos que o experimento E4 prejudicou o modelo, com quedas de 11,23 pontos percentuais em precisão, refletindo em uma perda de 4,52 em *f1-score*. Assim, deduzimos que a escolha adequada do método de aumento de dados tem um papel essencial

Tabela 6.12: Resultados da ViT nos experimentos E1, E2, E3, E4 e E5. A região de cima da tabela apresenta as precisões, *recalls* e *f1-scores* específicos por classe e a debaixo as médias, considerando todas as classes. Os valores destacados em azul mostram melhorias e os valores em vermelho indicam piora em relação aos resultados do experimento E1. Os resultados são referentes às médias das cinco pastas do conjunto de teste dos Receptores de Progesterona (PR). Fonte: elaborado pela autora.

	PR - Métricas ViT														
	Precisão (%)					Recall (%)					F1-Score (%)				
Classe	E1	E2	E3	E4	E5	E1	E2	E3	E4	E5	E1	E2	E3	E4	E5
0	98,38	98,63	98,59	99,13	98,37	98,47	97,55	98,60	97,05	98,53	98,42	98,08	98,59	98,08	98,45
1+	**90,50**	87,29	92,31	87,86	**91,89**	**86,39**	84,39	87,20	82,72	**88,57**	**88,15**	85,64	89,47	84,90	**90,03**
2+	**83,37**	76,67	80,66	**72,14**	83,88	86,02	83,40	86,70	89,69	84,86	**84,26**	79,75	83,10	**79,74**	84,05
3+	96,40	96,30	96,67	97,79	96,31	95,97	94,83	95,09	92,69	96,42	96,17	95,55	95,85	95,16	96,35
Média	92,16	89,72	92,06	89,23	92,61	91,71	90,04	91,90	90,54	92,10	**91,75**	**89,76**	**91,76**	**89,47**	**92,22**
DP	06,25	04,81	06,16	04,85	06,89	06,27	05,59	06,51	05,91	06,87	06,35	05,09	06,26	05,59	06,53

para a classificação de imagens, e que nem sempre transformações simples serão suficientes e confiáveis para trazer uma maior variedade e generalização do processo de aprendizado.

Em relação aos resultados gerais de *f1-score*, vemos que o experimento E3 manteve os resultados do experimento E1 (com um aumento de 0,01) e o experimento E5 atingiu 92,22%, melhorando a métrica em 0,47 pontos percentuais. Já os experimentos E2 e E4 reduziram as métricas em relação ao experimento com os dados originais (E1), apresentando quedas de 1,99 pontos percentuais em E2 e de 2,28 em E4 ao considerarmos os *f1-scores* gerais. Dessa forma, concluímos que o nosso melhor classificador para as categorias de intensidade dos Receptores de Progesterona (PR) foi o ViT. Acreditamos que sua arquitetura inovadora baseada em Transformers e a sua capacidade de aprender características locais, globais e as interações entre elas tenham sido os fatores principais para o seu alto desempenho.

Para resumir os resultados desta seção, criamos a Tabela 6.13 com os *f1-scores* gerais para todos os métodos de classificação em cada um dos experimentos. Assim como nos dados de Receptores de Estrogênio (ER), alcançamos as maiores métricas com os experimentos E3 e E5, prevalecendo o experimento E4 com os piores desempenhos. Sobre as melhorias, conseguimos passar do limite dos 90% de *f1-score* neste conjunto de dados, combinando o ViT com o balanceamento de classes das imagens de treino trazido pela StyleGAN2ADA. Acreditamos que esta alta performance tenha relação com as inovações da arquitetura tanto do ViT quanto das otimizações da StyleGAN2ADA, reforçando que a escolha de um método inadequado de aumento de dados pode prejudicar o aprendizado dos modelos. Essas questões mostram a importância do investimento na pesquisa e na atualização dos modelos de inteligência artificial para auxiliar nos diagnósticos da área médica.

Novamente, quando levamos em consideração o balanceamento das classes com o AutoAugment, os resultados apresentam quedas de *f1-score* em todos os modelos ao compararmos com E1, exceto com a CNN. Como mencionamos, com uma arquitetura mais simples e sem o uso de pesos pré-treinados, a rede pode ter se beneficiado pelo aumento da quantidade de dados sintéticos para estabilizar o seu treinamento (ao Huang et al., 2022). Para aprofundar o nosso estudo, planejamos realizar as análises quantitativas e qualitativas para ambos métodos de geração de imagens neste conjunto de dados no futuro. Mas por questões de tempo e disponibilidade dos especialistas, focamos somente nas análises visuais e nos resultados na classificação.

Tabela 6.13: Resultados dos métodos de classificação nos experimentos E1, E2, E3, E4 e E5, considerando a média dos *f1-scores* de todas as classes. Os valores destacados em azul informam os melhores resultados de cada método e os valores em vermelho indicam os piores desempenhos. Os resultados são referentes às médias das cinco pastas do conjunto de teste dos Receptores de Progesterona (PR). Fonte: elaborado pela autora.

	PR - F1-Score Médio (%)				
Modelo	E1	E2	E3	E4	E5
Met. Rogalsky	76,36 ± 4,91	69,56 ± 6,81	**78,63** ± 4,13	**68,23** ± 6,83	77,86 ± 3,01
CNN	**57,16** ± 2,80	58,84 ± 3,05	64,69 ± 5,57	69,36 ± 6,49	**71,31** ± 3,95
DenseNet	80,47 ± 4,48	80,42 ± 2,77	**82,00** ± 2,82	**79,34** ± 2,98	81,25 ± 3,37
ViT	91,75 ± 6,35	89,76 ± 5,09	91,76 ± 6,26	**89,47** ± 5,59	**92,22** ± 6,53

Por fim, mostramos as acurácias gerais para todos os métodos de classificação em cada um dos experimentos na Tabela 6.14. Podemos ver que experimentos que alcançaram *f1-scores* baixos na tabela anterior, apresentam acurácias altas, próximas à 90% (como a CNN em E1 e os resultados de E4, por exemplo). Além disso, essa métrica não demonstrou diferenças significativas entre os três primeiros modelos no experimento E5. Assim, reiteramos as conclusões da seção anterior, enfatizando que a acurácia não relata adequadamente o comportamento dos classificadores em um conjunto de teste desbalanceado, sendo de extrema importância a compreensão aprofundada da métrica para evitar análises equivocadas (Thölke et al., 2023).

Tabela 6.14: Resultados dos métodos de classificação nos experimentos E1, E2, E3, E4 e E5, considerando a média das acurácias de todas as classes. Os valores destacados em azul informam os melhores resultados de cada método e os valores em vermelho indicam os piores desempenhos. Os resultados são referentes às médias das cinco pastas do conjunto de teste dos Receptores de Progesterona (PR). Fonte: elaborado pela autora.

	PR - Acurácia Média (%)				
Modelo	E1	E2	E3	E4	E5
Met. Rogalsky	91,85 ± 2,10	89,26 ± 2,33	**92,55** ± 1,60	**88,03** ± 4,19	92,06 ± 1,30
CNN	**90,31** ± 1,02	90,92 ± 1,09	91,28 ± 1,39	**91,60** ± 1,31	91,12 ± 1,44
DenseNet	93,63 ± 1,50	93,75 ± 0,97	**94,03** ± 0,96	**93,20** ± 1,00	93,75 ± 1,44
ViT	97,12 ± 2,24	96,44 ± 1,87	97,03 ± 2,25	**96,23** ± 2,01	**97,27** ± 2,28

7 CONSIDERAÇÕES FINAIS

Com a difusão das WSIs, a automatização confiável de diagnósticos na área médica se tornou essencial para reduzir o tempo e esforço dos especialistas em tarefas demoradas, repetitivas e exaustivas. Nesta dissertação, vimos que muitos progressos foram alcançados, principalmente com a aplicação de *Deep Learning* nos problemas de classificação e segmentação. Como discutimos, modelos baseados em redes neurais profundas necessitam de uma grande quantidade de dados com um alto padrão de qualidade para serem treinados. Assim, as limitações em relação à escassez de dados, desbalanceamento de classes e falta de disponibilidade pública dessas informações médicas ainda são barreiras para resultados confiáveis e robustos.

Visando lidar com essas questões, criamos uma metodologia capaz de aumentar a assertividade de métodos de classificação envolvendo a adição de imagens médicas sintéticas ao processo de treinamento. Para isso, nos concentramos em classificar a pontuação de intensidade (IS) celular de *patches* extraídos das imagens de biomarcadores de Receptores de Estrogênio (ER) e Progesterona (PR), responsáveis pela detecção e caracterização do câncer de mama. Nossa solução envolveu o treinamento e a geração de imagens sintéticas com a rede StyleGAN2ADA e o modelo AutoAugment para incorporar novas imagens ao conjunto de treinamento dos métodos de classificação: Metodologia Rogalsky, CNN, DenseNet e ViT.

Como primeira etapa, focamos no treinamento e avaliação das imagens produzidas pela rede generativa StyleGAN2ADA. Ao final do processo de treino, a rede apresentou valores abaixo de 4% em relação às distâncias entre as distribuições de características das imagens geradas e das imagens reais, mostrando a convergência do treinamento em ambos conjuntos de dados. Na segunda etapa, implementamos avaliações quantitativas e qualitativas para comprovar a qualidade do modelo generativo no contexto dos biomarcadores de Receptores de Estrogênio (ER).

Para a avaliação quantitativa, aplicamos a métrica DreamSIM, que é um método pertencente ao estado da arte capaz de medir a similaridade entre duas imagens. Nós alteramos o contexto de aplicação da métrica para calcular a similaridade entre as imagens sintéticas e reais, alcançando um valor de 0,091. Com este resultado, demonstramos a qualidade das imagens sintéticas, visto que elas são similares às imagens reais ao mesmo tempo que trazem variações em iluminação, contraste, formato e organização das células.

Sobre a análise qualitativa, nós convidamos três especialistas em patologia para participarem de um experimento de Categorização Rápida de Cenas (CRC). Para isso, desenvolvemos uma interface para os médicos avaliarem se as imagens apresentadas eram reais ou sintéticas, estabelecendo um curto período de visualização de três segundos. Como resultado, todos os patologistas informaram que mais da metade das imagens geradas apresentadas eram reais, demonstrando a similaridade visual das imagens sintéticas em relação às imagens reais.

Em relação a classificação dos níveis de câncer das imagens de Receptores de Estrogênio (ER) e Progesterona (PR), nós implementamos cinco experimentos: treinamos os classificadores com as imagens originais (E1), adicionamos 100 imagens sintéticas ao processo de treinamento com as técnicas AutoAugment e StyleGAN2ADA (E2 e E3) e balanceamos as classes do conjunto de treino e inserimos mais 100 imagens de cada classe com as duas técnicas (E4 e E5). Nos dados de ER, alcançamos os melhores resultados de classificação com a DenseNet, com 87,46% de *f1-score* no experimento E3, melhorando a métrica em 10,58 pontos percentuais em relação ao experimento com os dados originais. No conjunto de dados de PR, atingimos um *f1-score* de 92,22% com o ViT no experimento E5, além de alcançarmos um aumento de 14,15 pontos percentuais em relação à E1 com a CNN no experimento E5.

Em ambos conjuntos de dados, os experimentos que envolveram a adição de imagens sintéticas da técnica AutoAugment não alcançaram bons resultados, levando a pioras de até 8 pontos percentuais em relação ao experimento com os dados originais. Dessa forma, concluímos que técnicas simples de aumento de dados podem não ser adequadas para todos os tipos de imagens médicas, pois rotações, translações e, principalmente, mudanças de contraste podem interferir em características essenciais das patologias, prejudicando o aprendizado do classificador. Já as imagens da StyleGAN2ADA contribuíram para melhores resultados na maioria dos experimentos, trazendo uma maior variabilidade ao conjunto de treinamento e ajudando na generalização dos modelos de classificação.

Com estes resultados, mostramos a viabilidade de uso da StyleGAN2ADA no cenário de imagens médicas, garantindo uma maior assertividade e robustez ao processo de classificação. Além disso, destacamos a importância da escolha da métrica de avaliação dos resultados de classificação em problemas desbalanceados. Mostramos que métricas como a acurácia podem ser afetadas pelo desbalanceamento, levando a conclusões errôneas dos desempenhos dos classificadores. Definindo o *f1-score* como métrica principal, descobrimos que as imagens sintéticas da StyleGAN2ADA apresentaram um maior impacto nas classes minoritárias (1+ e 2+), aumentando em até 39 pontos percentuais a classe 1+ nos dados de ER.

Até onde sabemos, somos os primeiros a utilizar a StyleGAN2ADA no contexto de imagens de IHQ e a aplicar a métrica DreamSIM para avaliar a qualidade das imagens produzidas pela rede. Por fim, apresentamos diversas alternativas para trabalhos futuros, como a aplicação da metodologia em outros conjuntos de dados com diferentes patologias; o uso da normalização de cores; a otimização da quantidade de imagens sintéticas inseridas no conjunto de treino e estudos de interpretabilidade dos modelos para facilitar a adoção desta proposta em ambientes clínicos. Assim, concluímos que apesar de existirem muitos desafios, a nossa metodologia alcançou bons resultados, com *f1-scores* próximos à 90% em dois conjuntos de dados, além de comprovar a qualidade das imagens sintéticas. Acreditamos que a nossa proposta seja só o começo, com potencial de contribuição para o avanço científico em diversas áreas.

7.1 TRABALHOS FUTUROS

Com base no trabalho desenvolvido, concluímos que o uso de imagens sintéticas geradas pela StyleGAN2ADA pode ser uma abordagem promissora para aumentar a assertividade dos diagnósticos assistidos por inteligência artificial no ambiente médico. Portanto, elaboramos este capítulo com o objetivo de discutir algumas ideias que poderiam ser aplicadas para aprimorar os resultados de nossa pesquisa no futuro, proporcionando maior robustez e confiabilidade à metodologia proposta.

- **Realizar todos os testes quantitativos e qualitativos:** em razão de questões de tempo e disponibilidade dos patologistas, realizamos testes quantitativos e qualitativos apenas para as imagens produzidas pela StyleGAN2ADA no conjunto de dados de Receptores de Estrogênio (ER). Para complementar os nossos resultados e trazer uma maior confiabilidade para nossas análises, planejamos executar estes experimentos também para as imagens sintéticas do AutoAugment e para o conjunto de dados de Receptores de Progesterona (PR). Além disso, pensamos em ampliar a análise qualitativa, de forma a incorporar o método de Julgamento por Preferência para categorizar as imagens sintéticas nas classes “boa”, “média” e “ruim” (Borji, 2018).

- **Avaliar a metodologia em outros conjunto de dados:** nossa metodologia foi estruturada de forma a permitir facilmente a adaptação do domínio do problema, podendo ser

utilizada para diversos tipos de imagens. Dessa forma, um dos caminhos futuros de pesquisa poderia abranger a aplicação dos métodos propostos em diferentes conjuntos de dados, compreendendo outras imagens de câncer de mama (como Aksac et al. (2019) e Spanhol et al. (2016b), apresentados nas figuras 5.1 e 7.1 respectivamente), ou até mesmo de patologias distintas, como classificações de demência (Tang et al., 2019) ou doenças pulmonares (Wang et al., 2023).

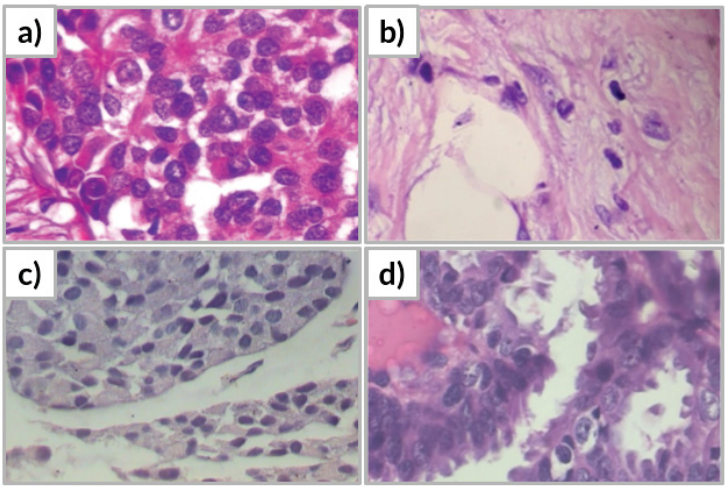


Figura 7.1: Exemplos de cada classe do conjunto de dados *BreaKHis* (Spanhol et al., 2016b). Em *a)* é apresentada a classe *ductal carcinoma*; em *b)* a classe *lobular carcinoma*; em *c)* a classe *mucinous carcinoma* e em *d)* a classe *papillary carcinoma*. Fonte: elaborado pela autora.

- **Aplicar uma normalização de cores:** durante o desenvolvimento do trabalho, observamos que as imagens de IHQ podem apresentar diferenças de coloração de uma lâmina para outra. Além disso, características como brilho e contraste podem variar durante a fase de aquisição dos dados, impactando diretamente na qualidade das WSIs (Tang et al., 2019). Na Figura 7.1, apresentamos exemplos dessas variações, onde podemos notar as diferenças de cores, iluminação e contraste entre as lâminas. Para mitigar esses problemas, poderiam ser aplicadas técnicas de ajuste de contraste e normalização de cores, como a CLAHE (Yadav et al., 2014) e a Reinhard (Tang et al., 2019). Em estudos iniciais, aplicamos estas técnicas no contexto de quantificação do índice proliferativo Ki-67 do câncer de mama, propiciando melhorias na etapa de segmentação (Ribeiro et al., 2024). Assim, uma maior qualidade nas imagens de entrada poderia ser alcançada no contexto dos receptores de hormônio, impactando em melhores resultados nas etapas de geração e classificação dos níveis de câncer de mama.

- **Selecionar as melhores imagens sintéticas:** na fase de avaliação quantitativa, descobrimos que algumas imagens sintéticas apresentam uma maior similaridade com as imagens do conjunto de treinamento. Embora a geração de imagens diversificadas possa ser um fator contribuinte para uma melhor generalização dos modelos de classificação, certas características das imagens sintéticas podem prejudicar o processo de aprendizado. Assim, uma ideia para o futuro seria definir um método de seleção para filtrar as imagens sintéticas, adicionando ao conjunto de treinamento somente dados de qualidade. Essa seleção poderia levar em consideração as características mais relevantes de acordo com os patologistas em conjunto com os valores da métrica DreamSIM, por exemplo.

- **Otimizar a quantidade de imagens sintéticas adicionadas ao treino:** durante os experimentos, consideramos apenas duas formas de aumento de dados, com o

balanceamento das classes e a adição de 100 imagens sintéticas ao conjunto de treinamento. Mas como vimos, cada tipo de dado e cada modelo de classificação pode se beneficiar de uma quantidade diferente de imagens sintéticas. Assim, pretendemos otimizar a quantidade de imagens geradas adicionadas ao conjunto de treino, investindo em técnicas de *undersampling*, variações nas proporções dos dados sintéticos e também na combinação das imagens do AutoAugment com as imagens da StyleGAN2ADA. Dessa forma, podemos aprofundar nosso entendimento sobre impacto dessas imagens na solução, descobrindo, por exemplo, se a adição de 300 imagens seria mais adequada do que o balanceamento dos exemplos de treino conforme a classe majoritária.

- **Otimizar os hiperparâmetros dos modelos:** devido a questões de tempo, realizamos pequenos ajustes nos hiperparâmetros dos modelos, sendo a maioria dos valores definidos conforme as recomendações das bibliotecas. Sabemos que a otimização de hiperparâmetros é de extrema importância no cenário de aprendizado de máquina, pois pode aumentar significativamente o desempenho dos modelos generativos e de classificação. Assim, pretendemos implementar técnicas como a *Grid Search*, a otimização Bayesiana ou os algoritmos evolutivos para encontrar os valores ótimos para os hiperparâmetros (Bischl et al., 2023), como, por exemplo, o valor de truncamento da StyleGAN2ADA e a taxa de *dropout* das redes neurais.

- **Implementar um sistema de votação:** a abordagem que apresentamos neste trabalho se concentrou na classificação dos *patches* extraídos das WSIs. Em contextos de análise de imagens imunohistoquímicas, os especialistas costumam rotular as WSIs com uma única classe, representando o estágio do câncer de mama presente na amostra. Assim, uma próxima fase de pesquisa envolveria o desenvolvimento de um sistema de votação, no qual a WSI de teste seria dividida em *patches* e cada um deles seria submetido à classificação pelos modelos treinados. A classe mais frequentemente atribuída dentre as classificações poderia ser designada como a classe da WSI. Contudo, seria crucial estabelecer a relevância das classes neste sistema de votação em colaboração com um especialista em patologia, garantindo uma classificação com maior precisão.

- **Estudar a interpretabilidade dos modelos:** os ambientes clínicos ainda encaram os modelos de *Deep Learning* com muita incerteza e insegurança. Para aumentar as chances de implantação dos sistemas de auxílio ao diagnóstico, pretendemos estudar a interpretabilidade desses modelos, incluindo tanto a StyleGAN2ADA quanto os classificadores baseados em redes neurais. Algumas alternativas incluem destacar as regiões relevantes da imagem de entrada utilizando o método *Guided Gradient-weighted Class Activation Mapping* (Guided Grad-CAM) e aplicar técnicas de oclusão de características para identificar quais coordenadas da imagem são cruciais para uma classificação de qualidade (Tang et al., 2019). Ambos métodos necessitam da validação dos especialistas para garantir que as regiões identificadas pelas redes correspondam aos conceitos de diagnóstico dos patologistas.

- **Publicar o conteúdo da pesquisa em uma revista:** durante o desenvolvimento deste trabalho, realizamos uma publicação internacional na conferência CMBES (*Canadian Medical and Biological Engineering Society*) no contexto de imagens de câncer de mama com o índice Ki-67 (Ribeiro et al., 2024). No cenário das imagens de receptores de hormônio, submetemos os resultados dos Receptores de Estrogênio (ER) para a revista *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*. Adicionalmente, planejamos trabalhar em uma futura publicação

contendo também os resultados advindos das imagens de progesterona (PR) com o processo de validação cruzada, cobrindo de maneira mais abrangente a nossa pesquisa.

REFERÊNCIAS

Ajit, A., Acharya, K. e Samanta, A. (2020). A Review of Convolutional Neural Networks. Em *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, páginas 1–5. https://doi.org/10.1109/ic-ETITE47903.2020.049. Acessado em: 01/04/2023.

Aksac, A., Demetrick, D. J., Ozyer, T. e Alhajj, R. (2019). Brecahad: a dataset for breast cancer histopathological annotation and diagnosis. *BMC Research Notes*, 12. https://doi.org/10.1186/s13104-019-4121-7. Acessado em: 30/04/2023.

American Cancer Society (2021a). Breast Cancer Hormone Receptor Status. https://www.cancer.org/cancer/breast-cancer/understanding-a-breast-cancer-diagnosis/breast-cancer-hormone-receptor-status.html. Acessado em: 11/03/2023.

American Cancer Society (2021b). Types of Breast Cancer. https://www.cancer.org/cancer/breast-cancer/about/types-of-breast-cancer.html. Acessado em: 11/03/2023.

American Cancer Society (2021c). What Is Breast Cancer? https://www.cancer.org/cancer/breast-cancer/about/what-is-breast-cancer.html. Acessado em: 11/03/2023.

ao Huang, Z., Sang, Y., Sun, Y. e Lv, J. (2022). A neural network learning algorithm for highly imbalanced data classification. *Information Sciences*, 612:496–513. https://doi.org/10.1016/j.ins.2022.08.074. Acessado em: 30/04/2023.

Arjovsky, M., Chintala, S. e Bottou, L. (2017). Wasserstein gan. https://doi.org/10.48550/arXiv.1701.07875, Acessado em: 11/05/2023.

Bai, Y., Yang, E., Han, B., Yang, Y., Li, J., Mao, Y., Niu, G. e Liu, T. (2021). Understanding and improving early stopping for learning with noisy labels.

Barnes, M., Srinivas, C., Bai, I., Frederick, J., Liu, W., Sarkar, A., Wang, X., Nie, Y., Portier, B., Kapadia, M., Sertel, O., Little, E., Sabata, B. e Ranger-Moore, J. (2017). Whole tumor section quantitative image analysis maximizes between-pathologists' reproducibility for clinical immunohistochemistry-based biomarkers. *Laboratory Investigation*, 97. https://doi.org/10.1038/labinvest.2017.82. Acessado em: 25/04/2023.

Bhattacharya, S., Kavousi, P., Carr, T. e Pantaleone, S. (2019). Application of predictive data analytics to model daily hydrocarbon production using petrophysical, geomechanical, fiber-optic, completions, and surface data: A case study from the marcellus shale, north america. *Journal of Petroleum Science and Engineering*, 176. http://dx.doi.org/10.1016/j.petrol.2019.01.013. Acessado em: 29/03/2023.

Bio-Techne (2022). Immunohistochemistry (IHC) Handbook. https://resources.bio-techne.com/bio-techne-assets/docs/literature/BT-Immunohistochemistry-Handbook.pdf. Acessado em: 11/03/2023.

Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., Thomas, J., Ullmann, T., Becker, M., Boulesteix, A.-L., Deng, D. e Lindauer, M. (2023). Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *WIREs Data Mining and Knowledge Discovery*, 13(2):e1484.

Bora, D. J., Gupta, A. K. e Khan, F. A. (2015). Comparing the performance of l*a*b* and hsv color spaces with respect to color image segmentation. https://doi.org/10.48550/arXiv.1506.01472. Acessado em: 18/03/2023.

Borji, A. (2018). Pros and Cons of GAN Evaluation Measures. https://doi.org/10.48550/arXiv.1802.03446. Acessado em: 27/04/2023.

Che, Y., Ren, F., Zhang, X., Cui, L., Wu, H. e Zhao, Z. (2023). Immunohistochemical HER2 Recognition and Analysis of Breast Cancer Based on Deep Learning. *Diagnostics*. https://doi.org/10.3390/diagnostics13020263. Acessado em: 08/02/2023.

Cordeiro, C. Q. (2019). An Automatic Patch-Based Approach for HER-2 Scoring in Immunohistochemical Breast Cancer Images. https://acervodigital.ufpr.br/handle/1884/66131. Acessado em: 08/02/2023.

Cordeiro, C. Q., Ioshii, S. O., Alves, J. H. e de Oliveira, L. F. (2018). An Automatic Patch-based Approach for HER-2 Scoring in Immunohistochemical Breast Cancer Images Using Color Features. *XVIII Simpósio Brasileiro de Computação Aplicada à Saúde*. https://doi.org/10.5753/sbcas.2018.3685. Acessado em: 08/02/2023.

Cortes, C. e Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20. https://doi.org/10.1007/BF00994018. Acessado em: 27/03/2023.

Cubuk, E. D., Zoph, B., Mané, D., Vasudevan, V. e Le, Q. V. (2019). Autoaugment: Learning augmentation strategies from data. Em *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, páginas 113–123.

Cunningham, P., Cord, M. e Delany, S. J. (2008). *Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval*, páginas 21–49. Springer Berlin Heidelberg, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-75171-7_2. Acessado em: 28/03/2023.

Data Science Academy (2022). Deep learning book. https://www.deeplearningbook.com.br/. Acessado em: 11/04/2023.

Denton, E., Chintala, S., Szlam, A. e Fergus, R. (2015). Deep generative image models using a laplacian pyramid of adversarial networks. Em *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, página 1486–1494, Cambridge, MA, USA. MIT Press.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. e Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale.

Fu, S., Tamir, N., Sundaram, S., Chai, L., Zhang, R., Dekel, T. e Isola, P. (2023). Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv:2306.09344*.

Garcea, F., Serra, A., Lamberti, F. e Morra, L. (2023). Medical image data augmentation: techniques, comparisons and interpretations. *Artificial Intelligence Review*, 56.

Gonzalez, R. C. e Woods, R. E. (2008). *Digital image processing*. Prentice Hall.

Gulakala, R., Markert, B. e Stoffel, M. (2022). Generative adversarial network based data augmentation for cnn based detection of covid-19. *Scientific Reports*, 12.

Han, Z., Wei, B., Zheng, Y., Yin, Y., Li, K. e Li, S. (2017). Breast cancer multi-classification from histopathological images with structured deep learning model. *Scientific Reports*, 7. https://doi.org/10.1038/s41598-017-04075-z. Acessado em: 25/04/2023.

Harvey, J. M., Clark, G. M., Osborne, C. K. e Allred, D. C. (1999). Estrogen Receptor Status by Immunohistochemistry Is Superior to the Ligand-Binding Assay for Predicting Response to Adjuvant Endocrine Therapy in Breast Cancer . *Journal of Clinical Oncology*, 17. https://doi.org/10.1200/JCO.1999.17.5.1474. Acessado em: 25/04/2024.

He, K., Zhang, X., Ren, S. e Sun, J. (2015). Deep Residual Learning for Image Recognition. https://doi.org/10.48550/arXiv.1512.03385. Acessado em: 19/04/2023.

He, K., Zhang, X., Ren, S. e Sun, J. (2016). Deep residual learning for image recognition. Em *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, páginas 770–778.

Hemanth, D., Gupta, D. e Balas, V. (2019). *Intelligent Data Analysis for Biomedical Applications: Challenges and Solutions*. Intelligent Data-Centric Systems. Elsevier Science.

Herrero-Lopez, S. (2011). Chapter 20 - Multiclass Support Vector Machine. Em mei W. Hwu, W., editor, *GPU Computing Gems Emerald Edition*, Applications of GPU Computing Series, páginas 293–311. Morgan Kaufmann, Boston. https://doi.org/10.1016/B978-0-12-384988-5.00020-6. Acessado em: 29/03/2023.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B. e Hochreiter, S. (2018). GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. https://doi.org/10.48550/arXiv.1706.08500. Acessado em: 28/04/2023.

Hira, Z. M. e Gillies, D. F. (2015). A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data. *Adv Bioinformatics*, 2015. https://doi.org/10.1155/2015/198363. Acessado em: 27/03/2023.

Hong, S., Marinescu, R., Dalca, A. V., Bonkhoff, A. K., Bretzner, M., Rost, N. S. e Golland, P. (2021). 3D-StyleGAN: A Style-Based Generative Adversarial Network for Generative Modeling of Three-Dimensional Medical Images. *arXiv*. https://doi.org/10.48550/arxiv.2107.09700. Acessado em: 08/02/2023.

Horak, D., Yu, S. e Salimi-Khorshidi, G. (2020). Topology Distance: A Topology-Based Approach For Evaluating Generative Adversarial Networks. https://doi.org/10.48550/arXiv.2002.12054. Acessado em: 29/04/2023.

Horev, R. (2018). Explained: A style-based generator architecture for gans - generating and tuning realistic artificial faces. https://doi.org/10.1159/000442389 Acessado em: 15/04/2023.

Huang, G., Liu, Z., Maaten, L. V. D. e Weinberger, K. Q. (2017). Densely connected convolutional networks. Em *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, páginas 2261–2269, Los Alamitos, CA, USA. IEEE Computer Society.

IARC (2023). Cancer Today. https://gco.iarc.fr/today/online-analysis-multi-bars?v=2020. Acessado em: 09/05/2023.

INCA (2020). Atlas On-line de Mortalidade. https://www.inca.gov.br/MortalidadeWeb/pages/Modelo03/consultar.xhtml#panelResultado. Acessado em: 09/05/2023.

INCA (2022). Outubro Rosa 2022. https://www.gov.br/inca/pt-br/assuntos/campanhas/2022/outubro-rosa. Acessado em: 09/05/2023.

Islam, J. e Zhang, Y. (2020). GAN-based synthetic brain PET image generation. *Brain Informatics*. https://doi.org/10.1186/s40708-020-00104-2. Acessado em: 08/02/2023.

Islam, S., Elmekki, H., Elsebai, A., Bentahar, J., Drawel, N., Rjoub, G. e Pedrycz, W. (2024). A comprehensive survey on applications of transformers for deep learning tasks. *Expert Systems with Applications*, 241:122666. https://doi.org/10.1016/j.eswa.2023.122666. Acessado em: 30/04/2023.

Karnewar, A. e Wang, O. (2020). MSG-GAN: Multi-Scale Gradients for Generative Adversarial Networks. https://doi.org/10.48550/arXiv.1903.06048. Acessado em: 19/04/2023.

Karras, T., Aila, T., Laine, S. e Lehtinen, J. (2018). Progressive Growing of GANs for Improved Quality, Stability, and Variation. https://doi.org/10.48550/arXiv.1710.10196. Acessado em: 14/04/2023.

Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J. e Aila, T. (2020a). Training Generative Adversarial Networks with Limited Data. https://doi.org/10.48550/arXiv.2006.06676. Acessado em: 24/04/2023.

Karras, T., Laine, S. e Aila, T. (2019). A Style-Based Generator Architecture for Generative Adversarial Networks. https://doi.org/10.48550/arXiv.1812.04948. Acessado em: 14/04/2023.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J. e Aila, T. (2020b). Analyzing and Improving the Image Quality of StyleGAN. https://doi.org/10.48550/arXiv.1912.04958. Acessado em: 19/04/2023.

Kim, S.-W., Roh, J. e Park, C.-S. (2016). Immunohistochemistry for Pathologists: Protocols, Pitfalls, and Tips. *Journal of Pathology and Translational Medicine 2016; 50: 411-418*. https://doi.org/10.4132/jptm.2016.08.08. Acessado em: 16/03/2023.

Krinski, B. A., Ruiz, D. V., Laroca, R. e Todt, E. (2023). DACov: a deeper analysis of data augmentation on the computed tomography segmentation problem. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 0(0):1–18. https://doi.org/10.1080/21681163.2023.2183807. Acessado em: 29/05/2023.

Laurinavicius, A., Plancoulaine, B., Herlin, P. e Laurinaviciene, A. (2016). Comprehensive Immunohistochemistry: Digital, Analytical and Integrated. *Pathobiology 2016;83:156-163*. https://doi.org/10.1159/000442389. Acessado em: 08/02/2023.

LeCun, Y., Kavukcuoglu, K. e Farabet, C. (2010). Convolutional networks and applications in vision. Em *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, páginas 253–256.

Ledley, R., Buas, M. e Golab, T. (1990). Fundamentals of true-color image processing. Em *[1990] Proceedings. 10th International Conference on Pattern Recognition*, volume i, páginas 791–795 vol.1.

Liao, Y.-T., Lee, C.-H., Chen, K. S., Chen, C.-P. e Pai, T.-W. (2021). Data augmentation based on generative adversarial networks to improve stage classification of chronic kidney disease. *Applied Sciences*, 12:352.

Magalhães, G. M., Terra, E. M., Calazans, S. G., de O. Vasconcelos, R. e Alessi, A. C. (2014). Avaliação da imunomarcação de células-tronco tumorais em carcinossarcomas mamários e carcinomas em tumores mistos em cadelas. *Pesquisa Veterinária Brasileira, 2014, 455-461*. http://dx.doi.org/10.1590/S0100-736X2014000500012. Acessado em: 16/03/2023.

Maleki, F., Muthukrishnan, N., Ovens, K., Md, C. e Forghani, R. (2020). Machine Learning Algorithm Validation. *Neuroimaging Clinics of North America*, 30:433–445. http://dx.doi.org/10.1016/j.nic.2020.08.004. Acessado em: 25/04/2023.

M.Ghandour, S. e Turab, N. (2016). Geographic Maps Classification Based on L*A*B Color System. *International journal of Computer Networks & Communications*, 8:81–92. http://dx.doi.org/10.5121/ijcnc.2016.8306. Acessado em: 17/03/2023.

Mouelhi, A., Rmili, H., Ali, J. B., Sayadi, M., Doghri, R. e Mrad, K. (2018). Fast unsupervised nuclear segmentation and classification scheme for automatic allred cancer scoring in immunohistochemical breast tissue images. *Computer Methods and Programs in Biomedicine*. https://doi.org/10.1016/j.cmpb.2018.08.005. Acessado em: 08/02/2023.

Mridha, M. F., Morol, M. K., Ali, M. A. e Shovon, M. S. H. (2022). convoHER2: A Deep Neural Network for Multi-Stage Classification of HER2 Breast Cancer. *arXiv*. https://doi.org/10.48550/arxiv.2211.10690. Acessado em: 08/02/2023.

Mukherkjee, D., Saha, P., Kaplun, D., Sinitca, A. e Sarkar, R. (2022). Brain tumor image generation using an aggregation of GAN models with style transfer. *Sci Rep. 2022; 12: 9141*. https://doi.org/10.1038%2Fs41598-022-12646-y. Acessado em: 08/02/2023.

Mukundan, R. (2017). A robust algorithm for automated her2 scoring in breast cancer histology slides using characteristic curves. Em Valdés Hernández, M. e González-Castro, V., editores, *Medical Image Understanding and Analysis*, páginas 386–397, Cham. Springer International Publishing.

Mutlag, W. K., Ali, S. K., Aydam, Z. M. e Taher, B. H. (2020). Feature Extraction Methods: A Review. *Journal of Physics: Conference Series*, 1591. http://dx.doi.org/10.1088/1742-6596/1591/1/012028. Acessado em: 27/03/2023.

National Cancer Institute (2021). What Is Cancer? https://www.cancer.gov/about-cancer/understanding/what-is-cancer. Acessado em: 11/03/2023.

Osuala, R., Kushibar, K., Garrucho, L., Linardos, A., Szafranowska, Z., Klein, S., Glocker, B., Diaz, O. e Lekadir, K. (2023). Data synthesis and adversarial networks: A review and

meta-analysis in cancer imaging. *Medical Image Analysis*. https://doi.org/10.1016/j.media.2022.102704. Acessado em: 08/02/2023.

Otsu, N. (1979). A threshold selection method from gray level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9:62–66. https://doi.org/10.1109/TSMC.1979.4310076. Acessado em: 25/03/2023.

Pecha, M. e Horák, D. (2020). *Analyzing L1-loss and L2-loss Support Vector Machines Implemented in PERMON Toolbox*, páginas 13–23. International Conference on Advanced Engineering Theory and Applications. http://dx.doi.org/10.1007/978-3-030-14907-9_2. Acessado em: 29/03/2023.

Peng, Y.-T., Chen, Y.-R., Chen, Z., Wang, J.-H. e Huang, S.-C. (2022). Underwater Image Enhancement Based on Histogram-Equalization Approximation Using Physics-Based Dichromatic Modeling. *Sensors (Basel).*, 8. https://doi.org/10.3390/s22062168. Acessado em: 20/03/2023.

Ponti, M. A., Ribeiro, L. S. F., Nazare, T. S., Bui, T. e Collomosse, J. (2017). Everything You Wanted to Know about Deep Learning for Computer Vision but Were Afraid to Ask. Em *2017 30th SIBGRAPI Conference on Graphics, Patterns and Images Tutorials (SIBGRAPI-T)*, páginas 17–41. https://doi.org/10.1109/SIBGRAPI-T.2017.12. Acessado em: 02/04/2023.

Qaiser, T., Mukherjee, A., Pb, C. R., Munugoti, S. D., Tallam, V., Pitkäaho, T., Lehtimäki, T., Naughton, T., Berseth, M., Pedraza, A., Mukundan, R., Smith, M., Bhalerao, A., Rodner, E., Simon, M., Denzler, J., Huang, C.-H., Bueno, G., Snead, D., Ellis, I. O., Ilyas, M. e Rajpoot, N. (2017). HER2 challenge contest: a detailed assessment of automated HER2 scoring algorithms in whole slide images of breast cancer tissues. *Histopathology. 2018 Jan;72(2):227-238*. https://doi.org/10.1111/his.13333. Acessado em: 08/02/2023.

Radford, A., Metz, L. e Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. https://doi.org/10.48550/arXiv.1511.06434, Acessado em: 11/05/2023.

Ribeiro, K., Guimarães Moreira, N., Train, G., Massaneiro, A. N., Ossamu Ioshii, S. e de Oliveira, L. (2024). Semi-automation of the ki-67 proliferative index quantification method in breast cancer. *CMBES Proceedings*, 46.

Rmili, H., Mouelhi, A., Solaiman, B., Doghri, R. e Labidi, S. (2022). A novel pre-processing approach based on colour space assessment for digestive neuroendocrine tumour grading in immunohistochemical tissue images. *Pol J Pathol. 2022;73(2):134-158*. https://doi.org/10.5114/pjp.2022.119841. Acessado em: 08/02/2023.

Rogalsky, J. E. (2021). Semi-automatic ER and PR scoring in immunohistochemistry H-BAD breast cancer images. https://acervodigital.ufpr.br/handle/1884/73470. Acessado em: 08/02/2023.

Rogalsky, J. E., Ioshii, S. O. e de Oliveira, L. F. (2021). Automatic ER and PR scoring in Immunohistochemistry H-DAB Breast Cancer images. *XXI Simpósio Brasileiro de Computação Aplicada à Saúde*. https://doi.org/10.5753/sbcas.2021.16075. Acessado em: 08/02/2023.

Ruiz, D. V., Krinski, B. A. e Todt, E. (2019). ANDA: A Novel Data Augmentation Technique Applied to Salient Object Detection. https://doi.org/10.48550/arXiv.1910.01256. Acessado em: 04/04/2023.

Ruiz, D. V., Krinski, B. A. e Todt, E. (2020a). IDA: Improved Data Augmentation Applied to Salient Object Detection. Em *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. IEEE. https://doi.org/10.48550/arXiv.2009.08845. Acessado em: 07/04/2023.

Ruiz, D. V., Salomon, G. e Todt, E. (2020b). Can Giraffes Become Birds? An Evaluation of Image-to-image Translation for Data Generation. https://doi.org/10.48550/arXiv.2001.03637. Acessado em: 08/04/2023.

Salahuddin, Z., Woodruff, H. C., Chatterjee, A. e Lambin, P. (2022). Transparency of deep neural networks for medical image analysis: A review of interpretability methods. *Computers in Biology and Medicine*. https://doi.org/10.1016/j.compbiomed.2021.105111. Acessado em: 08/02/2023.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A. e Chen, X. (2016). Improved Techniques for Training GANs. https://doi.org/10.48550/arXiv.1606.03498. Acessado em: 27/04/2023.

Santos, M. d. O., Lima, F. C. d. S. d., Martins, L. F. L., Oliveira, J. F. P., Almeida, L. M. d. e Cancela, M. d. C. (2023). Estimativa de incidência de câncer no brasil, 2023-2025. *Revista Brasileira de Cancerologia*, 69(1):e–213700. https://doi.org/10.32635/2176-9745.RBC.2023v69n1.3700. Acessado em: 09/05/2023.

Schutte, K., Moindrot, O., Hérent, P., Schiratti, J.-B. e Jégou, S. (2021). Using StyleGAN for Visual Interpretability of Deep Learning Models on Medical Images. *arXiv*. https://doi.org/10.48550/arxiv.2101.07563. Acessado em: 08/02/2023.

Shylasree, T. S., Mahajan, D., Chaturvedi, A., Menon, S., Gupta, S., Thakur, M., Poddar, P. e Maheshwari, A. (2024). Clinicopathological and oncological outcomes of borderline mucinous tumours of ovary: a large case series. *Indian Journal of Surgical Oncology*, 15.

Signaevsky, M., Prastawa, M., Farrell, K., Tabish, N., Baldwin, E., Han, N., Iida, M. A., Koll, J., Bryce, C., Purohit, D., Haroutunian, V., McKee, A. C., Stein, T. D., III, C. L. W., Walker, J., Richardson, T. E., Hanson, R., Donovan, M. J., Cordon-Cardo, C., Zeineh, J., Fernandez, G. e Crary, J. F. (2019). Artificial intelligence in neuropathology: deep learning-based assessment of tauopathy. *Laboratory Investigation volume 99, 1019–1029 (2019)*. https://doi.org/10.1038/s41374-019-0202-4. Acessado em: 08/02/2023.

Sokolova, M. e Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437. https://doi.org/10.1016/j.ipm.2009.03.002. Acessado em: 25/04/2023.

Sonka, M., Hlavac, V. e Boyle, R. (1993). *Image pre-processing*, páginas 56–111. Springer US, Boston, MA.

Spanhol, F. A., Oliveira, L. S., Petitjean, C. e Heutte, L. (2016a). Breast cancer histopathological image classification using convolutional neural networks. Em *2016 International Joint Conference on Neural Networks (IJCNN)*, páginas 2560–2567.

Spanhol, F. A., Oliveira, L. S., Petitjean, C. e Heutte, L. (2016b). A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering*, 63(7):1455–1462. https://doi.org/10.1109/TBME.2015.2496264. Acessado em: 30/04/2023.

Srivastava, R. K., Greff, K. e Schmidhuber, J. (2015). Training very deep networks. Em *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, página 2377–2385, Cambridge, MA, USA. MIT Press.

Stephan, P. e Nelson, D. A. (2021). Hormone Receptor Status in Breast Cancer. https://www.verywellhealth.com/hormone-receptor-status-and-diagnosis-430106. Acessado em: 11/03/2023.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. e Rabinovich, A. (2015). Going deeper with convolutions. Em *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, páginas 1–9.

Tamaki, J. V. K., Rogalsky, J. E., Ioshii, S. O. e de Oliveira, L. F. (2021). Método de delimitação celular em imagens ER/PR de câncer de mama. *XXI Simpósio Brasileiro de Computação Aplicada à Saúde*. https://doi.org/10.5753/sbcas.2021.16089. Acessado em: 08/02/2023.

Tan, P.-N., Steinbach, M. e Kumar, V. (2005). *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., USA.

Tang, Z., Chuang, K. V., DeCarli, C., Jin, L.-W., Beckett, L., Keiser, M. J. e Dugger, B. N. (2019). Interpretable classification of Alzheimer's disease pathologies with a convolutional neural network pipeline. *Nature Communications 10*. https://doi.org/10.1038/s41467-019-10212-1. Acessado em: 08/02/2023.

Thölke, P., Mantilla-Ramos, Y.-J., Abdelhedi, H., Maschke, C., Dehgan, A., Harel, Y., Kemtur, A., Mekki Berrada, L., Sahraoui, M., Young, T., Bellemare Pépin, A., El Khantour, C., Landry, M., Pascarella, A., Hadid, V., Combrisson, E., O'Byrne, J. e Jerbi, K. (2023). Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data. *NeuroImage*, 277:120253.

Vasuki, P., Kanimozhi, J. e Devi, M. B. (2017). A survey on image preprocessing techniques for diverse fields of medical imagery. Em *2017 IEEE International Conference on Electrical, Instrumentation and Communication Engineering (ICEICE)*, páginas 1–6. https://doi.org/10.1109/ICEICE.2017.8192443, Acessado em: 11/05/2023.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u. e Polosukhin, I. (2017). Attention is all you need. Em Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. e Garnett, R., editores, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Vujovic, Z. (2021). Classification Model Evaluation Metrics. *International Journal of Advanced Computer Science and Applications*, Volume 12:599–606. http://dx.doi.org/10.14569/IJACSA.2021.0120670. Acessado em: 25/04/2023.

Wang, H., Zhu, H., Ding, L. e Yang, K. (2023). A diagnostic classification of lung nodules using multiple-scale residual network. *Scientific reports*, 13. https://doi.org/10.1038/s41598-023-38350-z. Acessado em: 30/04/2023.

Wang, Z., Bovik, A., Sheikh, H. e Simoncelli, E. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612. https://doi.org/10.1109/TIP.2003.819861. Acessado em: 27/04/2023.

Weidner, N., Cote, R. J., Suster, S. e Weiss, L. M. (2009). *Modern Surgical Pathology*. E-book. Elsevier Health Sciences. Acessado em: 16/03/2023.

WHO (2021). Breast cancer. https://www.who.int/news-room/fact-sheets/detail/breast-cancer. Acessado em: 09/05/2023.

WHO (2022). Cancer. https://www.who.int/news-room/fact-sheets/detail/cancer. Acessado em: 09/05/2023.

WHO (2023). Cancer. https://www.who.int/health-topics/cancer#tab=tab_1. Acessado em: 11/03/2023.

Yadav, G., Maheshwari, S. e Agarwal, A. (2014). Contrast limited adaptive histogram equalization based enhancement for real time video system. Em *2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, páginas 2392–2397.

Yang, S., Xiao, W., Zhang, M., Guo, S., Zhao, J. e Shen, F. (2022). Image Data Augmentation for Deep Learning: A Survey. https://doi.org/10.48550/arXiv.2204.08610. Acessado em: 09/04/2023.

Yip, C.-H. e Rhodes, A. (2014). Estrogen and progesterone receptors in breast cancer. *Future Oncology, 10(14), 2293-2301*. https://doi.org/10.2217/fon.14.110. Acessado em: 11/03/2023.

Yu, S., Zhang, S., Wang, B., Dun, H., Xu, L., Huang, X., Shi, E. e Feng, X. (2021). Generative adversarial network based data augmentation to improve cervical cell classification model. *Mathematical Biosciences and Engineering*, 18:1740–1753.

Zaitoun, N. M. e Aqel, M. J. (2015). Survey on Image Segmentation Techniques. *Procedia Computer Science*, 65:797–806. https://doi.org/10.1016/j.procs.2015.09.027. Acessado em: 23/03/2023.