UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Guilherme Rafael Cerqueira Dias

# Classificação de Lesões Histológicas Baseada em Modelos Híbridos de CNNs e ViT

**Uberlândia, Brasil**

**2025**

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Guilherme Rafael Cerqueira Dias

# Classificação de Lesões Histológicas Baseada em Modelos Híbridos de CNNs e ViT

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 Sistemas de Informação.

Orientador: Marcelo Zanchetta do Nascimento

Universidade Federal de Uberlândia – UFU

Faculdade de Computação

Bacharelado em Sistemas de Informação

Uberlândia, Brasil

2025

# Resumo

A classificação de imagens histológicas tem sido amplamente explorada em pesquisas de visão computacional, devido ao seu potencial de auxiliar no diagnóstico precoce de alterações celulares. Este trabalho apresenta um estudo sobre a aplicação de aprendizado profundo na classificação automatizada de lesões histológicas da cavidade oral, propondo um modelo híbrido que integra redes Convolucionais (CNNs) e *Vision Transformers* (ViT).O uso isolado de CNNs nesse tipo de tarefa apresenta limitações, como a dificuldade de generalização entre diferentes padrões morfológicos e a alta dependência de grandes bases de dados. Para contornar essas restrições, a arquitetura proposta combina a *EfficientNetV2B0*, responsável por extrair características locais (como bordas, texturas e estruturas finas), com o ViT, capaz de capturar relações globais e dependências de longo alcance entre diferentes regiões da imagem. Essa integração permite uma representação mais completa e informativa dos tecidos analisados.Os experimentos foram realizados com o objetivo de classificar os diferentes graus de displasia epitelial (leve, moderada e severa) em imagens histológicas da cavidade oral de camundongos, coradas com hematoxilina e eosina. Para essa tarefa, este estudo apresenta um modelo híbrido que integra *EfficientNet* e *Vision Transformer* (ViT), combinando suas habilidades de extração de características locais e globais, respectivamente. Essas informações foram agregadas em uma camada totalmente conectada para a classificação binária das lesões e demonstraram alta capacidade de distinção em casos de maior contraste morfológico: alcançaram AUCs de 1,00 na distinção entre tecido saudável e displasia leve e 0,97 na distinção entre tecido saudável e displasia severa, apresentando resultados promissores.Os resultados obtidos indicam que a combinação entre CNNs e ViT constitui uma abordagem robusta e eficaz para a classificação de imagens histológicas, contribuindo para o desenvolvimento de sistemas de diagnóstico assistido por computador na detecção de lesões epiteliais da cavidade oral.

**Palavras-chave**: Displasia; CAD; EfficientNet; ViT; Modelo Híbrido; Imagens Histológicas.

# Lista de ilustrações

# Lista de abreviaturas e siglas

| | |
|---|---|
| ac | Acurácia |
| CNN | Convolutional Neural Network |
| f1 | F1-Score |
| FN | Falso Negativo |
| FP | Falso Positivo |
| VP | Verdadeiro Positivo |
| VN | Verdadeiro Negativo |
| ID | Imagens Digitais (ID) histológicas de displasia |
| EN | EfficientNet |
| ViT | Vision Transformer |
| OMS | Organização Mundial da Saúde |
| ReLU | Função de ativação linear retificada |
| DA | Data Augmentation |

# Sumário

# 1 Introdução

A classificação de imagens histológicas de lesões epiteliais tem sido objeto de ampla pesquisa em visão computacional, representando uma tarefa essencial para o diagnóstico médico. No entanto, a análise manual dessas imagens é um processo demorado e sujeito a variações de precisão, que dependem diretamente da experiência do profissional. Para mitigar essas limitações, sistemas de auxílio ao diagnóstico (do inglês, *computer-aided diagnosis* — CAD) têm sido desenvolvidos com o propósito de apoiar especialistas, proporcionando maior rapidez e acurácia. Nesse contexto, as Redes Neurais Convolucionais (do inglês, *Convolutional Neural Networks* — CNNs) têm demonstrado eficácia na extração de padrões visuais locais. Arquiteturas mais recentes, como os *Vision Transformers* (ViT), destacam-se por uma abordagem distinta: a imagem é dividida em pequenos blocos não sobrepostos (*patches*), que são processados sequencialmente. Isso permite a captura de relações globais e dependências de longo alcance, ou seja, o ViT consegue identificar a conexão entre características morfológicas presentes em regiões distantes da lâmina (e.g., o padrão de invasão em uma extremidade e o grau de diferenciação celular em outra). O modelo EfficientNet (EN), por sua vez, complementa essa abordagem ao capturar características locais e otimizar a eficiência computacional.

Diante da necessidade de superar as limitações individuais dessas arquiteturas, este trabalho propõe a investigação de um modelo híbrido de comitê de classificadores. Essa abordagem integra as capacidades de CNNs, ViT e EN na classificação de lesões histológicas, visando alcançar uma arquitetura robusta e eficaz que combine a precisão na identificação de detalhes finos com a interpretação global da imagem. O desenvolvimento deste estudo busca contribuir para a classificação de lesões de displasia com maior precisão e eficiência, auxiliando o diagnóstico especializado.

O câncer é atualmente uma das principais causas de morte no mundo, configurando-se como um dos maiores desafios de saúde pública em escala global e responsável por grande parte das mortes prematuras na maioria dos países (Instituto Nacional de Câncer, 2023). Segundo estimativas da INCA para o triênio 2023-2025, são esperados cerca de 15.100 novos casos de câncer de cavidade oral e orofaringe por ano no Brasil, posicionando essa neoplasia entre os tipos de câncer de maior incidência no país (Instituto Nacional de Câncer, 2023). Em estudos populacionais, observou-se que, entre os homens brasileiros, o câncer de boca figura entre os seis tipos com maior frequência (LIMA et al., 2021), com uma taxa estimada de 10,30 novos casos para cada 100 mil indivíduos do sexo masculino, enquanto entre as mulheres a taxa é de 3,83 para cada 100 mil. Em 2020, o país registrou 6.192 óbitos associados a esse tipo de câncer, evidenciando a importância da detecção precoce de lesões displásicas para evitar sua progressão a estágios malignos.

As displasias são lesões pré-invasivas que comprometem o desenvolvimento normal das células epiteliais, sendo mais comumente observadas na cavidade oral. Podem ser identificadas por características morfológicas como crescimento celular anormal, variação no tamanho, hipercromasia, entre outros sinais (KUMAR, 2010). Segundo (INCA - Instituto Nacional de Câncer, 2002), a detecção precoce dessas lesões é essencial, pois, quando diagnosticadas em estágios iniciais, as displasias podem ser tratadas e revertidas antes de evoluírem para graus mais severos. Entretanto, à medida que atingem estágios avançados, tendem a evoluir para o câncer invasivo, no qual a metástase se torna descontrolada e as alterações morfológicas ainda mais acentuadas.

A classificação manual de imagens histológicas é uma tarefa trabalhosa e demorada (CHEN; CHEN, 2022), cuja qualidade depende da experiência do profissional responsável. Especialistas mais experientes tendem a apresentar maior consistência e precisão nas classificações. Para enfrentar essas limitações, os sistemas CAD buscam automatizar parte desse processo, permitindo diagnósticos mais rápidos e confiáveis, reduzindo erros humanos e a carga de trabalho dos patologistas. O uso de inteligência artificial em aplicações médicas remonta às origens da computação, mas foi apenas nas últimas décadas que sua aplicação se consolidou em diagnósticos clínicos (LISBOA, 2002).

Diversas técnicas têm sido aplicadas em sistemas CAD, como árvores de decisão, regressão logística, florestas aleatórias e redes neurais artificiais, que utilizam modelos matemáticos para realizar predições (PAIXãO et al., 2022). No entanto, esses sistemas ainda apresentam limitações que podem ser aprimoradas. Modelos de aprendizado profundo (do inglês, *Deep Learning* — DL) tornaram-se o núcleo de muitos sistemas CAD modernos voltados à patologia digital, demonstrando desempenho consistente na identificação de estruturas teciduais e na classificação de padrões. Contudo, redes CNN tradicionais ainda possuem restrições na modelagem de relações espaciais amplas. Nesse sentido, a utilização do modelo ViT contribui para uma representação mais abrangente e contextual na classificação de lesões histológicas. Assim, investigar novas associações entre ViT e CNNs, em uma abordagem híbrida, configura-se como um desafio relevante e promissor para o avanço do diagnóstico assistido de displasias da cavidade oral.

## 1.1 Objetivo Geral

O objetivo geral deste trabalho é propor e investigar um modelo híbrido de aprendizado profundo, que associa a capacidade de extração de características locais de CNNs, incluindo a EN, com a capacidade de processamento de contexto global dos ViTs. Essa abordagem é aplicada na classificação automatizada de lesões histológicas da cavidade oral, visando auxiliar especialistas no diagnóstico de displasia com maior precisão e eficiência.

### 1.1.1 Objetivos Específicos

- Desenvolver uma arquitetura híbrida de comitê, integrando EN, ViT e uma CNN ,simplificada para a identificação de padrões discriminativos em imagens histológicas de displasia;
- Avaliar o desempenho e a parametrização do modelo híbrido em cenários de classificação binária de diferentes graus de displasia epitelial;
- Investigar e quantificar o impacto de diferentes fatores de replicação do aumento de dados na capacidade de generalização e nas métricas de desempenho do modelo proposto;
- Realizar a análise das métricas estatísticas de desempenho (Acurácia, AUC, F1-Score) de cada arquitetura individualmente (EN, ViT e CNN simplificada) e no modelo híbrido final;
- Conduzir uma análise comparativa aprofundada entre os resultados do método proposto e as abordagens consolidadas presentes na literatura para classificação de lesões da cavidade oral.

## 1.2 Organização do Trabalho de Conclusão de Curso

Neste capítulo, foram apresentadas as considerações iniciais, motivação e objetivos do desenvolvimento deste trabalho. O restante deste trabalho possui a seguinte organização:

- **Capítulo 2: Fundamentação Teórica e Trabalhos Correlatos.** Apresenta os conceitos fundamentais para a compreensão e validação do trabalho, incluindo uma revisão detalhada sobre CNNs, ViTs e arquiteturas híbridas.
- **Capítulo 3: Revisão da Literatura.** Apresenta os principais trabalhos da literatura relacionados à classificação de imagens histológicas, destacando as abordagens, estratégias e técnicas complementares utilizadas para aprimorar o desempenho dos modelos.
- **Capítulo 4: Metodologia Proposta.** Descreve a base de dados utilizada, as técnicas de pré-processamento, a estratégia de aumento de dados e o detalhamento do modelo híbrido de aprendizado profundo desenvolvido.
- **Capítulo 5: Resultados e Discussão.** Apresenta os resultados alcançados nos cenários de classificação binária e multi-classe, bem como as respectivas análises, discussões e o comparativo com o Estado da Arte.

- **Capítulo 6: Considerações Finais.** São apresentadas as contribuições obtidas, as considerações finais sobre o trabalho e propostas para trabalhos futuros.

# 2 Fundamentação Teórica

Nesta seção, serão introduzidos os conceitos teóricos relacionados às principais etapas deste trabalho. Dentre os temas abordados, encontram-se: Displasia, Redes Neurais Convolucionais do inglês*convolutional neural networks*, EfficientNet (EN), *Vision Transformers* (ViT) e Processamento de Imagens Histológicas.

## 2.1 Displasias

As displasias são lesões pré-invasivas que prejudicam o desenvolvimento das células de órgãos ou tecidos, sendo mais comuns na cavidade oral. Essas lesões podem ser identificadas por meio de características morfológicas, como o crescimento anormal das células, diferença em seu tamanho, hipercromasia e outras alterações celulares (KUMAR, 2010).

De acordo com o (INCA - Instituto Nacional de Câncer, 2002), a detecção precoce dessas lesões é essencial, pois, quando descobertas em estágio inicial, as displasias podem ser revertidas e tratadas antes que evoluam para graus mais severos. Em estágios avançados, essas lesões tendem a progredir para um câncer, com alterações morfológicas ainda mais intensas e potencial de metástase. Na Figura 1 são apresentados dois casos de carcinoma de células escamosas na região de língua e de gengiva.

(a) Exemplos de carcinoma celular escamoso na língua. Fonte: (CARLSON et al., 2023).

(b) Um exemplo de leucoplasia homogênea da mucosa bucal esquerda. Fonte: (CARLSON et al., 2023).

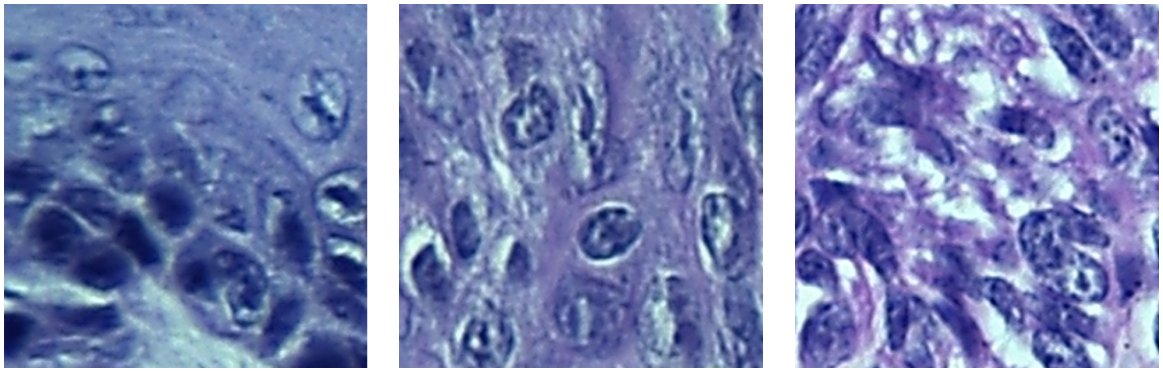
Figura 1 – Exemplos de lesões orais.

De acordo com (WARNAKULASURIYA et al., 2021), os graus de displasia são geralmente divididos em três categorias: leve, moderada e severa. Esses graus descrevem

o quanto as características das lesões se tornam proeminentes, sendo que, quanto maior o grau, mais evidentes elas se tornam.

Quando detectadas, essas lesões devem ser tratadas o mais rapidamente possível para evitar que evoluam para graus mais avançados, pois, uma vez em grau severo, torna-se difícil o tratamento para reversão para tecido saudável (INCA - Instituto Nacional de Câncer, 2002). Além disso, existe uma grande chance de evolução para um câncer, onde as características apresentadas na displasia tornam-se ainda mais evidentes e a multiplicação das células com a condição se torna descontrolada.

Na Figura 2 são apresentados exemplos de amostras de tecidos saudáveis corados com os corantes hematoxilina e eosina, com os diferentes níveis de displasia e tecido saudável.



(a) Amostra de tecido saudável.

(b) Amostra de tecido com displasia moderada.

(c) Amostra de tecido com displasia severa.

Figura 2 – Comparação entre tecido saudável, moderado e displasia severa. Fonte: (SILVA, 2019).

Segundo (WARNAKULASURIYA et al., 2008), a classificação da displasia oral nem sempre é unânime entre os profissionais da área. Enquanto algumas características são amplamente reconhecidas, outras podem ser interpretadas de maneiras distintas, resultando em divergências na avaliação.

## 2.2 Redes Neurais Convolucionais (CNNs)

As CNNs são modelos de *deep learning* amplamente empregados em tarefas de classificação de imagem devido à sua capacidade de extrair padrões visuais complexos, como bordas, texturas e formas. Desde sua aplicação revolucionária no *ImageNet Challenge*, as CNNs têm sido utilizadas em diagnósticos médicos, especialmente na análise de imagens histológicas, como as de displasia oral (KIM et al., 2016). O EN é um exemplo de modelo CNN otimizado que proporciona alta precisão com eficiência computacional (PAIXãO et al., 2022). Na prática, as CNNs facilitam a identificação de características morfológicas em imagens histológicas, auxiliando no reconhecimento de padrões especí-

ficos e na diferenciação entre tecidos saudáveis e lesões (GONZALEZ; WOODS, 2010; SOLOMON; BRECKON, 2000).

A convolução é uma operação matemática que combina funções para produzir uma terceira. No processamento de imagens, trata-se de uma técnica utilizada para modificar ou transformar uma imagem por meio da aplicação de um pequeno filtro, também denominado *kernel* ou máscara. Esse filtro, geralmente representado por uma matriz de dimensões reduzidas (como $3 \times 3$ ou $5 \times 5$), percorre a imagem de entrada. Cada elemento do filtro possui um valor numérico e, em cada posição, os valores correspondentes da imagem de entrada são multiplicados pelos valores do filtro. Os produtos resultantes são somados, gerando um único valor. Esse processo é repetido ao longo de toda a imagem, deslocando o filtro em uma posição por vez. A configuração desse deslocamento pode variar conforme o método de convolução empregado, como a convolução completa ou a convolução com preenchimento (*padding*), que influencia parâmetros como a posição inicial e o passo (*stride*). Na Figura 3 ilustra esse processo aplicado a uma imagem de entrada. A região destacada em laranja na matriz de entrada representa a parte que está sendo multiplicada pelo filtro. O resultado dessa operação é atribuído à célula correspondente no mapa de ativação, também destacada em laranja. Esse procedimento é repetido para todas as demais posições do mapa, gerando a representação final.

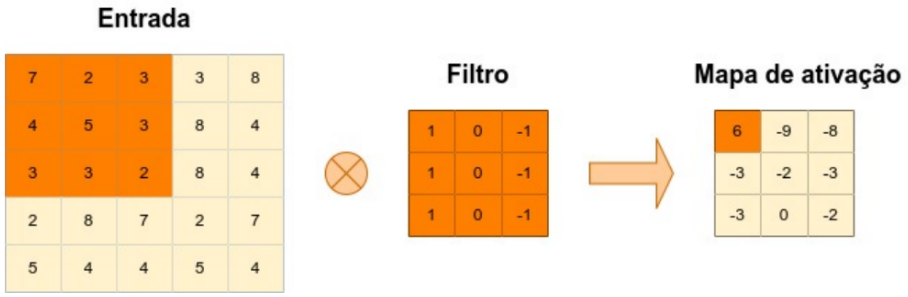



Figura 3 – Exemplo do desenvolvimento de uma convolução. Fonte: (ALZUBAIDI et al., 2021)

O resultado desse processo é uma nova imagem denominada mapa de características (*feature map*), que representa informações extraídas dos dados de entrada. No contexto das CNNs, esses mapas capturam diferentes níveis de abstração. Em estágios iniciais, identificam padrões de baixo nível, como bordas e texturas; em camadas mais profundas, reconhecem características mais complexas, como formas e padrões estruturais (HAYKIN, 2009). Na Figura 4 é uma representação da estrutura básica de uma CNN, usando uma matriz ($28 \times 28$).

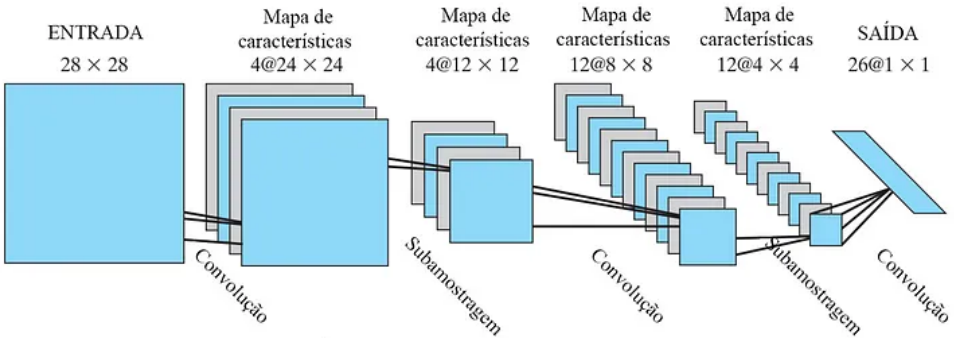



Figura 4 – Estrutura de uma rede convolucional. Fonte: (HAYKIN, 2009)

## 2.3 EfficientNet

O modelo EN é uma arquitetura de CNN projetada para otimizar o desempenho em tarefas de classificação de imagens. Ela introduz um novo método de escalonamento que equilibra a largura, profundidade e resolução da rede de maneira sistemática. Com isso, o EN consegue capturar características locais, como bordas e texturas, que são cruciais em tarefas de classificação de imagens médicas. Seu design eficiente permite que a rede extraia detalhes finos das imagens histológicas, melhorando a acurácia do diagnóstico em comparação com arquiteturas anteriores. Este modelo é derivado de uma família de arquiteturas de redes neurais convolucionais que foram projetadas para serem altamente eficientes em termos de consumo de recursos computacionais enquanto mantêm ou melhoram o desempenho em tarefas de visão computacional. Proposta por (TAN; LE, 2019) em um artigo intitulado *"EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks"*, publicado em 2019. Para conseguir fazer isso a EN foi a estrutura definida na Figura 5.

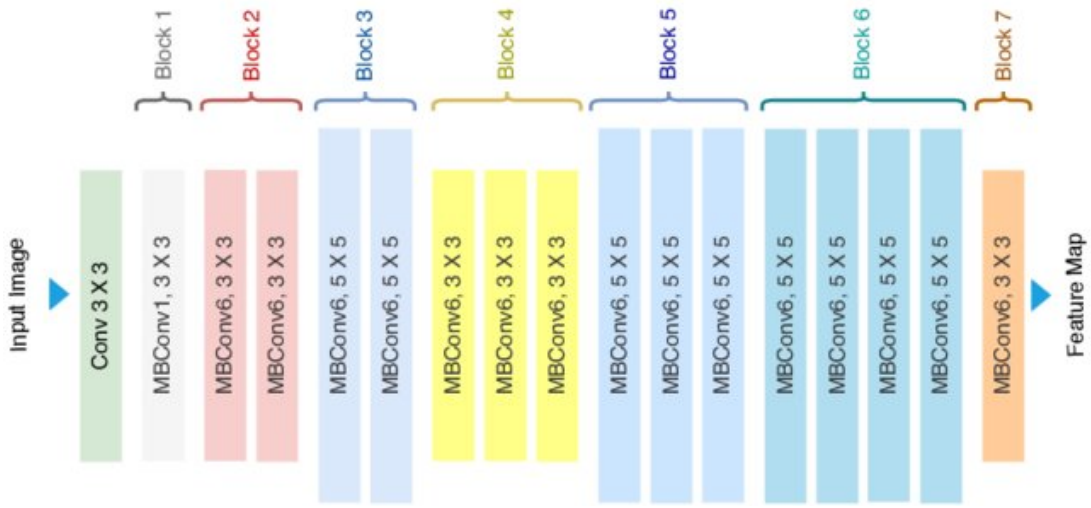



Figura 5 – Arquitetura do EfficientNet-B0 com MBConv como blocos de construção básicos. Fonte: (AHMED et al., 2024).

## 2.4 Vision Transformer

O ViT representa uma das principais inovações recentes em visão computacional, ao transpor os mecanismos originalmente desenvolvidos para processamento de linguagem natural para o domínio das imagens. Proposto por (DOSOVITSKIY et al., 2021), o ViT rompe com a abordagem tradicional das redes convolucionais ao substituir as convoluções por mecanismos de atenção, permitindo capturar relações de longo alcance entre diferentes regiões da imagem.

O funcionamento do ViT inicia com a divisão da imagem em pequenos blocos fixos denominados *patches*, que são linearmente projetados em vetores de mesma dimensão e tratados como uma sequência, de maneira análoga aos *tokens* em modelos de linguagem. A esses vetores são adicionados *embeddings* posicionais, que fornecem informação sobre a localização de cada *patch* dentro da imagem. Essa sequência é processada por camadas do tipo *Transformer*, compostas por mecanismos de autoatenção multi-cabeça (*multi-head self-attention*) e camadas totalmente conectadas, possibilitando que o modelo aprenda dependências globais de forma direta (YIN et al., 2022).

Uma das principais vantagens do ViT está na sua capacidade de modelar contextos globais sem depender de filtros convolucionais fixos. Enquanto as CNNs concentram-se principalmente em capturar padrões locais, como texturas ou bordas, o ViT é capaz de compreender como regiões espacialmente distantes de uma imagem se relacionam entre si, o que se mostra crucial em domínios biomédicos, onde a morfologia celular e a organização tecidual frequentemente dependem de padrões distribuídos por toda a lâmina histológica (Viso.ai, 2024).

Contudo, o ViT apresenta desafios, como a necessidade de grandes volumes de dados para treinamento eficaz, devido à ausência de viés indutivo característico das convoluções. Para mitigar essa limitação, estratégias como o pré-treinamento em grandes bases de imagens (por exemplo, *ImageNet*) seguido de *fine-tuning* em conjuntos específicos têm sido amplamente utilizadas.

No contexto da classificação de imagens histológicas, o ViT traz um diferencial ao complementar arquiteturas convolucionais como a EN. Enquanto a EN captura detalhes finos e padrões locais das imagens, o ViT agrega informações globais e relações estruturais de maior alcance. Essa complementaridade tem motivado a construção de modelos híbridos, nos quais CNNs e ViT são combinados para explorar simultaneamente informações locais e globais, resultando em classificadores mais robustos e precisos.

## 2.5 Processamento de Imagens Histológicas

O processamento de imagens histológicas desempenha um papel fundamental na análise computacional, sendo responsável por preparar as imagens para etapas subsequentes de extração de características e classificação. Essa etapa visa reduzir a variabilidade inerente às amostras, causada por fatores como diferenças na coloração, presença de ruídos, artefatos de digitalização e até divergências na interpretação entre patologistas (IRSHAD et al., 2014; CHEN; CHEN, 2022). Um *pipeline* bem estruturado de processamento é essencial para garantir maior consistência nos dados e, consequentemente, maior robustez nos modelos de aprendizado profundo.

O pré-processamento constitui a fase inicial, cujo objetivo é padronizar as condições de entrada. Técnicas como normalização de corantes são amplamente aplicadas em lâminas coradas por Hematoxilina e Eosina (H&E), a fim de reduzir discrepâncias de tonalidade entre amostras e minimizar variações causadas por protocolos laboratoriais distintos. Além disso, filtros de suavização e remoção de ruídos auxiliam na correção de artefatos, enquanto ajustes de contraste e realce de bordas permitem destacar estruturas celulares relevantes (GONZALEZ; WOODS, 2010; CHEN; CHEN, 2022).

Na etapa de segmentação, busca-se isolar regiões de interesse, como núcleos celulares, membranas ou áreas epiteliais. Diversos métodos podem ser empregados, desde técnicas clássicas de limiarização e watershed até abordagens mais avançadas baseadas em CNNs (ex.: U-Net). A segmentação é crucial para que apenas as estruturas biologicamente relevantes sejam analisadas, evitando que informações irrelevantes interfiram no processo de classificação. Trabalhos recentes destacam o impacto positivo dessa etapa na melhoria dos índices de acurácia de classificadores aplicados à displasia oral (SILVA et al., 2022a).

O pós-processamento atua refinando os resultados obtidos, eliminando regiões espúrias ou corrigindo falhas de segmentação. Técnicas de reconstrução de contornos e filtragem morfológica são comumente aplicadas para assegurar maior precisão na definição das estruturas celulares.

A etapa seguinte, de extração de características, tem como finalidade transformar as regiões segmentadas em representações quantitativas. Tradicionalmente, são utilizados descritores manuais, como estatísticas de textura (Haralick, LBP), medidas de forma e informações espectrais. No entanto, com a evolução do aprendizado profundo, modelos como CNNs, EN e ViT passaram a desempenhar papel central na extração automática de atributos discriminativos, superando limitações das abordagens manuais e permitindo que as redes aprendam representações mais complexas e informativas diretamente a partir das imagens (YIN et al., 2022).

Por fim, a classificação combina as características extraídas com algoritmos de aprendizado, atribuindo rótulos de acordo com o grau de displasia ou tipo de tecido ana-

lisado. Diferentes estratégias têm sido exploradas, incluindo classificadores tradicionais (SVM, *Random Forest*) e redes profundas, isoladas ou em comitês de modelos. Estudos recentes evidenciam que a associação entre um pipeline de processamento consistente e arquiteturas híbridas de aprendizado profundo resulta em ganhos significativos de desempenho, refletindo em maior confiabilidade no sistema de auxílio ao diagnóstico (LONGO et al., 2024; TENGUAM et al., 2024). Um sistema CAD tem como base as fases: pré-processamento, segmentação, descrição e classificação (HE et al., 2022). Na Figura 6, é apresentada as etapas dos processos aplicados as imagens histopatológicas.

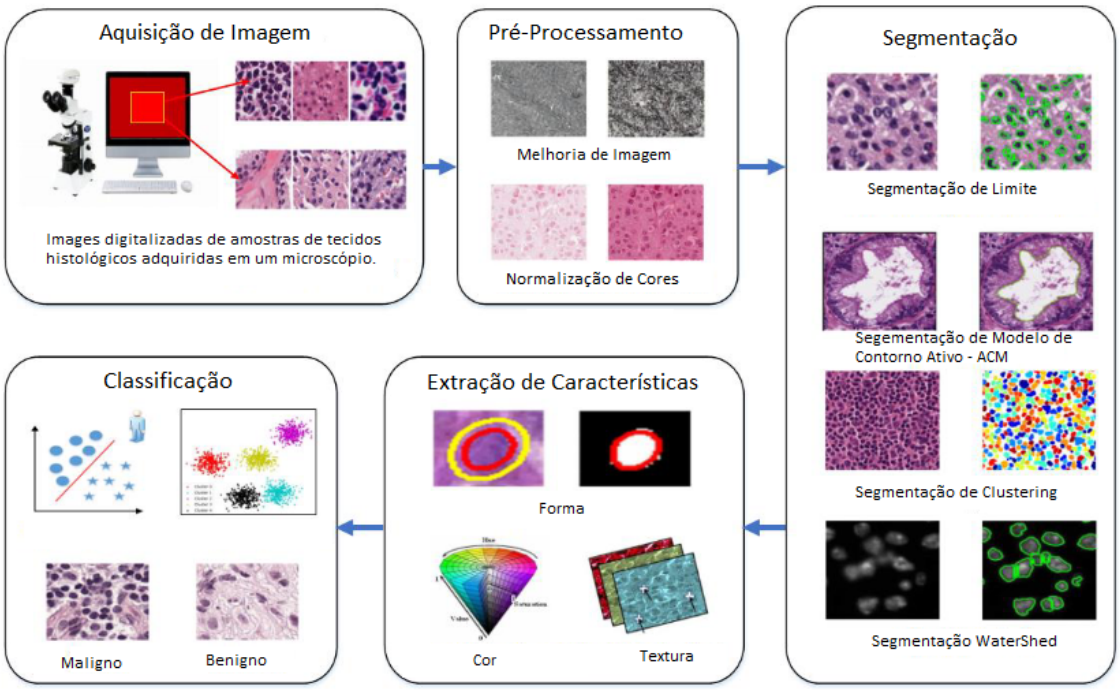



Figura 6 – Sistema para análise e detecção de células em histopatologia. **Adaptado de**: (HE et al., 2022)

Em síntese, o processamento de imagens histológicas por meio do CAD não deve ser visto apenas como uma etapa preliminar, mas como um pilar central para o sucesso de modelos de classificação. A integração entre técnicas potencializa a eficiência, possibilitando avanços significativos na detecção precoce de lesões e displasias.

# 3 Revisão Bibliográfica

De acordo com a pesquisa de (YIN et al., 2022), o Vision Transformer (ViT) representa uma abordagem distinta na classificação de imagens, introduzindo o uso da arquitetura *Transformer*, originalmente desenvolvida para processamento de linguagem natural, em tarefas de visão computacional. Diferentemente das redes neurais convolucionais (do inglês, *convolutional neural networks* - CNNs), o ViT fragmenta as imagens em pequenos blocos denominados *patches* e os processa como uma sequência de *tokens*, permitindo capturar relações globais entre as partes da imagem de maneira mais direta. Embora as CNNs tenham demonstrado grande eficácia em identificar padrões locais por meio de convoluções (CHEN et al., 2021), o ViT explora a atenção global em imagens, superando, em alguns casos, as limitações das CNNs em lidar com dependências de longo alcance. Essa complementaridade entre as duas abordagens abre caminhos para modelos híbridos e avanços significativos na precisão de classificações em diferentes cenários, especialmente em cenários que haja a necessidade de diagnósticos mais precisos, uma vez que o erro de classificação afeta diretamente a taxa de mortalidade.

Já a arquitetura *EfficientNet*, introduzida por Tan e Le (TAN; LE, 2019), complementa os avanços em CNNs e ViT ao propor um escalonamento eficiente e balanceado de redes neurais. Enquanto as CNNs se destacam na identificação de padrões locais (CHEN et al., 2021) e o ViT captura relações globais nas imagens (YIN et al., 2022), a EN aprimora esse processo ao utilizar blocos MBConv (*Mobile Inverted Bottleneck Convolution*), inspirados na arquitetura *MobileNet*, projetada para dispositivos móveis e plataformas com recursos computacionais limitados. Dessa forma, a combinação da extração precisa de características locais com a atenção a relações globais torna-se essencial para minimizar erros de classificação e reduzir taxas de mortalidade.

Além das abordagens baseadas em CNNs e ViTs, diversos trabalhos recentes têm explorado metodologias para a classificação de lesões histológicas da cavidade oral, cada qual adotando estratégias distintas de pré-processamento, extração de atributos e escolha de classificadores. Entre esses, o estudo de (SILVA et al., 2022a) propôs tanto um classificador polinomial baseado em descritores morfológicos e texturais quanto um modelo com *deep features* extraídas de arquiteturas como *AlexNet* e *ResNet-50*. Os autores relataram desempenhos consistentes em classificações binárias, com valores de AUC superiores a 0,90 e acurácia próxima a 98%, destacando a eficácia da combinação entre atributos manuais e redes profundas. Já (LONGO et al., 2024) investigaram a utilização de arquiteturas avançadas, como *Inception-V3* e *VGG-19*, integradas em um comitê homogêneo de classificadores. Esse método alcançou acurácia de 97,97% em cenários binários, evidenciando o impacto positivo do uso de *ensembles* de redes convolucionais no diagnóstico assistido.

Outro trabalho importante com avaliação das lesões de displasia, foi apresentado por (TENGUAM et al., 2024), em que os autores adotaram uma abordagem baseada na seleção de atributos (*Handcrafted Features*) por meio do algoritmo *ReliefF*, combinada a técnicas de otimização como *Particle Swarm Optimization* (PSO). Essa estratégia alcançou valores de acurácia próximos a 100%, demonstrando a força de métodos baseados em atributos explícitos para discriminação de padrões sutis em estágios intermediários da displasia. Além disso, (SILVA et al., 2024) introduziram o banco de dados *OralEpitheliumDB*, utilizado nesta pesquisa, e aplicaram um fluxo experimental que incluía normalização de cor, segmentação nuclear e extração de *deep features* com *ResNet-50*, combinadas com *Random Forest* para classificação multiclasse. Esse estudo obteve acurácia de 94,22%, destacando-se por avaliar simultaneamente os quatro graus de displasia. Por fim, (FERNANDES et al., 2024) exploraram técnicas de compressão de redes convolucionais via decomposição de Tucker em arquiteturas *ResNet*, com validação 10-*fold* e normalização de coloração H&E. Os autores atingiram acurácia de 100% em múltiplos cenários, sugerindo que estratégias de compressão e padronização de imagens podem potencializar o desempenho mesmo em bases de dados reduzidas.

Esses trabalhos ilustram a diversidade de abordagens investigadas no estado da arte, desde abordagens baseadas em descritores morfológicos clássicos até arquiteturas híbridas, ressaltando tanto o potencial das redes profundas quanto os benefícios de técnicas complementares de seleção de atributos, segmentação e *ensembles*. No entanto, em nenhum caso foram avaliadas as associações de modelos híbridos para classificação de displasia com as imagens histológicas da cavidade oral.

# 4 Metodologia

Neste capítulo são descritas as etapas e estratégias adotadas para o desenvolvimento do modelo proposto. Inicialmente, é apresentada a base de dados utilizada, incluindo as características das imagens histológicas e o processo de pré-processamento aplicado. Em seguida, são detalhadas as arquiteturas de redes neurais utilizadas — incluindo modelos de CNNs e uma abordagem baseada no ViT — bem como o processo de treinamento e aumento de dados. Também são explicadas as métricas de avaliação aplicadas para comparar as abordagens testadas.

## 4.1 Introdução

Na Figura 7 apresenta um diagrama de blocos que ilustra as principais etapas metodológicas desenvolvidas neste trabalho. O processo foi dividido em: i) preparação e organização do banco de dados histológico, contendo imagens de tecidos da cavidade oral com diferentes graus de displasia; ii) aplicação de técnicas de aumento de dados (*data augmentation*) para contornar a limitação do número de amostras; iii) treinamento e avaliação de modelos de classificação baseados nas arquiteturas *EfficientNet* (EN), *Vision Transformer* (ViT) e uma rede neural convolucional (do inglês, *convolutional neural networks* - CNNs) de arquitetura simplificada (modelo reduzido), projetada com poucos parâmetros para atuar como um extrator de características leve; iv) análise comparativa de desempenho entre diferentes configurações de treinamento com e sem aumento de dados, incluindo variações no fator de replicação dos dados.

Todas as implementações foram realizadas utilizando a linguagem Python, com o apoio das bibliotecas Keras, TensorFlow e scikit-learn. Os experimentos foram executados em um ambiente computacional com processador AMD Ryzen 5 4500 (6 núcleos, 3.60 GHz), GPU NVIDIA GeForce RTX 3060 com 12 GB de memória VRAM e 16 GB de RAM.

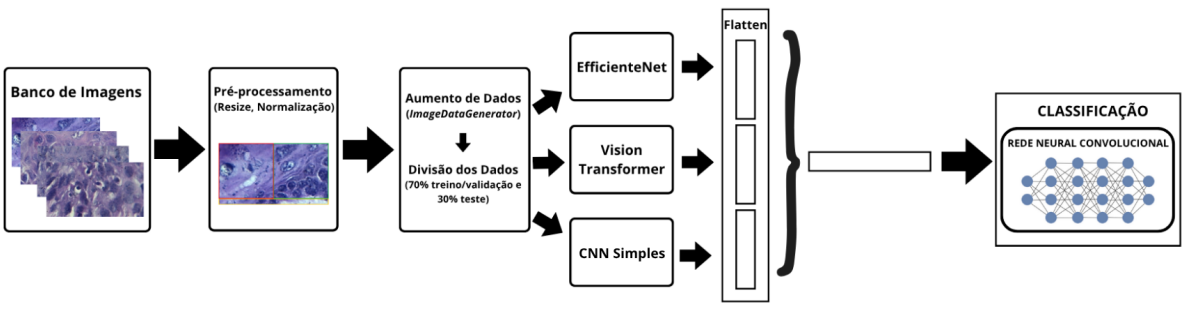


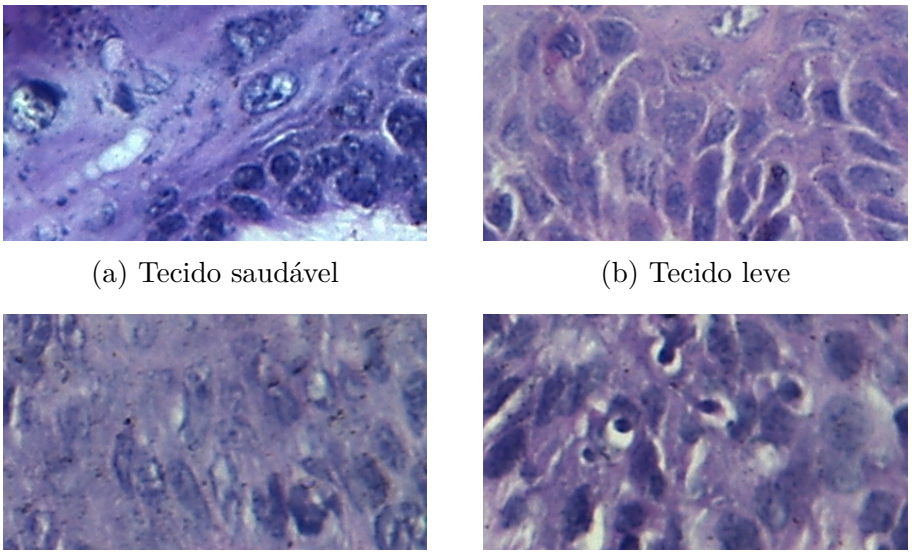

Figura 7 – Fluxo metodológico adotado neste trabalho, desde a coleta dos dados até a avaliação comparativa dos modelos.

## 4.2 Banco de Imagens

Neste trabalho, foi utilizada uma base de dados histológicos conhecida como *Oral Epithelium DB*, disponível publicamente no repositório GitHub <https://github.com/LIPAIGroup/OralEpitheliumDB>, composta por imagens de tecidos da cavidade oral de camundongos da linhagem C57BL/6, que foram submetidos experimentalmente ao agente carcinógeno 4-nitroquinolina-N-óxido (4NQO), dissolvido na água fornecida aos animais. O experimento foi conduzido sob aprovação do Comitê de Ética na Utilização de Animais da Universidade Federal de Uberlândia (nº 038/09), e consistiu em dois grupos com 15 camundongos cada. O primeiro grupo foi eutanasiado após 16 semanas de exposição ao 4NQO, enquanto o segundo grupo permaneceu vivo por mais 16 semanas com acesso apenas à água filtrada, totalizando 32 semanas até o sacrifício (SILVA et al., 2024).

Após o término do experimento, as línguas dos animais foram removidas, processadas e incluídas em lâminas histológicas coradas, que foram digitalizadas por meio de um microscópio óptico Leica DM500 com aumento de 400x. As imagens resultantes foram analisadas por dois patologistas especialistas, que classificaram as amostras de acordo com o grau de displasia, segundo os critérios estabelecidos por Lumerman, Freedman e Kerpel (1995). Para este estudo, foram selecionadas as amostras das classes tecido saudável, leve, moderado e severo sendo 114 imagens para cada classe de displasia leve, moderado e severo.

Na Figura 8 são apresentados exemplos de imagens das classes de imagens investigadas neste trabalho.



(a) Tecido saudável (b) Tecido leve

(c) Tecido moderada (d) Tecido severa

Figura 8 – Exemplos de imagens representativas das quatro classes do *Oral Epithelium DB*.

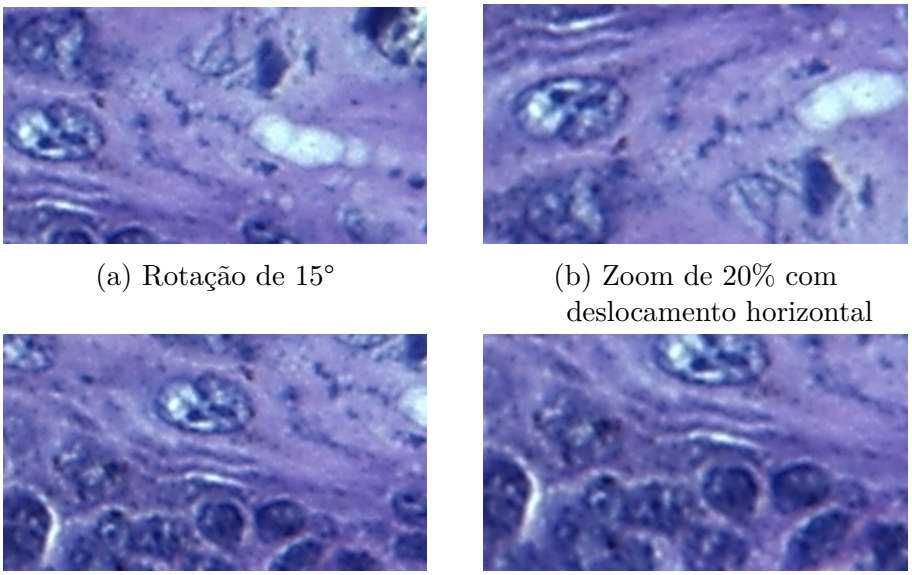
## 4.3 Aumento de Dados

Devido à quantidade limitada de amostras no conjunto de dados, foi aplicado o aumento de dados (do inglês, *data augmentation* – DA) com o intuito de melhorar a capacidade de generalização da rede treinada e mitigar o sobreajuste (*overfitting*).Para garantir a validade e imparcialidade da avaliação do modelo, a base de dados foi primeiramente dividida de forma estratificada para manter a proporção de classes: 70% das imagens foram destinadas ao conjunto de Treinamento e Validação (com 70% e 30% deste subconjunto, respectivamente, resultando em 49% para Treinamento e 21% para Validação do total) e 30% os restantes foram reservados para o conjunto de Teste. O aumento de dados foi aplicado exclusivamente sobre o conjunto de Treinamento.

As operações de aumento aplicadas incluíram rotações aleatórias de até 20°, deslocamentos horizontais e verticais limitados a 10% do tamanho da imagem, além de cisalhamento (*shear*) com intensidade de até 30%. Também foi aplicado um fator de *zoom* aleatório de até 30%, simulando aproximações nas imagens. As áreas faltantes geradas por essas transformações foram preenchidas utilizando o modo *reflect*, o qual reflete os pixels vizinhos para preencher os espaços vazios, mantendo a coerência visual. Para a implementação, foi utilizada a classe `ImageDataGenerator` da biblioteca Keras, que permite realizar transformações em tempo real durante o treinamento.

O aumento foi implementado de forma iterativa, com o fator de replicação sendo variado para análise. Foram testados fatores de multiplicação do conjunto de Treinamento de 1× (sem replicação adicional), 2×, 3× e 4×. Dessa forma, para um fator 4×, por exemplo, cada imagem do conjunto de treinamento foi transformada em três versões distintas adicionais, totalizando quatro variações (a original mais três aumentadas). Essa abordagem possibilitou a ampliação substancial do conjunto de treinamento, aumentando a robustez do modelo frente a variações nas imagens e permitindo a identificação do ponto ótimo de saturação dos dados.

Na Figura 9 são apresentadas amostras dessas transformações aplicadas sobre uma mesma imagem de um tecido saudável.



(a) Rotação de 15°

(b) Zoom de 20% com deslocamento horizontal

(c) Cisalhamento vertical com deslocamento

(d) Transformações combinadas: rotação, zoom e cisalhamento

Figura 9 – Exemplos de transformações aplicadas sobre a mesma imagem original com o uso do `ImageDataGenerator`.

## 4.4 Modelo de Aprendizagem Profunda

O Modelo de aprendizagem profunda proposto adota uma Arquitetura Híbrida que funciona como um Comitê de Classificadores. O objetivo é que cada componente analise a imagem histológica sob uma perspectiva diferente, direcionadas aos detalhes finos, visão panorâmica do tecido e extração de padrões básicos, e combine essas informações para tomar uma decisão mais robusta sobre o grau de displasia.

Na Figura 10 é ilustrado o fluxo metodológico adotado. A arquitetura integra três módulos principais, cujas saídas são combinadas para a classificação final:

**i)** Módulo EfficientNet (EN): Atua como o especialista em características locais detalhadas. Ele identifica e extrai padrões finos (bordas, texturas) de maneira eficiente.

**ii)** Módulo Vision Transformer (ViT) Adaptado: Responsável pela interpretação do contexto global. Este módulo busca conexões entre regiões distantes da imagem (dependências de longo alcance), essenciais para compreender a arquitetura geral do tecido.

**iii)** Módulo CNN Simplificado: Uma Rede Neural Convolucional (CNN) de arquitetura enxuta, atuando como um extrator de padrões básicos e texturas fundamentais.

As saídas vetoriais dos três caminhos são unidas por uma operação de concatenação para formar uma representação unificada da imagem, que é então processada pelas camadas densas finais para determinar a classificação.

### Caminho 1: EN — O Foco nos Detalhes (*Bottleneck*)

O primeiro bloco utiliza a arquitetura `EfficientNetV2B0`, aproveitando pesos previamente treinados no grande conjunto de dados *ImageNet*. Essa rede é empregada como um extrator de características, tendo sua camada final de classificação removida (`include_top=False`).

A saída da EN é resumida espacialmente por meio de uma operação de *GlobalAveragePooling2D* (média global das ativações), estabilizada com ***BatchNormalization*** para acelerar o treinamento. Em seguida, uma camada densa com 256 unidades e ativação ReLU refina essas representações. O caminho se conclui com a operação de ***Flatten***, que transforma o resultado em um vetor unidimensional de características para a concatenação.

### Caminho 2: ViT Adaptado — A Visão Global

Este caminho simula a filosofia do ViT, focada em relações de longo alcance. A imagem de entrada é inicialmente reestruturada (***Reshape***) para simular a divisão em *patches* (pequenas regiões) e seus valores de *pixels* são normalizados entre 0 e 1, garantindo a consistência dos dados.

A estrutura passa por uma sequência de camadas convolucionais (com 64 e 128 filtros e ativação ReLU), que extraem padrões de complexidade crescente. Para reduzir a dimensionalidade e focar nas regiões mais informativas, são aplicadas camadas de ***MaxPooling2D***. Por fim, a operação de ***Flatten*** converte a saída em um vetor de características para a junção final.

### Caminho 3: CNN Simplificado — O Extrator Básico (*Neck*)

Representado pelo *Neck Section*, este caminho é projetado para ser leve. Ele se inicia com uma camada convolucional de 256 filtros e *kernel* 7×7, com *stride* de 2, que promove uma redução espacial agressiva para capturar um resumo da estrutura global da imagem.

Para otimizar o aprendizado, são aplicadas uma normalização por lote e uma ativação PReLU (*Parametric ReLU*), que permite maior expressividade ao modelo. O processo é finalizado por uma camada de ***MaxPooling2D*** e a operação de ***Flatten***, que transforma os dados em um vetor unidimensional. Este bloco é crucial para garantir a captura de texturas e padrões básicos.

As saídas vetoriais dos três caminhos (EN, ViT e CNN) são combinadas em uma única estrutura vetorial, representando as informações extraídas das imagens histológicas. Essa concatenação é então utilizada como entrada para camadas densas totalmente conectadas, responsáveis pela tarefa de classificação final. Essa abordagem híbrida visa capturar tanto padrões globais quanto locais, aproveitando as forças individuais de cada arquitetura para obter um desempenho superior em contextos de base de dados limitada e variabilidade morfológica, como é o caso das imagens de displasias.

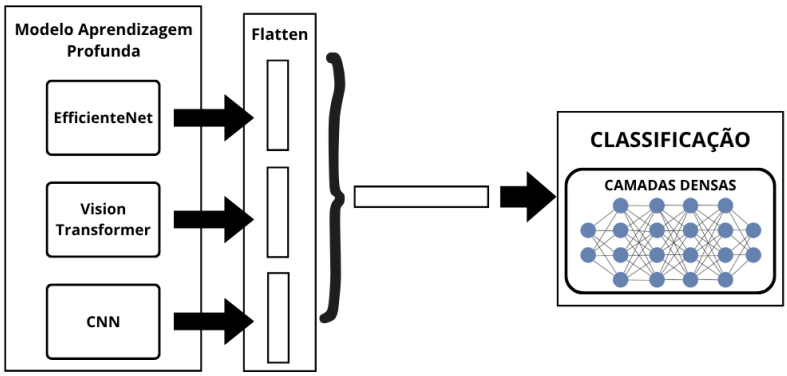



Figura 10 – Fluxo da metodologia adotada na etapa de aprendizado profundo, após a coleta dos dados até a classificação do modelo.

## 4.5 Arquitetura EfficientNet

Para a tarefa de classificação de imagens histológicas, adotou-se a arquitetura EfficientNet-B0 proposta (TAN; LE, 2019). EfficientNet é reconhecida por seu *método de scaling composto*, que expande simultaneamente a profundidade, a largura e a resolução da rede de forma balanceada, controlada por um único coeficiente de escala $\phi$. Isso permite aumentar a capacidade representacional do modelo sem um crescimento desproporcional do custo computacional.

Em termos práticos, a EfficientNet-B0 foi pré-treinada na base *ImageNet* e posteriormente ajustada (*fine-tuned*) sobre nosso conjunto de imagens extraídas do *Oral Epithelium DB*. Cada bloco da rede utiliza a estrutura *Mobile Inverted Bottleneck Convolution* (MOBConv), composta por:

1. Uma convolução 1×1 para expandir o número de canais de entrada;
2. Uma convolução *depthwise*, de tamanho 3×3 pixels, para extrair características espaciais de baixo custo;
3. Uma projeção, de tamanho 1×1 pixel, para reduzir novamente o número de canais;
4. Um módulo *Squeeze-and-Excitation* (SE) para recalibrar a atenção por canal.

Além disso, a EN incorpora o módulo SE dentro de cada bloco MBConv, reforçando dinamicamente os recursos mais informativos para a classificação. A escolha da versão B0 deveu-se ao seu equilíbrio entre acurácia e eficiência, fator determinante para aplicações em imagens biomédicas, onde o poder de processamento pode ser limitado (LI et al., 2022).

Na Figura 11 ilustra a arquitetura do módulo MBConv.


Figura 11 – Arquitetura com as camadas conectadas entre os módulos MBConv. Fonte: (LI et al., 2022)

Durante os testes, o modelo foi treinado utilizando imagens aumentadas, com otimização por Adam e redução da taxa de aprendizado em função da melhora da validação. Essa combinação de *compound scaling* e blocos MBConv-SE possibilitou ao modelo aprender representações discriminativas mesmo com o número restrito de amostras de células.

## 4.6 Vision Transformer

Embora a proposta do Vision Transformer (ViT) original utilize mecanismos de atenção multi-cabeça e *tokens* posicionais sobre sequências de *patches* extraídos das imagens (DOSOVITSKIY et al., 2021), neste trabalho foi implementada uma abordagem inspirada conceitualmente nessa estrutura, porém adaptada para maior simplicidade computacional.

A arquitetura adotada consiste na divisão da imagem em regiões menores, simulando os *patches* do ViT. A imagem é primeiramente reformatada para esse padrão via operação de `Reshape`, seguida de uma normalização dos *pixels*. Posteriormente, esses *patches* são processados por uma sequência de camadas convolucionais com ativação ReLU e operações de `MaxPooling`, com o objetivo de extrair características locais de cada região da imagem. Ao final, os mapas de ativação são achatados (`Flatten`) e direcionados à etapa de classificação. Na Figura 12 é ilustrado a estrutura geral de um ViT.

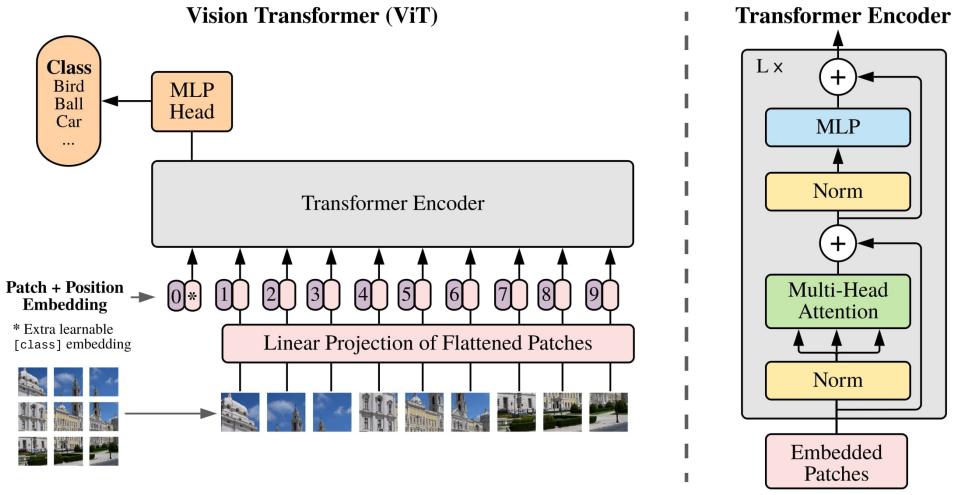



Figura 12 – Arquitetura do ViT. Adaptado de (DOSOVITSKIY et al., 2021)

A arquitetura utilizada no modelo mantém parte da filosofia dos ViTs, ou seja, o processamento localizado de regiões da imagem, mas sem aplicar mecanismos de atenção, *embeddings* posicionais ou *tokens* de classe. Em vez disso, foi utilizado uma abordagem baseada em convoluções sucessivas sobre os *patches*, tornando o modelo mais leve e acessível computacionalmente, especialmente útil em cenários com bases de dados reduzidas e infraestrutura computacional limitada.

## 4.7 Rede Neural Convolucional

A arquitetura baseada em Redes Neurais Convolucionais (do inglês, *convolutional neural networks* - CNNs) utilizada neste trabalho foi projetada como um dos caminhos paralelos dentro do modelo híbrido proposto. Essa estrutura, representada pelo bloco *neck_section*, tem por objetivo atuar como um extrator de características visuais fundamentais, como texturas e padrões locais presentes nas imagens histológicas.

Inicialmente, as imagens de entrada são submetidas a uma convolução com filtro de $7 \times 7$ *pixels* e *stride* de 2, o que permite uma rápida redução da resolução espacial, ao mesmo tempo em que amplia a área de percepção do filtro. Em seguida, é aplicada a técnica de normalização por lote (*BatchNormalization*), que contribui para acelerar o treinamento e mitigar problemas de covariância interna ao longo das camadas. Essa função foi empregada uma vez que a função de ativação (PReLU) pode favorecer a convergência e aumentar a acurácia em arquiteturas menores, contribuindo para a detecção de padrões mais sutis presentes nos dados.

Após essa etapa, é aplicado um *MaxPooling2D* com filtro de $3 \times 3$ *pixels* e *stride* de 2, que tem como função reduzir a dimensionalidade dos mapas de ativação e reforçar as características mais relevantes. A saída do bloco é então convertida em um vetor

unidimensional por meio da operação de *Flatten*, de modo a preparar os dados para sua posterior fusão com as demais representações extraídas pelos outros caminhos (EN e ViT).

A escolha de uma arquitetura CNN simples, com poucos parâmetros e baixa profundidade, visa garantir a eficiência computacional do modelo e ao mesmo tempo contribuir de forma complementar na representação da imagem. Ao atuar de forma paralela às demais arquiteturas, a CNN oferece uma perspectiva mais tradicional da extração de características, baseada em convoluções e *pooling*, que se mostrou útil especialmente na captura de padrões morfológicos consistentes em imagens histológicas.

Na Figura 13 é ilustrado a sequência de operações empregadas nos blocos do modelo simplificado da CNN. O esquema visual permite observar de forma clara as etapas de processamento aplicadas às imagens, desde a convolução inicial até a vetorização dos mapas de ativação. Destacando como as transformações espaciais são aplicadas antes da integração com os demais módulos da rede híbrida.

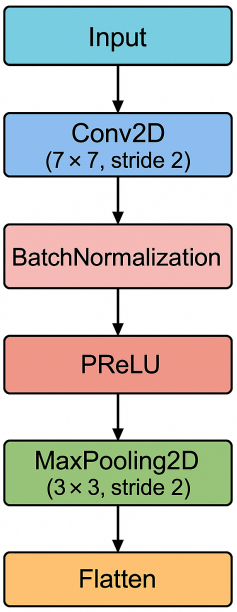



Figura 13 – Arquitetura de CNN utilizado neste trabalho.

## 4.8 Métricas de Avaliação

A seleção e análise das métricas de desempenho são etapas cruciais para a validação, especialmente em tarefas de classificação em contextos biomédicos, onde a precisão e a confiabilidade das predições são críticas (TENGUAM et al., 2024). Uma estratégia fundamental para compreender o comportamento do modelo e identificar oportunidades de melhoria é a análise das métricas geradas ao longo do treinamento e da validação. Neste trabalho, essas métricas também foram utilizadas para avaliar o impacto da normalização e do aumento de dados sobre a capacidade preditiva do modelo.

Para a realização das avaliações, foram consideradas quatro quantidades básicas derivadas da matriz de confusão, obtidas a partir da comparação entre os rótulos reais e os rótulos previstos pelo modelo em cada teste. A interpretação dessas quantidades é contextualizada pelo problema da classificação de displasia:

- Verdadeiros Positivos (VP): Uma amostra com displasia é corretamente identificada como tendo displasia.
- Verdadeiros Negativos (VN): Uma amostra saudável é corretamente identificada como saudável.
- Falsos Positivos (FP): Uma amostra saudável é incorretamente classificada como tendo displasia (falso alarme).
- Falsos Negativos (FN): Uma amostra com displasia é incorretamente classificada como saudável (erro de diagnóstico grave).

Com base nesses valores, são calculadas as seguintes métricas principais, utilizadas na avaliação dos modelos e na comparação com o estado da arte: Acurácia (*Acc*), Precisão (*Prec*), Revocação (*Recall*), F1-Score e Área sob a Curva ROC (*AUC*).

O cálculo das métricas foi realizado com o auxílio da biblioteca `scikit-learn`, por meio da função `classification_report`, que fornece as métricas para cada classe, além de médias ponderadas (*weighted*) e médias macro. Além disso, a métrica *AUC* foi obtida através da função `roc_auc_score`, que avalia a separabilidade entre as classes para diferentes limiares de decisão.A seguir, estão descritas as principais métricas utilizadas:

A métrica *Acc* mede a proporção de previsões corretas em relação ao total de amostras avaliadas:

$$Acc = \frac{VP + VN}{VP + VN + FP + FN}. \tag{4.1}$$

Já a métrica *Prec* avalia quantas das amostras classificadas como positivas realmente pertencem à classe positiva:

$$Prec = \frac{VP}{VP + FP}. \tag{4.2}$$

A medida *Recall* mede a proporção de amostras da classe positiva que foram corretamente identificadas:

$$Recall = \frac{VP}{VP + FN}. \tag{4.3}$$

O *F1-Score* representa a média harmônica entre precisão e revocação, sendo especialmente útil quando há desbalanceamento entre as classes:

$$\text{F1-Score} = 2 \cdot \frac{Prec \cdot Rev}{Prec + Rev}. \tag{4.4}$$

A métrica *AUC* quantifica a capacidade do modelo em distinguir entre as classes. Quanto mais próxima de 1 for a AUC, maior a separabilidade entre classes:

$$AUC = \int_0^1 TPR(FPR^{-1}(x))dx. \tag{4.5}$$

A ACC foi inicialmente utilizada como métrica de avaliação, calculada como a razão entre o número de previsões corretas e o total de amostras do conjunto de testes. Embora forneça uma visão geral do desempenho, sua eficácia é limitada em cenários com classes desbalanceadas, como no caso deste trabalho, que envolve imagens de tecido saudável e severo. Por isso, a avaliação foi complementada com métricas mais robustas como *F1-Score*, *AUC*, precisão e revocação, permitindo uma análise mais abrangente do comportamento do classificador.

Além disso, foram conduzidas análises comparativas entre diferentes configurações do modelo, variando-se o fator de aumento de dados. As métricas de F1-Score, AUC e acurácia foram utilizadas para a construção de gráficos que auxiliaram na identificação da configuração mais robusta e com maior capacidade de generalização.

# 5 Experimentos e Análise dos Resultados

Neste capítulo, apresentam-se os experimentos e resultados do modelo híbrido proposto. Os testes foram realizados com imagens histológicas da cavidade oral do conjunto *OralEpitheliumDB*, considerando inicialmente as classes “saudável” e “severa” e, posteriormente, combinações binárias entre os diferentes níveis de lesão, “leve“ e “moderada“ .

São descritos os efeitos da variação do *batch size* e do aumento de dados no desempenho do modelo, seguidos da análise de métricas como Acurácia, *F1-Score*, *Precision* e AUC, além de curvas de aprendizado e matrizes de confusão. Por fim, os resultados são comparados com estudos da literatura para posicionar o modelo no contexto do estado da arte.

## 5.1 Avaliação do modelo com variação do Batch Size

Para investigar o impacto do tamanho do lote (*batch size*) no desempenho do modelo híbrido proposto, foram realizados experimentos com valores de 1, 8, 16, 32 e 64, mantendo fixos os demais parâmetros: 100 épocas de treinamento, ausência de aumento de dados e classificação binária apenas entre as classes “saudável” e “severa”.

Na Figura 14 é observado a variação das métricas AUC, Acurácia, F1-Score e Precisão no conjunto de testes para cada configuração.

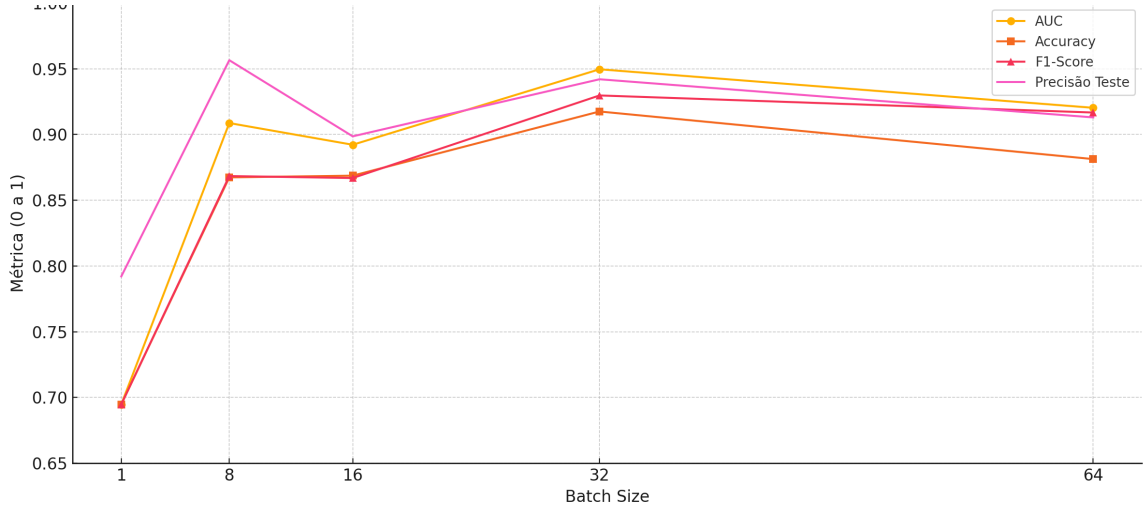



Figura 14 – Desempenho do modelo híbrido proposto em função do *batch size*, para a classificação binária entre tecidos saudáveis e severos, com 100 épocas e sem aumento de dados.

Observa-se que valores muito baixos de classificação pelas métricas quando confi-

gurado *batch size* como 1, já com aumento do *batch size* para 8 proporcionou um salto significativo de desempenho, especialmente no *F1-Score* e na AUC. Para valores intermediários, como 16 e 32, o modelo manteve métricas elevadas e mais estáveis, alcançando seu melhor desempenho geral com *batch size* igual a 32, onde AUC e *F1-Score* ultrapassaram 0,94. Já o valor de 64 levou a uma leve queda de desempenho, possivelmente associada à redução da variabilidade do gradiente, o que pode dificultar a generalização.

Além da análise das métricas em função do *batch size*, é igualmente relevante entender como as métricas evoluem ao longo do treinamento. A observação das curvas de acurácia, AUC e *F1-Score* para os conjuntos de treino e validação permitiu identificar padrões de convergência, eventuais sinais de *overfitting* e a estabilidade do aprendizado. Essa análise temporal forneceu evidências adicionais sobre a adequação do *batch size* escolhido e sua influência na capacidade de generalização do modelo. Na Figura 15, observa-se que as evoluções de cada métrica com o passar das 100 épocas.

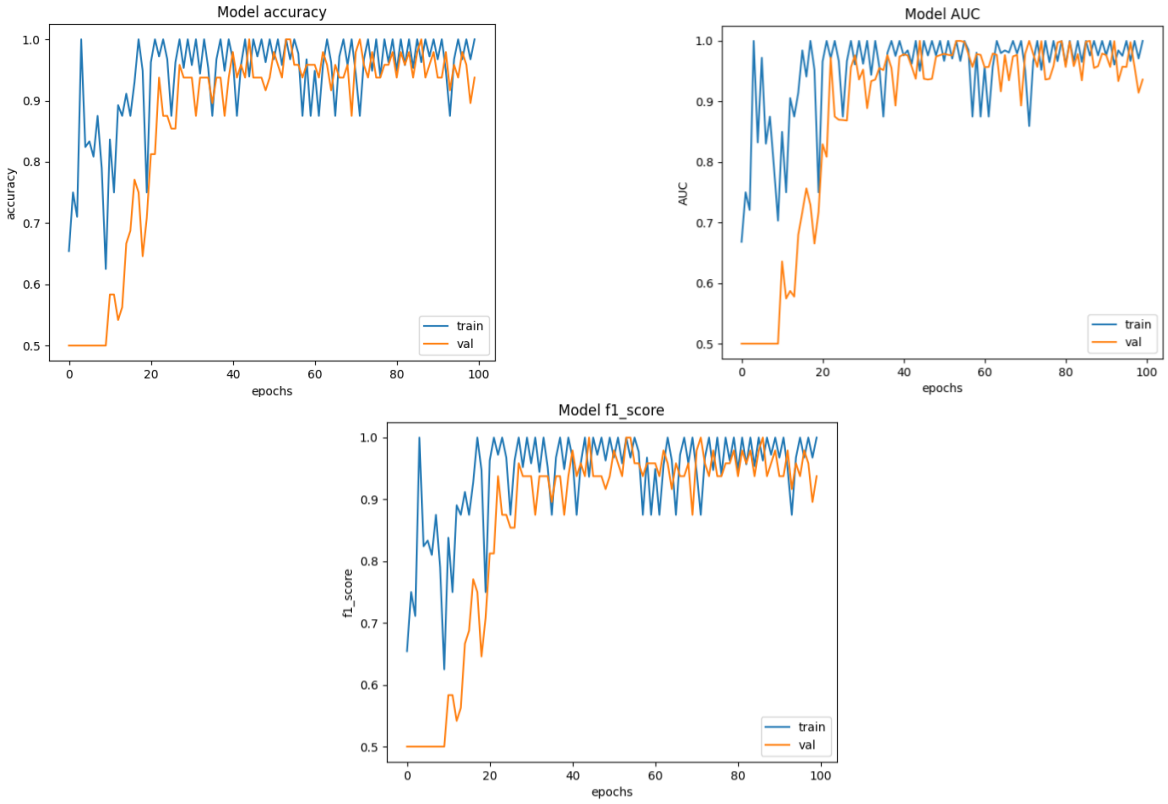



Figura 15 – Curvas de evolução das métricas: (a) Acurácia; (b) AUC e (c) *F1-Score*. Para treino e validação ao longo de 100 épocas, considerando a configuração de *batch size* que apresentou melhor desempenho.

Observa-se que após as primeiras 20 épocas, o modelo converge para valores elevados e estáveis nas três métricas analisadas. No início do treinamento, a discrepância entre treino e validação é maior, refletindo o ajuste inicial dos pesos; porém, ao longo das épocas, as curvas se aproximam, indicando que o modelo conseguiu generalizar bem para o conjunto de validação. A acurácia (ver Figura 15(a)) estabiliza acima de 0,92, en-

quanto a AUC (Figura 15(b)) mantém-se próxima de 1,0, demonstrando alta capacidade discriminativa. O F1-Score (Figura 15(c)) segue a mesma tendência, superando 0,93 e permanecendo estável até o final do treino, evidenciando um equilíbrio entre precisão e sensibilidade.

Por fim, para validar o desempenho obtido com o *batch size* de 32, foi gerada a matriz de confusão referente à classificação binária entre tecidos saudáveis e severos, conforme mostrado na Figura 16.

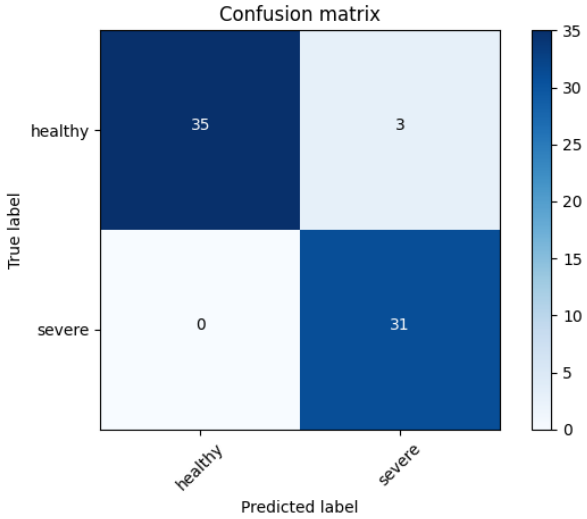



Figura 16 – Matriz de confusão obtida na configuração com *batch size* de 32 para a classificação binária no conjunto de teste.

O resultado indica que o modelo classificou corretamente 35 amostras saudáveis e 31 severas, apresentando apenas 3 classificações incorretas, todas referentes a amostras saudáveis identificadas como severas. A ausência de falsos negativos para a classe "severa" é um ponto relevante, pois indica que, no conjunto de teste, nenhuma lesão severa foi erroneamente classificada como saudável, o que é desejável em aplicações clínicas. Esses resultados indicam que, no contexto deste trabalho, valores intermediários de *batch sizes* tendem a oferecer o melhor equilíbrio entre estabilidade do treinamento e capacidade de generalização, enquanto valores extremos (muito baixos ou muito altos) apresentam desvantagens evidentes.

## 5.2 Análise após Aumento de Dados

Considerando a baixa disponibilidade de imagens para realizar a classificação — limitação já discutida anteriormente neste trabalho na Seção 4.2, que decorre da natureza restrita do banco de dados utilizado (*Oral Epithelium DB*), composto apenas por amostras das classes tecido saudável, leve, moderado e severo sendo 114 imagens para cada classe e diplasia nos níveis classificados como leve, moderado e severo — torna-se necessário

adotar estratégias para ampliar a variabilidade do conjunto de treino sem a necessidade de coleta de novos dados. Essa escassez não apenas restringe a diversidade de padrões morfológicos disponíveis para o aprendizado do modelo, mas também aumenta o risco de *overfitting*, comprometendo sua capacidade de generalização para dados inéditos.

Para mitigar esse problema e avaliar o impacto do aumento de dados (*data augmentation*) no desempenho do modelo proposto, foram conduzidos experimentos considerando diferentes fatores de replicação sintética das amostras originais. Foram testados quatro níveis de aumento, correspondendo à multiplicação do conjunto original de treino por 1× (sem replicação adicional), 2×, 3× e 4×. O fator 1× corresponde ao treinamento sem aplicação de aumento de dados, enquanto os demais representam a replicação do conjunto, aplicando as transformações descritas na Seção 4.3, como rotações, deslocamentos, cisalhamentos e *zoom*.

Ainda utilizando as mesmas métricas para análise de desempenho do modelo. O resultado comparativo encontra-se na Figura 17.

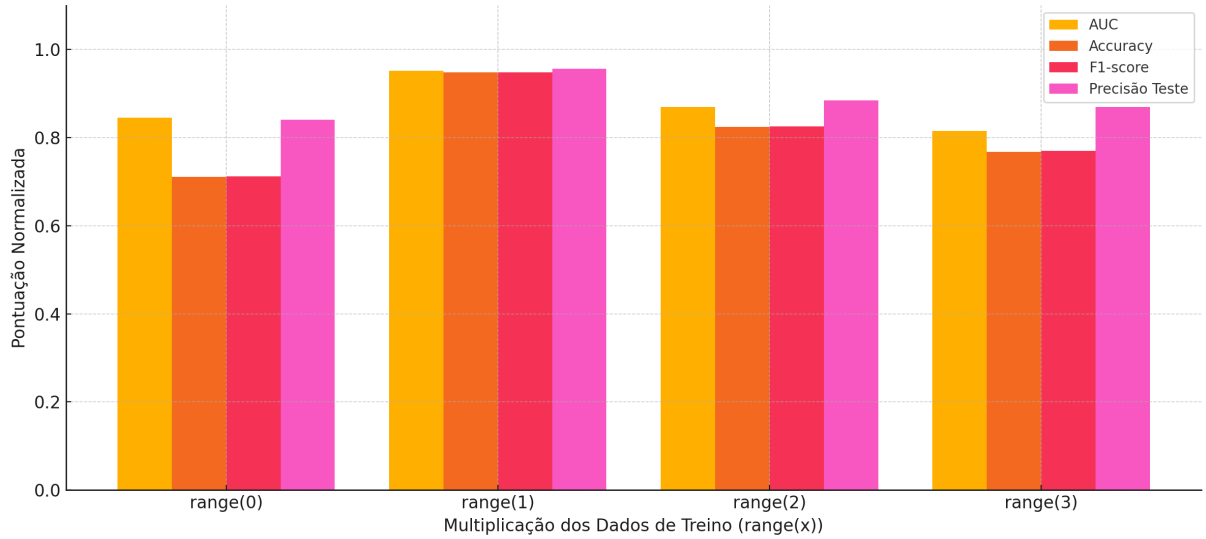



Figura 17 – Comparativo de desempenho por nível de aumento de dados para a classificação binária entre tecidos saudáveis e severos.

Observa-se que o uso moderado de aumento de dados (`range(1)`) proporcionou uma elevação significativa no desempenho do modelo em todas as métricas, atingindo valores próximos ao máximo obtido nos experimentos. No entanto, níveis mais elevados de replicação sintética (`range(2)` e `range(3)`) não resultaram em ganhos adicionais, e em alguns casos ocasionaram redução no desempenho, sugerindo que o excesso de amostras sintéticas pode introduzir variações irrelevantes ou redundantes para a tarefa de classificação.

Esses resultados reforçaram a importância de um balanceamento adequado no uso do aumento de dados: enquanto níveis moderados contribuem para melhorar a capacidade

de generalização. A aplicação de um número alto de amostras sintéticas degradou o desempenho, possivelmente devido à introdução de ruído que não contribui para a discriminação entre classes.

## 5.3 Classificação binária dos níveis de displasia

Com o objetivo de compreender o desempenho do modelo proposto em diferentes cenários de classificação binária, foram realizados experimentos considerando combinações par-a-par entre os quatro níveis de severidade presentes no conjunto de dados. Dessa forma, buscou-se avaliar como o modelo responde à tarefa quando a distinção entre classes apresenta maior ou menor complexidade visual, variando de contrastes acentuados (exemplo: saudável vs. severo) a distinções mais sutis (exemplo: leve vs. moderado).

Na Figura 18 é representado o desempenho do modelo em termos de AUC, Acurácia, F1-Score, Precisão por classe e Precisão no conjunto de teste para cada cenário analisado.

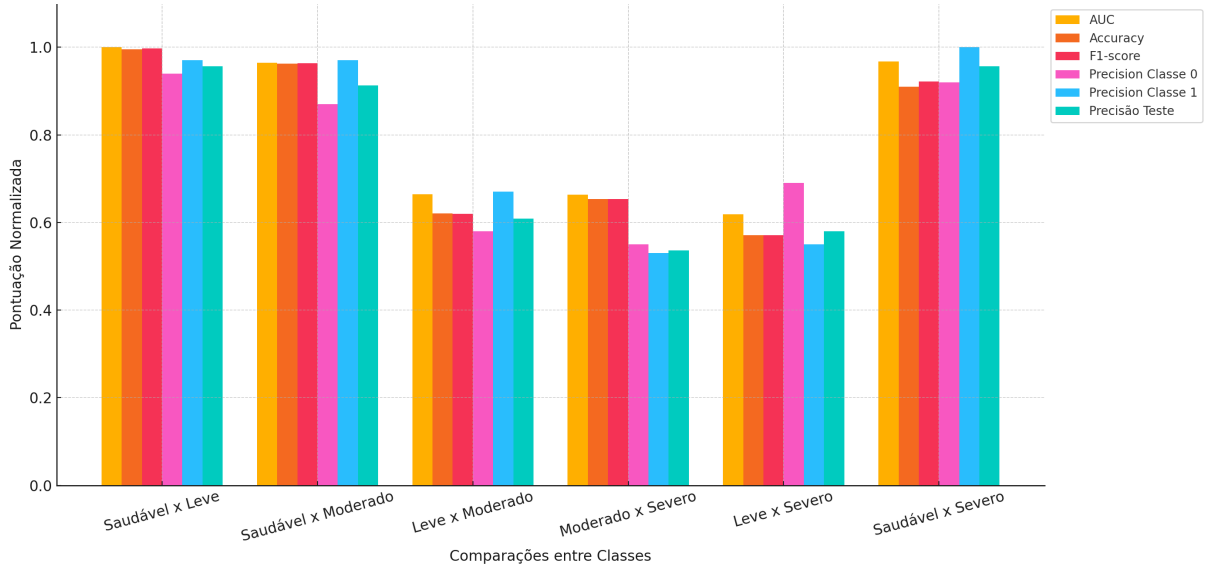



Figura 18 – Comparativo de desempenho do modelo em classificações binárias entre diferentes pares de classes.

Observa-se que as combinações que envolvem a classe saudável, em especial *saudável vs. severo*, apresentaram desempenho consistentemente elevado, com métricas próximas ou iguais a 1,0. Esse resultado indica que o contraste visual e estrutural entre essas classes é suficiente para permitir ao modelo separar as amostras com alta confiança. Por outro lado, os cenários *leve vs. moderado* e *leve vs. severo* obtiveram valores inferiores, sugerindo maior sobreposição de características entre as classes e, consequentemente, maior dificuldade de separação pelo modelo.

Na Tabela 1 são apresentados os valores médios das principais métricas de avaliação — AUC, Acurácia e *F1-Score* — obtidos para cada cenário de classificação binária, considerando pares de classes distintas. Nesta análise, a métrica AUC foi utilizada para avaliar a capacidade discriminativa do modelo independentemente do ponto de corte adotado, enquanto a acurácia e o *F1-Score* medem, respectivamente, a proporção de acertos e o equilíbrio entre precisão e revocação.

Tabela 1 – Métricas médias obtidas para cada cenário de classificação binária.

| **Par de Classes** | **AUC** (%) | **Acurácia** (%) | **F1-Score** (%) |
|---|---|---|---|
| Saudável x Leve | 100 | 99,71 | 99,56 |
| Saudável x Moderado | 96,45 | 96,28 | 96,33 |
| Leve x Moderado | 66,48 | 62,12 | 61,97 |
| Moderado x Severo | 66,33 | 65,38 | 65,39 |
| Leve x Severo | 61,90 | 57,08 | 57,13 |
| Saudável x Severo | 96,80 | 90,96 | 92,19 |
| **Média ± DP** | 81,99 ± 16,25 | 78,59 ± 17,15 | 78,43 ± 17,17 |

Esses resultados evidenciam que a separação entre classes com maior distância de severidade resulta em métricas elevadas, indicando que o modelo consegue identificar de forma consistente as diferenças visuais mais marcantes. Em contraste, a classificação entre classes intermediárias apresenta desempenho inferior, possivelmente devido à maior similaridade morfológica e à menor presença de padrões discriminativos claros nas imagens dessas categorias. Essa tendência reforça a hipótese de que o grau de distinção visual entre classes está diretamente relacionado ao desempenho do classificador.

## 5.4 Discussão e Análise Comparativa com o Estado da Arte

Com o objetivo de posicionar o modelo no contexto de pesquisas recentes, realizou-se uma análise de desempenho do modelo com trabalhos da literatura que abordam problemas de classificação binária e multi-classes de imagens histológicas da cavidade oral. Os trabalhos utilizados adotam bases de dados comparáveis, métricas como a AUC e acurácia, e configuram experimentos voltados à diferenciação entre grupos. A Tabela 2 apresenta os valores de acurácia de diferentes métodos da literatura, uma análise mais aprofundada evidenciando diferenças metodológicas relevantes que ajudam a contextualizar os resultados.

Tabela 2 – Análise de desempenho entre diferentes métodos da literatura e o modelo proposto para classificação binária de imagens histológicas da cavidade oral.

| **Estudo** | **Atributos** | **Classificador** | **Classificação** | $A_{CC}$(%) |
|---|---|---|---|---|
| (SILVA et al., 2022a) | Características de textura e morfológicas | HOP | multi-classe | 92,40 |
| (LONGO et al., 2024) | Inception-V3 e VGG-19 | Comitê Homogêneo | binário | 97,97 |
| (SILVA et al., 2022b) | *Deep features* com AlexNet e ResNet-50 | HOP | binário | 98,00 |
| (TENGUAM et al., 2024) | HFs selecionados via ReliefF + PSO | HFs | binário | 100 |
| (SILVA et al., 2024) | Segmentação de Imagem e Normalização com ResNet50 (*Deep features*) | RF | multi-classe | 94,22 |
| **Método Proposto** | EfficientNet + ViT + CNN tradicional | Modelo Híbrido | binário | **90,96** |

O trabalho de (SILVA et al., 2022a) adota um pipeline clássico baseado em segmentação de núcleos (via *Mask R-CNN*) e extração de atributos morfológicos e texturais. A classificação foi realizada com o classificador polinomial (PL), resultando em acurácia de 92,4% em um cenário multiclasse (saudável, leve, moderado e severo). Essa abordagem fortemente guiada por informações morfológicas mostrou-se relevante na diferenciação entre graus intermediários de displasia, com pequenas variações estruturais sendo decisivas. Em contraste, (LONGO et al., 2024) exploraram *deep features* extraídas de CNNs pré-treinadas (*Inception-V3* e *VGG-19*), combinadas em um comitê homogêneo e reduzidas via seleção de atributos. A estratégia permitiu alcançar 97,97% de acurácia no cenário binário (*saudável* × *severo*), demonstrando o potencial de *ensembles* de representações extraídas de modelos profundos quando aplicados a bases de imagens pequenas. Já em outro estudo de (SILVA et al., 2022b), os autores investigaram o uso de *deep features* extraídas por redes pré-treinadas *AlexNet* e *ResNet-50*, aplicando o algoritmo *ReliefF* para seleção de atributos e, posteriormente, utilizando o classificador polinomial *Hermite Orthogonal Polynomial* (HOP). Essa abordagem alcançou desempenho médio elevado, com AUC superior a 0,98 nas classificações binárias, incluindo valores de até 1,00 em cenários com maior separação entre classes (como *saudável* × *severo*). O resultado evidencia o potencial da combinação entre representações profundas e técnicas clássicas de seleção de atributos, que auxiliam na eliminação de redundâncias e no aumento da robustez do classificador. No estudo de (TENGUAM et al., 2024), os autores ampliaram essa abordagem incorporando descritores manuais (fractais, *Haralick*, LBP) e aplicando um esquema

de seleção em duas etapas (*ReliefF* + PSO). Curiosamente, nesse estudo a melhor performance para displasia oral foi obtida exclusivamente com atributos manuais, atingindo 100% de acurácia no par *saudável* × *severo*. Esse resultado sugere que, para esse conjunto em particular, os padrões morfológicos e texturais capturados pelos descritores tradicionais já são suficientemente discriminativos, dispensando *deep features*. Por outro lado, o estudo (SILVA et al., 2024) introduziu o banco de imagens *OralEpitheliumDB* e explorou um fluxo experimental baseado em normalização de corantes H&E, segmentação nuclear e associação de *Deep features* (*ResNet50*) com *Random Forest*. Diferente dos anteriores, este trabalho focou no cenário multiclasse, alcançando 94,22% de acurácia e destacando-se como referência por avaliar simultaneamente os quatro estágios da displasia epitelial (saudável, leve, moderado e severo). A incorporação da etapa de segmentação mostrou-se crucial para elevar o desempenho e sustentar a aplicabilidade clínica do método. No contexto desses trabalhos, o método híbrido proposto neste estudo obteve 90,96% de acurácia no par *saudável* × *severo*. Embora esse valor seja inferior aos reportados por (LONGO et al., 2024), (SILVA et al., 2022b), (TENGUAM et al., 2024) e (SILVA et al., 2024), ele se mostra competitivo ao considerar a ausência de etapas de segmentação e seleção de atributos com diferenças nos protocolos experimentais (validação, aumento de dados e normalização). Além disso, os resultados do híbrido foram superiores ou equivalentes em cenários de maior contraste morfológico frente ao classificador polinomial de (SILVA et al., 2022a). Também é uma das primeiras abordagens a explorar a combinação híbrida entre modelos CNNs e *Transformers* em um contexto de avaliação de tecidos histológicos de displasia.

# 6 Conclusão

Neste trabalho foi desenvolvido e avaliado um modelo híbrido para classificação de imagens histológicas da cavidade oral, associando as CNNs e ViT para construção de um modelo híbrido. A proposta buscou explorar a complementaridade entre a captura de padrões locais, dependências globais e representações convolucionais clássicas, compondo uma arquitetura de comitê robusta.

Os experimentos realizados evidenciaram que o modelo híbrido atinge desempenho competitivo frente ao estado da arte. Em especial, nos cenários com maior contraste morfológico — como *Saudável* × *Leve* e *Saudável* × *Severo* — o método proposto alcançou AUCs de 1,00 e 0,97, superando ou igualando abordagens consolidadas como o classificador polinomial (PL). Contudo, nas comparações entre classes adjacentes (*Leve* × *Moderado*, *Moderado* × *Severo*), modelos baseados em atributos morfológicos e texturais mostraram melhor desempenho, sugerindo que padrões sutis ainda são mais bem explorados por descritores manuais.

Na comparação com a literatura, destacam-se diferentes estratégias. O trabalho de (SILVA et al., 2022b) demonstrou a eficácia da combinação entre *deep features* e o PL, alcançando acurácia de 98%. Já os estudos de (LONGO et al., 2024) e (TENGUAM et al., 2024) evidenciaram o potencial dos *ensembles*, seja com CNNs pré-treinadas ou com atributos manuais selecionados, obtendo até 100% de acurácia no cenário *Saudável* × *Severo*. Essas comparações permitem posicionar o modelo híbrido proposto como uma alternativa viável e eficaz no espectro de soluções para a classificação de displasia oral. Apesar de ainda existir um *gap* frente a abordagens com seleção de atributos e *ensembles* especializados, o método aqui apresentado mostrou promissor em cenários de maior contraste e oferece um ponto de partida sólido para investigações futuras. Como perspectivas, destaca-se a incorporação de atributos morfológicos explícitos, o uso de estratégias de seleção em múltiplas fases, e a adoção de protocolos experimentais mais robustos, como validação estratificada e normalização de cor. Esses avanços podem contribuir para reduzir as diferenças frente ao estado da arte e aprimorar a capacidade preditiva em cenários mais desafiadores, como os estágios intermediários de displasia.

O trabalho desenvolvido demonstra que a arquitetura híbrida é robusta na classificação de displasias orais. No entanto, esta abordagem abre oportunidades para investigações futuras. Uma direção promissora seria a aplicação do modelo a um conjunto de dados maior e mais diversificado para avaliar e aprimorar sua capacidade de generalização para diferentes tipos de imagens histológicas. Outro ponto de aprimoramento reside na investigação de métodos de aprendizagem em comitê (*ensemble learning*) para refinar a

combinação dos modelos e buscar um desempenho ainda superior. Por fim, a metodologia proposta pode ser validada em outras aplicações de imagens biomédicas, como a classificação de outros tipos de tumores ou anomalias celulares, a fim de testar a robustez e a adaptabilidade da arquitetura. O código-fonte completo e os *scripts* utilizados para a execução dos experimentos deste trabalho estão disponíveis publicamente para fins de reprodutibilidade no seguinte repositório GitHub:

<https://github.com/GuilhermeRafaell/OralEpithelium>

# Referências

AHMED, T.; ALEX, S.; MUSTAFA, A.; AWAIS, M.; JACKSON, P. J. Max-ast: Combining convolution, local and global self-attentions for audio event classification. In: **ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. [S.l.: s.n.], 2024. p. 1061–1065. Citado 2 vezes nas páginas 3 e 13.

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. 1–74, 2021. Citado 3 vezes nas páginas 25, 26 e 28. Disponível em: <https://doi.org/10.1186/s40537-021-00444-8>. Citado 2 vezes nas páginas 3 e 12.

CARLSON, E. R.; KADEMANI, D.; WARD, B. B.; OREADI, D. American association of oral and maxillofacial surgeon's position paper on oral mucosal dysplasia. **Journal of Oral and Maxillofacial Surgery**, Elsevier, v. 81, n. 8, p. 1042–1054, 2023. ISSN 0278-2391. Disponível em: <https://doi.org/10.1016/j.joms.2023.04.017>. Citado na página 10.

CHEN, J.; CHEN, Z. Tensor decomposition based networks for nuclei segmentation and classification. **Electronics Letters**, v. 58, n. 25, p. 975–977, 2022. Citado na página 10. Disponível em: <https://doi.org/10.1049/ell2.12668>. Citado 2 vezes nas páginas 7 e 15.

CHEN, L.; LI, S.; BAI, Q.; YANG, J.; JIANG, S.; MIAO, Y. Review of image classification algorithms based on convolutional neural networks. **Remote Sensing**, v. 13, p. 4712, 2021. Disponível em: <https://doi.org/10.3390/rs13224712>. Citado na página 17.

DOSOVITSKIY, A.; BEYER, L.; KOLESNIKOV, A.; WEISSENBORN, D.; ZHAI, X.; UNTERTHINER, T.; DEHGHANI, M.; MINDERER, M.; HEIGOLD, G.; GELLY, S.; USZKOREIT, J.; HOULSBY, N. An image is worth 16×16 words: Transformers for image recognition at scale. **International Conference on Learning Representations (ICLR)**, 2021. Disponível em: <https://arxiv.org/abs/2010.11929>. Citado 3 vezes nas páginas 14, 25 e 26.

FERNANDES, V.; SILVA, A.; PEREIRA, D.; CARDOSO, S.; FARIA, P. R. de; LOYOLA, A.; TOSTA, T.; NEVES, L.; NASCIMENTO, M. Z. do. Investigation of deep neural network compression based on tucker decomposition for the classification of lesions in cavity oral. In: INSTICC. **Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 3: VISAPP**. [S.l.]: SciTePress, 2024. p. 516–523. ISBN 978-989-758-679-8. Citado na página 18.

GONZALEZ, R. C.; WOODS, R. **Processamento Digital de Imagens**. [S.l.]: Pearson Prentice Hall, 2010. Citado 2 vezes nas páginas 12 e 15.

HAYKIN, S. **Neural Networks and Learning Machines**. 3rd. ed. Upper Saddle River, NJ: Pearson Education, 2009. Disponível em: <https://lps.ufrj.br/~caloba/Livros/Haykin2009.pdf>. Citado 3 vezes nas páginas 3, 12 e 13.

HE, W.; HAN, Y.; MING, W.; DU, J.; LIU, Y.; YANG, Y.; WANG, L.; WANG, Y.; JIANG, Z.; CAO, C. et al. Progress of machine vision in the detection of cancer cells in histopathology. **IEEE Access**, IEEE, v. 10, p. 46753–46771, 2022. Citado 2 vezes nas páginas 3 e 16.

INCA - Instituto Nacional de Câncer. **Falando sobre o câncer da boca**. Rio de Janeiro, RJ: Engenho e Arte: [s.n.], 2002. Disponível em: <https://bvsms.saude.gov.br/bvs/publicacoes/falando_sobre_cancer_boca.pdf>. Citado 3 vezes nas páginas 7, 10 e 11.

Instituto Nacional de Câncer. **Câncer: o que é, causas, sintomas e prevenção**. 2023. Acesso em: 6 out. 2024. Disponível em: <https://www.gov.br/inca/pt-br/assuntos/cancer>. Citado na página 6.

IRSHAD, H.; VEILLARD, A.; ROUX, L.; RACOCEANU, D.; LE, T. A. Methods for nuclei detection, segmentation, and classification in digital histopathology: a review—current status and future potential. **IEEE Reviews in Biomedical Engineering**, IEEE, v. 7, p. 97–114, 2014. Disponível em: <https://doi.org/10.1109/RBME.2013.2295804>. Citado na página 15.

KIM, Y.; PARK, E.; YOO, S.; CHOI, T.; YANG, L.; SHIN, D. Compression of deep convolutional neural networks for fast and low power mobile applications. In: BENGIO, Y.; LECUN, Y. (Ed.). **4th International Conference on Learning Representations, ICLR 2016, Conference Track**. San Juan, Puerto Rico: [s.n.], 2016. Citado na página 11.

KUMAR, V. **Robbins & Cotran - Patologia: Bases Patológicas das Doenças**. 8. ed. Elsevier Brasil: [s.n.], 2010. Disponível em: <https://books.google.com.br/books?id=2O_jYLsNxrYC>. ISBN 9788535246339. Citado 2 vezes nas páginas 7 e 10.

LI, X.; QIU, B.; CAO, G.; WU, C.; ZHANG, L. A novel method for ground-based cloud image classification using transformer. **Remote Sensing**, MDPI, v. 14, n. 16, p. 3978, 2022. Disponível em: <https://doi.org/10.3390/rs14163978>. Citado na página 25.

LIMA, A. L. A.; SILVA, A. M.; SOARES, M. S.; FIGUEIREDO, N.; OLIVEIRA, B.; SANTOS, A. C.; JúNIOR, H. M. Trend of hospitalized cases of oral cancer in brazil and its relationship with oral health coverage in the public health system between 2009 and 2017. **BMC Oral Health**, v. 21, n. 1, p. 1–10, 2021. Citado na página 6.

LISBOA, P. J. G. A review of evidence of health benefit from artificial neural networks in medical intervention. **Neural Networks: The Official Journal of the International Neural Network Society**, v. 15, n. 1, p. 11–39, January 2002. ISSN 0893-6080. Citado 2 vezes nas páginas 10 e 15. Disponível em: <https://doi.org/10.1016/s0893-6080(01)00111-3>. Citado na página 7.

LONGO, L. H. da C.; ROBERTO, G. F.; TOSTA, T. A. A.; FARIA, P. R. de; LOYOLA, A. M.; CARDOSO, S. V.; SILVA, A. B.; NASCIMENTO, M. Z. do; NEVES, L. A. Classification of multiple h&e images via an ensemble computational scheme. **Entropy**, v. 26, n. 1, p. 34, 2024. Citado 5 vezes nas páginas 16, 17, 36, 37 e 38.

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 20.

PAIXãO, G. M. d. M.; SANTOS, B. C.; ARAUJO, R. M. d.; RIBEIRO, M. H.; MORAES, J. L. d.; RIBEIRO, A. L. Machine learning in medicine: Review and applicability. **Arquivos Brasileiros de Cardiologia**, v. 118, n. 1, p. 95–102, January 2022. ISSN 0066-782X. Citado 2 vezes nas páginas 10 e 14. Disponível em: <https://doi.org/10.36660/abc.20200596>. Citado 2 vezes nas páginas 7 e 11.

SILVA, A. B. **Métodos computacionais para análise e classificação de displasias em imagens da cavidade bucal**. 2019. Disponível em: <http://dx.doi.org/10.14393/ufu.di.2019.2390>. Citado na página 31. Disponível em: <http://dx.doi.org/10.14393/ufu.di.2019.2390>. Citado 2 vezes nas páginas 3 e 11.

SILVA, A. B.; MARTINS, A. S.; TOSTA, T. A. A.; NEVES, L. A.; SERVATO, J. P. S.; de Araújo, M. S.; de Faria, P. R.; NASCIMENTO, M. Z. do. Computational analysis of histological images from hematoxylin and eosin-stained oral epithelial dysplasia tissue sections. **Expert Systems with Applications**, v. 193, p. 116456, 2022a. ISSN 0957-4174. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0957417421017401>. Citado 4 vezes nas páginas 15, 17, 36 e 37.

SILVA, A. B.; MARTINS, A. S.; TOSTA, T. A. A.; LOYOLA, A. M.; CARDOSO, S. V.; NEVES, L. A.; FARIA, P. R. d.; NASCIMENTO, M. Z. d. N. do. Oralepitheliumdb: A dataset for oral epithelial dysplasia image segmentation and classification. **Journal of Imaging Informatics in Medicine**, Springer, 2024. Citado 4 vezes nas páginas 18, 20, 36 e 37.

SILVA, A. B.; OLIVEIRA, C. I. de; PEREIRA, D. C.; TOSTA, T. A. A.; MARTINS, A. S.; LOYOLA, A. M.; CARDOSO, S. V.; FARIA, P. R. de; NEVES, L. A.; NASCIMENTO, M. Z. do. Assessment of the association of deep features with a polynomial algorithm for automated oral epithelial dysplasia grading. In: **2022 35th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)**. [S.l.: s.n.], 2022b. p. 264–268. Citado 3 vezes nas páginas 36, 37 e 38.

SOLOMON, C.; BRECKON, T. **Fundamentos de processamento digital de imagens: uma abordagem prática com exemplos em Matlab**. [S.l.]: Grupo Gen-LTC, 2000. Citado na página 12.

TAN, M.; LE, Q. V. Efficientnet: Rethinking model scaling for convolutional neural networks. In: **Proceedings of the 36th International Conference on Machine Learning (ICML)**. PMLR, 2019. v. 97, p. 6105–6114. Disponível em: <https://arxiv.org/pdf/1905.11946.pdf>. Citado 3 vezes nas páginas 13, 17 e 24.

TENGUAM, J. J.; LONGO, L. H. da C.; ROBERTO, G. F.; TOSTA, T. A. A.; FARIA, P. R. de; LOYOLA, A. M.; CARDOSO, S. V.; SILVA, A. B.; NASCIMENTO, M. Z. do; NEVES, L. A. Ensemble learning-based solutions: An approach for evaluating multiple features in the context of h&e histological images. **Applied Sciences**, v. 14, n. 3, p. 1084, 2024. Citado 6 vezes nas páginas 16, 18, 27, 36, 37 e 38.

Viso.ai. **Vision Transformer (ViT): A Complete Guide**. 2024. <https://viso.ai/deep-learning/vision-transformer-vit/>. Acesso em: 8 set. 2025. Citado na página 14.

WARNAKULASURIYA, S. et al. Oral epithelial dysplasia classification systems: predictive value, utility, weaknesses and scope for improvement. **Oral Oncology**, Elsevier, v. 44, n. 4, p. 309–320, 2008. Citado na página 11.

WARNAKULASURIYA, S.; REIBEL, J.; BOUQUOT, J.; DABELSTEEN, E. Oral epithelial dysplasia classification systems: predictive value, utility, weaknesses and scope for improvement. **Review Article**, 2021. Citado na página 10.

YIN, H.; VAHDAT, A.; ALVAREZ, J. M.; MALLYA, A.; KAUTZ, J.; MOLCHANOV, P. A-vit: Adaptive tokens for efficient vision transformer. In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2022. p. 10809–10818. Citado 3 vezes nas páginas 14, 15 e 17.