UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
DEPARTAMENTO DE INFORMÁTICA APLICADA
CURSO DE ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS

DIEINE ESTELA BERNIERI SCHIAVON DE ANDRADE

Classificação Automática de Lesões Bucais no Apoio ao Diagnóstico: uma Abordagem de Aprendizado Profundo e Inteligência Artificial Explicável

Monografia de Conclusão de Curso apresentada como requisito parcial para a obtenção do grau de Especialista em Engenharia de Software e Ciência de Dados

Orientador: Prof. Dr. Anderson Rocha Tavares

Porto Alegre
2026

CIP — CATALOGAÇÃO NA PUBLICAÇÃO

Andrade, Dieine Estela Bernieri Schiavon de

Classificação Automática de Lesões Bucais no Apoio ao Diagnóstico: uma Abordagem de Aprendizado Profundo e Inteligência Artificial Explicável / Dieine Estela Bernieri Schiavon de Andrade. – Porto Alegre: PPGC da UFRGS, 2026.

73 f.: il.

Monografia (especialização) – Universidade Federal do Rio Grande do Sul. Curso de Especialização em Ciência de Dados, Porto Alegre, BR–RS, 2026. Orientador: Anderson Rocha Tavares.

1. Câncer bucal. 2. Deep learning. 3. Inteligência Artificial Explicável. I. Tavares, Anderson Rocha. II. Título.

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
Reitora: Prof[a]. Marcia Barbosa
Vice-Reitor: Prof. Pedro de Almeida Costa
Pró-Reitor de Pós-Graduação: Prof[a]. Claudia Wasserman
Diretor do Instituto de Informática: Prof. Luciano Paschoal Gaspary
Coordenadores da Especialização: Prof[a]. Dra. Karin Becker e Prof. Dr. Marcelo Pimenta
Bibliotecário-chefe do Instituto de Informática: Alexsander Borges Ribeiro

“It does not matter how slowly you go as long as you do not stop”

— CONFUCIUS

AGRADECIMENTOS

Agradeço aos meus filhos, Mariana e Eduardo, e ao meu marido, Paulo, pelo amor, apoio, compreensão e incentivo constante; por vibrarem a cada conquista e pelas palavras de conforto nos momentos mais difíceis.

Ao meu orientador, Anderson da Rocha Tavares, por compartilhar suas experiências e conhecimentos, os quais me fizeram crescer academicamente e me guiaram na construção deste trabalho.

Aos meus colegas e professores, pela oportunidade de convívio e aprendizado ao longo do curso. Cada momento compartilhado foi fundamental para minha formação.

Por fim, gostaria de expressar minha sincera gratidão aos meus pais, Nelson e Edy, às minhas irmãs, Nina e Nara, e ao meu irmão, João, que sempre torceram por mim e acreditaram nos meus sonhos.

Muito obrigada!

RESUMO

O câncer bucal representa um desafio significativo para a saúde pública global, exigindo métodos de diagnóstico precoce que sejam precisos e acessíveis. Recentemente, modelos de aprendizado profundo têm demonstrado potencial na análise de imagens intraorais. No entanto, a presença de artefatos visuais em fotografias de face total e a ausência de integração com dados clínicos multimodais limitam a eficácia de abordagens puramente automatizadas. Esta pesquisa propõe uma abordagem multimodal baseada em Redes Neurais Convolucionais para a classificação de lesões orais, integrando informações visuais de fotografias intraorais e metadados clínicos dos pacientes. A metodologia consistiu no desenvolvimento de um pipeline de processamento para a extração da Região de Interesse (ROI), visando mitigar o impacto de elementos irrelevantes presentes nas imagens completas. Foram avaliadas quatro arquiteturas, incluindo MobileNetV2, MobileNetV3 Small, EfficientNetB0 e ConvNeXt Tiny. Foi adotado um protocolo rigoroso de separação de subconjuntos de treinamento, validação e teste por paciente, mitigando o vazamento de dados e garantindo uma avaliação realista da capacidade de generalização do modelo. Os resultados demonstram que o recorte da ROI elevou significativamente o desempenho das redes, com a MobileNetV3 Small atingindo 84,81% de acurácia. A integração multimodal, combinando a extração de características visuais com variáveis clínicas, alcançou o desempenho superior de 88,52% de acurácia e *Loss* de 0,3472 nos dados de teste, um F1-score macro de 0,8601 e macro *recall* de 83,29%. A interpretabilidade das decisões foi validada por meio de técnicas de Grad-CAM e LIME, confirmando que os modelos concentram a atenção em regiões clinicamente relevantes. A comparação com o estado da arte revelou que, embora estudos relacionados relatem acurácias superiores (acima de 95%), tais pesquisas frequentemente não mencionam o rigor da separação por paciente. Em contraste, este trabalho supera as abordagens multimodais validadas sob protocolos similares, consolidando-se como uma solução robusta para o auxílio ao diagnóstico clínico. Conclui-se que a combinação de segmentação estratégica, integração de dados heterogêneos e transparência algorítmica constitui um avanço essencial para a aplicação ética e eficaz da inteligência artificial na oncologia bucal.

Palavras-chave: Câncer bucal. Deep learning. Inteligência Artificial Explicável.

ABSTRACT

Oral cancer represents a significant challenge to global public health, requiring accurate and accessible early diagnostic methods. Recently, deep learning models have shown promise in analyzing intraoral images. However, the presence of visual artifacts in full-face photographs and the lack of integration with multimodal clinical data limit the effectiveness of purely automated approaches. This research proposes a multimodal approach based on Convolutional Neural Networks for the classification of oral lesions, integrating visual information from intraoral photographs and clinical metadata from patients. The methodology consisted of developing a processing pipeline for Region of Interest (ROI) extraction to mitigate the impact of irrelevant elements in complete images. Four architectures were evaluated, including MobileNetV2, MobileNetV3 Small, EfficientNetB0, and ConvNeXt Tiny. A rigorous protocol for separating training, validation, and patient-based testing subsets was adopted, mitigating data leakage and ensuring a realistic assessment of the model's generalization capacity. The results demonstrate that ROI clipping significantly improved network performance, with MobileNetV3 Small achieving 84.81% accuracy. Multimodal integration, combining visual feature extraction with clinical variables, achieved superior performance with 88.52% accuracy and a loss of 0.3472 in the test data, a macro F1-score of 0.8601, and a macro recall of 83.29%. The interpretability of the decisions was validated using Grad-CAM and LIME techniques, confirming that the models focus attention on clinically relevant regions. Comparison with the state of the art revealed that although related studies report superior accuracies (above 95%), these studies often fail to specify the rigor of patient separation. In contrast, this work surpasses validated multimodal approaches under similar protocols, establishing itself as a robust solution to aid clinical diagnosis. It is concluded that the combination of strategic segmentation, heterogeneous data integration, and algorithmic transparency constitutes an essential advance for the ethical and effective application of artificial intelligence in oral oncology.

Keywords: Oral cancer. Deep learning. Explainable Artificial Intelligence.

LISTA DE ABREVIATURAS E SIGLAS

AUC *Area Under the Receiver Operating Characteristic Curve*

CAM *Class Activation Mapping*

COCO *Common Objects in Context*

CEC Carcinoma Espinocelular

CNN *Convolutional Neural Networks*

DL *Deep Learning*

FID *Fréchet Inception Distance*

FLOPS *Floating Point Operations*

FN *False Negative*

FP *False Positive*

GeLU *Gaussian Error Linear Unit*

GPU *Graphics Processing Unit*

HPV *Human Papiloma Virus*

IA Inteligência Artificial

LIME *Local Interpretable Model-agnostic Explanations*

MANOVA *Multivariate Analysis of Variance*

OCA *Oral Cancer*

OML *Oral Mucosal Lesion*

OPMD *Oral Potentially Malignant Disorder*

OSEMN *Obtain, Scrub, Explore, Model and Interpret*

PCA *Principal Component Analysis*

PR *Precision-Recall*

QNCC *Quantum Convolutional Neural Network*

ReLU *Rectified Linear Unit*

RGB *Red, Green and Blue*

RNC Rede Neural Convolucional

ROC *Receiver Operating Characteristics*

ROI *Region of Interest*

SAM *Segment Anything Model*

SEER *Surveillance, Epidemiology, and End Results Program,*

SIM Sistema de Informações sobre Mortalidade

SOTA *State-of-the-Art*

SSIM *Structural Similarity Index*

SVM *Support Vector Machine*

TN *True Negative*

TP *True Positive*

XAI *Explainable Artificial Intelligence*

LISTA DE FIGURAS

Figura 2.1 Estratégias de transferência de aprendizado em rede pré-treinada VGG16 ..18
Figura 2.2 Curva ROC.....21

Figura 3.1 Estratégia de busca e seleção dos artigos 23

Figura 4.1 Distribuição das classes do *dataset* 32
Figura 4.2 Exemplo de imagem anotada formato COCO destacando a região oral (verde) e a região da lesão (roxo).....33
Figura 4.3 Metadados dos Indivíduos 33
Figura 4.4 Exemplos de imagens após o recorte (crop) das lesões e redimensionamento35
Figura 4.5 Distribuição das imagens por classe 36
Figura 4.6 Modelo MobileNetV2.....38
Figura 4.7 Parte final do modelo multimodal.....39

Figura 5.1 Curvas ROC e de Aprendizado - MobileNetV2 - Full Frame 45
Figura 5.2 Matrizes de Confusão - MobileNetV2 - Full Frame.....45
Figura 5.3 Exemplos com LIME e Grad-CAM - MobileNetV2 - Full Frame.....46
Figura 5.4 Curvas ROC e de aprendizado - MobileNetV2 - Crop I 47
Figura 5.5 Matrizes de Confusão - MobileNetV2 - Crop I.....47
Figura 5.6 Curvas ROC e de aprendizado - MobileNetV2 - Crop II 48
Figura 5.7 Matrizes de Confusão - MobileNetV2 - Crop II.....48
Figura 5.8 Predições do modelo 2 - Crop com LIME.....49
Figura 5.9 Matrizes de Confusão do Modelo MobilenetV3 - Crop 52
Figura 5.10 Curvas de Treinamento e Validação - MobileNetV3 - Crop.....52
Figura 5.11 Curva ROC do Modelo MobileNetV3 - Crop.....53
Figura 5.12 Predições do modelo MobileNetV3 - Crop com LIME 53
Figura 5.13 Imagens originais completas com as anotações COCO. 54
Figura 5.14 Curvas ROC e PR do Modelo Multimodal 55
Figura 5.15 Curvas de Aprendizagem completas e um recorte até a 10ª época do Modelo Multimodal 55
Figura 5.16 Matrizes de Confusão do Modelo Multimodal.....56
Figura 5.17 Explicabilidade Classe Benign - Imagem e Dados Clínicos 57
Figura 5.18 Explicabilidade Classe OPMD - Imagem e Dados Clínicos 58
Figura 5.19 Explicabilidade Classe OCA - Imagem e Dados Clínicos 59
Figura 5.20 Explicabilidade de um erro de predição - Imagem e Dados Clínicos.....60

Figura 6.1 Arquitetura de integração e fluxo operacional da solução multimodal proposta.....66

SUMÁRIO

1 INTRODUÇÃO **11**
1.1 Objetivos **12**
1.1.1 Geral 12
1.1.2 Específicos 12
1.2 Organização do trabalho **12**
2 FUNDAMENTAÇÃO TEÓRICA **14**
2.1 Câncer Bucal **14**
2.2 Inteligência Artificial em Saúde **15**
2.2.1 Deep Learning 16
2.3 Explicabilidade em IA **18**
2.4 Métricas de Desempenho **19**
3 TRABALHOS RELACIONADOS **22**
4 MATERIAIS E MÉTODOS **31**
4.1 Ambiente de Desenvolvimento **31**
4.2 Conjunto de Dados **32**
4.3 Pré-processamento **33**
4.4 Divisão dos dados: conjuntos de Treinamento, Validação e Teste **34**
4.5 Arquitetura e Configuração dos Modelos **35**
4.6 Experimentos **38**
4.7 Técnicas de IA Explicável **40**
4.8 Resumo da metodologia **40**
4.9 Considerações Éticas **41**
5 RESULTADOS **42**
5.1 Modelo MobileNetV2 - *Full Frame* **42**
5.2 Modelo MobileNetV2 - Crop **46**
5.3 Avaliação das arquiteturas MobileNetV3 Small, EficientNetB0 e ConvNeXt Tiny - Entrada Full Frame **49**
5.4 Avaliação das arquiteturas MobileNetV3 Small, EficientNetB0 e ConvNeXt Tiny - Entrada Crop da Lesão **50**
5.5 Modelo Multimodal **54**
5.6 Análise de Explicabilidade Multimodal: Interpretação Local via LIME (Visual e Tabular) e Grad-CAM **55**
6 DISCUSSÃO **61**
6.1 Análise dos resultados experimentais **61**
6.2 Comparação do desempenho do modelo multimodal com os artigos do Estado da Arte **64**
6.3 Integração do modelo multimodal com a prática clínica **65**
7 CONCLUSÃO **67**
7.1 Limitações e trabalhos futuros **68**
REFERÊNCIAS **69**

1 INTRODUÇÃO

O câncer bucal é uma neoplasia (proliferação celular anormal) maligna que afeta a mucosa oral e representa um importante problema de saúde pública devido à sua alta incidência e mortalidade (SANKARANARAYANAN et al., 2015). Dada a gravidade da doença, o diagnóstico precoce é essencial para aumentar as chances de sobrevida e reduzir os impactos clínicos e sociais (RETHMAN et al., 2010). A prevenção e a antecipação do diagnóstico começam com a identificação de lesões potencialmente malignas da mucosa oral, que podem preceder o desenvolvimento do câncer propriamente dito. O reconhecimento clínico de alterações epiteliais é uma estratégia relevante, pois pode detectar casos de cânceres bucais e lesões pré-malignas quando realizado adequadamente (SANKARANARAYANAN et al., 2015).

Conforme o *National Institute of Dental and Craniofacial Research (2020)*, qualquer lesão suspeita que não regredir em até duas semanas, após a remoção de possíveis fatores locais de irritação, deve ser submetida a biópsia cirúrgica, que continua sendo o padrão de referência para o diagnóstico do câncer bucal (RETHMAN et al., 2010). No entanto, apesar dos avanços tecnológicos e do conhecimento médico, muitos casos de câncer bucal ainda são diagnosticados em estágios avançados, quando o prognóstico é mais reservado. O atraso no diagnóstico está associado a fatores como a subnotificação, a dificuldade de acesso a especialistas e a baixa adesão a programas de rastreamento populacional (LIMA et al., 2021; Instituto Nacional de Câncer, 2023).

Nesse cenário, a área de visão computacional aplicada à saúde tem avançado significativamente, sobretudo com o uso de técnicas de aprendizado profundo que permitem a análise automatizada de imagens clínicas, histopatológicas e fotográficas da cavidade oral, visando detectar precocemente lesões potencialmente malignas e apoiar a tomada de decisão clínica (DEVINDI et al., 2024; MUKHERJEE et al., 2025) . Assim, a questão norteadora deste estudo é: “Considerando o desbalanceamento de classes em cenários clínicos reais, como diferentes arquiteturas de aprendizado profundo e estratégias de mitigação se comparam na classificação de lesões bucais, e em que medida a Inteligência Artificial Explicável valida ou refuta a confiabilidade sugerida pelas métricas de desempenho?”.

Diante desse contexto, o presente trabalho investiga o potencial de arquiteturas de aprendizado profundo na análise de imagens clínicas como apoio ao diagnóstico precoce do câncer bucal. Além do desempenho preditivo, este estudo prioriza a eficiência

computacional, visando à futura incorporação do modelo em um aplicativo móvel, considerando as limitações inerentes a dispositivos móveis, como memória, capacidade de processamento e consumo energético.

Nota Ética: devido à natureza clínica desta pesquisa, este estudo contém imagens de patologias orais que podem ser consideradas desconfortáveis para pessoas sensíveis. Tais registros são necessários para a demonstração visual dos achados e validação das técnicas de inteligência artificial propostas.

1.1 Objetivos

1.1.1 Geral

Desenvolver um modelo baseado na arquitetura de redes neurais convolucionais (RNCs), utilizando *transfer learning*, para a classificação automática de lesões bucais em imagens clínicas.

1.1.2 Específicos

- Revisar a literatura científica sobre a utilização de RNCs aplicadas ao reconhecimento de lesões orais;
- Implementar modelos de classificação baseados em *transfer learning* Avaliar o desempenho dos modelos utilizando métricas formais, tais como acurácia, sensibilidade, especificidade e F1-score;
- Verificar, através de técnicas de explicabilidade as predições dos modelos;
- Examinar a aplicabilidade clínica e os desafios da utilização de imagens obtidas com dispositivos móveis nesse contexto.

1.2 Organização do trabalho

Este trabalho está organizado em sete capítulos, conforme descrito a seguir:

Capítulo 1 – Introdução: Apresenta a contextualização do tema, a justificativa

da pesquisa e a definição dos objetivos geral e específicos.

Capítulo 2 – Fundamentação Teórica: Estabelece a base conceitual sobre câncer bucal, as técnicas de *Deep Learning* empregadas e os princípios de explicabilidade em inteligência artificial.

Capítulo 3 – Trabalhos Relacionados Realiza uma revisão sistemática e crítica da literatura, consolidando as principais abordagens e lacunas identificadas em estudos correlatos.

Capítulo 4 – Metodologia: Detalha o protocolo experimental, incluindo a descrição da base de dados, o pré-processamento, a arquitetura dos modelos e as métricas de avaliação.

Capítulo 5 – Resultados: Reporta os dados brutos e os desempenhos obtidos pelos modelos durante as etapas de treinamento e teste.

Capítulo 6 – Discussão: Analisa os resultados com base na fundamentação teórica e do estado da arte, interpretando o comportamento dos modelos e sua relevância clínica.

Capítulo 7 – Considerações Finais: Encerra o estudo com a síntese das conclusões, as limitações encontradas e as propostas para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta a fundamentação teórica que embasa este trabalho, abordando inicialmente o câncer bucal e, em seguida, o uso da Inteligência Artificial na saúde. São discutidos conceitos de *Deep Learning*, Redes Neurais Convolucionais e modelos pré-treinados. Também é apresentada a explicabilidade em IA no contexto da saúde odontológica, bem como as métricas de desempenho utilizadas na avaliação dos modelos.

2.1 Câncer Bucal

As neoplasias malignas que se desenvolvem nos lábios e na cavidade oral são, em sua maioria, classificadas como Carcinoma Espinocelular (CEC). Esse subtipo responde por cerca de 90% dos casos, devido à origem histológica nas células escamosas da mucosa oral (WARNAKULASURIYA, 2009; ABATI et al., 2020). Epidemiologicamente, o câncer bucal integra o grupo dos cânceres de cabeça e pescoço, representando aproximadamente 85% dessa categoria (Oral Cancer Foundation, 2025)

A prevalência é significativamente maior em homens, com incidência de duas a três vezes superior em comparação às mulheres, em diferentes grupos étnicos (SEER – Surveillance, Epidemiology, and End Results Program, 2020). No Brasil, segundo o Instituto Nacional de Câncer (2023), o número estimado de casos novos de câncer da cavidade oral para cada ano do triênio de 2023 a 2025 é de 15.100 casos, correspondendo a um risco estimado de 6,99 por 100 mil habitantes, sendo 10.900 em homens e 4.200 em mulheres. Esses valores correspondem a um risco estimado de 10,30 casos novos por 100 mil homens e de 3,83 por 100 mil mulheres. Em 2019, o SIM – Sistema de Informações sobre Mortalidade (2019) apontou 6.605 óbitos (5.120 homens e 1.485 mulheres).

Em termos globais, os cânceres de cabeça e pescoço correspondem a 10% de todos os tumores malignos, dos quais cerca de 40% ocorrem na cavidade oral. A distribuição geográfica é heterogênea: em países asiáticos, como a Índia, o câncer da boca é altamente prevalente e associado a fatores culturais, como o uso de *betel quid*(uma mistura mastigada que inclui noz de areca (*areca nut*), envolta em folha de *betel* (*Piper betle*), frequentemente associada a outros componentes, como cal e tabaco) (IARC Working Group on the Evaluation of Carcinogenic Risks to Humans, 2004). A folha de *Piper betle* isoladamente não é considerada carcinogênica; entretanto, a noz de areca é classificada como carcinogênica para humanos, sendo o principal fator de risco associado ao consumo dessa

preparação (IARC Working Group on the Evaluation of Carcinogenic Risks to Humans, 2004). Altas taxas também são observadas em regiões específicas da França e da Itália (World Health Organization, 2023).

A elevada taxa de mortalidade do câncer bucal está fortemente associada ao diagnóstico tardio. Estudos de sobrevida mostram variação entre 46% nos anos 1950 e cerca de 55% a partir da década de 1990 nos Estados Unidos, enquanto no Brasil, pesquisa realizada no Hospital Erasto Gaertner (Curitiba), entre 1990 e 1992, apontou sobrevida de 50,1% em cinco anos (WARNAKULASURIYA, 2009).

Entre os fatores de risco, destacam-se o tabagismo e o etilismo, presentes em até 90% dos casos, com efeito sinérgico no aumento da carcinogênese (WARNAKULASURIYA, 2009; KOONTONGKAEW, 2013). Outros fatores incluem a infecção por papilomavírus humano (HPV), em especial os subtipos HPV16 e HPV18, associados a carcinomas da orofaringe e da cavidade oral (KANG; KIESS; CHUNG, 2015), além da radiação ultravioleta, particularmente relacionada ao câncer de lábio. O HPV contribui para a carcinogênese por meio das proteínas virais E6 e E7, que desregulam o ciclo celular ao inativar genes supressores tumorais, como p53 e pRb, resultando na superexpressão do marcador p16 (ANG et al., 2010).

O prognóstico para pacientes com câncer bucal ainda é considerado desfavorável, apesar dos avanços terapêuticos. O diagnóstico precoce permanece essencial para a melhora da sobrevida (WARNAKULASURIYA, 2009). Além dos impactos clínicos, o câncer bucal gera danos funcionais significativos na fala, deglutição e mastigação, sendo a dor um dos principais sintomas, frequentemente exacerbada por alterações no microambiente tumoral, inervação densa e efeitos adversos do tratamento (VIET; SCHMIDT, 2011).

Estratégias de prevenção e rastreamento incluem o exame visual da cavidade oral, que pode reduzir a mortalidade em populações de risco (BROCKLEHURST et al., 2013). Além disso, programas de educação em saúde, com foco na redução de fatores de risco como tabaco e álcool, são fundamentais para a prevenção da doença.

2.2 Inteligência Artificial em Saúde

A integração da Inteligência Artificial (IA) na área da saúde tem promovido uma mudança de paradigma no diagnóstico e prognóstico de patologias. De acordo com Topol (2019), a IA na medicina não visa a substituição do julgamento clínico, mas sim o seu

aprimoramento, permitindo o processamento de grandes volumes de dados que excedem a capacidade analítica humana individual. No contexto do diagnóstico por imagem, essa evolução é impulsionada principalmente pelo aprendizado profundo, uma subárea da IA que utiliza redes neurais artificiais para extrair padrões complexos de dados não estruturados (LECUN; BENGIO; HINTON, 2015).

A aplicação de modelos computacionais em imagens clínicas permite a identificação de biomarcadores visuais sutis, muitas vezes imperceptíveis ao olho humano em estágios precoces. Especificamente na odontologia e estomatologia, a IA tem demonstrado eficácia na triagem automatizada de lesões orais, auxiliando na redução de diagnósticos tardios e na otimização do fluxo de encaminhamento para especialistas (ESTEVA et al., 2019).

Entretanto, a implementação dessas tecnologias em ambientes de saúde impõe desafios rigorosos. A eficácia de um sistema de suporte à decisão diagnóstica depende não apenas da acurácia bruta do modelo, mas de sua capacidade de generalização em cenários reais e de sua interpretabilidade. A busca por modelos que conciliam alto desempenho preditivo com baixo custo computacional é fundamental para democratizar o acesso a essas ferramentas, permitindo seu uso em dispositivos móveis ou em regiões com infraestrutura de saúde e tecnológica limitadas.

2.2.1 Deep Learning

O *Deep Learning* (DL) constitui um subconjunto do aprendizado de máquina baseado em redes neurais artificiais profundas, caracterizado pela capacidade de extrair representações hierárquicas a partir de dados complexos. Sua principal vantagem está na capacidade de aprender automaticamente características relevantes diretamente dos dados, reduzindo a necessidade da extração manual de atributos.

Em aplicações de visão computacional, técnicas de DL, como as Redes Neurais Convolucionais (RNCs), apresentaram desempenho superior em tarefas de classificação, detecção e segmentação de imagens (LECUN; BENGIO; HINTON, 2015).

As Redes Neurais Convolucionais são arquiteturas compostas por diferentes tipos de camadas organizadas hierarquicamente. Essas redes são estruturadas pelo empilhamento de camadas convolucionais, responsáveis pela extração de características, camadas de *pooling* (ou *subsampling*), que realizam a redução dimensional, e camadas totalmente conectadas, cuja combinação define a arquitetura da rede (LECUN et al., 2002; CIRESAN

et al., 2011).

As camadas convolucionais atuam como extratoras de características e, assim, aprendem as representações das características dos *inputs* de entrada. Os neurônios nas camadas convolucionais são organizados em mapas de características (*feature maps*). Os filtros conectam cada neurônio em um mapa de características aos neurônios da camada anterior. Os filtros de convolução detectam características cada vez mais relevantes da imagem/sinal. Várias camadas convolucionais são agrupadas e empilhadas umas sobre as outras para formar um modelo profundo e recuperar representações de características mais abstratas.

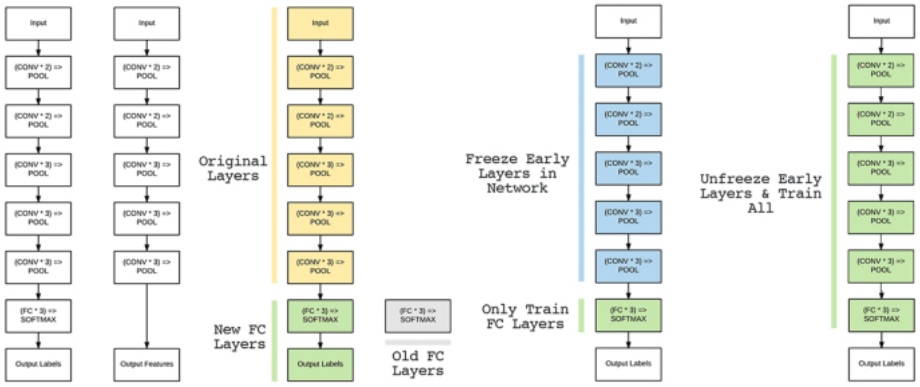
Uma camada convolucional contém um conjunto de filtros ou *kernels*, cujos parâmetros devem ser aprendidos ao longo do treinamento. O tamanho do filtro é geralmente menor do que a imagem real. Cada filtro envolve a imagem e cria um mapa de ativação. Para a convolução, o filtro desliza na altura e na largura da imagem e o produto escalar é calculado entre cada elemento do filtro e a entrada. Ao final, um mapa de características é criado, esses valores são então, passados para a próxima camada.

Entretanto, os mapas de características produzidos pelas camadas convolucionais podem ser sensíveis a pequenas variações espaciais, como deslocamentos, rotações ou recortes. Para mitigar esse efeito e reduzir a dimensionalidade dos dados, são utilizadas camadas de *pooling*, responsáveis pelo processo de *downsampling*. Esse procedimento preserva as informações mais relevantes enquanto reduz a complexidade computacional do modelo (SREENIVAS et al., 2020).

As técnicas de *pooling* mais empregadas são o Max *Pooling* e o *Average Pooling*. O *Max Pooling* seleciona o valor máximo em uma determinada região do mapa de características, enquanto o *Average Pooling* calcula a média dos valores dessa região. Ambas contribuem para tornar a rede mais robusta a pequenas variações espaciais, além de reduzir o risco de sobreajuste.

Entre as técnicas de aprendizado de máquina, destaca-se o aprendizado por transferência (*transfer learning*), amplamente utilizado em modelos de aprendizado profundo. Como o próprio nome sugere, essa abordagem consiste em extrair e transferir conhecimento informativo de um domínio de dados de origem (*source*) para um novo domínio de destino (*target*). Quando bem-sucedida, essa técnica melhora significativamente o desempenho preditivo no domínio de destino (YANG et al., 2019). Essencialmente, trata-se de um método de reutilização de modelos pré-treinados para novas tarefas, sendo aplicável a problemas de classificação, regressão e agrupamento (TAMMINA et al., 2019).

Figura 2.1: Estratégias de transferência de aprendizado em rede pré-treinada VGG16



Fonte: Adaptado de (ROSEBROCK, 2019)

Conforme Chollet (2021), existem duas formas principais de utilizar uma rede pré-treinada (Figura 2.1):

- Extração de características (*feature extraction*): consiste em utilizar a base convolucional de uma rede previamente treinada para processar novos dados, mantendo seus pesos congelados, e treinar um novo classificador sobre as representações extraídas.

- Ajuste fino (*fine-tuning*): consiste em descongelar parte das camadas superiores da base do modelo pré-treinado e treiná-las conjuntamente com as novas camadas adicionadas. Essa estratégia permite adaptar as representações abstratas previamente aprendidas, tornando-as mais específicas e relevantes para o problema em análise (CHOLLET, 2021).

2.3 Explicabilidade em IA

A explicabilidade em Inteligência Artificial, frequentemente referida pelo termo *eXplainable AI (XAI)*, compreende o conjunto de métodos e técnicas que permitem aos humanos compreender e confiar nos resultados e decisões gerados por modelos de aprendizado de máquina. No campo da visão computacional, a explicabilidade é essencial para mitigar a natureza de caixa preta das redes neurais profundas, garantindo a confiabilidade de sistemas que auxiliam em tarefas críticas, como o diagnóstico de patologias orais.

Os métodos de explicabilidade têm como objetivo fornecer interpretações visuais ou numéricas que elucidem o mapeamento entre os dados de entrada e a predição final (GILPIN et al., 2018). Na prática clínica, essa transparência é o que permite ao profis-

sional de saúde validar se o modelo está baseando sua decisão em características morfológicas relevantes da lesão ou em artefatos presentes na imagem que são clinicamente irrelevantes.

Diversas técnicas foram propostas para tornar os modelos de DL mais interpretáveis. Entre elas destacam-se os métodos baseados em visualização, como o *Gradient-weighted Class Activation Mapping* (Grad-CAM) (SELVARAJU et al., 2017), e os métodos de perturbação local, como o *Local Interpretable Model-agnostic Explanations* (LIME) (RIBEIRO; SINGH; GUESTRIN, 2016). Essas abordagens permitem identificar quais regiões da imagem foram determinantes para a predição realizada, transformando modelos opacos em sistemas auditáveis

O Grad-CAM é uma técnica que utiliza os gradientes das camadas convolucionais para gerar mapas de ativação que destacam as regiões mais relevantes de uma imagem para a tomada de decisão do modelo. Essa técnica permite visualizar, de forma intuitiva, o que a rede neural considerou ao realizar a classificação, auxiliando na interpretação e na validação dos resultados obtidos (SELVARAJU et al., 2017).

Diferente de métodos globais que tentam explicar o comportamento de todo o modelo, o LIME aborda a explicabilidade local. A técnica opera perturbando a imagem de entrada ao ocultar sistematicamente segmentos da imagem denominados *superpixels*, observando como essas mudanças afetam a predição final do modelo. A partir disso, o LIME treina um modelo linear simples (interpretável) ao redor da predição específica para identificar quais regiões da imagem foram determinantes para a decisão da rede.

No contexto de imagens odontológicas, as técnicas de explicabilidade permitem validar se a rede neural está direcionando a atenção para características patológicas reais, como a textura da mucosa ou bordas de uma ulceração, ou se está sendo influenciada por artefatos irrelevantes, como a presença de dentes, instrumentos ou variações de iluminação. Essa transparência é vital para a aceitação clínica, pois permite que o profissional de saúde verifique a decisão da inteligência artificial, garantindo que a predição esteja ancorada em biomarcadores visuais clinicamente relevantes.

2.4 Métricas de Desempenho

A avaliação do modelo é indicada para medir a capacidade de generalização do algoritmo e é realizada por meio de métricas de desempenho, que avaliam a proporção de classes corretamente classificadas.

Uma matriz de confusão compara as previsões de um modelo com o padrão real. A diagonal principal (TN e TP) da matriz representa as classes corretamente preditas, enquanto a diagonal secundária (FP e FN) representa as classes classificadas incorretamente. Com base na matriz de confusão, calculam-se métricas para avaliar se o modelo apresenta bons resultados.

A utilização de métricas de avaliação é crucial para determinar a eficácia e as limitações de um modelo. Nos experimentos realizados, os modelos são avaliados com um conjunto de métricas, incluindo acurácia, precisão, sensibilidade, especificidade, F1-score e curva ROC (*Receiver Operating Characteristics*), para fornecer uma compreensão abrangente do desempenho.

A acurácia (Equação 2.1) avalia o percentual de acertos; ela é obtida pela razão entre o número de acertos e o total de entradas. A medida informa a razão entre as classificações corretas e todas as classificações realizadas.

$$\text{Acurácia} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} \tag{2.1}$$

A precisão (Equação 2.2) é uma métrica que avalia a proporção de verdadeiros positivos em relação à soma de todos os valores positivos. Também é conhecida como valor preditivo positivo.

$$\text{Precisão} = \frac{TP}{TP + FP} \tag{2.2}$$

A sensibilidade, ou *recall* (Equação 2.3), avalia a capacidade do método de detectar com sucesso resultados classificados como positivos.

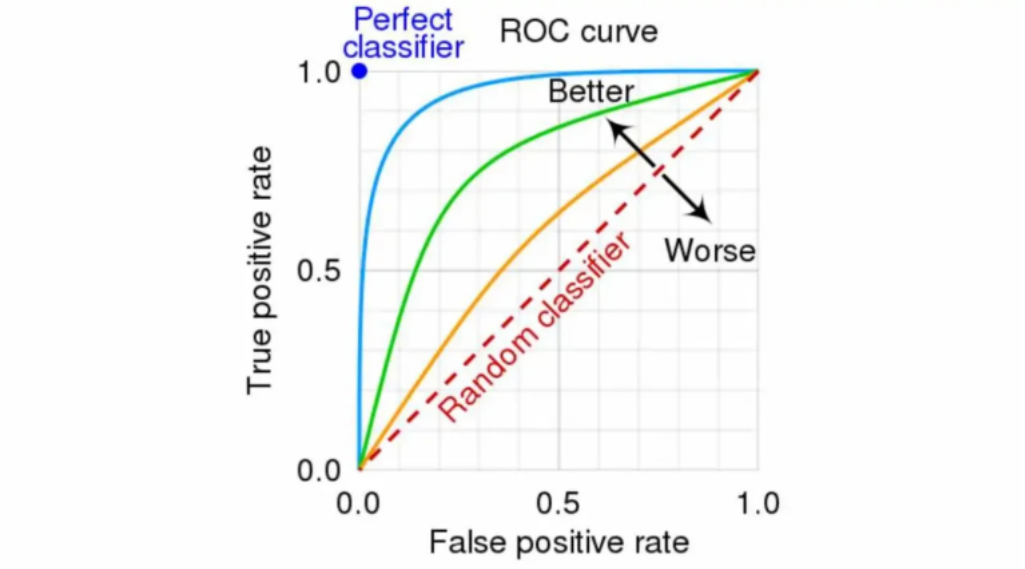
$$\text{Sensibilidade} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.3}$$

A especificidade (Equação 2.4) avalia a capacidade do método de detectar resultados negativos.

$$\text{Especificidade} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{2.4}$$

A curva ROC (Figura 2.2) é amplamente utilizada em tarefas de classificação para avaliar o desempenho de um modelo. Ela representa a taxa de verdadeiros positivos no eixo Y e a taxa de falsos positivos no eixo X. O canto superior esquerdo do gráfico corresponde ao ponto ideal, com taxa de falsos positivos igual a 0 e taxa de verdadeiros positivos igual a 1.

Figura 2.2: Curva ROC



Fonte: Figura adaptada de Thoma and Lee (2018)

De modo geral, quanto maior a área sob a curva ROC (*Area Under the Curve* – AUC), melhor o desempenho do classificador. A linha tracejada representa o desempenho de um classificador aleatório (AUC = 0,5). Curvas posicionadas abaixo dessa linha indicam que o modelo apresenta desempenho inferior ao acaso, ou seja, suas predições são sistematicamente incorretas.

Em problemas que envolvem múltiplas classes, a avaliação do desempenho é frequentemente estendida por meio da Média Macro (*macro-average*). Essa abordagem atribui o mesmo peso a todas as classes, calculando métricas como AUC-ROC, precisão, *recall* e F1-score de forma independente para cada categoria e, posteriormente, obtendo-se a média aritmética desses valores. Tal estratégia é particularmente relevante em cenários com distribuição desigual entre as classes, pois permite avaliar se o desempenho do sistema se mantém consistente entre as diferentes categorias diagnósticas, independentemente da frequência com que cada uma ocorre na amostra.

Além da curva ROC, a curva *Precision-Recall* (PR) é amplamente utilizada na avaliação de modelos de classificação, especialmente em cenários com forte desbalanceamento de classes. Diferentemente da ROC, a curva PR baseia-se na relação entre Precisão e Sensibilidade, sendo particularmente informativa quando o interesse está na correta identificação da classe minoritária (SAITO; REHMSMEIER, 2015).

3 TRABALHOS RELACIONADOS

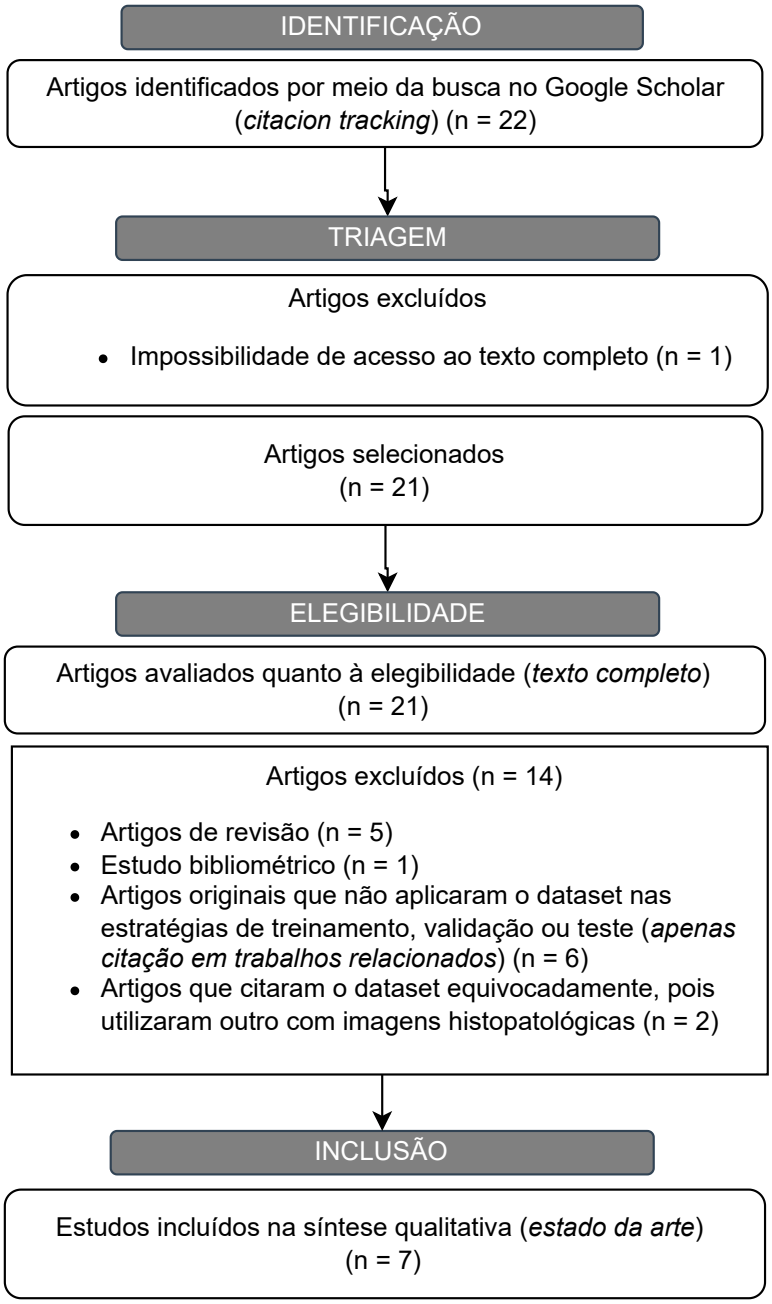
Os estudos apresentados neste capítulo fornecem evidências sobre o uso de técnicas de aprendizado de máquina e extração de características aplicadas a imagens clínicas para a identificação precoce de lesões bucais potencialmente malignas. Todos os trabalhos selecionados utilizaram o conjunto de dados *Dataset of Annotated Oral Cavity Images for Oral Cancer Detection* (PIYARATHNE et al., 2024), que compreende 3.000 imagens de alta resolução da cavidade oral, capturadas com *smartphones* e anotadas por especialistas. As classes que compõem o dataset são: saudável (*Healthy*), lesões orais potencialmente malignas (OPMD - *Oral Potentially Malignant Disorders*), benigno (*Benign*) e câncer bucal (OCA - *Oral Cancer*). Este mesmo conjunto de dados será utilizado no presente trabalho, buscando a comparabilidade dos resultados obtidos com os encontrados na literatura recente.

A identificação dos trabalhos foi realizada entre agosto e dezembro de 2025, com o objetivo de mapear a produção científica relacionada ao *dataset* de interesse. Foram incluídos estudos que empregaram o referido conjunto de dados de forma exclusiva ou em combinação com outras bases de dados. A estratégia de busca baseou-se no rastreamento prospectivo de citações (*forward citation tracking*), utilizando o artigo de lançamento do *dataset* como estudo-semente. A partir deste, foram identificadas, por meio do *Google Scholar*, as citações subsequentes, com o objetivo de localizar estudos que apresentassem implementações algorítmicas incorporando esses dados.

Foram considerados elegíveis os estudos que aplicaram o *dataset* de forma exclusiva ou em conjunto com outras bases de dados, desde que o *dataset* tenha sido efetivamente utilizado nos experimentos. A busca resultou em 22 artigos. Desses, um estudo foi excluído por não haver acesso ao texto integral. Os 21 artigos restantes foram submetidos à análise de elegibilidade. Nessa etapa, 14 estudos foram excluídos por não aplicarem o *dataset* nas estratégias de treinamento, validação ou teste, incluindo 5 artigos de revisão, 1 estudo bibliométrico, 6 estudos originais que mencionaram o conjunto de dados apenas como trabalho relacionado e 2 estudos que citaram equivocadamente o *dataset*, pois utilizaram outro com imagens histopatológicas. Assim, o corpo final de análise foi composto por 7 estudos originais, que constituem o estado da arte deste trabalho. A estratégia de busca está apresentada na Figura 3.1.

Devindi et al. (2024) propõem um pipeline multimodal para detecção precoce de desordens orais potencialmente malignas (OPMD), integrando imagens da cavidade oral

Figura 3.1: Estratégia de busca e seleção dos artigos



Fonte: Figura elaborada pela autora (2026).

obtidas por *smartphones* e metadados clínicos. O conjunto de dados inclui 2.271 imagens de lesões benignas (*Benign*), OPMD e OCA, com informações demográficas e de hábitos como uso de tabaco e ácool, sendo realizada a divisão por paciente para evitar vazamento de dados. A metodologia combina segmentação via U-Net (RONNEBERGER; FISCHER; BROX, 2015), extração de características com *encoders* SOTA (*State-of-the-Art*) e processamento de metadados com XGBoost (CHEN; GUESTRIN, 2016), utilizando fusão precoce. A abordagem multimodal com MobileNetV3-Large (HOWARD et al., 2019) alcançou 81% de acurácia. Como limitações, os autores destacam o tamanho reduzido do *dataset* e a restrição populacional. Além disso, observa-se na metodologia adotada que a tarefa foi conduzida como classificação binária (*Benign* vs. OPMD), com exclusão da classe saudável presente no *dataset* original e agrupamento das classes OPMD e OCA, o que pode introduzir limitações clínicas ao não distinguir categorias distintas. Adicionalmente, o estudo não incorporou técnicas de explicabilidade em sua metodologia, mas os autores reconhecem o seu valor e as propõe como parte de futuras melhorias.

O estudo de Yadav et al. (2025) propõe uma rede leve explicável com um *encoder* baseado em axial attention, originalmente introduzido por Ho et al. (2019), e um módulo de atenção guiado por rótulos desenvolvido no próprio estudo, visando aprimorar a discriminação de regiões relevantes. Para interpretabilidade, foi incorporado o Grad-CAM, permitindo a visualização das áreas mais relevantes para a decisão do modelo. Os resultados demonstraram acurácia média de 98,36%, F1-score de 97,98%, redução de 43% no número de parâmetros e de 35% no tempo de inferência, em comparação com RNCs tradicionais, além de uma forte correlação entre as regiões destacadas e as áreas patológicas reais. Entre as limitações, o estudo destaca a avaliação restrita a um único conjunto de dados, a ausência de validação externa, bem como a falta de validação clínica da interpretabilidade em ambiente real. Ressalta-se que o artigo não explicita se a separação das amostras foi realizada com base no paciente, o que poderia contribuir para mitigar o vazamento de dados e permitir uma avaliação mais realista do desempenho do modelo.

Amien et al. (2025) apresentam uma análise comparativa de arquiteturas convolucionais modernas para a classificação de doenças orais, priorizando o desempenho preditivo e na eficiência computacional. O estudo segue o framework OSEMN (*Obtain, Scrub, Explore, Model, and Interpret*) (MASON; WIGGINS, 2010), contemplando a coleta, a limpeza, a exploração, a modelagem e a interpretação dos dados, além de aplicar *data augmentation* para mitigar o desbalanceamento. Foram avaliadas as arquiteturas

ConvNeXt Tiny (LIU et al., 2022), MobileNetV3 Large (HOWARD et al., 2019) e EfficientNet B0 (TAN; LE, 2019), todas com pesos pré-treinados no ImageNet, considerando cenários com e sem *fine-tuning*. Os modelos foram testados em diferentes configurações de treino/validação/teste e avaliados quanto à acurácia e ao tempo de treinamento. Os resultados indicaram que a EfficientNet B0 alcançou a maior acurácia (98,21%) com menor complexidade computacional, enquanto a ConvNeXt Tiny obteve 96,96% com custo computacional mais elevado. A MobileNetV3 Large apresentou o menor tempo de treinamento (28s) e desempenho competitivo (96,44%). Em todos os modelos, a combinação de *data augmentation* e *fine-tuning* resultou em ganhos consistentes de desempenho. Entre as limitações, o estudo aponta a avaliação restrita a um único *dataset*, a ausência de validação externa e a falta de testes em ambiente clínico, além do alto custo computacional do ConvNeXt, que pode limitar sua aplicação prática. Ressalta-se que os autores não explicitam se a separação das amostras foi realizada com base no paciente, o que pode ter contribuído para uma avaliação mais otimista do desempenho dos modelos. Embora tenham priorizado métricas como acurácia e eficiência computacional, não foi indicado o uso de técnicas de explicabilidade. Considerando o contexto da saúde, a interpretabilidade é um aspecto fundamental, uma vez que favorece a compreensão das decisões do modelo por profissionais humanos.

Lee, Kim and Lee (2025) investigam o uso de modelos de difusão para gerar imagens sintéticas realistas de lesões orais, com o objetivo de aprimorar a acurácia diagnóstica em tarefas de classificação e detecção. A abordagem utiliza um *Stable Diffusion* (ROMBACH et al., 2022) ajustado via *DreamBooth* (RUIZ et al., 2023) para síntese de lesões por *inpainting* (LIU et al., 2018), a partir de máscaras geradas automaticamente pelo *Segment Anything Model*(SAM) (KIRILLOV et al., 2023). Os experimentos foram conduzidos em um conjunto interno (777 imagens) e em bases externas, incluindo dados do Hospital Odontológico da *Kyung Hee University* e o *dataset* de Piyarathne et al. (2024), que foi utilizado como parte do *external dataset* para gerar lesões sintéticas usando o modelo de difusão por inpainting, demonstrando a capacidade de generalização do modelo, já que não foi usado para o ajuste fino. As imagens sintéticas foram incorporadas ao treinamento de modelos de classificação (ResNet-50) (HE et al., 2016) e de detecção (YOLOv8) (JOCHER; CHAURASIA; QIU, 2023), resultando em melhorias consistentes de desempenho, com acurácia média de 97% e aumento do AUROC (*Area Under the Receiver Operating Characteristic curve*) para 0,9586 na classificação. Avaliações de qualidade indicaram SSIM (*Structural Similarity Index*) elevado, redução do FID (*Fré-*

chet Inception Distance) e maior foco das regiões patológicas, confirmadas por análises com Grad-CAM. Entre as limitações, os autores destacam a ausência de refinamento manual das máscaras, possíveis variações na fidelidade das imagens devido a diferenças de resolução e iluminação entre as bases de dados, a dependência de dados limitados para o ajuste do modelo e a falta de validação clínica em ambiente real. Adicionalmente, o estudo não detalha se a separação dos dados foi realizada por paciente, o que constitui uma ressalva importante para evitar o vazamento de dados.

O estudo de Aparna et al. (2025) investiga uma abordagem híbrida que combina aprendizado profundo clássico e computação quântica para a segmentação de lesões da mucosa oral, utilizando o *Oral Mucosal Lesion Dataset*, com 3.000 imagens anotadas em duas classes (lesão e cavidade oral). Ressalta-se que os autores utilizaram o mesmo *dataset* alvo deste trabalho, ainda que sob uma denominação distinta; contudo, sua descrição é apresentada de forma incompleta. A versão integral do *dataset*, disponível de forma restrita no Zenodo (2013), inclui marcações (lesão e cavidade oral) e quatro classes, sendo necessário solicitar autorização aos autores para acesso completo. Os autores avaliaram duas arquiteturas: uma U-Net clássica e uma rede neural convolucional quântica (QCNN) (CONG; CHOI; LUKIN, 2019), alimentada por versões reduzidas das imagens por meio da Análise de Componentes Principais (Principal Component Analysis – PCA)(JOLLIFFE; CADIMA, 2016). Os resultados indicaram melhor desempenho da U-Net (acurácia de 95,02%), enquanto a QCNN apresentou desempenho inferior, embora com potencial para modelagem de padrões complexos. Entre as limitações informadas pelos autores, destacam-se as restrições atuais do hardware quântico, a necessidade de redução de dimensionalidade e a condução dos experimentos em uma formulação binária. Observa-se ainda que, devido ao acesso restrito ao conjunto de dados completo, os autores não detalham aspectos relevantes, como a separação por paciente, o que pode impactar a robustez da avaliação. Embora o estudo reconheça a importância da interpretabilidade, não são descritos métodos específicos de explicabilidade para análise do comportamento dos modelos.

Gürses et al. (2025) investigam uma abordagem clássica, baseada na análise de textura superficial, para a diferenciação entre lesões orais benignas e malignas, explorando um caminho alternativo aos métodos baseados em *deep learning*. A análise é conduzida com base em 275 imagens intraorais provenientes de um banco institucional e do *dataset* alvo deste estudo, com as lesões manualmente delimitadas por especialistas. Foi realizado um processo de seleção rigoroso em duas etapas. Do conjunto de dados

público, de acesso aberto porém restrito, originalmente composto por 3.000 imagens, foram selecionados 82 registros após avaliação por um painel de especialistas. Esses foram combinados a 193 registros provenientes do arquivo institucional, que inicialmente continha 405 imagens. As imagens são convertidas para escala de cinza e representações tridimensionais da textura superficial são geradas, permitindo a extração de características estatísticas como entropia, curtose, coeficiente de variação e medidas de dispersão. Essas características são analisadas estatisticamente por meio de MANOVA (HAIR et al., 2018) e utilizadas para treinar um classificador *Support Vector Machine* (SVM) (CORTES; VAPNIK, 1995), que alcança sensibilidade elevada (99,2%), acurácia de 90,5% e AUC de 0,939, com a entropia destacando-se como um forte marcador de malignidade. Apesar do desempenho robusto, os autores indicam limitações relacionadas ao caráter retrospectivo dos dados, à ausência de geração de novos dados e à restrição ao diagnóstico baseado apenas na textura, o que indica a necessidade de integração com outros parâmetros clínicos e metodologias mais escaláveis. Como limitações observadas, destaca-se a condução do estudo como um problema de classificação binária (lesões benignas vs. malignas), o que pode restringir a capacidade do modelo em diferenciar categorias clínicas mais específicas. Além disso, não é informado se a separação das amostras foi realizada com base no paciente, o que pode torná-lo suscetível ao vazamento de dados e, consequentemente, a uma avaliação potencialmente otimista do desempenho.

Mukherjee et al. (2025) utilizaram um conjunto de dados público de lesões orais, inicialmente composto por 3.000 imagens de 714 pacientes, posteriormente revisado por especialista, resultando em 2.438 imagens de 653 pacientes com ROIs anotadas. O estudo propõe uma abordagem multimodal para classificação em quatro classes (saudável, benigno, OPMD e OCA), integrando *embeddings* morfológicos extraídos por ConvNeXtV2 (WOO et al., 2023), descritores espectro-texturais derivados de reconstrução hiperespectral de imagens RGB e dados clínicos em nível de paciente. As características foram combinadas por meio de um Meta-Aprendiz Heurístico Incremental (IHML), com validação cruzada em 5-folds e separação de 15% dos pacientes para teste não visto. O modelo alcançou F1 macro de 66,23% e AUC-ROC de 84,45%, superando métodos tradicionais como Regressão Logística, Random Forest, SVM e XGBoost. Entre as limitações destacadas pelos autores estão a reconstrução hiperespectral a partir de imagens RGB, que pode comprometer a fidelidade espectral, e a dependência de dados anotados. Adicionalmente, o estudo não contempla o uso de técnicas de explicabilidade, o que representa uma lacuna relevante, especialmente no contexto clínico.

A Tabela 3.1 apresenta uma síntese comparativa dos estudos incluídos, destacando as arquiteturas empregadas, o uso do *dataset*-alvo e os principais resultados reportados.

Os sete estudos analisados neste capítulo demonstram a predominância de abordagens de *deep learning*, empregando desde redes convolucionais e modelos de difusão até arquiteturas multimodais e híbridas, incluindo experimentações com computação quântica. Observa-se a aplicação dessas técnicas tanto na classificação e segmentação de lesões, quanto na geração de imagens sintéticas para enriquecer os *datasets*. Apesar dos avanços significativos em acurácia e eficiência computacional, uma lacuna crítica observada é que somente Devindi et al. (2024) e Mukherjee et al. (2025) detalharam a estratégia de divisão dos dados a nível de paciente, um aspecto essencial para evitar o *data leakage* e a superestimação do desempenho, especialmente considerando que o *dataset* principal (PIYARATHNE et al., 2024) é composto por 3.000 imagens de apenas 714 indivíduos. Além disso, a explicabilidade dos modelos, um ponto essencial para a confiança e adoção de IA na saúde, é pouco explorada, apenas Yadav et al. (2025) e Lee, Kim and Lee (2025) relataram a utilização de Grad-CAM para essa finalidade. Outras limitações importantes comuns a esses trabalhos incluem a variabilidade inerente às imagens geradas por diferentes aparelhos de *smartphones*, e o número ainda reduzido de dados anotados, o que pode impactar a robustez e a generalização dos modelos. Além disso, limitações recorrentes apontadas pelos próprios autores são a necessidade de validação clínica em ambientes reais e a diversificação dos conjuntos de dados para aprimorar a sua capacidade de generalização.

A Tabela 3.2 sintetiza as tendências, desafios e potenciais caminhos futuros para pesquisas sobre detecção de lesões bucais usando aprendizado de máquina conforme os estudos revisados.

Tabela 3.1: Resumo das principais características dos estudos quanto à metodologia e aos resultados

Estudo	*Metodologia*	*Resultados*
Devindi et al. (2024)	*Deep learning* multimodal baseado na MobileNetV3-Large, integrando imagens clínicas e metadados dos pacientes. Estratégia de divisão dos dados a nível de paciente para evitar vazamento de dados.	Acurácia: 81%.
Yadav et al. (2025)	Arquitetura leve combinando CNNs, *Axial Transformer Encoder* e *Label-Guided Attention Module*, com interpretabilidade via Grad-CAM.	Acurácia média: 98,36%; F1-score: 97,98%.
Amien et al. (2025))	Análise comparativa entre EfficientNet B0, ConvNeXt Tiny e MobileNetV3 Large, com *data augmentation* e *fine-tuning*.	EfficientNet B0 alcançou acurácia de 98,21% com menor complexidade computacional; ConvNeXt Tiny obteve 96,96% com maior custo; MobileNetV3 Large apresentou menor tempo de treinamento (28 s) e desempenho competitivo (96,44%).
Lee, Kim and Lee (2025)	Geração de imagens sintéticas por modelos de difusão ajustado via *DreamBooth* aplicada à classificação (ResNet-50) e detecção (YOLOv8). Utilização de Grad-CAM para explicabilidade	Acurácia média (classificação): 97%; AUROC: 0,9586; ganhos relevantes nas métricas de detecção.
Aparna et al. (2025)	Abordagem híbrida para classificação binária, combinando U-Net clássica e Rede Neural Convolucional Quântica (QCNN) para segmentação.	U-Net: Acurácia de 95,02%.
Gürses et al. (2025)	Análise clássica de textura superficial com extração de características estatísticas e classificação via SVM.	Acurácia: 90,5%; Sensibilidade: 99,2%; AUC: 0,939.
Mukherjee et al. (2025)	Pipeline multimodal combinando *embeddings* morfológicos, características espectrais e dados clínicos, utilizando Meta-Aprendizagem Heurística Incremental (IHML). Estratégia de divisão dos dados a nível de paciente para evitar vazamento de dados.	F1 macro: 66,23%; Acurácia: 64,56%; PR-AUC: 69,10%; AUC-ROC: 84,45% no conjunto de teste.

Fonte: Elaborada pela autora com base na literatura revisada.

Tabela 3.2: Síntese dos Trabalhos Relacionados

Categoria	Principais Aspectos
Tendências Metodológicas	• Uso de RNCs leves e eficientes. • Abordagens multimodais: combinação de imagens com metadados clínicos (DEVINDI et al., 2024) e dados espectrais, morfológicos e demográficos (MUKHERJEE et al., 2025). • Geração e aumento de dados: uso de modelos de difusão (LEE; KIM; LEE, 2025) e *data augmentation*. • Explicabilidade: uso de Grad-CAM (YADAV et al., 2025; LEE; KIM; LEE, 2025).
Limitações Relatadas	• Escassez de *datasets* anotados. • Falta de validação externa e clínica. • Tamanho e diversidade ainda restritos. • Dependência de reconstruções espectrais ou dados sintéticos. • Custo computacional elevado.
Limitações Observadas	• Ausência de detalhamento sobre divisão por paciente, com potencial risco de vazamento de dados. • Ausência ou uso limitado de técnicas de explicabilidade em alguns estudos.
Direções Futuras Relatadas	• Ampliação e diversificação dos *datasets*. • Ensaios clínicos e implementações em ambiente real. • Otimização entre precisão, eficiência e interpretabilidade. • Integração multimodal com dados histopatológicos e genéticos.

Fonte: Elaborada pela autora (2026).

4 MATERIAIS E MÉTODOS

Este capítulo apresenta os materiais e métodos utilizados na condução deste estudo, contemplando o ambiente de desenvolvimento computacional adotado, a descrição do conjunto de dados, os procedimentos de preparação e processamento das imagens e as estratégias de modelagem. São detalhados os critérios de divisão dos dados, o treinamento e validação dos modelos, além dos aspectos éticos relacionados ao uso de dados e ferramentas de apoio.

4.1 Ambiente de Desenvolvimento

Para o desenvolvimento e a execução dos algoritmos de manipulação de dados e implementação dos modelos de classificação, foi utilizado o ambiente de computação em nuvem *Google Colab Pro*. A escolha desta plataforma foi pela oferta de recursos computacionais de alto desempenho, incluindo acesso a Unidades de Processamento Gráfico (GPUs) de alta capacidade, essenciais para acelerar o treinamento dos modelos.

- As implementações foram realizadas na linguagem Python (v.3.12), utilizando bibliotecas especializadas:

- TensorFlow (v.2.19) e Keras (v.3.10): para a estruturação, treinamento e validação das arquiteturas de Aprendizado Profundo.

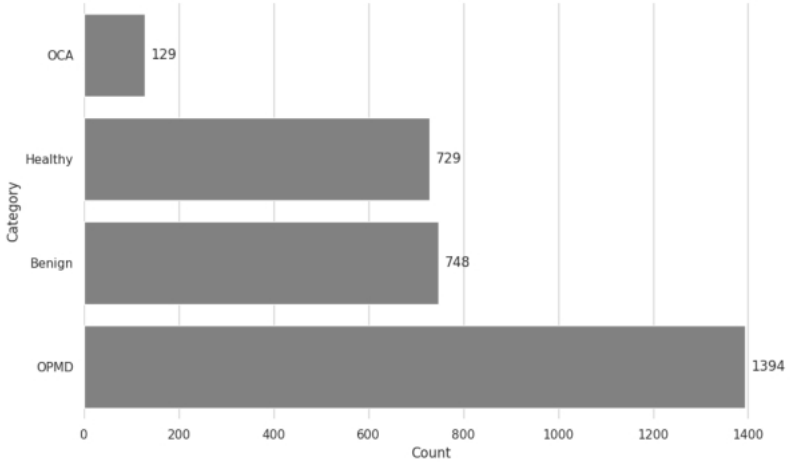
- OpenCV (v.4.13) e Pillow (v.11.3): foram utilizadas no pré-processamento das imagens, incluindo operações de redimensionamento e aplicação de corte.

- Pandas (v.2.2) e NumPy (v.2.0): utilizadas para a manipulação de estruturas de dados e execução de operações matemáticas.

- Scikit-learn (v.1.6): aplicada para a extração de métricas de desempenho e divisão dos conjuntos de treino e teste.

- Matplotlib (v.3.10) e Seaborn (v.0.13): para geração de gráficos e visualização das curvas de aprendizado.

Figura 4.1: Distribuição das classes do *dataset*



Fonte: Autora (2026).

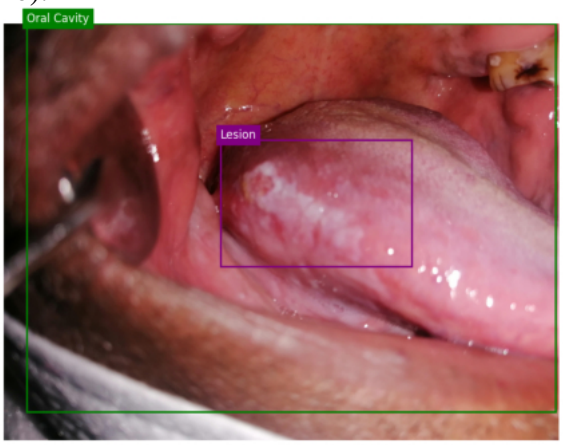
4.2 Conjunto de Dados

Para a realização do estudo, foi utilizado o *Dataset of Annotated Oral Cavity Images for Oral Cancer Detection* (PIYARATHNE et al., 2024). O conjunto é composto por 3.000 imagens da cavidade oral, capturadas com câmeras de *smartphones* em ambientes clínicos do Sri Lanka, refletindo condições reais de iluminação, variabilidade anatômica e diferenças entre pacientes. Essa característica torna o dataset particularmente adequado para o desenvolvimento e validação de modelos voltados à prática clínica. As imagens estão distribuídas em quatro categorias, definidas e validadas por especialistas: Saudável (*Healthy*), Benigna (*Benign*), Desordens orais potencialmente malignas (OPMD) e Câncer oral (OCA). A Figura 4.1 apresenta a distribuição das classes e evidencia o desequilíbrio entre elas.

Cada imagem possui anotações detalhadas no formato COCO (*Common Objects in Context*) (Figura 4.2), introduzido por Lin et al. (2014) , incluindo contornos da cavidade oral e das regiões lesionadas, o que possibilita aplicações em tarefas de classificação, detecção e segmentação.

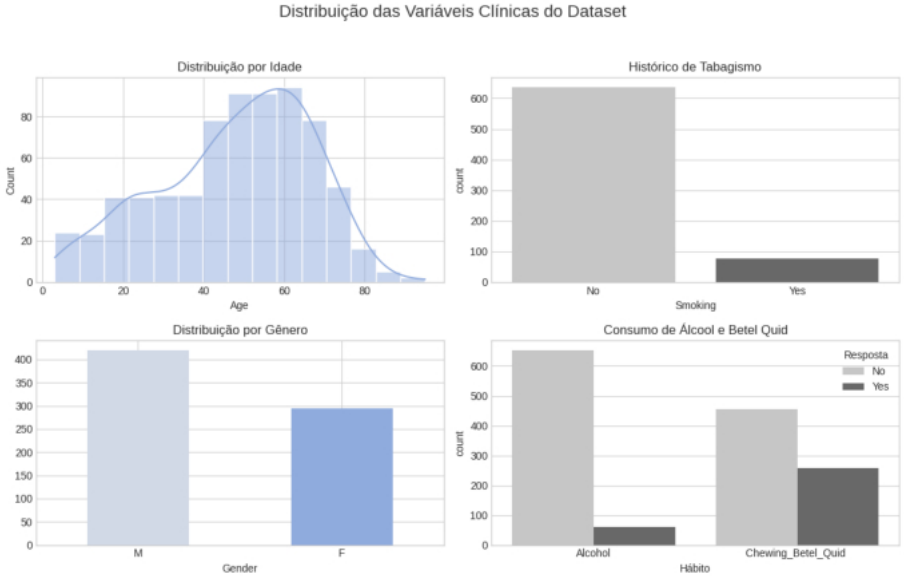
O *dataset* acompanha arquivos de metadados essenciais para análises clínicas. Dentre eles, é possível encontrar informações sobre o ID da imagem, a categoria atribuída, o diagnóstico clínico e o número de regiões anotadas, como também, idade, sexo, histórico de tabagismo, consumo de álcool, de *betel quid* (uma mistura mastigada que inclui noz de areca (*areca nut*), envolta em folha de *betel* (*Piper betle*), frequentemente associada a outros componentes, como cal e tabaco) (IARC Working Group on the Evaluation of Carcinogenic Risks to Humans, 2004) (Figura 4.3) e o total de imagens capturadas por

Figura 4.2: Exemplo de imagem anotada formato COCO destacando a região oral (verde) e a região da lesão (roxo).



Fonte: Piyarathne et al. (2024).

Figura 4.3: Metadados dos Indivíduos



Fonte: Autora (2026).

indivíduo.

4.3 Pré-processamento

A divisão foi implementada utilizando o *StratifiedGroupKfold* da biblioteca *scikit-learn*, permitindo a separação estratificada (mantendo a proporção entre as classes) e agrupada por paciente. Após o particionamento, os dados foram estruturados em três subconjuntos: *train, val* e *test*. Cada subconjunto recebeu apenas as imagens dos pacientes pertencentes ao respectivo *split*, garantindo a separação entre indivíduos nas três pastas.

Após a divisão por paciente, a condução dos experimentos baseou-se em dois pi-

pelines distintos de processamento. No primeiro, as imagens originais (*full-frame*) foram utilizadas integralmente, sendo apenas convertidas para o espaço de cores RGB e redimensionadas para 224 X 224 *pixels*. O segundo pipeline foi direcionado à extração da Região de Interesse (ROI, do inglês *Region of Interest*), utilizando anotações no padrão *Common Objects in Context* (COCO).

Para as imagens contendo anotações manuais, a extração da região de interesse (lesão) seguiu um pipeline estruturado. Em casos com múltiplas marcações, o algoritmo seleciona automaticamente a *bounding box* de maior área, assumindo que esta corresponde à lesão principal e garantindo a captura do maior volume de características morfológicas relevantes. Em seguida, aplicou-se um *padding* de 10 *pixels* ao redor da região delimitada, com tratamento das coordenadas para evitar extrapolação dos limites da matriz da imagem. Esse processo assegura a preservação de bordas diagnósticas e do tecido perilesional adjacente.

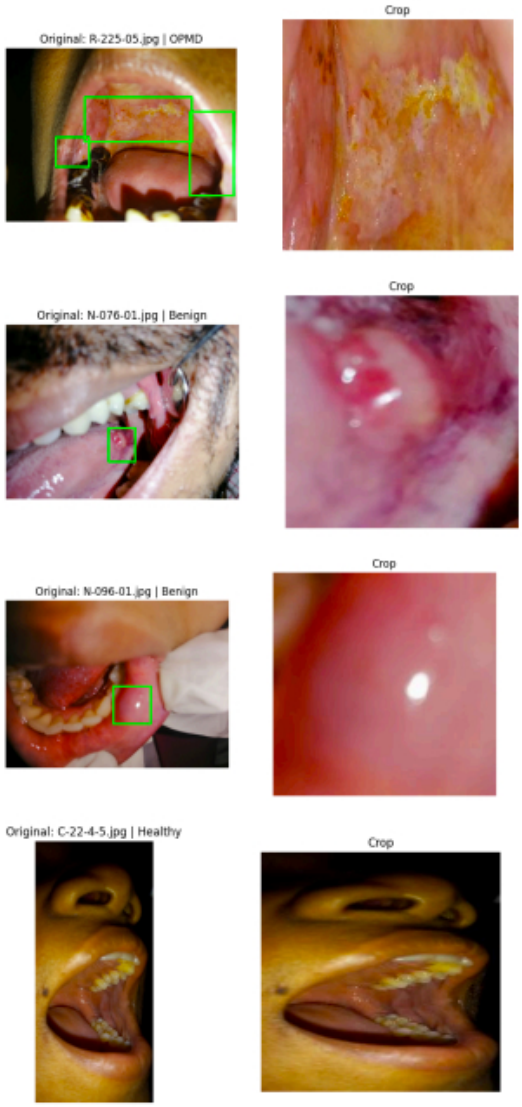
Para as imagens sem anotação, geralmente pertencentes à classe saudável, optou-se por não aplicar *crop*. Essas imagens já apresentam naturalmente a cavidade oral enquadrada, e a ausência de uma região patológica específica tornaria arbitrária a definição de um recorte. A utilização da imagem integral evita a introdução de padrões artificiais, preservando a representação anatômica natural da mucosa. Essa decisão metodológica busca prevenir a introdução de viés no pré-processamento e garantir consistência entre o conteúdo clínico e as operações aplicadas a cada classe.

Por fim, todas as amostras foram redimensionadas para a dimensão de 224×224 *pixels*, garantindo a uniformidade dos tensores de entrada para as redes neurais. Embora esse processo possa introduzir leves distorções geométricas na representação visual da lesão, ele é necessário para a padronização exigida pelos modelos. A Figura 4.4 ilustra, respectivamente, a imagem original, a aplicação da *bounding box* com *padding* e o resultado final da ROI segmentada e redimensionada.

4.4 Divisão dos dados: conjuntos de Treinamento, Validação e Teste

O conjunto de dados foi dividido em três subconjuntos independentes (treinamento, validação e teste) utilizando a estratégia *Stratified Group Split*. A divisão foi realizada de forma estratificada por paciente, garantindo que todas as imagens pertencentes a um mesmo indivíduo fossem mantidas no mesmo subconjunto. Essa abordagem foi adotada para evitar *data leakage*, assegurando que o modelo não tivesse acesso, durante o

Figura 4.4: Exemplos de imagens após o recorte (crop) das lesões e redimensionamento



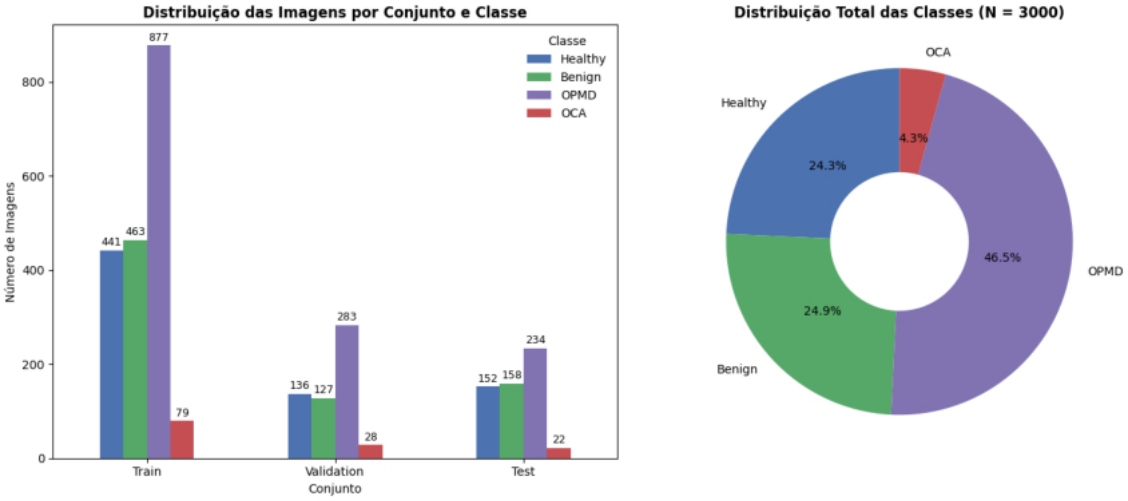
Fontes: Piyarathne et al. (2024) e Autora (2026).

treinamento, a imagens do mesmo paciente presentes na validação ou no teste. Para todas as arquiteturas avaliadas, foi adotada a proporção $\approx 60\%$ para treinamento, $\approx 20\%$ para validação e $\approx 20\%$ para teste, garantindo comparabilidade entre os modelos. A Figura 4.5 apresenta a quantidade e distribuição das imagens por classe.

4.5 Arquitetura e Configuração dos Modelos

Para a classificação das lesões orais, foram selecionadas arquiteturas de aprendizado profundo amplamente utilizadas na literatura, com diferentes características arquiteturais e custos computacionais. A escolha dessas arquiteturas foi fundamentada em

Figura 4.5: Distribuição das imagens por classe



Fonte: Autora (2026)

estudos prévios, em especial no trabalho de (AMIEN et al., 2025), que empregou as arquiteturas MobileNetV3 *Large*, EfficientNetB0 e ConvNeXt *Tiny*, sendo adotada neste estudo a variante *Small* da MobileNetV3 por razões de eficiência computacional.

- MobileNetV2 e V3: a MobileNetV2 (SANDLER et al., 2018) é utilizada como o linha de base deste estudo devido à sua introdução de blocos residuais invertidos e convoluções separáveis em profundidade, que otimizam o custo computacional. A MobileNetV3 (HOWARD et al., 2019) evolui essa linhagem através da Busca de Arquitetura Neural (NAS), incorporando mecanismos de *Squeeze-and-Excitation* e a função de ativação *hard-swish*, sendo voltada para cenários de alta eficiência em dispositivos de borda (*edge computing*), como *smartphones*, onde há limitação de processamento e memória.

- EfficientNetB0: esta arquitetura fundamenta-se na técnica de *Compound Scaling* (TAN; LE, 2019), que propõe o escalonamento balanceado entre a profundidade, a largura e a resolução da rede. Tal abordagem permite que a EfficientNet alcance níveis de precisão superiores a arquiteturas mais densas, mantendo uma carga computacional reduzida, o que é ideal para a análise de detalhes morfológicos em imagens clínicas.

- ConvNeXt *Tiny*: representando o estado da arte das redes convolucionais, a ConvNeXt (LIU et al., 2022) "moderniza"as CNNs ao integrar conceitos dos *Vision Transformers* (ViT). Através do uso de *kernels* de maior dimensão (7 X 7), camadas de *LayerNorm* e ativações GeLU, o modelo oferece uma capacidade de aprendizado

superior para padrões complexos, mantendo a simplicidade operacional das convoluções tradicionais.

- Proposta Multimodal: a arquitetura multimodal foi desenvolvida utilizando uma estratégia de fusão tardia (*late fusion*). O design compõe-se de dois ramos principais:
  - Ramo de Imagem: utiliza o modelo com melhor desempenho no estudo, a rede MobileNetV3 *Small*, como extrator de características, onde a última camada de *pooling* gera um vetor representativo do conteúdo visual da lesão.
  - Ramo de Metadados: consiste em uma sub-rede densa (*fully connected*) que recebe as variáveis clínicas normalizadas, processando-as através de camadas lineares com ativação ReLU.

A integração ocorre na camada de concatenação que une os vetores de ambos os ramos em uma única representação híbrida. Este vetor resultante é alimentado em camadas densas de classificação final, onde a Softmax determina a probabilidade de cada uma das quatro categorias patológicas.

O treinamento dos modelos experimentais foi conduzido em duas etapas complementares. Na primeira fase, denominada extração de características, o *backbone* convolucional foi mantido totalmente congelado, sendo treinadas apenas as camadas densas da cabeça classificadora (*classification head*). Na segunda fase, foi realizado o *fine-tuning* parcial do modelo, liberando uma fração das camadas finais do *backbone* para adaptação ao domínio clínico.

Inicialmente, foi adotada uma estratégia padronizada de *fine-tuning*, liberando 20% finais das camadas para todas as arquiteturas avaliadas. Posteriormente, foi realizado um ajuste específico para cada modelo. A MobileNetV3 apresentou melhor desempenho com o descongelamente de apenas 15% das camadas finais, enquanto a MobileNetV2 e a EfficientNetB0 obtiveram melhores resultados ao liberar aproximadamente 35% das camadas finais. Para a arquitetura ConvNext optou-se pela estratégia de descongelamento por bloco (*Stage* 4). Esse refinamento evidenciou a importância de adaptar o *fine-tuning* à profundidade e à complexidade de cada *backbone*.

Além do ajuste das camadas convolucionais finais, o *classification head* foi mantido treinável durante toda a fase de *fine-tuning*. Esse bloco foi composto por uma camada de *Global Average Pooling*, seguida de uma camada densa com ativação ReLU, duas camadas de *Dropout* (taxa = 0.4 e 0.3) para regularização e, por fim, a camada de saída com ativação softmax para classificação multiclasse. A Figura 4.6 apresenta a configuração da

Figura 4.6: Modelo MobileNetV2

Layer (type)	Output Shape	Param #
input_layer_1 (InputLayer)	(None, 224, 224, 3)	0
mobilenetv2_1.00_224 (Functional)	(None, 7, 7, 1280)	2,257,984
global_average_pooling2d (GlobalAveragePooling2D)	(None, 1280)	0
dropout (Dropout)	(None, 1280)	0
dense (Dense)	(None, 256)	327,936
dropout_1 (Dropout)	(None, 256)	0
dense_1 (Dense)	(None, 4)	1,028

Total params: 2,586,948 (9.87 MB)
Trainable params: 328,964 (1.25 MB)
Non-trainable params: 2,257,984 (8.61 MB)

Fonte: Autora (2026)

rede base MobileNetV2 utilizada. A mesma estratégia foi aplicada às demais arquiteturas avaliadas.

A arquitetura proposta para o modelo multimodal (arquitetura com o melhor desempenho nos experimentos + dados clínicos), combina um vetor de características visuais (576 dimensões) com um vetor clínico (128 dimensões), totalizando um espaço de características de 704 neurônios. A Figura 4.7 apresenta a parte final (*late fusion*) da rede multimodal.

4.6 Experimentos

A fase experimental inicial foi estruturada em dois pipelines principais para identificar a melhor estratégia de pré-processamento e manejo de dados, utilizando a MobileNetV2 como arquitetura base:

- Pipeline A (Análise de Escopo): Comparação entre o uso de imagens recortadas na região de interesse (*Crop*) versus o uso da imagem *Full Frame*. O objetivo foi avaliar se a preservação do contexto global interfere na capacidade de localização da patologia pela rede.
- Pipeline B (Otimização de Treino): Estratégias para lidar com desbalanceamento de classes (*Class Weights Balanced* vs. *Class Weights Smoothing*(sqrt)) e o impacto da aplicação de *data augmentation*.

Essa abordagem possibilitou a avaliação isolada do impacto de cada técnica, antes da aplicação em arquiteturas mais complexas. Uma vez definida a melhor combinação de

Figura 4.7: Parte final do modelo multimodal

conv_1_bn (BatchNormalizatio…	(None, 7, 7, 576)	2,304	conv_1[0][0]
activation_17 (Activation)	(None, 7, 7, 576)	0	conv_1_bn[0][0]
tab (InputLayer)	(None, 9)	0	-
gap (GlobalAveragePool…	(None, 576)	0	activation_17[0]…
tab_fc (Dense)	(None, 128)	1,280	tab[0][0]
drop_img (Dropout)	(None, 576)	0	gap[0][0]
drop_tab (Dropout)	(None, 128)	0	tab_fc[0][0]
fusion (Concatenate)	(None, 704)	0	drop_img[0][0], drop_tab[0][0]
fc1 (Dense)	(None, 256)	180,480	fusion[0][0]
drop1 (Dropout)	(None, 256)	0	fc1[0][0]
saida (Dense)	(None, 4)	1,028	drop1[0][0]

Total params: 2,299,550 (8.77 MB)
Trainable params: 588,820 (2.25 MB)
Non-trainable params: 533,088 (2.03 MB)
Optimizer params: 1,177,642 (4.49 MB)

Fonte: Autora (2026).

estratégias nos Pipelines A e B, essa configuração foi replicada nas arquiteturas MobileNetV3, EfficientNetB0, ConvNeXt *Tiny* e no modelo multimodal.

A escolha da arquitetura MobileNetV2 como modelo base para o desenvolvimento dos pipelines experimentais justifica-se por seu equilíbrio otimizado entre eficiência computacional e capacidade de representação em tarefas de visão computacional médica. Por ser uma rede consolidada na literatura e projetada para dispositivos de baixo recurso, a MobileNetV2 permite estabelecer uma linha de base de desempenho robusto, facilitando a análise do impacto de variáveis críticas como o uso de *class weights* e *data augmentation* em imagens completas.

Uma vez definida a configuração ideal de hiperparâmetros e pré-processamento nesta arquitetura, os resultados serviram como métrica comparativa para a transição para modelos de maior complexidade, como a ConvNeXt *Tiny* e a EfficientNetB0, garantindo que ganhos de acurácia sejam atribuídos à superioridade arquitetural e não a variações no treinamento.

Para mitigar o risco de sobreajuste, foi empregado o *callback Early Stopping* exclusivamente na etapa de *fine-tuning* por 15 épocas e monitoramento da perda de validação (val_loss). Já a fase de extração de características foi conduzida por 30 épocas para as imagens *Full Frame* e 20 épocas para as imagens *Crop*, sem a interrupção prematura do *Early Stopping*, a fim de garantir a convergência inicial do classificador. Adicionalmente,

utilizou-se o *callback Model Checkpoin* para salvar automaticamente os pesos que apresentaram o melhor desempenho no conjunto de validação, sendo esse modelo restaurado e avaliado no conjunto de teste.

Todos os modelos experimentais foram treinados com *batch size* de 32 imagens e entrada padronizada de 224×224 pixels (IMG_SIZE = 224). A função de perda utilizada foi a *Sparse Categorical Crossentropy*, e o otimizador adotado foi o Adam (DIEDERIK, 2014), com taxa de aprendizado inicial de 1×10^{-4} e estratégia de agendamento adaptativo baseada na validação (*ReduceLROnPlateau*). Para assegurar a reprodutibilidade dos experimentos, foi definida uma semente fixa (SEED = 42) em todas as etapas de treinamento e validação.

4.7 Técnicas de IA Explicável

A técnica LIME (RIBEIRO; SINGH; GUESTRIN, 2016) foi utilizada como etapa intrínseca a cada experimento. Esta abordagem foi aplicada sistematicamente após o treinamento de cada arquitetura, permitindo uma análise qualitativa comparativa entre os modelos. O objetivo foi verificar se a redução do ruído visual e a inclusão dos metadados clínicos na rede multimodal resultaram em uma atribuição de relevância mais precisa sobre as características morfológicas das lesões. Dessa forma, o LIME atuou como um critério de validação, assegurando que o melhor modelo final fosse escolhido não apenas pela maior acurácia, mas pela sua capacidade de identificar corretamente os biomarcadores visuais de relevância clínica, conferindo maior interpretabilidade ao processo de diagnóstico auxiliado por computador. Complementarmente, a técnica Grad-CAM (SELVARAJU et al., 2017) foi aplicada especificamente às arquiteturas na modalidade imagens completas (Full Frame) em sua melhor configuração, utilizando pesos de classe suavizados e *data augmentation* para corroborar a localização das ativações neurais.

4.8 Resumo da metodologia

A metodologia adotada neste estudo consiste no desenvolvimento de um sistema de classificação de lesões bucais com base em imagens clínicas. O conjunto de dados, composto por imagens intraorais anotadas, foi previamente organizado e submetido a um rigoroso protocolo de separação por paciente, a fim de evitar vazamento de dados.

Os experimentos foram realizados com a imagem completa (*F*ull Frame) e com imagens recortadas na região de interesse (lesão) - (*Crop*). As regiões de interesse foram obtidas a partir das anotações em formato COCO, utilizando *bounding boxes* com aplicação de *padding*, seguidas de redimensionamento das imagens para 224 × 224 *pixels*. Essa padronização permitiu a utilização de arquiteturas de Redes Neurais Convolucionais pré-treinadas por meio de aprendizado por transferência.

Foram avaliadas abordagens baseadas em imagens utilizando inicialmente a MobileNetV2 como rede de base. Posteriormente, foram analisadas outras arquiteturas, incluindo MobileNetV3, EfficientNetB0 e ConvNeXt *Tiny*. Por fim, foi implementada uma estratégia multimodal, integrando informações clínicas associadas aos pacientes.

Adicionalmente, foram incorporadas técnicas de Inteligência Artificial Explicável (GradCAM e LIME), visando tornar as decisões dos modelos mais interpretáveis no contexto clínico.

4.9 Considerações Éticas

O presente estudo utilizou um conjunto de dados público, porém de acesso restrito, disponibilizado mediante solicitação aos autores, respeitando as condições de uso estabelecidas. As imagens utilizadas não continham informações identificáveis dos indivíduos.

Adicionalmente, foram empregados modelos da família *Gemini (Google)* como ferramentas de apoio: a versão *Gemini* 3 foi utilizada para o refinamento textual, verificação de ortografia e clareza do conteúdo elaborado pela autora; enquanto a versão *Gemini* 2.5 *Flash*, integrada ao ambiente *Google Colab*, auxiliou na identificação de erros sintáticos e depuração dos códigos. Todo o conteúdo gerado ou sugerido pelas ferramentas foi posteriormente revisado, editado e validado, assegurando a integridade técnica e a autoria do trabalho.

5 RESULTADOS

Este capítulo apresenta os resultados obtidos a partir dos experimentos realizados conforme a metodologia proposta, incluindo a avaliação das arquiteturas tanto em imagens completas (*Full Frame*) quanto em regiões recortadas (*crop*), além da estratégia multimodal. São apresentadas as métricas de desempenho dos modelos, bem como gráficos, tabelas comparativas e exemplos de explicabilidade.

5.1 Modelo MobileNetV2 - *Full Frame*

O experimento inicial utilizando as imagens completas, denominadas neste trabalho como *Full Frame*, foi conduzido sem *data augmentation*, visando estabelecer uma linha de base e avaliar o comportamento do modelo diante do severo desbalanceamento das classes. Conforme apresentado na Tabela 5.1, os resultados foram limitados, com acurácia de 50,35% e uma *Loss* elevada de 1,1338, evidenciando a dificuldade da arquitetura em convergir sob tais condições.

Tabela 5.1: Desempenho do modelo MobileNetV2 (Full Frame) no conjunto de teste sem *Data Augmentation* e com *Class Weights* balanceada

Classe	*Precision*	*Recall*	*F1-score*	*Support*
Healthy	0.54	0.62	0.57	152
Benign	0.56	0.36	0.44	158
OPMD	0.59	0.55	0.57	234
OCA	0.08	0.27	0.13	22
Accuracy		0.5035		
Loss		1.1338		
Macro Avg	0.44	0.45	0.43	566
Weighted Avg	0.55	0.50	0.52	566

Fonte: Autora (2026)

Em uma etapa subsequente, foi conduzido o treinamento do segundo modelo, mantendo-se a ausência de técnicas de aumento de dados, porém introduzindo o balanceamento de classes suavizado. Os resultados quantitativos detalhados na Tabela 5.2 demonstram uma evolução no desempenho: a acurácia aumentou para 52,47%, acompanhada por uma redução na função de perda (1,0711). Este comportamento sugere que o ajuste nos pesos das classes contribuiu para uma melhora discreta na generalização inicial do modelo frente ao desbalanceamento do *dataset*.

Tabela 5.2: Desempenho do modelo MobileNetV2 (Full Frame) no conjunto de teste sem *Data Augmentation* e com *Class Weights* suavizada

Classe	*Precision*	*Recall*	*F1-score*	*Support*
Healthy	0.56	0.57	0.57	152
Benign	0.61	0.32	0.42	158
OPMD	0.53	0.67	0.59	234
OCA	0.09	0.14	0.11	22
Accuracy		0.5247		
Loss		1.0711		
Macro Avg	0.45	0.42	0.42	566
Weighted Avg	0.54	0.52	0.52	566

Fonte: Autora (2026)

Nos dois experimentos subsequentes, foi introduzida a técnica de aumento de dados durante a etapa de treinamento. A configuração adotada incluiu transformações espaciais moderadas, como rotações de até 15°, variações de zoom de 10% e espelhamento horizontal (*horizontal flip*). Estas operações visam expandir a variabilidade do conjunto de treino, permitindo que o modelo aprenda características invariantes à orientação e escala, mitigando o risco de *overfitting*. A técnica de aumento de dados foi aplicada de maneira uniforme a todas as categorias do *dataset*. Embora essa estratégia tenha expandido a variabilidade morfológica das amostras, a proporcionalidade original entre as classes foi mantida, sendo o desbalanceamento numérico tratado especificamente através da técnica de pesos de classe (*class weights*). As Tabelas 5.3 e 5.4 detalham os resultados obtidos após a integração do aumento de dados, variando a estratégia de compensação do desbalanceamento. A primeira apresenta o desempenho utilizando pesos estritamente balanceados, enquanto a segunda sintetiza os indicadores sob uma abordagem de balanceamento suavizado, visando uma distribuição de gradientes menos agressiva durante o treinamento.

Os dois experimentos conclusivos desta etapa utilizaram a técnica de aumento de dados associada a diferentes regimes de pesos de classe. O primeiro, utilizando pesos estritamente balanceados, obteve uma perda de 1,1342 e acurácia de 50,35%, conforme detalhado na Tabela 5.3.

O quarto modelo apresentou o desempenho mais consistente ao combinar o aumento de dados com o regime de pesos suavizados. Sob esta configuração, observou-se uma redução na função de perda para 1,0812 e um incremento na acurácia para 52,65% (Tabela 5.4). A dinâmica de treinamento (Figura 5.1) revela uma convergência progressiva da acurácia até a 8ª época, seguida de sobreajuste a partir da 10ª época. Em con-

Tabela 5.3: Desempenho do modelo MobileNetV2 (Full Frame) no conjunto de teste com *Data Augmentation* e com *Class Weights* balanceada

Classe	*Precision*	*Recall*	*F1-score*	*Support*
Healthy	0.54	0.62	0.58	152
Benign	0.56	0.36	0.44	158
OPMD	0.59	0.55	0.57	234
OCA	0.08	0.27	0.13	22
Accuracy		0.5035		
Loss		1.1342		
Macro Avg	0.44	0.45	0.43	566
Weighted Avg	0.55	0.50	0.52	566

Fonte: Autora (2026)

trapartida, a curva de perda manteve sua trajetória descendente até aproximadamente *ReduceLROnPlateau*, que, ao reduzir a taxa de aprendizado após a estabilização da métrica principal, permitiu um refinamento contínuo dos pesos. Nota-se que a l*loss* de validação manteve-se inferior à de treinamento, podendo ocorrer devido ao fato do aumento de dados tornar o conjunto de treino propositalmente mais complexo que o de validação. Por fim, a Figura 5.2 apresenta as matrizes de confusão, detalhando a capacidade discriminatória, mesmo que limitada, da rede entre as quatro classes.

Tabela 5.4: Desempenho do modelo MobileNetV2 (Full Frame) no conjunto de teste com *Data Augmentation* e *Class Weights* suavizada

Classe	*Precision*	*Recall*	*F1-score*	*Support*
Healthy	0.58	0.54	0.56	152
Benign	0.59	0.37	0.45	158
OPMD	0.54	0.66	0.60	234
OCA	0.08	0.14	0.10	22
Accuracy		0.5265		
Loss		1.0812		
Macro Avg	0.45	0.43	0.43	566
Weighted Avg	0.55	0.53	0.53	566

Fonte: Autora (2026)

Em síntese, os modelos com os desempenhos mais consistentes foram com a configuração dos pesos suavizados, com acurácias de 52,47% (sem *data augmentation*) e 52,65% (com *data augmentation*), apresentando *loss* de 1,0711 e 1,0812, respectivamente. A Tabela 5.5 resume os resultados obtidos nos dados de teste conforme a configuração de cada modelo.

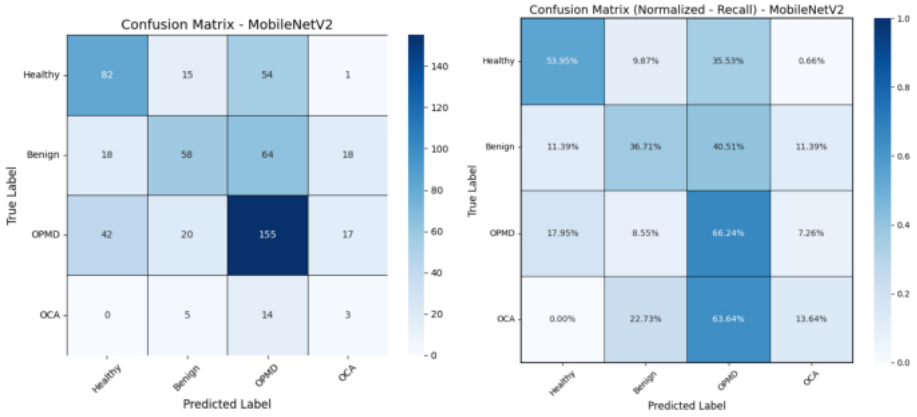
Embora os resultados globais reflitam a complexidade inerente à classificação de

Figura 5.1: Curvas ROC e de Aprendizado - MobileNetV2 - Full Frame



Fonte: Autora (2026).

Figura 5.2: Matrizes de Confusão - MobileNetV2 - Full Frame



Fonte: Autora (2026)

lesões orais, a análise de explicabilidade, apresentadas na Figura 5.3, revelou que o modelo configurado com aumento de dados e balanceamento de classes suavizado (Modelo 4) demonstrou uma capacidade de generalização mais robusta.

Para este último experimento, conduzido em modo *Full Frame*, foi realizada uma análise das imagens associada aos mapas de relevância, evidenciando o potencial da rede em identificar padrões clinicamente relevantes. Para isso, foram empregadas as técnicas LIME e Grad-CAM: enquanto o LIME permite identificar as regiões da imagem que mais contribuíram para a decisão final do modelo, o Grad-CAM destaca as áreas de maior

Tabela 5.5: Desempenho do modelo MobileNetV2 no conjunto de teste sob diferentes configurações de treinamento

Configuração	*Loss*	*Acc*	*Recall por classe*			
			Healthy	Benign	OPMD	OCA
Sem DA + CW Balanced	1.1338	50.35%	0.62	0.36	0.55	0.27
Sem DA + CW Suavizada	1.0711	52.47%	0.57	0.32	0.67	0.14
Com DA + CW Balanced	1.1342	50.35%	0.62	0.36	0.55	0.27
Com DA + CW Suavizada	1.0812	52.65%	0.54	0.37	0.66	0.14

DA = *Data Augmentation*; CW = *Class Weights*.

Fonte: Autora (2026)

ativação nas camadas convolucionais profundas, associadas ao processo de classificação.

Figura 5.3: Exemplos com LIME e Grad-CAM - MobileNetV2 - Full Frame



Fonte: Autora (2026).

5.2 Modelo MobileNetV2 - Crop

Diferentemente da avaliação em modo *Full Frame*, na qual foram exploradas quatro combinações de estratégias de treinamento, os experimentos realizados com o recorte da lesão (*crop*) foram conduzidos com apenas duas configurações:

I. Treinamento com *class weights balanced*, sem aplicação de *data augmentation*.

II. Treinamento com *class weights smoothing*, associado ao uso de *data augmentation*.

Essa escolha foi motivada pelos resultados previamente observados no cenário *Full Frame*, no qual a combinação entre *data augmentation* e *class weights smoothing* apresentou desempenho superior. Dessa forma, buscou-se otimizar o tempo computacional e direcionar a análise comparativa entre configurações com e sem *data augmentation*, bem como entre pesos balanceados e suavizados.

No primeiro experimento (I) utilizando as lesões segmentadas (ROI), os indicadores de desempenho apresentaram uma evolução significativa em comparação ao modelo *Full Frame*. Conforme detalhado na Tabela 5.6, a acurácia atingiu 74,20%, acompanhada por uma redução expressiva na função de perda, que se estabilizou em 0,6052. As Figuras 5.4 e 5.5 apresentam, respectivamente, as curvas ROC, de aprendizado e as matrizes de confusão do modelo MobileNetV2 - Crop I.

Tabela 5.6: Desempenho do modelo MobileNetV2 Crop no conjunto de teste - Estratégia I

Classe	*Precision*	*Recall*	*F1-score*	*Support*
Healthy	0.91	0.95	0.93	152
Benign	0.70	0.65	0.67	158
OPMD	0.86	0.68	0.76	234
OCA	0.19	0.68	0.30	22
Accuracy		0.7420		
Macro Avg	0.67	0.74	0.67	566
Weighted Avg	0.80	0.74	0.76	566

Fonte: Autora (2026).

Figura 5.4: Curvas ROC e de aprendizado - MobileNetV2 - Crop I

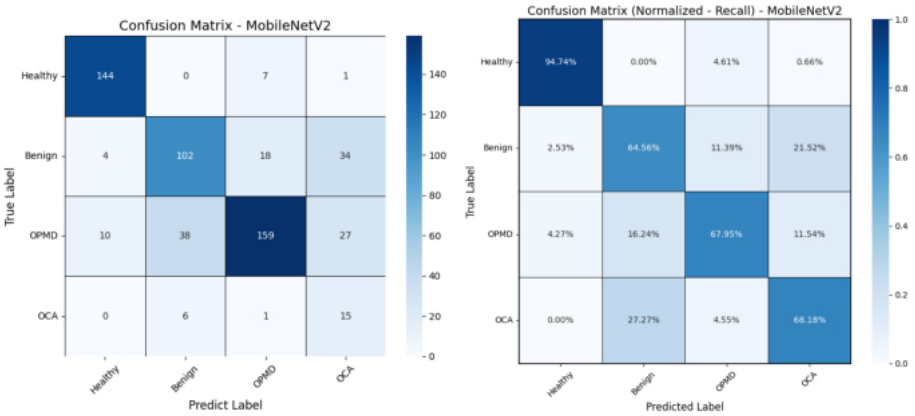


Fonte: Autora (2026).

A implementação da estratégia (II), que integrou o aumento de dados ao balanceamento de classes suavizado, resultou no melhor desempenho deste conjunto de experimentos. O modelo alcançou uma acurácia de 77,39% e uma redução da perda para 0,5283. A Tabela 5.7 apresenta as métricas detalhadas deste modelo. Na Figura 5.6 estão as curvas ROC e de aprendizado e na Figura 5.7 as matrizes de confusão do experimento.

A interpretação visual de predições selecionadas, analisadas por meio da técnica LIME, encontra-se detalhada na Figura 5.8. Pode-se obervar que o modelo apresenta um

Figura 5.5: Matrizes de Confusão - MobileNetV2 - Crop I



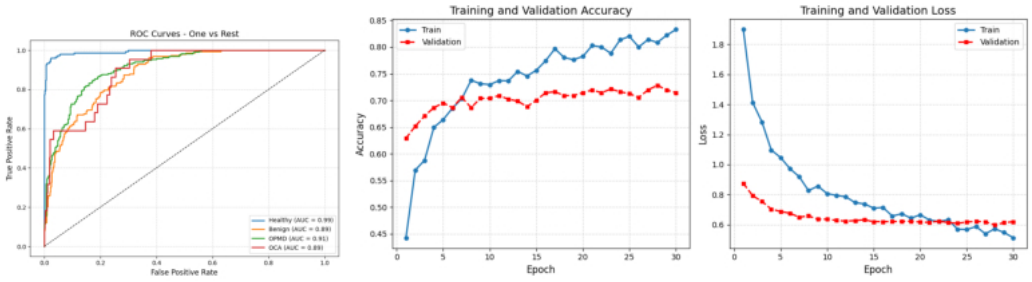
Fonte: Autora (2026).

Tabela 5.7: Desempenho do modelo MobileNetV2 Crop no conjunto de teste - Estratégia II

Classe	*Precision*	*Recall*	*F1-score*	*Support*
Healthy	0.95	0.94	0.94	152
Benign	0.62	0.76	0.68	158
OPMD	0.82	0.69	0.75	234
OCA	0.52	0.59	0.55	22
Accuracy		0.7739		
Macro Avg	0.73	0.75	0.73	566
Weighted Avg	0.79	0.77	0.78	566

Fonte: Autora (2026).

Figura 5.6: Curvas ROC e de aprendizado - MobileNetV2 - Crop II

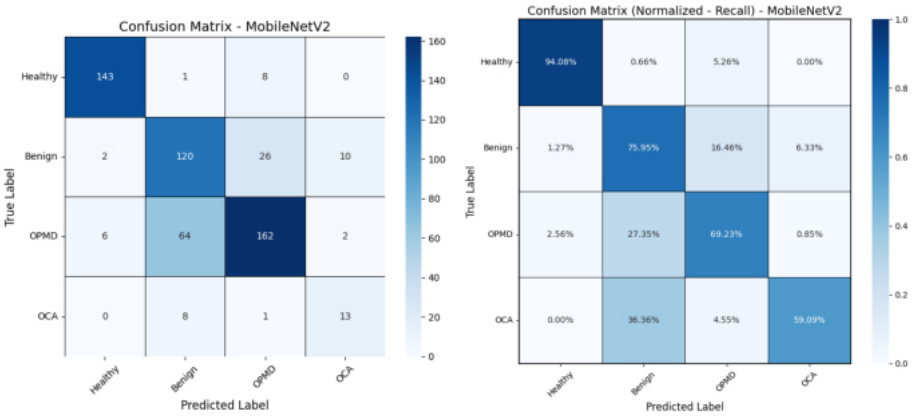


Fonte: Autora (2026).

elevado grau de confiança (superior a 90%) em diversas amostras de lesões, sendo que as regiões destacadas pela técnica de explicabilidade convergem precisamente com as características morfológicas de interesse clínico. Por outro lado, a figura também ilustra um caso de erro condizente com as métricas da matriz de confusão: uma amostra da classe OPMD foi incorretamente classificada como **Benign** com 79,26% de confiança. Esse tipo de ocorrência evidencia os desafios impostos pela similaridade visual entre certas condições patológicas e lesões benignas.

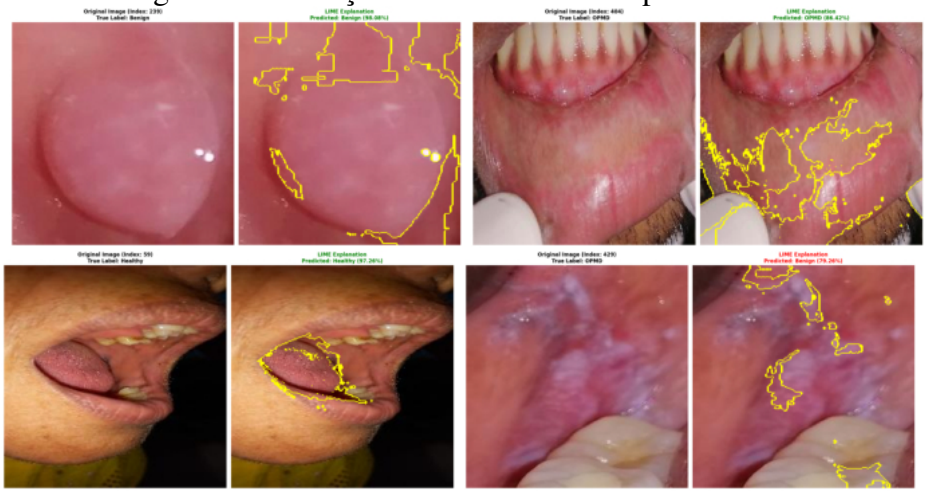
Os experimentos realizados com a arquitetura de referência (MobileNetV2) per-

Figura 5.7: Matrizes de Confusão - MobileNetV2 - Crop II



Fonte: Autora (2026).

Figura 5.8: Predições do modelo 2 - Crop com LIME



Fonte: Autora (2026).

mitiram consolidar a configuração de treinamento mais robusta para este domínio. Verificou-se que a combinação de aumento de dados e balanceamento de classes suavizado minimiza o impacto do desequilíbrio do *dataset*, independentemente da técnica de entrada utilizada.

Dessa forma, para os experimentos conduzidos com as arquiteturas MobileNetV3 Small, EfficientNetB0 e ConvNeXt V2 Tiny, adotou-se exclusivamente essa configuração otimizada. Tal padronização foi aplicada tanto para as avaliações em modo *Full Frame* quanto para os testes com imagens segmentadas (*Crop*), permitindo uma comparação entre as arquiteturas.

5.3 Avaliação das arquiteturas MobileNetV3 Small, EficientNetB0 e ConvNeXt Tiny - Entrada Full Frame

Os resultados consolidados na Tabela 5.8 revelam um desempenho equilibrado, porém limitado, entre as três arquiteturas testadas em modo *Full Frame*. Observa-se que a acurácia oscilou em uma faixa estreita, entre 52,30% (MobileNetV3 Small) e 54,42% (EfficientNetB0), com valores de *loss* superiores a 1,04 em todos os casos.

Embora a ConvNeXt Tiny tenha apresentado métricas de Macro F1-score (45,02%) e Macro Precision (46,41%) ligeiramente superiores, a proximidade dos resultados indica que a presença de elementos irrelevantes nas imagens completas (ruído de fundo) impôs uma barreira de aprendizado para todas as arquiteturas. Essa estagnação das métricas em patamares próximos a 54% confirma a necessidade da estratégia de recorte da Região de Interesse e da posterior integração de dados clínicos, visando elevar a sensibilidade e a

especificidade do diagnóstico assistido.

Tabela 5.8: Comparação de desempenho entre modelos Full Frame para classificação multiclasse no conjunto de teste

Métrica	*MobileNetV3 Small*	*EfficientNetB0*	*ConvNeXt Tiny*
Accuracy	0.5230	**0.5442**	0.5424
Loss	1.0465	**1.0430**	1.0504
Macro Precision	0.4047	0.4338	**0.4641**
Macro Recall	0.3906	0.4247	**0.4442**
Macro F1-score	0.3871	0.4279	**0.4502**
Weighted Precision	0.5160	**0.5435**	0.5362
Weighted Recall	0.5230	**0.5442**	0.5424
Weighted F1-score	0.5061	**0.5421**	0.5377

Fonte: Autora (2026).

5.4 Avaliação das arquiteturas MobileNetV3 Small, EficientNetB0 e ConvNeXt Tiny - Entrada Crop da Lesão

Os resultados obtidos com a estratégia de recorte da ROI, apresentados na Tabela 5.9, demonstram uma evolução substancial em todas as métricas de desempenho. Diferente do cenário *Full Frame*, a remoção de ruídos nas imagens permitiu que as redes extraíssem características mais discriminantes das lesões.

A arquitetura MobileNetV3 Small apresentou o desempenho superior, atingindo uma acurácia de 84,81% e o menor índice de perda (0,4177). O modelo se mostrou mais equilibrado, com um *recall* de 77,07%, F1-score de 77,89% e uma especificidade macro de 94,27%, indicando uma capacidade de distinguir corretamente entre as diferentes classes.

Embora as arquiteturas EfficientNetB0 e ConvNeXt V2 Tiny também tenham apresentado resultados interessantes (acurácias de 80,04% e 79,86%, respectivamente), a MobileNetV3 Small superou-as em todas as métricas apuradas. O desempenho robusto, aliado à sua eficiência computacional, indicou a rede como sendo a melhor escolha para a integração com os dados clínicos no modelo multimodal.

A análise da matriz de confusão da MobileNetV3 Small (Figura 5.9) indica a alta capacidade do modelo em distinguir tecidos saudáveis de lesões, com 143 acertos na classe *Healthy*. O modelo demonstrou também robustez na identificação de OPMD, classificando corretamente 205 instâncias, o que é importante para o rastreio precoce de

Tabela 5.9: Comparação de desempenho entre modelos Crop para classificação multiclasse no conjunto de teste

Métrica	*MobileNetV3 Small*	*EfficientNetB0*	*ConvNeXt Tiny*
Accuracy	**0.8481**	0.8004	0.7986
Loss	**0.4177**	0.4832	0.5228
Macro Precision	**0.7884**	0.7261	0.7480
Macro Specificity	**0.9427**	0.9258	0.9245
Macro Recall	**0.7707**	0.7001	0.7088
Macro F1-score	**0.7789**	0.7100	0.7225
Weighted Precision	**0.8487**	0.8029	0.7947
Weighted Recall	**0.8481**	0.8004	0.7986
Weighted F1-score	**0.8479**	0.8004	0.7955

Fonte: Autora (2026).

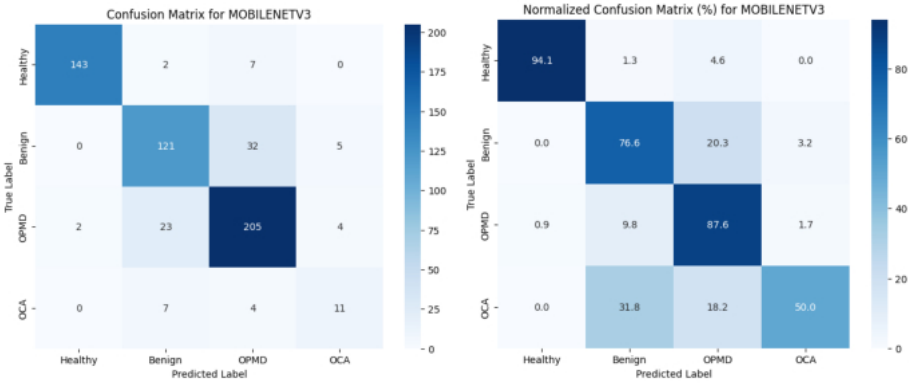
patologias orais potencialmente malignas.

Observa-se que a maior parte das classificações incorretas ocorre entre as classes *Benign* e OPMD (32 amostras benignas classificadas como OPMD e 23 OPMDs classificadas como benignas). Esse fenômeno é justificável pela intersecção morfológica e similaridade visual de certas características clínicas dessas lesões. No entanto, o baixo índice de confusão entre tecidos saudáveis (*Healthy*) e câncer (OCA) reforça o desempenho promissor do modelo proposto.

A análise conjunta das matrizes de confusão absoluta e normalizada (Figura 5.9) permite uma compreensão detalhada da eficácia do classificador. A matriz normalizada destaca a alta taxa de acerto para as classes *Healthy*(94,1%) e OPMD (87,6%). Particularmente para a classe OCA, embora o número absoluto de amostras seja reduzido, a matriz normalizada revela que o modelo obteve uma taxa de acerto de 50% na classe minoritária. Embora este valor seja bem inferior às demais classes, ele reflete as limitações impostas pelo reduzido número de amostras positivas de carcinoma no *dataset* original. É importante notar que, nos casos de erro, o modelo teve por tendência a classificação do carcinoma como OPMD ou *Benign*, evidenciando a possível complexidade de distinção em estágios iniciais. Este resultado justifica a etapa subsequente de integração de dados clínicos, uma vez que fatores como histórico de tabagismo e idade são variáveis decisivas que podem auxiliar a rede a diferenciar casos malignos de lesões benignas morfologicamente semelhantes.

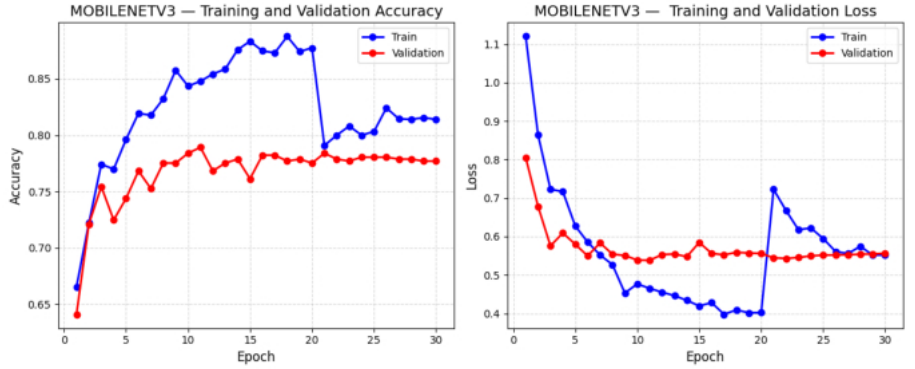
As curvas de aprendizado (Figura 5.10) da MobileNetV3 Small revelam um cenário de *overfitting*, caracterizado pela divergência entre as métricas de treino e validação após a época 7. No entanto, apesar dessa especialização excessiva nos dados de treina-

Figura 5.9: Matrizes de Confusão do Modelo MobilenetV3 - Crop



Fonte: Autora (2026).

Figura 5.10: Curvas de Treinamento e Validação - MobileNetV3 - Crop



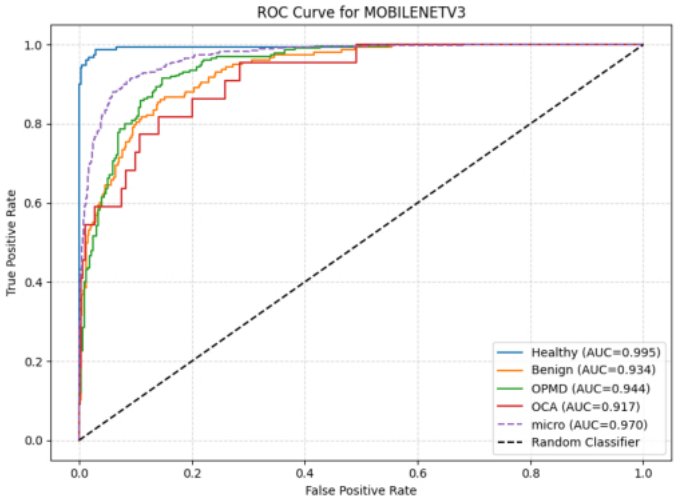
Fonte: Autora (2026).

mento, o modelo demonstrou uma alta capacidade de generalização, refletida no desempenho robusto alcançado no conjunto de teste. Esse comportamento indica que, mesmo com a memorização de certas nuances do *dataset*, a rede foi capaz de extrair padrões clínicos relevantes para a classificação das patologias orais.

A Curva ROC (Figura 5.11) corrobora essa eficácia, apresentando uma área sob a curva (AUC) elevada (entre 0,91 a 0,99). Indicando que a rede possui uma elevada capacidade de distinguir entre as diferentes patologias, mesmo em cenários de desbalanceamento. Para a classe OCA, o valor de AUC de 0,91 demonstra que, apesar da taxa de acerto direta de 50% observada na matriz de confusão, o modelo mantém uma ordenação probabilística eficiente na identificação de casos de carcinoma.

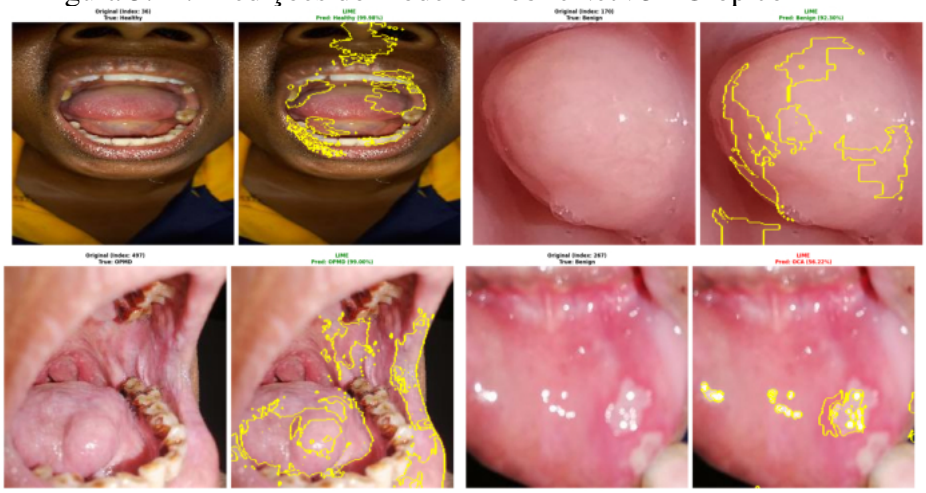
Para compreender os critérios de decisão da MobileNetV3 Small, aplicou-se a técnica de explicabilidade LIME 5.12. Nos casos de acerto, o modelo demonstrou alta confiabilidade, apresentando predições para as classes *Healthy* (99,98%) e OPMD (99%), e 92,30% para a classe *Benign*. A análise visual dos *superpixels* destacados pelo LIME confirmou que a rede está com a atenção em regiões de lesões que são clinicamente relevantes.

Figura 5.11: Curva ROC do Modelo MobileNetV3 - Crop



Fonte: Autora (2026).

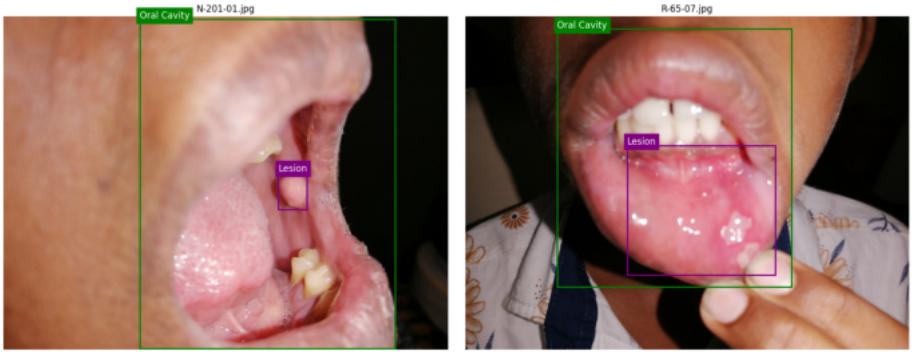
Figura 5.12: Predições do modelo MobileNetV3 - Crop com LIME



Fonte: Autora (2026).

Por outro lado, a análise de um erro de classificação, onde uma lesão*Benign* foi erroneamente rotulada como OCA, revelou uma confiança de apenas 56,22%. Este patamar próximo à zona de incerteza (limiar de decisão) pode ser um indicador positivo, pois demonstra que, embora o modelo tenha falhado, ele não o fez com alta convicção. Esse tipo de erro com menos confiança reforça a importância de um sistema de apoio à decisão, onde casos de baixa confiança sejam sinalizados para revisão profissional ou confrontados com os dados clínicos do paciente. A Figura 5.13 mostra as imagens originais anotadas por especialista do recorte de lesões da classe *Benign* (92,30%) e da lesão *Benign* classificado erroneamente como OCA (56,22%).

Figura 5.13: Imagens originais completas com as anotações COCO.



Fonte: Autora (2026).

5.5 Modelo Multimodal

A integração de dados multimodais apresentou um desempenho superior em comparação à análise isolada de imagens. O modelo baseado exclusivamente em características visuais (IMG Only) atingiu uma acurácia de 84,72% e uma função de perda de 0,3856. Com a inclusão dos dados clínicos (variáveis tabulares) através de uma arquitetura de fusão tardia (*late fusion*), foi observada uma melhora significativa das métricas em todas as classes (Tabela 5.10). A acurácia final foi de **88,52%** (um ganho absoluto de 3,8%) e a *Loss* Final foi de **0,3472**.

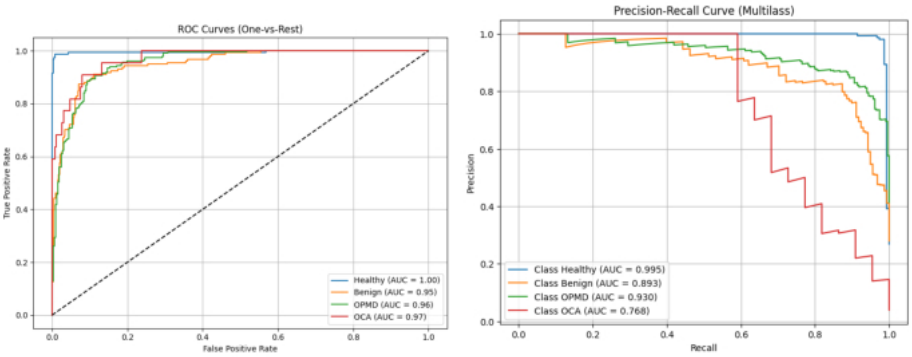
As métricas globais obtidas durante as etapas de treinamento, validação e teste estão consolidadas na Tabela 5.10. Adicionalmente, o desempenho detalhado por classe para o conjunto de teste é apresentado na Tabela 5.11. A Figura 5.14 ilustra as curvas ROC e a PR (*Precision-Recall*), enquanto a Figura 5.15 exibe as curvas de aprendizado do modelo multimodal. Por fim, as matrizes de confusão geradas a partir da avaliação do conjunto de teste podem ser visualizadas na Figura 5.16.

Tabela 5.10: Desempenho global no conjunto de teste do modelo multimodal

Métrica	**Valor**
Test Accuracy (IMG Only)	0.8472
Test Loss (IMG Only)	0.3856
Test Accuracy (Multimodal)	0.8852
Test Loss (Multimodal)	0.3472
Precision (macro)	0.9019
Recall (macro)	0.8329
F1-score (macro)	0.8601
Training Accuracy	0.9677
Validation Accuracy	0.8031
Training Loss	0.1065
Validation Loss	0.4820

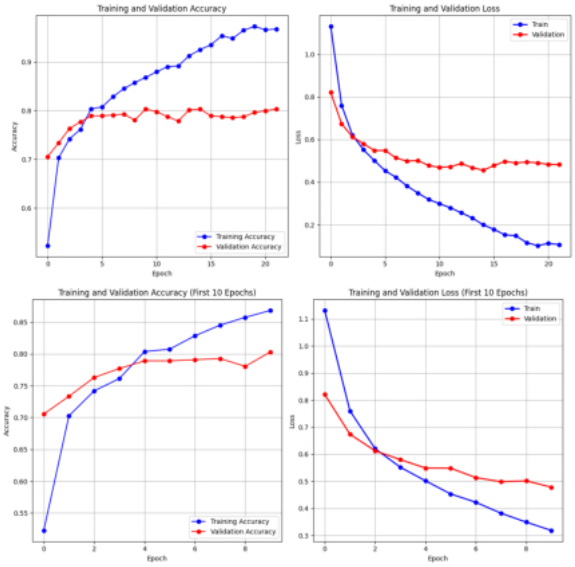
Fonte: Autora (2026).

Figura 5.14: Curvas ROC e PR do Modelo Multimodal



Fonte: Autora (2026).

Figura 5.15: Curvas de Aprendizagem completas e um recorte até a 10ª época do Modelo Multimodal



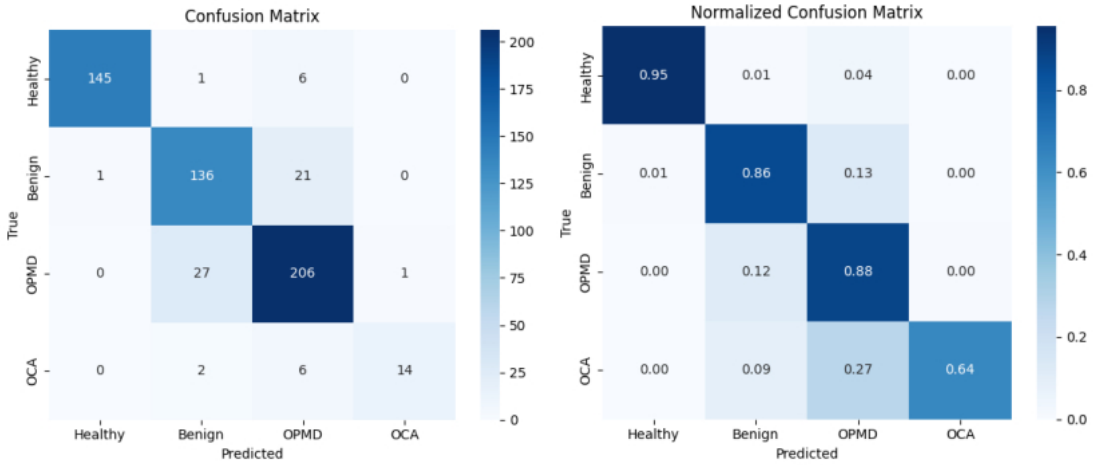
Fonte: Autora (2026).

5.6 Análise de Explicabilidade Multimodal: Interpretação Local via LIME (Visual e Tabular) e Grad-CAM

A explicabilidade do modelo foi analisada sob múltiplas perspectivas para garantir a transparência do processo de decisão. Primeiramente, aplicou-se o método Grad-CAM para validar a atenção da rede neural em regiões morfologicamente relevantes. Complementarmente, utilizou-se o LIME tanto para o domínio visual quanto para o tabular; enquanto o primeiro destaca *superpixels* determinantes na imagem, o segundo permite quantificar o impacto individual das variáveis clínicas nas predições. Essa abordagem integrada permitiu confirmar que o ganho de desempenho decorre de uma fusão coerente entre padrões visuais e o contexto clínico do paciente.

Os resultados dessas análises para casos representativos de cada classe estão con-

Figura 5.16: Matrizes de Confusão do Modelo Multimodal



Fonte: Autora (2026).

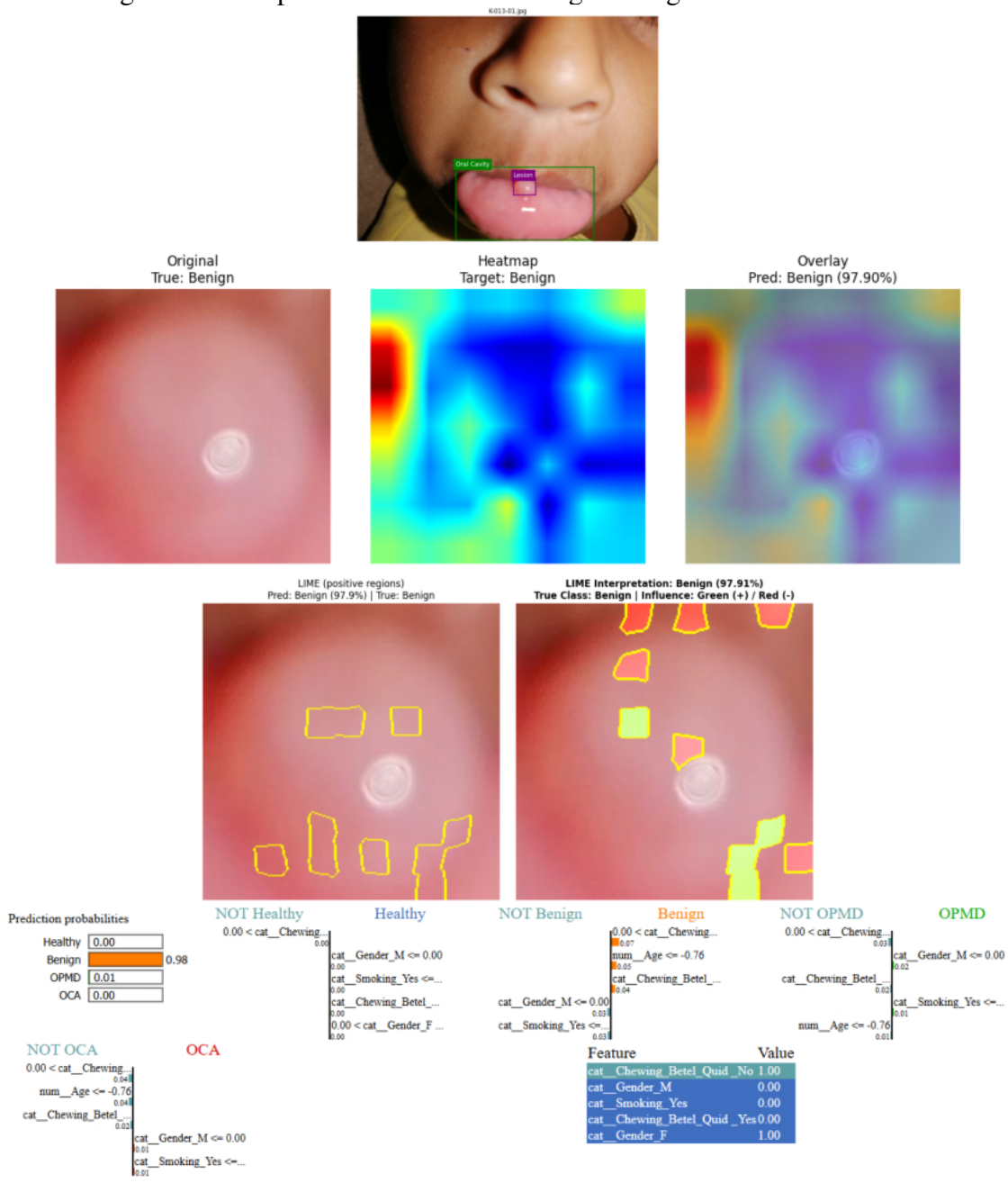
Tabela 5.11: Desempenho por classe no conjunto de teste — Modelo Multimodal

Classe	**Precision**	**Recall**	**F1-score**	**Specificity**	**Support**	**AUC**
Healthy	0.99	0.95	0.97	0.99	152	0.9957
Benign	0.82	0.86	0.84	0.93	158	0.9513
OPMD	0.86	0.88	0.87	0.90	234	0.9555
OCA	0.93	0.64	0.76	0.99	22	0.9706
Macro Avg	0.90	0.83	0.86	0.96	566	0.9683
Weighted Avg	0.89	0.89	0.89	0.94	566	–

Fonte: Autora (2026).

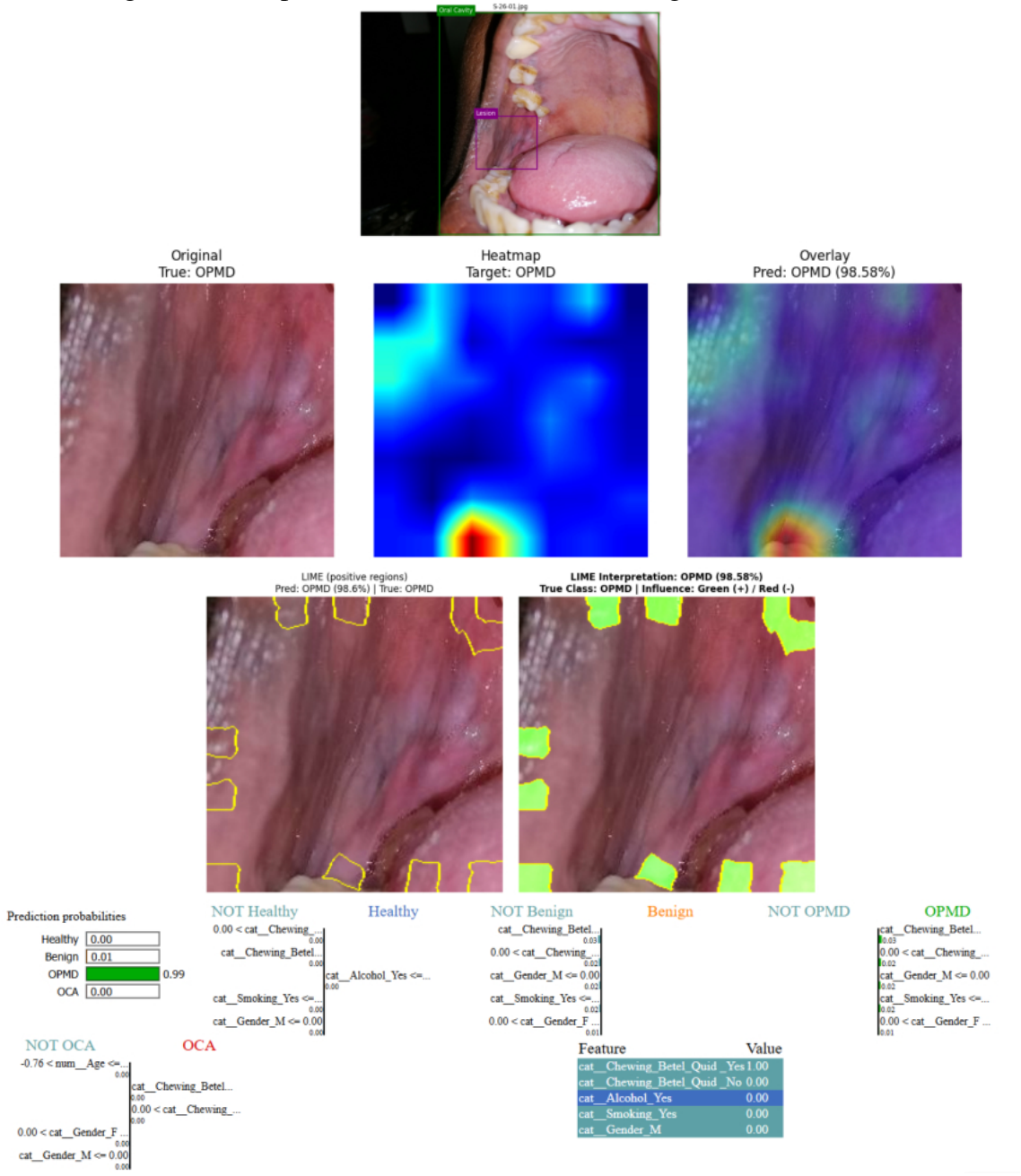
solidados nas Figuras 5.17, 5.18 e 5.19. Em cada conjunto, o painel superior apresenta a imagem original com as respectivas anotações; a segunda linha exibe o mapa de calor Grad-CAM; a terceira linha demonstra a segmentação de *superpixels* via LIME Imagem; e, por fim, a seção inferior quantifica a importância das variáveis clínicas por meio do LIME Tabular. Adicionalmente, a Figura 5.20 detalha uma investigação sobre um erro de predição frequente do modelo (OPMD por *Benign*).

Figura 5.17: Explicabilidade Classe Benign - Imagem e Dados Clínicos



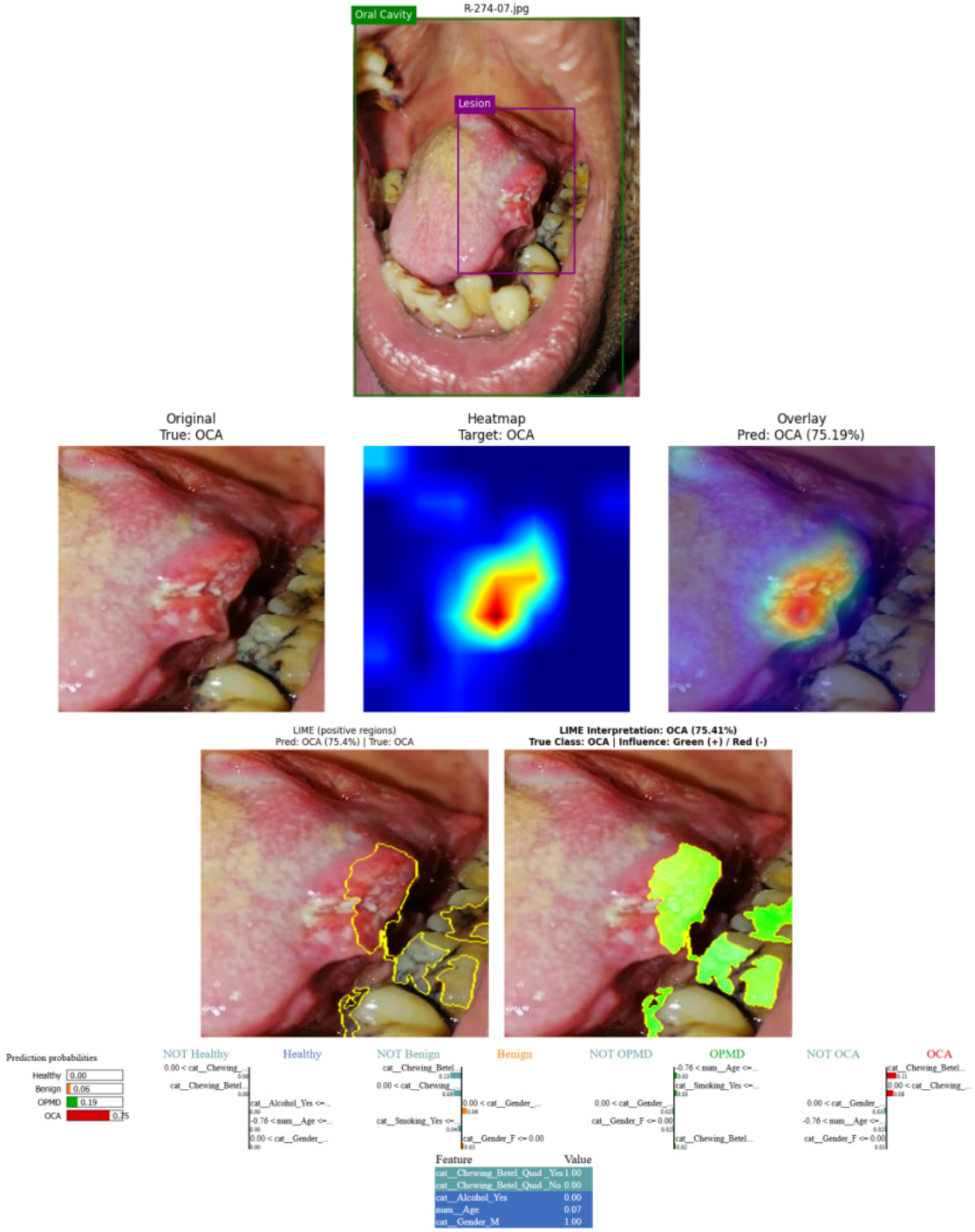
Fonte: Autora (2026).

Figura 5.18: Explicabilidade Classe OPMD - Imagem e Dados Clínicos



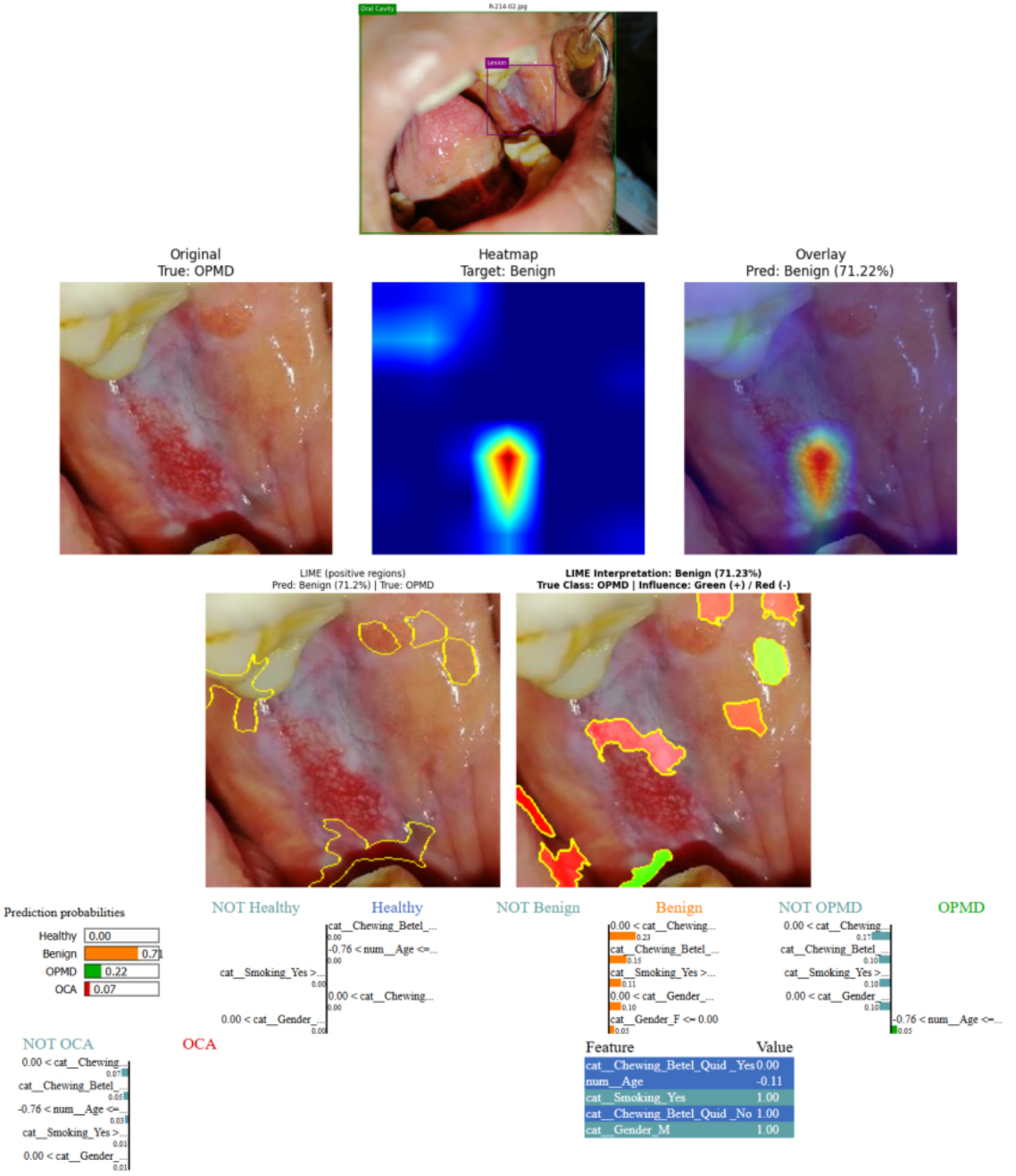
Fonte: Autora (2026).

Figura 5.19: Explicabilidade Classe OCA - Imagem e Dados Clínicos



Fonte: Autora (2026).

Figura 5.20: Explicabilidade de um erro de predição - Imagem e Dados Clínicos



Fonte: Autora (2026).

6 DISCUSSÃO

Este capítulo tem como objetivo analisar criticamente os resultados apresentados no Capítulo 5, considerando os métodos adotados e as métricas empregadas.

Na Seção 6.1, a discussão concentra-se na interpretação dos principais achados experimentais, nos padrões de desempenho observados entre as diferentes arquiteturas avaliadas e no impacto das estratégias adotadas para lidar com o desbalanceamento de classes.

Na Seção 6.2, os resultados são comparados aos de estudos anteriores apresentados no Capítulo 3 (Estado da Arte), especificamente aqueles que utilizaram o mesmo conjunto de dados para tarefas de classificação, permitindo uma análise mais direta e consistente do desempenho do modelo proposto.

Por fim, a Seção 6.3 aborda uma perspectiva de integração dos resultados com a prática clínica.

6.1 Análise dos resultados experimentais

Os resultados preliminares com o modelo MobileNetV2 utilizando imagens completas (*Full Frame*) revelaram uma acurácia limitada em torno de 50-52%, evidenciando a baixa capacidade de generalização diante do desbalanceamento original dos dados. A utilização de pesos de classe estritamente equilibrados penalizava excessivamente classes importantes, como OPMD e *Benign*, que possuíam maior densidade de amostras.

A adoção de uma estratégia combinada de *data augmentation* e pesos de classe ponderados, suavizados pela raiz quadrada (sqrt), foi essencial para mitigar esse viés e elevar a performance preditiva. O balaceamento de classes suavizado, mesmo sem aumento de dados, já demonstrou uma melhora discreta na acurácia para 52,47% e uma redução na perda para 1,0711. A combinação do aumento de dados com pesos suavizados resultou no desempenho mais consistente para o cenário *Full Frame*, atingindo 52,65% de acurácia e perda de 1,0812.

A análise das curvas de aprendizado do MobileNetV2 (Full Frame) com aumento de dados e pesos suavizados indicou uma convergência progressiva da acurácia até a 8ª época, seguida de *overfitting* a partir da 10ª época. Contudo, a curva de perda manteve sua trajetória descendente até a época 25, atribuída ao uso do *ReduceLROnPlateau*, que permitiu um refinamento contínuo dos pesos. Curiosamente, a perda de validação se

manteve inferior à de treinamento, sugerindo que o aumento de dados tornou o conjunto de treino propositalmente mais complexo que o de validação.

As matrizes de confusão iniciais (Figura 5.2) detalham a capacidade discriminatória da rede, apontando para a complexidade inerente à classificação de lesões orais. As técnicas de explicabilidade, como LIME e Grad-CAM (Figura 5.3), já demonstravam o potencial da rede em identificar padrões clinicamente pertinentes, focando em regiões da imagem que mais contribuíam para a classificação final.

Os experimentos iniciais com *Full Frame*, incluindo MobileNetV3 Small, EfficientNetB0 e ConvNeXt Tiny, mostraram desempenho limitado e equilibrado, com acurácia oscilando entre 52,30% e 54,42%. Isso confirmou que a presença de elementos irrelevantes (ruído de fundo) nas imagens completas impôs uma barreira de aprendizado para todas as arquiteturas, ressaltando a necessidade de uma estratégia de recorte da Região de Interesse.

A adoção da estratégia de recorte da lesão resultou em uma evolução substancial em todas as métricas de desempenho. A remoção de ruídos permitiu que as redes extraíssem características mais discriminantes das lesões. A arquitetura MobileNetV3 Small se destacou neste cenário, alcançando uma acurácia de 84,81% e a menor perda (0,4177), superando as outras arquiteturas em todas as métricas apuradas. Este modelo demonstrou-se mais equilibrado, com um *recall* de 77,07%, F1-score de 77,89% e uma especificidade macro de 94,27%, indicando uma elevada capacidade de distinção entre as classes.

A matriz de confusão da MobileNetV3 Small (Figura 5.9) revelou alta capacidade de distinguir tecidos saudáveis (94,10% de acerto) e uma robustez na identificação de OPMD (87,60% de acerto), essencial para o rastreio precoce de patologias potencialmente malignas. A principal área de confusão ocorreu entre as classes *Benign* e OPMD, com 32 amostras benignas classificadas como OPMD e 23 OPMDs como benignas. Esse fenômeno é justificável pela intersecção morfológica e similaridade visual entre certas características clínicas dessas lesões. O baixo índice de confusão entre *Healthy* e Câncer (OCA) reforça o desempenho promissor do modelo. Para a classe OCA, apesar do número reduzido de amostras e da taxa de acerto direta de 50%, o modelo apresentou um AUC de 0,91, indicando uma ordenação probabilística eficiente na identificação de casos de carcinoma, com tendência a classificar erros como OPMD ou *Benign*.

Apesar do cenário de *overfitting* nas curvas de aprendizado da MobileNetV3 Small (Crop) após a época 7, o modelo demonstrou uma alta capacidade de generalização no conjunto de teste. As Curvas ROC com AUCs entre 0,91 e 0,99 corroboram essa eficácia.

A interpretabilidade via LIME confirmou a alta confiabilidade das predições, destacando os *superpixels* relevantes nas imagens.

A integração de dados multimodais, combinando características visuais com as variáveis clínicas tabulares (fusão tardia), representou um salto significativo no desempenho. O modelo exclusivamente baseado em imagens (IMG Only) alcançou 84,72% de acurácia, enquanto o modelo multimodal elevou a acurácia para 88,52% (um ganho absoluto de 3,8%) e reduziu a perda para 0,3472.

As métricas detalhadas por classe para o modelo multimodal (Tabela 5.11) demonstraram melhorias em todas as categorias. Para a classe OCA, que era particularmente desafiadora devido ao seu baixo número de amostras, o modelo multimodal alcançou uma precisão de 0,93, *recall* de 0,64, F1-score de 0,76 e uma especificidade de 0,99, com AUC de 0,9706. Isso evidencia que a inclusão de dados clínicos permitiu ao modelo superar as limitações da classificação puramente visual, especialmente em casos complexos e classes minoritárias onde fatores como histórico de tabagismo e idade são decisivos.

A análise de explicabilidade multimodal, utilizando Grad-CAM para validar a atenção da rede a regiões morfologicamente relevantes e LIME (visual e tabular) para quantificar o impacto de *superpixels* e variáveis clínicas, confirmou que o ganho de desempenho decorre de uma fusão coerente entre padrões visuais e o contexto clínico do paciente. Essa abordagem integrada garante a transparência do processo de decisão, evidenciando como a rede combina informações visuais e clínicas para chegar às suas predições (Figuras 5.17 a 5.19. A investigação de erros comuns, como OPMD por *Benign* (Figura 5.20), demonstra o potencial de entender as falhas do modelo, o que é importante para refinamentos futuros e aumento da confiança clínica.

Essa abordagem integrada, combinando Grad-CAM para atenção visual, LIME imagem e tabular, permitiu confirmar que o ganho de desempenho do modelo multimodal decorre de uma fusão coerente e interpretável entre os padrões visuais e o contexto clínico do paciente. Por exemplo, em casos de dúvida morfológica (como entre *Benign* e OPMD, ou OPMD e OCA), as variáveis clínicas podem fornecer o peso decisivo para uma classificação mais precisa.

6.2 Comparação do desempenho do modelo multimodal com os artigos do Estado da Arte

Os resultados obtidos neste estudo, particularmente mediante a abordagem multimodal e a utilização de imagens com recorte da Região de Interesse (ROI), demonstram um avanço expressivo em relação aos modelos de base *Full Frame*. Embora alguns estudos relacionados apresentem métricas superiores, a análise comparativa revela nuances metodológicas fundamentais, conforme discutido a seguir.

Enquanto modelos treinados neste estudo com imagens completas apresentaram acurácia limitada (50-54%), a estratégia de segmentação da ROI elevou o desempenho do MobileNetV2 para 77,39% e do MobileNetV3 Small para 84,81%. Este fenômeno corrobora a literatura, indicando que o ruído de fundo em fotografias intraorais é um ponto crítico no aprendizado de máquina (SCHWENDICKE; SAMEK; KROIS, 2020). O modelo multimodal proposto, integrando dados clínicos e visuais, alcançou 88,52% de acurácia, superando pipelines multimodais recentes como os de Devindi et al. (2024) (81%) e Mukherjee et al. (2025) (64,56%).

Ao comparar os resultados deste trabalho com os de Yadav et al. (2025) (98,36%) e Amien et al. (2025) (98,21%), observa-se uma lacuna nas acurácias relatadas. Entretanto, é importante destacar uma diferença rigorosa no protocolo experimental:

- **Rigor Metodológico deste Estudo**: adotou-se a separação estrita de subconjuntos (treino, validação e teste) por paciente. Isso garante que imagens de um mesmo indivíduo não apareçam em grupos distintos, prevenindo o *data leakage* e garantindo a capacidade de generalização para novos pacientes.

- **Viés no Estado da Arte**: grande parte dos estudos que relatam acurácias próximas a 100% não explicitam a separação por paciente. Nestes casos, a rede neural pode aprender características específicas do fundo da imagem, iluminação ou textura da pele de um paciente presente em ambos os *sets*, resultando em métricas infladas que dificilmente se repetiriam em ambiente clínico real.

Assim como em Yadav et al. (2025), este trabalho utilizou Grad-CAM e LIME para conferir explicabilidade ao modelo. A melhora da sensibilidade alcançada para classes críticas como *Bening* (86%), OPMD (88%)e OCA (64%) demonstra que a inclusão de variáveis clínicas permitiu ao modelo refinar padrões morfológicos semelhantes entre OPMD e lesões benignas, um desafio amplamente reconhecido no estado da arte.

Em síntese, embora os números absolutos de acurácia deste documento pareçam inferiores a estudos que utilizam imagens sintéticas (LEE; KIM; LEE, 2025) ou módulos de atenção complexos, o desempenho de 88,52% é obtido sob um protocolo de teste mais rigoroso e realista. O foco no manejo do desbalanceamento de classes e na validação por paciente alinha este trabalho às tendências mais recentes de robustez e ética em IA na saúde, sugerindo que os modelos aqui desenvolvidos possuem maior potencial de aplicabilidade prática do que arquiteturas com métricas superiores, porém validadas sob protocolos menos restritivos.

Uma tabela comparativa completa 6.1 é apresentada para fins de comparação.

Tabela 6.1: Comparação entre o desempenho deste trabalho e o Estado da Arte

Estudo	*Ano*	*Arq./Entrada*	*Acurácia*	*Split por Paciente*	*Observação*
Yadav et al.	2025	CNN + Axial Att (*Full Frame*).	98,36%	Não	Risco elevado de *data leakage*.
Amien et al.	2025	EfficientNet B0 (*Full Frame*)	98,21%	Não	Resultados potencialmente inflados.
Lee et al.	2025	Modelo de Difusão (ROI)	97,00%	Não	Utilização de dados não reais.
Aparna et al.	2025	Híbrida (U-Net) (ROI)	95,02%	Não	Abordagem binária.
Gürses et al.	2025	Textura + SVM (ROI)	90,50%	Não	Risco de *data leakage*.
Deivindi et al.	2024	Multimodal (ROI)	81,00%	Sim	Abordagem binária e desempenho inferior.
Mukerjee et al.	2025	Multimodal (ROI)	64,56%	Sim	Baixa sensibilidade observada.
Este Trabalho	**2026**	**Multimodal (ROI)**	**88,52%**	**Sim**	**Generalização realista e robusta.**

Fonte: Elaborada pela autora (2026).

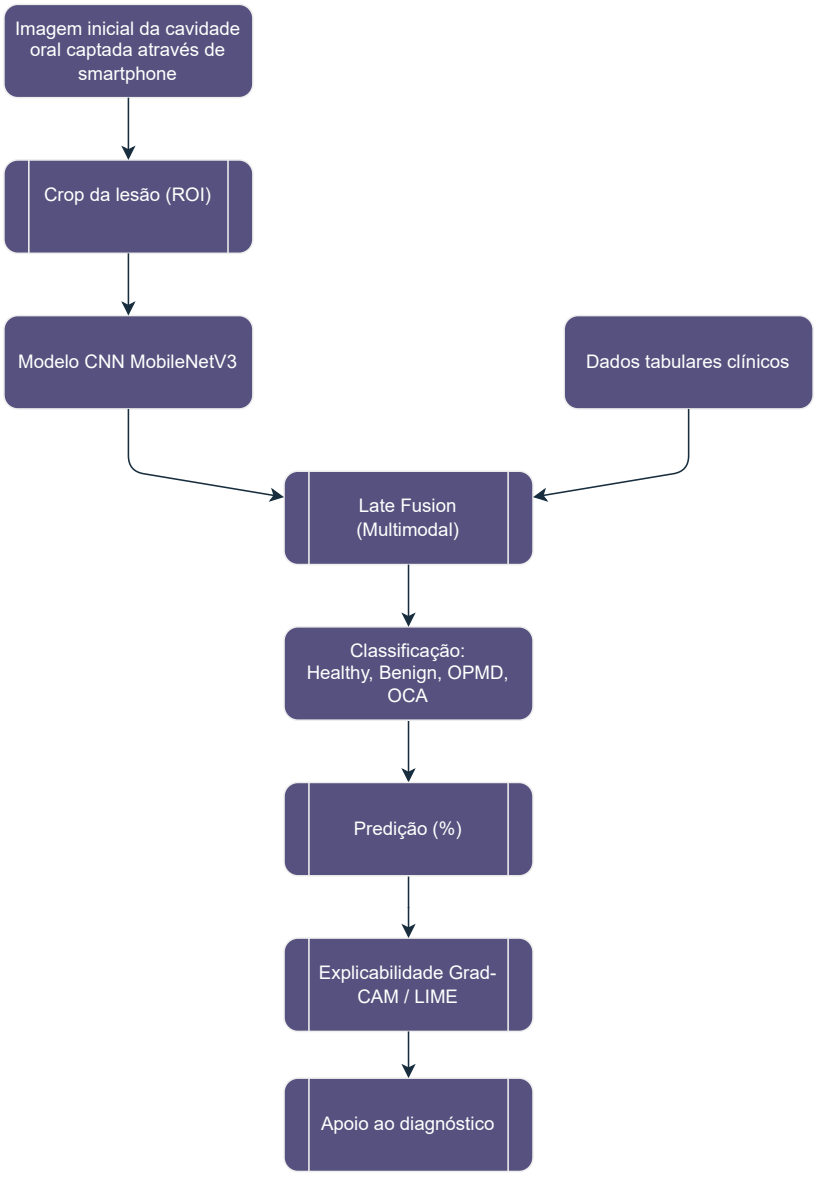
6.3 Integração do modelo multimodal com a prática clínica

A Figura 6.1 apresenta o fluxograma simplificado da solução multimodal proposta, ilustrando a integração entre a engenharia de software e os processos de ciência de dados, visando sua viabilidade e validação em ambiente clínico.

Além de auxiliar o especialista em suas decisões, a ferramenta pode ampliar o alcance do diagnóstico precoce, servindo como suporte para profissionais não especia-

listas que atuam em áreas geográficas remotas e no atendimento a populações carentes. Portanto, a solução reafirma o potencial da inteligência artificial como ferramenta de auxílio, otimizando o fluxo de encaminhamento e aumentando a precisão diagnóstica em diferentes níveis de atenção à saúde.

Figura 6.1: Arquitetura de integração e fluxo operacional da solução multimodal proposta



Fonte: Figura elaborada pela autora (2026).

7 CONCLUSÃO

Este trabalho investigou a aplicação de modelos de aprendizado profundo, associados a técnicas de Inteligência Artificial Explicável, para a classificação automática de lesões bucais a partir de imagens clínicas, considerando cenários realistas caracterizados pelo desbalanceamento de classes. A análise desenvolvida contemplou a avaliação rigorosa do desempenho dos modelos por meio de métricas robustas e clinicamente relevantes, bem como a investigação da confiabilidade das predições com base nas evidências fornecidas pelas técnicas de explicabilidade.

Os resultados obtidos demonstraram que a escolha das métricas exerce papel fundamental na interpretação do desempenho dos modelos, especialmente em contextos de desbalanceamento, nos quais métricas globais como a acurácia podem conduzir a conclusões equivocadas. Métricas como *recall* por classe, macro-F1 e *balanced accuracy* mostraram-se mais adequadas para capturar o comportamento dos modelos nas classes de maior interesse clínico. Além disso, a comparação entre diferentes arquiteturas e estratégias de mitigação evidenciou variações relevantes no desempenho, reforçando a importância de uma avaliação sistemática e padronizada.

A incorporação de técnicas de IA Explicável permitiu complementar a análise quantitativa, fornecendo subsídios para avaliar a coerência espacial das decisões dos modelos. Em diversos casos, observou-se que modelos com métricas semelhantes apresentaram padrões de atenção distintos, indicando que bons valores numéricos nem sempre refletem decisões visualmente plausíveis do ponto de vista clínico. Esse achado reforça o papel da explicabilidade como ferramenta de validação e não apenas como recurso ilustrativo.

A comparação direta com estudos anteriores apresentados nos trabalhos relacionados (Capítulo 3), que utilizaram o mesmo conjunto de dados, possibilitou uma análise metodologicamente consistente. Os resultados alcançados mostraram desempenho competitivo em relação às abordagens previamente reportadas, ao mesmo tempo em que avançam ao incorporar uma análise sistemática de métricas sob desbalanceamento e uma avaliação integrada da explicabilidade, aspectos ainda pouco explorados de forma conjunta na literatura.

Assim, conclui-se que a combinação entre modelos de aprendizado profundo, estratégias adequadas de avaliação e técnicas de IA Explicável constitui uma abordagem promissora para o apoio ao diagnóstico automático de lesões bucais. Os resultados in-

dicam que soluções desse tipo podem contribuir para sistemas de triagem clínica mais confiáveis e transparentes, sem substituir a avaliação especializada, mas atuando como ferramenta de suporte à tomada de decisão.

7.1 Limitações e trabalhos futuros

Como limitações deste estudo, destaca-se a dependência de um único conjunto de dados, bem como o tamanho relativamente reduzido da amostra, fatores que podem contribuir para o risco de *overfitting* e restringir a capacidade de generalização dos modelos avaliados. Além disso, as imagens clínicas foram obtidas em condições diversas e não controladas, resultando em elevada variabilidade quanto à iluminação, enquadramento e qualidade visual.

Embora essas características representem desafios do ponto de vista metodológico, elas também aproximam o estudo de cenários clínicos reais de aplicação, reforçando o caráter de prova de conceito da abordagem proposta. A heterogeneidade inerente às imagens reflete as condições encontradas na prática clínica, o que tende a favorecer o desenvolvimento de modelos mais robustos e potencialmente mais adaptáveis a sistemas de saúde reais.

No processo de modelagem, foram adotadas estratégias voltadas à mitigação desses desafios, incluindo o balanceamento de classes, a utilização de parada antecipada (*early stopping*) durante o treinamento e a avaliação da generalização em um conjunto de teste previamente separado, garantindo a ausência de vazamento de dados. Ainda assim, é importante ressaltar que o balanceamento de classes não elimina completamente os vieses associados a conjuntos de dados desbalanceados. Da mesma forma, a parada antecipada, embora eficaz, não constitui garantia absoluta contra o *overfitting*, especialmente quando o modelo é aplicado a domínios externos ao conjunto de dados utilizado.

Trabalhos futuros podem explorar a ampliação do conjunto de dados, a validação externa em bases independentes e a integração do modelo em ambientes clínicos reais.

REFERÊNCIAS

ABATI, S. et al. Oral cancer and precancer: a narrative review on the relevance of early diagnosis. **International Journal of Environmental Research and Public Health**, v. 17, n. 24, p. 9160, 2020.

AMIEN, F. M. et al. A comparative study of cnn architectures: Convnext, mobilenetv3, and efficientnet for oral disease diagnosis. **Jurnal Pepadun**, v. 6, n. 1, p. 81–91, 2025.

ANG, K. K. et al. Human papillomavirus and survival of patients with oropharyngeal cancer. **New England Journal of Medicine**, Mass Medical Soc, v. 363, n. 1, p. 24–35, 2010.

APARNA, M. et al. Interpretable analysis of oral mucosal lesions using u-net and quantum approaches. In: IEEE. **2025 International Conference on Inventive Computation Technologies (ICICT)**. [S.l.], 2025. p. 1019–1024.

BROCKLEHURST, P. et al. Screening programmes for the early detection and prevention of oral cancer. **Cochrane database of systematic reviews**, John Wiley & Sons, Ltd, n. 11, 2013.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: **Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining**. [S.l.: s.n.], 2016. p. 785–794.

CHOLLET, F. **Deep learning with Python**. [S.l.]: simon and schuster, 2021.

CIRESAN, D. C. et al. Flexible, high performance convolutional neural networks for image classification. In: BARCELONA, SPAIN:. **IJCAI proceedings-international joint conference on artificial intelligence**. [S.l.], 2011. v. 22, n. 1, p. 1237.

CONG, I.; CHOI, S.; LUKIN, M. D. Quantum convolutional neural networks. **Nature Physics**, Nature Publishing Group UK London, v. 15, n. 12, p. 1273–1278, 2019.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine learning**, Springer, v. 20, n. 3, p. 273–297, 1995.

DEVINDI, G. et al. Multimodal deep convolutional neural network pipeline for ai-assisted early detection of oral cancer. **IEEE Access**, IEEE, 2024.

DIEDERIK, K. Adam: A method for stochastic optimization. **(No Title)**, 2014.

ESTEVA, A. et al. A guide to deep learning in healthcare. **Nature Medicine**, v. 25, n. 1, p. 24–29, 2019.

GILPIN, L. H. et al. Explaining explanations: An overview of interpretability of machine learning. **Proceedings of the IEEE International Conference on Data Science and Advanced Analytics (DSAA)**, 2018.

GÜRSES, B. O. et al. Differentiation of benign and malignant oral lesions through surface texture analysis and svm modeling. **Clinical Oral Investigations**, Springer, v. 29, n. 9, p. 431, 2025.

HAIR, J. F. et al. **Multivariate Data Analysis**. 8. ed. Boston: Cengage Learning, 2018. Available from Internet: <https://eli.johogo.com/Class/CCU/SEM/_Multivariate%20Data%20Analysis_Hair.pdf>.

HE, K. et al. 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.

HO, J. et al. Axial attention in multidimensional transformers. **arXiv preprint arXiv:1912.12180**, 2019.

HOWARD, A. et al. Searching for mobilenetv3. In: **Proceedings of the IEEE International Conference on Computer Vision (ICCV)**. [S.l.: s.n.], 2019.

IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. **Betel-quid and Areca-nut Chewing and Some Areca-nut-derived Nitrosamines**. Lyon, France: International Agency for Research on Cancer, 2004. (IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, v. 85). Available from Internet: <https://www.ncbi.nlm.nih.gov/books/NBK316574/>.

Instituto Nacional de Câncer. **Síntese de Resultados e Comentários**. 2023. Available from Internet: <https://www.gov.br/inca/pt-br/assuntos/cancer/numeros/estimativa/sintese-de-resultados-e-comentarios>.

JOCHER, G.; CHAURASIA, A.; QIU, J. **YOLO by Ultralytics (Version 8)**. 2023. <https://github.com/ultralytics/ultralytics>.

JOLLIFFE, I. T.; CADIMA, J. Principal component analysis: a review and recent developments. **Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences**, the Royal Society publishing, v. 374, n. 2065, p. 20150202, 2016.

KANG, H.; KIESS, A.; CHUNG, C. H. Emerging biomarkers in head and neck cancer in the era of genomics. **Nature reviews Clinical oncology**, Nature Publishing Group UK London, v. 12, n. 1, p. 11–26, 2015.

KIRILLOV, A. et al. Segment anything. In: **Proceedings of the IEEE/CVF international conference on computer vision**. [S.l.: s.n.], 2023. p. 4015–4026.

KOONTONGKAEW, S. The tumor microenvironment contribution to development, growth, invasion and metastasis of head and neck squamous cell carcinomas. **Journal of Cancer**, v. 4, n. 1, p. 66, 2013.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature**, v. 521, n. 7553, p. 436–444, 2015.

LECUN, Y. et al. Gradient-based learning applied to document recognition. **Proceedings of the IEEE**, Ieee, v. 86, n. 11, p. 2278–2324, 2002.

LEE, Y. O.; KIM, J.; LEE, J. W. Improving diagnostic accuracy for oral cancer with inpainting synthesis lesions generated using diffusion models. **arXiv preprint arXiv:2508.06151**, 2025.

LIMA, A. M. da C. et al. Delay in diagnosis of oral cancer: a systematic review. **Medicina oral, patologia oral y cirugia bucal**, v. 26, n. 6, p. e815, 2021.

LIN, T.-Y. et al. Microsoft coco: Common objects in context. In: SPRINGER. **European conference on computer vision**. [S.l.], 2014. p. 740–755.

LIU, G. et al. Image inpainting for irregular holes using partial convolutions. In: **Proceedings of the European conference on computer vision (ECCV)**. [S.l.: s.n.], 2018. p. 85–100.

LIU, Z. et al. A convnet for the 2020s. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2022.

MASON, H.; WIGGINS, C. A taxonomy of data science. **Dataists. Last modified September**, v. 25, 2010.

MUKHERJEE, R. et al. Multimodal rgb-hsi feature fusion with patient-aware incremental heuristic meta-learning for oral lesion classification. **arXiv preprint arXiv:2511.12268**, 2025.

National Institute of Dental and Craniofacial Research. **Detecting oral cancer: a guide for health care professionals**. U.S. Department of Health and Human Services, 2020. Accessed: 2025-08-18. Available from Internet: <https://www.nidcr.nih.gov/health-info/publications/detecting-oral-cancer-guide-healthcare-professionals>.

Oral Cancer Foundation. **Oral cancer facts**. 2025. Accessed: 2025-08-18. Available from Internet: <https://oralcancerfoundation.org/facts/>.

PIYARATHNE, N. S. et al. A comprehensive dataset of annotated oral cavity images for diagnosis of oral cancer and oral potentially malignant disorders. **Oral Oncology**, v. 156, p. 106946, 2024.

RETHMAN, M. P. et al. Evidence-based clinical recommendations regarding screening for oral squamous cell carcinomas. **The Journal of the American Dental Association**, Elsevier, v. 141, n. 5, p. 509–520, 2010.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "why should i trust you?"explaining the predictions of any classifier. In: **Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining**. [S.l.: s.n.], 2016. p. 1135–1144.

ROMBACH, R. et al. High-resolution image synthesis with latent diffusion models. In: **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**. [S.l.: s.n.], 2022. p. 10684–10695.

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

ROSEBROCK, A. **Fine-tuning with Keras and Deep Learning**. 2019. Accessed: 2026-02-15. Available from Internet: <https://pyimagesearch.com/2019/06/03/fine-tuning-with-keras-and-deep-learning/>.

RUIZ, N. et al. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**. [S.l.: s.n.], 2023. p. 22500–22510.

SAITO, T.; REHMSMEIER, M. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. **PloS one**, Public Library of Science, v. 10, n. 3, p. e0118432, 2015.

SANDLER, M. et al. Mobilenetv2: Inverted residuals and linear bottlenecks. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2018.

SANKARANARAYANAN, R. et al. Oral cancer: prevention, early detection, and treatment. **Cancer: disease control priorities. 3rd ed. Washington, DC: The International Bank for Reconstruction and Development/The World Bank**, v. 3, p. 85–99, 2015.

SCHWENDICKE, F. a.; SAMEK, W.; KROIS, J. Artificial intelligence in dentistry: chances and challenges. **Journal of dental research**, SAGE Publications Sage CA: Los Angeles, CA, v. 99, n. 7, p. 769–774, 2020.

SEER – Surveillance, Epidemiology, and End Results Program. **Cancer of the Oral Cavity and Pharynx - Cancer Stat Facts**. 2020. <https://seer.cancer.gov/statfacts/html/oralcav.html>. Accessed: 2025-10-07.

SELVARAJU, R. R. et al. 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.

SIM – Sistema de Informações sobre Mortalidade. **Atlas de Mortalidade por Câncer**. 2019. <https://mortalidade.inca.gov.br>. Ministério da Saúde. Accessed: 2025-10-07.

SREENIVAS, A. et al. Indian sign language communicator using convolutional neural network. **International Journal of Advanced Science and Technology**, v. 29, n. 3, 2020.

TAMMINA, S. et al. Transfer learning using vgg-16 with deep convolutional neural network for classifying images. **International Journal of Scientific and Research Publications (IJSRP)**, v. 9, n. 10, p. 143–150, 2019.

TAN, M.; LE, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In: **Proceedings of the International Conference on Machine Learning (ICML)**. [S.l.: s.n.], 2019.

THOMA, M.; LEE, C. M. G. **Receiver Operating Characteristic (ROC) Curve with False Positive Rate and True Positive Rate**. 2018. <https://commons.wikimedia.org/wiki/File:Roc_curve.svg#filelinks>. Illustration showing ROC curve and performance of example classifiers. Wikimedia Commons. Accessed: 2025-02-07.

TOPOL, E. J. High-performance medicine: the convergence of human and artificial intelligence. **Nature Medicine**, v. 25, n. 1, p. 44–56, 2019.

VIET, C. T.; SCHMIDT, B. L. Biologic mechanisms of oral cancer pain and implications for clinical therapy. **Journal of Dental Research**, v. 91, n. 5, p. 447–453, 2011.

WARNAKULASURIYA, S. Global epidemiology of oral and oropharyngeal cancer. **Oral Oncology**, v. 45, n. 4–5, p. 309–316, 2009.

WOO, S. et al. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In: **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**. [S.l.: s.n.], 2023. p. 16133–16142.

World Health Organization. **Cancer Today**. 2023. Available from Internet: <https://gco.iarc.fr/today/en>.

YADAV, D. P. et al. Explainable label guided lightweight network with axial transformer encoder for early detection of oral cancer. **Scientific Reports**, Nature Publishing Group UK London, v. 15, n. 1, p. 6391, 2025.

YANG, X. et al. Concepts of artificial intelligence for computer-assisted drug discovery. **Chemical reviews**, ACS Publications, v. 119, n. 18, p. 10520–10594, 2019.

Zenodo. **Zenodo**. 2013. <https://zenodo.org>.