INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO PIAUÍ
CAMPUS PIRIPIRI
TECNÓLOGO EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

JOSÉ NILTON SILVA LIMA

COMPARAÇÃO DE ARQUITETURAS DE REDES NEURAIS CONVOLUCIONAIS PARA CLASSIFICAÇÃO DE CÁRIES EM RADIOGRAFIAS PANORÂMICAS

PIRIPIRI/PI
2026

JOSÉ NILTON SILVA LIMA

COMPARAÇÃO DE ARQUITETURAS DE REDES NEURAIS CONVOLUCIONAIS PARA CLASSIFICAÇÃO DE CÁRIES EM RADIOGRAFIAS PANORÂMICAS

Trabalho de Conclusão de Curso (artigo científico) apresentado como exigência parcial para obtenção do diploma do Curso de Tecnólogo em Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia do Piauí, Campus Piripiri.

Orientador(a): Prof. Dr. Iallen Gábio de Sousa Santos

PIRIPIRI/PI
2026

COMPARAÇÃO DE ARQUITETURAS DE REDES NEURAIS CONVOLUCIONAIS PARA CLASSIFICAÇÃO DE CÁRIES EM RADIOGRAFIAS PANORÂMICAS

José Nilton Silva Lima[1]
Prof. Dr. Iallen Gábio de Sousa Santos[2]

RESUMO

A cárie dentária afeta 2,5 bilhões de pessoas com lesões não tratadas em dentes permanentes, segundo relatório mais recente da OMS publicado em 2022. Nos serviços odontológicos, seu diagnóstico em radiografias panorâmicas depende da leitura visual do profissional, uma etapa sujeita a variação entre examinadores e sensível à carga de trabalho. Modelos baseados em redes neurais convolucionais têm mostrado desempenho relevante nessa tarefa, mas a literatura carece de comparações controladas entre famílias arquiteturais distintas sobre o mesmo conjunto de dados. Este trabalho compara quatro arquiteturas, Inception-v3, InceptionResNet-v2, Xception e EfficientNetV2-S, na classificação binária de recortes individuais de dentes extraídos de radiografias panorâmicas. O conjunto de dados reúne 7.304 imagens balanceadas dos repositórios públicos InReDD e DENTEX. Um terço foi destinado ao treinamento com validação cruzada estratificada de cinco *folds* e todas as arquiteturas passaram por *fine-tuning* sobre pesos pré-treinados no ImageNet, enquanto os dois terços restantes (4.870 imagens), mantidos fora do ciclo de treinamento e validação, foram reservados para avaliação por *ensemble* dos cinco modelos de cada arquitetura. O Xception registrou o maior *F1-score* médio na validação cruzada (0,9458) e a maior precisão (0,9818). O Inception-v3 foi o mais estável entre *folds*. O InceptionResNet-v2 obteve o maior *recall* médio (0,9268), perdendo menos lesões por *fold*. O EfficientNetV2-S ficou abaixo dos demais em todas as métricas, com acurácia média de 91,37% e desvio padrão do *recall* de 4,14 pontos percentuais. No conjunto de inferência, o padrão observado na validação cruzada se confirmou, o Xception manteve o melhor equilíbrio entre precisão e *recall*, com acurácia de *ensemble* de 94,52% e *F1-score* de 0,9433, enquanto o Inception-v3 registrou a menor taxa de falsos positivos, apenas 46 sobre as 4.870 imagens avaliadas. As três arquiteturas da família Inception foram mais robustas nesse regime de dados, e o Xception apresentou o perfil de erro mais adequado para uso em triagem odontológica.

Palavras-chave: Cárie Dentária; Radiografia Panorâmica; *Transfer Learning*; Redes Reurais Convolucionais; Aprendizado Profundo.

ABSTRACT

Dental caries affects 2.5 billion people with untreated lesions in permanent teeth, according to the most recent WHO report published in 2022. In dental care services, its diagnosis on panoramic radiographs relies on the practitioner's visual reading, a step subject to inter-examiner variation and sensitive to workload. Convolutional neural network-based models have shown relevant performance on this task, but the literature lacks controlled comparisons between distinct architectural families on the same dataset. This work compares four architectures,

[1]Discente do curso Tecnólogo em Análise e Desenvolvimento de Sistemas. Instituto Federal de Educação, Ciência e Tecnologia do Piauí - Campus Piripiri. e-mail: jniltonsilval35@gmail.com

[2]Docente orientador do Instituto Federal de Educação, Ciência e Tecnologia do Piauí - Campus Piripiri. e-mail: iallen@ifpi.edu.br

Inception-v3, InceptionResNet-v2, Xception, and EfficientNetV2-S, in the binary classification of individual tooth crops extracted from panoramic radiographs. The dataset comprises 7,304 balanced images from the public InReDD and DENTEX repositories. One third was allocated to training with five-fold stratified cross-validation, and all architectures underwent fine-tuning on ImageNet-pretrained weights, while the remaining two thirds (4,870 images), kept outside the training and validation cycle, were reserved for evaluation via ensembling of the five models from each architecture. Xception achieved the highest mean F1-score in cross-validation (0.9458) and the highest precision (0.9818). Inception-v3 was the most stable across folds. InceptionResNet-v2 obtained the highest mean recall (0.9268), missing fewer lesions per fold. EfficientNetV2-S underperformed relative to the others across all metrics, with a mean accuracy of 91.37% and a recall standard deviation of 4.14 percentage points. In the inference set, the pattern observed in cross-validation was confirmed: Xception maintained the best balance between precision and recall, with an ensemble accuracy of 94.52% and an F1-score of 0.9433, while Inception-v3 recorded the lowest false positive rate, only 46 out of the 4,870 images evaluated. The three architectures from the Inception family were more robust under this data regime, and Xception showed the most suitable error profile for use in dental screening.

Keywords: Dental Caries; Panoramic Radiograph; Transfer Learning; Convolutional Neural Networks; Deep Learning.

Data de aprovação: 10/07/2026

1 INTRODUÇÃO

A cárie dentária é a condição oral mais prevalente no mundo. O Relatório Global de Saúde Bucal publicado pela Organização Mundial da Saúde em 2022 estimou que aproximadamente 3,5 bilhões de pessoas sofriam de alguma doença bucal, sendo a cárie não tratada em dentes permanentes a mais comum entre elas, com 2,5 bilhões de casos (World Health Organization, 2022). No Brasil, a situação é acompanhada por meio da Pesquisa Nacional de Saúde Bucal, denominada SB Brasil. A edição mais recente, teve a coleta de dados concluída no primeiro semestre de 2024 e o relatório final publicado em 2024 (Brasil, 2024).

Vinculada à Política Nacional de Saúde Bucal, essa pesquisa avalia, entre outros indicadores, o índice CPOD, que registra o número de dentes cariados, perdidos e restaurados por indivíduo. Na população adulta, entre 35 e 44 anos, o CPOD médio caiu de 20,13 em 2003 para 16,75 em 2010 e para 10,70 no SB Brasil 2023, redução observada em todas as macrorregiões do país (Brasil, 2024). Isso significa que um CPOD de 10,70 indica que, em média, cada pessoa do grupo avaliado apresenta 10,70 dentes, de um total de 32, afetados pela cárie ao longo da vida, seja porque estão cariados no momento do exame, foram extraídos ou já foram restaurados.

A identificação precoce da lesão cariosa enfrenta limitações clínicas objetivas. A desmineralização inicial do esmalte não produz alteração de coloração perceptível ao exame visual direto, e as radiografias convencionais têm sensibilidade reduzida para lesões incipi-

entes em superfícies proximais. Essas restrições, somadas à alta demanda por atendimento nos serviços públicos de saúde bucal, fundamentam o interesse em sistemas computacionais capazes de auxiliar no rastreamento diagnóstico em larga escala (Negi *et al.*, 2024).

A Inteligência Artificial, em especial o *Deep Learning*, tem revolucionado a análise de imagens médicas ao utilizar redes neurais artificiais complexas que mimetizam o processamento cognitivo humano para identificar padrões em dados brutos. No campo odontológico, essa evolução manifesta-se principalmente através das redes neurais convolucionais (CNNs), utilizadas para a análise de imagens complexas e na segmentação de estruturas (Tavares *et al.*, 2025).

Assim, diante dos avanços recentes em aprendizado profundo e visão computacional, emerge como problemática central desta pesquisa a seguinte questão: "Qual arquitetura CNN apresenta melhor desempenho na classificação binária de lesões cariosas em radiografias panorâmicas?"

O diagnóstico de cáries em radiografias panorâmicas é uma tarefa rotineira em serviços odontológicos, mas permanece dependente da análise visual manual, sujeita a variabilidade entre examinadores e a erros relacionados à carga de trabalho clínico. Modelos baseados em CNNs têm demonstrado capacidade de operar nesse tipo de tarefa com acurácia relevante, porém a literatura ainda apresenta comparações incompletas entre famílias arquiteturais distintas aplicadas ao mesmo problema e ao mesmo conjunto de dados.

O objetivo geral deste trabalho é avaliar e comparar o desempenho de quatro arquiteturas de redes neurais convolucionais, Inception-v3, InceptionResNet-v2, Xception e EfficientNetV2-S, na classificação binária de lesões cariosas em recortes individuais de dentes, extraídos de radiografias panorâmicas. Os objetivos específicos são: compor um conjunto de dados balanceado a partir de duas fontes públicas anotadas, o InReDD e o DENTEX; definir um *pipeline* de pré-processamento com extração de recortes individuais, expansão de contexto anatômico e normalização de escala; treinar cada arquitetura com *fine-tuning* sobre pesos pré-treinados no ImageNet; avaliar o desempenho por validação cruzada estratificada de cinco partições.

A avaliação das arquiteturas foi conduzida por meio de validação cruzada estratificada de cinco partições sobre um conjunto balanceado de 2.434 imagens de dentes extraídos de radiografias panorâmicas. Os resultados demonstraram que as arquiteturas da família Inception atingiram acurácias médias entre 93,96% e 94,78%, com o Xception registrando o maior *F1-score* médio (0,9458) e o Inception-v3 apresentando o menor desvio padrão de acurácia ao longo dos *folds*, indicando maior estabilidade preditiva. O EfficientNetV2-S registrou acurácia média de 91,37%, com intervalos de desempenho que não se sobrepõem aos das demais arquiteturas, configurando inferioridade estatisticamente sustentada nas condições experimentais adotadas. Os modelos foram ainda avaliados sobre um subconjunto independente de 4.870 imagens, correspondente a dois terços do conjunto total e mantido fora de todo o ciclo de treinamento e validação cruzada. Nessa etapa, o Xception manteve o

melhor desempenho, com acurácia de *ensemble* de 94,52% e *F1-score* de 0,9433, enquanto o Inception-v3 apresentou a menor taxa de falsos positivos, com apenas 46 alarmes incorretos sobre as 4.870 imagens avaliadas.

Este trabalho está organizado em cinco seções. A Seção 2 apresenta o referencial teórico sobre aprendizado profundo, cárie dentária, radiografia panorâmica e diagnóstico assistido por inteligência artificial. A Seção 3 descreve a metodologia, incluindo os dados, o pré-processamento, as arquiteturas avaliadas e o protocolo de validação cruzada. A Seção 4 apresenta e discute os resultados experimentais. A Seção 5 expõe as considerações finais, as limitações e as direções para trabalhos futuros.

2 REFERENCIAL TEÓRICO

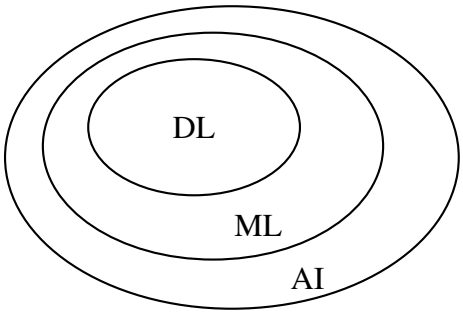
Este capítulo apresenta os fundamentos teóricos que sustentam a pesquisa. Nele, é abordado os conceitos de aprendizado profundo e redes neurais convolucionais. Em seguida, é discutido os aspectos odontológicos relacionados às lesões cariosas e ao uso de radiografias panorâmicas, bem como os desafios do diagnóstico radiográfico apoiado por inteligência artificial. Ao final, são apresentadas as contribuições e as limitações do presente trabalho.

2.1 *DEEP LEARNING*

A Inteligência Artificial (IA) é um campo científico e de engenharia que busca construir sistemas capazes de perceber, raciocinar e agir de forma autônoma em ambientes complexos (Russell; Norvig, 2013). Como pode ser visto na Figura 1, o Aprendizado de Máquina (*Machine Learning* – ML) é uma subárea da IA que permite que sistemas computacionais aprimorem seu desempenho a partir da experiência e da análise de dados, empregando algoritmos capazes de identificar padrões com base em conjuntos de dados rotulados ou não rotulados, sem depender de regras explícitas. Entre as abordagens de ML, o Aprendizado Profundo (*Deep Learning* – DL) se distingue pela aprendizagem hierárquica de representações, estruturando o conhecimento em múltiplos níveis de abstração por meio de transformações não lineares sucessivas (Goodfellow; Bengio; Courville, 2016).

As Redes Neurais Convolucionais (*Convolutional Neural Networks* – CNNs) são uma família de arquiteturas de DL desenvolvidas para o processamento de dados com estrutura em grade, como imagens bidimensionais (LeCun; Bengio; Hinton, 2015). Elas utilizam a operação de convolução, o que permite interações esparsas e compartilhamento de parâmetros, reduzindo o custo computacional sem comprometer a capacidade de representação (Goodfellow; Bengio; Courville, 2016). As CNNs são empregadas em três categorias de tarefas: a classificação, que atribui um rótulo a uma imagem ou região; a detecção, que localiza objetos por meio de caixas delimitadoras (Zancan *et al.*, 2025; Ramírez-Pedraza *et al.*, 2025); e a segmentação, que realiza a delimitação ao nível de pixel (Costa *et al.*, 2024; Gao *et al.*, 2025).

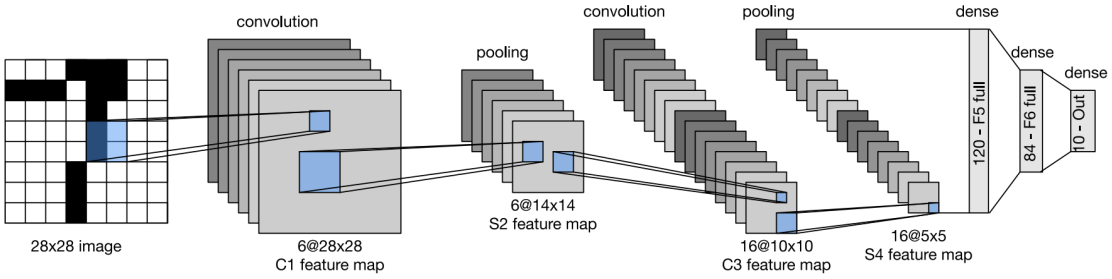
Figura 1 – Hierarquia entre Inteligência Artificial (AI), Aprendizado de Máquina e Aprendizado Profundo.



Fonte: Adaptado de Goodfellow, Bengio e Courville (2016).

A operação de convolução funciona por meio de um filtro, também chamado de núcleo (*kernel*), que desliza sobre a imagem de entrada e, a cada posição, computa o produto escalar entre seus pesos e os valores de pixel cobertos por ele (Figura 2). O resultado acumulado para todas as posições espaciais forma um *mapa de características* (*feature map*), que indica onde e com que intensidade o padrão aprendido pelo filtro aparece na imagem. Uma camada convolucional aplica simultaneamente dezenas ou centenas de filtros distintos, gerando um volume de mapas de características como saída. O compartilhamento de parâmetros torna esse processo eficiente: um filtro de 3×3 pixels tem apenas 9 pesos, independentemente do tamanho da imagem (Goodfellow; Bengio; Courville, 2016).

Figura 2 – Fluxo de dados em uma Rede Neural Convolucional.



Fonte: Adaptado de Zhang *et al.* (2023).

Intercaladas às camadas convolucionais, as camadas de *pooling* reduzem a resolução espacial dos mapas de características. A operação mais comum, o *max pooling*, divide o mapa em regiões e retém apenas o valor máximo de cada uma, reduzindo o volume de dados que flui para as camadas seguintes e introduzindo tolerância a pequenas variações de posição (LeCun; Bengio; Hinton, 2015). Nas camadas iniciais da rede, os filtros aprendem características elementares como bordas e texturas simples. À medida que o sinal avança pelas camadas mais profundas, essas características são combinadas em representações progressivamente mais abstratas, que correspondem a estruturas complexas específicas do domínio do problema (Goodfellow; Bengio; Courville, 2016).

O treinamento de CNNs do zero, no entanto, requer grandes volumes de dados rotulados e alto custo computacional, o que restringe sua aplicação em domínios com bases de dados menores. A utilização de redes pré-treinadas surge como resposta a essa limitação: ao reutilizar representações aprendidas em grandes conjuntos de dados, é possível reduzir a demanda por dados de treinamento e atingir desempenho competitivo mesmo em tarefas com acervos menores (Zhang *et al.*, 2023). Este trabalho adota essa abordagem, usando pesos pré-treinados como ponto de partida para todas as arquiteturas avaliadas.

2.2 ASPECTOS ODONTOLÓGICOS RELACIONADOS AO DIAGNÓSTICO DE CÁRIES

Esta subseção contextualiza os elementos clínicos e radiográficos necessários à compreensão do problema investigado. Para isso, são apresentados a definição e o impacto epidemiológico da cárie dentária, as características da radiografia panorâmica odontológica e os principais desafios associados ao diagnóstico radiográfico e às aplicações de inteligência artificial nesse domínio.

2.2.1 Cárie Dentária: Definição e Impacto Epidemiológico

A cárie dentária (Figura 3) é uma doença infecciosa decorrente da desmineralização progressiva das estruturas calcificadas do dente, compreendendo esmalte, dentina e cemento, pela ação de ácidos orgânicos produzidos durante a fermentação de carboidratos por microorganismos presentes no biofilme dental (Fejerskov; Nyvad; Kidd, 2015). A desmineralização inicial do esmalte, reversível mediante aplicação de flúor e controle de higiene, antecede a formação de cavidade, que avança em direção à dentina e pode atingir a polpa dentária nos casos não tratados, exigindo intervenções progressivamente mais invasivas (Negi *et al.*, 2024).

Figura 3 – Exemplos de diferentes estágios da cárie dentária.

Fonte: Roboflow (2025).

A análise do *Global Burden of Disease* em 2019 conduzida por Li *et al.* (2022) estimou 2,03 bilhões de casos prevalentes de cárie não tratada em dentes permanentes no mundo naquele ano, representando aumento de 46% em relação a 1990. Em 2021, 27,54% da população adulta mundial apresentava cárie em dentes permanentes, segundo projeções

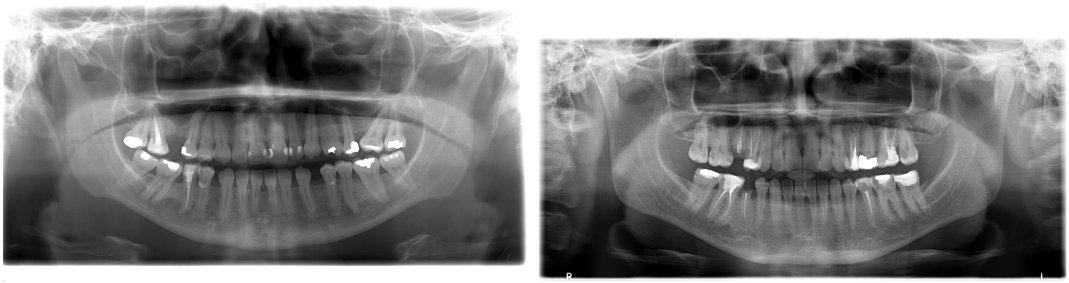
baseadas no mesmo banco de dados epidemiológicos (Wang *et al.*, 2025). Esses valores posicionam a cárie entre as doenças crônicas de maior prevalência global em adultos, reforçando a necessidade de ferramentas diagnósticas que ampliem a capacidade de detecção precoce em contextos de atenção em saúde coletiva (Li *et al.*, 2022).

A detecção precoce da cárie tem relevância clínica direta. Lesões diagnosticadas na fase de mancha branca, antes da formação de cavidade, são passíveis de remineralização por meio de intervenções não invasivas. Lesões cavitadas requerem remoção de tecido e restauração, cujos custo e complexidade aumentam com a profundidade da lesão e o envolvimento pulpar. A adoção de sistemas capazes de identificar alterações sutis de densidade óssea em imagens radiográficas antes que se tornem perceptíveis ao exame clínico responde, portanto, a uma lacuna diagnóstica de consequências concretas sobre desfechos em saúde (Negi *et al.*, 2024).

2.2.2 Radiografia Panorâmica Odontológica

O diagnóstico por imagem é indispensável na prática odontológica clínica porque possibilita a visualização de estruturas não acessíveis ao exame direto, entre as quais o interior dos dentes, o tecido radicular, o osso alveolar e as regiões de contato proximal entre dentes adjacentes. Entre as modalidades disponíveis, a radiografia panorâmica, também denominada ortopantomografia ou PAN, ocupa posição central tanto na prática clínica quanto em estudos de base epidemiológica.

Figura 4 – Radiografias dos *datasets* InReDD e DENTEX, respectivamente.



Fonte: Uehara Martins *et al.* (2025) e Hamamci *et al.* (2023).

A PAN é um exame extraoral que registra a arcada dentária completa em uma única imagem, abrangendo maxila, mandíbula, articulações temporomandibulares e estruturas adjacentes. O equipamento realiza um movimento de rotação ao redor da cabeça do paciente durante a aquisição, gerando uma projeção bidimensional de estruturas tridimensionais com dimensões típicas de 2.903 × 1.536 *pixels*. Sua ampla adoção na prática clínica decorre da cobertura completa da dentição em um único exame, do conforto do paciente durante a aquisição e do menor custo operacional em relação a tomografias computadorizadas de feixe cônico, modalidade de maior resolução mas menos acessível em contextos de atenção básica (Hung *et al.*, 2025).

A principal limitação das radiografias panorâmicas em relação ao diagnóstico de cárie está na resolução espacial inferior à das radiografias intraorais. Apesar dessa limitação, sistemas de inteligência artificial aplicados a radiografias panorâmicas têm demonstrado desempenho diagnóstico próximo ao de especialistas para cáries moderadas e avançadas, que correspondem ao perfil de lesão predominante nos conjuntos de dados publicamente disponíveis para pesquisa (Hung *et al.*, 2025).

2.2.3 Desafios do Diagnóstico Radiográfico e Aplicações de Inteligência Artificial

A interpretação de radiografias odontológicas está sujeita a variabilidade interobservador. Fadiga profissional, nível de experiência clínica, condições do ambiente de trabalho e a natureza inerentemente subjetiva da interpretação visual são fontes documentadas de inconsistência diagnóstica. A taxa de desacordo entre radiologistas experientes na avaliação de lesões cariosas em radiografias panorâmicas pode variar entre 10% e 30%, conforme o estágio da lesão e a qualidade da imagem (Hung *et al.*, 2025).

Lesões incipientes são as mais suscetíveis a erros de detecção. A diferença de tonalidade entre esmalte saudável e esmalte desmineralizado é sutil e frequentemente abaixo do limiar de discriminação visual humana. Sistemas computacionais operam com representações numéricas dos valores de pixel e podem ser treinados para identificar padrões de variação de densidade que não são perceptíveis ao exame visual direto (Negi *et al.*, 2024).

A pesquisa sobre uso de inteligência artificial no diagnóstico radiográfico odontológico tem crescido de forma expressiva na última década. Hung *et al.* (2025) revisaram estudos de IA aplicada especificamente a radiografias panorâmicas e documentaram que os modelos de maior desempenho atingem acurácia entre 0,90 e 0,95 para detecção de cáries, com especificidade de 0,98 no melhor resultado reportado. Uma análise publicada por Alsubiheen *et al.* (2025), consolidando dados de 14 revisões sistemáticas que abrangeram 137 estudos primários, estimou sensibilidade agrupada de 0,85 e especificidade de 0,90 para algoritmos de inteligência artificial na detecção de cáries em diversas modalidades radiográficas.

Esses resultados indicam que sistemas automatizados apresentam desempenho comparável ao de especialistas humanos em tarefas de classificação estruturada. Contudo, a variabilidade entre estudos permanece elevada em função da heterogeneidade dos conjuntos de dados e dos protocolos de avaliação adotados (Negi *et al.*, 2024). A integração clínica desses sistemas ainda é incipiente, e a maioria dos modelos publicados foi avaliada em condições experimentais, sem validação em ambientes de prática real. O desenvolvimento de conjuntos de dados anotados por especialistas, como o InReDD, é condição necessária para avançar na validação desses modelos para uso clínico (Uehara Martins *et al.*, 2025).

2.3 CONTRIBUIÇÕES DO PRESENTE TRABALHO

A literatura sobre aplicação de redes neurais convolucionais ao diagnóstico odontológico tem crescido de forma expressiva nos últimos anos, mas ainda apresenta lacunas metodológicas recorrentes. Estudos que avaliam uma única arquitetura em condições não padronizadas são predominantes, o que dificulta a comparação entre abordagens distintas e a identificação do que, de fato, explica as diferenças de desempenho observadas.

Dayı *et al.* (2023) propôs a DCDNet, uma arquitetura própria para segmentação de cáries oclusais, proximais e cervicais em radiografias panorâmicas, avaliada em 504 imagens provenientes de uma única instituição, sem comparação com outras arquiteturas sob o mesmo protocolo experimental. Vinayahalingam *et al.* (2021) avaliou exclusivamente MobileNet V2 para classificar cáries em terceiros molares, com apenas 400 imagens de treinamento e protocolo específico para esse grupo dentário. Haghanifar *et al.* (2023) propôs o PaXNet, avaliado em 470 radiografias panorâmicas de uma única fonte, sem comparação direta com outras arquiteturas sob o mesmo protocolo. Esta pesquisa posiciona-se nessa lacuna ao adotar um protocolo experimental rigoroso e uniforme para quatro arquiteturas avaliadas sob as mesmas condições.

A primeira contribuição diz respeito ao escopo da comparação arquitetural. Trabalhos como os de Zancan *et al.* (2025), Meseci (2026) e Bayraktar e Ayan (2023) avaliaram arquiteturas de aprendizado profundo para a detecção de cáries em radiografias panorâmicas com metodologias próximas à deste estudo. O primeiro avaliou Inception-v3 e InceptionResNet-v2 com os mesmos conjuntos de dados desta pesquisa; o segundo investigou DenseNet121 e Mask R-CNN sobre um conjunto de 14.498 recortes individuais de dentes; o terceiro comparou EfficientNet-B0, DenseNet-121 e ResNet-50 em classificação binária de cáries com mapas de ativação para interpretabilidade clínica.

O presente trabalho amplia esse escopo com a inclusão do Xception e do EfficientNetV2-S, arquiteturas com fundamentos de projeto distintos. Essa ampliação não é apenas quantitativa, pois comparar modelos de famílias diferentes sob o mesmo protocolo permite verificar se as vantagens observadas para as arquiteturas Inception decorrem de características específicas ao domínio odontológico ou se aparecem também em arquiteturas concorrentes.

A segunda contribuição está no protocolo experimental robusto. O conjunto de dados foi dividido em três partições com papéis distintos: aproximadamente um terço (2.434 imagens) foi reservado ao treinamento com validação cruzada estratificada de cinco partições em proporção 80/20, e os dois terços restantes (4.870 imagens) foram separados para inferência em dados completamente fora do ciclo de treinamento. Essa separação estrutural, verificada por rotina automatizada de integridade antes de cada execução, garante que as métricas reportadas reflitam a capacidade de generalização dos modelos a dados não observados durante o ajuste dos pesos, e não apenas a memorização do conjunto de treinamento.

Como garantia desse rigor metodológico, todos os experimentos são estruturados de

forma estritamente reproduzível. O código completo, cobrindo a montagem do conjunto de dados e o treinamento das quatro arquiteturas sob validação cruzada, está disponível em repositório do GitHub (Lima, 2026).

2.4 LIMITAÇÕES DO PRESENTE TRABALHO

O tamanho do conjunto utilizado no treinamento é a limitação mais direta deste trabalho. O subconjunto de treinamento e validação contém 2.434 imagens, o que representa um terço do total disponível após o balanceamento. Essa redução não foi motivada apenas pela separação de um conjunto de inferência independente. Treinar com as 7.304 imagens completas exigia um custo computacional alto demais para o ambiente disponível.

Além disso, o volume de dados produzia um efeito indesejado, ou seja, ao final da primeira época, o modelo já acumulava informação suficiente de cada *batch* para atingir acurácias elevadas, e o treinamento convergia rápido demais para que os *folds* fossem comparáveis entre si de forma confiável. Reduzir o conjunto de treinamento ao terço menor resolveu esses dois problemas, mas criou outro. Com cerca de 1.947 imagens por *fold* de treinamento, os modelos foram ajustados com menos exemplos do que seria adequado, e as arquiteturas de maior capacidade, como o InceptionResNet-v2, são as mais afetadas por essa restrição.

Arquiteturas como transformadores de visão e detectores de objetos mais recentes, não foram avaliadas. O custo para treiná-las sob validação cruzada, no mesmo ambiente e com o mesmo protocolo, seria inviável. Os resultados deste trabalho dizem o que é possível dentro de restrições de infraestrutura, não o que seria possível com recursos maiores.

A ausência de análise de interpretabilidade é outra limitação relevante. Nenhum modelo foi submetido a técnicas como Grad-CAM ou mapas de ativação para identificar quais regiões da imagem influenciaram cada classificação. Sem isso, não é possível verificar se os modelos aprenderam padrões radiográficos com sentido clínico ou se estão respondendo a diferenças de contraste e artefatos entre as duas fontes de dados. Para qualquer aplicação clínica futura, essa questão precisa ser respondida antes da acurácia.

A avaliação restringiu-se à classificação binária de imagens de dentes já recortadas, tendo a certeza que a etapa de extração dos dentes individuais a partir das radiografias panorâmicas foi realizada de forma adequada. O desempenho do sistema completo em um cenário de uso clínico real depende também da qualidade dessa etapa anterior.

3 METODOLOGIA

Esta seção descreve os procedimentos adotados para a realização dos experimentos. São apresentados os conjuntos de dados utilizados, as etapas de pré-processamento, as arquiteturas de redes neurais convolucionais avaliadas, o protocolo de treinamento e validação e as métricas empregadas na análise do desempenho dos modelos.

3.1 CONJUNTOS DE DADOS

Esta pesquisa utiliza imagens panorâmicas odontológicas provenientes de duas fontes distintas: o InReDD (*Intraoral and Radiographic Dental Dataset*) (Uehara Martins *et al.*, 2025), construído a partir de atendimentos clínicos realizados na Faculdade de Odontologia de Ribeirão Preto (FORP-USP), e o DENTEX (*Dental Enumeration and Diagnosis on Panoramic X-Rays*) (Hamamci *et al.*, 2023), disponível publicamente sob licença CC BY-SA 4.0.

O InReDD reúne 924 radiografias panorâmicas de pacientes com idades entre 14 e 81 anos (mediana de 35 anos), sendo aproximadamente 60% do sexo feminino e 40% do masculino, obtidas entre 2014 e 2020. O conjunto está disponível no repositório PhysioNet (Uehara Martins *et al.*, 2025). As anotações foram produzidas por três radiologistas dentomaxilofaciais, cada um com uma década de experiência, mediante protocolo de revisão independente entre anotadores; os casos discordantes foram arbitrados por um terceiro especialista, resultando em taxa de desacordo inferior a 1% do total de anotações.

Cada dente recebe um rótulo de condição clínica conforme nomenclatura padronizada, sendo as mais relevantes para este estudo: saudável (H), cárie (C) e destruição da coroa (Dc). O conjunto apresenta 20.957 anotações em formato de *bounding box* cobrindo todas as 924 imagens, e 4.621 máscaras de segmentação distribuídas sobre um subconjunto de 200 imagens. Na versão pública utilizada, foram identificadas 10.665 anotações de dentes saudáveis e 468 anotações de dentes com alteração coronária nas categorias C e Dc.

O DENTEX foi concebido como *benchmark* para detecção simultânea de múltiplas patologias em radiografias panorâmicas de pacientes com idade mínima de 12 anos, provenientes de hospitais na Turquia. O esquema de anotação codifica, para cada dente, o quadrante, a posição segundo a notação FDI e a condição patológica. As condições disponíveis incluem cárie, cárie profunda, lesão periapical e dente impactado. Neste trabalho, foram aproveitadas apenas as instâncias rotuladas como cárie ou cárie profunda, abrangendo os *folds* de treinamento, validação e teste do conjunto original. A Tabela 1 apresenta as estatísticas descritivas de cada conjunto antes do balanceamento.

Tabela 1 – Estatísticas descritivas dos conjuntos de dados.

Dataset	**PANs**	**Cariados**	**Saudáveis**
InReDD	924	468	10.665
DENTEX	1.005	3.647	–
Total	**1.929**	**4.115**	**10.665**

Fonte: Elaborado pelos autores (2026).

O DENTEX não registra dentes saudáveis nas anotações, pois seu protocolo marca apenas estruturas com patologia confirmada. Por essa razão, todos os exemplos negativos (dentes não cariados) foram extraídos do InReDD.

O InReDD foi o conjunto de dados empregado por Zancan *et al.* (2025) no desenvolvimento de um sistema modular de análise de radiografias panorâmicas que abrange numeração

de dentes e detecção de lesões cavitadas. O presente trabalho se vale das mesmas imagens para investigar a capacidade de quatro arquiteturas de redes neurais na mesma tarefa de classificação de cáries, incluindo duas arquiteturas não avaliadas no estudo original.

3.2 PRÉ-PROCESSAMENTO

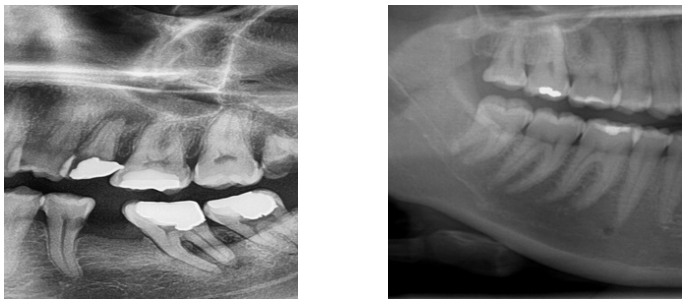
O pré-processamento foi realizado com o objetivo de transformar as radiografias panorâmicas e suas anotações em um conjunto adequado ao treinamento e à avaliação dos modelos. Essa etapa compreendeu a extração das imagens individuais de dentes, a composição e o balanceamento das classes e o particionamento primário dos dados.

3.2.1 Extração de imagens individuais de dentes

Cada radiografia panorâmica contém a arcada dentária completa em uma única imagem de 2.903 × 1.536 *pixels*, armazenada no formato JPEG a 95 dpi (Uehara Martins *et al.*, 2025). Para que os modelos operem sobre regiões clinicamente relevantes, cada dente foi extraído individualmente a partir das coordenadas de delimitação presentes nas anotações.

No InReDD, as anotações de dentes individuais são fornecidas como polígonos de segmentação no subconjunto de 200 imagens com máscaras disponíveis, e como caixas delimitadoras retangulares nas demais. A caixa delimitadora de cada dente foi derivada dos valores mínimo e máximo das coordenadas horizontal e vertical dos vértices do polígono quando a segmentação estava disponível, ou utilizada diretamente nas demais. As coordenadas seguem o padrão COCO: [x, y, largura, altura], onde x e y correspondem ao canto superior esquerdo da caixa. No DENTEX, as anotações já estão no formato de caixa retangular em coordenadas absolutas.

Figura 5 – Exemplos de imagens individuais de dentes cariados e não cariados, respectivamente, após o pré-processamento.



Fonte: Adaptado de Uehara Martins *et al.* (2025) e Hamamci *et al.* (2023).

A caixa delimitadora obtida foi expandida em 300% sobre sua altura e largura, mantendo o centro geométrico do dente, como pode ser visto na Figura 5. Essa expansão preserva o contexto anatômico adjacente ao dente de interesse, que inclui tecido periodontal, osso alveolar e dentes vizinhos. Zancan *et al.* (2025) demonstraram que incluir a região circun-

dante melhora a acurácia em classificação de dentes em radiografias. A caixa expandida foi truncada nos limites da imagem original para evitar artefatos de borda.

O recorte resultante foi redimensionado para 299 × 299 pixels por interpolação de Lanczos, resolução de entrada das quatro arquiteturas avaliadas. A normalização dos valores de pixel para o intervalo [0, 1] ocorreu em tempo de execução durante o carregamento de cada lote, via *ImageDataGenerator* do Keras.

3.2.2 Composição e balanceamento do conjunto

As 468 instâncias de dentes com alteração coronária do InReDD e as 3.647 do DENTEX foram reunidas em um conjunto de 4.115 exemplos positivos. Como o alvo era 3.652 imagens por classe, realizou-se amostragem aleatória sem reposição das imagens cariadas até atingir esse limite.

Para as instâncias não cariadas, foram selecionadas aleatoriamente 3.652 imagens dentre as 10.665 disponíveis no InReDD. O conjunto final contém 7.304 imagens, com distribuição igualitária entre as duas classes. A Tabela 2 apresenta a composição detalhada após o balanceamento.

Tabela 2 – Composição e balanceamento do conjunto de classificação.

Classe	**Fonte**	**Instâncias disponíveis**	**Instâncias utilizadas**
Cariado	InReDD (C + Dc)	468	3.652
	DENTEX (treinamento + validação)	2.900	
	DENTEX (teste)	747	
Não cariado	InReDD (H)	10.665	3.652
Total		**14.780**	**7.304**

Fonte: Elaborado pelos autores (2026). C = cárie; Dc = destruição da coroa; H = saudável.

Dentes com restaurações extensas, fraturas ou artefatos metálicos foram excluídos na etapa de anotação do InReDD e não exigiram filtragem adicional. No DENTEX, apenas as anotações rotuladas como cárie inicial ou cárie profunda foram aproveitadas; lesões periapicais e dentes impactados foram descartados.

3.2.3 Particionamento primário do conjunto

O conjunto balanceado de 7.304 imagens foi submetido a uma divisão primária de caráter estratificado, na proporção de um terço para treinamento e validação e dois terços reservados para inferência e *stratify* aplicado sobre os rótulos de classe, de modo a preservar a proporção 50%/50% em ambas as frações resultantes.

O subconjunto de treinamento e validação reuniu 2.434 imagens, sendo 1.217 da classe cariado e 1.217 da classe não cariado. O subconjunto reservado para inferência totalizou 4.870 imagens, com 2.435 exemplares por classe. Essa decisão foi motivada por

restrições de custo computacional, que ao confinar o ciclo de treinamento supervisionado a aproximadamente um terço do acervo total, cada experimento de validação cruzada consumiu consideravelmente menos memória de vídeo (VRAM) e tempo de processamento por época. O subconjunto de inferência foi armazenado separadamente e não participou de nenhuma etapa de treinamento nem de ajuste de hiperparâmetros. A Tabela 3 resume a distribuição por classe obtida após a divisão estratificada.

Tabela 3 – Distribuição por classe após o particionamento primário do conjunto.

Subconjunto	**Cariados**	**Não cariados**	**Total**
Treinamento e validação (1/3)	1.217	1.217	2.434
Inferência (2/3)	2.435	2.435	4.870
Total	**3.652**	**3.652**	**7.304**

Fonte: Elaborado pelos autores (2026).

3.3 CLASSIFICAÇÃO DE LESÕES CARIOSAS EM RADIOGRAFIAS PANORÂMICAS

A tarefa investigada é a classificação binária de imagens de dentes individuais extraídas de radiografias panorâmicas, com duas categorias possíveis: cariado e não cariado. Essa formulação opera em nível de dente inteiro, diferindo da detecção de objetos, que requer a predição de coordenadas de localização, e da segmentação semântica, que classifica em nível de pixel.

3.3.1 Arquiteturas de Redes Neurais Convolucionais Avaliadas

Como já foi destacado, As CNNs levam em conta a relação espacial entre *pixels* vizinhos, o que as torna mais eficientes do que as redes neurais convencionais para detectar padrões visuais localizados em regiões específicas da imagem. Nesta pesquisa, quatro arquiteturas de CNN foram selecionadas com base em três critérios: compatibilidade nativa com a resolução de entrada de 299 × 299 *pixels*; disponibilidade de pesos pré-treinados no ImageNet[3]; e cobertura de distintas abordagens arquiteturais, de modo que os resultados permitam comparações metodologicamente informativas. A seguir, cada arquitetura é descrita com ênfase em seus princípios de projeto e no que a diferencia das demais.

Inception-v3: Arquitetura proposta por Szegedy, Vanhoucke *et al.* (2016) com foco na redução do custo computacional sem perda de capacidade representacional. A arquitetura conta com 23 milhões de parâmetros e tem como principal inovação a fatoração de convoluções: filtros $n \times n$ são decompostos em pares assimétricos $1 \times n$ e $n \times 1$, reduzindo o número de operações em aproximadamente 33%. Cada módulo aplica filtros de múltiplas escalas em paralelo e concatena suas saídas antes de propagar o sinal adiante.

InceptionResNet-v2: Szegedy, Ioffe *et al.* (2017) incorporaram conexões residuais aos mó-

[3]https://www.image-net.org/

dulos de múltiplas escalas do Inception. Essas conexões, que somam a entrada de um bloco à sua saída, atenuam o problema de desvanecimento do gradiente em redes profundas e aceleram a convergência do treinamento. Com 55 milhões de parâmetros, o InceptionResNet-v2 é a arquitetura de maior capacidade entre as avaliadas nesta pesquisa.

Xception: Chollet (2017) reformulou os módulos Inception a partir da hipótese de que correlações espaciais e correlações entre canais de ativação são suficientemente independentes para serem tratadas por operações separadas. A arquitetura substitui os módulos Inception por convoluções separáveis por profundidade, nas quais um filtro espacial opera sobre cada canal individualmente, seguido de uma convolução pontual que combina os resultados entre canais. Com 22 milhões de parâmetros, o Xception supera o Inception-v3 no conjunto de validação do ImageNet com custo computacional equivalente. Não foram identificados estudos anteriores que tenham avaliado essa arquitetura especificamente para detecção de cáries em radiografias panorâmicas.

EfficientNetV2-S: Tan e Le (2021) investigaram como profundidade, largura e resolução de entrada devem ser escalonadas de forma conjunta para maximizar a acurácia dado um orçamento computacional fixo. A família EfficientNetV2 acrescenta à geração anterior um protocolo de treinamento progressivo, no qual a resolução das imagens e a intensidade da regularização aumentam gradualmente ao longo das épocas. A variante S conta com 21 milhões de parâmetros e foi incluída nesta pesquisa para avaliar se arquiteturas mais recentes e compactas são competitivas com modelos mais profundos no domínio odontológico.

Em todas as arquiteturas, os pesos pré-treinados no ImageNet, uma coleção de mais de 14 milhões de imagens organizadas em aproximadamente 22.000 categorias, foram usados como inicialização e todas as camadas foram liberadas para atualização durante o treinamento, configuração conhecida como *fine-tuning*. Sobre a saída da base convolucional foram adicionadas uma camada de *Global Average Pooling*, uma camada de *dropout* e uma camada densa com ativação sigmoide para a classificação binária. O otimizador Adam foi adotado em todos os experimentos, com parâmetros padrão ($\beta_1 = 0{,}9$; $\beta_2 = 0{,}999$; $\epsilon = 10^{-7}$). A função de perda foi a entropia cruzada binária. A Tabela 4 apresenta os hiperparâmetros específicos de cada arquitetura.

Tabela 4 – Hiperparâmetros de treinamento por arquitetura.

Arquitetura	***Dropout***	**Taxa de aprendizado**	**Épocas máx.**	**Lote**
Inception-v3	0,2	$1{,}0 \times 10^{-5}$	50	32
InceptionResNet-v2	0,2	$1{,}0 \times 10^{-5}$	50	32
Xception	0,2	$1{,}0 \times 10^{-5}$	50	32
EfficientNetV2-S	0,2	$1{,}0 \times 10^{-5}$	50	32

Fonte: Elaborado pelos autores (2026).

3.4 PROTOCOLO DE TREINAMENTO E VALIDAÇÃO

O protocolo de treinamento e validação foi definido para permitir uma comparação uniforme entre as arquiteturas avaliadas e reduzir a influência de uma única divisão dos dados sobre os resultados. Para isso, foi adotada a validação cruzada estratificada com cinco partições, mantendo-se a distribuição das classes durante as diferentes rodadas experimentais.

3.4.1 Validação cruzada estratificada com cinco *folds*

A validação cruzada foi aplicada exclusivamente sobre o subconjunto de 2.434 imagens separado para treinamento e validação, conforme descrito na Seção 3.2.3. O método adotado foi o *StratifiedKFold*, implementado por meio da biblioteca *scikit-learn* (Pedregosa *et al.*, 2018). A função de estratificação garante que cada partição mantenha a mesma proporção entre classes do subconjunto de origem, aproximadamente 50% de exemplos cariados e 50% de não cariados em treino e em validação.

O funcionamento do *StratifiedKFold* com cinco *folds* consiste em dividir o subconjunto de 2.434 exemplos em cinco frações de tamanho aproximadamente igual. A cada rodada, quatro frações são combinadas para compor o conjunto de treinamento (cerca de 80% do subconjunto) e a fração restante forma o conjunto de validação (cerca de 20%). O processo se repete por cinco rodadas, cada uma usando uma fração diferente como validação, de forma que ao final cada imagem tenha feito parte do conjunto de validação uma vez.

As métricas foram calculadas sobre o conjunto de validação de cada rodada, e o desempenho final de cada arquitetura corresponde à média e ao desvio padrão obtidos nas cinco rodadas. Ao todo foram executados vinte experimentos, resultado do produto entre as quatro arquiteturas e os cinco *folds* de validação cruzada. O modelo de cada *fold* foi salvo a cada época em que a perda de validação atingiu novo mínimo, garantindo que os pesos restaurados pela parada antecipada correspondam ao melhor estado observado na validação.

Além disso, dois mecanismos foram aplicados em todos os experimentos. O primeiro é o *early stopping*, onde o treinamento é encerrado quando a perda observada no conjunto de validação não apresenta melhora por seis épocas consecutivas, restaurando automaticamente os pesos da época de menor perda observada. O segundo é o *ReduceLROnPlateau*, na qual a taxa de aprendizado é reduzida à metade quando a perda observada no conjunto de validação não melhora em uma época.

3.5 MÉTRICAS DE AVALIAÇÃO

O desempenho dos modelos foi mensurado por cinco métricas calculadas sobre o conjunto de validação de cada *fold*, todas derivadas da matriz de confusão binária onde a classe positiva corresponde ao dente cariado. A precisão avalia a proporção de dentes identificados como cariados que, de fato, apresentam a lesão ($TP/(TP + FP)$). O *recall* corresponde à proporção de dentes cariados que o modelo conseguiu identificar corretamente

(TP/(TP + FN)). A acurácia expressa a proporção total de classificações corretas sobre o conjunto avaliado. A especificidade indica a proporção de dentes saudáveis corretamente identificados (TN/(TN + FP)). O *F1-score* corresponde à média harmônica entre precisão e *recall*, sendo particularmente informativo quando há interesse em equilibrar os erros de falso positivo e falso negativo. Os resultados finais de cada arquitetura correspondem à média e ao desvio padrão dos valores obtidos nos cinco *folds*.

Todos os experimentos foram executados na plataforma Google Colaboratory (Colab) com aceleração por GPU Tesla T4 e RAM alta, TensorFlow 2.20.0, CUDA 12.5.1, cuDNN 9 e Keras como *framework* de aprendizado profundo.

4 RESULTADOS E DISCUSSÃO

Os vinte experimentos conduzidos, quatro arquiteturas treinadas e avaliadas em cinco partições de validação cruzada cada, produziram resultados consistentes o suficiente para traçar comparações confiáveis entre os modelos. A Tabela 5 apresenta as médias e os desvios padrão das cinco métricas calculadas sobre o conjunto de validação de cada *fold*.

Tabela 5 – Desempenho médio das quatro arquiteturas na classificação binária de cáries (média ± desvio padrão dos 5 *folds*).

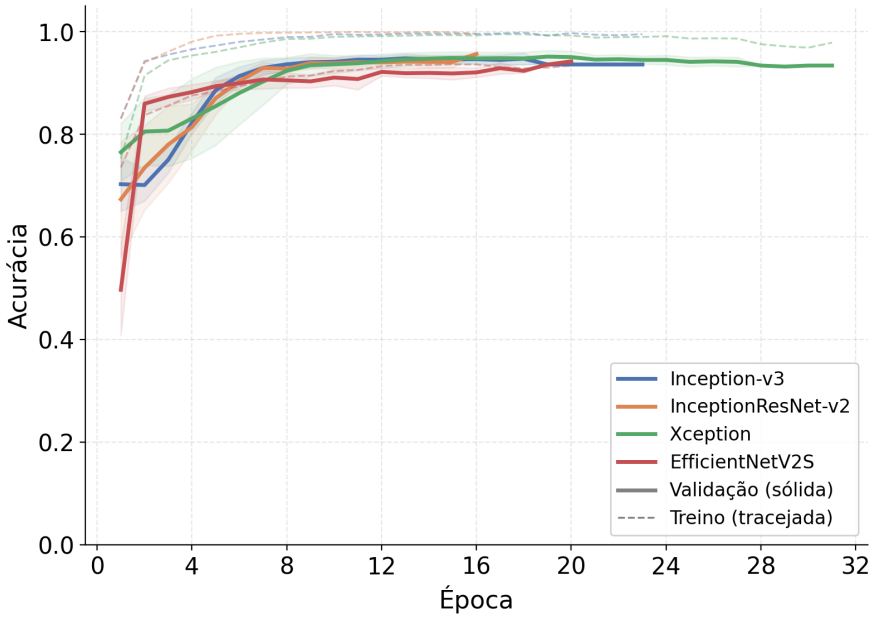
Métrica	**Inception-v3**	**InceptionResNet-v2**	**Xception**	**EfficientNetV2-S**
Precisão	0,9758 ± 0,0119	0,9515 ± 0,0147	0,9818 ± 0,0136	0,9456 ± 0,0133
Recall	0,9112 ± 0,0218	0,9268 ± 0,0247	0,9129 ± 0,0252	0,8784 ± 0,0414
Acurácia	0,9441 ± 0,0075	0,9396 ± 0,0104	0,9478 ± 0,0126	0,9137 ± 0,0183
Especificidade	0,9770 ± 0,0118	0,9523 ± 0,0155	0,9827 ± 0,0131	0,9490 ± 0,0142
F1-*score*	0,9421 ± 0,0085	0,9387 ± 0,0114	0,9458 ± 0,0137	0,9101 ± 0,0214

Fonte: Elaborado pelos autores (2026).

4.1 DESEMPENHO GERAL DAS ARQUITETURAS

O primeiro dado que chama atenção ao observar a Figura 6 é a proximidade entre as três arquiteturas do grupo Inception. Xception, Inception-v3 e InceptionResNet-v2 alcançaram acurácias médias de 94,78%, 94,41% e 93,96%, respectivamente, com intervalos de desvio padrão que se sobrepõem amplamente, o que impede qualquer distinção conclusiva entre elas. O EfficientNetV2-S registrou acurácia média de 91,37%, valor inferior ao do grupo quando comparadas as médias isoladas. No entanto, seu intervalo de desvio padrão (89,54%–93,20%) apresenta sobreposição marginal com o do InceptionResNet-v2 (92,92%–95,00%), o que sugere cautela antes de afirmar diferença estatisticamente significativa em relação a essa arquitetura. Em relação ao Inception-v3 e ao Xception, a separação entre os intervalos é clara e a inferioridade observada nas médias é mais difícil de atribuir apenas à variação amostral.

Figura 6 – Comparativo entre Modelos (média ± desvio padrão dos 5 *folds*).



Fonte: Elaborado pelos autores (2026).

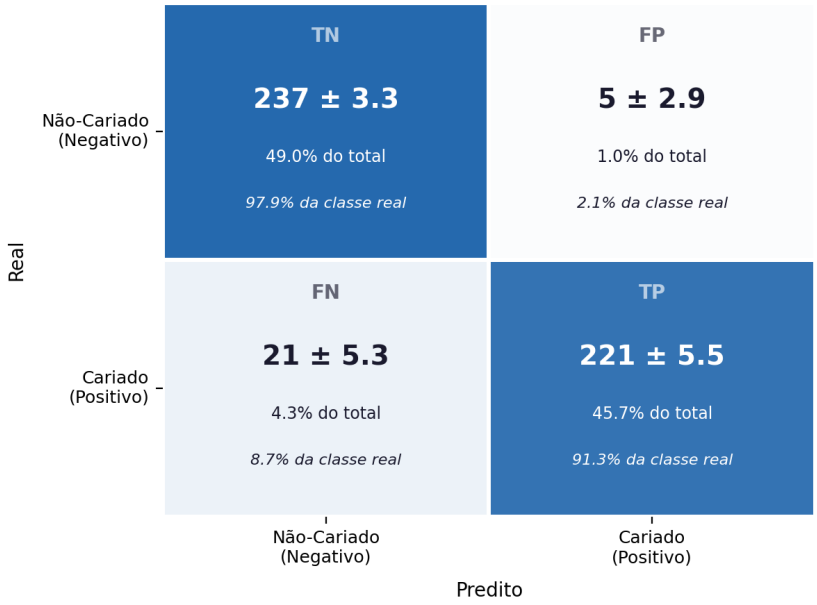
Em todas as arquiteturas, o número de falsos negativos foi superior ao de falsos positivos. Esse resultado indica maior ocorrência de lesões cariosas não detectadas, uma vez que um falso negativo corresponde à falha do modelo em identificar uma lesão existente. Dessa forma, no contexto da triagem odontológica, esse tipo de erro é considerado o mais grave, pois a lesão permanece sem diagnóstico e, consequentemente, sem tratamento, o que favorece sua progressão sem detecção.

4.1.1 Inception-v3

O Inception-v3 encerrou os cinco *folds* com acurácia média de 94,41% e o menor desvio padrão de acurácia entre todos os modelos avaliados (0,75 pontos percentuais). Esses dois números juntos dizem algo relevante sobre o comportamento do modelo ao longo da validação cruzada: ele não apenas acertou bem, mas fez isso de forma previsível em cada partição. Para um conjunto de apenas 2.434 imagens, essa estabilidade não era garantida.

A matriz de confusão média, apresentada na Figura 7, revela o que está por trás da acurácia. O modelo identificou corretamente 237 ± 3,3 dentes saudáveis por *fold*, o que corresponde a 97,9% da classe negativa real. Os erros sobre dentes saudáveis foram raros: apenas 5 ± 2,9 falsos positivos por *fold*, representando 2,1% da classe e 1,0% do total de predições. Entre os dentes cariados, 221 ± 5,5 foram classificados corretamente (91,3% da classe positiva real), enquanto 21 ± 5,3 casos foram perdidos como falsos negativos (8,7% da classe cariada).

Figura 7 – Matriz de confusão média do Inception-v3 obtida pela validação cruzada estratificada de cinco partições.



Fonte: Elaborado pelos autores (2026).

Esse padrão revela uma preferência clara do modelo pela precisão sobre o *recall*. Quando o Inception-v3 classifica um dente como cariado, ele quase sempre está certo. Quando erra, o faz principalmente por omissão, já que 21 lesões passaram sem identificação. Em triagem odontológica, esse tipo de erro tem peso clínico maior do que um falso alarme, pois cáries não detectadas evoluem sem intervenção. O modelo, portanto, favorece a confiança nas suas predições positivas em detrimento da sensibilidade total.

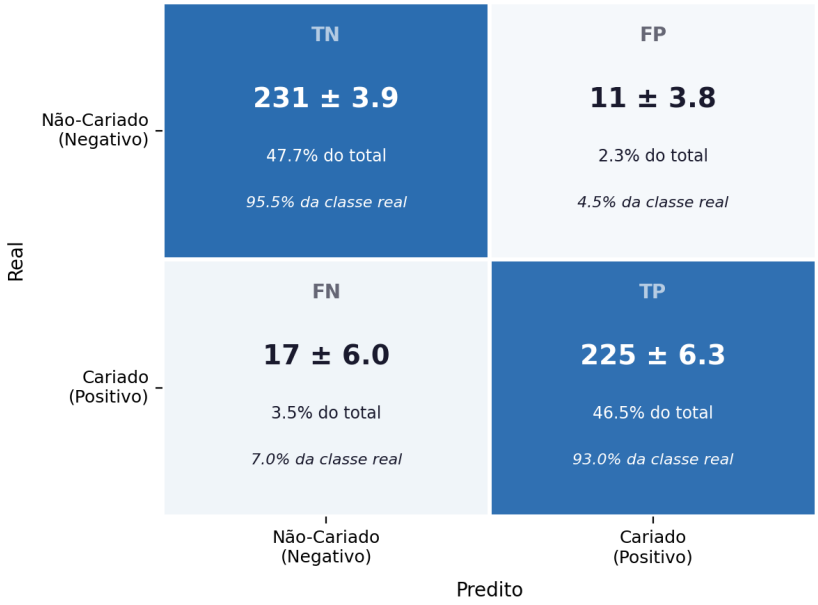
O desvio padrão dos falsos negativos (±5,3) foi o mais baixo entre os quatro modelos nessa categoria, indicando que a quantidade de lesões perdidas variou pouco entre os *folds*. Isso é um traço de consistência que vai além do que a média sozinha transmite.

4.1.2 InceptionResNet-v2

O InceptionResNet-v2 apresentou o perfil oposto ao do Inception-v3. Sua acurácia média (93,96%) ficou ligeiramente abaixo das outras duas arquiteturas do grupo Inception, mas seu *recall* médio de 92,68% foi a mais alta entre todos os modelos avaliados. Esses números, tomados juntos, indicam um modelo que prefere não deixar lesões passarem, mesmo que isso signifique acionar mais alarmes sobre dentes saudáveis.

A matriz de confusão deixa isso explícito. O número médio de falsos negativos por *fold* foi de $17 \pm 6{,}0$, o menor entre todos os modelos. Em termos práticos, o InceptionResNet-v2 perdeu em média quatro lesões a menos por *fold* do que o Inception-v3 e o Xception. Em contrapartida, seus falsos positivos foram $11 \pm 3{,}8$ por *fold*, mais que o dobro do Inception-v3 e quase três vezes os do Xception. A especificidade caiu para 95,23%, a segunda mais baixa da comparação.

Figura 8 – Matriz de confusão média do InceptionResNet-v2 obtida pela validação cruzada estratificada de cinco partições.



Fonte: Elaborado pelos autores (2026).

O desvio padrão dos falsos positivos (±3,8) merece atenção. Ele foi o mais alto entre as três arquiteturas do grupo Inception e indica que a propensão do modelo a classificar incorretamente dentes saudáveis variou com mais intensidade dependendo de quais imagens compunham cada *fold*. Há uma relação provável com a complexidade da arquitetura: com 55 milhões de parâmetros, o InceptionResNet-v2 é mais de duas vezes maior que o Inception-v3 e o Xception. Nesse volume de parâmetros, capturar padrões espúrios dentro de um subconjunto de treinamento pequeno é mais fácil do que capturar padrões generalizáveis, e os falsos positivos oscilantes ao longo dos *folds* podem ser o reflexo disso.

Para aplicações clínicas onde o custo de um falso negativo supera o de um falso positivo, o InceptionResNet-v2 continua sendo uma escolha defensável. O profissional recebe mais alertas para revisar, mas perde menos lesões reais. A decisão de preferir esse modelo ou outro depende de onde o custo de erro é mais tolerável no fluxo de trabalho clínico.

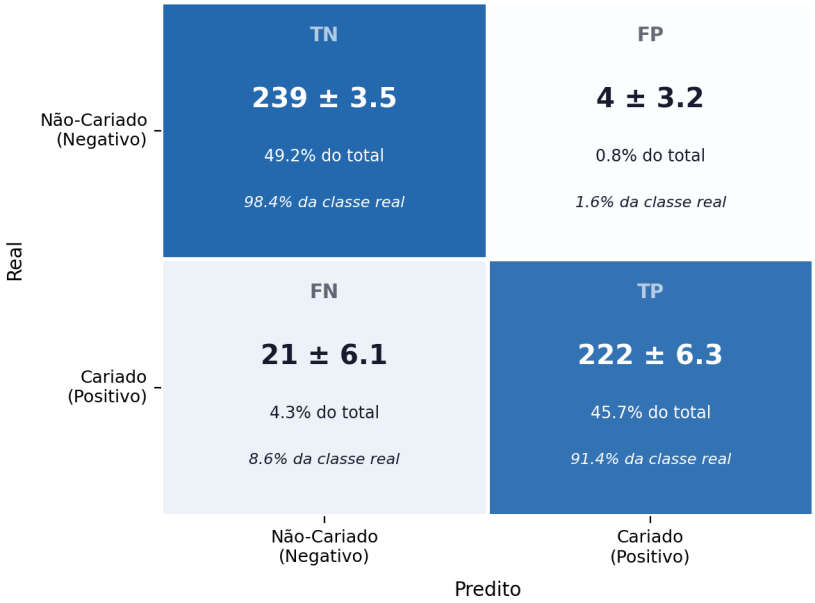
4.1.3 Xception

O Xception registrou a maior precisão média entre todos os modelos avaliados (0,9818) e a maior especificidade (98,27%), combinadas com a segunda melhor acurácia geral (94,78%). Em termos de quadrante da matriz de confusão, o modelo quase nunca errou sobre dentes saudáveis: apenas 4 ± 3,2 falsos positivos por *fold*, o que representa 0,8% do total de predições e 1,6% da classe negativa real. Nenhuma outra arquitetura chegou perto desse nível de proteção sobre a classe saudável.

O número de verdadeiros positivos foi de 222 ± 6,3 por *fold*, correspondendo a 91,4% da classe cariada real. Os falsos negativos ficaram em 21 ± 6,1 por *fold*, valor praticamente

idêntico ao do Inception-v3 (21 casos), mas com desvio padrão 15% maior. Essa variabilidade maior dos falsos negativos no Xception, em relação ao Inception-v3, é um detalhe que a média por si só não transmite: o modelo foi tão conservador em alguns *folds* que perdeu mais lesões do que em outros.

Figura 9 – Matriz de confusão média do Xception obtida pela validação cruzada estratificada de cinco partições.



Fonte: Elaborado pelos autores (2026).

Há algo interessante nessa assimetria. O Xception foi concebido como uma generalização extrema do Inception-v3 com convoluções separáveis por profundidade, projetadas para capturar correlações espaciais e entre canais de forma mais eficiente. Na prática, com imagens radiográficas em escala de cinza e um conjunto de treinamento relativamente pequeno, essa eficiência se traduz em um modelo que aprende fronteiras de decisão mais rígidas para a classe positiva. O resultado é uma taxa de falsos positivos excepcionalmente baixa, ao custo de uma sensibilidade que oscila mais do que a do Inception-v3.

O *F1-score* médio do Xception (0,9458) foi o mais alto entre os quatro modelos, resultado que reflete o equilíbrio entre sua precisão elevada e um *recall* que, apesar de ligeiramente menor que a do InceptionResNet-v2, foi obtida com muito menos falsos positivos. Para contextos onde tanto a confiança nas predições positivas quanto a limitação de alarmes falsos são relevantes, o Xception apresenta o perfil mais equilibrado do grupo.

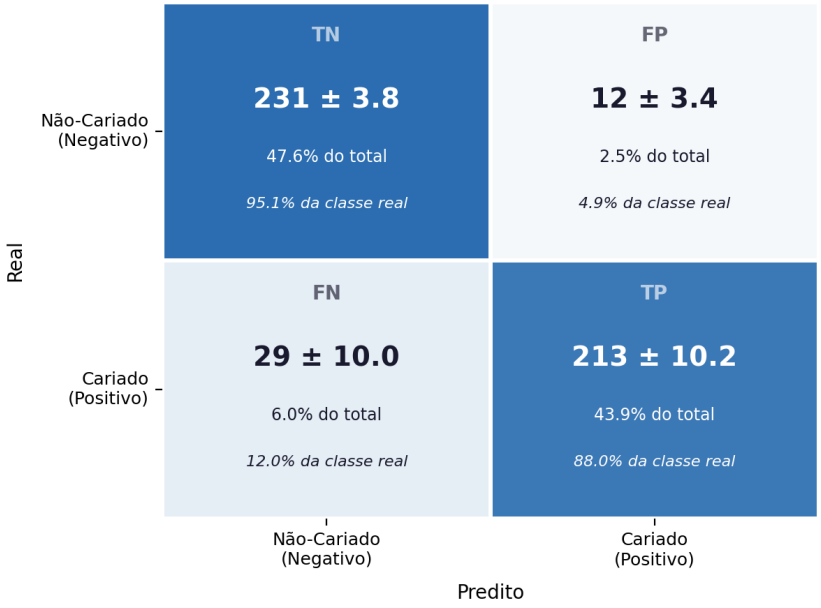
4.1.4 EfficientNetV2-S

O EfficientNetV2-S ficou consistentemente abaixo dos demais modelos em todas as métricas avaliadas. A acurácia média foi de 91,37%, uma diferença de três pontos percentuais em relação ao Xception e de mais de dois pontos em relação ao InceptionResNet-v2. O

recall médio de 87,84% foi a mais baixa da comparação, e o desvio padrão de 4,14 pontos percentuais sinalizou que esse resultado foi também o mais instável entre os *folds*.

A matriz de confusão média revela a magnitude dos problemas. Os falsos negativos chegaram a 29 ± 10,0 por *fold*, número que supera em 38% os falsos negativos do Inception-v3 e do Xception. O desvio padrão de ±10,0 é o dado mais preocupante: enquanto nos demais modelos esse valor variou entre 5 e 6, no EfficientNetV2-S ele foi o dobro, indicando que o número de lesões perdidas variou dramaticamente dependendo de quais imagens compunham cada *fold* de treinamento. Em um *fold*, o modelo pode ter perdido 19 lesões; em outro, 39. Essa variância não é aceitável em qualquer aplicação que exija previsibilidade.

Figura 10 – Matriz de confusão média do EfficientNetV2-S obtida pela validação cruzada estratificada de cinco partições.



Fonte: Elaborado pelos autores (2026).

Os falsos positivos do EfficientNetV2-S foram de 12 ± 3,4 por *fold*, similares aos do InceptionResNet-v2. A diferença é que o InceptionResNet-v2 troca esses falsos positivos por um *recall* alto. O EfficientNetV2-S gerou praticamente o mesmo volume de alarmes incorretos sobre dentes saudáveis sem a compensação de detectar mais lesões reais.

Uma hipótese para esse desempenho está na relação entre a arquitetura e o tamanho do conjunto de treinamento. O EfficientNetV2-S foi projetado com um mecanismo de aprendizado progressivo que aquece gradualmente a resolução das imagens e a intensidade da aumentação ao longo das épocas de treinamento. Com 1.947 imagens por *fold* de treinamento e sem aumentação de dados aplicada neste experimento, parte da lógica de regularização interna da arquitetura nunca chegou a operar nas condições para as quais foi desenvolvida. As três arquiteturas da família Inception, por sua vez, foram concebidas em um período em que os conjuntos de treinamento típicos eram menores e as técnicas de regularização externa

mais modestas, o que pode explicar sua maior robustez nesse regime de dados reduzido.

4.2 AVALIAÇÃO NO SUBCONJUNTO DE INFERÊNCIA

Esta subseção apresenta os resultados das quatro arquiteturas avaliadas sobre o subconjunto de 4.870 imagens mantido completamente fora do ciclo de treinamento e validação cruzada. Cada um dos cinco modelos treinados por arquitetura foi aplicado sobre esse conjunto de forma independente, produzindo cinco avaliações distintas por modelo. Além das métricas individuais por modelo, foi computada uma predição por conjunto (*ensemble*), que combina as probabilidades geradas pelos cinco modelos por média aritmética simples antes da limiarização em 0,5. A Tabela 6 apresenta as médias das cinco avaliações por arquitetura, e a Tabela 7 apresenta os resultados do *ensemble*.

Tabela 6 – Desempenho médio das quatro arquiteturas no subconjunto de inferência (média ± desvio padrão dos cinco modelos treinados, $n = 4.870$).

Métrica	**Inception-v3**	**InceptionResNet-v2**	**Xception**	**EfficientNetV2-S**
Precisão	0,9720 ± 0,0078	0,9381 ± 0,0108	0,9720 ± 0,0059	0,9330 ± 0,0078
Recall	0,9062 ± 0,0038	0,9092 ± 0,0084	0,9116 ± 0,0056	0,8758 ± 0,0206
Acurácia	0,9400 ± 0,0027	0,9246 ± 0,0074	0,9427 ± 0,0015	0,9065 ± 0,0111
Especificidade	0,9738 ± 0,0075	0,9399 ± 0,0111	0,9737 ± 0,0058	0,9371 ± 0,0075
F1-score	0,9379 ± 0,0026	0,9234 ± 0,0075	0,9408 ± 0,0015	0,9034 ± 0,0125

Fonte: Elaborado pelos autores (2026).

Tabela 7 – Desempenho do *ensemble* de cinco modelos no subconjunto de inferência ($n = 4.870$).

Métrica	**Inception-v3**	**InceptionResNet-v2**	**Xception**	**EfficientNetV2-S**
Precisão	0,9795	0,9630	0,9775	0,9389
Recall	0,9035	0,9092	0,9113	0,8891
Acurácia	0,9423	0,9372	0,9452	0,9156
Especificidade	0,9811	0,9651	0,9791	0,9421
F1-score	0,9400	0,9354	0,9433	0,9133

Fonte: Elaborado pelos autores (2026).

O padrão geral observado na validação cruzada se sustentou no subconjunto de inferência. O grupo formado pelo Inception-v3, pelo Xception e pelo InceptionResNet-v2 permaneceu claramente acima do EfficientNetV2-S em todas as métricas, com o Xception registrando a maior acurácia de *ensemble* (94,52%), o maior *F1-score* (0,9433). O Inception-v3 reteve a maior precisão entre todos os modelos, chegando a 0,9795 no *ensemble*, com apenas 46 falsos positivos sobre 4.870 imagens.

O aspecto mais relevante dos resultados de inferência não é necessariamente o valor absoluto das métricas, mas sua estabilidade em relação à validação cruzada. Os modelos foram treinados sobre no máximo 1.948 imagens por *fold* (uma fração do conjunto de

treinamento) e avaliados aqui sobre as 4.870 imagens do conjunto de inferência, mais que o dobro do volume total usado em todo o ciclo de treinamento e validação (2.434 imagens). O fato da acurácia do *ensemble* do Xception ter caído apenas 0,26 pontos percentuais em relação à média de validação cruzada (de 94,78% para 94,52%) indica que as representações aprendidas generalizaram sem degradação expressiva.

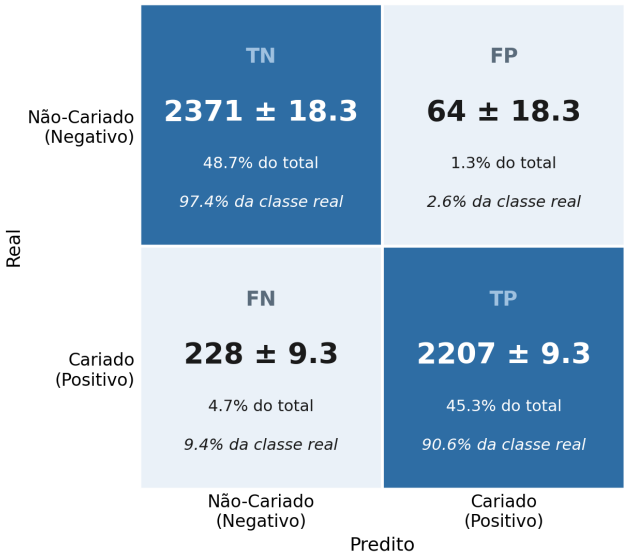
Cabe notar que a comparação contrapõe uma média de cinco execuções independentes (validação cruzada) a um resultado único e determinístico (*ensemble* de inferência), de modo que a variação reportada reflete diferença de valor central, não de dispersão estatística entre as duas fases.

4.2.1 Inception-v3 na inferência

O Inception-v3 confirmou na inferência o traço mais consistente observado na validação cruzada que é o baixo número de falsos positivos. Nas matrizes de confusão médias sobre os cinco modelos, o valor foi de 64 ± 18,3 por avaliação, o que representa 1,3% do total de predições. No *ensemble*, esse número caiu para 46, correspondendo a apenas 0,9% do total, a menor taxa de falsos positivos entre todas as combinações avaliadas. Sua precisão média individual ($0,9720 \pm 0,0078$) acompanhou de perto a do Xception, reforçando que o baixo volume de falsos positivos não foi um resultado isolado de um único modelo, mas um traço estável entre os cinco treinados.

O desvio padrão de falsos negativos (±9,3) foi também o menor do grupo entre os três modelos Inception, reforçando a estabilidade que o Inception-v3 já havia demonstrado na fase de validação. A matriz de confusão média pode ser vista na Figura 11.

Figura 11 – Matriz de confusão média do Inception-v3 no subconjunto de inferência, calculada a partir dos cinco modelos individuais e não do *ensemble* ($n = 4.870$).



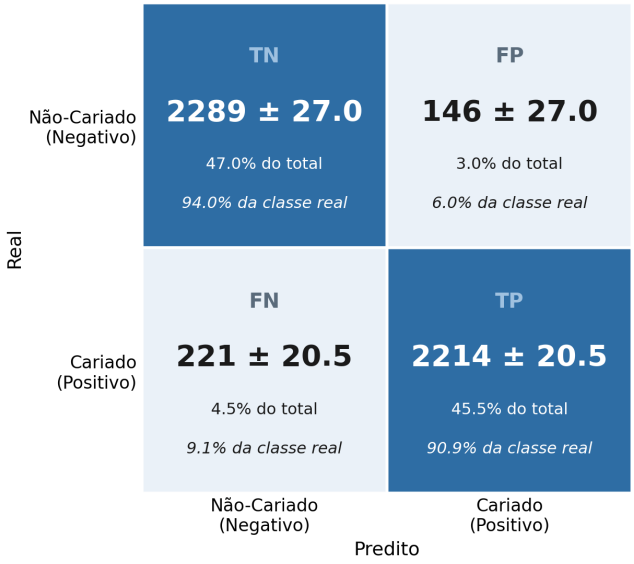
Fonte: Elaborado pelos autores (2026).

4.2.2 InceptionResNet-v2 na inferência

O InceptionResNet-v2 apresentou a variação de comportamento mais notável entre as fases de validação cruzada e inferência. Na validação, esse modelo registrou o maior *recall* médio entre as quatro arquiteturas (0,9268). Na inferência, seu *recall* caiu para $0,9092 \pm 0,0084$ na média dos cinco modelos, valor inferior ao do Xception ($0,9116 \pm 0,0056$) e próximo ao do Inception-v3 ($0,9062 \pm 0,0038$). A vantagem que diferenciava o InceptionResNet-v2 na detecção de lesões reduziu-se ao ponto de não mais o distinguir dos demais modelos do grupo.

Esse deslocamento tem uma interpretação possível, com apenas 1.947 imagens por *fold* de treinamento, o modelo pode ter ajustado seu limiar de decisão de forma excessivamente adaptada às características específicas de cada partição. No conjunto de inferência, mais amplo e com distribuição independente, esse ajuste não se transferiu com a mesma eficiência. A Figura 12 ilustra a matriz de confusão média, com $221 \pm 20,5$ falsos negativos, desvio padrão mais alto entre os três modelos Inception.

Figura 12 – Matriz de confusão média do InceptionResNet-v2 no subconjunto de inferência, calculada a partir dos cinco modelos individuais e não do *ensemble* ($n = 4.870$).



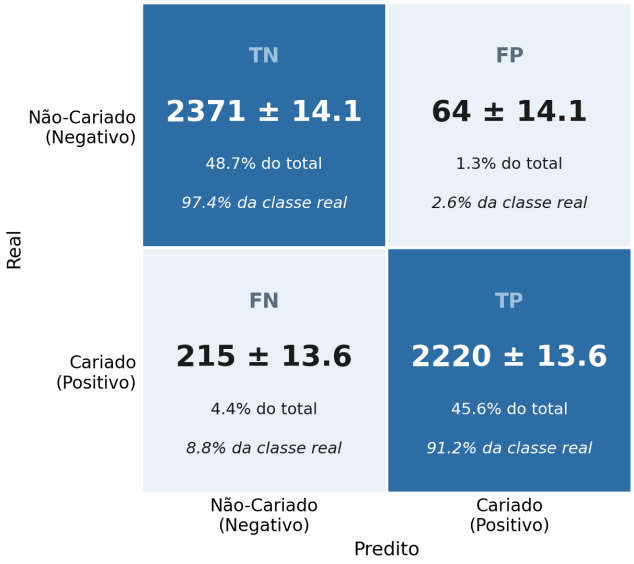
Fonte: Elaborado pelos autores (2026).

4.2.3 Xception na inferência

O Xception foi o modelo que apresentou o desempenho mais equilibrado entre as fases de validação e inferência. Seu *recall* médio nos cinco modelos de inferência foi de $0,9116 \pm 0,0056$, ligeiramente superior à observada na validação cruzada (0,9129), diferença dentro da margem de variabilidade estatística. Ao mesmo tempo, manteve a maior especificidade do grupo ($0,9737 \pm 0,0058$) e a maior precisão ($0,9720 \pm 0,0059$, empatada com o Inception-v3), o que indica que o modelo se manteve robusto em dados completamente novos.

A matriz de confusão média, apresentada na Figura 13, mostra 215 ± 13,6 falsos negativos, o menor valor médio entre todas as arquiteturas na fase de inferência, superando o InceptionResNet-v2 que havia liderado essa métrica na validação cruzada. O desvio padrão de 13,6 é também moderado, indicando comportamento estável entre os cinco modelos avaliados, o que se reflete também no *F1-score* médio mais estreito do grupo ($0,9408 \pm 0,0015$).

Figura 13 – Matriz de confusão média do Xception no subconjunto de inferência, calculada a partir dos cinco modelos individuais e não do *ensemble* ($n = 4.870$).



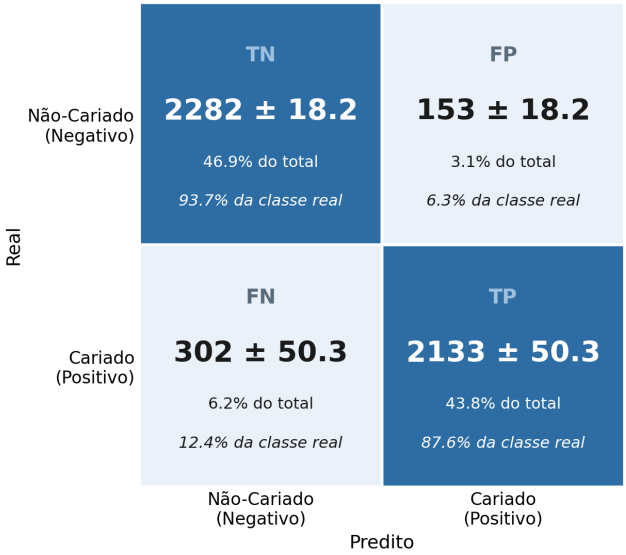
Fonte: Elaborado pelos autores (2026).

4.2.4 EfficientNetV2-S na inferência

O EfficientNetV2-S manteve na inferência o padrão observado na validação cruzada, porém com um agravante relevante, o desvio padrão de falsos negativos foi de ±50,3, mais que o triplo do segundo pior resultado do grupo (InceptionResNet-v2, ±20,5). Esse comportamento já era visível no próprio *recall* médio, que ficou em $0,8758 \pm 0,0206$, o maior desvio padrão entre as quatro arquiteturas e mais de duas vezes o do segundo colocado nesse quesito (InceptionResNet-v2, ±0,0084). Em termos concretos, dependendo de qual dos cinco modelos é utilizado, o número de lesões não detectadas pode variar entre aproximadamente 252 e 352 sobre as mesmas 4.870 imagens. Essa variância compromete qualquer tentativa de caracterizar o modelo por um único valor médio.

O *ensemble* atenuou parcialmente o problema: com 270 falsos negativos, o resultado combinado ficou abaixo da média individual de $302 \pm 50,3$, demonstrando que a combinação de modelos corrige parte da instabilidade de cada modelo isolado. Ainda assim, a acurácia de *ensemble* de 91,56% permanece cerca de três pontos percentuais abaixo do Xception. A matriz de confusão média está apresentada na Figura 14.

Figura 14 – Matriz de confusão média do EfficientNetV2-S no subconjunto de inferência, calculada a partir dos cinco modelos individuais e não do *ensemble* (n = 4.870).



Fonte: Elaborado pelos autores (2026).

4.2.5 Efeito do *ensemble* e comparação entre fases

A combinação dos cinco modelos por média de probabilidades produziu ganhos consistentes de precisão em todas as arquiteturas. O número de falsos positivos caiu do valor médio individual para o valor de *ensemble* em todos os casos: de 64 para 46 no Inception-v3, de 146 para 85 no InceptionResNet-v2, de 64 para 51 no Xception e de 153 para 141 no EfficientNetV2-S. Esse padrão é esperado, já que modelos treinados em diferentes partições tendem a errar em imagens distintas, e a média das probabilidades reduz os alarmes que não são sustentados pela maioria dos modelos.

A Tabela 8 compara os resultados de *F1-score* e acurácia entre a validação cruzada e o *ensemble* de inferência, ilustrando a variação entre as duas fases para cada arquitetura.

Tabela 8 – Comparação de *F1-score* e acurácia entre validação cruzada e *ensemble* de inferência.

Arquitetura	**Validação cruzada**		***Ensemble* de inferência**	
	F1	**Acurácia**	**F1**	**Acurácia**
Inception-v3	0,9421	0,9441	0,9400	0,9423
InceptionResNet-v2	0,9387	0,9396	0,9354	0,9372
Xception	0,9458	0,9478	0,9433	0,9452
EfficientNetV2-S	0,9101	0,9137	0,9133	0,9156

Fonte: Elaborado pelos autores (2026).

As quedas observadas são pequenas e distribuídas de forma consistente, o Xception perdeu 0,25 pontos percentuais de acurácia, o Inception-v3 perdeu 0,18, e o InceptionResNet-v2 perdeu 0,24. O EfficientNetV2-S, em sentido contrário, registrou um leve aumento de 0,19

pontos percentuais no *ensemble* de inferência em relação à validação, o que não é suficiente para alterar a hierarquia entre os modelos, mas indica que seu comportamento no conjunto independente foi ligeiramente mais favorável do que na validação cruzada.

Em conjunto, os resultados de inferência confirmam que as três arquiteturas da família Inception aprendem representações que generalizam bem para dados completamente novos, e que o Xception, em particular, manteve o melhor equilíbrio entre precisão e *recall* sobre o maior conjunto de dados avaliado neste trabalho.

5 CONSIDERAÇÕES FINAIS

Este trabalho investigou o desempenho de quatro arquiteturas de redes neurais convolucionais, Inception-v3, InceptionResNet-v2, Xception e EfficientNetV2-S, na classificação binária de lesões cariosas em recortes individuais de dentes extraídos de radiografias panorâmicas. Os experimentos foram conduzidos sobre 2.434 imagens balanceadas provenientes de dois *datasets* com origens distintas, o InReDD e o DENTEX, com validação cruzada estratificada de cinco partições aplicada a cada arquitetura.

As três arquiteturas da família Inception demonstraram capacidade de aprender representações úteis mesmo com um subconjunto de treinamento relativamente pequeno. O Xception registrou o melhor *F1-score* médio (0,9458) e a maior precisão (0,9818), com apenas 4 ± 3,2 falsos positivos por *fold*. O Inception-v3 apresentou o menor desvio padrão de acurácia entre os cinco *folds*, indicando maior estabilidade preditiva. O InceptionResNet-v2 obteve o maior *recall* média (0,9268), perdendo menos lesões por *fold*, o que pode ser relevante em contextos de triagem onde falsos negativos têm custo clínico mais alto. O EfficientNetV2-S ficou abaixo dos demais em todas as métricas, com acurácia média de 91,37% e desvio padrão do *recall* de 4,14 pontos percentuais, resultado provavelmente relacionado à incompatibilidade entre seu mecanismo de aprendizado progressivo e o volume reduzido de dados disponível neste experimento.

Um padrão observado em todos os modelos merece atenção. Em todas as arquiteturas, o número de falsos negativos foi superior ao de falsos positivos. Esse resultado indica maior ocorrência de lesões cariosas não detectadas, uma vez que um falso negativo corresponde à falha do modelo em identificar uma lesão existente. No contexto da triagem odontológica, esse tipo de erro é considerado o mais grave, pois a lesão permanece sem diagnóstico e, consequentemente, favorece sua progressão sem detecção.

O Xception demonstrou desempenho competitivo com as arquiteturas já estabelecidas na literatura para essa tarefa. O EfficientNetV2-S, apesar dos resultados inferiores nas condições deste experimento, permanece como candidato para investigação futura em protocolos com maior volume de dados. A avaliação sobre os dois terços do conjunto reservados para inferência, 4.870 imagens completamente fora do ciclo de treinamento e validação cruzada, confirmou os padrões observados nas partições de validação. O Xception manteve o melhor

equilíbrio entre precisão e revocação, com acurácia de *ensemble* de 94,52% e *F1-score* de 0,9433, enquanto o Inception-v3 registrou a menor taxa de falsos positivos, chegando a 46 alarmes incorretos sobre 4.870 imagens no *ensemble*. As quedas de desempenho em relação à validação cruzada foram pequenas e consistentes, não ultrapassando 0,26 pontos percentuais de acurácia para nenhuma das arquiteturas do grupo Inception, o que indica que as representações aprendidas generalizaram sem degradação expressiva para dados completamente novos.

Para trabalhos futuros, três caminhos podem ser considerados. O primeiro é a incorporação de técnicas de explicabilidade, como o Grad-CAM (*Gradient-weighted Class Activation Mapping*), que gera mapas de ativação evidenciando as regiões da imagem que influenciaram a classificação (Selvaraju *et al.*, 2019). Em um domínio clínico, isso tem valor prático direto, já que um mapa que coincide com a lesão reforça a confiança do profissional, enquanto ativações em regiões irrelevantes levantam suspeita sobre o aprendizado do modelo.

O segundo é migrar da classificação binária por recorte para detecção de objetos diretamente sobre a radiografia panorâmica completa, eliminando a etapa de extração manual dos dentes individuais. Os conjuntos InReDD e DENTEX já fornecem a base necessária para esse experimento. Os resultados obtidos aqui, com arquiteturas treinadas sobre pouco mais de duas mil imagens alcançando acurácia acima de 93%, indicam que a direção é viável e que o principal gargalo é a disponibilidade de dados anotados com qualidade uniforme, não a capacidade das arquiteturas.

O terceiro é a otimização do limiar de decisão (*threshold*). Todos os experimentos usaram o valor padrão de 0,5, sem ajuste específico ao problema. Como os falsos negativos superaram os falsos positivos em todas as arquiteturas, um erro mais grave no contexto clínico, a curva de *Precision-Recall* permitiria identificar um limiar que privilegie *recall* sem comprometer precisão de forma significativa. Esse ajuste é independente da arquitetura e pode ser aplicada posteriormente, sem novo treinamento.

REFERÊNCIAS

ALSUBIHEEN, Abdulrahman M. *et al.* Examining the diagnostic accuracy of artificial intelligence for detecting dental caries across a range of imaging modalities: An umbrella review with meta-analysis. **PLOS ONE**, Public Library of Science, v. 20, n. 8, e0329986, 2025. DOI: 10.1371/journal.pone.0329986.

BAYRAKTAR, Yavuz; AYAN, Enes. An Explainable Deep Learning Model to Prediction Dental Caries Using Panoramic Radiograph Images. **Diagnostics**, MDPI, v. 13, n. 2, p. 226, 2023. DOI: 10.3390/diagnostics13020226.

BRASIL. **SB Brasil 2023: Pesquisa Nacional de Saúde Bucal. Relatório Final**. 1. ed. rev. Brasília, 2024. Disponível em: https://www.gov.br/saude/pt-br/composicao/saps/brasil-

sorridente/sb-brasil/publicacoes/primeiros-resultados-sb-brasil-2023/. Acesso em: 25 maio 2026.

CHOLLET, François. Xception: Deep Learning with Depthwise Separable Convolutions. *In:* PROCEEDINGS of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [*S. l.: s. n.*], jul. 2017. p. 1251–1258. DOI: 10.1109/CVPR.2017.195.

COSTA, Eliana Dantas *et al.* Development of a dental digital data set for research in artificial intelligence: the importance of labeling performed by radiologists. **Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology**, v. 138, n. 1, p. 205–213, 2024. ISSN 2212-4403. DOI: https://doi.org/10.1016/j.oooo.2023.12.006. Disponível em: https://www.sciencedirect.com/science/article/pii/S2212440323007654.

DAYI, Burak *et al.* A Novel Deep Learning-Based Approach for Segmentation of Different Type Caries Lesions on Panoramic Radiographs. **Diagnostics**, v. 13, n. 2, p. 202, 2023. DOI: 10.3390/diagnostics13020202.

FEJERSKOV, Ole; NYVAD, Bente; KIDD, Edwina (ed.). **Dental Caries: The Disease and Its Clinical Management**. 3. ed. Oxford: Wiley-Blackwell, 2015. ISBN 9781118935828. Disponível em: https://library.knu.edu.af/opac/temp/13861.pdf.

GAO, S. *et al.* Artificial Intelligence in Dentistry: A Narrative Review of Diagnostic and Therapeutic Applications. **Medical Science Monitor**, v. 31, e946676, abr. 2025. DOI: 10.12659/MSM.946676.

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. **Deep Learning**. Cambridge, MA: MIT Press, 2016. ISBN 9780262035613. Disponível em: http://www.deeplearningbook.org.

HAGHANIFAR, Arman *et al.* PaXNet: Tooth Segmentation and Dental Caries Detection in Panoramic X-ray Using Ensemble Transfer Learning and Capsule Classifier. **Multimedia Tools and Applications**, Springer, v. 82, p. 27659–27679, 2023. DOI: 10.1007/s11042-023-14435-9.

HAMAMCI, Ibrahim Ethem *et al.* DENTEX: An Abnormal Tooth Detection with Dental Enumeration and Diagnosis Benchmark for Panoramic X-rays. **arXiv preprint arXiv:2305.19112**, 2023. Disponível em: https://doi.org/10.48550/arXiv.2305.19112.

HUNG, Man *et al.* Charting New Territory: AI Applications in Dental Caries Detection from Panoramic Imaging. **Dentistry Journal**, v. 13, n. 8, 2025. ISSN 2304-6767. DOI: 10.3390/dj13080366. Disponível em: https://www.mdpi.com/2304-6767/13/8/366.

LECUN, Yann; BENGIO, Yoshua; HINTON, Geoffrey. Deep learning. **nature**, Nature Publishing Group, v. 521, n. 7553, p. 436, 2015. Disponível em: https://www.bibsonomy.org/bibtex/2fcb4570f4481d2e36239cc683ce7261e/msteininger.

LI, Yan *et al.* Changes in the global burden of untreated dental caries from 1990 to 2019: A systematic analysis for the Global Burden of Disease study. **Heliyon**, Elsevier, v. 8, n. 9, e10714, 2022. DOI: 10.1016/j.heliyon.2022.e10714.

LIMA, José Nilton Silva. **caries-cnn-panoramic: código dos experimentos de comparação de arquiteturas CNN para classificação de cáries em radiografias panorâmicas**. [*S. l.*]: GitHub, 2026. Disponível em: https://github.com/j-nilton/caries-cnn-panoramic.

MESECI, Elif. An enhanced deep learning model for detection and classification of dental caries in panoramic radiographs. **Neural Computing and Applications**, Springer London, v. 38, n. 1, p. 3, 2026. DOI: 10.1007/s00521-025-11730-4.

NEGI, Sapna *et al.* Artificial Intelligence in Dental Caries Diagnosis and Detection: An Umbrella Review. **Clinical and Experimental Dental Research**, Wiley, v. 10, n. 4, e70004, 2024. DOI: 10.1002/cre2.70004.

PEDREGOSA, Fabian *et al.* **Scikit-learn: Machine Learning in Python**. [*S. l.: s. n.*], 2018. arXiv: 1201.0490 [cs.LG]. Disponível em: https://arxiv.org/abs/1201.0490.

RAMÍREZ-PEDRAZA, Alfonso *et al.* Deep learning in oral hygiene: automated dental plaque detection via YOLO frameworks and quantification using the O'Leary index. **Diagnostics**, MDPI, v. 15, n. 2, p. 231, 2025. DOI: https://doi.org/10.3390/diagnostics15020231.

ROBOFLOW. **Detecting Oral Diseaese Dataset**. [*S. l.*]: Roboflow, jul. 2025. https://universe.roboflow.com/sample-data-training/detecting-oral-diseaese. Disponível em: https://universe.roboflow.com/sample-data-training/detecting-oral-diseaese. Acesso em: 30 maio 2026.

RUSSELL, Stuart J.; NORVIG, Peter. **Inteligência artificial**. Tradução: Regina Célia Simille. Rio de Janeiro: Elsevier, 2013. Tradução de: Artificial intelligence, 3rd ed. ISBN 978-85-352-3701-6. Disponível em: https://www.kufunda.net/publicdocs/Intelig%C3%AAncia%20Artificial%20(Peter%20Norvig,%20Stuart%20Russell).pdf.

SELVARAJU, Ramprasaath R. *et al.* Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. **International Journal of Computer Vision**, Springer Science e Business Media LLC, v. 128, n. 2, p. 336–359, out. 2019. ISSN 1573-1405. DOI: 10.1007/s11263-019-01228-7.

SZEGEDY, Christian; IOFFE, Sergey *et al.* Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *In:* 1. PROCEEDINGS of the AAAI Conference on Artificial Intelligence. [*S. l.: s. n.*], 2017. v. 31. DOI: 10.1609/aaai.v31i1.11231.

SZEGEDY, Christian; VANHOUCKE, Vincent *et al.* Rethinking the Inception Architecture for Computer Vision. *In:* PROCEEDINGS of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [*S. l.: s. n.*], jun. 2016. p. 2818–2826. DOI: 10.1109/CVPR.2016.308.

TAN, Mingxing; LE, Quoc V. EfficientNetV2: Smaller Models and Faster Training. *In:* PROCEEDINGS of the 38th International Conference on Machine Learning (ICML). [*S. l.: s. n.*], 2021. DOI: 10.48550/arXiv.2104.00298. arXiv: 2104.00298.

TAVARES, Raul Felipe Almeida Guimarães *et al.* AVANÇOS NA DETECÇÃO PRECOCE DO CÂNCER BUCAL: TECNOLOGIAS E DESAFIOS. **Revista Ibero-Americana de Humanidades, Ciências e Educação**, v. 11, n. 6, p. 621–633, jun. 2025. DOI: 10.51891/rease.v11i6.19625. Disponível em: https://periodicorease.pro.br/rease/article/view/19625.

UEHARA MARTINS, Caio *et al.* InReDD-Dataset-PAN924. **PhysioNet**, nov. 2025. Version 1.0.0. DOI: 10.13026/r5nt-we67. Disponível em: https://doi.org/10.13026/r5nt-we67.

VINAYAHALINGAM, Shankeeth *et al.* Classification of caries in third molars on panoramic radiographs using deep learning. **Scientific Reports**, v. 11, p. 12609, 2021. DOI: 10.1038/s41598-021-92121-2.

WANG, Siwei *et al.* The global, regional, and national burden of oral disorders in 204 countries and territories, 1990–2021, and projections to 2050: A systematic analysis for the Global Burden of Disease Study 2021. **Regenesis Repair Rehabilitation**, v. 1, n. 3, p. 56–63, 2025. ISSN 2950-5755. DOI: https://doi.org/10.1016/j.rerere.2025.07.001.

WORLD HEALTH ORGANIZATION. **Global Oral Health Status Report: Towards Universal Health Coverage for Oral Health by 2030**. Geneva, 2022. Disponível em: https://www.who.int/publications/i/item/9789240061484. Acesso em: 25 maio 2026.

ZANCAN, B. G. *et al.* AI-powered precision in dental radiographic analysis using tailored CNNs for tooth numbering and cavity detection. **PLOS Digital Health**, v. 4, n. 11, e0001074, 2025. DOI: 10.1371/journal.pdig.0001074. Disponível em: https://doi.org/10.1371/journal.pdig.0001074.

ZHANG, Aston *et al.* **Dive into Deep Learning**. [*S. l.*]: Cambridge University Press, 2023. https://D2L.ai.