

UNIVERSIDADE FEDERAL DO PARANÁ

ALEX MATSUO

CLASSIFICAÇÃO DE CÁRIES DENTÁRIAS UTILIZANDO ENSEMBLE DE REDES  
NEURAIS CONVOLUCIONAIS

CURITIBA PR

2025

ALEX MATSUO

CLASSIFICAÇÃO DE CÁRIES DENTÁRIAS UTILIZANDO ENSEMBLE DE REDES  
NEURAIS CONVOLUCIONAIS

Trabalho apresentado como requisito parcial à conclusão  
do Curso de Bacharelado em Ciências da Computação,  
Setor de Ciências Exatas, da Universidade Federal do  
Paraná.

Área de concentração: *Computação*.

Orientador: Lucas Ferrari de Oliveira.

CURITIBA PR

2025





*Dedico a todos que me apoiaram  
durante essa jornada.*

## **AGRADECIMENTOS**

Agradeço ao meu orientador, Prof. Lucas Ferrari de Oliveira, pela orientação, paciência e conhecimento compartilhado durante o desenvolvimento deste trabalho.

Aos meus familiares e amigos pelo apoio incondicional e compreensão durante esta jornada acadêmica.

À Universidade Federal do Paraná e ao Departamento de Informática pela oportunidade de crescimento e aprendizado.

A todos que, direta ou indiretamente, contribuíram para a realização deste trabalho.

## **RESUMO**

Este trabalho apresenta o desenvolvimento de um sistema automatizado para classificação de cáries dentárias em imagens utilizando técnicas de deep learning. Foi implementado um modelo ensemble que combina duas arquiteturas de redes neurais convolucionais: ConvNeXt e YOLO11. O sistema é capaz de classificar imagens em cinco categorias distintas de cáries dentárias (bc, c4, c5, c6, hg), alcançando uma acurácia superior aos modelos individuais. A metodologia proposta demonstra o potencial de sistemas híbridos na assistência ao diagnóstico odontológico, oferecendo uma ferramenta auxiliar para profissionais da área.

Palavras-chave: Deep Learning, Classificação de Imagens, Cáries Dentárias, Ensemble Learning, ConvNeXt, YOLO11

## **ABSTRACT**

This work presents the development of an automated system for dental caries classification in radiographic images using deep learning techniques. An ensemble model was implemented combining two convolutional neural network architectures: ConvNeXt and YOLO11. The system is capable of classifying images into five distinct dental caries categories (bc, c4, c5, c6, hg), achieving higher accuracy than individual models. The proposed methodology demonstrates the potential of hybrid systems in assisting dental diagnosis, offering an auxiliary tool for dental professionals.

Keywords: Deep Learning, Image Classification, Dental Caries, Ensemble Learning, ConvNeXt, YOLO11

## LISTA DE FIGURAS

4.1	Acurácia geral do modelo ConvNeXt durante o treinamento e validação. Fonte: Alex Matsuo (2025) . . . . .	23
4.2	Acurácia do ConvNeXt por classe no conjunto de teste e validação. Fonte: Alex Matsuo (2025) . . . . .	24
4.3	Gráfico radar das métricas de desempenho do ConvNeXt por classe (F1-Score, Precision e Recall). Fonte: Alex Matsuo (2025) . . . . .	24
4.4	Matrix de Confusão do ConvNeXt no conjunto de teste. Fonte: Alex Matsuo (2025) . . . . .	25
4.5	Amostras de imagens classificadas pelo ConvNeXt. Fonte: Alex Matsuo (2025) .	26
4.6	Acurácia geral do modelo YOLO11 durante o treinamento e validação. Fonte: Alex Matsuo (2025) . . . . .	26
4.7	Acurácia do YOLO11 por classe no conjunto de teste e validação. Fonte: Alex Matsuo (2025) . . . . .	27
4.8	Gráfico radar das métricas de desempenho do YOLO11 por classe (F1-Score, Precision e Recall). Fonte: Alex Matsuo (2025) . . . . .	28
4.9	Matrix de Confusão do YOLO11 no conjunto de teste. Fonte: Alex Matsuo (2025)	29
4.10	Amostras de imagens classificadas pelo YOLO11. Fonte: Alex Matsuo (2025) . .	29
4.11	Comparação de desempenho entre os modelos individuais e métodos de ensemble. Fonte: Alex Matsuo (2025) . . . . .	30

## **LISTA DE QUADROS**

1.1	Estágios de cáries e suas características . . . . .	14
4.1	Comparação dos métodos de ensemble . . . . .	30

## **LISTA DE ACRÔNIMOS**

AI	Artificial Intelligence
CNN	Convolutional Neural Network
DInf	Departamento de Informática
GAP	Global Average Pooling
GELU	Gaussian Error Linear Unit
ICDAS	International Caries Detection and Assessment System
PPGINF	Programa de Pós-Graduação em Informática
ReLU	Rectified Linear Unit
SPPF	Spatial Pyramid Pooling - Fast
TCC	Trabalho de Conclusão de Curso
UFPR	Universidade Federal do Paraná
YOLO	You Only Look Once

## **LISTA DE SÍMBOLOS**

$K$	Kernel ou filtro convolucional
$p$	Probabilidade
$\epsilon$	Epsilon (valor pequeno para evitar divisão por zero)
*	Operação de convolução

## SUMÁRIO

<b>1</b>	<b>INTRODUÇÃO . . . . .</b>	<b>13</b>
1.1	CONTEXTUALIZAÇÃO . . . . .	13
1.2	JUSTIFICATIVA . . . . .	14
1.3	OBJETIVOS . . . . .	14
1.3.1	Objetivo Geral. . . . .	14
1.3.2	Objetivos Específicos . . . . .	14
<b>2</b>	<b>FUNDAMENTAÇÃO TEÓRICA. . . . .</b>	<b>16</b>
2.1	REDES NEURAIS CONVOLUCIONAIS . . . . .	16
2.1.1	Fundamentos das CNNs . . . . .	16
2.1.2	ConvNeXt. . . . .	16
2.1.3	YOLO 11 . . . . .	17
2.2	ENSEMBLE LEARNING . . . . .	19
2.2.1	Fundamentos . . . . .	19
2.2.2	Tipos de Ensemble . . . . .	19
<b>3</b>	<b>METODOLOGIA . . . . .</b>	<b>21</b>
3.1	DATASET. . . . .	21
3.1.1	Coleta e Organização dos Dados . . . . .	21
3.1.2	Pré-processamento . . . . .	21
3.2	IMPLEMENTAÇÃO DOS MODELOS . . . . .	21
3.2.1	ConvNeXt. . . . .	21
3.2.2	YOLO11 . . . . .	22
3.3	MÉTRICAS DE AVALIAÇÃO . . . . .	22
<b>4</b>	<b>RESULTADOS E DISCUSSÃO . . . . .</b>	<b>23</b>
4.1	DESEMPENHO DOS MODELOS INDIVIDUAIS . . . . .	23
4.1.1	ConvNeXt. . . . .	23
4.1.2	YOLO11 . . . . .	26
4.2	COMPARAÇÃO DOS MÉTODOS DE ENSEMBLE . . . . .	29
4.2.1	Análise Comparativa . . . . .	30
4.3	DISCUSSÃO . . . . .	31
4.3.1	Vantagens do Ensemble. . . . .	31
4.3.2	Limitações Identificadas . . . . .	31
4.3.3	Implicações Clínicas . . . . .	31

<b>5</b>	<b>CONCLUSÃO . . . . .</b>	<b>32</b>
5.1	SÍNTESE DOS RESULTADOS . . . . .	32
5.2	CONTRIBUIÇÕES . . . . .	32
5.2.1	Contribuições Técnicas . . . . .	32
5.2.2	Contribuições Metodológicas . . . . .	32
5.2.3	Contribuições Práticas . . . . .	32
5.3	LIMITAÇÕES E TRABALHOS FUTUROS . . . . .	33
5.3.1	Expansão e Diversificação . . . . .	33
5.3.2	Integração Clínica . . . . .	33
5.3.3	Extensão Diagnóstica . . . . .	33
5.4	CONSIDERAÇÕES FINAIS . . . . .	33
	<b>REFERÊNCIAS . . . . .</b>	<b>34</b>

## 1 INTRODUÇÃO

### 1.1 CONTEXTUALIZAÇÃO

A cárie dentária representa um dos problemas de saúde bucal mais prevalentes em todo o mundo, afetando bilhões de pessoas e gerando significativos custos de tratamento. O diagnóstico precoce e preciso das cárries é fundamental para o planejamento terapêutico adequado e a preservação da estrutura dentária.(Hung et al., 2019)

Tradicionalmente, o diagnóstico de cárries depende da experiência clínica do profissional e da interpretação visual, processo que pode ser subjetivo e sujeito a variações inter e intra-observador. Neste contexto, o desenvolvimento de sistemas automatizados de diagnóstico assistido por computador surge como uma alternativa promissora para aumentar a precisão e consistência do diagnóstico.

Para a identificação e classificação da doença, um dos índices que tem sido muito utilizado na atualidade é o Sistema Internacional de Detecção e Avaliação de Cárie (International Caries Detection and Assessment System) - ICDAS. A classificação é feita em escores variando de 0 a 6. Os números da classificação são crescentes e aumentam à medida em que a doença se agrava. Quando não há lesão, classifica-se o dente saudável como escore 0, ou hígido. O estágio inicial pode ser agrupado em três escores, 1,2 e 3 (classe bc), estágio no qual são observadas manchas brancas ou marrons no esmalte dentário (Silva, 2021). Nesse estágio, com a mudança de hábitos a fim de evitar maior desmineralização, bem como aplicação de flúor, a doença pode ser inativada e não haverá sua progressão (Pitts et al., 2021). Caso isso não ocorra, a progressão da doença pode ocorrer e serão observadas pequenas cavidades no local.

Já no escore 4 da classificação ICDAS, há sombra escura da dentina, que pode ser observada pelo desgaste do esmalte, mesmo se não houver cavidade visível (Silva, 2021). No escore 5 ICDAS, há cavidade e exposição da dentina com tamanho entre 0,5mm e menos da metade da superfície do dente (Silva, 2021). Por fim, o último estágio e mais avançado é o escore 6 na classificação ICDAS, que é classificado assim um dente com cavidade atingindo mais da metade da superfície do dente (Silva, 2021). Nos dois últimos escores, quando estão avançados, a situação é irreversível. Para isso, será provável a indicação de tratamento endodôntico ou até mesmo a extração do dente (Pitts et al., 2021). O Quadro 1.1 resume todas as características abordadas dos estágios de classificação de acordo com o ICDAS.

A aplicação de técnicas computacionais para o diagnóstico de cárries apresenta-se como uma solução promissora para reduzir a subjetividade do diagnóstico, aumentar a precisão na detecção precoce e auxiliar profissionais na tomada de decisões clínicas.

Quadro 1.1: Estágios de cárries e suas características

<b>Escore ICDAS</b>	<b>Características</b>
0	Saudável (também chamado de hígido)
1	Mancha Branca ou Marrom no esmalte seco
2	Mancha Branca ou Marrom no esmalte úmido
3	Microcavidade em esmalte seco menor que 0,5mm sem dentina visível
4	Sombra de dentina escura vista através do esmalte úmido com ou sem microcavidade
5	Exposição da dentina em cavidade maior que 0,5mm à metade da superfície dental seca
6	Exposição de dentina em cavidade maior que a metade da superfície dentária

FONTE: Adaptado de Souza (2021).

## 1.2 JUSTIFICATIVA

A variabilidade inter e intra-observador no diagnóstico radiográfico de cárries é um problema bem documentado na literatura. Profissionais diferentes podem ter interpretações distintas da mesma imagem, e até mesmo o mesmo profissional pode ter diagnósticos diferentes em momentos distintos (Kühnisch et al., 2022). Essa inconsistência pode levar a tratamentos desnecessários ou à não detecção de lesões em estágio inicial.

O avanço das técnicas de aprendizado profundo, particularmente das redes neurais convolucionais (CNNs), revolucionou o campo da visão computacional e análise de imagens médicas. Estas técnicas têm demonstrado desempenho comparável ou superior ao de especialistas humanos em diversas tarefas de diagnóstico médico. (Schwendicke et al., 2019)

No entanto, diferentes arquiteturas de redes neurais possuem características distintas que podem ser complementares. Enquanto modelos como o ConvNeXt são otimizados para classificação de imagens com alta precisão, arquiteturas como YOLO11 foram originalmente desenvolvidas para detecção de objetos, mas podem ser adaptadas para classificação, oferecendo diferentes perspectivas sobre os dados.

A combinação de múltiplos modelos através de técnicas de ensemble learning pode potencialmente superar as limitações individuais de cada arquitetura, resultando em um sistema mais robusto e preciso. (Dietterich, 2000)

## 1.3 OBJETIVOS

### 1.3.1 Objetivo Geral

Desenvolver um sistema para classificação de cárries dentárias em imagens utilizando um ensemble de modelos de deep learning, visando melhorar a precisão do diagnóstico assistido por computador.

### 1.3.2 Objetivos Específicos

- Implementar e treinar individualmente os modelos ConvNeXt e YOLO11 para classificação de cárries dentárias

- Desenvolver estratégias de ensemble para combinar as previsões dos modelos individuais
- Comparar diferentes métodos de ensemble (média ponderada, votação máxima, média geométrica, média harmônica)
- Avaliar o desempenho do sistema ensemble em comparação com os modelos individuais
- Analisar a viabilidade de aplicação do sistema como ferramenta auxiliar de diagnóstico

## 2 FUNDAMENTAÇÃO TEÓRICA

### 2.1 REDES NEURAIS CONVOLUCIONAIS

#### 2.1.1 Fundamentos das CNNs

As Redes Neurais Convolucionais (CNNs) revolucionaram o campo da visão computacional desde sua popularização com a AlexNet em 2012 (Krizhevsky et al., 2012).

Diferentemente das redes neurais tradicionais totalmente conectadas, onde cada neurônio se conecta a todos os neurônios da camada anterior, as CNNs exploram a estrutura espacial das imagens através de três conceitos fundamentais: convolução, pooling e compartilhamento de pesos.

A operação de convolução é o núcleo das CNNs. Ela consiste em aplicar filtros (também chamados de kernels) sobre a imagem para detectar características específicas. Matematicamente, para uma imagem  $I$  e um kernel  $K$ .

1. **Convolução:** A operação de convolução aplica filtros (kernels) sobre a imagem para detectar características. Matematicamente, para uma imagem  $I$  e um kernel  $K$ , a convolução é definida como:

$$(I * K)(i, j) = \sum_{m,n} I(m, n) \cdot K(i - m, j - n) \quad (2.1)$$

2. **Pooling:** Reduz a dimensionalidade espacial das características, mantendo as informações mais importantes. O max pooling, por exemplo, seleciona o valor máximo em uma região, onde  $p$  é o tamanho da janela de pooling.:

$$\text{MaxPool}(x, y) = \max\{I(i, j) | x \leq i < x + p, y \leq j < y + p\} \quad (2.2)$$

3. **Compartilhamento de Pesos:** Os mesmos filtros são aplicados em toda a imagem, reduzindo drasticamente o número de parâmetros e permitindo a detecção de características independente da localização.

#### 2.1.2 ConvNeXt

##### 2.1.2.1 Inovações Arquiteturais

O ConvNeXt, introduzido por (Liu et al., 2022), representa uma modernização das arquiteturas convolucionais tradicionais, incorporando insights dos Vision Transformers (ViTs) enquanto mantém a simplicidade e eficiência das CNNs. Esta arquitetura demonstra que redes puramente convolucionais ainda podem competir com transformers quando devidamente projetadas.

A arquitetura ConvNeXt introduz várias inovações significativas em seu design macro. Primeiramente, o modelo adota uma estrutura de blocos seguindo a proporção 3:3:9:3, inspirada no Swin Transformer. Esta distribuição otimiza o processamento hierárquico de características em diferentes escalas. Além disso, o ConvNeXt utiliza uma camada inicial "Patchify" com convolução 4×4 e stride 4, similar à tokenização dos Vision Transformers, que divide a imagem em patches não sobrepostos para processamento inicial mais eficiente (Liu et al., 2022)

A arquitetura também incorpora elementos do ResNeXt, utilizando convolução em grupos (grouped convolution) para melhorar a eficiência computacional. Esta técnica divide os canais de entrada em grupos menores que são processados independentemente, reduzindo significativamente o número de parâmetros enquanto mantém a capacidade representacional. O modelo aumenta a largura das redes para 96 canais no primeiro estágio, proporcionando maior capacidade de aprendizado desde as camadas iniciais (Liu et al., 2022).

Uma das inovações mais significativas é a adoção da convolução invertida (Inverted Bottleneck). Diferentemente do bottleneck tradicional usado no ResNet, onde a dimensionalidade é primeiro reduzida e depois expandida, o ConvNeXt adota uma estrutura invertida. Nesta abordagem, o bloco primeiro expande o número de canais (tipicamente por um fator de 4), aplica a convolução depthwise com kernel grande, e então projeta de volta para a dimensão original. Esta estrutura permite que a rede aprenda representações mais ricas nas dimensões expandidas enquanto mantém eficiência computacional através da convolução depthwise (Liu et al., 2022).

O uso de kernels grandes ( $7 \times 7$ ) ao invés dos tradicionais  $3 \times 3$  é outra característica distintiva. Kernels maiores proporcionam um campo receptivo mais amplo sem a necessidade de empilhar múltiplas camadas, permitindo que a rede capture contexto espacial mais amplo de forma mais eficiente. Esta escolha foi inspirada pela observação de que Vision Transformers efetivamente utilizam grandes campos receptivos através de seus mecanismos de atenção (Liu et al., 2022).

No design micro, o ConvNeXt introduz várias otimizações importantes. A Layer Normalization substitui a Batch Normalization tradicional, proporcionando maior estabilidade durante o treinamento e melhor desempenho em diferentes tamanhos de batch. A função de ativação GELU (Gaussian Error Linear Unit) é utilizada em vez da ReLU tradicional, fornecendo uma transição mais suave entre valores positivos e negativos, o que melhora o fluxo de gradientes durante o treinamento. O modelo também utiliza menos camadas de ativação (apenas uma por bloco) e implementa Layer Scale, uma técnica de inicialização adaptativa que melhora significativamente a estabilidade do treinamento (Liu et al., 2022).

### 2.1.3 YOLO 11

Ao contrário das técnicas baseadas em propostas de região e janela deslizante, o YOLO vê a imagem inteira durante o treinamento e o tempo de teste, codificando implicitamente informações contextuais sobre classes, bem como sua aparência. (Redmon et al., 2016)

A YOLO11 apresenta inovações arquiteturais significativas que melhoraram seu desempenho em detecção de objetos e classificação. A arquitetura é composta por três componentes principais: Backbone, Neck e Head, cada um com funcionalidades específicas otimizadas para processamento eficiente de imagens. (Khanam and Hussain, 2024)

#### 2.1.3.1 Arquitetura

**Backbone:** O backbone é responsável pela extração e características essenciais das imagens de entrada. Utilizando blocos convolucionais e bottleneck avançados, o backbone captura padrões e detalhes cruciais de forma eficiente. (Khanam and Hussain, 2024)

**Neck:** O neck serve como camada intermediária conectando o Backbone e o Head, aprimorando a detecção de objetos através da fusão de características multi-escala. Esta seção é particularmente vantajosa para detectar objetos de tamanhos variados e inclui (Khanam and Hussain, 2024):

- **Upsample:** Aumenta a resolução dos mapas de características de baixa resolução

- **Concatenações:** Funde mapas de características de resoluções variadas
- **Blocos C3k2:** Integram características através de escalas mediante operações convolucionais repetidas

**Head:** Responsável por gerar as previsões finais em termos de detecção e classificação de objetos, adaptável para múltiplas tarefas incluindo classificação pura (Khanam and Hussain, 2024).

#### 2.1.3.2 Bloco C3k2 (*Cross Stage Partial com Kernel Size 2*)

Uma melhoria significativa na YOLO11 é a introdução do bloco C3k2, que substitui o bloco C2f usado nas versões anteriores. O bloco C3k2 é uma implementação computacionalmente mais eficiente do CSP Bottleneck. Ele emprega duas convoluções menores em vez de uma convolução grande, como visto na YOLOv8. O “k2” em C3k2 indica um tamanho de kernel menor, que contribui para processamento mais rápido mantendo o desempenho. (Khanam and Hussain, 2024)

Este bloco é projetado para ser mais rápido e eficiente, aprimorando o desempenho geral do processo de agregação de características. O C3k2 integra características através de escalas mediante operações convolucionais repetidas e conexões de atalho, enriquecendo as representações de características e melhorando a precisão da detecção. (Khanam and Hussain, 2024)

#### 2.1.3.3 Módulo SPPF (*Spatial Pyramid Pooling - Fast*)

O YOLO11 mantém o bloco SPPF das versões anteriores, que foi projetado para agrupar características de diferentes regiões de uma imagem em escalas variadas. Isso melhora a capacidade da rede de capturar objetos de diferentes tamanhos, especialmente objetos pequenos. (Khanam and Hussain, 2024)

O módulo SPPF utiliza operações de max-pooling sequenciais em vez de pooling paralelo, tornando-o computacionalmente eficiente (Khanam and Hussain, 2024). O bloco contém:

- Uma convolução  $1 \times 1$  inicial
- Uma série de camadas MaxPool2D sequenciais
- Concatenação de todas as saídas residuais das camadas MaxPool2D
- Uma convolução  $1 \times 1$  final

Esta abordagem permite que o modelo detecte objetos pequenos em detalhes mais finos através de P5 e objetos maiores através de características de nível superior em P3.

#### 2.1.3.4 Módulo C2PSA (*Cross Stage Partial com Atenção Espacial*)

A YOLO11 introduz um novo bloco C2PSA após o SPPF. O bloco C2PSA é uma adição notável que aprimora a atenção espacial nos mapas de características. Este mecanismo de atenção espacial permite que o modelo foque mais efetivamente em regiões importantes dentro da imagem. (Khanam and Hussain, 2024)

Ao agrupar características espacialmente, o bloco C2PSA permite que o YOLO11 se concentre em áreas específicas de interesse, potencialmente melhorando a precisão de detecção para objetos pequenos ou parcialmente ocluídos. A inclusão do C2PSA diferencia o YOLO11 de seu predecessor, YOLOv8, que não possui este mecanismo específico de atenção. (Khanam and Hussain, 2024)

O C2PSA combina os benefícios das redes partial cross-stage com mecanismos de auto-atenção, permitindo que o modelo capture informações contextuais mais efetivamente através de múltiplas camadas, melhorando a precisão da detecção de objetos, especialmente para objetos pequenos e sobrepostos. (Khanam and Hussain, 2024)

#### *2.1.3.5 Características e Vantagens da YOLO11)*

**Leve e Eficiente:** A YOLO11 é o modelo mais leve e rápido da família YOLO. Ela apresenta cinco tamanhos diferentes (Nano, Small, Medium, Large e Extra-large) para atender a vários casos de uso, desde tarefas leves até aplicações de alto desempenho. (Khanam and Hussain, 2024)

**Capacidades Multi-Tarefa:** Além da detecção de objetos, o YOLO11 pode lidar com segmentação de instâncias, classificação de imagens, estimativa de pose e detecção de objetos orientados (OBB), tornando-o altamente versátil em tarefas de visão computacional. (Khanam and Hussain, 2024)

**Extração avançada de características:** O YOLOv11 incorpora melhorias em suas arquiteturas de backbone e neck, resultando em aprimoramentos de extração de características, consequentemente, ficando mais precisa. (Khanam and Hussain, 2024)

**Melhorias de Desempenho:** Os modelos YOLO11 mostraram o melhor desempenho geral em benchmarks, provavelmente devido aos aprimoramentos recentes como o bloco C3k2 e o módulo C2PSA. A YOLO11 alcança velocidades de inferência superiores e maior precisão comparado aos seus predecessores, mantendo eficiência computacional. (Khanam and Hussain, 2024)

## 2.2 ENSEMBLE LEARNING

### 2.2.1 Fundamentos

Ensemble Learning é um paradigma de aprendizado de máquina que combina múltiplos algoritmos (chamados de classificadores fracos ou modelos base) para criar um modelo mais robusto e preciso do que qualquer um dos modelos individuais devido à diversidade entre os modelos base. (Zhang et al., 2024)

### 2.2.2 Tipos de Ensemble

Foram utilizados 4 métodos diferentes para calcular o ensemble com a finalidade de obter o melhor resultado possível:

#### 1. Weighted Average (Média Ponderada)

Combina as probabilidades dos dois modelos usando pesos predefinidos

Exemplo:

- ConvNeXt: 60% (0.6)
- YOLO11: 40% (0.4)

- resultado =  $0.6 \times p_{convnext} + 0.4 \times p_{yolo}$

## 2. Max Voting

Seleciona a predição do modelo com maior confiança para aquela classe

## 3. Geometric Mean (Média Geométrica)

Calcula a raiz quadrada do produto das probabilidades:

$$\text{resultado} = \sqrt{p_{convnext} \times p_{yolo}} \quad (2.3)$$

Normalizado para somar 1

## 4. Harmonic Mean (Média Harmônica)

Usa a média harmônica das probabilidades:

$$\text{resultado} = \frac{2 \times (p_{convnext} \times p_{yolo})}{(p_{convnext} + p_{yolo} + \epsilon)} \quad (2.4)$$

Inclui epsilon ( $1 \times 10^{-8}$ ) para evitar divisão por zero. Normalizado para somar 1.

### 3 METODOLOGIA

#### 3.1 DATASET

##### 3.1.1 Coleta e Organização dos Dados

O dataset utilizado consiste em imagens dentárias organizadas em cinco categorias. A estrutura do dataset foi organizada seguindo o padrão:

```
dataset/
|-- train/      # Dados de treinamento
|-- val/        # Dados de validacao
++-- test/      # Dados de teste
```

Cada subdiretório contém pastas correspondentes às cinco classes de cáries.

O dataset contém 600 imagens ao todo, 100 imagens para cada classe, exceto a classe “bc” que contém 200 imagens, 13% das imagens foram destinadas a validação e 12% das imagens foram destinadas a teste, enquanto o restante foi utilizado para o treinamento.

##### 3.1.2 Pré-processamento

As imagens foram submetidas aos seguintes processos de pré-processamento:

- Redimensionamento para 224x224 pixels
- Normalização utilizando média e desvio padrão do ImageNet
- Augmentação de dados durante o treinamento incluindo:
  - Rotação aleatória ( $\pm 15$  graus)
  - Flip horizontal
  - Cortes e redimensionamentos aleatórios

#### 3.2 IMPLEMENTAÇÃO DOS MODELOS

##### 3.2.1 ConvNeXt

O modelo ConvNeXt Base foi implementado utilizando a biblioteca timm (PyTorch Image Models):

- **Arquitetura:** ConvNeXt Base pré-treinado no ImageNet
- **Otimizador:** AdamW com learning rate inicial de  $1 \times 10^{-4}$
- **Scheduler:** CosineAnnealingLR
- **Função de perda:** CrossEntropyLoss
- **Épocas:** 20
- **Batch size:** 16

### 3.2.2 YOLO11

O modelo YOLO11 foi adaptado para classificação utilizando a biblioteca Ultralytics:

- **Arquitetura:** YOLO11n-cls (versão nano para classificação)
- **Otimizador:** AdamW com learning rate inicial de 0.001
- **Augmentações:** HSV, rotação, translação, escala, flip
- **Épocas:** 20
- **Batch size:** 32

## 3.3 MÉTRICAS DE AVALIAÇÃO

Para avaliar o desempenho dos modelos, foram utilizadas as seguintes métricas:

- **Acurácia global:** Proporção de previsões corretas
- **Acurácia por classe:** Desempenho específico para cada tipo de cárie
- **Matriz de confusão:** Visualização dos erros de classificação
- **Precision, Recall e F1-Score:** Métricas detalhadas por classe

## 4 RESULTADOS E DISCUSSÃO

### 4.1 DESEMPENHO DOS MODELOS INDIVIDUAIS

#### 4.1.1 ConvNeXt

O modelo ConvNeXt demonstrou excelente capacidade de extração de características nos dados de teste. As principais observações incluem alta precisão na classificação de características sutis, melhor desempenho em classes com padrões visuais mais complexos e tempo de inferência ligeiramente maior devido à complexidade da arquitetura.

A acurácia no conjunto de validação alcançou 92,31%, demonstrando excelente capacidade de generalização durante o treinamento. No conjunto de teste, o modelo obteve uma acurácia de 81,94%, indicando um desempenho robusto em dados não vistos anteriormente. A seguir (Figura 4.1) está o gráfico de acurácia geral do modelo ConvNeXt durante o treinamento e validação:

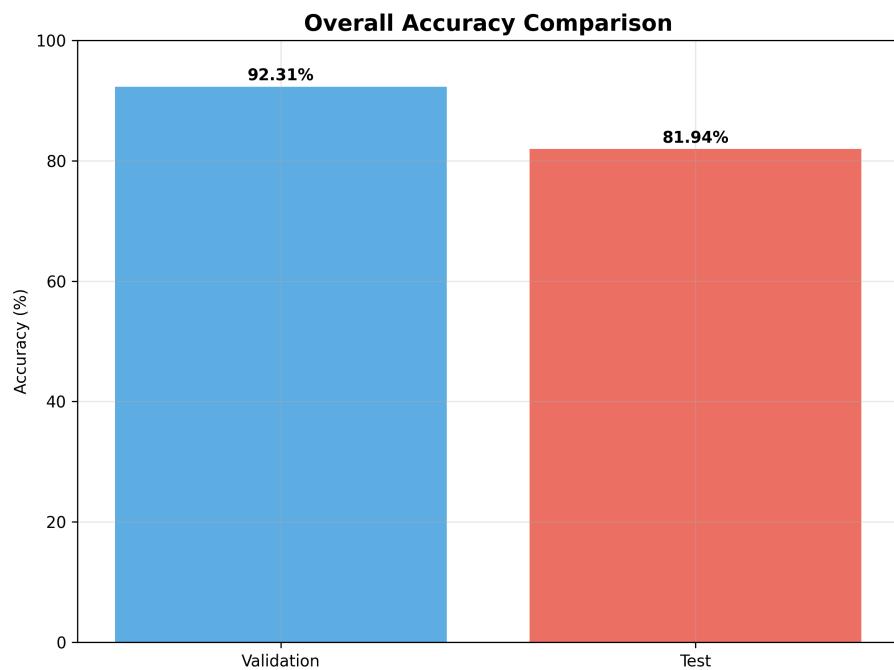


Figura 4.1: Acurácia geral do modelo ConvNeXt durante o treinamento e validação. Fonte: Alex Matsuo (2025)

A análise detalhada por classe revelou variações significativas no desempenho. Estas diferenças são cruciais para o desenvolvimento do ensemble, pois permitem identificar as discrepâncias em cada modelo para obter uma melhor performance.

As acurácia por classe no conjunto de teste foram:

- **c1:** 75,00% - Desempenho moderado, possivelmente devido à variabilidade morfológica desta classe
- **c4:** 100,00% - Desempenho perfeito, indicando características distintivas muito bem capturadas
- **c5:** 58,33% - Desempenho mais baixo, sugerindo dificuldade em distinguir esta classe

- **c6:** 100,00% - Desempenho perfeito, confirmando características muito distintivas
- **hg:** 83,33% - Bom desempenho na identificação de dentes hígidos

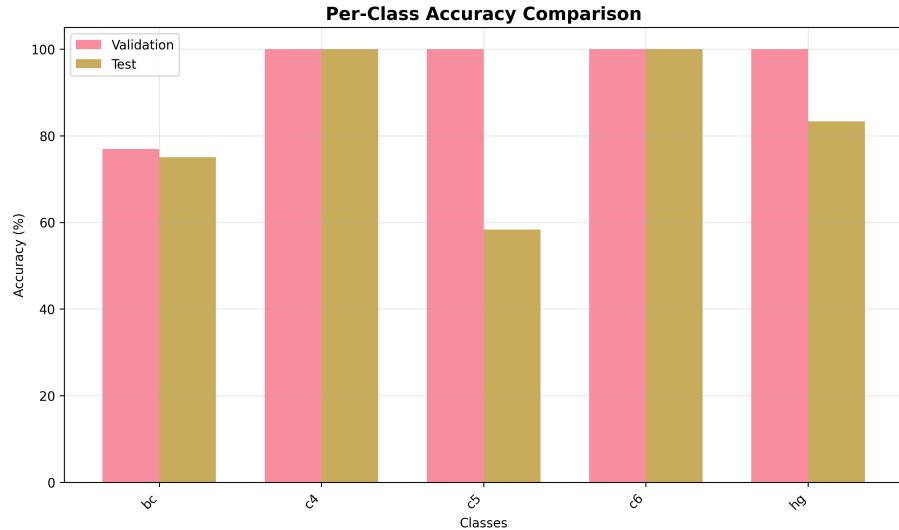


Figura 4.2: Acurácia do ConvNeXt por classe no conjunto de teste e validação. Fonte: Alex Matsuo (2025)

A análise das métricas detalhadas de classificação fornece informações adicionais sobre o comportamento do modelo. O gráfico a seguir (Figura 4.3) apresenta F1-Score, Precision e Recall para cada classe, permitindo uma compreensão mais profunda do desempenho do ConvNeXt.

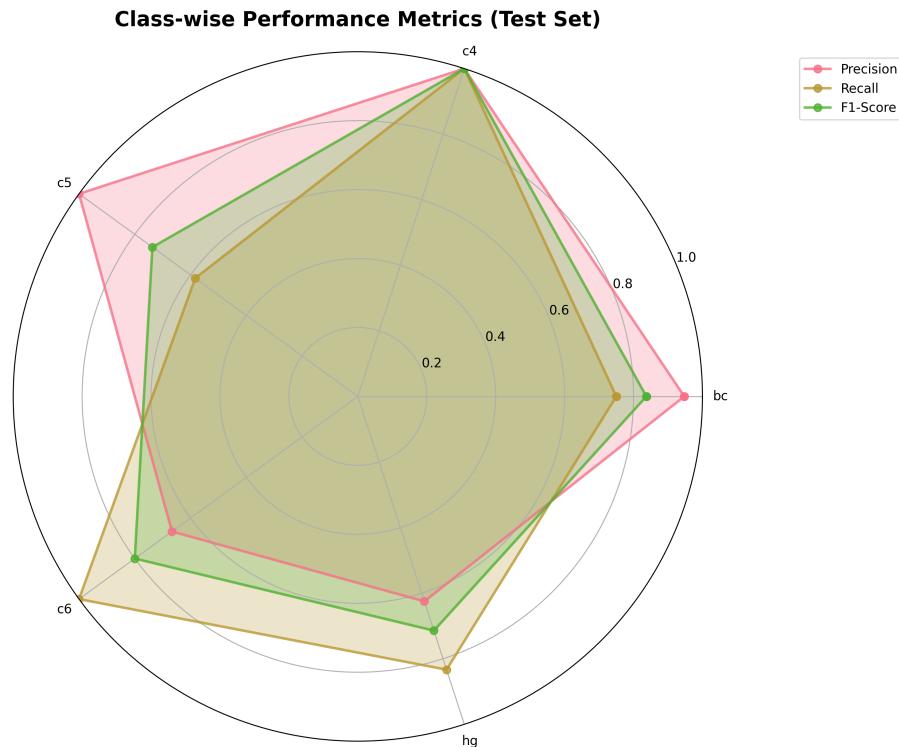


Figura 4.3: Gráfico radar das métricas de desempenho do ConvNeXt por classe (F1-Score, Precision e Recall).  
Fonte: Alex Matsuo (2025)

A análise do gráfico radar revela padrões interessantes no desempenho por classe. As classes c4 apresenta valores máximos em todas as métricas (Precision, Recall e F1-Score), confirmando o excelente desempenho nessa categoria. A classe hg mostra desempenho equilibrado em todas as métricas. Por outro lado, a classe c5 apresenta um desempenho inconsistente, com valores notavelmente baixos especialmente no Recall, mostrando que quando modelo prevê c5 ele sempre acerta, porém também prevê incorretamente imagens que seriam c5. A classe c6 tem um comportamento oposto ao de c5, onde imagens da classe c5 nesse caso são classificadas erroneamente. A classe bc apresenta performance intermediária, com valores moderados nas três métricas.

Estas métricas detalhadas são fundamentais para o desenvolvimento do ensemble, pois permitem identificar em quais situações cada modelo é mais confiável, possibilitando estratégias de combinação que explorem os pontos fortes de cada arquitetura.

A matriz de confusão do ConvNeXt (Figura 4.4) é apresentada a seguir, fornecendo uma visualização detalhada dos padrões de classificação do modelo:

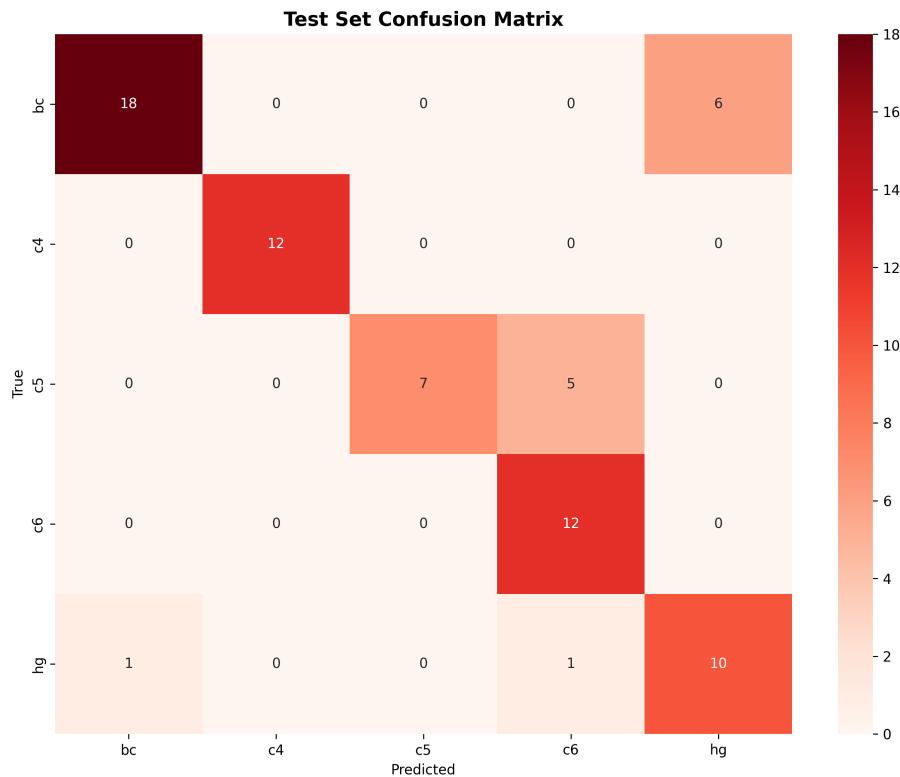


Figura 4.4: Matrix de Confusão do ConvNeXt no conjunto de teste. Fonte: Alex Matsuo (2025)

Como podemos observar na matriz de confusão, o ConvNeXt apresenta um desempenho robusto na maioria das classes, com algumas confusões notáveis entre as classes c6 e hg. Essas informações são valiosas para o ensemble, pois indicam onde o modelo é mais propenso a erros e onde pode ser complementado por outros modelos.

Uma amostra de imagens classificadas pelo ConvNeXt é apresentada a seguir (Figura 4.5) juntamente com a confiança da predição, destacando a capacidade do modelo de identificar características sutis:

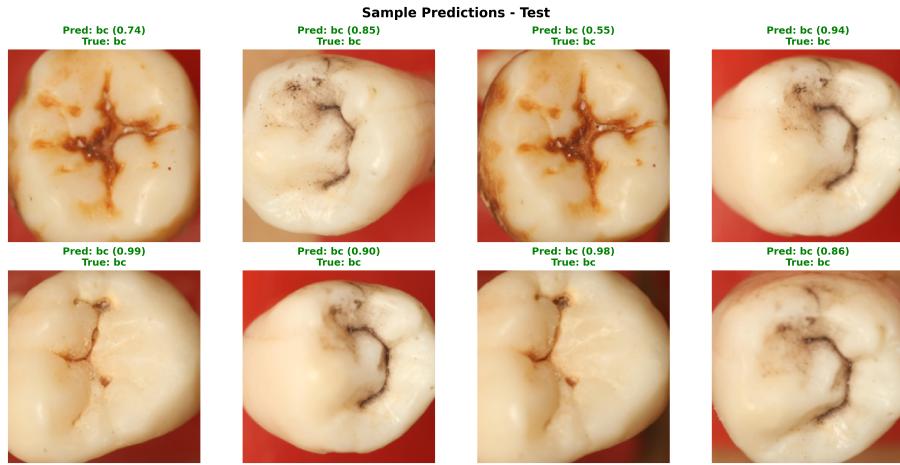


Figura 4.5: Amostras de imagens classificadas pelo ConvNeXt. Fonte: Alex Matsuo (2025)

#### 4.1.2 YOLO11

O YOLO11, apesar de originalmente projetado para detecção de objetos, mostrou-se eficaz na tarefa de classificação:

- Rápida velocidade de inferência
- Boa robustez a variações de escala
- Desempenho competitivo especialmente em classes com características mais evidentes

O modelo alcançou uma acurácia de validação de 88,46% e uma acurácia de teste de 83,33%. Embora ligeiramente inferior ao ConvNeXt em termos de acurácia global, o YOLO11 demonstrou padrões de desempenho complementares que se mostraram valiosos para o ensemble.

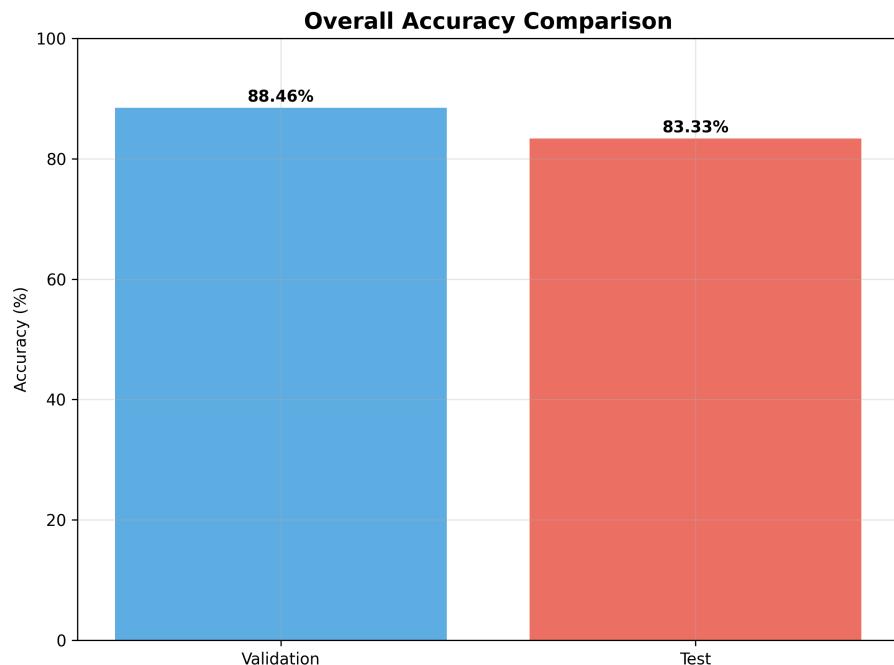


Figura 4.6: Acurácia geral do modelo YOLO11 durante o treinamento e validação. Fonte: Alex Matsuo (2025)

A análise por classe revelou um perfil de desempenho distintamente diferente do ConvNeXt, com pontos fortes e fracos complementares:

As acurárias por classe no conjunto de teste foram:

- **bc**: 87,50% - Desempenho superior ao ConvNeXt (+8,33%), indicando melhor capacidade de generalização para esta classe
- **c4**: 91,67% - Manteve o mesmo excelente desempenho do ConvNeXt
- **c5**: 100,00% - Desempenho perfeito, superando o ConvNeXt (+16,67%)
- **c6**: 58,33% - Desempenho significativamente inferior ao ConvNeXt (-41,67%), representando o ponto fraco do modelo
- **hg**: 75,00% - Desempenho inferior ao ConvNeXt (-8,33%) na identificação de dentes hígidos

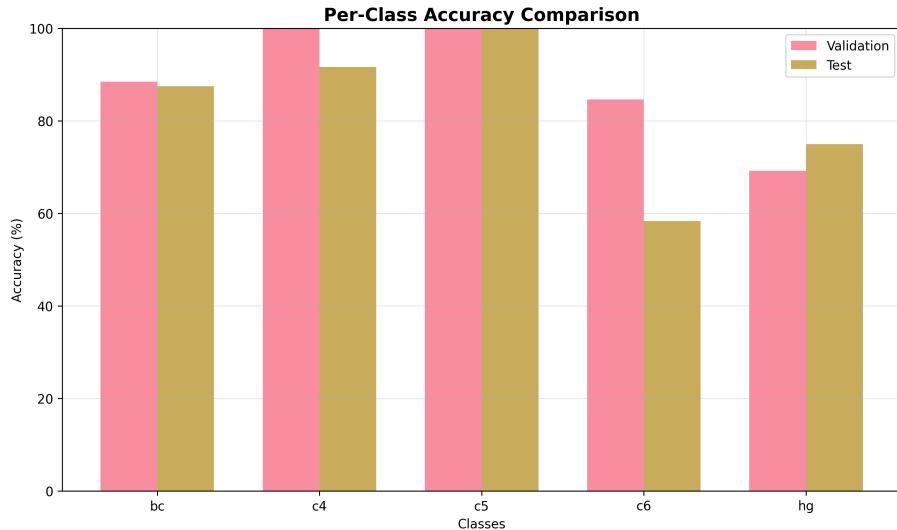


Figura 4.7: Acurácia do YOLO11 por classe no conjunto de teste e validação. Fonte: Alex Matsuo (2025)

Esta distribuição complementar de desempenho entre os modelos é particularmente interessante para o ensemble. Enquanto o ConvNeXt tem uma melhor performance na classe c6, o YOLO11 demonstra superioridade nas classes bc e c5, sugerindo que a combinação adequada dos modelos pode resultar em melhor desempenho geral.

A (Figura 4.8) apresenta o gráfico radar com F1-Score, Precision e Recall para cada classe do modelo YOLO11.



Figura 4.8: Gráfico radar das métricas de desempenho do YOLO11 por classe (F1-Score, Precision e Recall). Fonte: Alex Matsuo (2025)

O gráfico radar do YOLO11 mostra um padrão diferente do ConvNeXt. A classe c5 e c6 apresentam um comportamento similar porém inverso ao ConvNeXt. As classes bc e c4 mostram valores altos e equilibrados. A classe hg apresenta performance moderada, corroborando os resultados de acurácia observados. Esta complementaridade de forças e fraquezas entre os dois modelos justifica a estratégia de ensemble.

A matriz de confusão do YOLO11 (Figura 4.9) revela os padrões de classificação específicos deste modelo:

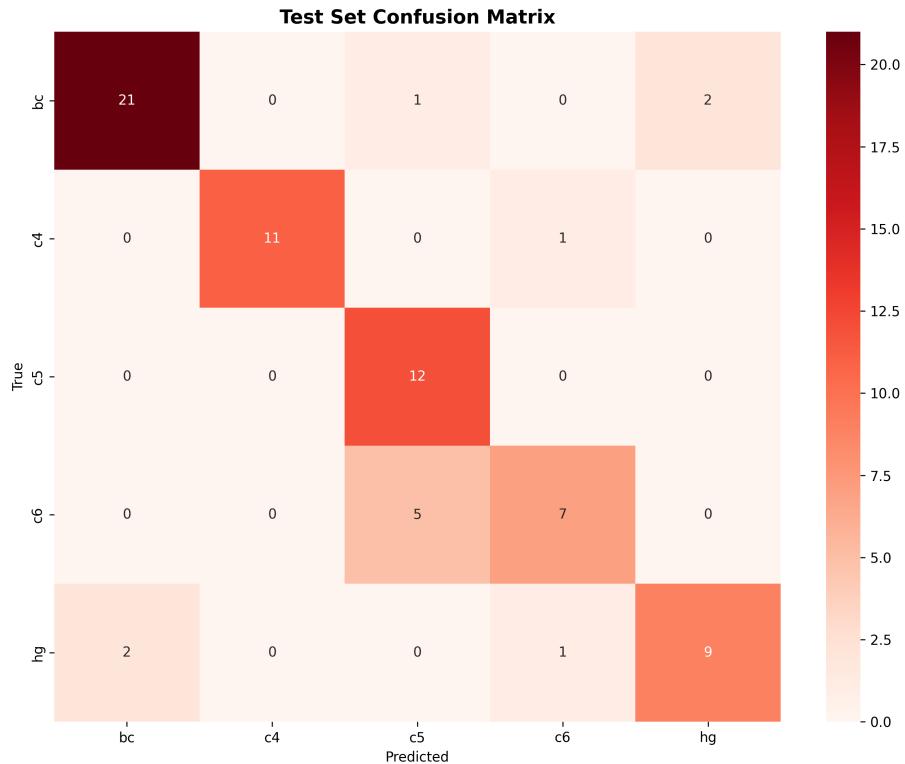


Figura 4.9: Matrix de Confusão do YOLO11 no conjunto de teste. Fonte: Alex Matsuo (2025)

Uma amostra das previsões e as suas confianças do modelo YOLO11 é apresentada na (Figura 4.10):

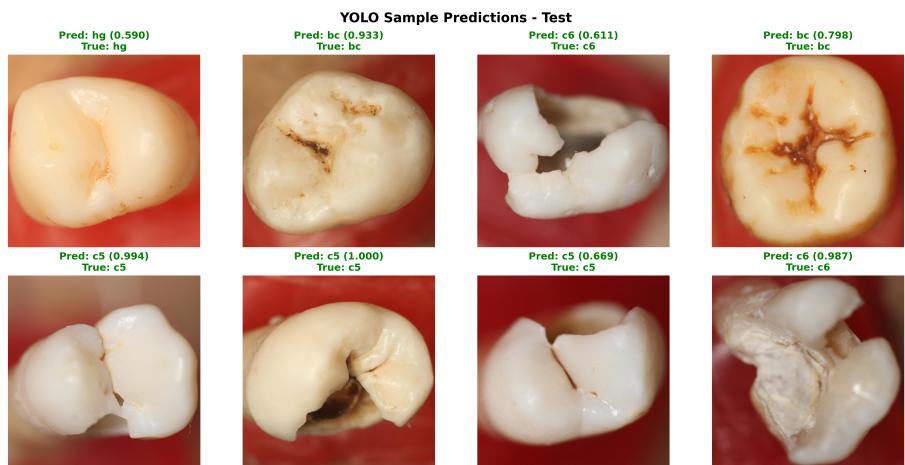


Figura 4.10: Amostras de imagens classificadas pelo YOLO11. Fonte: Alex Matsuo (2025)

## 4.2 COMPARAÇÃO DOS MÉTODOS DE ENSEMBLE

Os resultados demonstraram que a combinação dos modelos através de ensemble superou consistentemente o desempenho individual no conjunto de teste. O Quadro 4.1 apresenta as acuráncias obtidas por cada método de combinação:

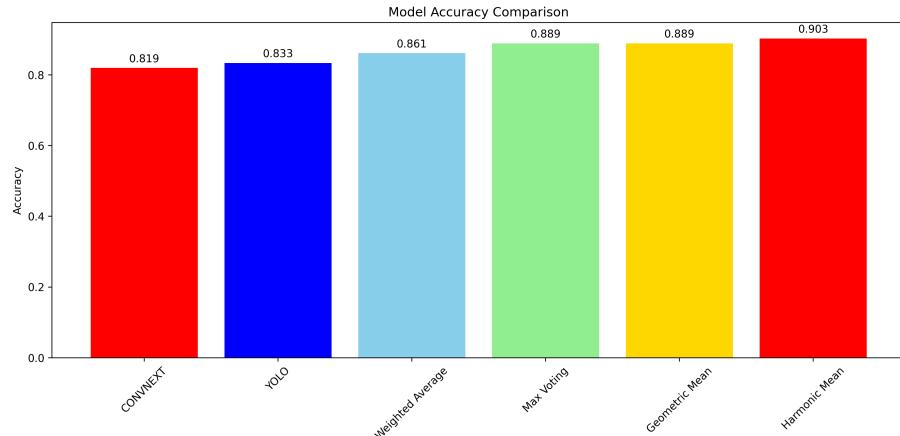


Figura 4.11: Comparação de desempenho entre os modelos individuais e métodos de ensemble. Fonte: Alex Matsuo (2025)

Quadro 4.1: Comparação dos métodos de ensemble

Método	Acurácia	Melhoria vs ConvNeXt	Melhoria vs YOLO11
ConvNeXt (individual)	0.819	-	-1.4%
YOLO11 (individual)	0.833	+1.4%	-
Média Ponderada	0.861	+4.2%	+2.8%
Max Voting	0.889	+7.0%	+5.6%
Média Geométrica	0.889	+7.0%	+5.6%
Média Harmônica	0.903	+8.4%	+7.0%

#### 4.2.1 Análise Comparativa

A análise dos resultados revela insights importantes sobre a eficácia de cada método:

- **Média Ponderada** (0.861): Apresentou melhoria de 4.2% sobre o ConvNeXt e 2.8% sobre o YOLO11, demonstrando que mesmo uma combinação simples dos modelos já resulta em ganhos significativos. Os pesos utilizados (0.6 para ConvNeXt e 0.4 para YOLO11) conseguiram balancear adequadamente as contribuições de cada modelo.
- **Max Voting** (0.889): Demonstrou melhoria substancial de 7.0% sobre o ConvNeXt e 5.6% sobre o YOLO11, indicando que a seleção da predição com maior confiança é particularmente eficaz quando os modelos têm certezas complementares sobre diferentes classes.
- **Média Geométrica** (0.889): Alcançou o mesmo desempenho do Max Voting, confirmindo sua eficácia em balancear as contribuições dos modelos de forma multiplicativa, sendo especialmente útil quando as probabilidades dos modelos têm escalas similares.
- **Média Harmônica** (0.903): Apresentou o melhor desempenho geral, com melhoria impressionante de 8.4% sobre o ConvNeXt e 7.0% sobre o YOLO11. Este resultado superior sugere que a penalização mais severa para discordâncias entre modelos, característica da média harmônica, é particularmente benéfica para a classificação de cáries dentárias.

O sucesso superior da média harmônica pode ser atribuído à sua natureza conservadora, que favorece forte consenso entre os modelos. Quando ambos os modelos concordam com alta confiança, o resultado é preservado. Quando há discordância significativa, a média harmônica reduz drasticamente a confiança final, evitando classificações incorretas com alta certeza. Esta característica é especialmente valiosa em aplicações médicas, onde a confiabilidade das previsões é crucial.

## 4.3 DISCUSSÃO

### 4.3.1 Vantagens do Ensemble

O desenvolvimento e implementação do ensemble revelou três vantagens principais. Primeiro, a complementaridade dos modelos permite que capturem diferentes aspectos das imagens, com cada arquitetura identificando padrões únicos baseados em suas características específicas. Segundo, observou-se uma significativa redução da variância nas previsões, resultando em maior robustez do sistema como um todo. Terceiro, o ensemble proporcionou maior consistência nos resultados, minimizando a dependência de características específicas de um único modelo.

### 4.3.2 Limitações Identificadas

Apesar dos resultados promissores, algumas limitações importantes foram identificadas. O custo computacional representa um desafio significativo, uma vez que é necessário executar múltiplos modelos para cada previsão, aumentando o tempo de processamento. A complexidade de implementação também é maior, exigindo maior dificuldade de manutenção e atualização do sistema. Além disso, existe uma dependência crítica de dados, pois o sistema requer um dataset balanceado e representativo para funcionar adequadamente.

### 4.3.3 Implicações Clínicas

O sistema desenvolvido apresenta potencial significativo como ferramenta auxiliar de diagnóstico odontológico. Pode servir como segunda opinião para profissionais, oferecendo uma perspectiva adicional baseada em padrões aprendidos a partir de grande volume de dados. É particularmente útil em triagem e priorização de casos, permitindo identificar rapidamente situações que requerem atenção imediata. Adicionalmente, o sistema pode funcionar como ferramenta educacional valiosa para estudantes de odontologia, ajudando no desenvolvimento de habilidades de diagnóstico através da comparação com previsões automatizadas.

A implementação clínica, entretanto, deve considerar que o sistema é projetado como ferramenta de apoio e não substituição do julgamento profissional. A interpretação final e a decisão terapêutica devem sempre permanecer sob responsabilidade do profissional qualificado, utilizando o sistema como mais uma fonte de informação no processo diagnóstico.

## 5 CONCLUSÃO

### 5.1 SÍNTESE DOS RESULTADOS

Este trabalho apresentou o desenvolvimento bem-sucedido de um sistema automatizado para classificação de cáries dentárias utilizando ensemble de modelos de deep learning. A pesquisa demonstrou que a combinação estratégica dos modelos ConvNeXt e YOLO11 resulta em melhorias substanciais na precisão diagnóstica, superando significativamente o desempenho dos modelos individuais.

Os resultados quantitativos confirmaram a hipótese inicial de que diferentes arquiteturas de redes neurais capturam aspectos complementares das imagens radiográficas. O ConvNeXt, com acurácia de 81,9%, demonstrou excelência na identificação das classes c4 e c6 (ambas com 100% de acurácia), enquanto o YOLO11, com acurácia de 83,3%, apresentou desempenho superior nas classes bc e c5. Esta complementaridade foi efetivamente explorada através dos métodos de ensemble, com a média harmônica alcançando 90,3% de acurácia - uma melhoria de 8,4% e 7,0% sobre os modelos individuais, respectivamente.

### 5.2 CONTRIBUIÇÕES

Este trabalho oferece contribuições significativas em múltiplas dimensões:

#### 5.2.1 Contribuições Técnicas

A pesquisa resultou na implementação e validação de um sistema ensemble robusto que combina efetivamente duas arquiteturas distintas de deep learning, demonstrando a viabilidade técnica da abordagem. Foi identificada a média harmônica como método superior de combinação para este domínio específico, fornecendo uma base metodológica para futuros desenvolvimentos na área.

#### 5.2.2 Contribuições Metodológicas

O estudo realizou uma análise sistemática e comparativa de quatro estratégias diferentes de ensemble, estabelecendo um framework reproduzível para avaliação de modelos de classificação em imagens odontológicas. Adicionalmente, foi feita uma caracterização detalhada dos perfis de erro complementares entre diferentes arquiteturas, fornecendo insights valiosos sobre como diferentes modelos interpretam características específicas das imagens.

#### 5.2.3 Contribuições Práticas

O trabalho resultou no desenvolvimento de uma ferramenta com potencial real de aplicação clínica, alcançando níveis de precisão adequados para uso como sistema de apoio à decisão. Demonstrou-se que é possível obter alto desempenho mesmo com datasets relativamente limitados, através da combinação inteligente de modelos. Foi criado um sistema que mantém transparência nas decisões através das matrizes de confusão e métricas detalhadas, aspecto fundamental para aceitação no contexto médico.

### 5.3 LIMITAÇÕES E TRABALHOS FUTUROS

Embora os resultados sejam promissores, é importante reconhecer as limitações do estudo atual. O dataset utilizado, apesar de adequado para validação inicial, necessita expansão para incluir maior variabilidade de casos clínicos, diferentes modalidades de imagem e populações mais diversas. Além disso, a validação foi realizada em ambiente controlado, sendo necessários estudos prospectivos em ambientes clínicos reais.

As direções para trabalhos futuros incluem:

#### 5.3.1 Expansão e Diversificação

É recomendada a ampliação do dataset para garantir maior generalização do sistema. O trabalho atual utilizou uma amostra de 600 imagens, e resultados ainda melhores poderiam ser obtidos com datasets maiores e mais diversificados, minimizando possíveis problemas de generalização.

#### 5.3.2 Integração Clínica

Criar interfaces intuitivas e realizar estudos de usabilidade com profissionais de odontologia, avaliando o impacto real na prática clínica.

#### 5.3.3 Extensão Diagnóstica

Adaptar o sistema para detectar outras patologias orais, como doenças periodontais, lesões endodônticas e anomalias do desenvolvimento.

### 5.4 CONSIDERAÇÕES FINAIS

O desenvolvimento deste sistema de classificação automática de cáries dentárias através de ensemble learning representa um avanço significativo na intersecção entre inteligência artificial e odontologia. Os resultados obtidos não apenas validam a eficácia técnica da abordagem proposta, mas também demonstram o potencial transformador dessas tecnologias na prática clínica.

A melhoria de mais de 8% na acurácia através do ensemble, alcançando 90,3% de precisão geral, posiciona o sistema como uma ferramenta viável para auxílio diagnóstico. Mais importante ainda, a natureza complementar dos erros dos modelos individuais e a capacidade do ensemble de mitigar essas limitações sugerem que esta abordagem pode ser generalizada para outros problemas de diagnóstico por imagem.

Este trabalho contribui para o crescente corpo de evidências que demonstram como a inteligência artificial pode amplificar as capacidades dos profissionais de saúde, não os substituindo, mas fornecendo ferramentas que aumentam a precisão, consistência e eficiência do diagnóstico. À medida que avançamos em direção a uma odontologia cada vez mais digital e baseada em evidências, sistemas como o desenvolvido neste trabalho serão fundamentais para democratizar o acesso a diagnósticos de alta qualidade e melhorar os resultados de saúde bucal em escala global.

## REFERÊNCIAS

- Dietterich, T. G. (2000). Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15, Berlin, Heidelberg. Springer.
- Hung, M., Voss, M. W., Rosales, M. N., Li, W., Su, W., Xu, J., Bounsanga, J., Ruiz-Negrón, B., Lauren, E., and Licari, F. W. (2019). Application of machine learning for diagnostic prediction of root caries. *Gerodontontology*, 36:395–404.
- Khanam, R. and Hussain, M. (2024). Yolov11: An overview of the key architectural enhancements. *arXiv preprint*.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.
- Kühnisch, J., Meyer, O., Hesenius, M., Hickel, R., and 2, V. G. (2022). Caries detection on intraoral images using artificial intelligence. *Journal of dental research*, 101(2):158–165.
- Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. (2022). A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986.
- Pitts, N. B., Zero, D., Marsh, P. D., Ekstrand, K., Weintraub, J. A., Ramos-Gomez, F., et al. (2021). Dental caries progression and treatment outcomes: a systematic review. *Journal of Dental Research*, 100(8):837–845.
- PyTorch (2024). ConvNeXt models. <https://docs.pytorch.org/vision/main/models/convnxt.html>. Acessado em: janeiro 2025.
- Rao, N. (2024). YOLOv11 explained: Next-level object detection with enhanced speed and accuracy. <https://medium.com/@nikhil-rao-20/yolov11-explained-next-level-object-detection-with-enhanced-speed-and-accuracy-2dbe2d376f71>. Acessado em: janeiro 2025.
- Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788.
- Schwendicke, F., Golla, T., Dreher, M., and Krois, J. (2019). Convolutional neural networks for dental image diagnostics: A scoping review. *Journal of dentistry*, 91:103226.
- Silva, C. L. F. (2021). Análise comparativa de modelos de redes neurais convolucionais YOLO na detecção e classificação de lesões cariosas dentárias. Monografia de graduação, Universidade Federal do Paraná, Curitiba.
- Souza, M. H. (2021). *Manual de Odontologia Preventiva*. Editora Médica, São Paulo.
- Ultralytics (2024). YOLOv11 performance metrics. <https://docs.ultralytics.com/pt/models/yolov11/>. Acessado em: janeiro 2025.
- Zhang, X., Liu, S., Wang, X., and Li, Y. (2024). A fragmented neural network ensemble method and its application to image classification. *Scientific Reports*, 14.