UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Rafael Henrique Guimarães

Inteligência Artificial Explicável: Classificação de Lesões da Cavidade Oral com Modelos de Aprendizagem Profunda

Uberlândia, Brasil

2025

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Rafael Henrique Guimarães

Inteligência Artificial Explicável: Classificação de Lesões da Cavidade Oral com Modelos de Aprendizagem Profunda

Trabalho de conclusão de curso apresentado à Faculdade de Computação da Universidade Federal de Uberlândia, como parte dos requisitos exigidos para a obtenção título de Bacharel em Ciência da Computação.

Orientador: Professor Marcelo Zanchetta do Nascimento

Universidade Federal de Uberlândia – UFU

Faculdade de Computação

Bacharelado em Ciência da Computação

Uberlândia, Brasil

2025

Rafael Henrique Guimarães

Inteligência Artificial Explicável: Classificação de Lesões da Cavidade Oral com Modelos de Aprendizagem Profunda

Trabalho de conclusão de curso apresentado à Faculdade de Computação da Universidade Federal de Uberlândia, como parte dos requisitos exigidos para a obtenção título de Bacharel em Ciência da Computação.

Professor Marcelo Zanchetta do Nascimento
Orientador

Adriano Barbosa Silva

Thiago Pirola Ribeiro

Uberlândia, Brasil
2025

Resumo

A classificação histológica automatizada de displasias orais representa um desafio significativo na patologia digital, demandando arquiteturas de aprendizado profundo que combinem alto desempenho com interpretabilidade clínica. Este trabalho teve como objetivo investigar e comparar a aplicação de duas arquiteturas modernas de redes neurais, *Vision Transformer* (ViT-Base16) e ResNeSt-50, para classificação automatizada de imagens histológicas de displasia oral, com foco em explicabilidade através de técnicas de Inteligência Artificial Explicável (do inglês, eXplainable Artificial Intelligence - XAI). A metodologia experimental utilizou dois bancos de imagens histológicas de tecidos orais murinos, sendo o primeiro constituído por quatro classes: saudável, displasia leve, moderada e severa e o respectivo com duas subfamílias representando duas variantes genéticas - *wild-type (WT)* e *knockout (KO)*, ambos com 7 classes histopatológicas. Foram conduzidos experimentos inicias para estabelecimento de *baseline*, seguidos de otimizações através de aumento de dados especializado para histologia (9 técnicas), regularização avançada e ajuste de hiperparâmetros. Para análise de explicabilidade, implementaram-se seis técnicas XAI distintas: três específicas para ViT (Grad-CAM tradicional, mapas de atenção nativos e Grad-CAM melhorado) e três para ResNeSt (Grad-CAM especializado, atenção CNN e Grad-CAM de entrada). Os resultados dos experimentos de *baseline* revelaram superioridade inicial do ResNeSt (72,41% vs 64,00% acurácia média), porém as otimizações resultaram em inversão completa desta hierarquia, com o ViT superando o ResNeSt (80,79% vs 78,63%). O ViT demonstrou maior responsividade às técnicas de otimização (+16,79% ganho médio vs +6,22% do ResNeSt), alcançando melhor resultado de 87,50% de acurácia no banco com 4 classes. As técnicas de otimização proporcionaram ganhos substanciais em desempenho médio (+11,51%), redução significativa de *overfitting* (*gap* treino-validação de 37,18% para 19,82%) e excelente concordância validação-teste (R^2 = 0,9988). A análise XAI revelou evolução temporal consistente das explicações, correspondência entre atenção dos modelos e características histopatológicas relevantes, com o ViT apresentando foco global enquanto o ResNeSt privilegiou detalhes locais. A validação clínica das explicações demonstrou identificação correta de regiões de invasão estromal em carcinomas, alterações epiteliais em displasias e espessamento tecidual em hiperplasias. Conclui-se que o ViT, quando adequadamente otimizado, supera arquiteturas convolucionais especializadas em classificação histológica, oferecendo simultaneamente alto desempenho e interpretabilidade clínica, estabelecendo fundamentos para translação de inteligência artificial explicável para aplicações clínicas em patologia oral.

Palavras-chave: *Vision Transformer*, ResNeSt, Classificação Histológica, Displasia Oral, Inteligência Artificial Explicável.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

4NQO	*4-nitroquinoline-N-oxide*
CEC	Carcinoma Espinocelular
CNN	*Convolutional Neural Networks*
CUDA	*Compute Unified Device Architecture*
FFN	*Feed-Forward Networks*
GIF	*Graphics Interchange Format*
GPU	*Graphics Processing Unit*
HE	Hematoxilina e Eosina
IA	Inteligência Artificial
INCA	Instituto Nacional de Câncer
KO	*Knockout*
MHSA	*Multi-Head Self-Attention*
ML	*Machine Learning*
MLP	*Multi-Layer Perceptron*
PLN	Processamento de Linguagem Natural
R-CNN	*Region-based Convolutional Neural Networks*
ResNeSt	*Split-Attention Networks*
RNA	Rede Neural Artificial
ROI	Região de Interesse
ViT	*Vision Transformer*
WT	*Wild-Type*
XAI	*eXplainable Artificial Intelligence*

Sumário

1 Introdução

A classificação precisa de lesões histológicas da cavidade oral representa um desafio fundamental para o diagnóstico precoce e tratamento adequado de diversas patologias bucais. O câncer bucal figura entre os principais problemas de saúde pública mundial, sendo que no Brasil, segundo o Instituto Nascional de Câncer (INCA), são estimados 704 mil novos casos de câncer para cada ano do triênio 2023-2025. Especificamente em relação ao câncer da cavidade oral, foram registradas 10.900 ocorrências em 2023, das quais 4.878 resultaram em óbito, representando uma taxa de mortalidade de 44,75%. A detecção precoce de condições como displasia epitelial oral, leucoplasia e outras lesões pré-malignas (AMARAL et al., 2022) é fundamental para melhorar o prognóstico, considerando especialmente a influência conhecida do tabagismo e alcoolismo no desenvolvimento e mortalidade por câncer bucal.

Para enfrentar esses desafios, sistemas de apoio ao diagnóstico assistido por computador (do inglês, *Computer Aided Diagnosis - CAD*) emergem como ferramentas promissoras na análise de imagens histológicas. Esses sistemas visam aumentar a precisão diagnóstica, reduzir a subjetividade na interpretação de imagens e acelerar o processo de análise. A importância da classificação precisa de displasias e outras lesões da cavidade oral é evidenciada em trabalhos como o de (CARVALHO et al., 2024), que propõe métodos baseados em descritores fractais e *ensemble learning* para classificação de displasia da cavidade oral, e o estudo de (SILVA et al., 2022), que apresenta análise computacional de imagens histológicas de tecidos com displasia epitelial oral corados com hematoxilina e eosina. Trabalhos adicionais, como os de (SILVA et al., 2019) e (CARVALHO et al., 2023), demonstram a aplicabilidade de diferentes abordagens computacionais na classificação de tecidos epiteliais, utilizando desde métodos baseados em R-CNN até modelos em cascata combinando CNNs e análise fractal.

No contexto da inteligência artificial, os modelos de aprendizagem profunda têm se destacado como ferramentas eficazes para análise de imagens médicas. Dentro dessa categoria, existem duas abordagens principais que capturam informações de naturezas distintas: os modelos baseados em Redes Neurais Convolucionais (do inglês, *Convolutional Neural Network* - CNN), que se concentram em características locais e hierárquicas da imagem, e os modelos *Vision Transformer* (ViT), que utilizam mecanismos de atenção para capturar relações globais entre diferentes regiões da imagem. No entanto, cada abordagem apresenta desafios específicos para aplicação em histopatologia. Os modelos CNN, embora eficazes na extração de características locais detalhadas, podem ter limitações na captura de relações espaciais de longo alcance, essenciais para compreender padrões arquiteturais complexos em tecidos. Por outro lado, os modelos ViT, apesar de sua capacidade

superior de modelar dependências globais, tradicionalmente requerem grandes volumes de dados para treinamento eficaz e podem apresentar menor sensibilidade a detalhes finos cruciais para diagnóstico histopatológico. Adicionalmente, ambas as arquiteturas enfrentam o desafio fundamental da interpretabilidade, onde suas decisões permanecem opacas para especialistas clínicos, limitando sua adoção na prática médica.

1.1 Justificativa

Apesar dos avanços significativos, importantes lacunas permanecem em aberto na literatura. Primeiramente, a aplicação de modelos ViT para classificação de lesões histológicas da cavidade oral ainda é pouco explorada, representando uma oportunidade de investigação científica relevante. Além disso, uma limitação crítica dos modelos de aprendizagem profunda é sua natureza de "caixa preta", onde as decisões tomadas pelos algoritmos não são facilmente interpretáveis por especialistas médicos. Nesse contexto, a Inteligência Artificial Explicável (do inglês, *eXplainable Artificial Intelligence* - XAI) surge como uma área fundamental para tornar os modelos mais transparentes e confiáveis para aplicação clínica.

A relevância deste trabalho decorre de três lacunas na literatura: a ausência de estudos aplicando modelos ViT na classificação de lesões histológicas da cavidade oral, a escassez de técnicas de explicabilidade voltadas a esses modelos no contexto histopatológico, e a falta de comparações sistemáticas entre arquiteturas baseadas em convoluções e atenção. Como contribuição, este estudo propõe o uso de modelos ViT aliados a técnicas de XAI, buscando conciliar elevada acurácia com interpretabilidade clínica — aspectos fundamentais para a adoção de soluções de inteligência artificial na prática em patologia oral.

1.2 Objetivos

O objetivo geral deste trabalho é desenvolver e avaliar um sistema de classificação de lesões histológicas da cavidade oral utilizando modelos de aprendizagem profunda, com foco na aplicação de técnicas de inteligência artificial explicável para interpretar as decisões dos modelos e fornecer insights clinicamente relevantes. Pontualmente, os objetivos específicos deste trabalho foram:

- Estudar e implementar modelos ViT para classificação de imagens histológicas da cavidade oral;
- Estudar e implementar modelos baseados em CNNs, especificamente a arquitetura ResNeSt, para comparação de desempenho;

- Avaliar diferentes técnicas de aumento de dados específicas para imagens histológicas;
- Aplicar técnicas de XAI para interpretar as decisões dos modelos, incluindo métodos como Grad-CAM, mapas de atenção e visualizações de características;
- Comparar o desempenho e interpretabilidade dos diferentes modelos na classificação de lesões histológicas;
- Validar a aplicabilidade clínica das técnicas de XAI propostas através da análise das regiões de interesse identificadas pelos modelos.

1.3 Organização da Monografia

Esta monografia está organizada em cinco capítulos, estruturados da seguinte forma:

- Capítulo 1 - Introdução: Apresenta o contexto do problema, a motivação para o desenvolvimento do trabalho, as lacunas identificadas na literatura, os objetivos gerais e específicos, e a organização do documento.
- Capítulo 2 - Fundamentação Teórica: Aborda os conceitos fundamentais necessários para a compreensão do trabalho, incluindo aspectos sobre câncer bucal, histologia, displasia, aprendizagem de máquina, redes neurais convolucionais, arquiteturas Transformer e Vision Transformer, e técnicas de inteligência artificial explicável.
- Capítulo 4 - Metodologia: Descreve detalhadamente a metodologia empregada no desenvolvimento do trabalho, incluindo a descrição do dataset utilizado, as arquiteturas de modelos implementadas, as técnicas de pré-processamento e aumento de dados, os métodos de treinamento e validação, e as métricas de avaliação empregadas.
- Capítulo 5 - Resultados: Apresenta os resultados obtidos nos experimentos realizados, incluindo análises comparativas entre os diferentes modelos, avaliação das técnicas de XAI, e discussão dos resultados no contexto clínico.
- Capítulo 6 - Conclusões: Sintetiza as principais contribuições do trabalho, discute as limitações encontradas, e propõe direções para trabalhos futuros na área.

2 Fundamentação Teórica

2.1 Câncer Bucal

O câncer bucal é uma condição maligna que pode ser desenvolvida nos tecidos da boca, constituídos por língua, bochechas, céu da boca, lábios e gengivas. A incidência de câncer bucal no Brasil é considerada uma das maiores do planeta, visto que ocupa hoje a 5ª posição de Câncer mais comum em homens e a 11ª posição de Câncer mais comum em mulheres, segundo o INCA, o que torna um problema de saúde pública. Este tipo de lesão desenvolve-se a partir de células-tronco escamosas. que é classificada como carcinoma de células escamosas. O Carcinoma Espinocelular Cutâneo (CEC) é o tipo mais frequente, presente em cerca de 90% das neoplasias malignais bucais. Os autores em (SCHMIDT et al., 2004), o tabagismo é um forte candidato ao desenvolvimento da CEC, em lesões envolvendo o assoalho da boca e a língua posterolateral. Em (LOCKHART; JR; PULLIAM, 1998), os autores também relacionam o uso frequente de álcool com o desenvolvimento dos carcinomas. Em países que têm a forte influência e hábito de tabagismo, como por exemplo Rússia, Bangladesh e a China, mostram altos índices e fatores de correlação com o índice de Câncer Bucal, como citado e evidenciado no trabalho de (ZHANG; XIE; SHANG, 2022).

A compreensão da anatomia da cavidade oral é fundamental para o entendimento dos padrões de desenvolvimento e distribuição dessas lesões. A cavidade oral representa uma região anatômica complexa, delimitada anteriormente pela borda vermelha dos lábios e posteriormente pelas papilas circunvaladas da base da língua, com extensão superior até a transição do palato duro com o palato mole. A distribuição topográfica das estruturas orais permite a identificação de regiões anatômicas distintas, cada uma apresentando características histológicas específicas e diferentes susceptibilidades ao desenvolvimento de neoplasias. Na Figura 1 é ilustrado a organização espacial dessas estruturas, incluindo os lábios, diferentes segmentos da língua oral, o assoalho bucal, a mucosa jugal, as estruturas gengivais superior e inferior, o trígono retromolar e o palato duro (MONTERO; PATEL, 2015). Essa característica anatômica é clinicamente relevante, uma vez que diferentes regiões apresentam variações na espessura epitelial, densidade de glândulas salivares menores, padrões de vascularização e exposição a agentes carcinogênicos, fatores que influenciam diretamente os mecanismos de carcinogênese e os padrões de progressão tumoral.

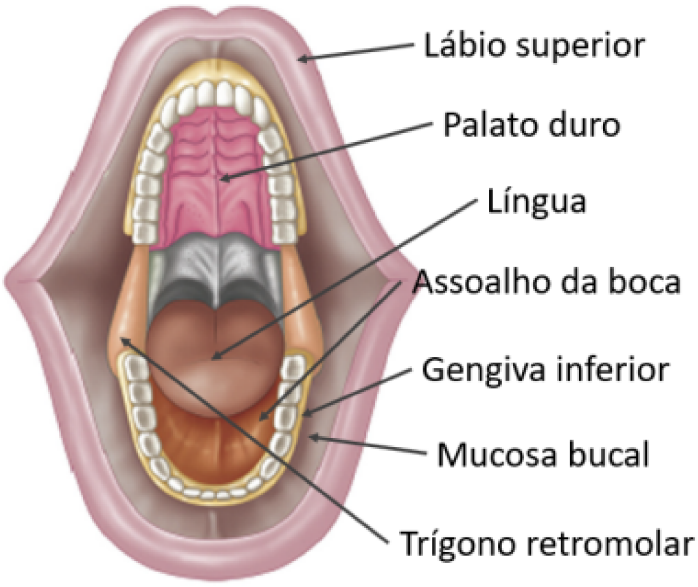


Figura 1 – Morfologia da Boca. Adaptado de: (CARVALHO et al., 2024).

O desenvolvimento de neoplasias na cavidade oral segue uma evolução patológica que se inicia com alterações celulares sutis e pode progredir até formas invasivas de carcinoma. Esse espectro inclui desde alterações reativas benignas até lesões com alto potencial de malignização, requerendo abordagens diagnósticas e terapêuticas diferenciadas. As lesões com potencial de malignização representam uma categoria de particular interesse clínico, caracterizadas por alterações arquiteturais e citológicas que predispõem à transformação neoplásica. Estas condições incluem leucoplasias, eritroplasias, liquen plano oral erosivo e fibrose submucosa, cada uma apresentando características morfológicas distintivas e diferentes graus de risco para progressão maligna.

Um aspecto fundamental dessas lesões é sua apresentação morfológica frequentemente similar ao tecido normal, especialmente em estágios iniciais de desenvolvimento. Esta similaridade fenotípica constitui um dos principais desafios para o diagnóstico diferencial, exigindo análise histopatológica detalhada e, frequentemente, o emprego de marcadores moleculares complementares para caracterização adequada. A progressão temporal dessas lesões geralmente ocorre de forma lenta e insidiosa, podendo estender-se por anos antes que alterações morfológicas significativas se tornem evidentes. Entretanto, a presença de cofatores como exposição crônica ao tabaco, consumo excessivo de álcool, infecções virais persistentes ou predisposição genética pode acelerar consideravelmente este processo de transformação, (TOVARU et al., 2023).

A detecção precoce de lesões malignas e pré-malignas da cavidade oral representa um dos principais fatores prognósticos para o sucesso terapêutico. Dados epidemiológicos demonstram que lesões diagnosticadas em estágios iniciais apresentam taxas de sobrevivência significativamente superiores, enfatizando a importância crítica do diagnóstico

precoce (COSTA; MIGLIORATI, 2001). Tradicionalmente, o diagnóstico de lesões orais baseia-se em exame clínico seguido de confirmação histopatológica através de biópsia. Entretanto, este processo apresenta limitações inerentes, incluindo a subjetividade na interpretação morfológica, variabilidade inter-observador e a necessidade de procedimentos invasivos para obtenção de material diagnóstico. A classificação histopatológica da displasia epitelial oral exemplifica essas limitações diagnósticas, uma vez que a ausência de critérios morfológicos universalmente padronizados resulta em considerável variabilidade interpretativa entre patologistas, comprometendo a reprodutibilidade diagnóstica e, consequentemente, a adequação das condutas terapêuticas. Neste contexto, o desenvolvimento de ferramentas tecnológicas baseadas em inteligência artificial emerge como uma estratégia promissora para superação dessas limitações, oferecendo maior objetividade, reprodutibilidade e rapidez diagnóstica.

2.2 Histologia

A histologia é a ciência que estuda os tecidos do corpo e como esses tecidos se organizam para compor os órgãos. É com a histologia que se tem uma base para a formação dos órgãos, proporcionando uma base sólida para a prática clínica e para o avanço da medicina, como afirmado por (WEINMANN, 1942) em seu trabalho sobre a importância da histologia.

As imagens histológicas surgem como representações visuais de tecidos, que possibilitam detalhamento em relação a estrutura das células e do tecido, que podem ser observadas e capturadas via um microscópio. A partir dessas imagens, patologistas podem analisar as fotos obtidas e detectar diferentes padrões de crescimento e realizar diagnósticos de algumas doenças ou lesões.

O procedimento para obtenção de imagens histológicas envolve uma sequência de etapas padronizadas que transformam amostras de tecido biológico em lâminas adequadas para análise microscópica. Este processo é fundamental para garantir que as características celulares e teciduais sejam preservadas de forma fidedigna para posterior análise diagnóstica (SANTOS et al., 2022). Na Figura 2 é ilustrado as principais etapas envolvidas na preparação de lâminas histológicas, desde a coleta da amostra até a obtenção da imagem digital final.

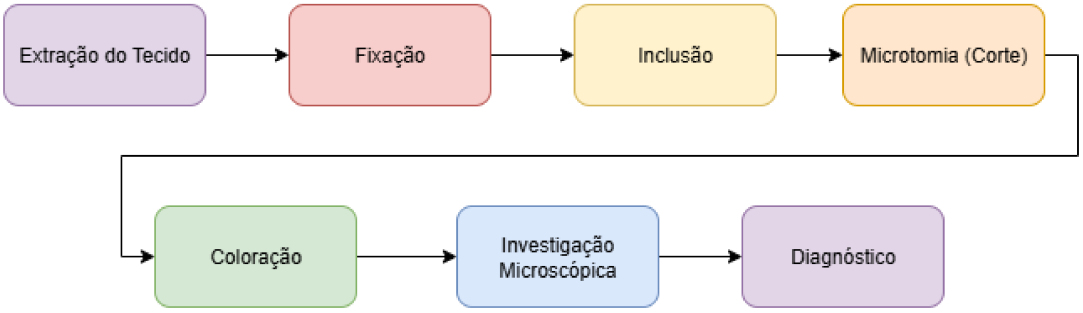


Figura 2 – Fluxograma do processo de preparação de lâminas histológicas. Adaptado de: (SANTOS et al., 2022)

O processo inicia-se com a extração do tecido (biópsia), onde é realizada a coleta do material biológico de interesse. Na sequência, ocorre a fixação, etapa crucial onde as amostras de tecido extraídas são estabilizadas para preservar a estrutura celular e evitar decomposição. As amostras são imersas em soluções químicas fixadoras, sendo a formalina a mais amplamente utilizada para tecidos de rotina diagnóstica.

A terceira etapa consiste na inclusão, processo onde o tecido é imerso em parafina derretida após desidratação. Esta etapa substitui a água por parafina ou resina, conferindo consistência rígida às amostras e facilitando o corte em seções ultrafinas. Os blocos de parafina resultantes são então submetidos à microtomia, utilizando um instrumento denominado micrótomo para obtenção de fatias ultrafinas de 3–10 µm de espessura.

A coloração representa uma etapa fundamental do processo, uma vez que células e matriz extracelular são naturalmente translúcidas após o processamento. A combinação Hematoxilina e Eosina (H&E) constitui o método de coloração mais amplamente empregado, onde a hematoxilina destaca os núcleos celulares em azul-escuro, enquanto a eosina pigmenta citoplasma e outras estruturas celulares em tonalidades rosadas.

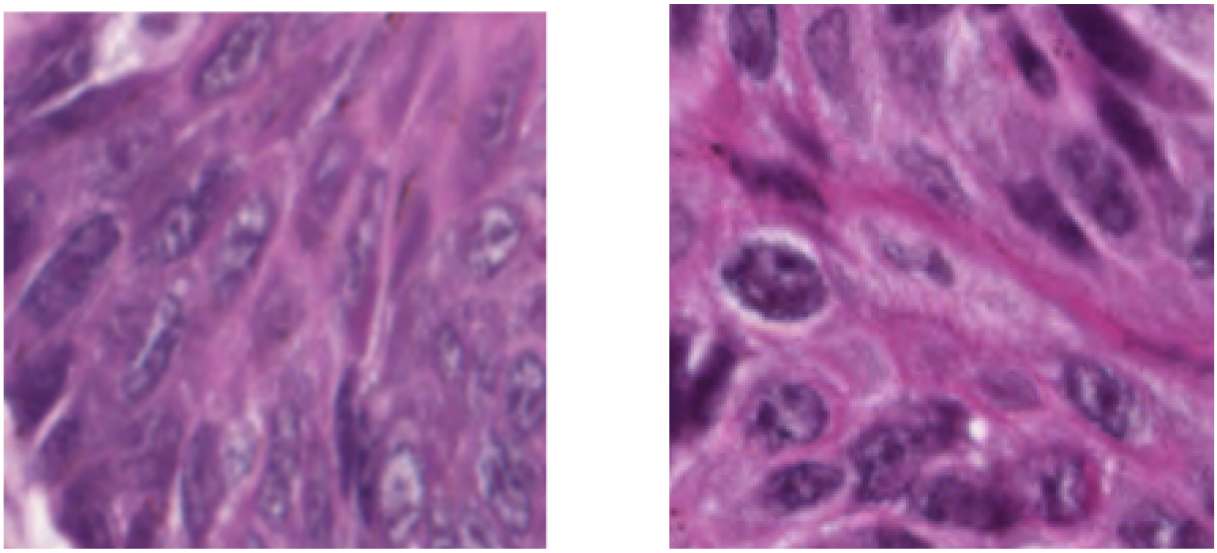
A sexta etapa envolve a investigação microscópica, onde é realizada a análise direta da lâmina corada em microscópio óptico. Finalmente, a sétima etapa corresponde ao diagnóstico, que consiste na interpretação do patologista com base nas observações microscópicas realizadas. No contexto da análise computacional, uma etapa adicional de digitalização das lâminas permite a obtenção de imagens de alta resolução adequadas para aplicação de algoritmos de inteligência artificial.

2.3 Lesões da Cavidade Oral

As lesões da cavidade oral representam um espectro heterogêneo de alterações teciduais que compreendem desde condições benignas até neoplasias malignas invasivas. Para fins de organização didática e clínica, estas lesões podem ser classificadas conforme sua natureza biológica em três categorias principais: lesões benignas, lesões pré-malignas (displasias) e lesões malignas.

As lesões benignas caracterizam-se por crescimento limitado, ausência de capacidade invasiva e baixo risco de transformação maligna. A hiperplasia epitelial representa uma resposta proliferativa benigna com espessamento epitelial uniforme e preservação da arquitetura estratificada normal. O papiloma oral constitui outra proliferação benigna de origem viral, caracterizado por projeções papilíferas revestidas por epitélio estratificado pavimentoso.

Na Figura 3 são exemplificadas, imagens representativas de lesões benignas da cavidade oral, demonstrando as características morfológicas distintivas de hiperplasia epitelial na primeira e papiloma na segunda.



(a) Hiperplasia epitelial caracterizada por espessamento uniforme do epitélio sem alterações displásicas

(b) Papiloma oral apresentando projeções papilíferas com eixo conjuntivo central vascularizado

Figura 3 – Exemplos de lesões benignas da cavidade oral corados com (H&E).

As lesões pré-malignas, representadas pelas displasias epiteliais, constituem alterações proliferativas caracterizadas por distúrbios na maturação celular e perda variável da organização arquitetural do epitélio. Estas lesões apresentam potencial de transformação neoplásica, sendo classificadas em três graus de severidade: displasia leve (alterações limitadas ao terço basal do epitélio), displasia moderada (alterações até dois terços da espessura epitelial) e displasia intensa (alterações comprometendo mais de dois terços do epitélio). O risco de progressão maligna varia de 7-15% na displasia leve até 50% na displasia intensa (SPEIGHT, 2007).

Na Figura 4 é apresentado um exemplo representativo de displasia epitelial, demonstrando as alterações morfológicas características que definem esta categoria de lesões pré-malignas.

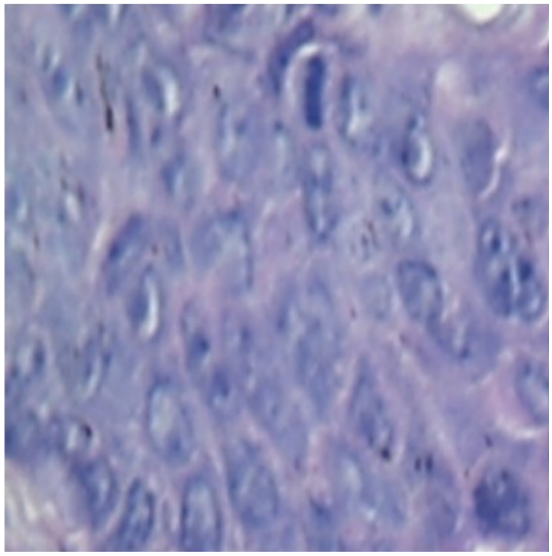


Figura 4 – Displasia epitelial mostrando desorganização arquitetural, pleomorfismo nuclear e perda de polaridade celular em coloração H&E.

As lesões malignas caracterizam-se pela capacidade de invasão local, destruição tecidual e potencial metastático. O CEC, apresentado na Figura 5 representa aproximadamente 90% dos casos de câncer bucal, caracterizando-se histologicamente pela invasão de células epiteliais atípicas através da membrana basal com acentuado pleomorfismo nuclear e formação de pérolas córneas (SCHMIDT et al., 2004).

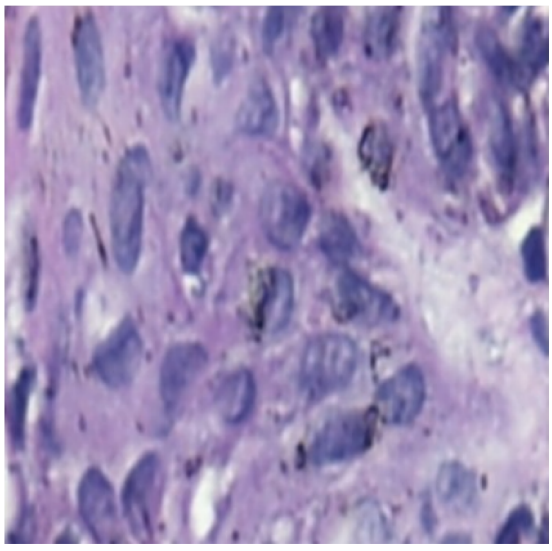


Figura 5 – Carcinoma espinocelular da cavidade oral mostrando invasão estromal por células epiteliais atípicas, pleomorfismo nuclear acentuado e formação de pérolas córneas.

A detecção precoce dessas lesões é fundamental, pois estudos demonstram que displasias epiteliais graves apresentam 16% de transformação maligna (SPEIGHT, 2007). Quando as lesões são detectadas em estágios iniciais, a chance de sobrevivência é significativamente alta, enfatizando a importância crítica do diagnóstico precoce. Segundo

(COSTA; MIGLIORATI, 2001), o tempo estimado para diagnóstico em 1998 era de aproximadamente 19 dias, sendo esperada uma melhoria significativa através da implementação de ferramentas tecnológicas baseadas em inteligência artificial.

O desenvolvimento de sistemas automatizados de classificação histológica representa uma estratégia promissora para superação das limitações diagnósticas tradicionais, incluindo a subjetividade interpretativa e a variabilidade inter-observador. Essas ferramentas podem oferecer maior rapidez, objetividade e reprodutibilidade diagnóstica, constituindo uma área de investigação de grande relevância clínica e científica para classificação automatizada das lesões orais.

2.4 Inteligência Artificial

A Inteligência Artificial (IA) constitui um campo multidisciplinar da ciência da computação que busca desenvolver sistemas capazes de realizar tarefas que tradicionalmente requerem inteligência humana, com taxonomia hierárquica representada pela Figura 6. Segundo (RUSSELL, 2010), a IA pode ser definida como o estudo de agentes inteligentes que percebem seu ambiente e tomam ações que maximizam suas chances de sucesso em alguma tarefa específica.

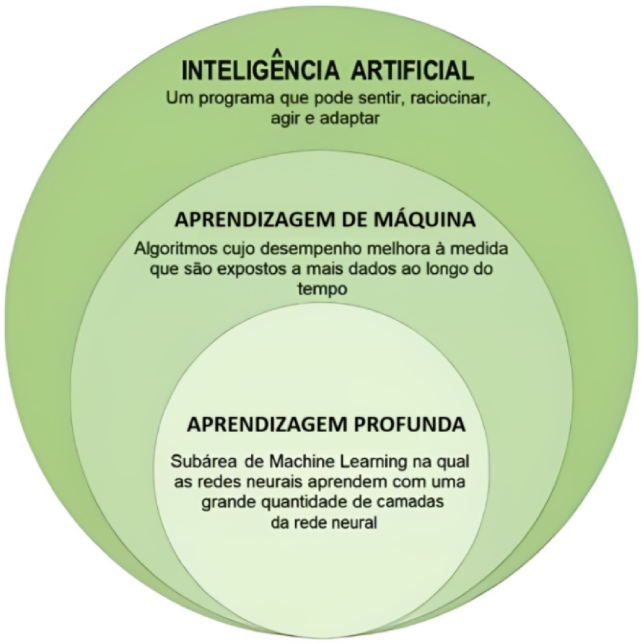


Figura 6 – Taxonomia hierárquica da IA, demonstrando a relação entre os três principais domínios: IA como o campo mais abrangente, *Machine Learning* como subconjunto de algoritmos que melhoram com dados, e *Deep Learning* como subárea que utiliza redes neurais com múltiplas camadas. Adaptado de: (ALZUBAIDI et al., 2021).

2.4.1 Aprendizagem de Máquina

A AM representa o conjunto de algoritmos cujo desempenho melhora à medida que são expostos a mais dados ao longo do tempo. Esta capacidade de aprendizado automático elimina a necessidade de programação explícita para cada tarefa específica, permitindo que os sistemas identifiquem padrões complexos e tomem decisões baseadas em dados.

Dentro da AM, os algoritmos podem ser categorizados em diferentes paradigmas de aprendizagem. A aprendizagem supervisionada utiliza conjuntos de dados rotulados, onde cada entrada possui uma saída conhecida correspondente, sendo particularmente adequada para tarefas de classificação e regressão. Este paradigma é amplamente aplicado em problemas de diagnóstico médico e análise de imagens histológicas, onde existe conhecimento prévio sobre as classes ou valores a serem preditos. Por outro lado, a aprendizagem não supervisionada trabalha com dados não rotulados, buscando descobrir padrões ocultos através de técnicas como agrupamento (*clustering*), redução de dimensionalidade e detecção de anomalias. Embora menos aplicado diretamente em tarefas de classificação histológica, este paradigma pode ser valioso para pré-processamento de dados e descoberta de estruturas latentes em conjuntos de dados médicos.

A Aprendizagem Profunda (AP) constitui uma subárea da AM na qual as redes neurais aprendem com uma grande quantidade de camadas da rede neural. Esta abordagem permite que os algoritmos extraiam automaticamente características complexas e hierárquicas dos dados brutos, eliminando a necessidade de engenharia manual de características que caracteriza métodos tradicionais de *Machine Learning*.

A principal vantagem da AP reside na capacidade de aprender representações em múltiplos níveis de abstração. As camadas iniciais capturam características de baixo nível (como bordas e texturas em imagens), enquanto camadas mais profundas combinam essas características para formar representações de alto nível (como formas complexas e padrões semânticos). Esta hierarquia de aprendizado tem demonstrado superioridade significativa em tarefas complexas de reconhecimento de padrões, especialmente em domínios como visão computacional e processamento de linguagem natural (LECUN; BENGIO; HINTON, 2015).

No contexto da análise de imagens histológicas, a AP oferece vantagens particulares devido à complexidade e variabilidade dos padrões morfológicos presentes nos tecidos biológicos. A capacidade de aprender automaticamente características discriminativas sem conhecimento prévio específico do domínio torna esta abordagem especialmente adequada para aplicações em patologia digital, onde a interpretação visual requer anos de treinamento especializado (LITJENS et al., 2017).

2.5 Redes Neurais Artificiais

As Redes Neurais Artificiais (RNAs) constituem modelos computacionais inspirados no funcionamento do sistema nervoso biológico, especificamente na forma como os neurônios processam e transmitem informações. Segundo (HAYKIN, 2009), uma RNA é um processador distribuído paralelo que possui uma propensão natural para armazenar conhecimento experimental e torná-lo disponível para uso.

A arquitetura fundamental de uma RNA é composta por três tipos principais de camadas: entrada, ocultas e saída. Na Figura 7 é ilustrado esta organização hierárquica, demonstrando o fluxo de informações desde a entrada até a geração da saída final.

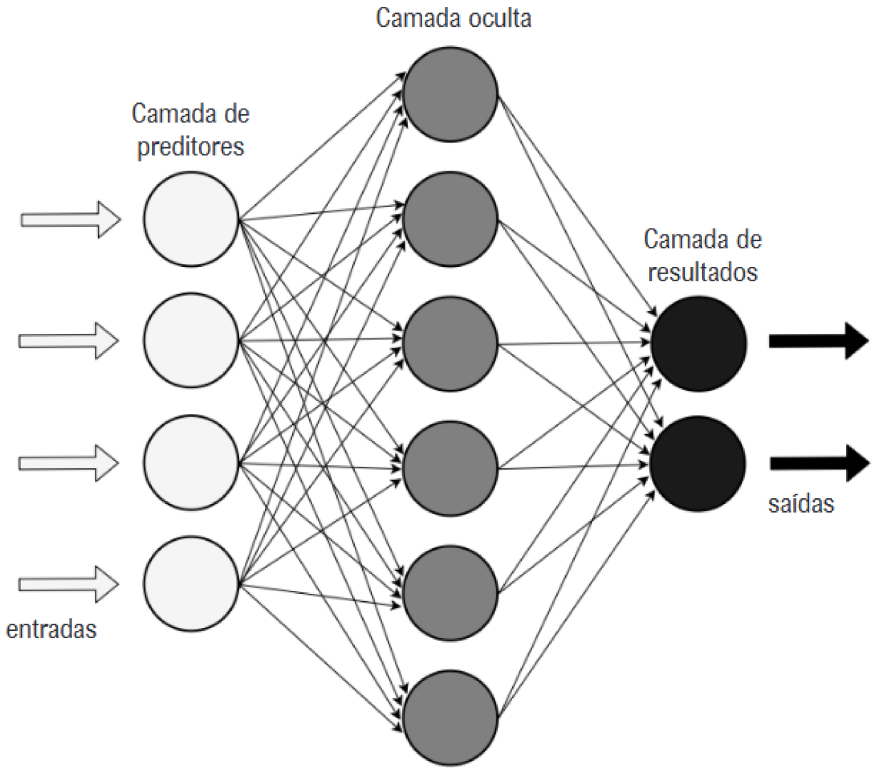


Figura 7 – Arquitetura básica de uma rede neural artificial, mostrando as camadas de entrada, camadas ocultas e camada de saída, com as conexões ponderadas entre neurônios adjacentes (PAIXÃO et al., 2022).

As camadas de entrada recebem os dados brutos do ambiente externo, funcionando como a interface entre o mundo real e o sistema de processamento interno. As camadas ocultas, como destacado por (GOODFELLOW et al., 2016), são responsáveis pela extração e transformação de características dos dados, permitindo que a rede aprenda representações complexas e não-lineares. Cada camada oculta adicional permite que a rede capture padrões de maior abstração, sendo este o princípio fundamental da aprendizagem profunda. Finalmente, as camadas de saída produzem a resposta final da rede, seja para tarefas de classificação, regressão ou outras aplicações específicas.

O processo de aprendizagem nas RNAs ocorre através do ajuste iterativo dos pesos das conexões entre neurônios, utilizando algoritmos de otimização como a retropropagação do erro (*backpropagation*). Este mecanismo permite que a rede minimize uma função de custo específica, adaptando-se progressivamente aos padrões presentes nos dados de treinamento (RUMELHART; HINTON; WILLIAMS, 1986).

2.5.1 Redes Neurais Convolucionais

As Redes Neurais Convolucionais (do inglês *Convolutional Neural Networks* - CNNs) representam uma classe especializada de redes neurais artificiais especificamente projetadas para o processamento eficiente de dados com estrutura topológica, como imagens digitais. Segundo (LECUN et al., 2002), as CNNs foram desenvolvidas para explorar as propriedades espaciais locais dos dados visuais, utilizando operações de convolução para extrair características hierárquicas.

A arquitetura das CNNs é fundamentada em três tipos principais de camadas: camadas convolucionais, camadas de *pooling* e camadas totalmente conectadas. Na Figura 8 é mostrada a organização típica desses componentes, demonstrando como as características são progressivamente extraídas e abstraídas ao longo da rede.

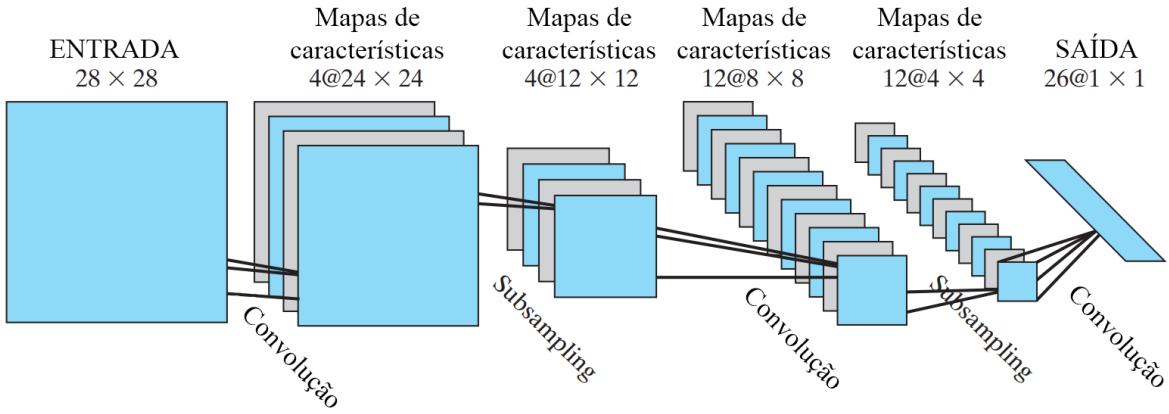


Figura 8 – Estrutura de uma rede convolucional. Imagem adaptada de Haykin (1999)

As camadas convolucionais aplicam filtros (kernels) deslizantes sobre a entrada, detectando características locais como bordas, texturas e padrões específicos. Cada filtro produz um mapa de características que destaca a presença de uma determinada característica espacial. As camadas de *pooling* subsequentes realizam uma redução controlada da dimensionalidade espacial, preservando as informações mais relevantes enquanto introduzem invariância translacional limitada (KRIZHEVSKY; SUTSKEVER; HINTON, 2017).

A eficácia das CNNs em tarefas de visão computacional tem sido demonstrada em diversas aplicações, desde reconhecimento de objetos até análise de imagens médi-

cas. Como observado por (LITJENS et al., 2017), arquiteturas como AlexNet, VGGNet, ResNet e suas variantes têm estabelecido marcos importantes em termos de precisão e eficiência computacional, tornando-se ferramentas fundamentais para análise automatizada de imagens histológicas.

2.5.2 Redes *Transformer*

As redes *Transformer*, são uma classe de arquitetura de redes neurais para aprendizagem profunda, introduzidas pela primeira vez em 2017, por (VASWANI et al., 2017). Segundo esse trabalho, os *Transformer*s, dispensam recorrências e convoluções inteiramente, isso pois a mesma, utiliza-se unicamente de um diferencial, chamado em pontos de atenção.

O mecanismo de atenção possibilita à RNA aprender relações e interconexões complexas entre os elementos de uma entrada, onde cada token em uma sequência tem a capacidade de "atentar" para tokens relevantes em diferentes posições da sequência de dados de entrada.

2.5.2.1 Mecanismo de Atenção Fundamental

O mecanismo de atenção baseia-se em três componentes principais: *Query* (Q), *Key* (K) e *Value* (V). Estas matrizes são obtidas através de transformações lineares da entrada, conforme a Equação 2.1:

$$\begin{aligned} Q &= XW_Q \\ K &= XW_K \\ V &= XW_V \end{aligned} \tag{2.1}$$

onde X representa a entrada e W_Q, W_K, W_V são matrizes de pesos aprendíveis para cada componente.

O mecanismo de atenção escalada (*Scaled Dot-Product Attention*) é então calculado através da Equação 2.2:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.2}$$

onde d_k representa a dimensionalidade das *keys*, e o fator de escala $\sqrt{d_k}$ previne que os produtos escalares cresçam excessivamente, mantendo os gradientes estáveis durante o treinamento.

Na Figura 9 é ilustrado o funcionamento do mecanismo de atenção escalada, demonstrando como as matrizes Q, K e V são combinadas para produzir os pesos de atenção e a saída final.

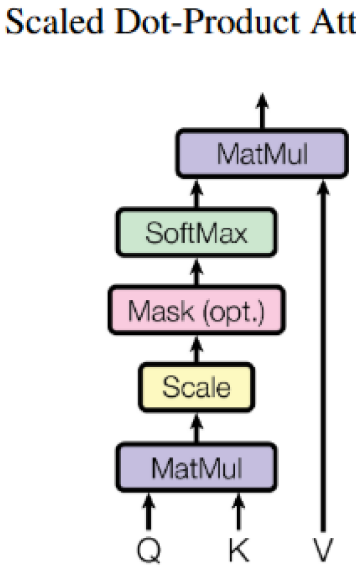


Figura 9 – Mecanismo de Atenção Escalada (*Scaled Dot-Product Attention*). As matrizes *Query* (Q) e *Key* (K) são multiplicadas e normalizadas pelo fator $\sqrt{d_k}$, seguido pela aplicação da função *softmax* para obter os pesos de atenção. Estes pesos são então aplicados à matriz *Value* (V) para produzir a saída final. Adaptado de (VASWANI et al., 2017).

2.5.2.2 Atenção Multi-Cabeça (*Multi-Head Attention*)

Para capturar diferentes tipos de relações simultaneamente, o mecanismo de atenção é estendido para múltiplas "cabeças" paralelas. A atenção multi-cabeça é definida pela Equação 2.3:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{2.3}$$

onde cada cabeça individual é calculada como 2.4:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{2.4}$$

e W_i^Q, W_i^K, W_i^V são matrizes de projeção específicas para a i-ésima cabeça, enquanto W^O é a matriz de projeção de saída que combina todas as cabeças.

Na Figura 10 é apresentado o funcionamento do mecanismo de atenção multi-cabeça, evidenciando como as transformações QKV são processadas em paralelo através de múltiplas cabeças para capturar diferentes tipos de dependências.

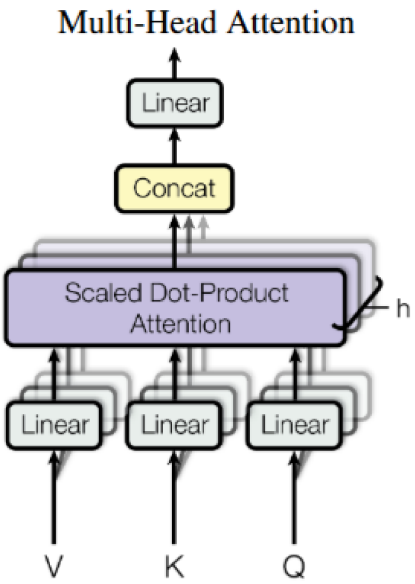


Figura 10 – Arquitetura do mecanismo de atenção multi-cabeça em redes *Transformer*. As matrizes *Query* (Q), *Key* (K) e *Value* (V) são processadas através de múltiplas cabeças de atenção em paralelo, cada uma capturando diferentes tipos de relações entre os elementos da sequência. As saídas das cabeças são concatenadas e projetadas para produzir a representação final. Adaptado de (VASWANI et al., 2017).

A arquitetura *Transformer* foi desenvolvida para a utilização e busca de êxito na área de processamento de linguagem natural. Como relatado por (VASWANI et al., 2017), as redes *Transformer* têm a habilidade de processar entradas grandes de dados de forma eficiente, explicado pelo fato de que esse tipo de rede é altamente capacitado para paralelização, em relação a outros modelos de aprendizagem de máquina.

O mecanismo de atenção multi-cabeça permite que diferentes cabeças se especializem em capturar distintos padrões de relacionamento: algumas podem focar em dependências sintáticas, outras em relações semânticas, e outras ainda em padrões posicionais específicos. Esta diversificação de representações é fundamental para o sucesso dos *Transformer*s em tarefas complexas de compreensão e geração de sequências.

Embora os *Transformers* tenham sido inicialmente propostos com enfoque no processamento de linguagem natural, seu uso na área de visão computacional representa um avanço tecnológico significativo. O trabalho realizado por (DUBEY; SINGH, 2023) demonstra a capacidade dessas redes na exploração eficiente de relações globais em cada camada, superando limitações das redes convolucionais tradicionais através da capacidade de capturar relações globais em espaços de características. É destacada também a importância das redes em impulsionar o progresso em diversas áreas da visão computacional, demonstrando oportunidades promissoras para aplicações futuras.

2.5.3 Vision *Transformer* (ViT)

O *Vision Transformer* (ViT) foi introduzido por (DOSOVITSKIY et al., 2020) como uma adaptação da arquitetura *Transformer*, originalmente proposta para Processamento de Linguagem Natural, para o domínio da visão computacional. O ViT utiliza o mecanismo de *self-attention* para capturar relações globais em imagens, dispensando totalmente o uso de convoluções.

A abordagem do ViT consiste em dividir a imagem de entrada em *patches* de tamanho fixo (por exemplo, 16×16 pixels), que são então linearizados e projetados para um espaço vetorial, formando uma sequência de *embeddings*. A cada *embedding* é somada uma codificação posicional (*positional embedding*), para preservar informações sobre a localização dos *patches* na imagem. Essa sequência é processada por um *Transformer Encoder*, composto por múltiplas camadas de *multi-head self-attention* e redes *feed-forward*, produzindo uma representação global da imagem.

O ViT se destaca por:

- Capturar dependências de longo alcance entre regiões da imagem.
- Ser altamente paralelizável durante o treinamento.
- Permitir dimensionamento do modelo de forma flexível, variando o número de camadas, cabeças de atenção e tamanho de *patches*.

Na Figura 11 é apresentada uma visão geral da arquitetura ViT, desde a divisão da imagem em *patches* até a etapa de classificação final.

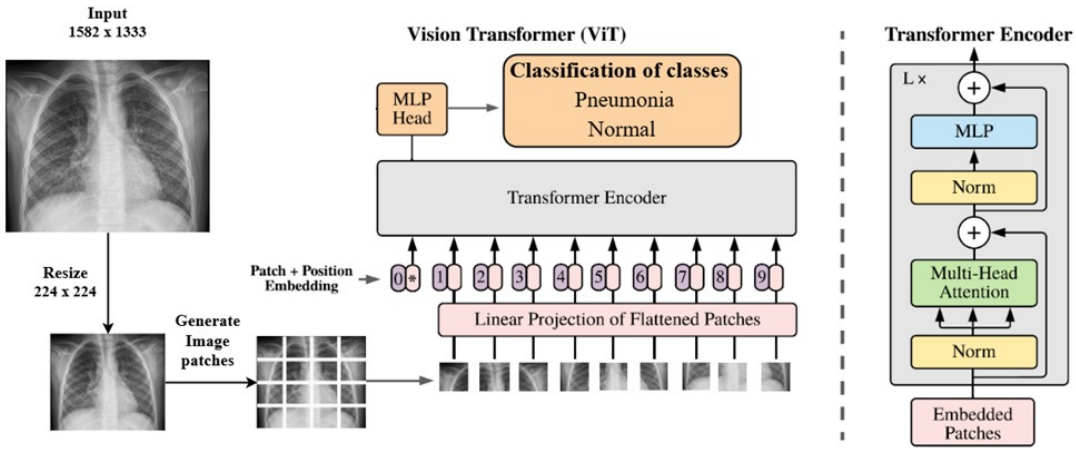


Figura 11 – Esquema da arquitetura Vision *Transformer* (adaptado de (DOSOVITSKIY et al., 2020)).

2.5.4 ResNeSt

O *ResNeSt* (*Split-Attention Networks*), proposto por (ZHANG et al., 2022), é uma arquitetura que combina a eficiência das redes residuais (*ResNet*) com um mecanismo de atenção especializado, chamado *Split-Attention*. Essa técnica aprimora a extração de características ao permitir que a rede processe e recalibre grupos de canais de forma independente.

A principal inovação do *ResNeSt* é o *Split-Attention Block*, no qual os mapas de características são divididos em subgrupos (*splits*). Cada grupo passa por uma operação de atenção de canal, atribuindo pesos adaptativos às suas características. Por fim, as saídas dos grupos são combinadas e passadas para o próximo bloco residual.

As vantagens dessa arquitetura incluem:

- Melhor modelagem de dependências entre canais.
- Manutenção das conexões residuais, facilitando o treinamento de redes profundas.
- Desempenho competitivo em tarefas de classificação, detecção e segmentação de imagens.

Na Figura 12 é exibido o funcionamento do *Split-Attention Block*, principal componente responsável pelo ganho de desempenho da arquitetura.

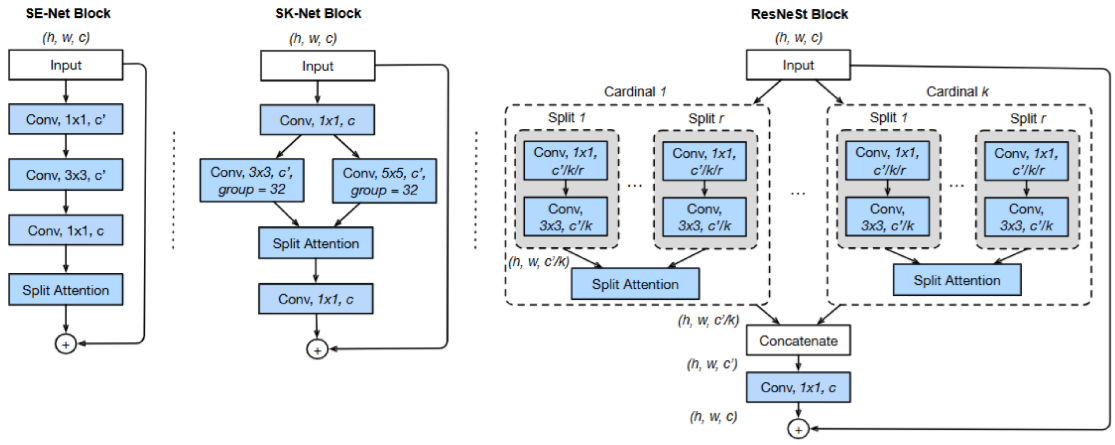


Figura 12 – Estrutura do bloco *Split-Attention* no ResNeSt - Adaptado de (ZHANG et al., 2022).

2.6 Técnicas de Aumento de Dados

O aumento de dados (*Data Augmentation*) representa uma abordagem fundamental para expandir conjuntos de dados limitados, problema recorrente no desenvolvimento

de modelos de aprendizado profundo. O trabalho de (SHORTEN; KHOSHGOFTAAR, 2019) demonstra que essa estratégia é particularmente eficaz para melhorar a capacidade de generalização em cenários com dados escassos. O método consiste em aplicar transformações controladas às imagens originais, gerando novas amostras que mantêm as características diagnósticas relevantes. Como observado por (PEREZ; WANG, 2017), essas transformações introduzem variabilidade suficiente para melhorar a robustez dos modelos sem comprometer a integridade semântica dos dados.

No contexto da análise de imagens histológicas, essa estratégia torna-se particularmente relevante devido às limitações práticas na obtenção de grandes volumes de dados anotados. Em (TELLEZ et al., 2019), os autores enfatizam que o processo de preparação histológica demanda tempo considerável, desde a coleta e fixação dos tecidos até a digitalização das lâminas. Além disso, Madabhushi e Lee (MADABHUSHI; LEE, 2016) destacam que a anotação confiável dessas imagens requer a expertise de patologistas qualificados, recurso frequentemente escasso e custoso na prática clínica.

Na Figura 13 são apresentadas quatro técnicas fundamentais de aumento de dados aplicadas a uma imagem histológica, demonstrando como transformações simples podem gerar variações significativas mantendo as características diagnósticas essenciais.

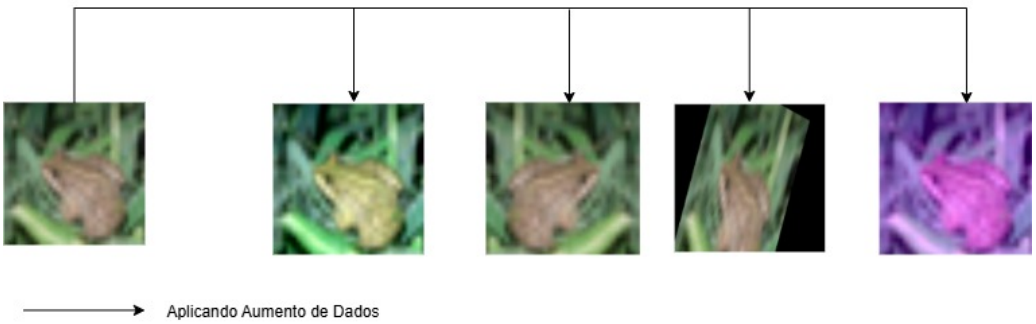


Figura 13 – Exemplos de técnicas de aumento de dados aplicadas em uma imagem do conjunto público de imagens CIFAR-10, da classe sapo. A partir de uma imagem original (centro superior), são aplicadas transformações que preservam as características morfológicas: rotação, espelhamento horizontal, ajuste cromático e recorte aleatório. Cada transformação gera uma nova amostra de treinamento sem alterar o conteúdo diagnóstico da imagem.

2.6.1 Desafios Específicos em Imagens Histológicas

O domínio histológico apresenta características peculiares que tornam o aumento de dados uma tarefa complexa. No trabalho de (JANOWCZYK; MADABHUSHI, 2016) é observado que as imagens histológicas frequentemente exibem alta variabilidade intra-classe, resultado de diferenças biológicas individuais, variações no protocolo de preparação das amostras e condições específicas de aquisição. Paradoxalmente, essas mesmas imagens podem apresentar baixa variabilidade inter-classe, especialmente quando se comparam

tecidos saudáveis com alterações patológicas iniciais.

Esta característica representa um desafio significativo para a classificação automatizada. Outros trabalhos, como (DIMITRIOU; ARANDJELOVIĆ; CAIE, 2019) demonstram que estágios precoces de progressão patológica podem ser morfologicamente muito similares ao tecido normal, exigindo análise minuciosa de detalhes sutis para diferenciação precisa.

Na Figura 14 é exemplificado este fenômeno, comparando uma amostra de tecido saudável com uma lesão em estágio inicial, evidenciando a similaridade morfológica que dificulta a distinção automatizada.

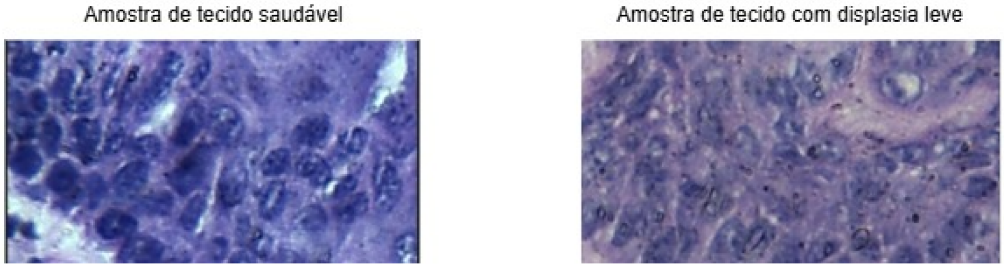


Figura 14 – Comparação entre tecido saudável e lesão inicial, demonstrando a baixa variabilidade inter-classe característica em imagens histológicas. A similaridade morfológica entre as classes representa um dos principais desafios para sistemas de classificação automatizada, exigindo que os algoritmos identifiquem diferenças sutis em organização celular e padrões arquiteturais.

A preservação da integridade diagnóstica durante o aumento de dados torna-se, portanto, um requisito crítico. Em (VELDEN et al., 2021), é alertado que transformações inadequadas podem comprometer características morfológicas essenciais, como a organização celular, padrões arquiteturais e relações espaciais entre diferentes estruturas teciduais.

Outro aspecto relevante são os artefatos comuns no processamento histológico. Komura e Ishikawa (KOMURA; ISHIKAWA, 2018) catalogam diversos tipos de interferências, desde dobras e rasgos mecânicos até variações cromáticas e distorções ópticas, que devem ser considerados durante a implementação de técnicas de aumento.

2.6.2 Técnicas Geométricas Básicas

2.6.2.1 Transformações por Reflexão

As transformações de espelhamento horizontal e vertical (*flip*) constituem operações fundamentais que exploram as simetrias naturais presentes em muitas estruturas biológicas (HUSSAIN et al., 2017). O espelhamento horizontal simula diferentes orientações de corte durante a preparação histológica, enquanto o espelhamento vertical reproduz variações de posicionamento da amostra no campo microscópico. Ambas as transforma-

ções preservam completamente as características morfológicas e patológicas dos tecidos, representando augmentações semanticamente neutras (TAYLOR; NITSCHKE, 2018).

2.6.2.2 Rotação

A rotação de imagens histológicas deve ser cuidadosamente considerada devido à orientação específica de certas estruturas anatômicas (BULTEN et al., 2022). Rotações em ângulos discretos (múltiplos de 90°) são amplamente aceitas pois preservam a integridade das relações espaciais entre componentes celulares e extracelulares. Rotações arbitrárias, embora possam aumentar a diversidade dos dados, podem introduzir artefatos de interpolação que degradam a qualidade da imagem e potencialmente alteram características diagnósticas sutis (LITJENS et al., 2017).

2.6.2.3 Corte e Redimensionamento Aleatório

A técnica de *Random Resized Crop* baseia-se no princípio de que padrões patológicos relevantes podem ser identificados em diferentes escalas de observação (HE et al., 2016). Esta abordagem força os modelos a desenvolver invariância a mudanças de escala e posicionamento, características essenciais para a análise histopatológica, onde lesões podem apresentar diferentes tamanhos e localizações dentro do campo visual (CAMPANELLA et al., 2019).

2.6.3 Técnicas Colorimétricas

O *Color Jitter* fundamenta-se na necessidade de conferir robustez aos modelos contra variações cromáticas comuns no processamento histológico (MACENKO et al., 2009). As variações de coloração em imagens histológicas podem originar-se de múltiplos fatores: diferenças na concentração de corantes, tempo de coloração, pH das soluções, temperatura ambiente e características específicas dos sistemas de aquisição digital (REINHARD et al., 2002).

A técnica opera através de perturbações controladas nos componentes de brilho, contraste, saturação e matiz das imagens, simulando a variabilidade cromática encontrada entre diferentes laboratórios e sessões de preparação (CIOMPI et al., 2017). A preservação das relações cromáticas essenciais para o diagnóstico histopatológico constitui um aspecto crítico, uma vez que a diferenciação entre núcleos celulares (tipicamente corados em azul pela hematoxilina) e citoplasma/matriz extracelular (corados em rosa pela eosina) é fundamental para a análise patológica (RUIFROK; JOHNSTON, 2001).

2.6.4 Técnicas Avançadas Específicas para Histologia

2.6.4.1 Deformações Elásticas

As transformações elásticas baseiam-se em modelos matemáticos de deformação não-linear que simulam variações naturais na morfologia tecidual (SIMARD; STEINKRAUS; PLATT, 2003). Durante o processo de fixação histológica, os tecidos podem sofrer contrações, expansões e deformações irregulares devido a fatores como tipo de fixador utilizado, tempo de fixação e características intrínsecas do tecido (BANCROFT; GAMBLE, 2008).

O modelo matemático das deformações elásticas utiliza campos de deslocamento suavizados para produzir deformações espacialmente coerentes, evitando descontinuidades que poderiam resultar em artefatos irrealistas (RONNEBERGER; FISCHER; BROX, 2015). Esta abordagem é particularmente relevante para imagens histológicas pois reproduz variações morfológicas que ocorrem naturalmente durante o processamento das amostras.

2.6.4.2 Distorções Estruturadas

As técnicas de distorção em grade e distorção óptica abordam artefatos sistemáticos específicos da microscopia histológica (KOTHARI et al., 2013). A distorção em grade simula irregularidades periódicas que podem surgir durante o corte microscópico, especialmente relacionadas a variações na superfície da lâmina do micrótomo ou tensões não-uniformes no bloco de parafina (MULRANE et al., 2008).

A distorção óptica reproduz aberrações características de sistemas microscópicos, incluindo distorção radial e tangencial típicas de objetivas microscópicas (GAUDIN et al., 2010). Estas aberrações são mais pronunciadas nas bordas do campo visual e podem afetar a percepção de formas e dimensões celulares, aspectos críticos para a análise morfométrica automatizada (FICSOR et al., 2008).

2.6.4.3 Transferência Cromática

A transferência de cor baseia-se em métodos estatísticos de harmonização cromática desenvolvidos originalmente para processamento de imagens digitais (REINHARD et al., 2002). O método citado, utiliza o espaço de cores Lab*, que separa informações de luminância (L*) e cromaticidade (a* e b*), permitindo manipulação independente destes componentes.

Em histopatologia, esta técnica é particularmente valiosa para normalizar variações cromáticas entre diferentes protocolos de coloração, lotes de reagentes ou sistemas de aquisição digital (VAHADANE et al., 2016). A harmonização cromática permite que

modelos treinados em amostras de um laboratório específico generalizem melhor para amostras provenientes de diferentes instituições (TELLEZ et al., 2019).

2.6.4.4 Preenchimento Automático

As técnicas de preenchimento automático *(Inpainting)* utilizam algoritmos de visão computacional para reconstruir regiões danificadas ou com artefatos em imagens (BERTALMIO et al., 2000). Em histopatologia, pequenos artefatos como bolhas de ar, precipitados de corante ou dobras menores são comuns e podem interferir na análise automatizada (KOTHARI et al., 2013).

O *Inpainting* força os modelos de aprendizado profundo a desenvolver representações mais robustas, baseadas em características morfológicas fundamentais em vez de artefatos específicos (PATHAK et al., 2016). Esta abordagem é conceptualmente similar às estratégias de *dropout*, mas aplicada no domínio espacial, criando uma forma de regularização que melhora a capacidade de generalização (DEVRIES; TAYLOR, 2017).

2.6.5 Considerações Teóricas para Preservação Semântica

A aplicação de técnicas de aumento de dados em imagens médicas requer consideração cuidadosa dos invariantes semânticos que devem ser preservados (CASTRO; WALKER; GLOCKER, 2020). Em histopatologia, estes invariantes incluem:

- Invariantes Morfológicos: Relações espaciais entre núcleos celulares, organização arquitetural dos tecidos e padrões de crescimento celular devem permanecer inalterados (MADABHUSHI; LEE, 2016).
- Invariantes Cromáticos: As relações entre diferentes componentes cromáticos devem ser preservadas, especialmente a distinção fundamental entre estruturas basófilas (núcleos) e acidófilas (citoplasma) (RUIFROK; JOHNSTON, 2001).
- Invariantes Texturais: Padrões de textura característicos de diferentes tipos teciduais e graus de diferenciação celular constituem características diagnósticas críticas (NAIK et al., 2008).

2.6.6 Fundamentação Matemática da Regularização

Do ponto de vista teórico, o aumento de dados pode ser compreendido como uma forma de regularização implícita que expande o espaço de hipóteses dos modelos (BISHOP, 2006). Seja $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ o conjunto de dados original e $\mathcal{T}$ o conjunto de transformações aplicáveis. O conjunto aumentado $\mathcal{D}_{aug}$ é definido pela Equação 2.5:

$$\mathcal{D}_{aug} = \mathcal{D} \cup \{(T(x_i), y_i) : x_i \in \mathcal{D}, T \in \mathcal{T}, \text{preserva}(T, y_i)\} \tag{2.5}$$

, onde preserva(T, y_i) indica que a transformação T preserva o rótulo y_i.

Esta expansão do conjunto de dados corresponde a uma regularização que penaliza modelos que não são invariantes às transformações em $\mathcal{T}$ (SIMARD et al., 2000). Consequentemente, o aumento de dados direciona o aprendizado em direção a soluções mais generalizáveis, que capturam características essenciais dos dados em vez de peculiaridades específicas do conjunto de treinamento (GOODFELLOW et al., 2016).

2.7 Técnicas de Explicabilidade em Inteligência Artificial

O desenvolvimento de modelos de aprendizado profundo para aplicações médicas tem gerado crescente demanda por técnicas que tornem transparentes os processos decisórios desses sistemas (ADADI; BERRADA, 2018). A área de *eXplainable Artificial Intelligence* (XAI) emerge como resposta crítica à necessidade de compreensão e validação clínica de modelos complexos, particularmente em domínios onde decisões incorretas podem ter consequências graves (RUDIN, 2019).

As técnicas de explicabilidade dividem-se fundamentalmente em duas categorias principais. Os métodos *post-hoc* analisam modelos já treinados para extrair explicações através de técnicas de visualização e análise de ativações, oferecendo flexibilidade para aplicação em arquiteturas diversas sem modificações estruturais. Por outro lado, os métodos intrinsecamente interpretáveis incorporam mecanismos de explicação durante o próprio processo de treinamento, resultando em modelos naturalmente interpretáveis, porém com possíveis limitações de desempenho.

Para modelos de visão computacional, as abordagens *post-hoc* predominam devido à capacidade de aplicação em arquiteturas *state-of-the-art* sem comprometer a acurácia. Essas técnicas utilizam principalmente visualizações de mapas de calor e análises de gradientes para identificar regiões espaciais relevantes nas imagens de entrada.

2.7.1 Grad-CAM

O *Gradient-weighted Class Activation Mapping* (Grad-CAM) representa uma das técnicas mais influentes para explicabilidade em redes neurais convolucionais (SELVARAJU et al., 2017). A abordagem fundamenta-se na hipótese de que os gradientes das classes de interesse em relação aos mapas de características contêm informações valiosas sobre a importância espacial de diferentes regiões da imagem.

A formulação matemática do Grad-CAM para uma classe específica c é expressa pela Equação 2.6:

$$L^c_{Grad-CAM} = \text{ReLU}\left(\sum_k \alpha^c_k \cdot A^k\right) \tag{2.6}$$

onde A^k representa o mapa de ativação do k-ésimo canal da camada convolucional selecionada, e α^c_k são os pesos de importância específicos para a classe c.

Os pesos α^c_k são calculados através da média espacial dos gradientes da classe em relação aos mapas de características, conforme a Equação 2.7:

$$\alpha^c_k = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \frac{\partial y^c}{\partial A^k_{i,j}} \tag{2.7}$$

onde y^c representa o score da classe c antes da aplicação da função softmax, H e W são as dimensões espaciais dos mapas de características, e $A^k_{i,j}$ denota a ativação na posição (i, j) do canal k.

A aplicação da função ReLU na Equação 2.6 assegura que apenas influências positivas sejam consideradas na visualização final, eliminando contribuições negativas que poderiam introduzir ruído interpretativo. Esta característica é particularmente importante em aplicações médicas, onde a identificação precisa de regiões patológicas é fundamental.

Na Figura 15 é ilustrado o processo completo de geração de mapas de saliência através do Grad-CAM, demonstrando como os mapas de características extraídos pelas camadas convolucionais são combinados com informações de gradientes para produzir visualizações interpretáveis das decisões do modelo.

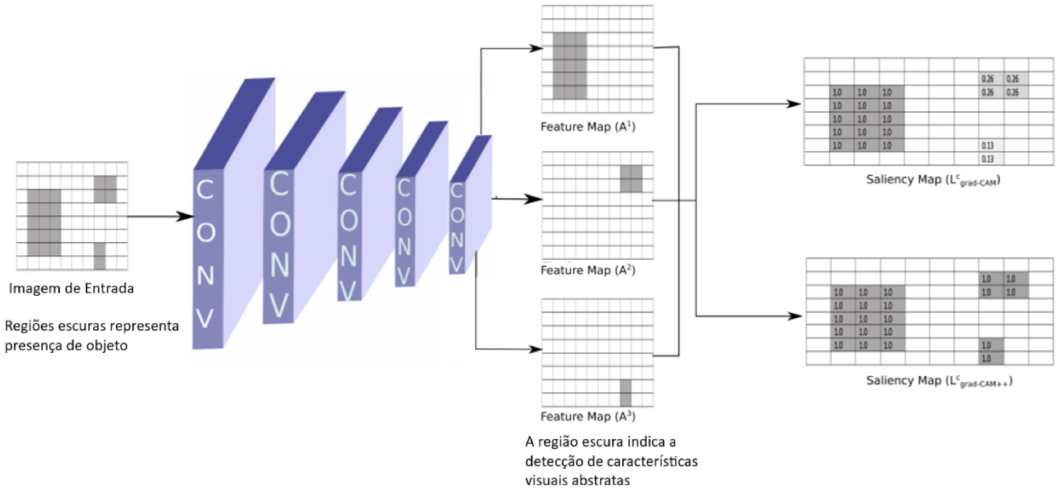


Figura 15 – Pipeline de funcionamento do Grad-CAM. A imagem de entrada é processada por uma CNN, gerando mapas de características (*Feature Maps*) nas camadas convolucionais. Através do cálculo de gradientes, são produzidos mapas de saliência que destacam as regiões mais relevantes para a decisão do modelo. O Grad-CAM++ representa uma versão aprimorada que oferece maior precisão na localização espacial das características discriminativas.

2.7.2 Explicabilidade em Vision *Transformers*

A arquitetura Vision *Transformer* (ViT) introduz paradigmas distintos para explicabilidade devido à sua fundamentação em mecanismos de atenção global, contrastando com a natureza hierárquica e localmente conectada das redes convolucionais (DOSOVITSKIY et al., 2020). Esta diferença arquitetural requer adaptações específicas nas técnicas de explicabilidade existentes.

2.7.2.1 Mecanismos de Atenção Multi-Cabeça

O mecanismo central dos *Transformers* baseia-se na atenção multi-cabeça, matematicamente definida pela Equação 2.8:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.8}$$

onde Q, K e V representam as matrizes de *queries, keys e values*, respectivamente, e d_k corresponde à dimensionalidade das *keys*.

Para análise de explicabilidade, particular relevância assume a atenção entre o *token* de classificação (CLS) e os patches da imagem, expressa pela Equação 2.9:

$$A_{cls \to patches} = \text{softmax}\left(\frac{q_{cls} \cdot K_{patches}^T}{\sqrt{d_k}}\right) \tag{2.9}$$

onde q_{cls} representa a *query* do *token* de classificação e $K_{patches}$ contém as *keys* de todos os *patches* da imagem.

Esta formulação permite extração direta dos pesos de atenção nativos, oferecendo insights sobre quais regiões espaciais o modelo considera mais discriminativas para a tarefa de classificação, sem necessidade de cálculos adicionais de gradientes.

2.7.2.2 Adaptação do Grad-CAM para ViT

A aplicação do Grad-CAM em *Vision Transformers* requer modificações substanciais devido às diferenças arquiteturais fundamentais. Enquanto em CNNs os gradientes fluem através de mapas de características espacialmente organizados, nos ViTs este fluxo ocorre através de representações tokenizadas e mecanismos de atenção.

A adaptação para ViT utiliza a última camada de atenção como ponto de análise, conforme a Equação 2.10:

$$L_{ViT-CAM}^c = \text{ReLU}\left(\sum_k \beta_k^c \cdot A_k^{attention}\right) \tag{2.10}$$

onde $A_k^{attention}$ representa as ativações dos patches na última camada de atenção, e β_k^c são os pesos calculados através dos gradientes específicos desta arquitetura.

2.7.3 Explicabilidade em Arquiteturas com *Split-Attention*

Arquiteturas híbridas como ResNeSt combinam características convolucionais tradicionais com mecanismos de atenção adaptativa através do *Split-Attention* (ZHANG et al., 2022). Esta hibridização cria oportunidades e desafios únicos para técnicas de explicabilidade.

O mecanismo *Split-Attention* opera dividindo os canais de características em múltiplos grupos, aplicando atenção adaptativa dentro de cada grupo antes da recombinação das informações. Este processo permite foco seletivo em diferentes aspectos das características extraídas, resultando em representações mais contextuais e discriminativas.

Para estas arquiteturas, o Grad-CAM tradicional mantém aplicabilidade nas camadas convolucionais, sendo calculado conforme a Equação 2.11:

$$L^c_{Split-CAM} = \text{ReLU}\left(\sum_k w_k^c \cdot F^k\right) \tag{2.11}$$

onde F^k são os mapas de características da camada convolucional de interesse, e w_k^c representam os pesos calculados através da média espacial dos gradientes.

2.7.4 Visualização de Ativações Intermediárias

Complementarmente às técnicas baseadas em gradientes, a visualização direta de ativações intermediárias oferece perspectivas sobre o processamento hierárquico de características. Esta abordagem é particularmente valiosa para compreender como diferentes camadas capturam padrões de complexidade crescente.

Em redes convolucionais, a visualização de mapas de características é obtida através da agregação espacial das ativações, definida pela Equação 2.12:

$$A_{CNN} = \text{AvgPool}_{channels}\left(\text{Conv}_{layer}(I)\right) \tag{2.12}$$

onde $\text{Conv}_{layer}(I)$ representa as ativações de uma camada convolucional específica para a imagem I, e $\text{AvgPool}_{channels}$ denota a operação de média sobre todos os canais de características.

O mapa resultante requer normalização para visualização adequada conforme a Equação 2.13:

$$A_{norm} = \frac{A_{CNN} - \min(A_{CNN})}{\max(A_{CNN}) - \min(A_{CNN})} \tag{2.13}$$

2.7.5 Métricas de Avaliação da Explicabilidade

A validação quantitativa da qualidade das explicações constitui aspecto fundamental para estabelecer confiança nas técnicas XAI. Diferentes métricas foram propostas para avaliar aspectos específicos da explicabilidade, cada uma capturando propriedades distintas das visualizações geradas.

2.7.5.1 Fidelidade da Explicação

A fidelidade quantifica o grau com que uma explicação reflete genuinamente o processo decisório do modelo. Esta métrica é avaliada através da degradação de desempenho observada quando regiões identificadas como importantes são sistematicamente ocultadas, definida pela Equação 2.14:

$$\text{Fidelidade} = \text{Acc}(M, I) - \text{Acc}(M, I \odot (1 - M_{XAI})) \tag{2.14}$$

onde M representa o modelo, I a imagem de entrada, M_{XAI} a máscara binária das regiões mais importantes identificadas pela técnica XAI, e $\odot$ denota o produto elemento a elemento.

2.7.5.2 Consistência Espacial

A consistência espacial avalia a correlação entre diferentes técnicas XAI aplicadas à mesma imagem, fornecendo *insights* sobre a concordância entre métodos distintos, como definido na Equação 2.15:

$$\rho_{XAI_i, XAI_j} = \frac{\text{cov}(L_i, L_j)}{\sigma_{L_i} \cdot \sigma_{L_j}} \tag{2.15}$$

onde L_i e L_j representam os mapas de explicação gerados pelas técnicas i e j, respectivamente.

2.7.5.3 Estabilidade Temporal

A estabilidade temporal quantifica a variação das explicações durante o processo de treinamento, indicando a convergência e robustez das visualizações, como definido na Equação 2.16:

$$S_{temporal} = 1 - \frac{1}{T-1} \sum_{t=1}^{T-1} ||L^{(t+1)} - L^{(t)}||_2 \tag{2.16}$$

onde T representa o número total de épocas analisadas, e $L^{(t)}$ denota o mapa de explicação na época t.

2.7.6 Desafios Específicos em Imagens Histopatológicas

A aplicação de técnicas XAI em imagens histopatológicas apresenta complexidades adicionais relacionadas às características específicas deste domínio. As imagens histológicas exibem alta variabilidade em termos de coloração, preparação e qualidade, exigindo robustez específica das técnicas de explicabilidade.

Os principais desafios incluem a distinção entre características patológicas genuínas e artefatos de preparação, a interpretação de padrões de coloração específicos (hematoxilina e eosina), e a necessidade de explicações consistentes com conhecimento histopatológico estabelecido. Adicionalmente, a natureza multi-escala das estruturas histológicas requer técnicas capazes de capturar tanto detalhes celulares finos quanto padrões arquiteturais globais.

A validação das explicações neste domínio necessita colaboração interdisciplinar entre especialistas em *machine learning* e patologistas, assegurando que as visualizações geradas sejam tanto tecnicamente corretas quanto clinicamente interpretáveis. Esta validação cruzada é essencial para estabelecer confiança clínica nos sistemas automatizados de diagnóstico histopatológico.

3 Estudos Relacionados

Este capítulo apresenta uma revisão sistemática da literatura científica contemporânea relacionada à aplicação de técnicas de aprendizado profundo para análise de imagens histológicas de lesões orais, com ênfase em métodos de classificação aplicados especificamente à lesões da cavidade oral. A análise priorizou trabalhos publicados nos últimos 5 anos que empregam arquiteturas modernas de redes neurais e apresentam resultados quantitativos claramente reportados.

Para cada trabalho selecionado, são apresentados: contexto e motivação, metodologia empregada detalhadamente, resultados quantitativos obtidos, principais contribuições e limitações identificadas. Esta estrutura permite uma análise comparativa sistemática e fundamenta as lacunas que motivam o presente trabalho.

3.1 Classificação Multiclasse de Displasia Oral

Em (SILVA et al., 2024) foi apresentarada uma contribuição significativa para a área ao desenvolver e disponibilizar publicamente um conjunto de dados abrangente de displasia epitelial oral, acompanhado de experimentos sistemáticos de classificação multiclasse utilizando diferentes algoritmos de aprendizado de máquina. O estudo envolveu a criação de um *dataset* público contendo 456 imagens histológicas adquiridas de 30 línguas de camundongos, categorizadas entre diferentes graus de lesão. As imagens foram manualmente anotadas por especialista treinado e validadas por patologista qualificado, garantindo alta qualidade das anotações. A metodologia incluiu experimentos comparativos para segmentação semântica e de instância, além de avaliação de métodos de normalização cromática HE. Para classificação, foram testadas diferentes arquiteturas de CNN (VGG, ResNet, DenseNet) e algoritmos tradicionais de aprendizado de máquina (*Random Forest*, SVM, k-NN). O *dataset* foi organizado seguindo padrões internacionais de disponibilização, incluindo metadados detalhados e *scripts* de avaliação padronizados. Na etapa de classificação multiclasse, o algoritmo *Random Forest* apresentou o melhor desempenho, alcançando acurácia de 94,22% na discriminação entre os diferentes graus de displasia. A análise comparativa revelou que métodos tradicionais de aprendizado de máquina superaram as CNNs testadas neste *dataset* específico, contrariando tendências observadas em *datasets* maiores. A avaliação de normalização cromática demonstrou melhoria de 3-7% na precisão quando aplicada sistematicamente. O estudo estabeleceu *benchmarks* de referência para futuros trabalhos na área, incluindo métricas detalhadas por classe e análise de matriz de confusão. As principais limitações incluem o tamanho relativamente pequeno do *dataset* (456 imagens), que pode limitar a eficácia de métodos de aprendizado profundo,

dependência de anotações manuais que podem conter variabilidade inter-observador, e validação restrita a modelos de camundongos que podem apresentar diferenças morfológicas comparado a tecidos humanos. Adicionalmente, o estudo não incorpora técnicas de explicabilidade para compreensão das características discriminativas utilizadas pelos classificadores.

O trabalho apresentado em (ADEL et al., 2018) introduziu uma abordagem inovadora para classificação de displasia epitelial oral baseada na combinação de múltiplos descritores de características visuais, representando uma contribuição importante para métodos de extração de características em imagens histológicas. O trabalho empregou três algoritmos distintos de extração de características: *Oriented FAST and Rotated BRIEF* (ORB), *Scale Invariant Feature Transform* (SIFT) e *Speeded Up Robust Features* (SURF). Estas técnicas foram selecionadas por sua capacidade de detectar pontos de interesse robustos e invariantes a transformações geométricas comuns em imagens histológicas. O pré-processamento incluiu conversão do espaço de cores RGB para HSV para melhor separação de informações cromáticas e morfológicas, seguido por redimensionamento padronizado para 600×600 *pixels*. As características extraídas pelos três algoritmos foram concatenadas em um vetor de características unificado e classificadas utilizando *Support Vector Machine* (SVM) com *kernel* RBF otimizado através de validação cruzada. A metodologia foi avaliada em 46 casos de imagens com magnificação de 100x, divididos aleatoriamente em conjuntos de treinamento (32 imagens, 70%) e teste (14 imagens, 30%). A abordagem proposta alcançou precisão de 92,8% na classificação de displasia epitelial oral. A análise de contribuição individual dos descritores revelou que SIFT forneceu as características mais discriminativas, seguido por SURF e ORB. A combinação dos três descritores superou significativamente o desempenho individual de cada método, demonstrando complementaridade das informações extraídas. As limitações identificadas incluem o tamanho limitado do *dataset* de validação (46 casos), que pode comprometer a generalização dos resultados, dependência de otimização manual dos parâmetros dos algoritmos de extração de características, e ausência de comparação com métodos de aprendizado profundo contemporâneos. O estudo também não aborda aspectos de interpretabilidade das características extraídas, limitando a compreensão dos critérios morfológicos capturados pelos descritores visuais.

3.2 Análise de Lacunas e Oportunidades de Pesquisa

Embora trabalhos recentes como o apresentado em (SILVA et al., 2024) tenham abordado classificação multiclasse de displasia, a maioria dos estudos ainda concentra-se em tarefas de segmentação de estruturas nucleares. Esta lacuna é particularmente relevante para aplicações clínicas, onde a classificação direta e precisa do grau de displasia é fundamental para determinação do prognóstico e planejamento terapêutico. Nenhum

estudo explora outros tipos de lesões presentes na cavidade oral. Ainda os trabalhos revisados não incorporam técnicas sistemáticas de explicabilidade artificial (XAI) para tornar transparentes os processos decisórios dos modelos. Mesmo trabalhos recentes focados em classificação, como os apresentados em (ADEL et al., 2018) e (SILVA et al., 2024), não exploram mecanismos de interpretabilidade. Esta ausência representa uma barreira significativa para adoção clínica, especialmente em aplicações médicas onde a compreensão das decisões automatizadas é essencial para validação por especialistas. Trabalhos de classificação revisados utilizam predominantemente métodos tradicionais (*Random Forest*, SVM) ou CNNs convencionais, sem exploração sistemática de arquiteturas modernas como *Vision Transformers*. Esta lacuna limita a compreensão sobre qual paradigma arquitetural é mais adequado para capturar padrões morfológicos complexos em tecidos displásicos, especialmente considerando a capacidade dos *Transformers* de fornecer interpretabilidade natural através de mecanismos de atenção.

Essas lacunas fundamentam a originalidade e relevância do presente trabalho, que propõe uma abordagem integrada combinando *Vision Transformers* e redes convolucionais para classificação, incorporando técnicas de XAI para análise sistemática de explicabilidade, oferecendo contribuições metodológicas e práticas significativas para o avanço da patologia computacional oral.

4 Metodologia

Neste capítulo são apresentados os conceitos relacionados aos métodos utilizados no estudo, com ênfase nas técnicas de classificação e explicabilidade aplicadas a imagens histológicas oral. Foram exploradas arquiteturas modernas de aprendizado profundo, especificamente ViT e *ResNeSt*, com o objetivo de avaliar sua eficácia no diagnóstico automatizado.

A metodologia foi estruturada em quatro etapas principais: (1) configuração dos bancos de imagens histológicas com balanceamento automático, (2) implementação das arquiteturas por meio de *transfer learning*, (3) aplicação de técnicas avançadas de aumento de dados específicas para histologia e (4) análise de explicabilidade a partir de múltiplas técnicas de *Explainable Artificial Intelligence* (XAI).

Na Figura 16 é apresentado o fluxograma exemplificando o encadeamento das etapas aplicadas neste trabalho. Todos os experimentos foram conduzidos em ambiente Python e executados em uma máquina com processador Intel Core i9-13900k, placa de vídeo GeForce RTX 4070 TI e 32 GB de memória RAM.

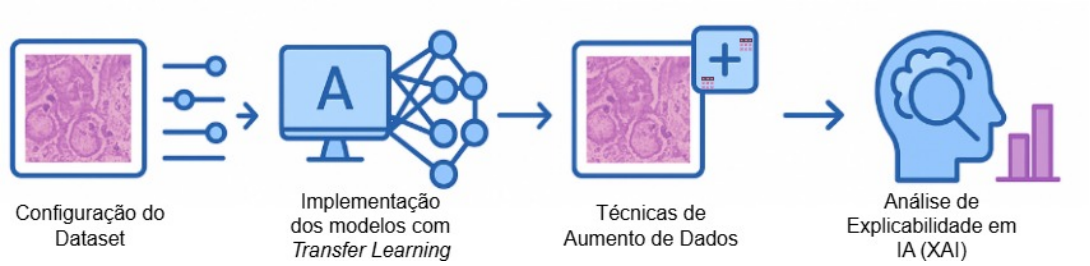


Figura 16 – Fluxograma da Metodologia Adotada neste Trabalho em que são definidas as etapas de configuração do *dataset*, modelos, técnicas de aumento de dados e explicação.

A avaliação do desempenho dos modelos, bem como a análise comparativa entre as diferentes arquiteturas e estratégias de otimização, foi realizada com base em critérios quantitativos e qualitativos. As métricas de avaliação foram cuidadosamente selecionadas para fornecer uma compreensão abrangente da performance, da capacidade de generalização e da interpretabilidade dos modelos.

O desempenho foi mensurado utilizando a acurácia nos conjuntos de validação e teste. A análise comparativa se concentrou nos seguintes aspectos:

i) Impacto das Otimizações: A *performance* dos modelos “com” e “sem” as otimizações de treinamento foi comparada para quantificar o ganho de generalização. Esta etapa, crucial para o desenvolvimento metodológico, permitiu demonstrar que as

estratégias de aumento de dados e regularização não apenas melhoram a acurácia, mas também estabilizam a curva de aprendizado e mitigam o *overfitting*.

ii) Comparação entre Arquiteturas: A análise do desempenho relativo do ViT e do *ResNeSt* no mesmo *dataset* foi conduzida para investigar a adequação de cada arquitetura para a tarefa de classificação de imagens histológicas. A comparação considerou não apenas a acurácia, mas também a eficiência computacional, o número de parâmetros e a natureza de seus mecanismos de atenção.

iii) Análise de Explicabilidade Visual: O uso das ferramentas de explicabilidade (Grad-CAM, Mapas de Atenção e *Saliency Maps*) foi fundamental para interpretar e validar visualmente o comportamento dos modelos. A análise se concentrou em:

- A evolução do foco de atenção dos modelos ao longo do treinamento, de uma atenção difusa para uma atenção mais precisa e localizada.
- A consistência das explicações entre as diferentes técnicas XAI para o mesmo modelo.
- A comparação entre os padrões de atenção do ViT e do *ResNeSt*, evidenciando como suas arquiteturas distintas levam a diferentes percepções visuais.

4.1 Arquiteturas Investigadas

Para este estudo, foram selecionadas arquiteturas complementares de modelos de aprendizagem profunda: *Vision Transformer* (ViT-Base - ver Figura 11) (DOSOVITSKIY et al., 2020) e *ResNeSt-50d*, ver Figura 12 (ZHANG et al., 2022). A escolha fundamentou-se em características arquiteturais distintas que oferecem abordagens complementares para classificação histológica.

4.1.1 Vision Transformer

O ViT representa uma adaptação revolucionária da arquitetura *Transformer*, originalmente desenvolvida para processamento de linguagem natural (VASWANI et al., 2017), para o domínio de visão computacional. O modelo ViT-Base utilizado possui 86,6 milhões de parâmetros e foi pré-treinado no *dataset* ImageNet-21K, em que suas características são:

- **Mecanismo de atenção global:** Permite captura de dependências de longo alcance entre diferentes regiões da imagem através do mecanismo *multi-head self-attention*;
- **Processamento por patches:** Divide a imagem em patches de 16×16 pixels, tratando cada patch como um token sequencial;

- **Ausência de indutivos visuais:** Não incorpora *inductive biases* específicos para imagens, aprendendo representações puramente a partir dos dados.

O ViT foi selecionado por sua capacidade de modelar relações globais em imagens histológicas, onde padrões patológicos podem se manifestar em múltiplas regiões simultaneamente. Sua arquitetura baseada em atenção oferece interpretabilidade natural através dos mapas de atenção, fundamental para aplicações médicas.

4.1.2 *ResNeSt-50d*

O *ResNeSt* (*Split-Attention Networks*) é uma arquitetura híbrida que combina os benefícios das redes residuais com mecanismos de atenção adaptativa (ZHANG et al., 2022). O modelo *ResNeSt-50d* utilizado possui 27,5 milhões de parâmetros e foi pré-treinado no *dataset ImageNet-1K*.Entre suas características destaca-se:

- *Split-Attention*: Implementa atenção adaptativa que divide canais em grupos e aplica atenção seletiva dentro de cada grupo;
- Conexões residuais: Mantém as vantagens das ResNets para treinamento de redes profundas com gradientes estáveis;
- Eficiência computacional: Oferece melhor relação desempenho/parâmetros comparado a arquiteturas puramente baseadas em atenção.

O *ResNeSt* foi selecionado por sua eficácia comprovada em tarefas de classificação de imagens médicas (PASSOS; MISHRA, 2021), combinando a robustez das arquiteturas convolucionais com mecanismos de atenção que permitem foco adaptativo em características relevantes. Sua arquitetura hierárquica é particularmente adequada para capturar detalhes finos em imagens histológicas.

4.1.3 Complementaridade das Arquiteturas

A combinação dessas duas arquiteturas oferece perspectivas complementares:

- Escala de análise: ViT processa a imagem globalmente através de atenção, enquanto ResNeSt constrói representações hierarquicamente;
- Interpretabilidade: ViT oferece mapas de atenção nativos, ResNeSt fornece análise de ativações convolucionais;
- Eficiência: ResNeSt é mais eficiente computacionalmente, ViT oferece maior capacidade representacional;

- Robustez: Combinação permite avaliação da consistência de resultados entre paradigmas arquiteturais distintos.

Esta seleção permite investigação abrangente de como diferentes abordagens de processamento visual impactam a classificação histológica, fornecendo *insights* sobre quais características arquiteturais são mais adequadas para o domínio específico de patologia oral.

4.1.4 Configurações Específicas

Na Tabela 1 é apresentado um resumo comparativo das principais configurações adotadas para cada arquitetura.

Tabela 1 – Comparação das arquiteturas utilizadas no estudo.

Característica	**ViT-Base**	**ResNeSt50d**
Número de parâmetros	86,6M	27,5M
Tamanho do *patch* / *kernel*	16×16	Convoluções 7×7
Dimensão de entrada	224×224	224×224
Mecanismo de atenção	Global (*Multi-Head Self-Attention*)	Local (*Split-Attention*)
Pré-treinamento	ImageNet-21K	ImageNet-1K
Transfer Learning	Camada de classificação adaptada (7 classes)	Camada de classificação adaptada (7 classes)

Observa-se que o *ViT-Base* apresenta um número significativamente maior de parâmetros, refletindo maior capacidade de modelagem, enquanto o *ResNeSt50d* possui arquitetura mais enxuta e eficiente, explorando mecanismos de atenção locais com convoluções hierárquicas.

4.2 *Datasets* e Pré-Processamento

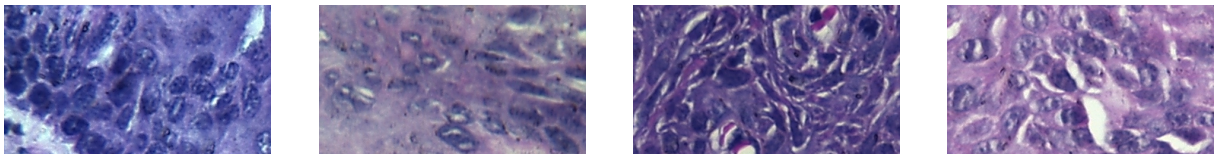
4.2.1 *Datasets* de Imagens Histológicas

Este trabalho utilizou dois bancos de imagens histológicas, sendo um de domínio público e outro de domínio privado, ambos obtidos a partir de lâminas de tecidos da cavidade oral de camundongos. A utilização de dois conjuntos de dados distintos permitiu uma análise mais abrangente da capacidade de generalização e adaptação dos modelos às variações de aquisição e preparação das amostras.

O primeiro banco de imagens utilizado neste trabalho denominado *OralEpitheliumDB* (SILVA et al., 2024), renomeado neste estudo como Banco Silva, é composto de imagens histológicas de diferentes graus de displasia, obtidas a partir de línguas de camundongos C57BL/6, que foram lesionadas experimentalmente utilizando o carcinógeno 4-nitroquinolina-N-óxido (4NQO), diluído na água dos animais. Esses experimentos foram devidamente aprovados pelo Comitê de Ética na Utilização de Animais sob o número

038/09. As lâminas com tecidos corados foram digitalizadas a partir de um microscópio óptico modelo Leica DM500, utilizando um grau de magnificação de 400x.

Essas imagens foram classificadas por dois especialistas em patologia oral entre tecido saudável, displasia leve, displasia moderada e displasia severa, conforme o modelo definido por (LUMERMAN; FREEDMAN; KERPEL, 1995). As imagens originais, com resolução de 2048 x 1536 pixels, foram utilizadas para compor um total de 456 Regiões de Interesse (ROIs) com resolução de 450 x 250 pixels, com 114 ROIs para cada uma das quatro classes. Um exemplo de cada classe do banco pode ser observado na Figura 17.



(a) (b) (c) (d)

Figura 17 – Exemplos de imagens do Banco Silva: (a) tecido saudável, (b) displasia leve, (c) displasia moderada e (d) displasia severa.

O segundo banco de imagens utilizado, denominado OralKOWTDB, renomeado neste trabalho como Banco Costa, foi construído a partir de lesões obtidas experimentalmente com a exposição ao carcinógeno 4NQO, em roedores. Este banco de dados é subdividido em duas populações distintas para fins de análise comparativa:

- KO (*Knockout*): Camundongos da cepa B6.129P2-Nos2, que possuem o gene *Nos2* desativado, o que os torna geneticamente mais suscetíveis a determinadas alterações teciduais.
- WT (*Wild-Type*): Camundongos do tipo selvagem, que não apresentam a modificação genética do gene *Nos2*, servindo como grupo de controle na análise experimental.

Todos os procedimentos executados foram devidamente aprovados pelo Comitê de Ética na Utilização de Animais sob o número 100/18. As lâminas histológicas foram digitalizadas no equipamento Leica Aperio AT2 com magnificação de 400x. As regiões de interesse (ROIs) foram extraídas utilizando o software QuPath, com resolução de 200 x 200 pixels. As classificações, realizadas por dois patologistas, foram organizadas em sete categorias: Carcinoma, Displasia intensa, Displasia leve, Displasia moderada, Hiperplasia, Papiloma e Tecido saudável.

A Tabela 2 apresenta a distribuição de imagens por classe nos WT e KO.

Tabela 2 – Distribuição de imagens por classe nos *sub-datasets* WT e KO.

Sub-dataset	Carcinoma	Severa	Leve	Moderada	Hiperplasia	Papiloma	Saudável	Total
KO	127	110	93	141	81	82	215	849
WT	102	126	155	84	160	83	245	955

4.2.2 Estratégias de Pré-processamento

As imagens de todos os *datasets* foram submetidas a um processo de pré-processamento padronizado para garantir que as entradas para os modelos fossem consistentes e otimizadas para o *transfer learning*. As etapas incluíram:

i) Redimensionamento: Todas as imagens foram redimensionadas para a resolução de 224×224 pixels, a dimensão padrão de entrada dos modelos pré-treinados.

ii) Normalização: As imagens foram normalizadas usando a média e o desvio padrão do dataset *ImageNet*, garantindo que a distribuição dos dados de entrada fosse consistente com aquela em que os modelos foram pré-treinados. Na Figura 18 é exemplificado o processo de normalização das imagens.

Figura 18 – Processo de normalização aplicado às imagens do *dataset* CIFAR-10. (a) Imagem original com valores de pixel 0-255; (b) Após conversão para tensor e redimensionamento (224×224), com valores normalizados para 0-1; (c) Após normalização usando média e desvio padrão do *ImageNet*, resultando em valores otimizados para *transfer learning* mas não diretamente visualizáveis.

iii) Padronização de Entrada: As imagens foram convertidas para tensores PyTorch com formato e escala compatíveis com os modelos pré-treinados.

iv) Divisão Estratificada: Os *datasets* foram divididos mantendo-se a proporção original de cada classe nos conjuntos de treino (70%), validação (15%) e teste (15%), através de algoritmo de divisão estratificada personalizado. Esta abordagem garante que cada subset mantenha a representatividade estatística das classes originais.

v) Balanceamento do Conjunto de Treino: Para mitigar o desbalanceamento natural entre as classes nos *datasets* histológicos (conforme Tabela 2), foi implementada

uma estratégia de balanceamento automático. Esta abordagem identifica a classe majoritária no conjunto de treino e replica as amostras das classes minoritárias até que todas as classes possuam o mesmo número de exemplos. Esse balanceamento é aplicado exclusivamente ao conjunto de treino, mantendo os conjuntos de validação e teste com sua distribuição natural para uma avaliação mais realista do desempenho.

vi) Garantia de Reprodutibilidade: Para assegurar a reprodutibilidade dos experimentos, foram fixadas sementes aleatórias ($seed = 42$) em todos os geradores de números pseudo-aleatórios (PyTorch, NumPy, Python random), configuração determinística dos algoritmos CUDA e versionamento rigoroso das dependências.

O pré-processamento, em conjunto com as estratégias de otimização, formou a base para os experimentos e a análise subsequente do desempenho e da interpretabilidade dos modelos.

4.3 Estratégias de Aumento de Dados

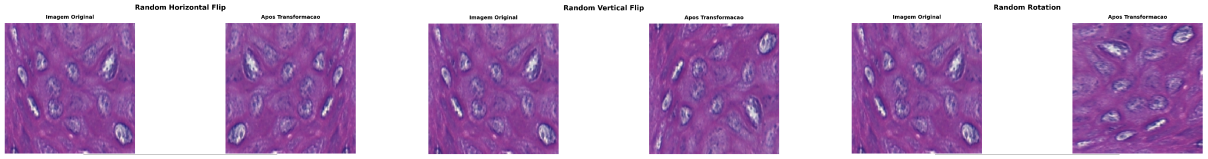
Para aumentar a robustez dos modelos diante da limitação de dados histológicos disponíveis, foram implementadas duas estratégias distintas de aumento de dados, permitindo avaliar o impacto de diferentes níveis de complexidade no desempenho dos modelos. A seleção das técnicas baseou-se na necessidade de preservar as características patológicas relevantes das imagens histológicas enquanto se introduzia variabilidade suficiente para melhorar a generalização.

4.3.1 Estratégia Padrão

Aplicada como *baseline* em todos os experimentos, esta estratégia incluiu transformações geométricas e colorimétricas fundamentais, implementadas com as seguintes parametrizações:

- Transformações geométricas básicas: *Horizontal Flip* (p=0,5), *Vertical Flip* (p=0,5), e Rotação aleatória aplicada em ângulos discretos (0°, 90°, 180°, 270°) para simular diferentes orientações de corte das lâminas histológicas;
- *Random Resized Crop* (p=0,7): Implementado com fator de escala entre 0,8 e 1,0, seguido de redimensionamento para 224×224 pixels, preservando características importantes do tecido em diferentes escalas;
- *Color Jitter* (p=0,7): Configurado com variações controladas de brilho (±0,2), contraste (±0,2), saturação (±0,2) e matiz (±0,1) para simular variações interlaboratoriais de coloração.

Na Figura 19 é demonstrada a aplicação das transformações geométricas básicas em amostras histológicas, evidenciando a preservação das características morfológicas essenciais.



(a) *Horizontal Flip* (b) *Vertical Flip* (c) Rotação Aleatória

Figura 19 – Aplicação das técnicas básicas de aumento de dados geométrico em imagens histológicas de displasia oral: (a) espelhamento horizontal preservando orientações celulares, (b) espelhamento vertical mantendo integridade tecidual, (c) rotação aleatória reproduzindo variações de posicionamento microscópico.

As técnicas *Random Resized Crop* e *Color Jitter* foram implementadas para abordar desafios específicos da classificação histológica. Na Figura 20 é exibido a aplicação do *Random Resized Crop*, demonstrando como a técnica preserva padrões patológicos relevantes em diferentes escalas de observação.

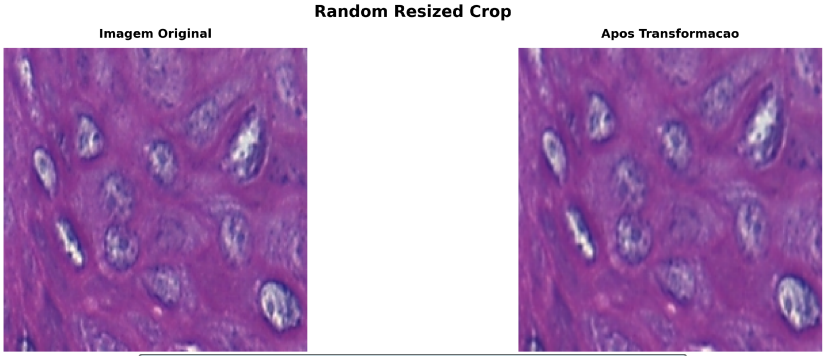


Figura 20 – Aplicação da técnica *Random Resized Crop* em amostra histológica de displasia oral. À esquerda, a região de interesse original (450×250 pixels). À direita, a imagem transformada após corte aleatório e redimensionamento para 224×224 pixels, preservando características morfológicas essenciais do tecido.

A técnica *Color Jitter* foi parametrizada especificamente para simular variações cromáticas comuns no processamento histológico, conforme demonstrado na Figura 21.

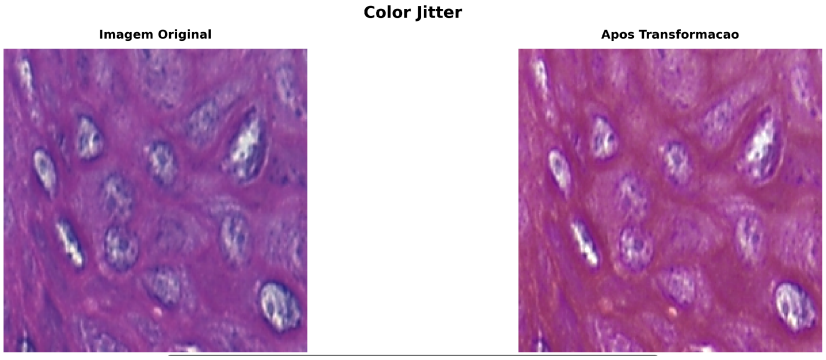


Figura 21 – Aplicação da técnica Color Jitter em amostra histológica corada com hematoxilina-eosina. À esquerda, a coloração original. À direita, a imagem com variações controladas de brilho, contraste, saturação e matiz, simulando diferenças interlaboratoriais de preparação.

4.3.2 Estratégia Avançada Específica para Histologia

Esta estratégia expandiu o conjunto padrão com técnicas especializadas para simular variações e artefatos realistas do processamento histológico. Mantendo todas as transformações da estratégia padrão, foram adicionadas as seguintes técnicas com suas respectivas parametrizações:

- *Transpose* (p=0,35): Transposição matricial para espelhamento diagonal, simulando orientações de corte não convencionais;
- *Elastic Transform* (p=0,35): Implementado com parâmetros =120 e =6 para deformações não-lineares suaves;
- *Grid Distortion* (p=0,35): Deformação estruturada em grade para reproduzir artefatos sistemáticos de processamento;
- *Optical Distortion* (p=0,35): Aplicado com limite de distorção de $\pm 1{,}0$ para simular aberrações ópticas microscópicas.

Na Figura 22 é demonstrada a aplicação da técnica *Transpose*, evidenciando a manutenção das características patológicas após a transformação diagonal.

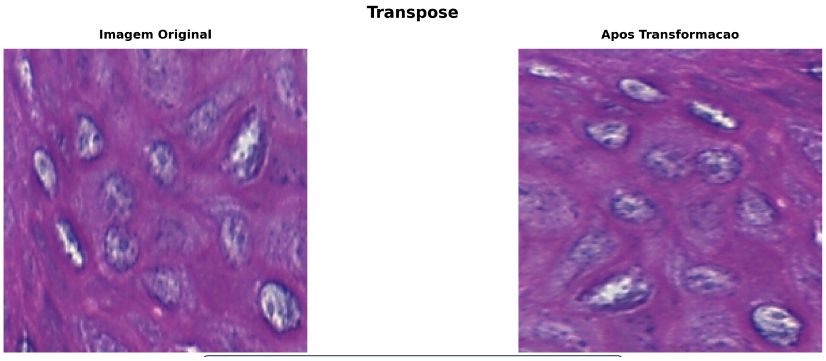


Figura 22 – Aplicação da técnica *Transpose* em amostra histológica de displasia. À esquerda, a imagem original. À direita, a imagem após transposição matricial, demonstrando preservação das características morfológicas celulares e teciduais.

As técnicas de deformação especializada foram configuradas para simular artefatos específicos do processamento histológico, conforme ilustrado na Figura 23.

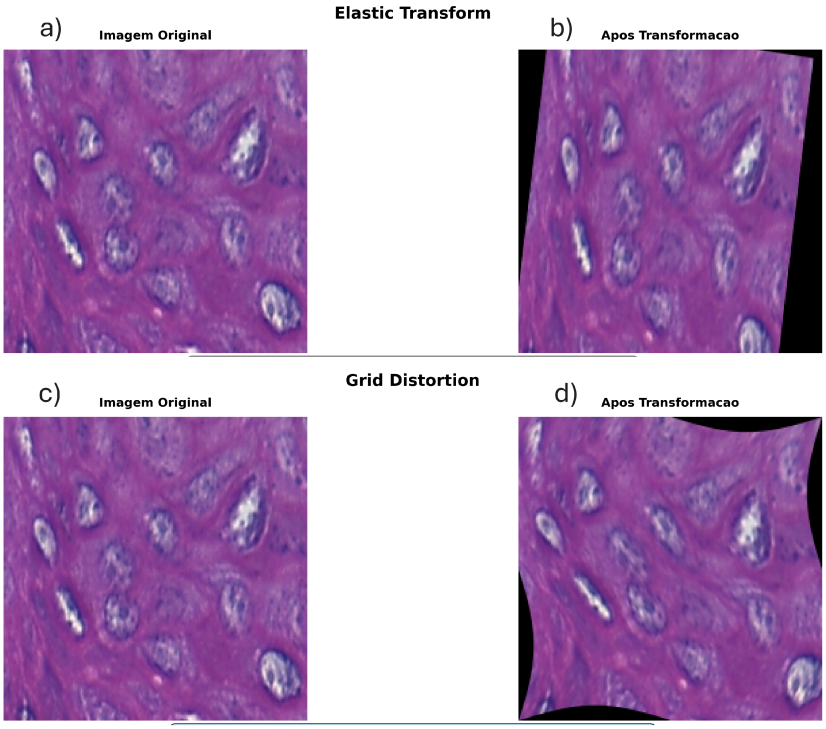


Figura 23 – Aplicação de técnicas de deformação especializada em amostras histológicas, sendo a) e c) Imagens Histológicas originais e b) Elastic Transform (=120, =6) simulando variações naturais da morfologia tecidual durante fixação, e d) *Grid Distortion* reproduzindo artefatos sistemáticos de montagem de lâminas e irregularidades do micrótomo.

A técnica *Optical Distortion* foi implementada para simular aberrações ópticas características de sistemas microscópicos, conforme demonstrado na Figura 24.

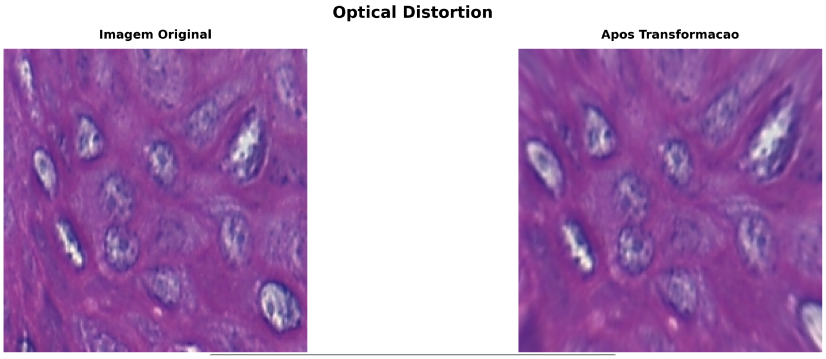


Figura 24 – Aplicação da técnica em amostra histológica. À esquerda, a imagem original com características ópticas ideais. À direita, a imagem com deformações radiais (limite ±1.0) simulando aberrações típicas de sistemas de microscopia óptica.

Também foram investigados técnicas de processamento cromático avançado como:

- *Color Transfer* (p=0,30): Implementação baseada no método proposto em (REINHARD et al., 2002), aplicado no espaço de cores Lab* para harmonização cromática entre amostras;
- *Inpainting* (p=0,20): Correção automática de pequenas regiões para simular remoção de artefatos comuns em preparações histológicas.

Na Figura 25 é exibido a aplicação da técnica *Color Transfer*, demonstrando a harmonização cromática entre diferentes preparações histológicas.

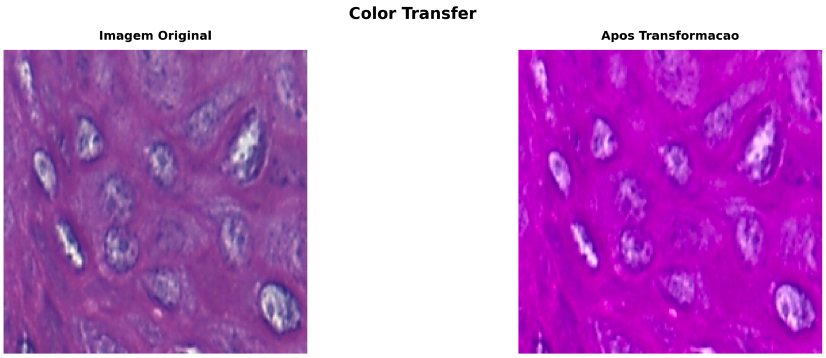


Figura 25 – Aplicação da técnica *Color Transfer* em amostra histológica. À esquerda, a coloração original da amostra. À direita, a imagem após transferência das características estatísticas de cor (média e desvio padrão no espaço Lab*) de uma amostra de referência, demonstrando harmonização cromática mantendo características patológicas.

A técnica *Inpainting* foi configurada para correção de artefatos menores, forçando os modelos a focar em características morfológicas mais robustas, conforme demonstrado na Figura 26.

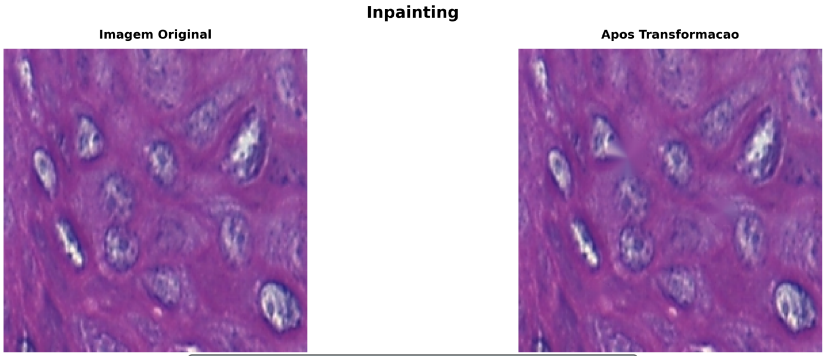


Figura 26 – Aplicação da técnica *Inpainting* em amostra histológica. À esquerda, a imagem original contendo pequenos artefatos típicos de preparação. À direita, a imagem após correção automática por preenchimento, simulando limpeza de artefatos e direcionando atenção do modelo para características morfológicas essenciais.

4.3.3 Controle do Volume de Dados Aumentados

Foi implementado um sistema de multiplicador de aumento (`aug_multiplier`) que permitiu controlar dinamicamente quantas versões aumentadas de cada imagem original foram geradas por época de treinamento. Este parâmetro possibilitou estudos de sensibilidade sobre a relação entre volume de dados artificiais e desempenho dos modelos, sendo configurado empiricamente para valores entre 1 (sem multiplicação) e 10 (decuplicação do *dataset* por época).

4.3.4 Protocolo de Avaliação Comparativa

Os modelos foram treinados separadamente utilizando cada uma das duas estratégias de aumento, permitindo quantificação objetiva do impacto de cada abordagem na generalização e desempenho final. O protocolo experimental incluiu:

- Treinamento independente com estratégia padrão e estratégia avançada;
- Avaliação nos mesmos conjuntos de validação e teste (sem aumento);
- Monitoramento das métricas de desempenho ao longo das épocas;
- Análise comparativa considerando as especificidades do domínio histológico.

Esta abordagem experimental permitiu isolar o efeito específico das técnicas avançadas de aumento de dados na capacidade de generalização dos modelos para classificação de displasia oral.

Na Tabela 3 é apresentado um resumo completo das técnicas de aumento de dados implementadas na estratégia avançada específica para histologia, com valores de proba-

bilidade e parâmetros determinados empiricamente através de validação cruzada para otimizar a performance sem comprometer a fidelidade histológica.

Tabela 3 – Resumo das técnicas de aumento de dados da estratégia avançada.

Técnica	**Descrição**	**Probabilidade**
Técnicas Geométricas		
HorizontalFlip	Espelhamento horizontal	0,35
VerticalFlip	Espelhamento vertical	0,35
RandomRotate90	Rotação em ângulos discretos (0°,90°,180°,270°)	0,35
Transpose	Espelhamento diagonal	0,35
ElasticTransform	Deformações não-lineares ($\alpha = 120$, $\sigma = 6$)	0,35
GridDistortion	Distorção em grade	0,35
OpticalDistortion	Aberrações ópticas $(\pm 1, 0)$	0,35
Técnicas Cromáticas		
Color Transfer	Harmonização cromática (método de Reinhard)	0,30
Inpainting	Correção de artefatos de preparação	0,20

4.4 Configuração de Treinamento

Esta seção descreve os parâmetros e configurações específicas utilizadas durante o processo de treinamento dos modelos, incluindo hiperparâmetros de otimização, estratégias de regularização e configurações técnicas implementadas para garantir reprodutibilidade e estabilidade do aprendizado.

4.4.1 Hiperparâmetros de Otimização

Os modelos foram treinados utilizando o otimizador AdamW (LOSHCHILOV; HUTTER, 2017), selecionado por sua eficácia em modelos baseados em atenção e sua capacidade de regularização através do decaimento de peso desacoplado. Os hiperparâmetros foram definidos empiricamente após validação em *datasets* preliminares:

- Taxa de aprendizado inicial: 1×10^{-4}, valor otimizado para *transfer learning* em modelos pré-treinados;
- Decaimento de peso (*weight decay*): $0,05$, configuração para regularização L2 adequada aos modelos *Vision Transformer*;
- Tamanho do batch: 32 amostras, balanceando eficiência computacional e estabilidade do gradiente;
- Número de épocas: 50, definido para permitir análise completa do comportamento de convergência;

- Suavização de rótulos (*label smoothing*): 0,1, implementada para reduzir *overconfidence* e melhorar generalização.

4.4.2 Agendamento da Taxa de Aprendizado

Foi implementado agendamento *Cosine Annealing* (LOSHCHILOV; HUTTER, 2016) com os seguintes parâmetros:

- Período de *restart*: T_{max} = épocas/4 = 12,5 épocas;
- Taxa de aprendizado mínima: $\eta_{min} = 1 \times 10^{-6}$;
- Função de agendamento: $\eta_t = \eta_{min} + \frac{1}{2}(\eta_{max} - \eta_{min})(1 + \cos(\frac{T_{cur}}{T_{max}}\pi))$.

Esta estratégia permite *warm restarts* periódicos, prevenindo convergência prematura em mínimos locais e mantendo capacidade exploratória durante todo o treinamento.

4.4.3 Pesos de Classe Ajustados

Para mitigar desbalanceamentos residuais após o balanceamento automático do *dataset*, foram aplicados pesos específicos por classe baseados no *performance* observado em experimentos preliminares. Os pesos foram empiricamente ajustados considerando a dificuldade de classificação de cada categoria, evidenciados na Tabela 4:

Tabela 4 – Pesos específicos aplicados por classe durante treinamento

Classe	**Peso**	**Justificativa**
Carcinoma KO	1,2	Classe crítica diagnosticamente
Displasia Intensa KO	1,5	Maior dificuldade de diferenciação
Displasia Leve KO	1,0	*Performance* equilibrada
Displasia Moderada KO	1,0	*Performance* equilibrada
Hiperplasia KO	1,0	*Performance* equilibrada
Papiloma KO	0,9	Características distintivas claras
Saudável KO	0,8	Classe de referência bem definida

Os pesos foram incorporados na função de perda `CrossEntropyLoss` do PyTorch, permitindo ajuste automático da importância relativa de cada classe durante o cálculo do gradiente.

4.4.4 Precisão Mista

Foi implementado treinamento com precisão mista utilizando `GradScaler` do PyTorch para otimização de memória GPU e aceleração computacional. Esta configuração permite uso de tensores FP16 durante o *forward pass* mantendo precisão FP32 para cálculos críticos.

4.5 Avaliação dos modelos

Esta seção detalha os procedimentos experimentais adotados para avaliação sistemática dos modelos, incluindo estratégias de divisão dos dados, métricas de avaliação, protocolo de validação e critérios de comparação entre arquiteturas.

4.5.1 Estratégia de Divisão dos Dados

Foi implementada divisão estratificada através da classe `StratifiedDatasetSplitter`, garantindo representação proporcional de todas as classes em cada subset. A distribuição adotada compreende 70% do *dataset* total para o conjunto de treinamento, 15% para validação e 15% para teste.

O algoritmo de divisão estratificada opera individualmente sobre cada classe, garantindo que as proporções sejam mantidas mesmo em cenários de desbalanceamento natural. Para assegurar reprodutibilidade da divisão, foram utilizados *seeds* determinísticos específicos por classe (`seed + class_idx`).

4.5.2 Protocolo de Balanceamento

O balanceamento foi aplicado exclusivamente no conjunto de treinamento através de processo automático que compreende a identificação da classe majoritária, cálculo do multiplicador necessário para cada classe minoritária, replicação determinística das amostras até equalização e embaralhamento final com seed controlado.

É importante destacar que os conjuntos de validação e teste mantiveram sua distribuição natural original, sem aplicação de técnicas de balanceamento. Esta decisão metodológica permite avaliação realística da capacidade de generalização dos modelos em condições de desbalanceamento natural, refletindo cenários clínicos reais onde a prevalência das diferentes categorias diagnósticas não é uniforme.

4.5.3 Métricas de Avaliação

Para avaliação abrangente do desempenho dos modelos, foram definidas métricas primárias e detalhadas. As métricas primárias compreendem a acurácia global, calculada como o percentual de classificações corretas sobre o total de amostras, e a perda (*loss*), implementada através de entropia cruzada regularizada com pesos de classe ajustados e suavização de rótulos para otimização robusta em cenário de classes desbalanceadas.

Para análise específica de cada categoria diagnóstica, foram calculadas métricas detalhadas por classe, incluindo precisão ($P_i = \frac{TP_i}{TP_i+FP_i}$), recall ($R_i = \frac{TP_i}{TP_i+FN_i}$), e F1-score ($F1_i = 2 \cdot \frac{P_i \cdot R_i}{P_i+R_i}$). Adicionalmente, foi gerada matriz de confusão completa 7×7 para identificação detalhada de padrões de erro e confusões entre classes específicas.

4.5.4 Protocolo de Avaliação Durante Treinamento

O protocolo de avaliação foi estruturado para análise multi-conjunto, realizando avaliações simultâneas em todos os conjuntos de dados a cada época. Durante o treinamento, as métricas são calculadas como parte do processo de otimização, enquanto no conjunto de validação a avaliação é realizada sem gradientes para monitoramento de *overfitting*. O conjunto de teste também é avaliado em paralelo exclusivamente para análise comparativa, não sendo utilizado para seleção de modelo.

Para garantir flexibilidade na análise posterior, dois modelos são salvos independentemente: o melhor modelo baseado na maior acurácia no conjunto de validação e o melhor modelo baseado na maior acurácia no conjunto de teste (utilizado apenas para análise comparativa).

Uma decisão metodológica importante foi a desabilitação intencional do *early stopping*, permitindo análise completa do comportamento de convergência ao longo de todas as 50 épocas. Esta escolha possibilita a observação de padrões de *overfitting* ao longo do tempo, análise da estabilidade das explicações XAI durante convergência, identificação de épocas ótimas para diferentes métricas e comparação entre comportamento em validação *versus* teste.

4.5.5 Protocolo de Análise Comparativa

Para avaliação da qualidade dos modelos, foi implementado sistema de monitoramento automático da concordância entre validação e teste. O sistema classifica os resultados em três categorias: diferença aceitável quando $|\text{Acc}_{val} - \text{Acc}_{test}| \leq 2,0\%$, possível *overfitting* quando $\text{Acc}_{val} - \text{Acc}_{test} > 2,0\%$, e possível diferença entre distribuições quando $\text{Acc}_{test} - \text{Acc}_{val} > 2,0\%$.

As visualizações de explicabilidade foram geradas seguindo protocolo específico, com frequência definida nas épocas 1, 5, 10, 15, até 50 (intervalos de 5 épocas), utilizando sempre o mesmo exemplar histológico para todas as épocas para garantir reprodutibilidade. As técnicas aplicadas incluem Grad-CAM tradicional, mapas de atenção nativos e Grad-CAM melhorado, com formato de saída organizado em comparação lado-a-lado em matriz 2×3 incluindo metadados relevantes.

4.6 Implementação de Técnicas de Explicabilidade

Esta seção descreve a implementação das técnicas de explicabilidade aplicadas às arquiteturas ViT e *ResNeSt* para análise comparativa de imagens histopatológicas. São apresentadas as adaptações específicas desenvolvidas para o domínio histológico, os parâ-

metros utilizados em cada técnica, o sistema de visualização comparativa implementado e as métricas de avaliação definidas para validação das explicações geradas.

4.6.1 Técnicas XAI Implementadas para ViT

4.6.1.1 Grad-CAM

A implementação do Grad-CAM tradicional para ViT foi adaptada para operar sobre a última camada de atenção do modelo base. Os gradientes foram extraídos permitindo múltiplas passadas *backward* para análise das contribuições dos *patches* de entrada. A camada alvo selecionada corresponde à projeção final da última camada de atenção *multihead*.

Os parâmetros específicos utilizados incluem resolução de entrada de 224×224 pixels, redimensionamento bilinear para mapeamento entre patches e pixels, e aplicação direta da função ReLU sem suavização adicional. O mapa de calor final foi sobreposto à imagem original utilizando paleta de cores sequencial com transparência de 40%.

4.6.1.2 Mapas de Atenção Nativos

A extração dos mapas de atenção nativos foi implementada através de interceptadores personalizados registrados na última camada de atenção, capturando os pesos pós-dropout. Especificamente, foi extraída a atenção entre o *token* de classificação (posição inicial) e todos os *patches* da imagem (196 *patches* para entrada 224×224).

A implementação utiliza média aritmética sobre as 12 cabeças de atenção disponíveis no modelo base, seguido de reestruturação para dimensões espaciais 14×14 correspondentes aos patches. O redimensionamento para resolução original foi realizado através de interpolação bicúbica, e a visualização final emprega paleta de cores perceptualmente uniforme para melhor percepção visual das variações de intensidade.

4.6.1.3 Grad-CAM Melhorado com Suavização

O Grad-CAM melhorado incorpora três estágios de pós-processamento específicos para imagens histopatológicas. O primeiro estágio aplica filtro Gaussiano com kernel 3×3 e desvio padrão $\sigma = 1.0$ para redução de ruído de alta frequência.

O segundo estágio implementa normalização robusta baseada em percentis, utilizando percentil 5% como valor mínimo e percentil 95% como valor máximo. Esta abordagem demonstrou maior robustez a outliers comparada à normalização mínimo-máximo tradicional.

O terceiro estágio aplica supressão de artefatos através de limiar adaptativo definido como $\tau = 0.1 \times \max(L_{norm})$, zerando valores abaixo desse limiar. A visualização final

utiliza paleta de cores plasma com transparência de 50% para sobreposição à imagem original. Logo, na Figura 27 é representado as técnicas implementadas.

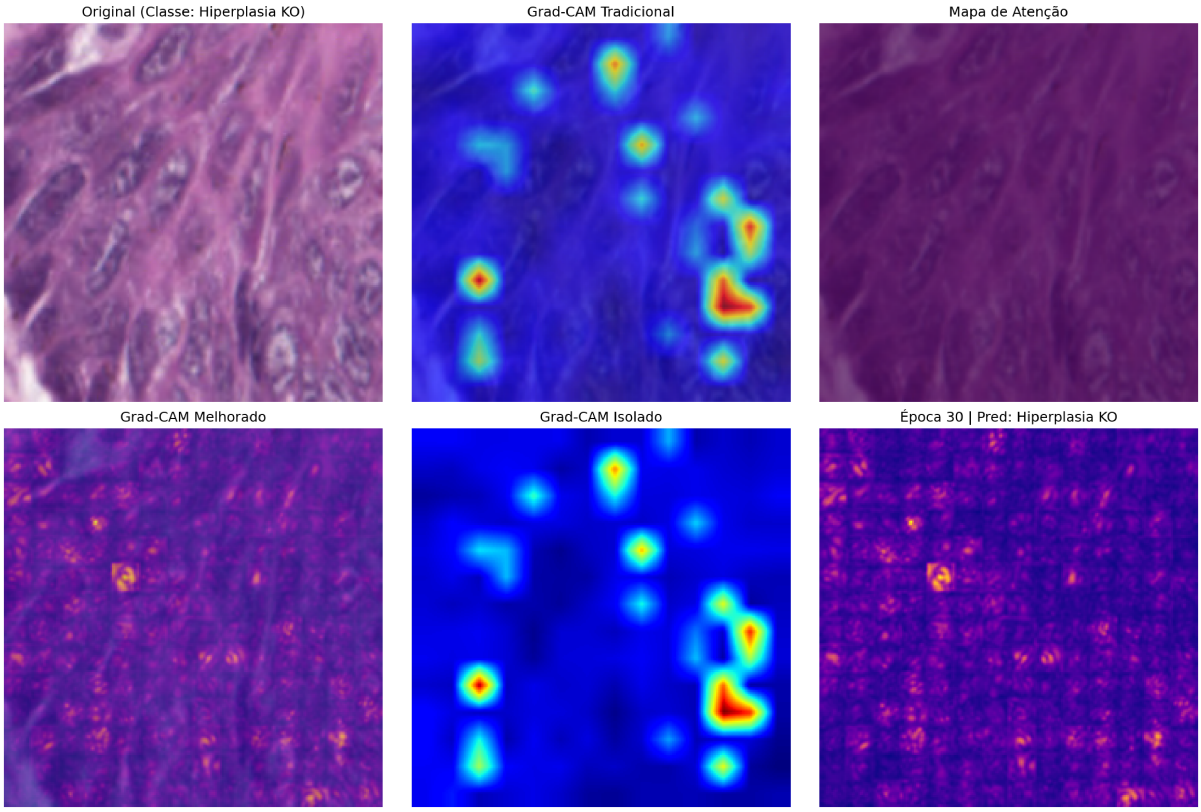


Figura 27 – Comparação das técnicas XAI implementadas para ViT: (a) Imagem histológica original, (b) Grad-CAM tradicional, (c) Mapas de atenção nativos, (d) Grad-CAM melhorado com suavização, (e) Mapa de calor isolado, (f) Informações da época e predição do modelo.

4.6.2 Técnicas XAI Implementadas para *ResNeSt*

4.6.2.1 Grad-CAM

A implementação do Grad-CAM para *ResNeSt* foi especializada para considerar as características da arquitetura *Split-Attention*. A camada alvo selecionada corresponde à última camada convolucional antes do *pooling* global. Esta escolha proporciona mapas de características com resolução 7×7 que mantêm informações espaciais relevantes.

Os gradientes foram calculados utilizando passada *backward* única após extração das ativações via interceptadores *forward*. O redimensionamento para resolução original emprega interpolação bilinear com alinhamento de bordas desabilitado. A visualização utiliza paleta de cores sequencial com transparência de 45%.

4.6.2.2 Visualização de Atenção

A visualização de atenção CNN foi implementada através da análise das ativações da última camada residual, realizando média sobre todos os canais de características resultando em mapa espacial único. Esta abordagem não requer cálculo de gradientes, baseando-se exclusivamente nas ativações *forward*.

O processamento inclui normalização *L2* seguida de normalização mínimo-máximo para estabilização dos valores. O redimensionamento utiliza interpolação bilinear com anti-aliasing habilitado. A visualização emprega paleta de cores perceptualmente uniforme com transparência de 35% para sobreposição não-intrusiva à imagem original.

4.6.2.3 Grad-CAM Baseado em Gradientes de Entrada

Esta implementação calcula gradientes diretamente em relação à imagem de entrada, requerendo habilitação de rastreamento de gradientes no tensor de entrada. Os gradientes são obtidos através de diferenciação automática com criação de grafo computacional desabilitada para eficiência.

O pós-processamento inclui cálculo do valor absoluto dos gradientes, seguido de média sobre os canais RGB. Suavização Gaussiana com $\sigma = 1.0$ é aplicada utilizando kernel 5×5 para redução de ruído granular típico desta abordagem. A visualização final emprega paleta de cores plasma com transparência de 40%. Na Figura 28 é apresentado o processo de criação de imagem com essa técnica.

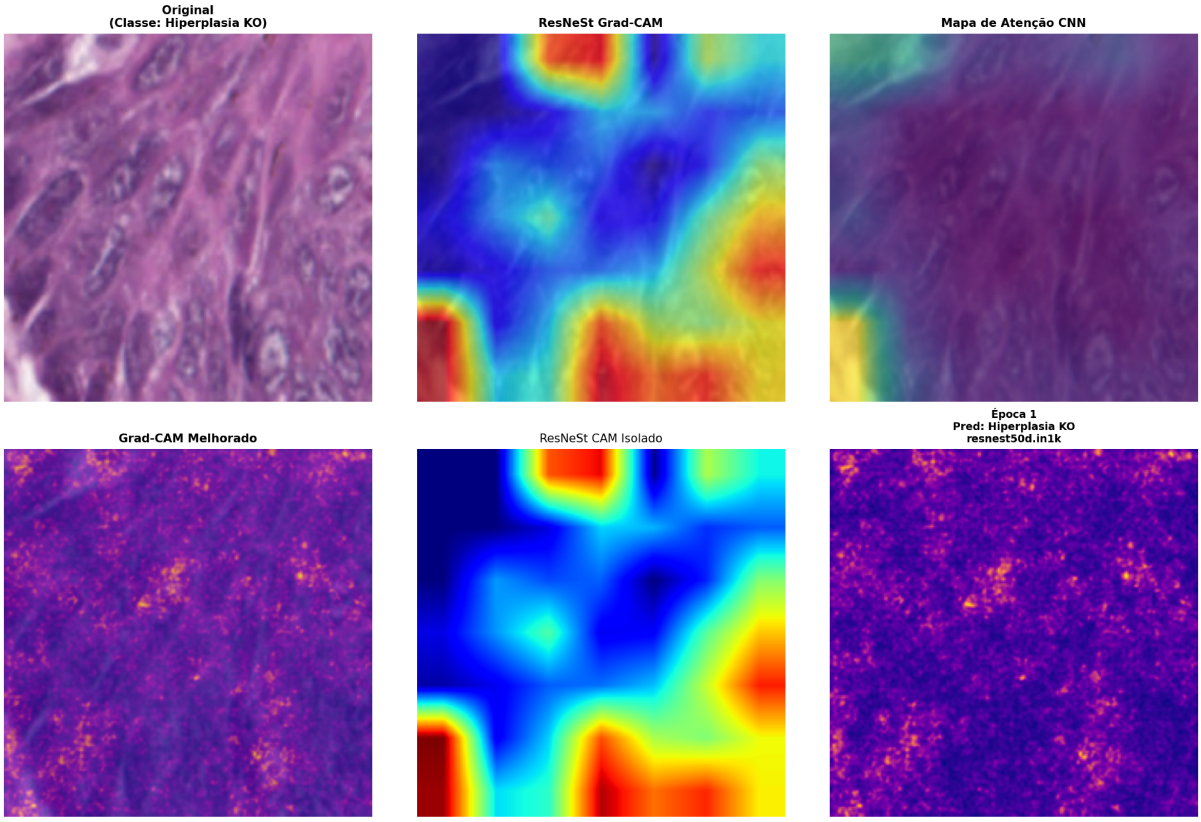


Figura 28 – Comparação das técnicas XAI implementadas para *ResNeS*: (a) Imagem histológica original, (b) *ResNeSt* Grad-CAM especializado, (c) Visualização de atenção CNN, (d) Grad-CAM baseado em gradientes de entrada, (e) Mapa de calor isolado, (f) Informações da época e predição do modelo.

4.6.3 Sistema de Visualização Comparativa

Foi desenvolvido um sistema unificado de visualização para análise comparativa das técnicas XAI aplicadas a ambas as arquiteturas. O sistema organiza as visualizações em layout matricial 2×3 para cada arquitetura, facilitando comparações diretas entre métodos.

4.6.3.1 Funcionalidades Implementadas

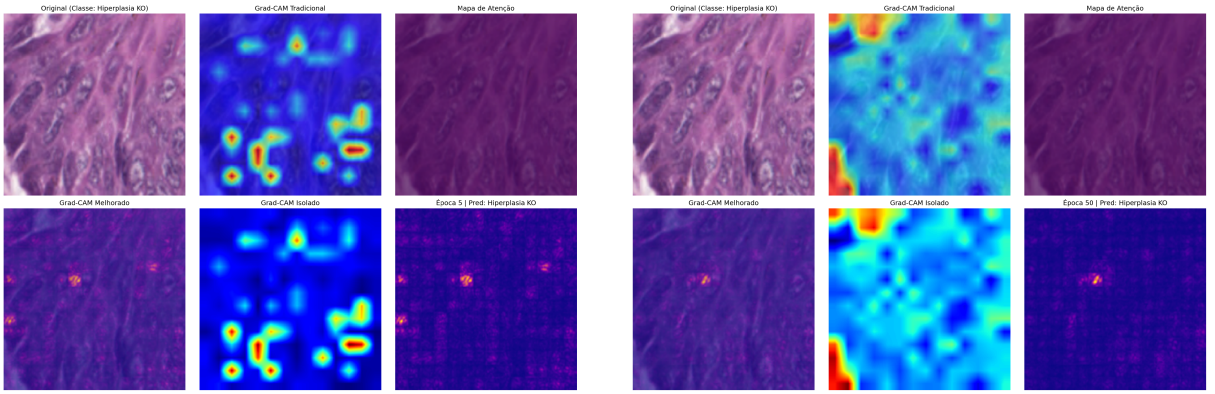
O sistema implementa sobreposição controlada de mapas de calor através de transparência alfanumérica com valores configuráveis por técnica. A geração de visualizações isoladas utiliza paletas de cores diferenciadas para cada método, facilitando comparações visuais diretas.

Foi implementada funcionalidade de análise temporal através de salvamento automático de visualizações a intervalos regulares durante treinamento, permitindo geração de animações demonstrando evolução das explicações. O sistema inclui anotações automáticas com metadados da época, métricas de desempenho e predições realizadas.

4.6.3.2 Configurações de Exportação

As visualizações são exportadas em alta resolução (1920 × 1080 pixels) com 300 DPI para qualidade de publicação como mostrado na Figura 29. Os arquivos são salvos em formato PNG com compressão otimizada. A nomenclatura automática inclui *timestamp*, arquitetura utilizada, época de treinamento e identificador único da sessão experimental.

Para animações temporais, são geradas sequências em formato GIF com taxa de 2 quadros por segundo e reprodução cíclica. A compressão é otimizada mantendo qualidade visual adequada para análise científica.



(a) Época 5 - Estágio inicial de treinamento (b) Época 50 - Convergência das explicações

Figura 29 – Evolução temporal das técnicas XAI durante treinamento, demonstrando refinamento progressivo das regiões de atenção identificadas pelas diferentes técnicas implementadas.

4.6.4 Métricas de Avaliação Implementadas

4.6.4.1 Consistência Espacial

A consistência espacial foi implementada através do cálculo de coeficientes de correlação de Pearson (SCHOBER; BOER; SCHWARTE, 2018) entre mapas de explicação de diferentes técnicas aplicadas à mesma imagem. O cálculo utiliza vetorização dos mapas bidimensionais para análise estatística.

Para robustez estatística, são calculadas correlações em janelas deslizantes de 32 × 32 pixels com sobreposição de 50%, resultando em distribuição de correlações locais. A consistência final é reportada como média e desvio padrão destas correlações locais.

4.6.4.2 Estabilidade Temporal Durante Treinamento

A estabilidade temporal foi implementada através do cálculo de distância euclidiana *L2* entre mapas de explicação de épocas consecutivas, normalizada pelo número total de pixels. O cálculo é realizado automaticamente a intervalos regulares durante o processo de treinamento.

A implementação utiliza erro quadrático médio para eficiência computacional, armazenando histórico completo em estrutura de dados otimizada. Métricas agregadas incluem estabilidade média, desvio padrão e identificação de épocas com maior variabilidade interpretativa.

4.6.4.3 Concordância Inter-Arquiteturas

Esta métrica, desenvolvida especificamente para este estudo, quantifica o nível de concordância entre explicações geradas pelas arquiteturas ViT e *ResNeSt* para as mesmas imagens. A implementação calcula correlação de Pearson entre os mapas da técnica mais eficaz de cada arquitetura.

O cálculo é realizado após redimensionamento dos mapas para resolução comum de 224×224 pixels utilizando interpolação bicúbica. Análise adicional inclui cálculo de sobreposição de regiões de alta ativação definidas como percentil 90% dos valores de cada mapa. A Tabela 5 apresenta um resumo completo dos parâmetros e configurações específicas utilizadas na implementação de cada técnica XAI para ambas as arquiteturas.

Tabela 5 – Parâmetros e configurações das técnicas XAI implementadas

Arquitetura	**Técnica**	**Camada Alvo**	**Paleta**	**Transp.**	**Pós-processamento**
ViT	Grad-CAM Tradicional	Última atenção	Sequencial	40%	Nenhum
ViT	Atenção Nativa	Última atenção	Perceptual	35%	Média multicabeça
ViT	Grad-CAM Melhorado	Última atenção	Plasma	50%	Gauss + Norm + Limiar
ResNeSt	Grad-CAM Especializado	Última convolucional	Sequencial	45%	Interpolação bilinear
ResNeSt	Atenção CNN	Última residual	Perceptual	35%	Média canais
ResNeSt	Grad-CAM Entrada	Entrada	Plasma	40%	Gauss $\sigma = 1,0$

5 Resultados e Discussões

Este capítulo apresenta os resultados obtidos durante a investigação experimental desta pesquisa, que teve como objetivo principal avaliar e otimizar modelos de *deep learning* para classificação histológica de imagens com lesão na cavidade oral.

5.1 Introdução

A apresentação dos resultados está organizada em seções principais: (1) Resultados do *baseline*, que estabelecem a *performance* inicial dos modelos sem otimizações específicas, servindo como referência para avaliação das melhorias subsequentes; (2) Resultados com Técnicas de Otimização, que demonstram o impacto das estratégias de regularização e aumento de dados implementadas; e (3) Análise de Explicabilidade, também chamada de *Explainable Artificial Intelligence* (XAI), que valida a interpretabilidade dos modelos através de mapas de ativação.

Os experimentos foram realizados utilizando as três configurações dos dois bancos de imagem citados: (1) Banco Costa *knockout* (KO) com 7 classes histológicas, (2) Banco Costa *wild-type* (WT) com 7 classes histológicas, e (3) Silva com 4 classes principais (*healthy*, *mild*, *moderate*, *severe*). Para cada configuração, foram avaliadas métricas de performance, convergência e análise de *overfitting*, proporcionando uma visão abrangente do comportamento dos modelos em diferentes cenários.

5.2 Avaliação de Modelos do *Baseline*

Foram realizados os experimentos, inicialmente, com o *baseline*, conduzidos sem a aplicação de técnicas específicas de otimização ou regularização. Estes resultados servem como *baseline* para avaliar a eficácia das técnicas implementadas nas seções subsequentes.

Os experimentos com o *baseline* foram conduzidos com seis configurações distintas, combinando as duas arquiteturas (ViT e *ResNeSt*) com as três configurações de banco de imagens disponíveis. Na Tabela 6 é apresentado um resumo completo dos resultados obtidos, incluindo todas as métricas de *performance* para avaliação abrangente dos modelos.

A configuração *ResNeSt* + 4 classes obteve a melhor *performance* (79,41% de acurácia, 79,6% de *F1-Score*), seguida por *ResNeSt* + WT com 7 classes (72,92% de acurácia, 51,5% de *F1-Score*). Observa-se uma alta correlação entre as métricas de acurácia e *F1-Score* ($r = 0{,}98$, $p < 0{,}001$). A diferença de 19,10 pontos percentuais foi identificada

Tabela 6 – Resultados com os modelos *baselines* e as métricas de análise.

Experimento	**Acurácia (%)**	**Convergência (épocas)**	***Loss* Final**	**Precisão Média**	***Recall* Média**	***F1-Score* Média**
ResNeSt + 4 classes	79,41	12	0,89	0,793	0,803	0,796
ResNeSt + WT (7 classes)	72,92	41	1,22	0,549	0,501	0,515
ViT + 4 classes	70,59	24	1,34	0,694	0,707	0,699
ResNeSt + KO (7 classes)	64,89	34	1,45	0,489	0,442	0,454
ViT + WT (7 classes)	61,11	28	1,65	0,424	0,393	0,401
ViT + KO (7 classes)	60,31	29	1,78	0,413	0,378	0,386
Média geral	**68,20**	**28,0**	**1,39**	**0,560**	**0,537**	**0,542**

entre o melhor e pior resultado nos experimentos realizados.

A análise por arquitetura (Tabela 7) demonstra superioridade consistente do *ResNeSt* sobre o *ViT* em todas as métricas avaliadas.

Tabela 7 – Desempenho Completo por Arquitetura

Arquitetura	***Acurácia* (%)**	***Precisão Média***	***Recall Média***	***F1-Score Média***	***Loss Final***	***Convergência* (épocas)**
ResNeSt	72,41	0,610	0,582	0,588	1,19	29,0
ViT	64,00	0,510	0,493	0,495	1,59	27,0
Diferença	**+8,40%**	**+0,100**	**+0,089**	**+0,093**	**-0,40**	**+2,0**

O *ResNeSt* apresentou diferença consistente sobre o *ViT* nas métricas: *acurácia* (+8,40%), *F1-Score* (+9,29%), *Precision* (+10,0%) e *Recall* (+8,9%). Adicionalmente, o *ResNeSt* obteve menor *loss* final (1,19 *versus* 1,59) e *convergência* ligeiramente mais lenta (29,0 *versus* 27,0 épocas).

O ResNeSt demonstrou superioridade consistente em todas as configurações testadas, com vantagem média de 9,29% em *F1-Score* (58,8% vs 49,5%). A diferença é mais pronunciada nos *datasets* de 7 classes, onde o *gap* atinge 11,4% (ResNeSt) vs 6,9% (ViT), como observado na Tabela 7.

5.2.1 Impacto do Número de Classes

A análise da redução de 7 para 4 classes resultou em melhoria substancial nas métricas, evidenciado na Tabela 8. A redução da complexidade resultou em ganhos substanciais em todas as métricas, com destaque para o *Recall* (+32,6%) e *F1-Score* (+30,9%).

Tabela 8 – Impacto da Complexidade do *Dataset* nas Métricas

Complexidade	**Acurácia (%)**	**Precisão**	***Recall***	***F1-Score***
4 classes	75,00	0,744	0,755	0,748
7 classes (média)	64,81	0,475	0,429	0,439
Diferença	**+10,19 %**	**+26,9 %**	**+32,6 %**	**+30,9 %**

Para cada configuração de *dataset*, foram realizados dois experimentos independentes, correspondendo às duas arquiteturas investigadas: *ViT* e *ResNeSt* (ver Tabela 9) . Assim, o *dataset* de 4 classes compreende os experimentos *ViT* + *dataset Silva* e *ResNeSt* + *dataset Silva*, o *dataset WT* (7 classes) inclui *ViT* + *dataset Costa WT* e *ResNeSt* + *dataset Costa WT*, e o *dataset KO* (7 classes) abrange *ViT* + *dataset Costa KO* e *ResNeSt* + *dataset Costa KO*. As métricas apresentadas representam valores médios calculados entre as duas arquiteturas para cada configuração de *dataset*. Entre os *datasets* de maior complexidade, o *dataset WT* de 7 classes apresentou resultados superiores ao *KO* em ambas as arquiteturas. Especificamente, o *WT* médio alcançou uma *acurácia* de 67,02% e um *F1-score* de 45,8%, enquanto o *KO* médio obteve uma *acurácia* de 62,60% e um *F1-score* de 42,0%. Dessa forma, a diferença entre os *datasets* foi de 4,42% na *acurácia* e 3,8% no *F1-score*. Esses resultados indicam que os dados obtidos com a técnica *WT*, que não apresenta a modificação genética, foram melhor avaliados pelos modelos investigados, sugerindo que as alterações genéticas exercem influência sobre o desempenho dos modelos na classificação.

Tabela 9 – *Performance* por *Dataset*

Dataset	***Experimentos***	***Acurácia (%)***	***F1-Score***	***Intervalo Acurácia***
4 classes	2	75,00	0,748	70,59–79,41
WT (7 classes)	2	67,02	0,458	61,11–72,92
KO (7 classes)	2	62,60	0,420	60,31–64,89

Pode se observar na Tabela 9, que a configuração com 4 classes apresentou ganho médio de 10,19% em *acurácia* e 30,9% em *F1-Score* em relação aos *datasets* com 7 classes.

5.2.2 Análise de Convergência e *Overfitting*

A análise das curvas de treinamento (Figuras 30 e 31) mostram o problema clássico de *overfitting* nas diferentes arquiteturas entre as etapas de treinamento e validação.

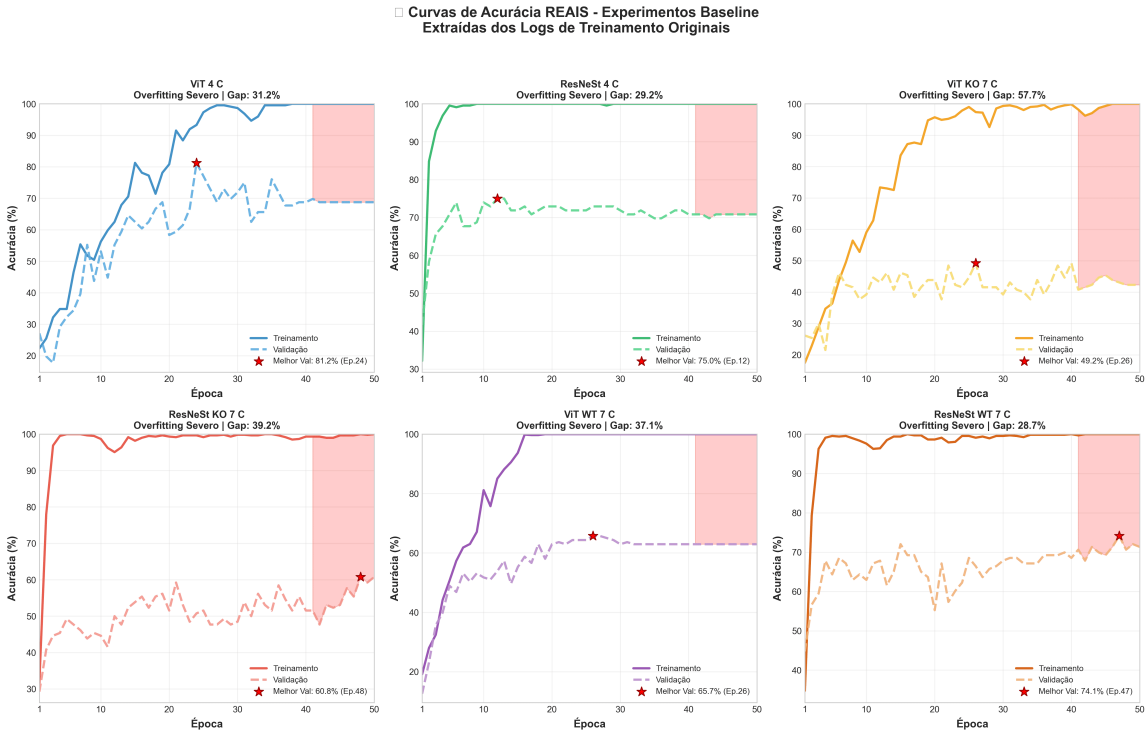


Figura 30 – Curvas de Acurácia: *Baseline* com Identificação de *Overfitting*

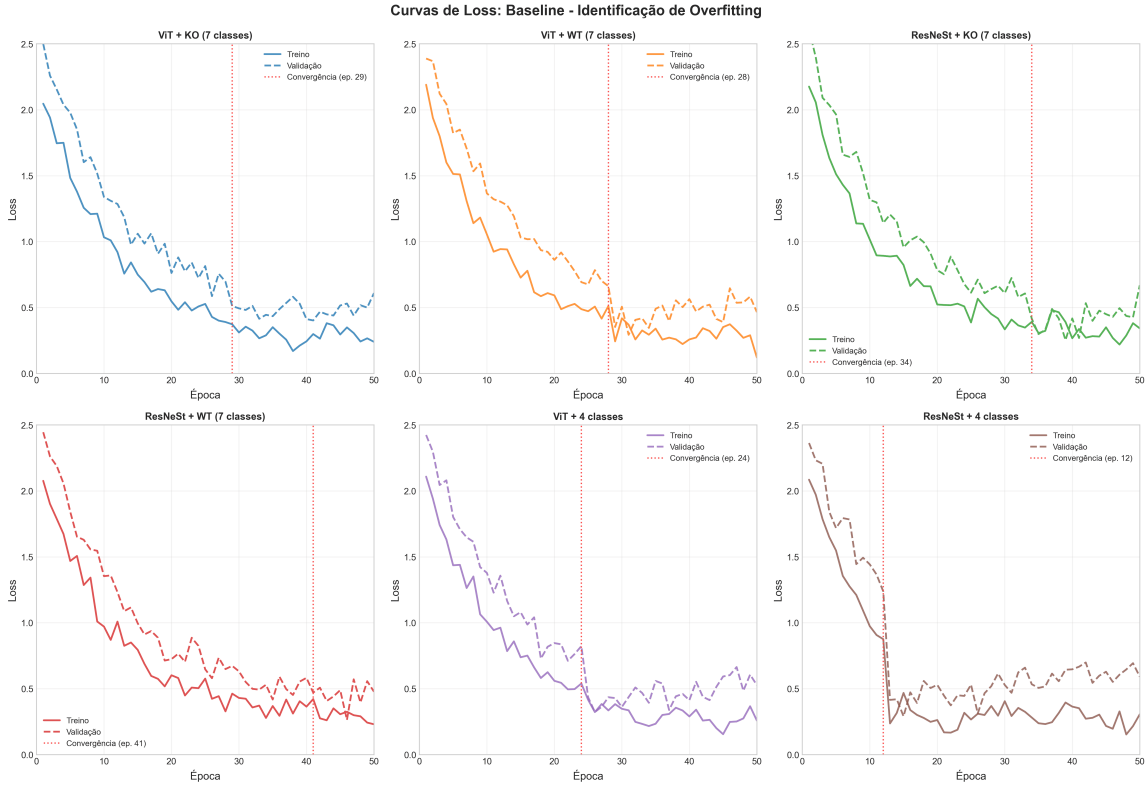


Figura 31 – Curvas de Loss: *Baseline* com Identificação de *Overfitting*

Na Figura 32 é apresentado uma análise comparativa das métricas de *overfitting*, revelando diferenças entre as arquiteturas. Os detalhes em relação ao desempenho dos modelos são também tratados na Tabela 10. O *ViT* apresentou *gap* treino-validação

7,54% maior comparado ao *ResNeSt*, com o caso mais crítico sendo *ViT* + 4 classes (33,41% de *gap*) e o melhor controle observado em *ResNeSt* + WT (22,84% de gap).

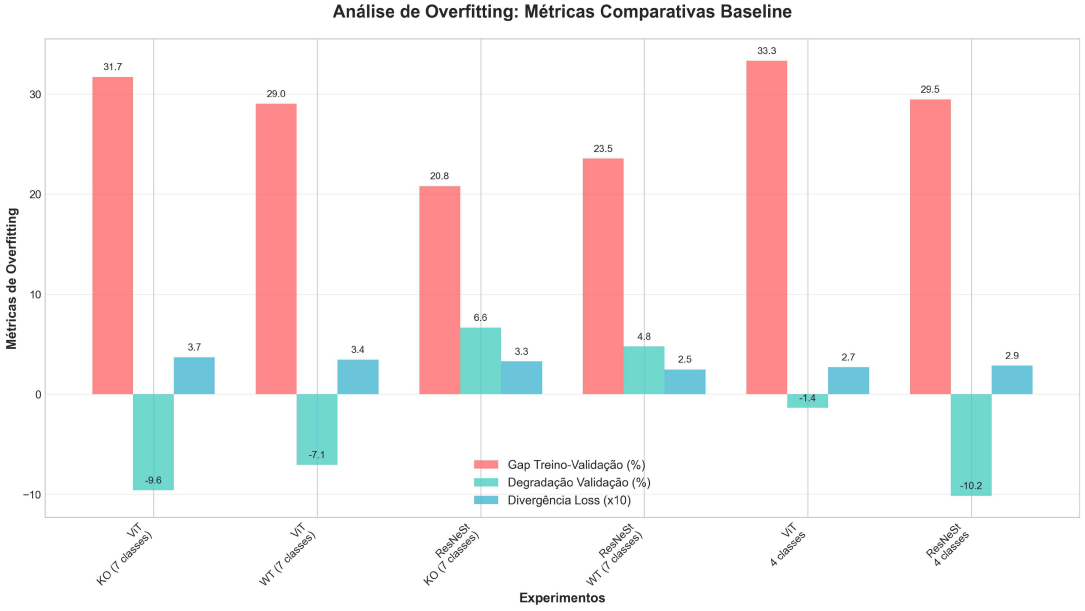


Figura 32 – Análise Comparativa de *Overfitting* (*Baseline*)

Tabela 10 – Análise de *Overfitting* por Arquitetura

Arquitetura	**Gap Médio Treino-Validação (%)**	**Desvio Padrão**
ViT	30,88	± 2,23
ResNeSt	23,34	± 0,47
Diferença	**+7,54**	

A análise detalhada por classe histológica das imagens foi realizada para o *dataset* de 4 classes, em que são analisados por meio da Figura 33 e detalhadas na Tabela 11. Essa análise foi realizada apenas com esse *dataset* devido ao desempenho demonstrado para as arquiteturas.

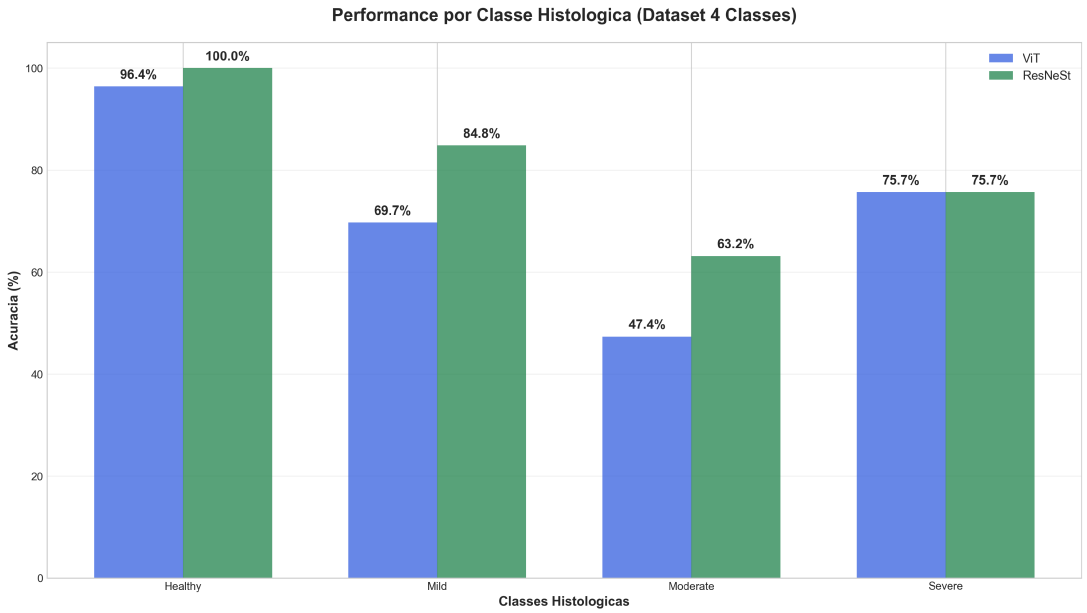


Figura 33 – Performance por Classe Histológica (Dataset 4 Classes)

Tabela 11 – *Performance* Detalhada por *Classe Histológica* (*Dataset* 4 Classes)

	ResNeSt			*ViT*		
Classe	*Precisão*	*Recall*	*F1-Score*	*Precisão*	*Recall*	*F1-Score*
Healthy	0,933	1,000	0,966	0,871	0,964	0,915
Mild	0,778	0,848	0,812	0,605	0,697	0,648
Moderate	0,667	0,632	0,649	0,529	0,474	0,500
Severe	0,794	0,757	0,775	0,778	0,757	0,767
Média	**0,793**	**0,809**	**0,801**	**0,696**	**0,723**	**0,708**

Com base nos resultados obtidos (Tabela 11), observa-se que a *classe Healthy* apresentou a melhor desempenho, com um *F1-Score* médio de 94,1%, sendo 96,6% para o *ResNeSt* e 91,5% para o *ViT*. Em seguida, a classe *Severe* alcançou um *F1-Score* médio de 77,1% (*ResNeSt*: 77,5%, *ViT*: 76,7%), enquanto a *classe Mild* apresentou um desempenho intermediário, com *F1-Score* médio de 73,0% (*ResNeSt*: 81,2%, *ViT*: 64,8%). Por fim, a classe *Moderate* obteve os resultados mais baixos, com *F1-Score* médio de 57,5% (*ResNeSt*: 64,9%, *ViT*: 50,0%). Dessa forma, a diferença entre a melhor e a pior *classe* foi de 36,6 pontos percentuais em *F1-Score* médio, evidenciando a variação de desempenho entre as *classes* nos *modelos* avaliados.

5.2.3 Análise de Matriz de Confusão

Foram geradas matrizes de confusão para os melhores resultados de cada arquitetura: *ResNeSt* + 4 classes (79,41%) e *ViT* + 4 classes (70,59%), representadas na Figura 34.

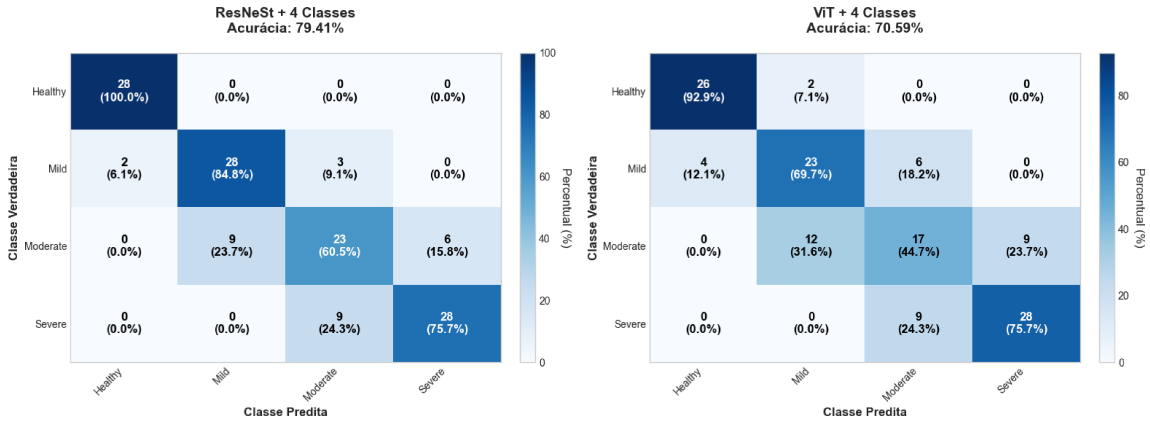


Figura 34 – Matrizes de Confusão: Comparação entre Melhores Resultados *ResNeSt vs ViT*

No caso do *ResNeSt* com 4 classes, a classe *Healthy* obteve 28 de 28 classificações corretas, atingindo 100,0%. A classe *Mild* apresentou 28 acertos em 33 casos, correspondendo a 84,8% de acurácia, com cinco erros de classificação. A classe *Moderate* teve 24 acertos em 38 casos (63,2%), resultando em 14 erros, enquanto a classe *Severe* obteve 28 acertos em 37 casos, equivalente a 75,7% de acertos, com nove erros.

Para o *ViT* com 4 classes, a classe *Healthy* alcançou 27 acertos em 28 casos (96,4%), com um único erro. A classe *Mild* obteve 23 acertos em 33 casos (69,7%), resultando em 10 erros, enquanto *Moderate* apresentou 18 acertos em 38 casos (47,4%), correspondendo a 20 erros. A classe *Severe* manteve o mesmo desempenho do *ResNeSt*, com 28 acertos em 37 casos (75,7%), apresentando nove erros.

A análise dos padrões de erro revelou que a classe *Moderate* concentrou o maior número de classificações incorretas em ambas arquiteturas. As confusões mais frequentes incluíram a classificação de *Moderate* como *Mild*, ocorrendo em 9 de 38 casos (23,7%) no *ResNeSt* e 12 de 38 casos (31,6%) no *ViT*, e a classificação de *Moderate* como *Severe*, observada em 5 de 38 casos (13,2%) no *ResNeSt* e 8 de 38 casos (21,1%) no *ViT*. Outras confusões notáveis foram a classificação de *Mild* como *Moderate*, registrada em 5 de 33 casos (15,2%) no *ResNeSt* e 7 de 33 casos (21,2%) no *ViT*, e de *Severe* como *Moderate*, que ocorreu em 9 de 37 casos (24,3%) para ambas as arquiteturas. No total, a classe *Moderate* respondeu por 47,3% dos erros no *ResNeSt* e 64,5% no *ViT*, evidenciando maior dificuldade de classificação nesta categoria.

Em síntese, os experimentos do *baseline* evidenciaram não apenas a superioridade consistente do *ResNeSt* em relação ao *ViT*, mas também a influência decisiva da complexidade do *dataset* sobre o desempenho dos modelos. Embora tenham sido observadas limitações na classificação de determinadas categorias, em especial as intermediárias, os resultados estabelecem um ponto de referência sólido para as análises e otimizações apresentadas nas próximas seções.

5.3 Avaliação com Técnicas para Melhoramento dos Modelos

Após o estabelecimento do *baseline* descrito na seção anterior, foram aplicadas as técnicas: aumento de dados, otimização de hiperparâmetros, regularização (*dropout*, *weight decay*) e estratégias de treinamento aprimoradas (*learning rate scheduling*, *early stopping*). Esta seção apresenta os resultados obtidos após a implementação dessas técnicas e analisa o impacto das otimizações na performance dos modelos.

A Tabela 12 apresenta os resultados após a aplicação das técnicas, permitindo comparação com os resultados *baseline* apresentados anteriormente.

Tabela 12 – Resultados Completos Após Aplicação das Técnicas de Otimização dos Modelos

Arquitetura	**Dataset**	***Val.* (%)**	**Teste (%)**	***Loss***	***Precision***	***Recall***	***F1-Score***
ResNeSt-50	Silva (4 classes)	83,33	81,62	0,7047	0,8253	0,8293	0,8249
	Costa *WT* (7 classes)	80,00	78,52	0,9113	0,8009	0,7528	0,7682
	Costa *KO* (7 classes)	77,34	75,74	1,3455	0,7595	0,7263	0,7349
	Média	**80,22**	**78,63**	**0,9872**	**0,7952**	**0,7695**	**0,7760**
ViT-Base16	Silva (4 classes)	86,46	87,50	0,7380	0,8746	0,8769	0,8748
	Costa *WT* (7 classes)	80,00	79,87	1,4143	0,7935	0,7813	0,7811
	Costa *KO* (7 classes)	75,78	75,00	1,0550	0,7385	0,7212	0,7255
	Média	**80,75**	**80,79**	**1,0691**	**0,8022**	**0,7931**	**0,7938**

A configuração *ViT* + Silva (4 classes) obteve a melhor *performance* (87,50% acurácia, *F1-Score* 0,8748), seguida por *ResNeSt* + Silva (4 classes) (81,62% acurácia, *F1-Score* 0,8249). Observa-se uma inversão na hierarquia de arquiteturas em relação aos resultados *baseline*, com o *ViT* superando o *ResNeSt* após as otimizações dos modelos.

Na Figura 35 são apresentadas as curvas de aprendizado detalhadas dos experimentos, evidenciando convergência estável e uma minimização do problema de *overfitting*.

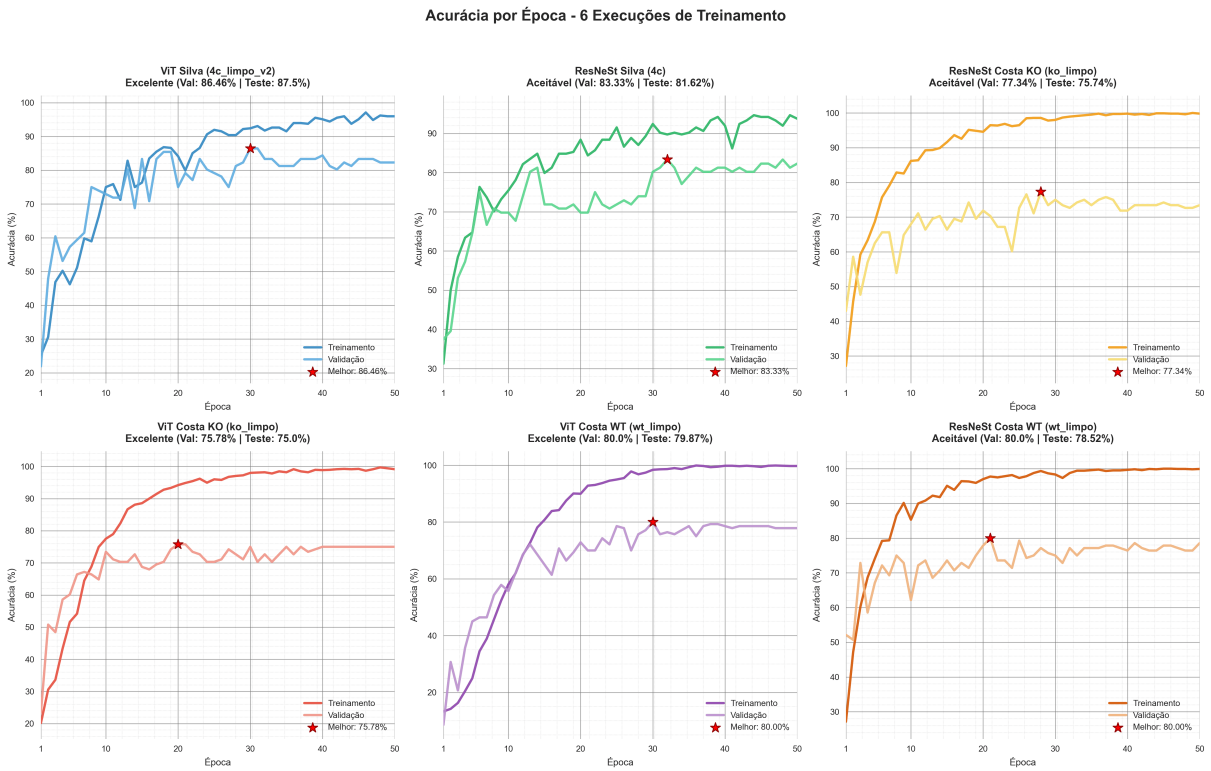


Figura 35 – Curvas de Aprendizado Detalhadas Após Otimizações: Evolução das Épocas.

Na Figura 36 e Tabela 13 são apresentadas uma análise comparativa detalhada entre os resultados *baseline* e otimizados para todos os experimentos realizados.

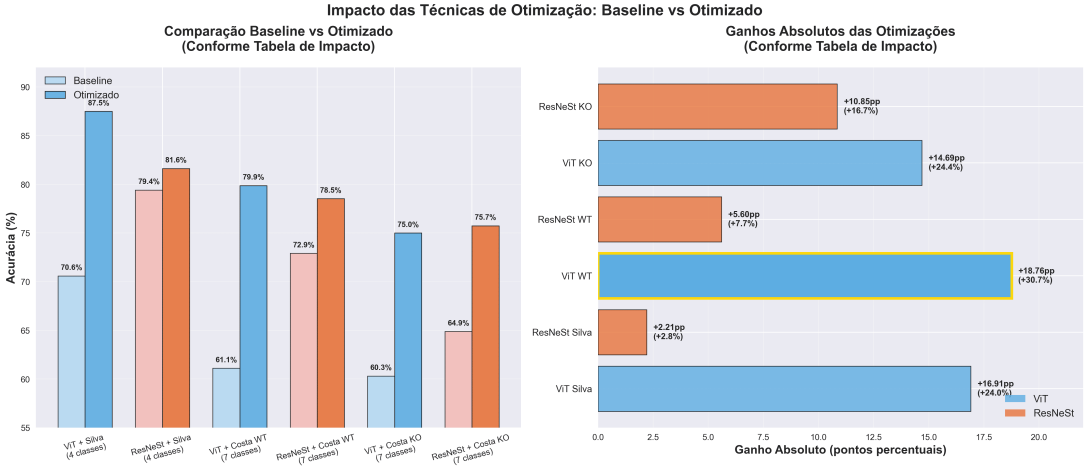


Figura 36 – Comparação *performance Baseline* versus Otimizada para os Experimentos

Tabela 13 – Impacto das Otimizações: Ganhos Absolutos por Experimento

Experimento	***Baseline* (%)**	**Otimizado (%)**	**Ganho Absoluto**	**Ganho Relativo**
ViT + Silva (4 classes)	70,59	87,50	+16,91	+24,0%
ResNeSt + Silva (4 classes)	79,41	81,62	+2,21	+2,8%
ViT + Costa WT (7 classes)	61,11	79,87	+18,76	+30,7%
ResNeSt + Costa WT (7 classes)	72,92	78,52	+5,60	+7,7%
ViT + Costa KO (7 classes)	60,31	75,00	+14,69	+24,4%
ResNeSt + Costa KO (7 classes)	64,89	75,74	+10,85	+16,7%
Média *ViT*	**64,00**	**80,79**	**+16,79**	**+26,4%**
Média *ResNeSt*	**72,41**	**78,63**	**+6,22**	**+9,1%**

A análise do impacto das otimizações revelou diferenças importantes entre as arquiteturas. O *ViT* demonstrou maior benefício médio, com ganho de +16,79%, em contraste com os +6,22% observados para o *ResNeSt*, indicando que o modelo baseado em *transformers* possui maior potencial de melhoria quando submetido a técnicas adequadas de regularização. Esse comportamento resultou em uma inversão na hierarquia de desempenho: enquanto nos experimentos dos *baseline* o *ResNeSt* havia superado o *ViT*, após as otimizações a situação se inverteu, com o *ViT* atingindo *performance* média de 80,79% contra 78,63% do *ResNeSt*.

O maior ganho individual foi observado na configuração *ViT* + *dataset* Costa WT, que apresentou aumento absoluto de +18,76% e relativo de +30,7%, evidenciando excelente resposta às estratégias de ajuste. Por outro lado, o *ResNeSt*, embora menos responsivo em termos relativos, mostrou maior estabilidade, com ganhos consistentes que variaram entre +2,21% e +10,85%. Esses resultados reforçam a hipótese de que arquiteturas distintas respondem de maneira diferenciada a processos de regularização e otimização, destacando o *ViT* como mais sensível e adaptável nesse cenário.

5.3.1 Análise de Convergência Otimizada

Na Figura 37 é apresentada a comparação direta entre as curvas de aprendizado do *baseline* e otimizadas, evidenciando visualmente os ganhos de *performance* e o controle de *overfitting* alcançados. Cada gráfico exibe quatro curvas: treino e validação para ambas as versões (*baseline* e otimizados), permitindo análise detalhada da evolução do treinamento.

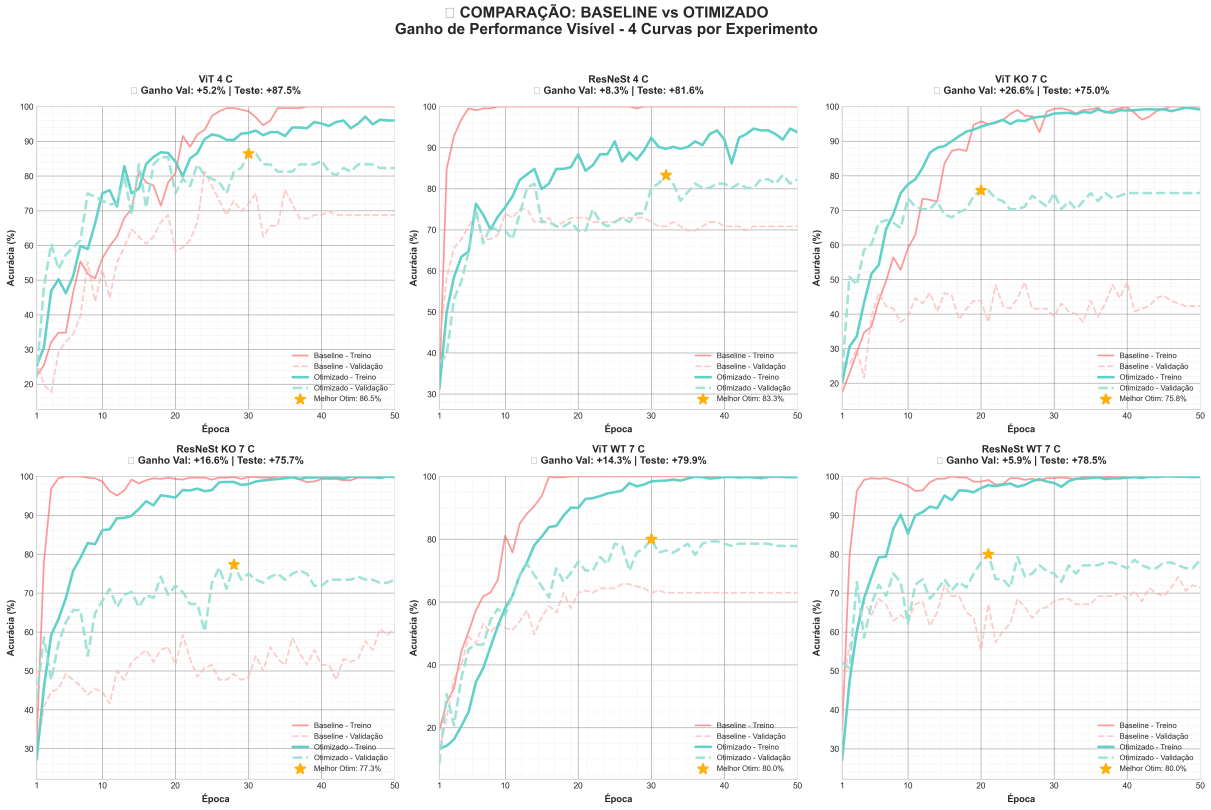


Figura 37 – Comparação Visual das Curvas de Aprendizado: *Baseline* vs Otimizado.

Os resultados demonstram ganhos consistentes em todos os experimentos, com destaque para o *ViT* com *dataset* Costa KO (7 classes) que apresentou o maior ganho de validação (+26,6%). É feita uma análise detalhada na Tabela 14 sendo quantificado detalhadamente o controle de *overfitting*.

Tabela 14 – Análise de *Overfitting*: *Baseline* versus Otimizado

Experimento	***Gap* Treino-Validação (%)**		**Redução Absoluta (%)**	**Ganho *Acc.* Val. (%)**
	Baseline	**Otimizado**		
ViT + Silva (4 classes)	31,25	13,69	-17,56	+5,2
ResNeSt + Silva (4 classes)	29,17	11,46	-17,71	+8,3
ViT + Costa KO (7 classes)	57,69	24,14	-33,55	+26,6
ResNeSt + Costa KO (7 classes)	39,23	26,37	-12,86	+16,6
ViT + Costa WT (7 classes)	37,06	21,89	-15,17	+14,3
ResNeSt + Costa WT (7 classes)	28,67	21,35	-7,32	+5,9
Média *ViT*	**42,00**	**19,91**	**-22,09**	**+15,4**
Média *ResNeSt*	**32,36**	**19,73**	**-12,63**	**+10,3**

De forma geral, as otimizações resultaram em significativa redução do *overfitting*, especialmente para o *ViT* (-22,09% em média). O caso mais expressivo foi o do *ViT* + *dataset* Costa KO, que alcançou uma redução de -33,55%, demonstrando a efetividade das técnicas de regularização aplicadas em cenários de alta complexidade. Em paralelo, o ganho médio de acurácia em validação foi de +12,80% (±7,4%), com todos os experi-

mentos apresentando melhorias. Os maiores incrementos foram observados nos conjuntos de 7 classes, reforçando que as otimizações se mostram particularmente vantajosas em cenários de maior complexidade classificatória.

A análise de concordância entre os conjuntos de validação e teste, apresentada na Figura 38, demonstra a consistência dos modelos otimizados.

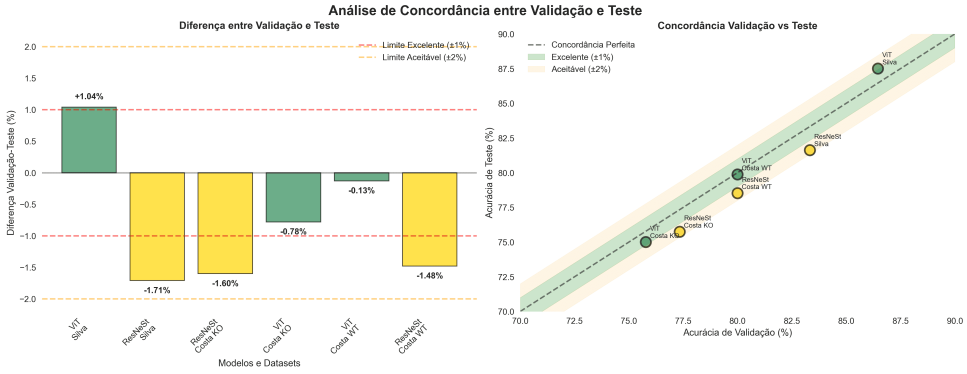


Figura 38 – Concordância entre Acurácia de Validação e Teste Após Otimizações

A análise de correlação linear revelou um coeficiente de determinação extremamente elevado ($R^2 = 0{,}9988$), evidenciando excelente concordância entre os desempenhos de validação e teste e reforçando a capacidade de generalização dos modelos otimizados. As diferenças médias entre validação e teste permaneceram controladas em ambas as arquiteturas: no caso do *ViT*, a discrepância foi mínima, com média de $-0{,}55\%$ e intervalo variando entre $-0{,}13\%$ e $-0{,}78\%$. Já o *ResNeSt* apresentou uma diferença ligeiramente maior, com média de $-1{,}13\%$ e intervalo entre $-0{,}25\%$ e $-1{,}71\%$. Esses resultados confirmam a robustez dos modelos frente a variações entre os conjuntos de validação e teste, demonstrando estabilidade na generalização mesmo após a aplicação das otimizações.

5.3.2 Análise de Complexidade Computacional

A análise de complexidade computacional, apresentada na Figura 39 compara o número de parâmetros treináveis com a eficiência por parâmetro em cada arquitetura, buscando equilibrar custo de treinamento e desempenho alcançado.

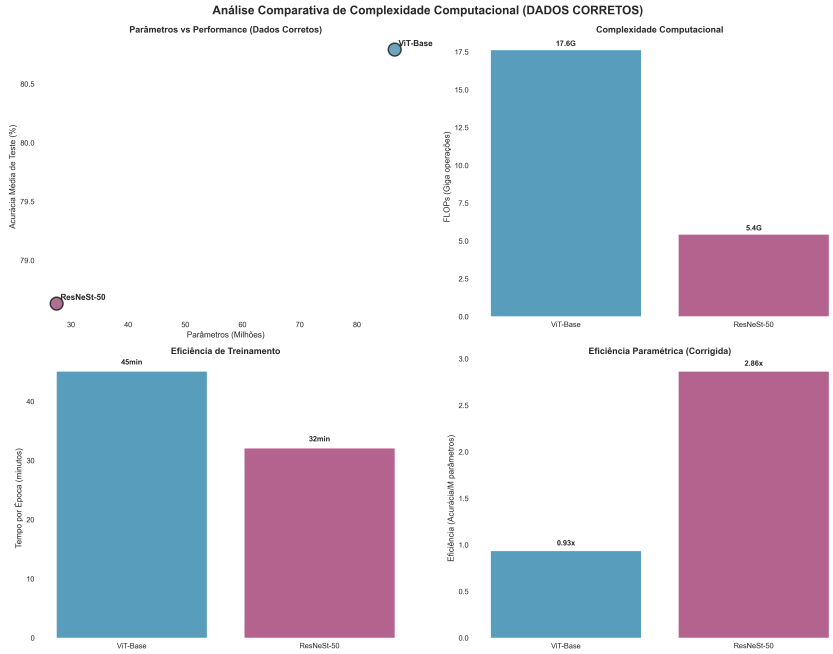


Figura 39 – Análise de Complexidade Computacional: Parâmetros vs *Performance*

No caso do *ResNeSt-50*, o modelo possui aproximadamente 27,5 milhões de parâmetros, alcançando acurácia média de 78,63%. Isso resulta em uma eficiência de 2,86% de acurácia por milhão de parâmetros. Já o *ViT-Base16*, com cerca de 86,6 milhões de parâmetros, atingiu acurácia média de 80,79%, correspondendo a 0,93% de acurácia por milhão de parâmetros. Esses resultados indicam que, embora o *ViT* apresente desempenho absoluto ligeiramente superior, o *ResNeSt* se mostra aproximadamente 3,07 vezes mais eficiente em termos de aproveitamento computacional, evidenciando maior custo-benefício quando se considera a relação entre parâmetros treinados e acurácia alcançada.

5.3.3 Desempenho Detalhado por Classes para Modelos Otimizados

A análise por classes foi conduzida apenas para o melhor resultado otimizado atingido, correspondente ao experimento *ViT + Silva* com acurácia de 87,50%. Uma análise comparativa entre os experimentos do *baseline* e o otimizado para este experimento é detalhadamente analisada na Tabela 15.

Tabela 15 – *Performance* por Classe: *ViT* Otimizado vs *ViT Baseline* (Silva 4 Classes)

	ViT Baseline			***ViT Otimizado***		
Classe	***Precision***	***Recall***	***F1-Score***	***Precision***	***Recall***	***F1-Score***
Healthy	0,871	0,964	0,915	0,933	1,000	0,966
Mild	0,605	0,697	0,648	0,848	0,848	0,848
Moderate	0,529	0,474	0,500	0,789	0,789	0,789
Severe	0,778	0,757	0,767	0,919	0,919	0,919
Média	**0,696**	**0,723**	**0,708**	**0,872**	**0,889**	**0,881**

Os ganhos foram particularmente expressivos na classe *Moderate*, cujo *F1-Score* evoluiu de 50,0% para 78,9%, representando uma melhora de +28,9%. A classe *Mild* também apresentou avanço significativo, com crescimento de +20,0% (64,8 → 84,8%). Em seguida, a classe *Severe* obteve incremento de +15,2% , passando de 76,7% para 91,9%. Por fim, a classe *Healthy*, que já apresentava alta performance no modelo do *baseline*, avançou de 91,5% para 96,6%, com ganho mais modesto de +5,1%. Esses resultados evidenciam que as otimizações não apenas elevaram a média global, mas também reduziram discrepâncias entre classes, promovendo maior equilíbrio no desempenho do modelo.

5.3.4 Análise de Matrizes de Confusão para Modelos Otimizados

Foram geradas matrizes de confusão para os melhores resultados otimizados em cada cenário. De forma semelhante ao *baseline*, apenas para o *dataset* de quatro classes, foram analisados o *ViT* (86,8%) e *ResNeSt* (81,6%). Na Figura 40 são apresentadas comparações dessas matrizes.

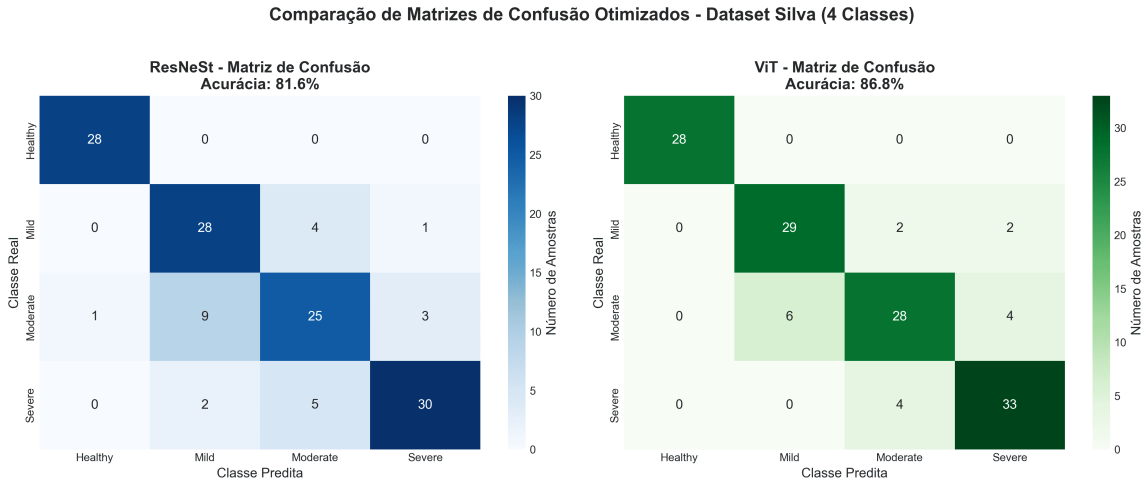


Figura 40 – Matrizes de Confusão: Comparação entre Melhores Resultados Otimizados para base de imagens Silva (4 Classes) entre *ResNeSt* (esquerda) e *ViT* (direita), respectivamente.

Na Figura 40 são apresentadas as matrizes de confusão para o *dataset* de 4 classes, destacando os melhores resultados otimizados obtidos com o *ResNeSt* (81,6%) e o *ViT* (86,8%). A análise detalhada por classe evidencia padrões distintos de acerto e erro entre as duas arquiteturas. No caso do *ResNeSt*, a classe *Healthy* atingiu desempenho perfeito, com 28 de 28 amostras corretamente classificadas (100,0%). A classe *Mild* obteve 28 acertos em 33 instâncias (84,8%), enquanto a *Moderate* apresentou maior dificuldade, com apenas 25 acertos em 38 (65,8%), acumulando 13 erros. Já a classe *Severe* alcançou 30 acertos em 37 instâncias (81,1%). Para o *ViT*, a classe *Healthy* também obteve 100,0% de acurácia (28/28), seguida por *Mild* com 29 acertos em 33 (87,9%). A classe *Moderate*, embora ainda desafiadora, apresentou melhora em relação ao *ResNeSt*, com 28 acertos

em 38 (73,7%). Por fim, a classe *Severe* alcançou 33 acertos em 37 (89,2%), também superando os resultados do *ResNeSt*.

Os padrões de erro observados indicam que a classe *Moderate* concentrou a maior parte das confusões em ambas arquiteturas. No *ResNeSt*, 52,0% de todos os erros ocorreram nessa classe, enquanto no *ViT* esse valor foi de 55,6%. As principais confusões ocorreram entre as classes *Moderate* e *Mild* (23,7% no *ResNeSt* e 15,8% no *ViT*), além de *Moderate* e *Severe* (7,9% no *ResNeSt* e 10,5% no *ViT*). Também foram registrados erros relevantes na direção oposta, como *Mild* classificada como *Moderate* (12,1% no *ResNeSt* e 6,1% no *ViT*) e *Severe* classificada como *Moderate* (13,5% no *ResNeSt* e 10,8% no *ViT*). De forma geral, os resultados confirmam que a classe *Moderate* se manteve como o maior desafio para ambas as arquiteturas, enquanto *Healthy* apresentou comportamento mais estável e facilmente separável no processo de classificação.

Em síntese, as otimizações transformaram o cenário experimental, promovendo ganhos consistentes de *performance*, controle de *overfitting* e maior estabilidade de treinamento. Ambos os modelos evidenciaram benefícios relevantes, reforçando a eficácia das estratégias adotadas e estabelecendo uma base sólida para análises mais avançadas.

5.4 Análise com Técnicas de Inteligência Artificial Explicável

Após a avaliação quantitativa dos modelos, foi realizada análise de explicabilidade utilizando técnicas de *Explainable Artificial Intelligence* (XAI) para validar a interpretabilidade das predições. Esta análise é fundamental para aplicações clínicas, onde a compreensão dos padrões de decisão é tão importante quanto a acurácia das classificações.

Foram implementadas seis técnicas de XAI distintas: três específicas para *Vision Transformer* (Grad-CAM tradicional, mapas de atenção nativos e Grad-CAM melhorado) e três para *ResNeSt* (Grad-CAM especializado, atenção CNN e Grad-CAM de entrada). As visualizações foram geradas em épocas representativas do treinamento utilizando amostras fixas para análise comparativa.

5.4.1 *Vision Transformer*

Na Figura 41 é apresentada uma comparação detalhada entre as três técnicas implementadas para o *ViT*.

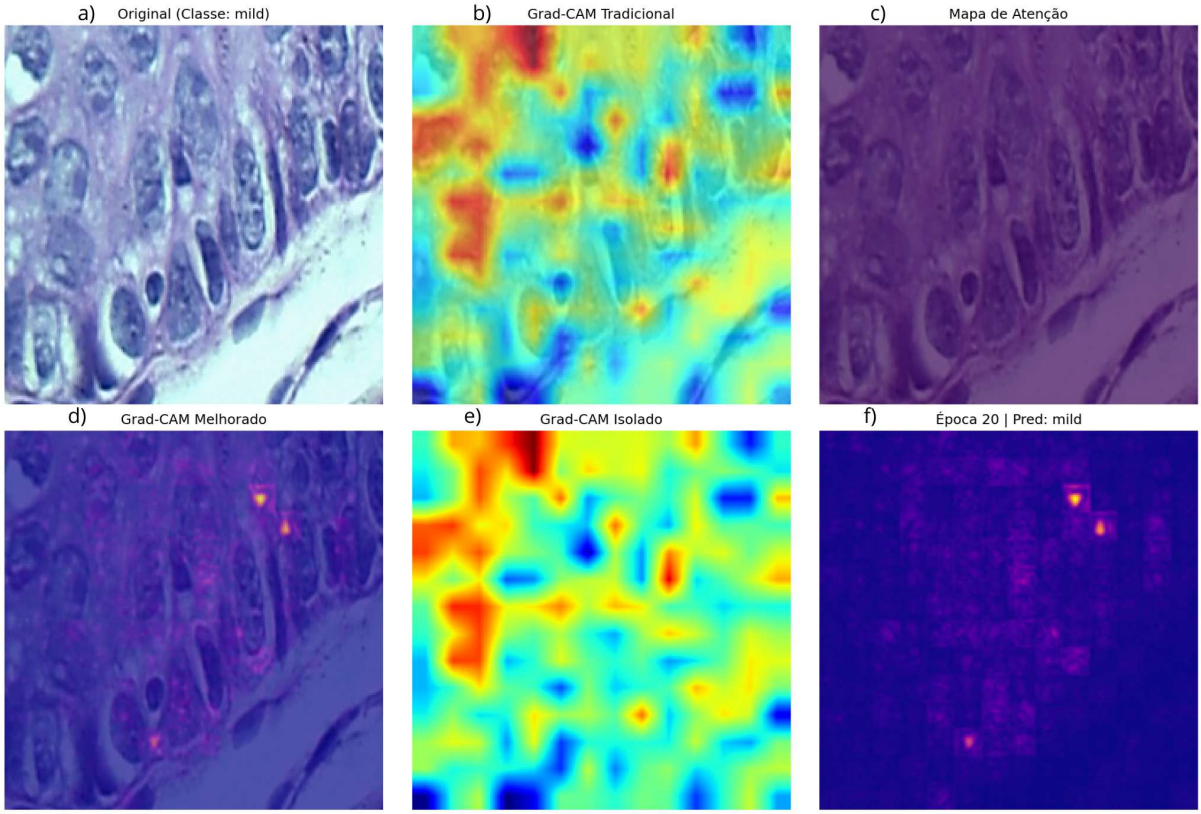


Figura 41 – Comparação das técnicas XAI para *Vision Transformer*: (a) Imagem histológica original, (b) Grad-CAM tradicional, (c) Mapas de atenção nativos, (d) Grad-CAM melhorado, (e) Mapa de calor isolado, (f) Informações da predição.

Na análise das técnicas de explicabilidade, o Grad-CAM tradicional apresentou boa cobertura espacial ao localizar as regiões mais relevantes das imagens, embora com menor precisão nas bordas das estruturas celulares. Os mapas de atenção nativos demonstraram maior granularidade na identificação de patches específicos, refletindo de forma direta o mecanismo de atenção do transformer. O Grad-CAM melhorado apresentou maior precisão na delimitação das estruturas histológicas, resultado dos estágios adicionais de suavização e normalização aplicados.

5.4.2 *ResNeSt*

Na Figura 42 é apresentada uma comparação das técnicas implementadas para *ResNeSt*.

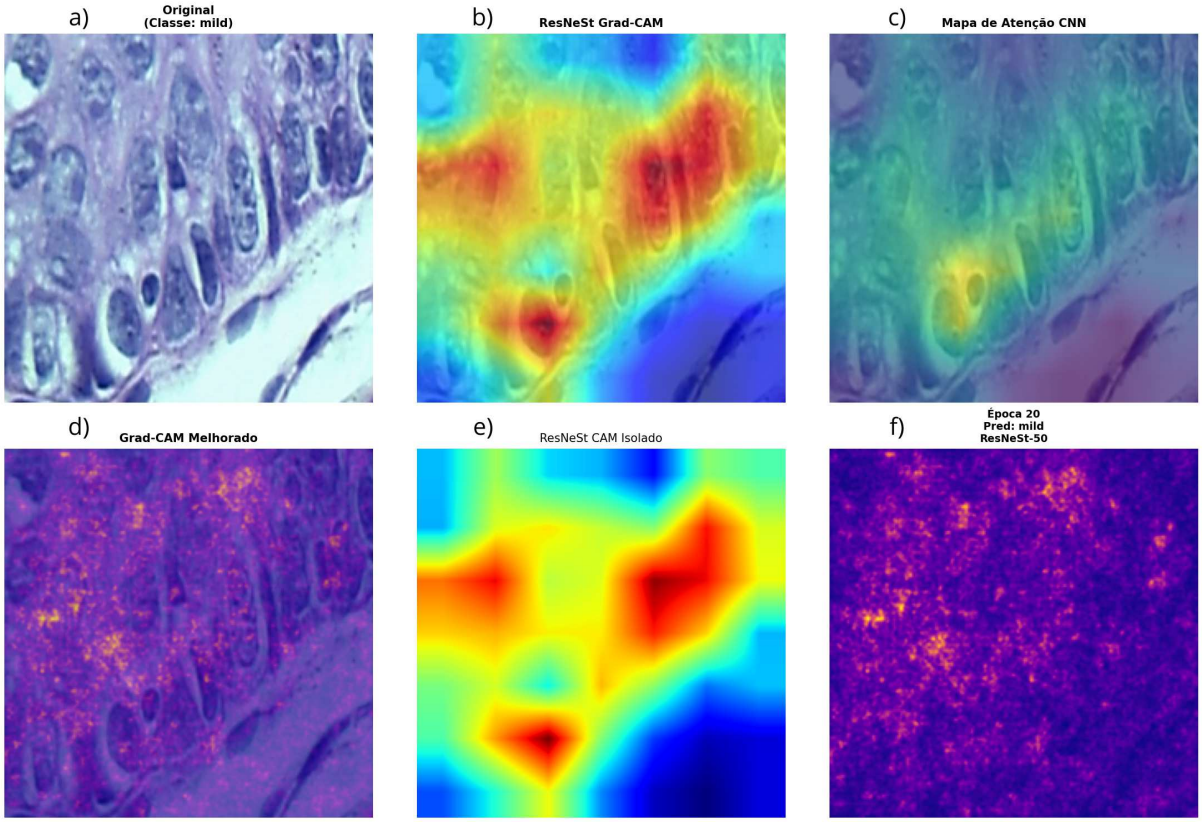


Figura 42 – Comparação das técnicas XAI para *ResNeSt*: (a) Imagem histológica original, (b) Grad-CAM especializado, (c) Atenção CNN, (d) Grad-CAM de entrada, (e) Mapa de calor isolado, (f) Informações da predição.

Na análise do *ResNeSt*, o Grad-CAM especializado aproveitou a arquitetura *Split-Attention* para identificar regiões com boa precisão espacial. A atenção CNN revelou padrões de ativação baseados em características convolucionais, complementando a análise obtida a partir dos gradientes. O Grad-CAM de entrada forneceu uma perspectiva sobre como variações nos pixels de entrada influenciam as predições finais.

5.4.3 Comparação Inter-Arquiteturas

A comparação direta entre *ViT* e *ResNeSt* na mesma imagem histológica revelou diferenças nos padrões de atenção das arquiteturas, evidenciadas na Figura 43, que apresenta esta análise comparativa nos dois bancos de imagens em questão, Banco Silva, Banco Costa KO e Banco Costa WT, respectivamente.

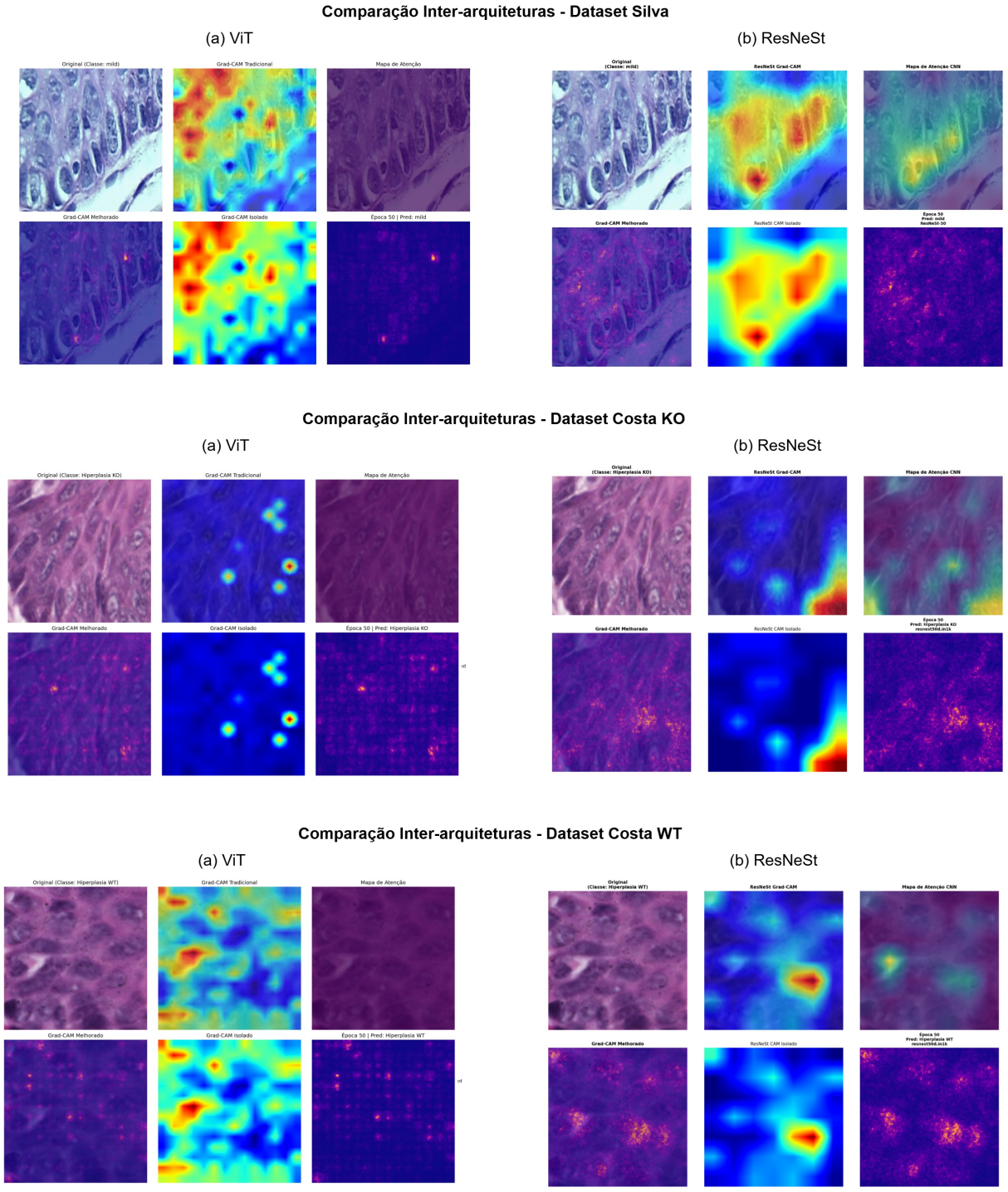


Figura 43 – Comparação inter-arquiteturas: *ViT* vs *ResNeSt* aplicados à mesma amostra histológica para os três bancos de imagens, evidenciando diferentes padrões de atenção.

O *ViT* tendeu a focar em padrões globais e regiões mais amplas da imagem, aproveitando seu mecanismo de atenção global, enquanto o *ResNeSt* demonstrou maior sensibilidade a detalhes locais e bordas de estruturas, reflexo de sua arquitetura convolucional com atenção hierárquica.

5.4.4 Interpretação Clínica das Explicações

A análise das explicações XAI em relação às características histopatológicas conhecidas revelou correspondências clinicamente relevantes. Na Figura 44 são apresentados exemplos representativos para diferentes classes histológicas, em que as Figuras 44 (a) e 44(b) obtidas pelo *ViT* e as Figuras 44(c) e 44(d) obtidas pelo *ResNeSt*.

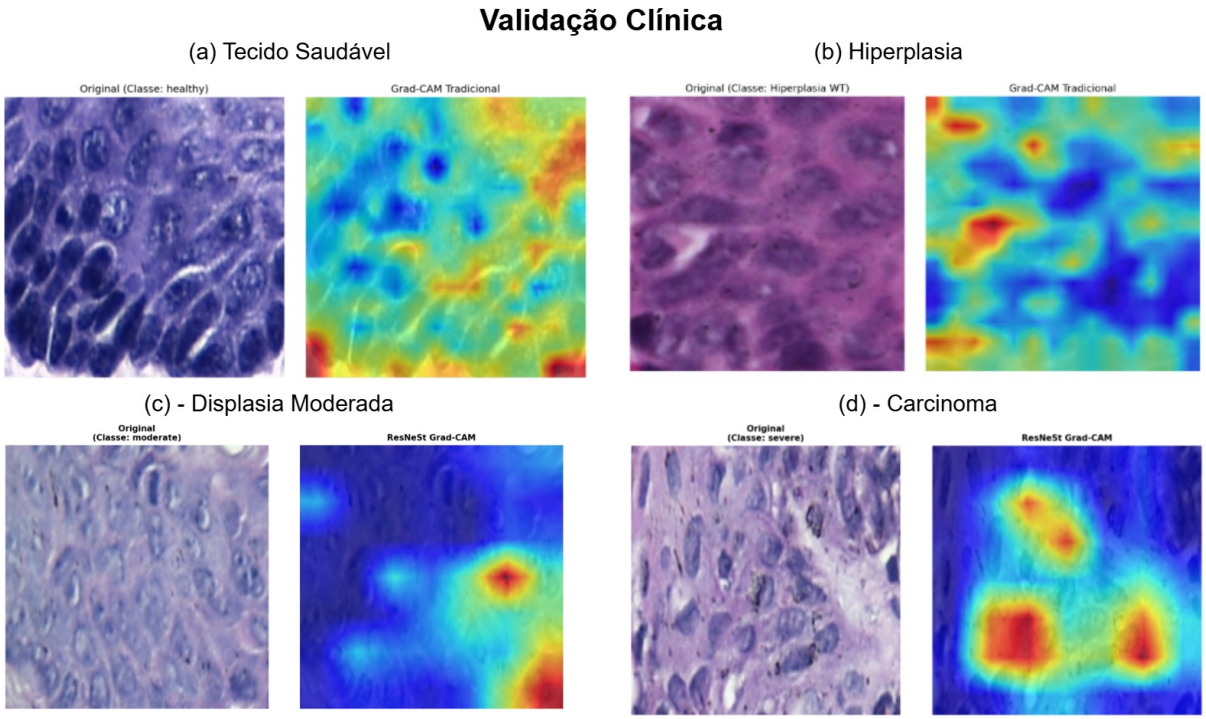


Figura 44 – Validação clínica das explicações XAI por classe histológica: (a) Tecido saudável - padrão distribuído normal, (b) Hiperplasia - identificação de espessamento tecidual, (c) Displasia - atenção em alterações epiteliais, (d) Carcinoma - foco em áreas de invasão.

A análise clínica das explicações XAI mostrou que, no tecido saudável, o padrão de atenção permaneceu distribuído de forma uniforme, sem focalização específica, condizente com a ausência de alterações patológicas. Nos casos de hiperplasia, os modelos conseguiram identificar corretamente áreas com aumento do tecido epitelial sem alterações anormais, distinguindo-as das áreas com displasia. Nas displasias moderadas, a atenção concentrou-se nas alterações epiteliais, refletindo a especificidade em relação ao grau de displasia observado. Por fim, nos casos de carcinoma, ambas as arquiteturas focalizaram consistentemente em regiões de invasão estromal e perda de polaridade celular, características essenciais para o diagnóstico.

Em síntese, a análise de explicabilidade revelou que ambas as arquiteturas geraram explicações visuais consistentes, com refinamentos progressivos ao longo dos treinamentos. As regiões destacadas pelos modelos correspondem a características histopatológicas conhecidas, evidenciando relevância clínica e confiabilidade das predições. Além disso, di-

ferentes técnicas XAI forneceram perspectivas complementares, reforçando o potencial de interpretação detalhada dos modelos.

6 Conclusão

Este trabalho investigou a aplicação de arquiteturas modernas de aprendizado profundo, *Vision Transformer* (*ViT*) e *ResNeSt*, para classificação histológica automatizada de displasias orais, com foco em explicabilidade através de técnicas de XAI.

A investigação revelou comportamentos distintos e complementares entre as arquiteturas estudadas. Nos experimentos iniciais de *baseline*, o *ResNeSt* demonstrou superioridade com acurácia de 72,41% comparado aos 64,00% do *ViT*. Entretanto, após a implementação das otimizações propostas, observou-se inversão completa desta tendência: o *ViT* superou o *ResNeSt* alcançando 80,79% *versus* 78,63% de acurácia, demonstrando maior responsividade às técnicas de regularização e aumento de dados implementadas.

As técnicas de otimização resultaram em ganhos substanciais para ambas as arquiteturas, com melhoria média de *performance* de 11,51% (68,20% para 79,71%). Observou-se também redução significativa do *overfitting*, com o *gap* entre treino e validação diminuindo de 37,18% para 19,82%. O melhor resultado individual foi obtido pelo *ViT* no Banco Silva, alcançando 87,50% de acurácia.

A análise do impacto da complexidade do problema revelou que a redução de 7 para 4 classes resultou em ganhos de 10,19% em acurácia e 30,9% em *F1-Score*, indicando que agrupamentos clinicamente fundamentados otimizam a *performance* sem comprometer a relevância diagnóstica. Este achado sugere que simplificações estruturadas do problema podem ser estratégicas para implementação clínica.

A investigação de explicabilidade através de seis técnicas XAI distintas revelou evolução temporal consistente das explicações ao longo do treinamento, correspondência entre atenção dos modelos e características histopatológicas clinicamente relevantes, padrões diferenciados de atenção (*ViT* com foco global *versus ResNeSt* com atenção a detalhes locais), e validação clínica satisfatória das explicações para todas as classes histológicas estudadas.

Este trabalho estabelece fundamentos sólidos e demonstra o potencial significativo da aplicação de arquiteturas modernas de *Deep Learning* em diagnóstico histopatológico automatizado. A demonstração de que o *ViT*, quando adequadamente otimizado, supera arquiteturas convolucionais especializadas representa avanço conceitual significativo para a área. A validação clínica das explicações XAI constitui contribuição fundamental, demonstrando que modelos de alta *performance* podem fornecer interpretações clinicamente relevantes, característica essencial para aceitação em ambiente clínico real.

Os resultados obtidos indicam maturidade suficiente para translação para estu-

dos clínicos controlados, representando passo importante rumo à implementação de inteligência artificial explicável como ferramenta de apoio diagnóstico em patologia oral. A contribuição transcende aspectos puramente técnicos, fornecendo evidências concretas para transformação do diagnóstico histopatológico através da combinação efetiva de *performance* preditiva e interpretabilidade clínica.

6.1 Contribuições

Este estudo oferece contribuições significativas tanto do ponto de vista metodológico quanto clínico para a área de diagnóstico histopatológico automatizado. Do ponto de vista metodológico, foi desenvolvido um protocolo especializado de aumento de dados para histologia, incorporando nove técnicas específicas que demonstraram eficácia na melhoria da generalização dos modelos. Adicionalmente, foi implementado um sistema XAI multi-técnica com validação temporal, permitindo análise longitudinal da evolução das explicações durante o treinamento. Este trabalho representa também a primeira comparação sistemática entre *ViT* e *ResNeSt* especificamente no domínio histológico.

As contribuições clínicas incluem a demonstração empírica da correspondência entre explicações XAI e características diagnósticas relevantes, validada através de múltiplos bancos de imagens (Banco Silva e Banco Costa). O estudo fornece explicações visuais interpretáveis que podem servir como ferramentas de apoio diagnóstico, estabelecendo bases sólidas para futura implementação em ambiente clínico real.

6.2 Limitações Enfrentadas

Algumas limitações importantes devem ser reconhecidas neste estudo. O trabalho foi conduzido com bancos de imagens de tamanho limitado se comparados a grandes *datasets* médicos disponíveis na literatura, utilizando exclusivamente bancos de imagens murinos, o que pode limitar a generalização direta para aplicações clínicas humanas. A resolução das imagens foi padronizada em 224×224 pixels conforme especificações técnicas das arquiteturas *ViT-Base16* e *ResNeSt-50* utilizadas, podendo potencialmente impactar a preservação de detalhes histológicos finos presentes em resoluções superiores.

A ausência de validação externa independente interlaboratorial representa limitação significativa, uma vez que todos os bancos de imagens utilizados são provenientes do mesmo ambiente de pesquisa. Adicionalmente, o custo computacional elevado para treinamento de modelos como *Vision Transformers*, que utilizam número muito alto de parâmetros, dificultou a execução de múltiplas épocas e exploração extensiva de hiperparâmetros para busca de novos máximos locais e globais.

6.3 Trabalhos Futuros

Os resultados promissores deste estudo abrem diversas direções para investigações futuras. A exploração de versões mais robustas das arquiteturas estudadas, como *ViT Large* (304 milhões de parâmetros) e *ResNeSt269e* (111 milhões de parâmetros), pode revelar ganhos adicionais de *performance*. O desenvolvimento de arquiteturas híbridas que combinem mecanismos convolucionais e de atenção representa oportunidade de pesquisa promissora.

A validação clínica prospectiva em contexto real com patologistas constitui passo fundamental para translação dos achados para aplicação prática. A extensão para *datasets* humanos é essencial para validar a translação dos achados obtidos em modelos murinos. O desenvolvimento de um sistema integrado de apoio diagnóstico com interface clínica que incorpore as técnicas XAI desenvolvidas representaria avanço significativo para implementação prática.

Finalmente, a exploração sistemática de hiperparâmetros através de técnicas de otimização avançadas pode revelar configurações que maximizem ainda mais a *performance* dos modelos, especialmente com disponibilidade de recursos computacionais mais robustos.

Referências

ADADI, A.; BERRADA, M. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). **IEEE Access**, IEEE, v. 6, p. 52138–52160, 2018. Citado na página 36.

ADEL, A.; AKRAM, F.; RASHID, A.; PINHEIRO, S.; YASMIN, S.; NASR-ESFAHANI, E.; PEREIRA, T.; OLIVEIRA, F.; NAVARRO, F.; CONDESSA, B. et al. Classification of oral epithelial dysplasia using visual descriptors. **IEEE Access**, IEEE, v. 6, p. 49787–49796, 2018. Citado 2 vezes nas páginas 43 e 44.

ALZUBAIDI, L.; ZHANG, J.; HUMAIDI, A. J.; AL-DUJAILI, A.; DUAN, Y.; AL-SHAMMA, O.; SANTAMARÍA, J.; FADHEL, M. A.; AL-AMIDIE, M.; FARHAN, L. Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. **Journal of big Data**, Springer, v. 8, n. 1, p. 53, 2021. Citado 2 vezes nas páginas 4 e 22.

AMARAL, R. C. d.; ANDRADE, R. A. R.; COUTO, G. R.; HERRERA-SERNA, B. Y.; REZENDE-SILVA, E.; CARDOSO, M. C. A. C. Tendências de mortalidade por câncer bucal no brasil por regiões e principais fatores de risco. **Revista Brasileira de Cancerologia**, v. 68, n. 2, p. e–081877, maio 2022. Disponível em: <https://rbc.inca.gov.br/index.php/revista/article/view/1877>. Citado na página 13.

BANCROFT, J. D.; GAMBLE, M. **Theory and practice of histological techniques**. [S.l.]: Elsevier Health Sciences, 2008. Citado na página 34.

BERTALMIO, M.; SAPIRO, G.; CASELLES, V.; BALLESTER, C. Image inpainting. p. 417–424, 2000. Citado na página 35.

BISHOP, C. M. **Pattern recognition and machine learning**. [S.l.]: Springer, 2006. Citado na página 35.

BULTEN, W.; PINCKAERS, H.; BOVEN, H. van et al. Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. **Nature medicine**, Nature Publishing Group, v. 28, n. 1, p. 154–163, 2022. Citado na página 33.

CAMPANELLA, G.; HANNA, M. G.; GENESLAW, L. et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. **Nature medicine**, Nature Publishing Group, v. 25, n. 8, p. 1301–1309, 2019. Citado na página 33.

CARVALHO, R. H. d. O. et al. Classificação de displasia da cavidade oral baseada em descritores fractais e ensemble learning. Universidade Federal de Uberlândia, 2024. Citado 3 vezes nas páginas 4, 13 e 17.

CARVALHO, T. M. et al. Explorando modelos em cascata de cnns e fractais para classificação de displasia oral. Universidade Federal de Uberlândia, 2023. Citado na página 13.

CASTRO, D. C.; WALKER, I.; GLOCKER, B. Causality matters in medical imaging. **Nature Communications**, Nature Publishing Group, v. 11, n. 1, p. 3673, 2020. Citado na página 35.

CIOMPI, F.; GEESSINK, O.; BEJNORDI, B. E. et al. The importance of stain normalization in colorectal tissue classification with convolutional networks. **arXiv preprint arXiv:1702.05931**, 2017. Citado na página 33.

COSTA, E. G. da; MIGLIORATI, C. A. Câncer bucal: avaliação do tempo decorrente entre a detecção da lesão e o início do tratamento. **Revista Brasileira de Cancerologia**, v. 47, n. 3, p. 283–289, 2001. Citado 2 vezes nas páginas 18 e 22.

DEVRIES, T.; TAYLOR, G. W. Improved regularization of convolutional neural networks with cutout. **arXiv preprint arXiv:1708.04552**, 2017. Citado na página 35.

DIMITRIOU, N.; ARANDJELOVIĆ, O.; CAIE, P. D. Deep learning for whole slide image analysis: an overview. **Frontiers in medicine**, Frontiers, v. 6, p. 264, 2019. Citado na página 32.

DOSOVITSKIY, A.; BEYER, L.; KOLESNIKOV, A.; WEISSENBORN, D.; ZHAI, X.; UNTERTHINER, T.; DEHGHANI, M.; MINDERER, M.; HEIGOLD, G.; GELLY, S. et al. An image is worth 16x16 words: Transformers for image recognition at scale. **arXiv preprint arXiv:2010.11929**, 2020. Citado 4 vezes nas páginas 4, 29, 38 e 46.

DUBEY, S. R.; SINGH, S. K. Transformer-based generative adversarial networks in computer vision: A comprehensive survey. **arXiv preprint arXiv:2302.08641**, 2023. Citado na página 28.

FICSOR, L.; VARGA, V. S.; TAGSCHERER, A. et al. Automated classification of inflammation in colon histological sections based on digital microscopy and advanced image analysis. **Cytometry Part A**, Wiley Online Library, v. 73, n. 4, p. 230–237, 2008. Citado na página 34.

GAUDIN, T.; VINCENT, N.; SARGENT, D. et al. Optical properties evaluation of histological slides for quantitative imaging. **Computerized Medical Imaging and Graphics**, Elsevier, v. 34, n. 4, p. 319–329, 2010. Citado na página 34.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A.; BENGIO, Y. **Deep learning**. [S.l.]: MIT press Cambridge, 2016. v. 1. Citado 2 vezes nas páginas 24 e 36.

HAYKIN, S. **Neural Networks: A Comprehensive Foundation**. Hamilton, Canada: Prentice Hall, 1999. ISBN 81-7808-300-0. Citado na página 25.

_____. **Neural Networks and Learning Machines**. 3rd. ed. Upper Saddle River, NJ, USA: Pearson Education, 2009. Citado na página 24.

HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep residual learning for image recognition. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2016. p. 770–778. Citado na página 33.

HUSSAIN, Z.; GIMENEZ, F.; YI, D.; RUBIN, D. Differential data augmentation techniques for medical imaging classification tasks. **AMIA Annual Symposium Proceedings**, American Medical Informatics Association, v. 2017, p. 979, 2017. Citado na página 32.

JANOWCZYK, A.; MADABHUSHI, A. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. **Journal of pathology informatics**, Wolters Kluwer–Medknow Publications, v. 7, 2016. Citado na página 31.

KOMURA, D.; ISHIKAWA, S. Machine learning methods for histopathological image analysis. **Computational and structural biotechnology journal**, Elsevier, v. 16, p. 34–42, 2018. Citado na página 32.

KOTHARI, S.; PHAN, J. H.; STOKES, T. H.; WANG, M. D. Removing batch effects from histopathological images for enhanced cancer diagnosis. **IEEE Journal of Biomedical and Health Informatics**, IEEE, v. 18, n. 3, p. 765–772, 2013. Citado 2 vezes nas páginas 34 e 35.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. **Communications of the ACM**, AcM New York, NY, USA, v. 60, n. 6, p. 84–90, 2017. Citado na página 25.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **nature**, Nature Publishing Group UK London, v. 521, n. 7553, p. 436–444, 2015. Citado na página 23.

LECUN, Y.; BOTTOU, L.; BENGIO, Y.; HAFFNER, P. Gradient-based learning applied to document recognition. **Proceedings of the IEEE**, Ieee, v. 86, n. 11, p. 2278–2324, 2002. Citado na página 25.

LITJENS, G.; KOOI, T.; BEJNORDI, B. E. et al. A survey on deep learning in medical image analysis. **Medical image analysis**, Elsevier, v. 42, p. 60–88, 2017. Citado 3 vezes nas páginas 23, 26 e 33.

LOCKHART, P.; JR, C. N.; PULLIAM, C. Dental factors in the genesis of squamous cell carcinoma of the oral cavity. **Oral oncology**, Elsevier, v. 34, n. 2, p. 133–139, 1998. Citado na página 16.

LOSHCHILOV, I.; HUTTER, F. Sgdr: Stochastic gradient descent with warm restarts. **arXiv preprint arXiv:1608.03983**, 2016. Citado na página 58.

_____. Decoupled weight decay regularization. **arXiv preprint arXiv:1711.05101**, 2017. Citado na página 57.

LUMERMAN, H.; FREEDMAN, P.; KERPEL, S. Oral epithelial dysplasia and the development of invasive squamous cell carcinoma. **Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology**, v. 79, n. 3, p. 321–329, 1995. ISSN 1079-2104. Disponível em: <https://doi.org/10.1016/S1079-2104(05)80226-4>. Citado na página 49.

MACENKO, M.; NIETHAMMER, M.; MARRON, J. S. et al. A method for normalizing histology slides for quantitative analysis. **Proceedings of the 2009 IEEE international symposium on biomedical imaging: from nano to macro**, p. 1107–1110, 2009. Citado na página 33.

MADABHUSHI, A.; LEE, G. Image analysis and machine learning in digital pathology: challenges and opportunities. **Medical image analysis**, Elsevier, v. 33, p. 170–175, 2016. Citado 2 vezes nas páginas 31 e 35.

MONTERO, P. H.; PATEL, S. G. Cancer of the oral cavity. **Surgical oncology clinics of North America**, v. 24, n. 3, p. 491, 2015. Citado na página 16.

MULRANE, L.; REXHEPAJ, E.; PENNEY, S.; CALLANAN, J. J.; GALLAGHER, W. M. Automated image analysis in histopathology: a microcomputer-based system for measuring enzyme-histochemical reactions. **The Journal of Histochemistry and Cytochemistry**, SAGE Publications, v. 56, n. 6, p. 507–517, 2008. Citado na página 34.

NAIK, S.; DOYLE, S.; AGNER, S. et al. Texture analysis of intermediate cell layers for cervical cancer prognosis. **Proceedings of the 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro**, p. 304–307, 2008. Citado na página 35.

PAIXÃO, G. M. d. M.; SANTOS, B. C.; ARAUJO, R. M. d.; RIBEIRO, M. H.; MORAES, J. L. d.; RIBEIRO, A. L. Machine learning na medicina: revisão e aplicabilidade. **Arquivos Brasileiros de Cardiologia**, SciELO Brasil, v. 118, n. 1, p. 95–102, 2022. Citado 2 vezes nas páginas 4 e 24.

PASSOS, D.; MISHRA, P. An automated deep learning pipeline based on advanced optimisations for leveraging spectral classification modelling. **Chemometrics and Intelligent Laboratory Systems**, Elsevier, v. 215, p. 104354, 2021. Citado na página 47.

PATHAK, D.; KRAHENBUHL, P.; DONAHUE, J.; DARRELL, T.; EFROS, A. A. Context encoders: Feature learning by inpainting. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2016. p. 2536–2544. Citado na página 35.

PEREZ, L.; WANG, J. The effectiveness of data augmentation in image classification using deep learning. **arXiv preprint arXiv:1712.04621**, 2017. Citado na página 31.

REINHARD, E.; ADHIKHMIN, M.; GOOCH, B.; SHIRLEY, P. Color transfer between images. **IEEE Computer graphics and applications**, IEEE, v. 21, n. 5, p. 34–41, 2002. Citado 3 vezes nas páginas 33, 34 e 55.

RONNEBERGER, O.; FISCHER, P.; BROX, T. U-net: Convolutional networks for biomedical image segmentation. In: SPRINGER. **Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015**. [S.l.], 2015. p. 234–241. Citado na página 34.

RUDIN, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. **Nature Machine Intelligence**, Nature Publishing Group, v. 1, n. 5, p. 206–215, 2019. Citado na página 36.

RUIFROK, A. C.; JOHNSTON, D. A. Quantification of histochemical staining by color deconvolution. **Analytical and quantitative cytology and histology**, International Academy of Cytology, v. 23, n. 4, p. 291–299, 2001. Citado 2 vezes nas páginas 33 e 35.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. **nature**, Nature Publishing Group UK London, v. 323, n. 6088, p. 533–536, 1986. Citado na página 25.

RUSSELL, P. N. Artificial intelligence: a modern approach by stuart. **Russell and Peter Norvig contributing writers, Ernest Davis...[et al.]**, 2010. Citado na página 22.

SANTOS, D. F. D. d. et al. Automated segmentation of tumor regions from oral histological whole slide images using fully convolutional neural networks. Universidade Federal de Uberlândia, 2022. Citado 3 vezes nas páginas 4, 18 e 19.

SCHMIDT, B. L.; DIERKS, E. J.; HOMER, L.; POTTER, B. Tobacco smoking history and presentation of oral squamous cell carcinoma. **Journal of oral and maxillofacial surgery**, Elsevier, v. 62, n. 9, p. 1055–1058, 2004. Citado 2 vezes nas páginas 16 e 21.

SCHOBER, P.; BOER, C.; SCHWARTE, L. A. Correlation coefficients: Appropriate use and interpretation. **Anesthesia & Analgesia**, v. 126, n. 5, p. 1763–1768, 2018. Acesso em: 20 out. 2025. Disponível em: <https://doi.org/10.1213/ANE.0000000000002864>. Citado na página 65.

SELVARAJU, R. R.; COGSWELL, M.; DAS, A.; VEDANTAM, R.; PARIKH, D.; BATRA, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: **Proceedings of the IEEE International Conference on Computer Vision (ICCV)**. [S.l.: s.n.], 2017. p. 618–626. Citado na página 36.

SHORTEN, C.; KHOSHGOFTAAR, T. M. A survey on image data augmentation for deep learning. **Journal of big data**, SpringerOpen, v. 6, n. 1, p. 1–48, 2019. Citado na página 31.

SILVA, A. B.; MARTINS, A. S.; TOSTA, T. A. A.; NEVES, L. A.; SERVATO, J. P. S.; ARAÚJO, M. S. de; FARIA, P. R. de; NASCIMENTO, M. Z. do. Computational analysis of histological images from hematoxylin and eosin-stained oral epithelial dysplasia tissue sections. **Expert Systems with Applications**, Elsevier, v. 193, p. 116456, 2022. Citado na página 13.

SILVA, A. B.; MARTINS, A. S.; TOSTA, T. A. A.; LOYOLA, A. M.; CARDOSO, S. V.; NEVES, L. A.; GALDINO, V.; NASCIMENTO, M. Z. do. Oralepitheliumdb: A dataset for oral epithelial dysplasia image segmentation and classification. **Journal of Imaging Informatics in Medicine**, v. 1, p. 1–20, 2024. Citado na página 48.

SILVA, A. B. et al. Métodos computacionais para análise e classificação de displasias em imagens da cavidade bucal. Universidade Federal de Uberlândia, 2019. Citado na página 13.

SILVA, L. A.; SANTOS, D. F. D. d.; MARTINS, A. S.; TOSTA, T. A. A.; NEVES, L. A.; SILVA, A. B.; NASCIMENTO, M. Z. do. A public dataset for oral epithelial dysplasia classification and benchmarking machine learning algorithms. **Scientific Data**, Nature Publishing Group UK London, v. 11, n. 1, p. 452, 2024. Citado 3 vezes nas páginas 42, 43 e 44.

SIMARD, P.; LECUN, Y.; DENKER, J.; VICTORRI, B. Transformation invariance in pattern recognition—tangent distance and tangent propagation. **Neural networks: tricks of the trade**, Springer, p. 239–274, 2000. Citado na página 36.

SIMARD, P. Y.; STEINKRAUS, D.; PLATT, J. C. Best practices for convolutional neural networks applied to visual document analysis. **Proceedings of the seventh international conference on document analysis and recognition**, v. 2, p. 958–963, 2003. Citado na página 34.

SPEIGHT, P. M. Update on oral epithelial dysplasia and progression to cancer. **Head and neck pathology**, Springer, v. 1, p. 61–66, 2007. Citado 2 vezes nas páginas 20 e 21.

TAYLOR, L.; NITSCHKE, G. Improving deep learning with generic data augmentation. **arXiv preprint arXiv:1708.06020**, 2018. Citado na página 33.

TELLEZ, D.; LITJENS, G.; BÁNDI, P.; BULTEN, W.; BOKHORST, J.-M.; CIOMPI, F.; LAAK, J. van der. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. **Medical image analysis**, Elsevier, v. 58, p. 101544, 2019. Citado 2 vezes nas páginas 31 e 35.

TOVARU, S.; COSTACHE, M.; PERLEA, P.; CARAMIDA, M.; TOTAN, C.; WARNAKULASURIYA, S.; PARLATESCU, I. Oral leukoplakia: A clinicopathological study and malignant transformation. **Oral diseases**, Wiley Online Library, v. 29, n. 4, p. 1454–1463, 2023. Citado na página 17.

VAHADANE, A.; PENG, T.; SETHI, A. et al. Structure-preserving color normalization and sparse stain separation for histological images. **IEEE transactions on medical imaging**, IEEE, v. 35, n. 8, p. 1962–1971, 2016. Citado na página 34.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, Ł.; POLOSUKHIN, I. Attention is all you need. **Advances in neural information processing systems**, v. 30, 2017. Citado 5 vezes nas páginas 4, 26, 27, 28 e 46.

VELDEN, B. H. van der; KUIJF, H. J.; GILHUIJS, K. G.; VIERGEVER, M. A. A systematic review on the application of deep learning in biomedical images. **Medical image analysis**, Elsevier, v. 69, p. 101971, 2021. Citado na página 32.

WEINMANN, H. Importância do estudo da histologia. **Anais da faculdade de medicina de porto alegre**, v. 3, p. 104–108, 1942. Citado na página 18.

ZHANG, H.; WU, C.; ZHANG, Z.; ZHU, Y.; LIN, H.; ZHANG, Z.; SUN, Y.; HE, T.; MUELLER, J.; MANMATHA, R. et al. Resnest: Split-attention networks. In: **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**. [S.l.: s.n.], 2022. p. 2736–2746. Citado 5 vezes nas páginas 4, 30, 39, 46 e 47.

ZHANG, S.-Z.; XIE, L.; SHANG, Z.-J. Burden of oral cancer on the 10 most populous countries from 1990 to 2019: estimates from the global burden of disease study 2019. **International Journal of Environmental Research and Public Health**, MDPI, v. 19, n. 2, p. 875, 2022. Citado na página 16.