REMUNOM
ISSN 2178-6925

Received: 19/04/2026 - **Accepted:** 06/05/2026
Vol: 13.09
DOI: 10.66104/eecb2s81
Pages: 1-34

AVALIAÇÃO DE ARQUITETURAS CONVOLUCIONAIS PRÉ-TREINADAS NA DETECÇÃO DE CÂNCER ORAL

EVALUATION OF PRE-TRAINED CONVOLUTIONAL ARCHITECTURES FOR ORAL CANCER DETECTION

EVALUACIÓN DE ARQUITECTURAS CONVOLUCIONALES PREENTRENADAS EN LA DETECCIÓN DE CÁNCER ORAL

Maria Isabelly de Brito Rodrigues
Graduanda em Tecnologia e Análise e Desenvolvimento de Sistemas
Instituto Federal do Piauí - IFPI, Brasil
E-mail: isabellybrt77@gmail.com

Wanderson de Vasconcelos Rodrigues da Silva
Doutor em Ciência da Propriedade Intelectual pela UFS
Docente do Instituto Federal do Piauí - IFPI, Brasil
E-mail: wanderson.vasconcelos@ifpi.edu.br

Ricardo Moura Sekeff Budaruiche
Doutor em Ciência da Computação pela USP
Instituto Federal do Piauí - IFPI, Brasil
E-mail: ricardo.sekeff@ifpi.edu.br

Iallen Gabio de Sousa Santos
Doutor em Ciência da Computação pela UFPE
Docente do Instituto Federal do Piauí - IFPI, Brasil
E-mail: iallen@ifpi.edu.br

Resumo

O câncer bucal permanece um dos principais desafios de saúde pública, especialmente em países em desenvolvimento, onde a detecção tardia contribui significativamente para a alta mortalidade associada à doença. Métodos automatizados baseados em Visão Computacional e Aprendizado Profundo têm demonstrado grande potencial no auxílio ao diagnóstico precoce, oferecendo suporte objetivo à avaliação clínica. Neste trabalho, investigou-se o desempenho de três arquiteturas convolucionais pré-treinadas amplamente consolidadas na literatura, DenseNet121, GoogLeNet e ResNet18, aplicadas à classificação binária de imagens clínicas da cavidade oral, utilizando um conjunto público contendo 950 amostras. O estudo emprega *transfer learning* e validação cruzada com 5 *folds*, permitindo uma análise da capacidade de generalização dos modelos. Os resultados indicam que todas as arquiteturas alcançaram métricas médias superiores a 0,92, incluindo acurácia, precisão, recall, F1-score e AUC, em todas as abordagens avaliadas. Embora a DenseNet121 tenha registrado as maiores médias pontuais entre os modelos (F1-score = 0.9476), a sobreposição das faixas de variação baseadas no desvio-padrão impede afirmar diferenças estatisticamente significativas de desempenho. Os achados reforçam o potencial das CNNs como ferramentas complementares no processo de triagem de lesões orais, apontando para a viabilidade da abordagem em cenários clínicos.

Palavras-chave: Aprendizado Profundo; Visão Computacional; Câncer Oral; Redes Neurais Convolucionais.

REMUNOM
ISSN 2178-6925

Received: 19/04/2026 - **Accepted:** 06/05/2026
Vol: 13.09
DOI: 10.66104/eecb2s81
Pages: 1-34

Abstract

Oral cancer remains one of the main public health challenges, especially in developing countries, where late detection contributes significantly to the high mortality associated with the disease. Automated methods based on Computer Vision and Deep Learning have demonstrated great potential in assisting early diagnosis, providing objective support to clinical evaluation. In this work, we investigated the performance of three widely established pre-trained convolutional architectures, DenseNet121, GoogLeNet, and ResNet18, applied to binary classification of clinical oral cavity images using a public dataset containing 950 samples. The study employs transfer learning and 5-fold cross-validation, allowing an analysis of the models' generalization capacity. The results indicate that all architectures achieved average metrics above 0.92 across all evaluated measures, including accuracy, precision, recall, F1-score, and AUC. Although DenseNet121 recorded the highest point estimates among the models (F1-score = 0.9476), the overlap of variability ranges based on standard deviation prevents asserting statistically significant performance differences. These findings reinforce the potential of CNNs as complementary tools in the screening of oral lesions, supporting the overall feasibility of the proposed approach in clinical scenarios.

Keywords: Deep Learning; Computer Vision; Oral Cancer; Convolutional Neural Network.

Resumen

El cáncer bucal sigue siendo uno de los principales desafíos de salud pública, especialmente en países en desarrollo, donde la detección tardía contribuye significativamente a la alta mortalidad asociada a la enfermedad. Los métodos automatizados basados en Visión Computacional y Aprendizaje Profundo han demostrado un gran potencial en el auxilio al diagnóstico temprano, ofreciendo soporte objetivo a la evaluación clínica. En este trabajo, se investigó el desempeño de tres arquitecturas convolucionales preentrenadas ampliamente consolidadas en la literatura, DenseNet121, GoogLeNet y ResNet18, aplicadas a la clasificación binaria de imágenes clínicas de la cavidad oral, utilizando un conjunto público que contiene 950 muestras. El estudio emplea transfer learning y validación cruzada con 5 folds, permitiendo un análisis de la capacidad de generalización de los modelos. Los resultados indican que todas las arquitecturas alcanzaron métricas promedio superiores a 0,92, incluyendo exactitud, precisión, recall, F1-score y AUC, en todos los enfoques evaluados. Aunque la DenseNet121 registró los mayores promedios puntuales entre los modelos (F1-score = 0.9476), la superposición de los rangos de variación basados en la desviación estándar impide afirmar diferencias estadísticamente significativas en el desempeño. Los hallazgos refuerzan el potencial de las CNNs como herramientas complementarias en el proceso de triaje de lesiones orales, apuntando hacia la viabilidad del enfoque en escenarios clínicos.

Palabras clave: Aprendizaje Profundo; Visión Computacional; Cáncer Oral; Redes Neuronales Convolucionales.

1. Introdução

O câncer oral representa um relevante problema de saúde pública, com cerca de 377 mil novos casos e 177 mil mortes registradas mundialmente a cada ano, segundo estimativas globais do GLOBOCAN (Sung *et al.*, 2021). A detecção precoce é crítica para melhoria das taxas de sobrevivência, visto que o diagnóstico em estágios avançados está associado a tratamentos mais agressivos e menor prognóstico (Warnakulasuriya, 2009). No entanto, a identificação clínica de lesões malignas ou potencialmente malignas depende fortemente da experiência do profissional, o que introduz variabilidade e limitações na sensibilidade diagnóstica

REMUNOM
ISSN 2178-6925

Received: 19/04/2026 - **Accepted:** 06/05/2026
Vol: 13.09
DOI: 10.66104/eecb2s81
Pages: 1-34

(Speight *et al*., 2018), uma lacuna que ferramentas computacionais de apoio ao diagnóstico têm potencial concreto de preencher.

Avanços recentes em Visão Computacional e Aprendizado Profundo (*Deep Learning*) têm impulsionado o desenvolvimento de sistemas automatizados capazes de analisar imagens bucais com alto desempenho. Redes Neurais Convolucionais (CNNs) têm se destacado por apresentar resultados promissores em diferentes tarefas médicas, incluindo classificação automatizada de imagens clínicas com suspeição de câncer oral (Welikala *et al*., 2020). Esses modelos demonstram capacidade de identificar padrões sutis que muitas vezes não são perceptíveis a olho nu, com potencial para auxiliar profissionais na triagem e reduzir a subjetividade do diagnóstico em cenários experimentais controlados, característica relevante para investigações futuras voltadas ao uso clínico.

Apesar desse avanço, ainda há escassez de estudos focados especificamente em câncer oral com uso de bases públicas e reprodutíveis, o que dificulta comparações e validação científica (Warin *et al.*, 2021). Para que soluções baseadas em CNNs avancem do ambiente de pesquisa para aplicações clínicas concretas, é necessário não apenas avaliar se esses modelos funcionam, mas identificar quais características arquiteturais, sensibilidade, estabilidade e capacidade de generalização, os tornam mais adequados para uso prático. Assim, investigações sistemáticas envolvendo múltiplas arquiteturas CNN pré-treinadas tornam-se relevantes para ampliar o conhecimento sobre o potencial desses modelos em ambientes clínicos reais.

Este trabalho tem como objetivo avaliar e comparar o desempenho de três arquiteturas convolucionais pré-treinadas - GoogLeNet, ResNet18 e DenseNet121 - aplicadas à classificação binária de imagens clínicas orais, com vistas a identificar qual perfil arquitetural apresenta maior adequação para investigações futuras em apoio ao diagnóstico. A escolha dessas três arquiteturas é deliberada e fundamentada em três critérios complementares.

Primeiro, custo computacional: as três redes figuram entre as arquiteturas com menor número de parâmetros entre os modelos pré-treinados consolidados, GoogLeNet com aproximadamente 6,8 milhões, DenseNet121 com 8 milhões e ResNet18 com 11,7 milhões (Bianco *et al.*, 2018), tornando-as viáveis para

REMUNOM
ISSN 2178-6925

Received: 19/04/2026 - **Accepted:** 06/05/2026
Vol: 13.09
DOI: 10.66104/eecb2s81
Pages: 1-34

inferência em ambientes com recursos limitados, como unidades básicas de saúde (Paramasivam *et al*., 2024). Segundo, disponibilidade e reprodutibilidade: as três arquiteturas estão disponíveis publicamente com pesos pré-treinados no ImageNet por meio de bibliotecas de código aberto amplamente adotadas, o que facilita a replicação dos experimentos e a comparação entre estudos. Terceiro, relevância na literatura médica: GoogLeNet (Szegedy *et al.*, 2015), ResNet (He *et al.*, 2016) e DenseNet (Huang *et al.,* 2017) são repetidamente empregadas como referências em estudos de classificação de imagens médicas (Tajbakhsh *et al.*, 2016; Devindi *et al.*, 2024), o que permite situar os resultados obtidos no contexto de evidências já acumuladas.

Os resultados obtidos contribuem com evidências quantitativas sobre a aplicabilidade de CNNs na análise automatizada de imagens clínicas orais e oferecem uma base metodológica reprodutível, tanto para estudos futuros com o mesmo *dataset* quanto para o desenvolvimento de soluções clínicas fundamentadas em dados reais.

Por fim, este artigo está estruturado da seguinte forma. Após esta introdução, apresenta-se, na Seção 2, a fundamentação teórica necessária para compreender os conceitos que embasam este estudo, incluindo aspectos do câncer oral, princípios de *Transfer Learning* e arquiteturas convolucionais. Em seguida, a Seção 3 discute os principais trabalhos relacionados e situa esta pesquisa no estado da arte, destacando as implicações práticas de cada estudo. A Seção 4 descreve detalhadamente a metodologia adotada, contemplando o *dataset* utilizado, o pré-processamento aplicado, a validação cruzada e a configuração dos experimentos. Os resultados obtidos e sua análise são discutidos na Seção 5, enquanto a Seção 6 oferece uma reflexão crítica sobre os achados, suas limitações e implicações. Por fim, a Seção 7 apresenta as conclusões e indica os próximos passos em direção à aplicação clínica dos modelos avaliados.

2. Fundamentação Teórica

2.1. Câncer Oral e Diagnóstico Auxiliado por Computador

O câncer oral é caracterizado pelo crescimento anormal de células nos tecidos da cavidade oral, abrangendo regiões como língua, gengivas, mucosa

REMUNOM
ISSN 2178-6925

Received: 19/04/2026 - **Accepted:** 06/05/2026
Vol: 13.09
DOI: 10.66104/eecb2s81
Pages: 1-34

bucal e lábios. É uma neoplasia frequentemente diagnosticada tardiamente devido à sutileza dos sinais iniciais, o que compromete o prognóstico e aumenta taxas de mortalidade (Warnakulasuriya, 2009).

A literatura aponta que exames clínicos tradicionais podem apresentar variação significativa entre profissionais, resultando em inconsistências na detecção de lesões potencialmente malignas (Speight *et al.*, 2018). Nesse contexto, Sistemas de Diagnóstico Auxiliado por Computador (CAD) vêm sendo explorados como ferramentas capazes de aumentar a sensibilidade diagnóstica e reduzir a subjetividade da análise clínica (Welikala *et al.*, 2020). Abordagens baseadas em aprendizado profundo têm se destacado por sua habilidade de capturar padrões complexos em imagens médicas, possibilitando triagens mais rápidas e precisas (Fu *et al*., 2020).

2.2. Transfer Learning

O *Transfer Learning* é uma abordagem que permite reutilizar modelos previamente treinados em bases extensas e gerais, como o ImageNet, composto por mais de um milhão de imagens distribuídas em mil classes (Deng *et al*., 2009). Essa estratégia é especialmente útil em aplicações médicas, onde *datasets* tendem a ser pequenos devido a limitações éticas e de disponibilidade. Tajbakhsh *et al*. (2016) demonstram que a adaptação (*fine-tuning*) de modelos pré-treinados oferece desempenho superior ao de redes treinadas do zero em cenários com poucos dados. Nesse sentido, as arquiteturas avaliadas neste estudo são empregadas como extratoras de características profundas, sendo as camadas finais modificadas para a tarefa de classificação binária.

2.3. Arquiteturas Avaliadas

As três arquiteturas investigadas neste trabalho possuem relevância consolidada na literatura de visão computacional.

GoogLeNet: A GoogLeNet propõe o módulo Inception, que combina convoluções de diferentes tamanhos executadas em paralelo, buscando eficiência computacional sem perda de profundidade. Essa abordagem reduz drasticamente o número de parâmetros, tornando a rede mais leve e adequada para aplicações em larga escala.

ResNet18: A ResNet introduz conexões residuais (*skip connections*) que

REMUNOM
ISSN 2178-6925

Received: 19/04/2026 - **Accepted:** 06/05/2026
Vol: 13.09
DOI: 10.66104/eecb2s81
Pages: 1-34

permitem o fluxo mais direto dos gradientes durante a retropropagação, evitando problemas de degradação em redes profundas (He *et al*., 2016). Esse mecanismo permitiu o treinamento de redes com centenas de camadas, consolidando as ResNets como uma das arquiteturas mais influentes da área.

DenseNet121: A DenseNet utiliza conexões densas em que cada camada recebe como entrada os mapas de características de todas as camadas anteriores (Huang *et al*., 2017). Esse design incentiva o reuso de características, reduz redundâncias e melhora o fluxo de gradientes, especialmente útil em *datasets* pequenos.

3. Trabalhos Relacionados

A pesquisa em detecção automatizada de câncer oral por meio de redes neurais convolucionais tem crescido nos últimos anos, com abordagens que exploram tanto imagens histopatológicas quanto imagens clínicas. A literatura evidencia avanços importantes, mas também desafios metodológicos que impactam diretamente a validade, a reprodutibilidade dos resultados e, consequentemente, a confiabilidade de sistemas construídos com base nessas abordagens.

Um dos trabalhos mais relevantes na área propõe uma investigação sistemática do uso de imagens histopatológicas para classificação binária de câncer oral, com ênfase especial em problemas de vazamento de dados decorrentes de estratégias inadequadas de *data augmentation*. Os autores demonstram que operações como rotações e *flips* aplicadas antes da divisão entre treino e teste podem levar a amostras quase idênticas presentes em ambos os conjuntos, elevando artificialmente a acurácia dos modelos, produzindo, na prática, sistemas que não generalizam diante de dados reais. Além de identificar esse problema, o estudo avalia diferentes arquiteturas baseadas em *transfer learning*, incluindo variantes da EfficientNet combinadas com mecanismos de atenção baseados em Transformer. Os resultados indicam que modelos híbridos desse tipo podem alcançar desempenho competitivo (por volta de 87% de acurácia), desde que o protocolo experimental seja rigorosamente controlado. O trabalho destaca ainda a necessidade de comparações justas entre estudos, sugerindo que somente

REMUNOM
ISSN 2178-6925

Received: 19/04/2026 - **Accepted:** 06/05/2026
Vol: 13.09
DOI: 10.66104/eecb2s81
Pages: 1-34

resultados obtidos sem evidências de vazamento devem ser utilizados como referência (Nogueira; Gomes, 2025). Essa lição orienta diretamente a metodologia do presente trabalho, que não emprega *data augmentation* antes da partição dos dados, garantindo que os resultados reflitam o comportamento real dos modelos diante de dados não vistos.

Outro estudo relevante investiga o uso de CNNs para detecção de carcinoma oral a partir de imagens de biópsia e imagens clínicas. Embora não disponibilize amplos detalhes metodológicos, o trabalho é citado na literatura como um exemplo de aplicação direta de redes convolucionais para predição de lesões potencialmente malignas, empregando modelos leves baseados em *transfer learning*. Os autores relatam métricas promissoras e reforçam a viabilidade de soluções computacionalmente eficientes para cenários de triagem clínica, onde o tempo de inferência e a simplicidade do modelo são fatores determinantes para a implantação em ambientes com recursos limitados, como unidades básicas de saúde (Paramasivam *et al*., 2024). Essa perspectiva reforça a relevância de avaliar arquiteturas enxutas e bem estabelecidas, como as investigadas neste trabalho, antes de recorrer a modelos mais complexos e custosos.

Para além das abordagens monocanal, estudos recentes têm explorado estratégias multimodais, combinando diferentes tipos de dados, incluindo histopatologia, radiologia e informações clínicas, com o objetivo de melhorar a robustez das predições. Pesquisas deste tipo avaliam múltiplas arquiteturas pré-treinadas, como ResNet-50, DenseNet-121 e Inception, e demonstram que a fusão de características provenientes de diferentes modalidades pode ampliar a precisão em comparação com modelos baseados apenas em imagem. Esses achados evidenciam que a variabilidade intrínseca aos dados clínicos de câncer oral pode ser mitigada por técnicas de agregação ou fusão de modelos (Devindi *et al*., 2024). Embora promissora, essa direção pressupõe infraestrutura de coleta e integração de dados ainda indisponível na maioria dos cenários de triagem de baixo custo. O presente trabalho, ao focar exclusivamente em imagens clínicas, representa uma etapa anterior e necessária nessa trajetória: estabelecer uma base sólida e reprodutível para sistemas monomodais, antes de avançar para arquiteturas mais complexas.

REMUNOM
ISSN 2178-6925

Received: 19/04/2026 - **Accepted:** 06/05/2026
Vol: 13.09
DOI: 10.66104/eecb2s81
Pages: 1-34

Em conjunto, a literatura evidencia que CNNs têm potencial real para integrar sistemas de apoio ao diagnóstico oral, mas que rigor metodológico e atenção às condições de implantação são fatores determinantes para que esse potencial se converta em valor clínico concreto. Este trabalho busca contribuir com evidências quantitativas e reprodutíveis nessa direção.

3.1. Contribuição

Os trabalhos analisados demonstram que CNNs pré-treinadas têm potencial real para classificação de lesões orais, mas raramente oferecem comparações sistemáticas entre múltiplas arquiteturas sob um protocolo experimental rigoroso e reprodutível. Estudos como Karthikeyan *et al*. (2024) e Ormeño-Arriagada *et al.* (2026) exploram o mesmo conjunto de dados utilizado neste trabalho, porém com foco em arquiteturas distintas ou sem validação cruzada formal. Este trabalho contribui com uma avaliação comparativa controlada entre GoogLeNet, ResNet18 e DenseNet121, empregando 5-Fold Cross-Validation e ausência intencional de *data augmentation*, o que permite isolar a capacidade discriminativa intrínseca de cada arquitetura. Visando facilitar a reprodutibilidade e o uso como linha de base em pesquisas futuras bem como em ambientes educacionais, os códigos fonte utilizados para a execução e análises de dados de todos os experimentos realizados neste trabalho foram disponibilizados em repositório público (Rodrigues, 2026).

3.2. Limitações do Presente Trabalho

A análise da literatura evidencia limitações recorrentes que afetam tanto os estudos anteriores quanto o presente trabalho, e que merecem reconhecimento explícito antes da descrição metodológica.

Um dos problemas mais documentados na área diz respeito à qualidade dos rótulos diagnósticos. Como apontado por Nogueira e Gomes (2025), a ausência de controle rigoroso sobre a procedência das anotações, incluindo a falta de confirmação histopatológica, pode comprometer a validade externa dos modelos, uma vez que a tarefa computacional passa a ser a discriminação de imagens segundo rótulos atribuídos por terceiros, e não a detecção de câncer oral em sentido anatomopatológico estrito. Nesse cenário, os modelos aprendem a separar imagens rotuladas como suspeitas daquelas rotuladas como saudáveis, o que é

REMUNOM
ISSN 2178-6925

Received: 19/04/2026 - **Accepted:** 06/05/2026
Vol: 13.09
DOI: 10.66104/eecb2s81
Pages: 1-34

distinto de reconhecer características morfológicas de uma entidade nosológica clinicamente definida. A ausência de discriminação do subtipo histológico específico por imagem reforça essa distinção: sem saber se cada lesão corresponde a um carcinoma espinocelular, leucoplasia com displasia ou outra alteração, não é possível afirmar, com plena segurança clínica, que o modelo aprendeu a reconhecer uma entidade nosológica específica.

Esse aspecto delimita a validade externa dos resultados obtidos neste trabalho: os modelos demonstraram boa capacidade de discriminação dentro deste conjunto específico, mas sua translação para contextos clínicos reais exige validação adicional com bases de dados que disponham de confirmação histopatológica documentada. Destaca-se que a responsabilidade sobre os rótulos utilizados é dos autores do dataset original (Zaidpy, 2022), e que o conjunto já foi empregado em outros estudos da área (Karthikeyan *et al.*, 2024; Ormeño-Arriagada *et al.,* 2026), atestando sua utilidade como base de referência mesmo diante dessas restrições.

A metodologia adotada neste trabalho, hiperparâmetros, estratégia de validação cruzada e conjunto de arquiteturas avaliadas, é independente do dataset. Desta forma, é possível replica-la em outros conjuntos de dados com maior grau de validação clínica construídos ou disponibilizados no futuro, aspecto retomado nas considerações finais.

Além das limitações relacionadas aos rótulos diagnósticos, o presente trabalho apresenta outras restrições metodológicas que devem ser explicitamente reconhecidas: (i) base única - todos os experimentos foram conduzidos em um único dataset, sem validação externa em conjuntos independentes; (ii) ausência de metadados clínicos - o conjunto não inclui informações como idade, sexo, método de aquisição ou características do equipamento utilizado, impedindo análises estratificadas ou controle de variáveis confundidoras; (iii) ausência de validação externa - os resultados refletem desempenho interno estimado por validação cruzada, sem teste em dados provenientes de outras instituições ou protocolos de aquisição; e (iv) restrições inferenciais do desenho experimental - a ausência de testes estatísticos inferenciais formais entre arquiteturas, justificada pelo reduzido número de partições (k=5), limita conclusões sobre superioridade estatística entre

REMUNOM
ISSN 2178-6925

Received: 19/04/2026 - **Accepted:** 06/05/2026
Vol: 13.09
DOI: 10.66104/eecb2s81
Pages: 1-34

os modelos avaliados.

4. Metodologia

4.1. Dataset

O presente estudo utiliza um único conjunto de dados público disponibilizado por Mohd Zaid Rashid (Zaidpy, 2022) na plataforma Kaggle, amplamente empregado em pesquisas sobre detecção automatizada de câncer oral. O dataset contém 950 imagens clínicas coloridas da cavidade oral, distribuídas em duas classes: "Câncer" (500 amostras) e "Não Câncer" (450 amostras). As imagens são fotografias clínicas da cavidade oral com a boca aberta, incluindo língua, gengivas e mucosa bucal, obtidas sem o uso de dispositivos odontológicos especializados de captura intraoral, apresentando variações naturais de iluminação, foco, resolução e posição anatômica, características comuns em cenários clínicos reais.

As amostras abrangem diferentes regiões da cavidade oral, com a classe "Não Câncer" representando tecidos saudáveis, enquanto a classe "Câncer" contém fotografias de lesões orais com aparência clínica compatível com alterações malignas ou potencialmente malignas. Os rótulos foram atribuídos pelos autores originais do conjunto, e o dataset não discrimina o subtipo histológico específico de cada lesão - ou seja, não é possível determinar, para cada imagem, se a alteração corresponde a um carcinoma espinocelular, leucoplasia com displasia ou outra forma de neoplasia maligna ou potencialmente maligna. As implicações metodológicas dessa característica são discutidas na Seção 3.2. Essa diversidade e ausência de padronização tornam o conjunto desafiador, ao mesmo tempo em que refletem a variabilidade presente no exame clínico de pacientes reais.

A Figura 1 apresenta quatro imagens exemplares provenientes do conjunto, ilustrando a diferença visual entre os tecidos saudáveis (parte superior) e as regiões lesionadas (parte inferior). Essas imagens são utilizadas apenas com finalidade ilustrativa, uma vez que o conjunto completo apresenta uma variabilidade muito mais ampla em termos de textura, coloração e severidade das lesões.

Todas as imagens foram fornecidas em formato RGB e não apresentam

REMUNOM
ISSN 2178-6925

Received: 19/04/2026 - **Accepted:** 06/05/2026
Vol: 13.09
DOI: 10.66104/eecb2s81
Pages: 1-34

subdivisões pré-definidas entre treino, validação e teste, cabendo ao pesquisador definir a estratégia de particionamento. Além disso, o conjunto não inclui informações adicionais como idade, sexo, método de aquisição ou características do equipamento, sendo composto exclusivamente pelas imagens. Essa ausência de metadados faz com que o processo de classificação dependa exclusivamente da informação visual, sem possibilidade de incorporação de atributos clínicos complementares.

Figura 1: Exemplos de imagens do dataset utilizado neste estudo: duas amostras da classe "Não Câncer" (superior) e duas amostras da classe "Câncer" (inferior).



Fonte: Imagem extraída do *Oral Cancer Dataset* por ZaidPy, licenciada sob Apache 2.0.

Para garantir consistência no pré-processamento, todas as imagens foram padronizadas quanto ao tamanho, normalização e orientações adequadas ao uso de redes neurais convolucionais pré-treinadas. A organização equilibrada do dataset, com leve predominância da classe "Câncer", permite a aplicação de estratégias de validação cruzada sem a necessidade de técnicas adicionais de balanceamento, embora abordagens complementares possam ser consideradas em investigações posteriores. Este conjunto, apesar de limitado em tamanho, possui qualidade visual suficiente para treinar modelos baseados em *transfer learning*.

REMUNOM
ISSN 2178-6925

Received: 19/04/2026 - **Accepted:** 06/05/2026
Vol: 13.09
DOI: 10.66104/eecb2s81
Pages: 1-34

4.2. Pré-processamento

O pré-processamento das imagens foi conduzido de forma a garantir consistência visual entre as amostras e compatibilidade com os requisitos das arquiteturas pré-treinadas utilizadas neste estudo. Todas as etapas aplicadas têm como objetivo padronizar as entradas, reduzir variações indesejadas e alinhar as imagens ao espaço estatístico no qual os modelos foram originalmente treinados.

A primeira etapa consistiu no redimensionamento de todas as imagens para uma resolução fixa de 224×224 pixels, tamanho convencionalmente adotado por redes pré-treinadas no ImageNet. Esse ajuste assegura que todas as amostras possuam dimensões idênticas, evitando distorções na camada de entrada e garantindo uniformidade arquitetural entre os modelos avaliados.

Em seguida, as imagens foram convertidas para uma representação numérica normalizada. Esse processo inclui a reorganização dos canais de cor para o formato utilizado pelas redes convolucionais e a escala dos valores de intensidade dos pixels para um intervalo contínuo. Tal transformação possibilita que as redes interpretem as imagens de forma coerente e estável durante o processo de aprendizado.

Por fim, aplicou-se uma normalização estatística baseada nas médias e desvios-padrão dos canais RGB do ImageNet. Essa etapa é essencial em cenários de *transfer learning*, pois ajusta a distribuição das imagens do dataset utilizado para que se aproximem da distribuição original empregada no treinamento das arquiteturas pré-treinadas. Dessa forma, reduz-se o impacto de discrepâncias entre bases distintas e facilita-se a adaptação dos modelos ao novo conjunto de dados.

Cabe destacar que nenhuma técnica de *data augmentation* foi empregada. Essa escolha teve como finalidade avaliar exclusivamente a capacidade intrínseca das redes em aprender padrões discriminativos a partir das imagens originais, eliminando possíveis interferências decorrentes de transformações artificiais das amostras.

4.3. Validação Cruzada Estratificada (5-Fold Stratified Cross-Validation)

A avaliação dos modelos neste estudo foi realizada por meio da técnica de Validação Cruzada Estratificada do tipo *K-Fold*, implementada com StratifiedKFold, conforme disponibilizado pela biblioteca Scikit-learn (Pedregosa *et al.,* 2011). Essa

REMUNOM
ISSN 2178-6925

Received: 19/04/2026 - **Accepted:** 06/05/2026
Vol: 13.09
DOI: 10.66104/eecb2s81
Pages: 1-34

abordagem é indicada em cenários com conjuntos de dados limitados e classes potencialmente desbalanceadas, sendo preferível ao *K-Fold* padrão sempre que a proporção entre classes deve ser preservada em cada partição. Neste trabalho, adotou-se a configuração K=5, resultando em um procedimento de *5-Fold Stratified Cross-Validation*.

A principal distinção da versão estratificada em relação ao *K-Fold* padrão reside no critério de particionamento: ao invés de dividir o conjunto de forma completamente aleatória, o StratifiedKFold garante que cada *fold* preserve a proporção original das classes tanto no subconjunto de treinamento quanto no de validação. No dataset utilizado, composto por 500 amostras da classe "Câncer" e 450 da classe "Não Câncer", essa estratégia assegura que cada fold de validação contenha aproximadamente 100 amostras positivas e 90 negativas. Isto visa evitar que variações na composição dos subconjuntos introduzam viés nas estimativas de desempenho, problema particularmente relevante em conjuntos de tamanho reduzido, onde divisões aleatórias podem gerar folds com distribuições muito distintas da distribuição global.

O método utilizado neste trabalho particiona o conjunto completo de imagens em cinco subconjuntos de tamanho aproximadamente igual, denominados *folds*, respeitando a distribuição proporcional de classes em cada partição. Em cada iteração, um desses subconjuntos é reservado exclusivamente para validação, enquanto os quatro *folds* restantes são utilizados para o treinamento do modelo. O processo é repetido cinco vezes, de modo que cada *fold* atua exatamente uma vez como conjunto de validação e quatro vezes como parte do conjunto de treinamento. Ao final das cinco iterações, todas as amostras do dataset terão sido utilizadas tanto para treinamento quanto para validação.

De forma mais específica, o procedimento de estratificação é realizado por meio do agrupamento inicial das amostras segundo suas respectivas classes. Em seguida, cada grupo é embaralhado de forma controlada, utilizando uma semente fixa para garantir reprodutibilidade, e particionado em subconjuntos aproximadamente equilibrados. Cada *fold* de validação é então construído pela combinação de uma partição de cada classe, enquanto as partições restantes compõem o conjunto de treinamento. Esse processo é repetido iterativamente até

REMUNOM
ISSN 2178-6925

Received: 19/04/2026 - **Accepted:** 06/05/2026
Vol: 13.09
DOI: 10.66104/eecb2s81
Pages: 1-34

que todas as partições tenham sido utilizadas como conjunto de validação.

A implementação adotada neste estudo segue essa metodologia, utilizando partições estratificadas consistentes ao longo das avaliações dos três modelos investigados (GoogLeNet, ResNet18 e DenseNet121). Para cada iteração do processo de validação cruzada, quatro dos cinco *folds* são utilizados para compor o conjunto de treinamento, enquanto o *fold* restante é reservado exclusivamente para validação. Dessa forma, o modelo é treinado com aproximadamente 80% dos dados disponíveis e avaliado nos 20% restantes, sendo esse procedimento repetido até que cada *fold* tenha sido utilizado uma única vez como validação. A acurácia é monitorada a cada época durante o treinamento, enquanto precisão, *recall*, F1-score e AUC são calculados ao final de cada *fold*, a partir da avaliação do modelo com os melhores pesos de validação registrados ao longo das épocas.

Concluídas as cinco iterações, calcula-se a média e o desvio padrão das métricas obtidas, fornecendo uma estimativa mais estável e estatisticamente confiável do desempenho real de cada modelo. Esse procedimento, combinado à estratificação, mitiga simultaneamente o risco de enviesamento decorrente de uma única separação treino–teste e o risco de que distribuições desequilibradas de classes em algum *fold* distorçam as estimativas de desempenho, fortalecendo a validade metodológica do estudo.

4.4. Configuração de Treinamento

O treinamento dos modelos foi conduzido seguindo uma configuração padronizada de hiperparâmetros e estratégias de otimização adequadas ao cenário de *transfer learning*. Todos os experimentos foram realizados utilizando as arquiteturas GoogLeNet, ResNet18 e DenseNet121, previamente treinadas no ImageNet e posteriormente ajustadas ao conjunto de imagens orais empregado neste estudo. Os valores específicos adotados para cada hiperparâmetro encontram-se resumidos na Tabela 1.

Tabela 1: Hiperparâmetros usados no treinamento

REMUNOM
ISSN 2178-6925

Received: 19/04/2026 - **Accepted:** 06/05/2026
Vol: 13.09
DOI: 10.66104/eecb2s81
Pages: 1-34

Parâmetro	Valor
Batch size	32
Number of epochs	30
Initial learning rate	1.5×10^{-5}
Learning rate step size	10 epochs
Decay factor	0.1
Cross-validation	5-Fold StratifiedKFold

Fonte: Elaborado pelos autores (2026).

A função de perda adotada foi CrossEntropyLoss, apropriada para tarefas de classificação binária baseadas em probabilidades. O otimizador escolhido foi Adam, devido à sua estabilidade e eficiência durante o processo de *fine-tuning* de modelos pré-treinados. A *initial learning rate* foi definida como 1,5 × 10^{-5}, um valor deliberadamente baixo para evitar modificações abruptas nos pesos derivados do ImageNet.

Para regular o ritmo de aprendizado ao longo do treinamento, utilizou-se o agendador *StepLR*, configurado conforme os hiperparâmetros apresentados na Tabela 1. Nesse esquema, a *learning rate* é reduzida em um fator de *decay* igual a 0,1 a cada 10 épocas, estratégia que permite atualizações mais sutis nas fases finais do treinamento e favorece a convergência estável dos modelos.

Em cada fold da validação cruzada, o modelo foi treinado durante 30 épocas, e os pesos associados ao melhor desempenho no conjunto de validação foram preservados. Esse procedimento evita o uso de estados sobreajustados (*overfitted*) ao final da última época, assegurando que a avaliação final reflita o melhor ponto de generalização obtido durante o processo de treinamento.

5. Resultados

A Tabela 2 apresenta o desempenho médio das arquiteturas avaliadas ao longo dos cinco *folds* da validação cruzada.

Tabela 2: Desempenho médio das arquiteturas avaliadas

Modelo	Acurácia	Precisão	Recall	Especificidade	F1	AUC
DenseNet121	$\mathbf{0.944 \pm 0.021}$	$\mathbf{0.938 \pm 0.023}$	$\mathbf{0.958 \pm 0.020}$	$\mathbf{0.929 \pm 0.027}$	$\mathbf{0.948 \pm 0.020}$	$\mathbf{0.985 \pm 0.007}$
ResNet18	0.933 ± 0.023	0.934 ± 0.022	0.938 ± 0.027	0.927 ± 0.025	0.936 ± 0.022	0.979 ± 0.011
GoogLeNet	0.932 ± 0.022	0.936 ± 0.027	0.934 ± 0.023	0.929 ± 0.031	0.935 ± 0.021	0.972 ± 0.014

Fonte: Elaborado pelos autores (2026).

Os resultados obtidos evidenciam que as três arquiteturas avaliadas

REMUNOM
ISSN 2178-6925

Received: 19/04/2026 - **Accepted:** 06/05/2026
Vol: 13.09
DOI: 10.66104/eecb2s81
Pages: 1-34

alcançaram métricas médias superiores a 0,92 em todas as medidas avaliadas - acurácia, precisão, recall, especificidade, F1-score e AUC. Observa-se, contudo, que os intervalos definidos pelo desvio-padrão se sobrepõem entre os modelos para a maioria das métricas, por exemplo, as acurácias de DenseNet121 (0,944 ± 0,021), ResNet18 (0,933 ± 0,023) e GoogLeNet (0,932 ± 0,022) compartilham faixas de variação que se cruzam, o que impede afirmar categoricamente que há diferença de desempenho estatisticamente significativa entre os modelos com base nos dados disponíveis.

A interpretação dos resultados deve, portanto, ser conduzida com cautela: as médias reportadas refletem o desempenho observado neste protocolo experimental específico, não constituindo evidência de superioridade absoluta de uma arquitetura sobre as demais.

Nesse contexto, o que os dados permitem afirmar é que todas as arquiteturas demonstraram viabilidade para a tarefa proposta. Valores de precisão acima de 0,92 em todos os modelos são consistentes com faixas reportadas em estudos de classificação de imagens médicas com conjuntos de dados de escala similar (Litjens *et al.*, 2017; Fawcett, 2006; Sokolova; Lapalme, 2009), o que aponta para a aplicabilidade geral da abordagem baseada em *transfer learning* com arquiteturas de baixo custo computacional neste domínio. As diferenças observadas entre os modelos são melhor interpretadas em termos de padrões de variância do que de hierarquia de desempenho.

A DenseNet121 registrou as médias mais altas em todas as métricas: acurácia (0,944 ± 0,021), precisão (0,938 ± 0,023), *recall* (0,958 ± 0,020), especificidade (0,929 ± 0,027), F1-score (0,948 ± 0,020) e AUC (0,985 ± 0,007). Além disso, apresentou os menores desvios-padrão em *recall* e AUC entre os três modelos, o que sugere que seu comportamento foi mais estável entre as partições estratificadas - característica possivelmente associada à sua estrutura de conexões densas, que favorece o reuso de características. A ResNet18 apresentou o maior desvio-padrão de *recall* (± 0,27), sugerindo maior sensibilidade à composição de cada *fold* nessa métrica específica, com especificidade de (0,927 ± 0,025). A GoogLeNet registrou o maior desvio-padrão em precisão (± 0,027), indicando variabilidade maior nessa dimensão entre as partições, e especificidade de (0,929

REMUNOM
ISSN 2178-6925

Received: 19/04/2026 - **Accepted:** 06/05/2026
Vol: 13.09
DOI: 10.66104/eecb2s81
Pages: 1-34

± 0,031). Essas diferenças de variância são relevantes para caracterizar o comportamento de cada arquitetura, ainda que não permitam conclusões sobre superioridade estatística.

5.1. Evolução da Acurácia Durante o Treinamento

As Figuras 2 e 3 apresentam, respectivamente, a evolução da acurácia média no conjunto de treinamento e no conjunto de validação.

Em ambos os gráficos, cada curva representa uma arquitetura distinta dentre as três avaliadas neste estudo: a curva azul corresponde ao modelo GoogLeNet, a curva vermelha representa a ResNet18 e a curva verde refere-se à DenseNet121. Essas cores são utilizadas consistentemente nos dois gráficos para facilitar a comparação direta entre as arquiteturas ao longo das épocas.

Cada ponto das curvas indica o valor médio de acurácia (μ) obtido a partir das cinco execuções provenientes da validação cruzada (5-Fold). Assim, a linha central de cada cor expressa a acurácia esperada do modelo em cada época, uma vez que suaviza oscilações individuais geradas por variações específicas de cada fold. Além disso, barras verticais de erro são exibidas ao redor de cada ponto das curvas. Essas barras representam o desvio-padrão (σ) calculado entre os cinco valores obtidos em cada época, funcionando como um intervalo de confiança simplificado que indica o grau de variação estatística entre diferentes divisões dos dados.

Valores reduzidos de σ (barras curtas) refletem alta estabilidade e invariância do modelo, significando que seu desempenho é consistente independentemente do subconjunto utilizado para validação. Por outro lado, barras mais longas indicam maior sensibilidade à divisão dos dados, sugerindo que o desempenho do modelo apresenta maior flutuação entre os folds.

5.1.1. Acurácia Média no Conjunto de Treinamento

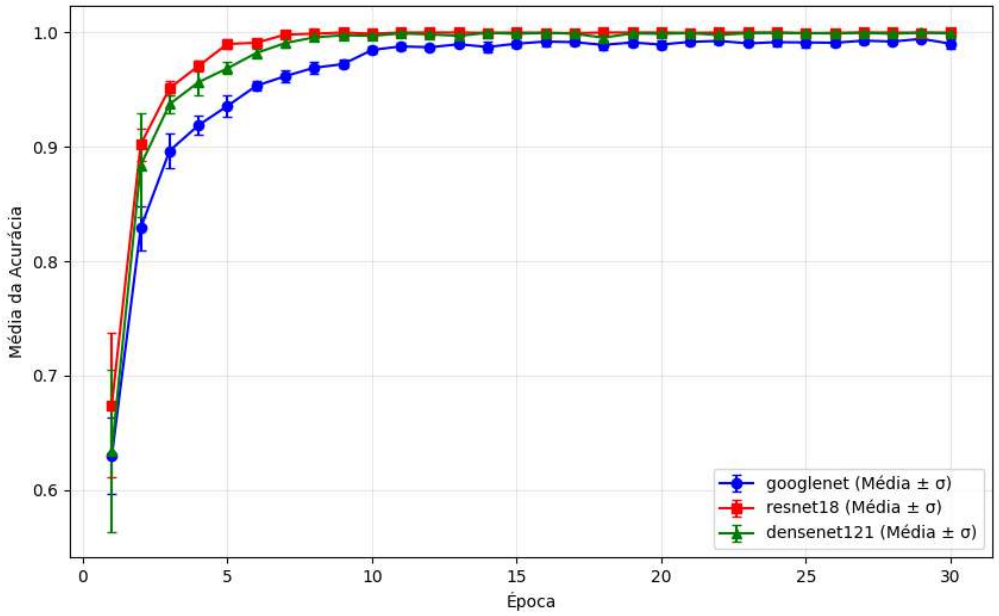
Na Figura 2, observa-se que todas as arquiteturas apresentam curvas ascendentemente suaves, com acréscimos mais pronunciados nas primeiras épocas. A curva vermelha (ResNet18) e a curva verde (DenseNet121) destacam-se por apresentar crescimento mais acelerado nas épocas iniciais, convergindo

REMUNOM
ISSN 2178-6925
Received: 19/04/2026 - **Accepted:** 06/05/2026
Vol: 13.09
DOI: 10.66104/eecb2s81
Pages: 1-34

para valores próximos de 1,0 a partir da época 15, de forma praticamente indistinguível. A curva azul (GoogLeNet) apresenta evolução similar, porém com valores iniciais mais baixos e convergência ligeiramente mais lenta, alcançando as demais na fase final do treinamento.

As barras de erro neste gráfico são visivelmente curtas para todas as arquiteturas, indicando que, independentemente do fold utilizado, os modelos apresentam comportamento muito semelhante durante o treinamento. Esse resultado demonstra que o processo de aprendizagem foi estável e pouco afetado pelas variações na partição dos dados. Além disso, a ausência de oscilações abruptas nas curvas e a redução progressiva do desvio-padrão ao longo das épocas sugerem um processo de otimização controlado, sem indícios de flutuação ou instabilidade numérica.

Figura 2: Evolução da acurácia média de treino (com desvio-padrão) ao longo das épocas para cada arquitetura.



Fonte: Elaborado pelos autores (2026).

5.1.2. Acurácia Média no Conjunto de Validação

A Figura 3, que apresenta a evolução da acurácia no conjunto de validação, revela um comportamento mais variável em comparação ao conjunto de treinamento. Essa maior variabilidade é esperada, uma vez que cada fold contém menos amostras e apresenta diferenças naturais relacionadas à distribuição das classes e à diversidade visual das imagens orais.

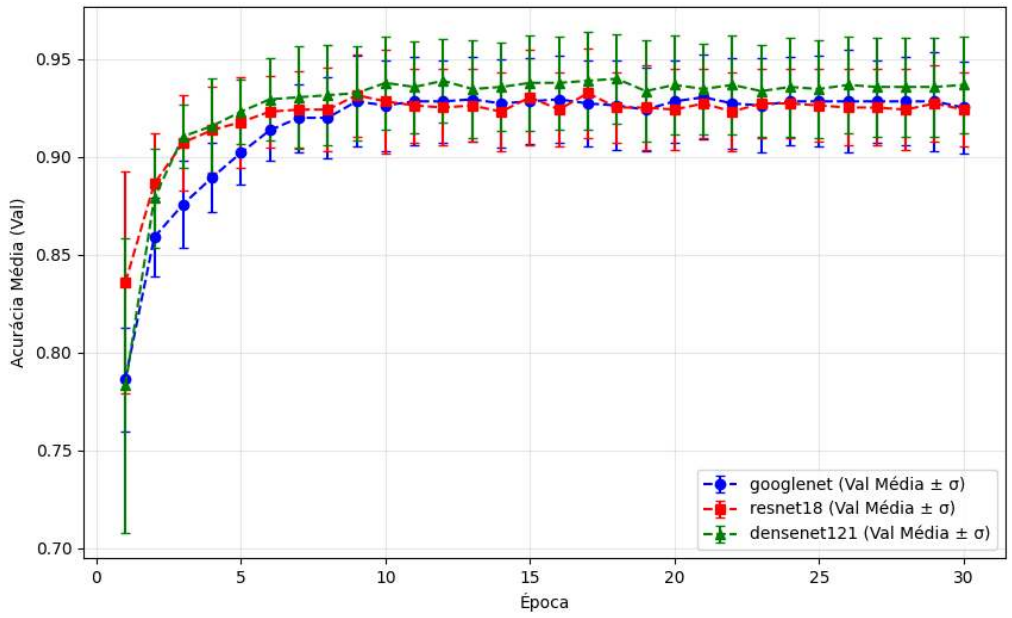
REMUNOM
ISSN 2178-6925

Received: 19/04/2026 - **Accepted:** 06/05/2026
Vol: 13.09
DOI: 10.66104/eecb2s81
Pages: 1-34

Ainda assim, observa-se que a curva verde (DenseNet121) registra, de forma consistente, as maiores médias pontuais de acurácia de validação ao longo das épocas, destacando-se visivelmente das demais arquiteturas, especialmente a partir das épocas intermediárias. As curvas vermelha (ResNet18) e azul (GoogLeNet), por sua vez, apresentam trajetórias bastante próximas entre si ao longo de todo o treinamento, com médias sistematicamente inferiores à DenseNet121 e sobreposição considerável de suas barras de erro, o que indica desempenho equivalente entre essas duas arquiteturas no conjunto de validação.

As barras de erro da validação fornecem informação crítica sobre a capacidade de generalização das arquiteturas. A DenseNet121 apresenta os menores valores de σ nas épocas finais em recall (±0,020) e AUC (±0,007) (conforme Tabela 2), sugerindo não apenas maior acurácia média, mas também maior estabilidade entre os folds. A ResNet18, apesar de apresentar desempenho próximo ao da GoogLeNet, exibe o maior desvio-padrão de recall entre as três arquiteturas (±0,037), evidenciando que sua identificação da classe positiva é mais sensível à composição específica de cada partição. A GoogLeNet mantém variância moderada em recall (±0,023), porém registra o maior desvio-padrão em especificidade (±0,031), indicando instabilidade na identificação da classe negativa entre os folds.

Figura 3: Evolução da acurácia média de validação (com desvio-padrão) ao longo das épocas para cada arquitetura.



REMUNOM
ISSN 2178-6925

Received: 19/04/2026 - **Accepted:** 06/05/2026
Vol: 13.09
DOI: 10.66104/eecb2s81
Pages: 1-34

Fonte: Elaborado pelos autores (2026).

5.1.3. Análise Comparativa entre Treinamento e Validação

Quando as curvas de treinamento e validação são comparadas, nota-se que todas as arquiteturas apresentam acurácia de treinamento superior à de validação, comportamento esperado em modelos baseados em aprendizado supervisionado. No entanto, a diferença entre ambas permanece relativamente estável ao longo das épocas, sugerindo que não houve sobreajuste severo. As três arquiteturas convergem para valores próximos de 1,0 no treinamento, enquanto no conjunto de validação os valores se estabilizam em patamares inferiores, com as curvas apresentando maior variabilidade entre os folds.

As análises estatísticas fornecidas pela média (μ) e pelo desvio-padrão (σ) demonstram que, embora a DenseNet121 registre os menores desvios-padrão em recall (±0,020) e AUC (±0,007), a sobreposição dos intervalos de variação entre as três arquiteturas impede inferências categóricas sobre superioridade de generalização. A ResNet18 revela maior variância no recall (±0,027), sugerindo que a composição específica de cada fold afeta de forma mais pronunciada sua capacidade de detectar casos positivos. A GoogLeNet, por sua vez, registra o maior desvio-padrão em especificidade (±0,031), indicando maior sensibilidade às partições na identificação da classe negativa.

Os gráficos de acurácia revelam que as arquiteturas convergiram para acurácia de treinamento próxima a 1,0 a partir da época 15 e apresentam desempenho de validação equivalente, dado que os intervalos de desvio-padrão se sobrepõem de forma considerável. As diferenças de variância entre os modelos são, portanto, mais informativas do que o ordenamento das médias pontuais para caracterizar o comportamento de generalização de cada arquitetura.

5.2. Análise das Matrizes de Confusão

As Figuras 4, 5 e 6 apresentam as matrizes de confusão obtidas a partir da média dos resultados de cada arquitetura ao longo do processo de cross-validation. Para construir essas matrizes finais, foram geradas cinco matrizes de confusão individuais por modelo, uma por fold, a partir da avaliação do modelo com os melhores pesos de validação de cada iteração. Para cada célula da matriz (TP, FN, FP e TN), calculou-se a média aritmética e o desvio-padrão (±σ) dos cinco

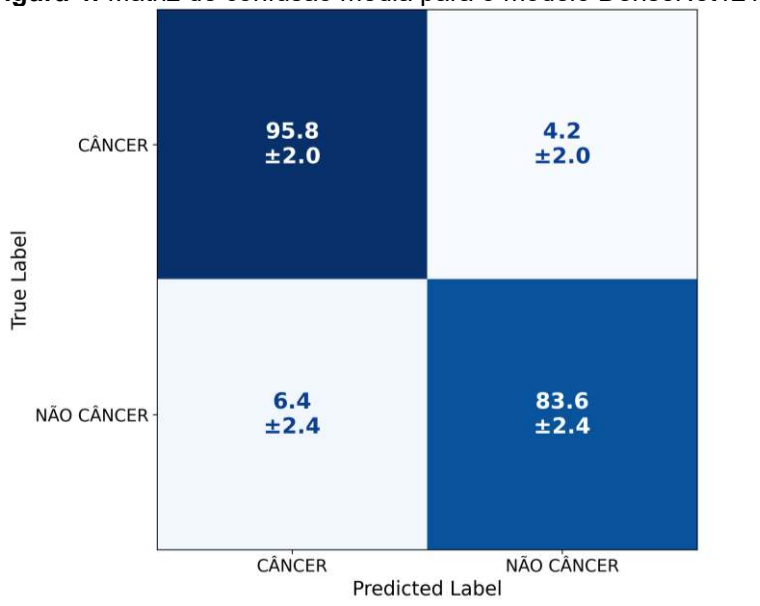
REMUNOM
ISSN 2178-6925
Received: 19/04/2026 - **Accepted:** 06/05/2026
Vol: 13.09
DOI: 10.66104/eecb2s81
Pages: 1-34

valores obtidos, sem arredondamento. Esse formato permite não apenas sintetizar o comportamento global de cada classificador, mas também visualizar diretamente a variabilidade das predições entre as diferentes partições estratificadas.

5.2.1. DenseNet121

A matriz de confusão média da DenseNet121 (Figura 4) revela um desempenho fortemente equilibrado entre as duas classes. O modelo obteve em média 95,8 ± 2,0 verdadeiros positivos e 83,6 ± 2,4 verdadeiros negativos, indicando alta capacidade de reconhecer corretamente imagens de ambas as classes. Os valores médios de FN (4,2 ± 2,0) e FP (6,4 ± 2,4) são os menores entre as três arquiteturas avaliadas em termos de FN absoluto, corroborando os resultados quantitativos da Tabela 2 e evidenciando um modelo simultaneamente sensível e preciso. A combinação desses fatores explica a liderança da DenseNet121 em recall, precisão e F1-score.

Figura 4: Matriz de confusão média para o modelo DenseNet121.



Fonte: Elaborado pelos autores (2026).

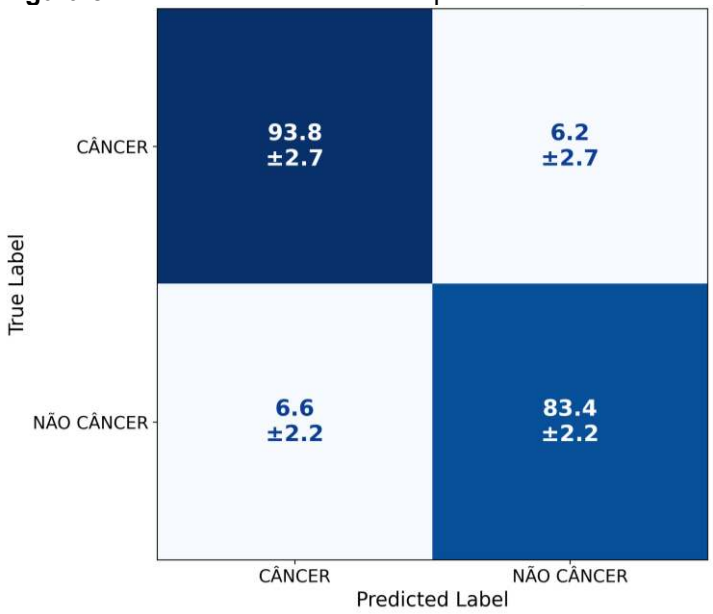
5.2.2. ResNet18

A ResNet18 (Figura 5) apresentou 93,8 ± 2,7 verdadeiros positivos e 83,4 ± 2,2 verdadeiros negativos. Com falso negativo médio de 6,2 ± 2,7 e falso positivo médio de 6,6 ± 2,2, o modelo exibe desempenho ligeiramente inferior à DenseNet121 na detecção da classe positiva, porém muito próximo ao da GoogLeNet em termos absolutos.

REMUNOM
ISSN 2178-6925

Received: 19/04/2026 - **Accepted:** 06/05/2026
Vol: 13.09
DOI: 10.66104/eecb2s81
Pages: 1-34

Figura 5: Matriz de confusão média para o modelo ResNet18.



Fonte: Elaborado pelos autores (2026).

Destaca-se que, nesta arquitetura, o desvio-padrão de FP (±2,2) é o menor entre os três modelos, indicando que a taxa de falsos positivos da ResNet18 é a mais estável entre as partições. A variância de FN (±2,7), por sua vez, é a maior entre os três modelos, comportamento consistente com o maior desvio-padrão de recall observado na Tabela 2 (±0,037).

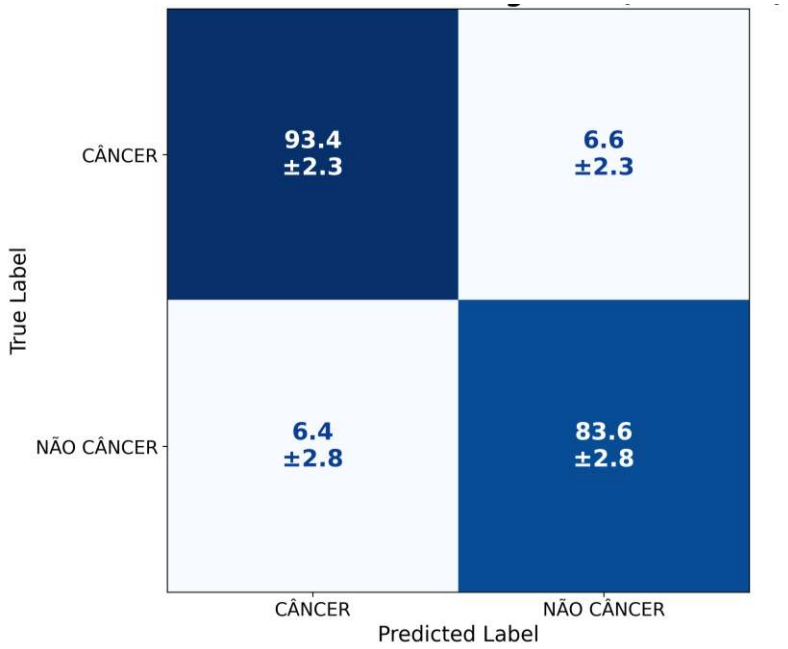
5.2.3. GoogLeNet

A GoogLeNet (Figura 6) registrou 93,4 ± 2,3 verdadeiros positivos e 83,6 ± 2,8 verdadeiros negativos. Com FN médio de 6,6 ± 2,3 e FP médio de 6,4 ± 2,8, a arquitetura apresentou o maior valor absoluto de FN entre as três arquiteturas, explicando seu menor recall geral. Destaca-se que a GoogLeNet registrou o maior desvio-padrão de FP (±2,8) entre os três modelos, indicando que sua taxa de falsos positivos é a mais sensível à composição específica de cada partição, comportamento consistente com o maior desvio-padrão de especificidade observado na Tabela 2 (±0,031).

Figura 6: Matriz de confusão média para o modelo GoogLeNet.

REMUNOM
ISSN 2178-6925

Received: 19/04/2026 - **Accepted:** 06/05/2026
Vol: 13.09
DOI: 10.66104/eecb2s81
Pages: 1-34



Fonte: Elaborado pelos autores (2026).

5.2.4. Discussão Geral

A análise comparativa das matrizes mostra que os três modelos atingiram desempenho sólido, com predominância de acertos em ambas as classes e variações relativamente pequenas entre TP e TN. No entanto, observam-se diferenças sutis que explicam o ranking final das métricas:

- A DenseNet121 apresentou o menor FN médio (4,2 ± 2,0) entre as três arquiteturas, justificando sua liderança em recall e F1-score com maior consistência entre os folds;
- A ResNet18 registrou FN médio de 6,2 ± 2,7, com a maior variância nessa célula entre os modelos, indicando maior instabilidade na detecção da classe positiva entre as partições estratificadas;
- A GoogLeNet apresentou o maior valor absoluto de FN (6,6 ± 2,3) e o maior desvio-padrão de FP (±2,8) entre as três arquiteturas, refletindo menor capacidade discriminativa global e maior sensibilidade das predições negativas à composição de cada fold.

Os três modelos apresentaram perfis de erro semelhantes em termos absolutos, com diferenças de apenas 1 a 2 unidades nas células fora da diagonal principal. Essas variações, em conjunto com a sobreposição dos desvios-padrão entre modelos, reforçam a interpretação de que as arquiteturas avaliadas exibem

REMUNOM
ISSN 2178-6925

Received: 19/04/2026 - **Accepted:** 06/05/2026
Vol: 13.09
DOI: 10.66104/eecb2s81
Pages: 1-34

capacidade discriminativa equivalente neste protocolo experimental, diferindo principalmente nos padrões de variância entre folds.

Do ponto de vista clínico, contudo, falsos negativos e falsos positivos não têm custo equivalente. Um falso negativo - classificar como saudável uma imagem de lesão suspeita - representa o risco de que um caso potencialmente maligno deixe de ser encaminhado para avaliação especializada, com consequências diretas sobre o prognóstico do paciente. Um falso positivo - classificar como suspeita uma imagem saudável - implica, no pior caso, um encaminhamento desnecessário, situação clinicamente gerenciável e de impacto consideravelmente menor.

Nesse contexto, o recall (sensibilidade para a classe positiva) é a métrica de maior relevância clínica, pois quantifica diretamente a proporção de casos suspeitos corretamente identificados. Importa destacar, no entanto, que o modelo proposto não se destina a substituir a avaliação profissional, mas a atuar como ferramenta de triagem auxiliar: nos casos classificados como suspeitos, a decisão diagnóstica permanece sob responsabilidade do profissional de saúde, o que atenua o impacto de eventuais falsos positivos e reforça a adequação de priorizar a minimização de falsos negativos como critério de desempenho. Sob essa perspectiva, a DenseNet121, que registrou o menor FN médio (4,2 ± 2,0) e o menor desvio-padrão dessa célula entre os três modelos, apresenta o perfil mais alinhado às exigências de um sistema de apoio ao diagnóstico orientado à segurança do paciente.

A especificidade - proporção de imagens saudáveis corretamente identificadas, calculada como TN / (TN + FP) a partir das matrizes de confusão médias - manteve-se elevada nos três modelos: DenseNet121 (0,929 ± 0,027), ResNet18 (0,927 ± 0,025) e GoogLeNet (0,929 ± 0,031), indicando que nenhuma das arquiteturas apresenta tendência sistemática a gerar falsos alarmes excessivos.

Além disso, para fortalecer a transparência e a reprodutibilidade dos experimentos, todo o código desenvolvido foi disponibilizado publicamente em: https://github.com/Isabellybrt/CancerOral-CNNs.

5.3. Análise Comparativa do AUC-ROC entre os Folds

REMUNOM
ISSN 2178-6925

Received: 19/04/2026 - **Accepted:** 06/05/2026
Vol: 13.09
DOI: 10.66104/eecb2s81
Pages: 1-34

As Figuras 7 e 8 apresentam uma análise complementar do desempenho dos modelos com base na métrica AUC-ROC. Enquanto a Figura 7 mostra a comparação entre os modelos considerando a média e o desvio padrão ao longo dos cinco folds, a Figura 8 ilustra a variação da AUC em cada partição individual da validação cruzada.

Observa-se que a DenseNet121 apresentou a maior média de AUC (0,985 ± 0,007), seguida pela ResNet18 (0,979 ± 0,011) e pela GoogLeNet (0,972 ± 0,014). No entanto, a sobreposição dos intervalos definidos pelo desvio-padrão indica que não é possível afirmar diferença estatisticamente significativa entre os modelos com base nos dados disponíveis.

As barras de erro reforçam a estabilidade dos modelos, indicando que o desempenho se mantém consistente mesmo diante das variações introduzidas pelas diferentes partições da validação cruzada.

A análise por fold evidencia que as três arquiteturas mantêm AUC superior a 0,95 em todas as partições, com variações entre folds inferiores a 0,03. Nota-se que a DenseNet121 tende a apresentar valores de AUC ligeiramente superiores na maioria dos folds, enquanto a ResNet18 exibe comportamento intermediário e a GoogLeNet apresenta maior dispersão em alguns casos. Ainda assim, essas diferenças devem ser interpretadas com cautela, uma vez que a variabilidade observada entre os folds limita inferências conclusivas sobre superioridade entre os modelos.

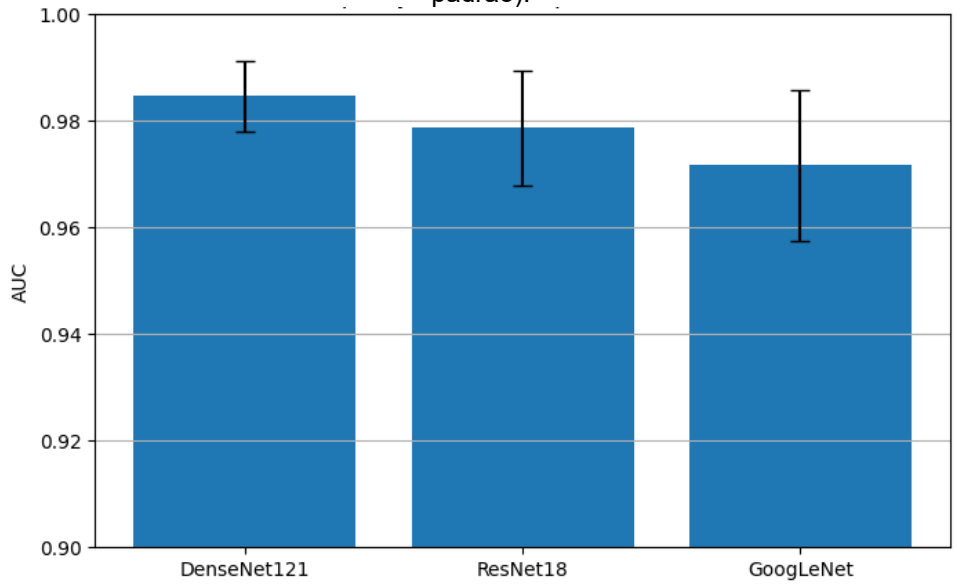
No caso da Figura 7, que apresenta a média da AUC com desvio-padrão, observa-se que esse tipo de visualização permite uma análise mais agregada do desempenho dos modelos. A principal vantagem desse gráfico é sintetizar, em uma única representação, tanto o desempenho médio quanto a estabilidade estatística de cada arquitetura. As barras centrais refletem o valor esperado da AUC ao longo dos folds, enquanto as barras de erro (±σ) fornecem uma estimativa da dispersão desses valores. Nota-se que todas as arquiteturas apresentam desvios-padrão relativamente baixos, o que indica consistência no comportamento dos modelos independentemente da partição dos dados. A DenseNet121 apresenta a maior média pontual e o menor desvio-padrão, seguida pela ResNet18, cujos intervalos se sobrepõem parcialmente aos da DenseNet121,

REMUNOM
ISSN 2178-6925

Received: 19/04/2026 - **Accepted:** 06/05/2026
Vol: 13.09
DOI: 10.66104/eecb2s81
Pages: 1-34

sugerindo desempenho próximo entre essas duas arquiteturas. A GoogLeNet, embora também opere em patamar elevado, apresenta a menor média e o maior desvio-padrão entre os três modelos, indicando maior variabilidade no desempenho entre as partições.

Figura 7: Comparação do desempenho dos modelos com base na AUC-ROC (média ± desvio padrão).



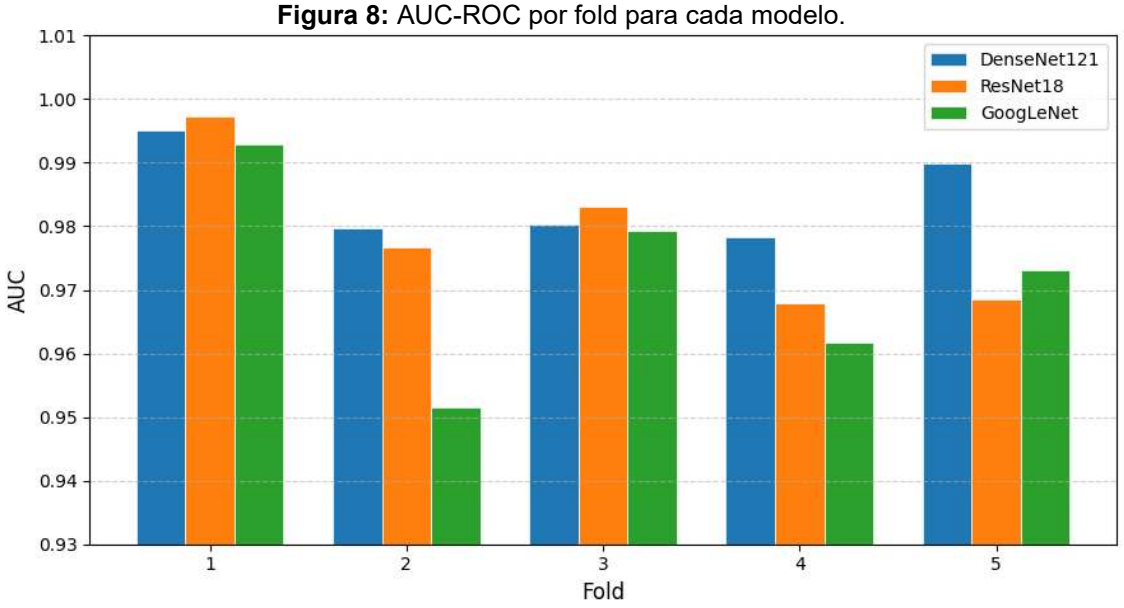
Fonte: Elaborado pelos autores (2026).

A Figura 8 oferece uma perspectiva mais detalhada e granular do comportamento dos modelos, ao apresentar explicitamente os valores de AUC obtidos em cada fold. Diferentemente do gráfico anterior, essa visualização permite identificar variações específicas entre as partições, evidenciando como cada arquitetura responde a diferentes subconjuntos do dataset. Observa-se que, embora todos os modelos apresentem AUC superior a 0,95 em todas as partições, há variações entre folds que refletem a sensibilidade às diferenças naturais de composição de cada subconjunto. A DenseNet121 registra os valores mais homogêneos entre os folds, indicando maior consistência no desempenho. A GoogLeNet apresenta a maior amplitude de variação, com o valor mais baixo no fold 2 (0,9516). A ResNet18 mantém comportamento intermediário, com valores mais reduzidos nos folds 4 e 5. Esse tipo de análise é particularmente relevante, pois evidencia que o desempenho não é uniforme em todas as divisões, reforçando a importância da validação cruzada para uma avaliação mais robusta.

REMUNOM
ISSN 2178-6925

Received: 19/04/2026 - **Accepted:** 06/05/2026
Vol: 13.09
DOI: 10.66104/eecb2s81
Pages: 1-34

Figura 8: AUC-ROC por fold para cada modelo.



Fonte: Elaborado pelos autores (2026).

De modo geral, os resultados reforçam que todas as arquiteturas apresentam elevado poder discriminativo (AUC superior a 0,97), diferindo principalmente em padrões de variabilidade, e não em desempenho estatisticamente distinto.

6. Discussão

A análise comparativa das arquiteturas avaliadas evidencia diferenças importantes na forma como cada modelo aprende e generaliza as características presentes nas imagens orais. Entre as redes estudadas, a DenseNet121 registrou menor variância em recall e AUC entre os folds, comportamento que pode estar associado à sua estrutura baseada em conexões densas, que promovem reutilização de mapas de características e facilitam o fluxo de gradientes ao longo da rede. Esse mecanismo se mostra particularmente vantajoso em conjuntos de dados reduzidos, como o utilizado neste estudo, pois reduz redundâncias internas e melhora a eficiência na captura de padrões sutis associados às lesões.

A GoogLeNet, por sua vez, apresentou desempenho equilibrado e estável, beneficiando-se da heterogeneidade dos blocos Inception, que permitem diferentes escalas de processamento dentro de uma mesma profundidade. Essa característica favorece a extração de detalhes que variam em granularidade, o que pode explicar sua boa estabilidade mesmo diante da variabilidade das imagens

REMUNOM
ISSN 2178-6925

Received: 19/04/2026 - **Accepted:** 06/05/2026
Vol: 13.09
DOI: 10.66104/eecb2s81
Pages: 1-34

orais. Já a ResNet18, embora mais enxuta, mostrou vantagens específicas, sobretudo na contenção de predições equivocadas da classe positiva, o que está relacionado ao uso de conexões residuais que facilitam a aprendizagem de mapeamentos mais simples e evitam degradação do desempenho em redes mais profundas.

A escolha de arquiteturas clássicas em detrimento de modelos mais recentes responde a uma lacuna específica da literatura: a ausência de comparações sistemáticas, metodologicamente controladas e reprodutíveis entre arquiteturas de baixo custo computacional aplicadas a imagens clínicas orais. Estudos recentes na área tendem a adotar modelos híbridos ou multimodais (Devindi *et al.*, 2024), que pressupõem infraestrutura de coleta e poder computacional indisponíveis na maioria dos cenários de triagem de baixo custo. Nesse contexto, estabelecer o comportamento de arquiteturas enxutas e acessíveis constitui uma etapa anterior e necessária: sem essa linha de base, não é possível quantificar o ganho real que arquiteturas mais complexas proporcionam. Adicionalmente, a disponibilidade pública dos pesos e das implementações utilizadas neste trabalho permite que os resultados aqui obtidos sirvam diretamente como referência comparativa para estudos futuros, independentemente do ambiente computacional disponível.

Outro ponto relevante diz respeito à estratégia metodológica adotada. A ausência intencional de técnicas de data augmentation permitiu uma avaliação mais isolada da capacidade intrínseca das arquiteturas, eliminando interferências de manipulações artificiais das imagens e possibilitando uma comparação mais precisa entre os modelos. Embora essa escolha limite o potencial máximo de generalização, ela proporciona um cenário experimental mais controlado, especialmente útil para compreender o comportamento estrutural de cada rede.

A validação cruzada aplicada garantiu maior robustez estatística às análises, reduzindo o impacto de flutuações decorrentes da divisão dos dados. A consistência observada entre os folds indica que os modelos responderam de forma previsível ao conjunto de treinamento, reforçando a credibilidade dos achados.

Por fim, as matrizes de confusão complementam essa interpretação,

REMUNOM
ISSN 2178-6925
Received: 19/04/2026 - **Accepted:** 06/05/2026
Vol: 13.09
DOI: 10.66104/eecb2s81
Pages: 1-34

revelando que todas as arquiteturas apresentaram padrões semelhantes de acertos e erros. Mesmo com pequenas variações entre elas, observa-se que os modelos foram mais eficientes em identificar casos negativos do que positivos. Essa tendência reflete desafios inerentes ao conjunto de dados, como a heterogeneidade visual das lesões e o reduzido número de amostras, fatores que podem dificultar a identificação de padrões associados à classe suspeita. Ainda assim, o comportamento geral indica capacidade discriminativa relevante e potencial aplicação em cenários reais de apoio ao diagnóstico.

6.1. Da prova de conceito à adoção clínica

É necessário distinguir com clareza três níveis de maturidade tecnológica que frequentemente se confundem no discurso sobre modelos de aprendizado profundo em saúde: desempenho experimental, potencial tecnológico e aplicabilidade clínica validada.

Este trabalho situa-se no primeiro nível. Os modelos foram avaliados em um cenário experimental controlado, com dataset único, sem validação externa, sem avaliação multicêntrica, sem teste em fluxos reais de triagem e sem análise de impacto clínico. Esse posicionamento não diminui a contribuição do trabalho, estudos nesse estágio compõem o substrato científico necessário para que o campo avance, mas impõe restrições precisas sobre o que os resultados permitem inferir.

A trajetória entre uma prova de conceito e uma solução clinicamente adotada envolve, de forma não exaustiva: (i) validação em datasets independentes com confirmação histopatológica; (ii) testes multicêntricos com variação de dispositivos e condições de aquisição; (iii) avaliação prospectiva com integração em fluxos reais de triagem; (iv) análise de interpretabilidade operacional, como mapas de ativação (Grad-CAM), essencial para a confiança do profissional de saúde; e (v) conformidade com requisitos regulatórios para dispositivos de diagnóstico assistido.

Dentre essas lacunas, a ausência de interpretabilidade merece reconhecimento explícito como limitação do presente trabalho. Sem evidência das regiões de ativação que fundamentam as predições de cada modelo - obtível, por exemplo, por meio de Grad-CAM -, não é possível aferir se os modelos

REMUNOM
ISSN 2178-6925

Received: 19/04/2026 - **Accepted:** 06/05/2026
Vol: 13.09
DOI: 10.66104/eecb2s81
Pages: 1-34

aprenderam padrões morfológicos clinicamente plausíveis, como bordas irregulares, alterações de coloração ou textura associadas às lesões, ou se suas predições estão ancoradas em artefatos contextuais da imagem, como condições de iluminação, posicionamento do dispositivo ou características de fundo. Em um conjunto heterogêneo e sem confirmação histológica, essa incerteza é particularmente significativa: métricas agregadas elevadas não garantem que o modelo esteja processando as regiões clinicamente relevantes da imagem.

Estas etapas não foram percorridas neste trabalho. O que este estudo oferece é uma base metodológica reprodutível e evidências quantitativas comparativas entre arquiteturas, contribuindo com um elo específico em uma cadeia científica mais ampla. A eventual viabilização de soluções clínicas de uso geral não decorre de um trabalho isolado, mas do amadurecimento coletivo do conhecimento que advém do conjunto de pesquisas na área.

7. Conclusões e Considerações finais

Este trabalho comparou o desempenho de diferentes arquiteturas de redes convolucionais aplicadas à detecção de lesões orais por meio de imagens clínicas, com vistas a identificar qual modelo apresenta o perfil mais adequado para compor futuramente um sistema computacional de apoio ao diagnóstico. Os resultados obtidos demonstraram que todas as redes apresentaram desempenho sólido, com todas as arquiteturas registrando médias superiores a 0,92 em todas as métricas avaliadas, resultado que aponta para a viabilidade geral da abordagem. A DenseNet121 registrou as maiores médias pontuais e os menores desvios-padrão em recall e AUC - F1-score de 0,948 ± 0,020 -, embora a sobreposição dos intervalos de variação entre os modelos não permita afirmar diferença de desempenho estatisticamente significativa - conjunto de características diretamente relevante em aplicações clínicas, onde identificar corretamente casos positivos tem peso maior do que a acurácia global isolada. Importa reiterar que esses resultados referem-se à classificação de imagens segundo rótulos definidos no dataset original, não constituindo, por si só, evidência de desempenho clínico para detecção de câncer oral confirmado histopatologicamente.

As demais arquiteturas, GoogLeNet e ResNet18, também apresentaram características favoráveis e complementares, contribuindo para uma compreensão

REMUNOM
ISSN 2178-6925

Received: 19/04/2026 - **Accepted:** 06/05/2026
Vol: 13.09
DOI: 10.66104/eecb2s81
Pages: 1-34

mais ampla sobre como diferentes estruturas de rede respondem ao mesmo problema e oferecendo alternativas para cenários com requisitos distintos.

Os achados reforçam o potencial das CNNs como ferramentas auxiliares na triagem e avaliação inicial de lesões orais, especialmente em ambientes clínicos onde a análise visual desempenha papel determinante na tomada de decisão. Ainda que o conjunto de dados apresente limitações em escala e variabilidade, os modelos foram capazes de aprender padrões relevantes e demonstraram capacidade de generalização compatível com o desafio proposto, indicando que, com expansão, validação clínica multicêntrica e avaliação prospectiva, essa base pode vir a sustentar investigações voltadas para cenários reais - etapas ainda não realizadas neste trabalho.

Para que esse potencial se converta em uma solução clínica concreta, os próximos passos incluem:

- incorporar técnicas robustas de data augmentation, aplicadas corretamente após a partição dos dados, para ampliar a diversidade visual e reduzir sobreajuste;
- aplicar a mesma metodologia de avaliação - incluindo os hiperparâmetros, a estratégia de validação cruzada e as arquiteturas comparadas - a outros datasets de lesões orais, públicos ou clínicos, com confirmação histopatológica, de modo a verificar a estabilidade dos resultados e a capacidade de generalização dos modelos além do conjunto utilizado neste estudo;
- empregar métodos de interpretabilidade, como Grad-CAM, para identificar as regiões das imagens que efetivamente fundamentam as predições de cada modelo. A ausência dessa análise no presente trabalho constitui uma limitação relevante: sem evidência visual das regiões de ativação, não é possível aferir se os modelos aprenderam padrões morfológicos clinicamente plausíveis ou se suas predições estão ancoradas em artefatos contextuais da imagem, sendo essa análise prioritária nos desdobramentos desta pesquisa;
- expandir o conjunto de dados com imagens anotadas e validadas clinicamente, incorporando confirmação histológica;
- explorar arquiteturas contemporâneas, incluindo modelos baseados em Vision Transformers e abordagens híbridas, avaliando seu custo-benefício

REMUNOM
ISSN 2178-6925

Received: 19/04/2026 - **Accepted:** 06/05/2026
Vol: 13.09
DOI: 10.66104/eecb2s81
Pages: 1-34

frente às arquiteturas investigadas neste estudo.

Esses aprimoramentos poderão fortalecer o corpo de evidências científicas sobre o uso de aprendizado profundo na odontologia, contribuindo para que, com o amadurecimento progressivo do conhecimento na área, soluções de apoio ao diagnóstico mais precisas e interpretáveis possam ser investigadas em contextos clínicos controlados. Os resultados obtidos neste trabalho, aliados ao código publicamente disponível, posicionam esta pesquisa como um ponto de partida concreto para iniciativas que busquem aproximar o aprendizado profundo da prática odontológica cotidiana.

Agradecimentos

Os autores agradecem ao Instituto Federal do Piauí (IFPI), Campus Piripiri, pelo suporte institucional oferecido durante o desenvolvimento desta pesquisa. Agradecemos também pelos recursos computacionais disponibilizados, essenciais para a realização dos experimentos. Por fim, reconhecemos a contribuição dos desenvolvedores do conjunto de dados público utilizado neste estudo, bem como da comunidade de software livre, cujas ferramentas possibilitaram a implementação dos modelos avaliados neste trabalho.

Referências

BIANCO, S. et al. Benchmark analysis of representative deep neural network architectures. **IEEE Access**, v. 6, p. 64270–64277, 2018. DOI: 10.1109/ACCESS.2018.2877890.

DENG, J. et al. ImageNet: a large-scale hierarchical image database. *In*: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR), 2009, Miami. **Proceedings [...]**. Piscataway: IEEE, 2009. p. 248–255. DOI: 10.1109/CVPR.2009.5206848.

DEVINDI, G. A. I. et al. Multimodal deep convolutional neural network pipeline for AI-assisted early detection of oral cancer. **IEEE Access**, v. 12, 2024. DOI: 10.1109/ACCESS.2024.3454338.

FAWCETT, T. An introduction to ROC analysis. **Pattern Recognition Letters**, v. 27, n. 8, p. 861–874, 2006. DOI: 10.1016/j.patrec.2005.10.010.

REMUNOM
ISSN 2178-6925

Received: 19/04/2026 - **Accepted:** 06/05/2026
Vol: 13.09
DOI: 10.66104/eecb2s81
Pages: 1-34

FU, Q. et al. A deep learning algorithm for detection of oral cavity squamous cell carcinoma from photographic images: a retrospective study. **EClinicalMedicine**, v. 27, p. 100558, 2020. DOI: 10.1016/j.eclinm.2020.100558.

HE, K. et al. Deep residual learning for image recognition. *In*: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR), 2016, Las Vegas. **Proceedings [...]**. Piscataway: IEEE, 2016. p. 770–778. DOI: 10.1109/CVPR.2016.90.

HUANG, G. et al. Densely connected convolutional networks. *In*: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR), 2017, Honolulu. **Proceedings [...]**. Piscataway: IEEE, 2017. p. 4700–4708. DOI: 10.1109/CVPR.2017.243.

KARTHIKEYAN, B. et al. INCS: design and development of an oral cancer identification methodology based on improved neural classification scheme. *In*: INTERNATIONAL CONFERENCE ON SOFT COMPUTING FOR SECURITY APPLICATIONS (ICSCSA), 4., 2024, Salem. **Anais [...]**. Salem: IEEE, 2024. p. 411–416. DOI: 10.1109/ICSCSA64454.2024.00072.

LITJENS, G. et al. A survey on deep learning in medical image analysis. **Medical Image Analysis**, v. 42, p. 60–88, 2017. DOI: 10.1016/j.media.2017.07.005.

NOGUEIRA, M.; GOMES, E. Histopathological imaging dataset for oral cancer analysis: a study with a data leakage warning. *In*: INTERNATIONAL JOINT CONFERENCE ON BIOMEDICAL ENGINEERING SYSTEMS AND TECHNOLOGIES – BIOSIGNALS, 18., 2025. **Proceedings [...]**. [S. l.]: SciTePress, 2025. p. 811–818. DOI: 10.5220/0013382100003911.

ORMEÑO-ARRIAGADA, P. et al. Comparison of deep learning and machine learning architectures for early oral cancer diagnosis. **PeerJ Computer Science**, v. 12, e3468, 2026. DOI: 10.7717/peerj-cs.3468.

PARAMASIVAM, M. E. et al. Oral cancer detection using convolutional neural network. *In*: INTERNATIONAL CONFERENCE ON INNOVATIVE PRACTICES IN TECHNOLOGY AND MANAGEMENT (ICIPTM), 4., 2024. **Anais [...]**. [S. l.]: IEEE, 2024. p. 1–6. DOI: 10.1109/ICIPTM59628.2024.10563232.

PEDREGOSA, F. et al. Scikit-learn: machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

RODRIGUES, Maria Isabelly de Brito. **CancerOral-CNNs**. GitHub, 2026. Disponível em: https://github.com/Isabellybrt/CancerOral-CNNs. Acesso em: 6 maio 2026.

SOKOLOVA, M.; LAPALME, G. A systematic analysis of performance measures for classification tasks. **Information Processing & Management**, v. 45, n. 4, p. 427–437, 2009. DOI: 10.1016/j.ipm.2009.03.002.

REMUNOM
ISSN 2178-6925

Received: 19/04/2026 - **Accepted:** 06/05/2026
Vol: 13.09
DOI: 10.66104/eecb2s81
Pages: 1-34

SPEIGHT, P. M. et al. Oral potentially malignant disorders: risk of progression to malignancy. **Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology**, v. 125, n. 6, p. 612–627, 2018. DOI: 10.1016/j.oooo.2017.12.011.

SUNG, H. et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. **CA: A Cancer Journal for Clinicians**, v. 71, n. 3, p. 209–249, 2021. DOI: 10.3322/caac.21660.

SZEGEDY, C. et al. Going deeper with convolutions. *In*: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR), 2015, Boston. **Proceedings [...]**. Piscataway: IEEE, 2015. p. 1–9. DOI: 10.1109/CVPR.2015.7298594.

TAJBAKHSH, N. et al. Convolutional neural networks for medical image analysis: full training or fine tuning? **IEEE Transactions on Medical Imaging**, v. 35, n. 5, p. 1299–1312, 2016. DOI: 10.1109/TMI.2016.2535302.

WARIN, K. et al. Automatic classification and detection of oral cancer in photographic images using deep learning algorithms. **Journal of Oral Pathology & Medicine**, v. 50, n. 9, p. 911–918, 2021. DOI: 10.1111/jop.13227.

WARNAKULASURIYA, S. Global epidemiology of oral and oropharyngeal cancer. **Oral Oncology**, v. 45, n. 4–5, p. 309–316, 2009. DOI: 10.1016/j.oraloncology.2008.06.002.

WELIKALA, R. A. et al. Automated detection and classification of oral lesions using deep learning for early detection of oral cancer. **IEEE Access**, v. 8, p. 132677–132693, 2020. DOI: 10.1109/ACCESS.2020.3010180.

ZAIDPY. **Oral cancer dataset**. Kaggle, 2022. Disponível em: https://www.kaggle.com/datasets/zaidpy/oral-cancer-dataset. Acesso em: 7 maio 2026.