



Universidade Federal do ABC  
Centro de Matemática, Computação e Cognição  
Bacharelado em Ciência da Computação

# **Detecção e Classificação Automática de Osteoartrite de Joelho em Radiografias Utilizando Visão Computacional**

**Guilherme de Sousa Santos**

**Santo André - SP, 15 de agosto de 2025**



Guilherme de Sousa Santos

# **Detecção e Classificação Automática de Osteoartrite de Joelho em Radiografias Utilizando Visão Computacional**

**Projeto de Graduação** apresentado como parte dos requisitos necessários para a obtenção do título de Bacharel em Ciência da Computação.

Universidade Federal do ABC – UFABC  
Centro de Matemática, Computação e Cognição  
Bacharelado em Ciência da Computação

Orientador: Hugo Puertas de Araújo

Santo André - SP  
15 de agosto de 2025

*Dedico este trabalho à minha família e amigos,  
pelo apoio incondicional e amor  
em todas as etapas da minha vida acadêmica.*

# Agradecimentos

Agradeço primeiramente a Deus, pela saúde, força e determinação para que eu pudesse concluir esta etapa muito importante. Aos meus pais, por todo amor e paciência nos últimos anos de graduação, além de todo apoio e exemplo de vida. Ao meu orientador, Prof. Dr. Hugo Puertas, por todos os conselhos, direcionamentos e ajuda durante todo o desenvolvimento deste trabalho. Aos colegas e amigos que levo para a vida, pela parceria, motivação e momentos de descontração que tornaram esta jornada mais leve. À Universidade Federal do ABC (UFABC), pelo suporte, excelência e infraestrutura disponibilizados. E a todos que, de forma direta ou indireta, contribuíram para a realização deste trabalho, meu muito obrigado.



# Resumo

A osteoartrite (OA) de joelho é uma das condições articulares mais comuns e incapacitantes no mundo, sendo caracterizada como uma doença progressiva que afeta principalmente a cartilagem do joelho. Embora não tenha cura, a detecção precoce é fundamental para prevenir sua progressão, e a radiografia é a principal técnica utilizada para o diagnóstico da OA e para sua classificação com base na escala de Kellgren/Lawrence (KL). No entanto, o diagnóstico radiológico depende da experiência, interpretação e tempo do profissional, o que pode gerar inconsistências ou erros. Nesse contexto, técnicas de aprendizado profundo oferecem uma alternativa mais rápida e eficiente, permitindo a automação da detecção e classificação da OA de joelho.

Este estudo propõe uma comparação entre modelos de redes neurais convolucionais (RNCs) e vision transformers (ViTs) na tarefa de classificar a severidade da OA de joelho, abrangendo os modelos ResNet-34, ResNet-50, ResNet-101, VGG-16, VGG-19, DenseNet-121, DenseNet-169, Inception-v3, DeiT, Swin Transformer, DaViT, MaxViT e GCViT. O treinamento dos modelos foi realizado com o uso de aprendizado por transferência, e a análise comparativa considera métricas de performance, consumo computacional, análise quantitativa de incerteza e interpretabilidade. Os resultados mostraram que as arquiteturas RNCs, especialmente aquelas da família DenseNet apresentaram o melhor desempenho geral, com o modelo DenseNet-169 alcançando uma acurácia de 78,85%. Em termos de eficiência computacional, as RNCs foram significativamente mais rápidas, com o DenseNet-121 oferecendo o melhor equilíbrio entre alto desempenho preditivo (QWK de 0,8878) e baixo custo de treinamento e inferência (3,11 ms/imagem). Os ViTs, apesar de competitivos, apresentaram um desempenho inferior com um custo computacional maior. Finalmente, a análise de interpretabilidade com Grad-CAM confirmou que os modelos de melhor desempenho baseiam suas decisões em marcadores patológicos relevantes, como o espaço articular e osteófitos.

**Palavras-chaves:** Classificação. osteoartrite de joelho. radiografias. redes neurais convolucionais. transfer-learning. vision transformers.



# Abstract

Knee osteoarthritis (OA) is one of the most common and disabling joint conditions worldwide. It is characterized as a progressive disease that primarily affects the knee cartilage. Although it has no cure, early detection is crucial to prevent its progression. Radiography is the main technique used for diagnosing OA and for classifying it based on the Kellgren/Lawrence (KL) scale. However, radiological diagnosis depends on the experience, interpretation, and time of the professional, which can lead to inconsistencies or errors. In this context, deep learning techniques offer a faster and more efficient alternative, enabling the automation of OA detection and classification.

This study proposes a comparison between convolutional neural networks (CNNs) and vision transformers (ViTs) for the task of classifying knee OA severity, including models such as ResNet-34, ResNet-50, ResNet-101, VGG-16, VGG-19, DenseNet-121, DenseNet-169, Inception-v3, DeiT, Swin Transformer, DaViT, MaxViT, and GCViT. The models were trained using transfer learning, and the comparative analysis considers performance metrics, computational cost, quantitative uncertainty analysis, and interpretability. The results showed that CNN architectures, particularly those from the DenseNet family, achieved the best overall performance, with the DenseNet-169 model reaching an accuracy of 78.85%. In terms of computational efficiency, CNNs were significantly faster, with DenseNet-121 offering the best balance between high predictive performance (QWK of 0.8878) and low training and inference cost (3.11 ms/image). Although competitive, ViTs showed lower performance and higher computational cost. Finally, the interpretability analysis using Grad-CAM confirmed that the top-performing models base their decisions on relevant pathological markers, such as joint space and osteophytes.

**Keywords:** Classification. convolutional neural networks. knee osteoarthritis. radiographs. transfer-learning. vision transformers.



# Lista de ilustrações

Figura 1 – Imagens de recuperação por inversão sagital (A–C) e eco de spin rápido coronal (D–F) ilustrando os achados da ressonância magnética na osteoartrite. (A) Sinovite reativa (seta branca espessa), (B) Formação de cistos subcondrais (seta branca), (C) Edema da medula óssea (setas brancas finas), (D) Desgaste parcial da cartilagem (seta preta espessa), (E–F) Desgaste total da cartilagem (setas pretas finas), esclerose subcondral (cabeça de seta) e formação de osteófitos marginais (seta dupla). Fonte: Loeser et al. (2012).	7
Figura 2 – Uma rede neural convolucional simples, composta por apenas cinco camadas. Fonte: Saxena (2022).	11
Figura 3 – Aprendizado residual introduzido pela arquitetura ResNet.	13
Figura 4 – Um bloco de cinco camadas de uma DenseNet. Cada camada recebe como entrada a saída de todas as camadas anteriores. Fonte: Huang et al. (2017)	15
Figura 5 – Um módulo Inception. Fonte: Szegedy et al. (2015)	16
Figura 6 – Um módulo Inception com fatoração de convoluções. Fonte: Szegedy et al. (2015)	17
Figura 7 – Arquitetura do vision transformer. Fonte: Dosovitskiy et al. (2021).	19
Figura 8 – Estratégia de distilação em transformers através da introdução de um token de distilação. Fonte: Touvron et al. (2021).	21
Figura 9 – (a) Mapa de características hierárquico do Swin Transformer. (b) Em contraste, o formato de resolução única dos mapas de características do ViT. Fonte: Liu et al. (2021).	22
Figura 10 – (a) A arquitetura do Swin Transformer (Swin-T); (b) Dois blocos Swin Transformer sucessivos. Fonte: Liu et al. (2021).	22
Figura 11 – (a) <i>Spatial window multihead self-attention</i> divide a dimensão espacial em janelas locais, onde cada janela contém múltiplos tokens espaciais. (b) <i>Channel group single-head self-attention</i> agrupa tokens de canal em múltiplos grupos. Fonte: Ding et al. (2022).	23
Figura 12 – Arquitetura DaViT do bloco <i>dual attention</i> . Fonte: Ding et al. (2022).	24
Figura 13 – Módulo de atenção multi-eixo do MaxViT (Max-SA). O módulo <i>block-attention</i> aplica atenção dentro das janelas, enquanto o módulo <i>grid-attention</i> atua globalmente no espaço 2D. Fonte: Tu et al. (2022).	25
Figura 14 – Arquitetura MaxViT. Fonte: Tu et al. (2022).	26

Figura 15 – Formulação da atenção no GCViT. A atenção local (esquerda) é restrita a uma janela local. Na atenção global (direita), um gerador de consultas extrai características de toda a imagem para formar tokens de consulta globais, que então interagem com os tokens de chave e valor locais, permitindo a captura de informações de longo alcance. Fonte: Hatamizadeh et al. (2022). . . . .	27
Figura 16 – Arquitetura do GCViT. A cada estágio, um gerador de tokens extrai consultas globais que interagem com as representações locais de chave e valor para capturar contexto de longo alcance. Fonte: Hatamizadeh et al. (2022). . . . .	27
Figura 17 – Arquitetura do framework CORN. Fonte: Shi, Cao e Raschka (2023). .	29
Figura 18 – Metodologia proposta por Tariq, Suhail e Nawaz (2023). . . . .	37
Figura 19 – Metodologia proposta por Mohammed et al. (2023). . . . .	38
Figura 20 – Metodologia proposta por Domingues et al. (2023). . . . .	38
Figura 21 – Metodologia proposta por Cueva et al. (2022). . . . .	39
Figura 22 – Metodologia proposta por Sekhri et al. (2023). . . . .	40
Figura 23 – Metodologia proposta por Wang et al. (2024b). . . . .	40
Figura 24 – Metodologia proposta por Apon et al. (2024). . . . .	41
Figura 25 – Visão geral da metodologia adotada neste estudo, desde a coleta de dados até a avaliação dos modelos. . . . .	43
Figura 26 – Distribuição das radiografias por classe KL nos subconjuntos de treino, teste, validação e calibração. . . . .	44
Figura 27 – Exemplo de equalização de histograma aplicada a uma radiografia de joelho. . . . .	45
Figura 28 – Distribuições de intensidade dos pixels antes e depois da equalização de histograma. . . . .	46

# Lista de tabelas

Tabela 1 – Escala de Kellgren/Lawrence para classificação da severidade de osteoartrite. . . . .	9
Tabela 2 – Configuração dos modelos VGG-16 e VGG-19. Fonte: Simonyan e Zisserman (2015). . . . .	12
Tabela 3 – Configuração dos modelos ResNet-34, ResNet-50 e ResNet-101. Fonte: He et al. (2016). . . . .	14
Tabela 4 – Configuração dos modelos DenseNet-121 e DenseNet-169. Fonte: Huang et al. (2017). . . . .	15
Tabela 5 – Configuração do modelo Inception-v3. Fonte Szegedy et al. (2016). . . . .	17
Tabela 6 – Número de radiografias por classe KL no conjunto de dados original. . . . .	44
Tabela 7 – Lista dos modelos utilizados neste estudo, com a fonte, o número de parâmetros e FLOPs estimados. . . . .	47
Tabela 8 – Lista das camadas escolhidas para a geração dos mapas de calor Grad-CAM. . . . .	50
Tabela 9 – Métricas de desempenho de cada modelo na tarefa de classificar a OA de joelho em cinco classes de severidade. . . . .	52
Tabela 10 – Curvas AUC-ROC para os cinco principais modelos. . . . .	53
Tabela 11 – Métrica F1-score para cada uma das cinco classes e modelo, considerando as funções de perda Entropia Cruzada e CORN. . . . .	55
Tabela 12 – Matriz de confusão para os cinco principais modelos. . . . .	56
Tabela 13 – Comparação entre os resultados de diferentes estudos que abordam o problema da OA de joelho. . . . .	57
Tabela 14 – Tempos de treinamento e inferência de cada modelo. . . . .	59
Tabela 15 – Valores de cobertura da predição conformal para cada modelo e função de perda. . . . .	60
Tabela 16 – Histograma do tamanho dos conjuntos de predição para cinco modelos relevantes. . . . .	62
Tabela 17 – Visualização Grad-CAM para os cinco principais modelos. . . . .	64
Tabela 18 – Resumo das métricas gerais de desempenho para modelos selecionados no cenário de 4 classes. . . . .	75
Tabela 19 – Resumo das métricas gerais de desempenho para modelos selecionados no cenário de 3 classes. . . . .	76
Tabela 20 – Resumo das métricas gerais de desempenho para modelos selecionados no cenário de classificação binária. . . . .	77



# Lista de abreviaturas e siglas

AUC-ROC	Area Under the Receiver Operating Characteristic Curve
CAM	Class Activation Mapping
CORN	Conditional Ordinal Regression for Neural Networks
GAP	Global Average Pooling
GBD	Global Burden of Disease
IA	Inteligência Artificial
KL	Kellgren/Lawrence
MAE	Mean Absolute Error
MLP	Multilayer Perceptron
NIH	National Institutes of Health
OA	Osteoartrite
OAI	Osteoarthritis Initiative
PLN	Processamento de Linguagem Natural
QWK	Quadratic Weighted Kappa
ReLU	Rectified Linear Unit
RNC	Rede Neural Convolucional
RMSProp	Root Mean Square Propagation
SGD	Stochastic Gradient Descent
TKA	Total Knee Arthroplasty
ViT	Vision Transformer
WHO	World Health Organization
YLD	Years Lived with Disability



# Sumário

<b>1</b>	<b>INTRODUÇÃO . . . . .</b>	<b>1</b>
<b>1.1</b>	<b>Objetivos . . . . .</b>	<b>3</b>
1.1.1	Objetivo Geral . . . . .	3
1.1.2	Objetivos Específicos . . . . .	3
<b>1.2</b>	<b>Organização do Trabalho . . . . .</b>	<b>4</b>
<b>2</b>	<b>FUNDAMENTAÇÃO TEÓRICA . . . . .</b>	<b>5</b>
<b>2.1</b>	<b>Osteoartrite de Joelho . . . . .</b>	<b>5</b>
2.1.1	Definição e Características Clínicas . . . . .	5
2.1.2	Mudanças Patológicas da OA de Joelho . . . . .	6
2.1.3	Impacto da OA na Qualidade de Vida . . . . .	7
2.1.4	Prevalência da OA . . . . .	8
2.1.5	Diagnóstico e Métodos de Avaliação da OA . . . . .	9
2.1.6	Classificação da OA . . . . .	9
<b>2.2</b>	<b>Rede Neural Convolucional (RNC) . . . . .</b>	<b>10</b>
2.2.1	Visual Geometry Group Network (VGG) . . . . .	11
2.2.2	Residual Network (ResNet) . . . . .	13
2.2.3	Densely Connected Convolutional Networks (DenseNet) . . . . .	14
2.2.4	Inception-v3 . . . . .	16
2.2.5	Aprendizado por Transferência . . . . .	17
<b>2.3</b>	<b>Vision Transformer (ViT) . . . . .</b>	<b>18</b>
2.3.1	Data-efficient image Transformer (DeiT) . . . . .	19
2.3.2	Swin Transformer . . . . .	20
2.3.3	Dual Attention Vision Transformers (DaViT) . . . . .	23
2.3.4	Multi-Axis Vision Transformer (MaxViT) . . . . .	25
2.3.5	Global Context Vision Transformer (GCViT) . . . . .	26
<b>2.4</b>	<b>Funções de Perda . . . . .</b>	<b>28</b>
2.4.1	Entropia Cruzada . . . . .	28
2.4.2	Conditional Ordinal Regression for Neural Networks (CORN) . . . . .	28
<b>2.5</b>	<b>Avaliação e métricas de desempenho . . . . .</b>	<b>30</b>
2.5.1	Acurácia . . . . .	30
2.5.2	Precisão . . . . .	30
2.5.3	Revocação . . . . .	31
2.5.4	F1-Score . . . . .	31
2.5.5	Quadratic Weighted Kappa (QWK) . . . . .	31

2.5.6	Matriz de Confusão . . . . .	31
2.5.7	AUC-ROC . . . . .	32
2.5.8	Eficiência computacional . . . . .	32
2.5.9	Predição Conformal . . . . .	33
2.5.9.1	Verificação de corretude . . . . .	34
2.5.10	Método de visualização . . . . .	34
<b>3</b>	<b>TRABALHOS RELACIONADOS . . . . .</b>	<b>37</b>
<b>4</b>	<b>METODOLOGIA . . . . .</b>	<b>43</b>
4.1	Coleta de dados . . . . .	43
4.2	Pré-processamento das imagens . . . . .	44
4.2.1	Equalização de Histograma . . . . .	45
4.2.2	Normalização . . . . .	45
4.2.3	Aumento de dados . . . . .	46
4.2.4	Subamostragem . . . . .	46
4.3	Treinamento dos modelos . . . . .	47
4.3.1	Hiperparâmetros . . . . .	48
4.3.2	Ambiente de execução . . . . .	48
4.4	Avaliação e Análise Complementar . . . . .	48
4.4.1	Predição Conformal . . . . .	49
4.4.2	Análise do Tempo de Inferência . . . . .	49
4.4.3	Análise de Interpretabilidade com Grad-CAM . . . . .	50
<b>5</b>	<b>RESULTADOS . . . . .</b>	<b>51</b>
5.1	Métricas Gerais de Desempenho . . . . .	51
5.2	Métricas de Desempenho por Classe . . . . .	54
5.3	Comparações com Trabalhos Relacionados . . . . .	54
5.4	Eficiência Computacional . . . . .	58
5.5	Análise Quantitativa . . . . .	58
5.6	Interpretabilidade dos Modelos . . . . .	61
<b>6</b>	<b>CONCLUSÃO . . . . .</b>	<b>65</b>
	<b>REFERÊNCIAS . . . . .</b>	<b>67</b>
	<b>APÊNDICES</b>	<b>73</b>
	<b>APÊNDICE A – RESULTADOS SUPLEMENTARES . . . . .</b>	<b>75</b>
A.1	Classificação em 4 Classes . . . . .	75

A.2	Classificação em 3 Classes	76
A.3	Classificação em 2 Classes	76



# 1 Introdução

A osteoartrite (OA), popularmente conhecida como artrose, é uma forma muito comum de doença reumática, caracterizada como uma condição multifatorial e degenerativa que afeta desde a cartilagem articular até os ossos adjacentes, resultando em sintomas de dor, deformidade e perda de função (KRAUS et al., 2015; PACCA et al., 2018). Esses impactos comprometem significativamente a qualidade de vida, especialmente em grupos mais afetados, como idosos, mulheres e indivíduos obesos (PACCA et al., 2018). Além de sua alta prevalência, a OA é uma das principais causas de incapacidade no mundo, com maior incidência na articulação do joelho, seguida pelo quadril e pela mão. Dados de 2020 apontam que a doença afeta cerca de 7,6% da população global, e projeções indicam um aumento de 60 a 100% até 2050 (COURTIES et al., 2024).

Exercícios de propriocepção e fortalecimento muscular, assim como terapias farmacêuticas, têm sido aplicados a pacientes diagnosticados com OA de joelho com o objetivo de controlar ou reduzir os sintomas de dor, uma vez que não existem medicamentos capazes de retardar o seu desenvolvimento (SARDIM; PRADO; PINFILDI, 2020; LIN et al., 2009). Essa abordagem é especialmente apropriada para pacientes em estágios iniciais da doença, quando a cartilagem ainda não foi completamente degradada (KANAMOTO et al., 2020). No entanto, o diagnóstico depende da experiência e do julgamento clínico do profissional na interpretação das radiografias, o que pode levar a inconsistências entre o grau previsto e o grau real, devido às mínimas diferenças entre os estágios adjacentes da doença (KELLGREN; LAWRENCE, 1957; MOHAMMED et al., 2023). Esses desafios têm impulsionado estudos sobre sistemas automáticos de detecção e classificação da OA de joelho.

A introdução de técnicas de inteligência artificial (IA) nos últimos anos tem permitido a automação de tarefas que antes eram realizadas manualmente, incluindo a interpretação de imagens médicas (WANG et al., 2024a). Alguns exemplos incluem a detecção de pneumonia Tilve et al. (2020), a identificação e classificação de câncer de pulmão em tomografias computadorizadas e a detecção de retinopatia diabética em imagens de fundo de olho (TEKADE; RAJESWARI, 2018; DAI et al., 2021).

No campo da reumatologia, a visão computacional também tem sido aplicada à detecção de OA de joelho a partir de radiografias, com o objetivo de automatizar o processo de diagnóstico, reduzir a subjetividade da interpretação humana e realizar a classificação da severidade da doença através da escala de Kellgren/Lawrence (KL) (MOHAMMED et al., 2023). Esses estudos têm se concentrado em utilizar arquiteturas de aprendizado profundo, como redes neurais convolucionais (RNCs), e compará-las entre si para identificar

qual abordagem oferece melhor desempenho na classificação da severidade da OA. No entanto, a operação de convolução limita o relacionamento entre pixels distantes em uma imagem, o que pode prejudicar a capacidade de captar dependências de longo alcance em radiografias (SHAMSHAD et al., 2023).

Como uma abordagem alternativa, ou até complementar, foram propostas arquiteturas baseadas em transformers, capazes de apresentar um excelente desempenho em tarefas de classificação, como é o caso do vision transformer (ViT) (DOSOVITSKIY et al., 2021). Essas arquiteturas têm sido aplicadas com sucesso em tarefas relacionadas à medicina, como o diagnóstico de COVID-19 a partir de radiografias, classificação de tumores e doenças de retina, tornando-se o estado da arte nesta área (SHAMSHAD et al., 2023).

Apesar dos avanços, persiste uma lacuna na literatura quanto a uma análise comparativa sistemática que avalie não apenas o desempenho preditivo, mas também a eficiência computacional e a interpretabilidade das RNCs em contraposição aos ViTs para a classificação ordinal da OA de joelho. Este trabalho, portanto, busca responder à seguinte questão de pesquisa: “Qual família de arquiteturas, RNC ou ViT, oferece o melhor balanço entre acurácia, robustez ordinal, eficiência e interpretabilidade para a classificação da severidade da OA de joelho a partir de radiografias?”

Para guiar esta investigação de forma objetiva, o estudo testará as seguintes hipóteses:

- **Hipótese 1 (desempenho base):** Modelos de RNC e ViT, quando treinados com a técnica de transferência de aprendizado, são capazes de classificar o grau de osteoartrite em radiografias de joelho com desempenho significativamente superior ao acaso, atingindo valores de F1-score macro superiores a 0,70 e de Quadratic Weighted Kappa (QWK) superiores a 0,80.
- **Hipótese 2 (comparação entre arquiteturas):** As RNCs, devido ao seu forte viés induutivo para o processamento de imagens, apresentarão desempenho preditivo (acurácia, QWK e F1-score macro) igual ou superior aos ViTs, com um custo computacional substancialmente menor, refletido em um tempo de inferência inferior a 50% do observado nos ViTs de capacidade similar.
- **Hipótese 3 (consistência ordinal):** A utilização de uma função de perda ordinal (CORN), que reconhece a relação de ordem entre os graus de severidade, resultará em modelos com maior consistência clínica em comparação com a abordagem categórica padrão (Entropia Cruzada). Essa melhoria será quantificada por um aumento no QWK e uma redução no *Mean Absolute Error* (MAE), mesmo que a acurácia geral não seja necessariamente superior.

- **Hipótese 4 (interpretabilidade clínica):** Os mapas de ativação (Grad-CAM) gerados pelos modelos de melhor desempenho destacarão predominantemente a região do espaço articular e as margens ósseas, áreas clinicamente relevantes para o diagnóstico da OA, demonstrando que o aprendizado não se baseia em características espúrias.

## 1.1 Objetivos

### 1.1.1 Objetivo Geral

O objetivo geral deste trabalho consiste em realizar uma análise comparativa completa entre modelos de RNC e ViT na tarefa de detectar e classificar a OA de joelho usando radiografias, com o intuito de validar as hipóteses propostas e identificar a abordagem mais adequada para uma potencial ferramenta de diagnóstico automatizado.

### 1.1.2 Objetivos Específicos

- Realizar uma revisão bibliográfica sobre a OA de joelho e as técnicas de visão computacional aplicadas à detecção de doenças reumáticas;
- Treinar os modelos propostos para classificar a severidade da OA de joelho a partir de um conjunto de dados público;
- Comparar os modelos de RNC e ViT com base em métricas de performance, eficiência computacional, incerteza preditiva e interpretabilidade;
- Analisar os resultados obtidos e discutir as vantagens e desvantagens de cada abordagem.

A metodologia proposta para atingir os objetivos deste trabalho consistiu nas seguintes etapas: coleta e pré-processamento de um conjunto de dados de radiografias de joelhos com diferentes graus de severidade da OA seguindo a escala KL; implementação da *pipeline* de treinamento dos modelos para classificar a severidade da OA de joelho com hiperparâmetros fixos; avaliação dos modelos com base em métricas de performance, tempos de treinamento e inferência; aplicação da predição conformal para análise quantitativa; interpretação visual dos mapas de ativação; análise dos resultados obtidos e discussão das vantagens e desvantagens de cada abordagem.

## 1.2 Organização do Trabalho

Este trabalho está organizado em seis capítulos, incluindo a introdução. No Capítulo 2, são apresentados os conceitos e definições necessárias para o entendimento deste trabalho, incluindo a osteoartrite de joelho e suas características clínicas, conceitos fundamentais das arquiteturas de aprendizado profundo, incluindo as RNCs e os ViTs. No Capítulo 3, são abordados os trabalhos relacionados. No Capítulo 4, é apresentada a metodologia proposta para atingir os objetivos deste trabalho, assim como a avaliação dos modelos. No Capítulo 5, são apresentados os resultados obtidos e discussões. Por fim, no Capítulo 6, são apresentadas as conclusões finais deste trabalho, apontando as contribuições, limitações e sugestões para trabalhos futuros.

## 2 Fundamentação Teórica

Neste capítulo, são apresentados os principais conceitos e definições que fundamentam este trabalho. A seção 2.1 aborda a osteoartrite de joelho, destacando suas características clínicas. A seção 2.2 introduz as redes neurais convolucionais (RNCs) e as arquiteturas exploradas neste estudo. Em seguida, a seção 2.3 apresenta os vision transformers (ViTs) e as arquiteturas escolhidas. A seção 2.4 descreve as funções de perda utilizadas para a comparação entre os modelos. Por fim, a seção 2.5 detalha as métricas de avaliação adotadas para comparar o desempenho das diferentes abordagens.

### 2.1 Osteoartrite de Joelho

#### 2.1.1 Definição e Características Clínicas

A osteoartrite (OA) é definida como uma doença heterogênea e degenerativa, que afeta as articulações e estruturas ósseas de pacientes, causando sintomas de dor, deformidade e perda de função (LOESER et al., 2012). Considerando os fenótipos da doença, ou seja, as características clínicas e radiográficas observáveis, a OA é considerada altamente heterogênea, podendo ser causada por diversos fatores, incluindo:

- **Idade:** a OA é mais comum em idosos, devido ao desgaste natural e inevitável das articulações ao longo do tempo (ANDERSON; LOESER, 2010).
- **Sexo:** mulheres têm maior risco de desenvolver OA do que homens, especialmente após a menopausa, devido à diminuição dos níveis de estrogênio, que protege a cartilagem articular (TSCHON et al., 2021).
- **Obesidade:** o excesso de peso também é uma condição de risco para a OA, pois aumenta a carga mecânica nas articulações, influenciando o início e a progressão da doença (PACCA et al., 2018).
- **Predisposição genética:** fatores genéticos também podem influenciar o desenvolvimento da OA, como a presença de mutações em genes relacionados à formação e manutenção da cartilagem articular (SPECTOR; MACGREGOR, 2004).
- **Outros fatores:** lesões articulares, atividade física intensa, doenças metabólicas, entre outros.

A OA pode afetar diversas articulações, como joelhos, quadris, mãos, ombros, entre outras. No entanto, a junção do joelho é a área mais afetada devido ao suporte do peso

corporal que está diretamente associado a movimentos essenciais, como caminhar, subir escadas e agachar (KANAMOTO et al., 2020). Portanto, tais fatores fazem com que a doença seja uma das principais causas de dor crônica e incapacidade funcional, levando a uma necessidade de identificar e classificar a OA de forma precisa e precoce, para que o tratamento seja iniciado o mais cedo possível a fim de retardar a progressão da doença e melhorar a qualidade de vida dos pacientes.

### 2.1.2 Mudanças Patológicas da OA de Joelho

Entre as mudanças patológicas observadas na OA, estão:

- **Degradação da cartilagem articular:** a cartilagem articular é um tecido que reveste as extremidades ósseas, permitindo movimentos suaves e absorção de impactos. Na OA, ocorre a perda progressiva da matriz cartilaginosa, onde as células da cartilagem, chamadas de condrócitos, se tornam “ativas” e aumentam a produção de enzimas que degradam a matriz (GOLDRING; MARCU, 2009).
- **Inflamação sinovial:** a membrana sinovial é um tecido que reveste as articulações e produz o líquido sinovial, que lubrifica e nutre a cartilagem. Na OA, ocorre a condição chamada sinovite, onde a membrana sinovial se torna inflamada, causando dano e destruição à cartilagem (PESSLER et al., 2008).
- **Degeneração dos ligamentos:** os ligamentos são estruturas que conectam os ossos e estabilizam as articulações. Na OA, os ligamentos podem sofrer rupturas e degeneração, afetando a mecânica articular. Essa degeneração aumenta a predisposição para o desenvolvimento da doença (LOESER et al., 2012).
- **Degeneração do menisco:** o menisco, estrutura fibrocartilaginosa que atua na absorção de choques e na estabilidade articular, também é afetado na OA. Sua degeneração leva à perda da função de amortecimento e à piora da sobrecarga nas superfícies articulares (LOESER et al., 2012).
- **Alterações ósseas:** o osso subcondral, localizado abaixo da cartilagem, também é afetado na OA, como a formação de osteófitos, que são projeções ósseas anormais, e a esclerose subcondral, que é o aumento da densidade óssea. Essas alterações podem causar dor e limitação de movimentos (KRAAN; BERG, 2007).

A Figura 1 ilustra as mudanças patológicas observadas na OA de joelho a partir de imagens de ressonância magnética.

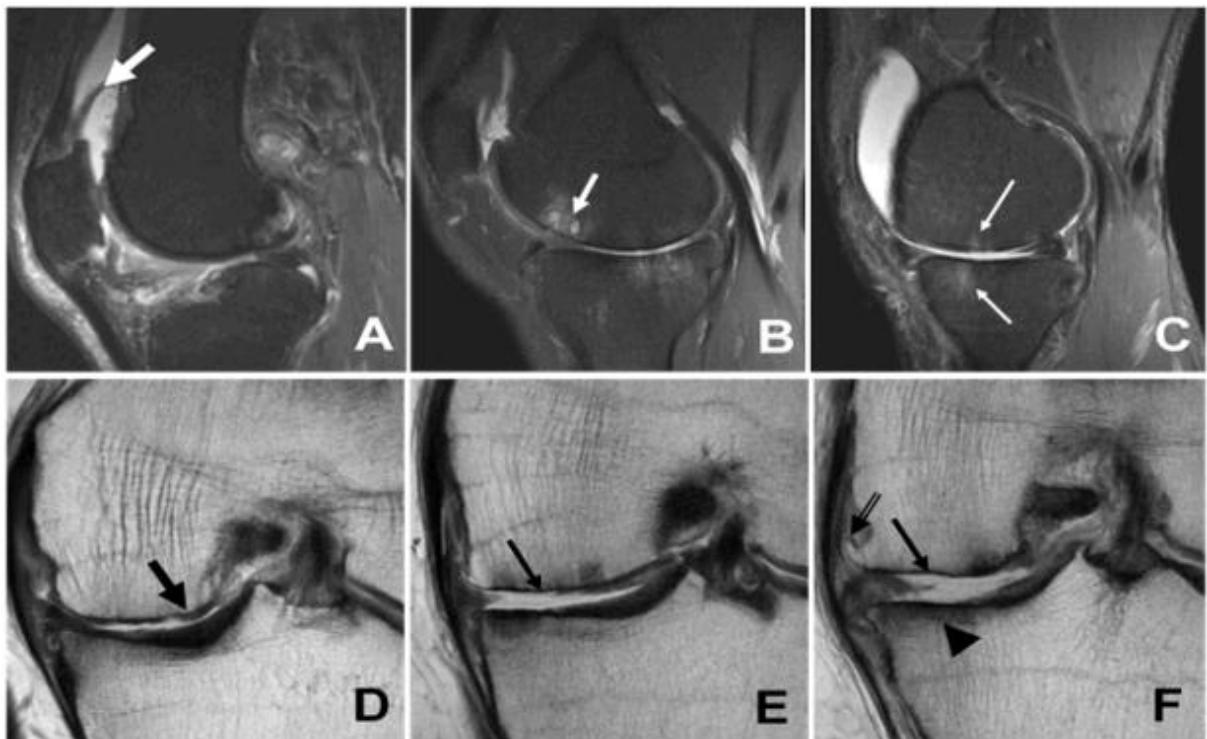


Figura 1 – Imagens de recuperação por inversão sagital (A–C) e eco de spin rápido coronal (D–F) ilustrando os achados da ressonância magnética na osteoartrite. (A) Sinovite reativa (seta branca espessa), (B) Formação de cistos subcondrais (seta branca), (C) Edema da medula óssea (setas brancas finas), (D) Desgaste parcial da cartilagem (seta preta espessa), (E–F) Desgaste total da cartilagem (setas pretas finas), esclerose subcondral (cabeça de seta) e formação de osteófitos marginais (seta dupla). Fonte: Loeser et al. (2012).

### 2.1.3 Impacto da OA na Qualidade de Vida

De acordo com o *World Health Organization* (WHO), a expressão “qualidade de vida” é definida como sendo a percepção do indivíduo sobre sua posição de vida no contexto da cultura e sistema de valores que ele vive e em relação aos seus objetivos, expectativas, padrões e preocupações (ORGANIZATION, 2012).

Existe um grande esforço de pesquisadores e especialistas para avaliar o grau de incapacidade física causado pela doença, além de avaliar os efeitos de diferentes tratamentos em aspectos como dor, função física e mobilidade. No entanto, tais manifestações físicas afetam diretamente outras áreas na vida dos pacientes, como interações sociais, saúde mental e qualidade do sono (FERREL, 1992). Em comparação com outras doenças crônicas, pacientes com doenças muscoesqueléticas, como a OA, são os mais afetados em termos de qualidade de vida. A OA de joelho, especificamente, tende a declinar progressivamente a qualidade de vida conforme a progressão da doença (HOOGEBOOM et al., 2013).

Desmeules et al. (2009) realizaram um estudo com 197 pacientes com cirurgia agendada para substituição total do joelho (*Total Knee Arthroplasty* - TKA) e avaliaram,

através da escala de qualidade de vida SF-36 (WARE; SHERBOURNE, 1992), a relação entre a OA de joelho e a qualidade de vida. Os resultados mostraram que a pontuação média da qualidade de vida dos pacientes era significativamente menor do que a população geral no Canadá ( $p < 0,05$ ). Outros estudos também mostraram resultados similares em pacientes esperando por TKA (SNIDER; MACDONALD; POTOTSCHNIK, 2005; KAPETANAKIS, 2011). É razoável, portanto, que pacientes com OA de joelho severa tenham baixos níveis de qualidade de vida comparado com a população geral.

Sutbeyaz et al. (2007) fizeram um estudo com 28 pacientes obesos com OA de joelho e também avaliaram a qualidade de vida através da escala SF-36. Os resultados mostraram que os pacientes obesos tiveram pontuações muito mais baixas em todos os domínios da escala SF-36, em comparação com o grupo de controle ( $p < 0,001$ ). Com isso, a obesidade foi associada a uma pior qualidade de vida em pacientes com OA de joelho, o que sugere que a perda de peso pode ser benéfica para melhorar o panorama desses pacientes.

Complementarmente, Kawano et al. (2015) mostraram que existe uma relação do nível de escolaridade com a capacidade funcional e dor em pacientes com OA de joelho. O estudo foi conduzido com 93 pacientes tratados no Serviço de Ortopedia e Traumatologia do Hospital Santa Izabel e Santa Casa da Misericórdia da Bahia, em Salvador, Brasil. A avaliação da qualidade de vida foi feita através do questionário SF-36 e mostrou que pacientes com níveis mais baixos de escolaridade tiveram pontuações mais baixas nos domínios de capacidade funcional ( $p < 0,001$ ), limitação funcional ( $p = 0,009$ ) e dor ( $p = 0,01$ ), em comparação com pacientes com níveis mais altos de escolaridade ( $p < 0,05$ ). Com isso, a escolaridade foi associada a uma melhor qualidade de vida em pacientes com OA de joelho, o que sugere que a educação também pode ser um fator importante nesse cenário.

#### 2.1.4 Prevalência da OA

Dados recentes do *Global Burden of Disease* (GBD) — o estudo epidemiológico observacional mais abrangente do mundo — revelaram que a prevalência da OA cresceu 132% entre 1990 e 2020, com projeções de crescimento de 60 a 100% até 2050, alcançando a marca de 1 bilhão de pessoas. Com uma prevalência de 7,6% da população global em 2020, o que equivale a aproximadamente 595 milhões de pessoas, a OA é mais comum em países desenvolvidos, devido à correlação com o status socieconômico, e contribui significativamente para os chamados “anos vividos com incapacidade” (*Years Lived with Disability* - YLD). Além disso, o estudo também aponta que a OA é mais comum em mulheres do que em homens, com prevalência de 8,0% e 5,8%, respectivamente, além de atingir principalmente idosos, especialmente aqueles acima de 70 anos, onde a OA assume a 7<sup>a</sup> posição entre as principais causas de incapacidade, primeiramente afetando a

articulação do joelho (COURTIES et al., 2024).

No Brasil, Senna et al. (2004) realizaram um estudo com mais de 3 mil pessoas e identificaram cerca de 7,2% com doenças reumáticas, sendo a OA a mais comum, com prevalência de 4,14%. Essa prevalência tende a aumentar visto que, além de existir uma correlação entre a OA e a obesidade, estima-se que o Brasil tenha uma taxa de sobrepeso e obesidade combinados de 68,1% em 2030 (BRASÍLIA, 2024).

### 2.1.5 Diagnóstico e Métodos de Avaliação da OA

O diagnóstico da OA normalmente é feito com base em exames clínicos, como a avaliação dos sintomas do paciente, exames de imagem, como radiografias e ressonâncias magnéticas, e exames laboratoriais, como a análise do líquido sinovial (KRAUS et al., 2015). Exames de raio-x têm sido o método mais comum para diagnosticar a OA, pois é uma abordagem acessível que permite visualizar o espaço articular, além de alterações ósseas e cartilaginosas nas articulações, como a formação de osteófitos.

Essa avaliação é tipicamente feita por radiologistas a partir de radiografias do joelho estendido ou flexionado, dependendo da necessidade de visualização intra-articular (BRAUN; GOLD, 2012). A partir dessas imagens, é possível fazer a classificação da severidade da OA e, em caso de diagnóstico, recomendar tratamentos farmacêuticos e não farmacêuticos, como exercícios de fortalecimento muscular e fisioterapia.

### 2.1.6 Classificação da OA

KELLGREN e LAWRENCE (1957) propuseram uma escala de classificação da OA baseada em radiografias e considerando fatores como a formação de osteófitos, estreitamento da cartilagem articular e esclerose subcondral. A escala de Kellgren/Lawrence (KL) classifica a OA em cinco estágios de progressão: 0 (saudável), 1 (duvidoso), 2 (mínimo), 3 (moderado) e 4 (grave) (Tabela 1). Como a classificação é comumente feita por radiologistas, estes avaliam as radiografias e atribuem um grau de acordo com a experiência e o julgamento clínico na interpretação das imagens.

No entanto, a classificação manual pode ser subjetiva e suscetível a erros, assim como foi observado pelos autores, o que pode levar a diagnósticos tardios ou incorretos em um cenário onde a detecção precoce é crucial para retardar a progressão da doença .

<b>0 (saudável)</b>	<b>1 (duvidoso)</b>	<b>2 (mínimo)</b>	<b>3 (moderado)</b>	<b>4 (severo)</b>

Tabela 1 – Escala de Kellgren/Lawrence para classificação da severidade de osteoartrite.

## 2.2 Rede Neural Convolucional (RNC)

Uma rede neural artificial é um modelo computacional inspirado no cérebro humano (MCCULLOCH; PITTS, 1943), onde neurônios artificiais recebem um conjunto de entradas ponderadas, realizam uma soma dessas entradas e aplicam uma função de ativação para produzir uma saída. Essa estrutura permite que as redes neurais aprendam padrões complexos a partir de dados, tornando-as adequadas para tarefas de processamento de linguagem natural, visão computacional, entre outras aplicações.

Em 2006, Hinton, Osindero e Teh (2006) propuseram o uso de redes neurais artificiais com múltiplas camadas com o objetivo de melhorar a capacidade dos modelos, o que levou a um renascimento do interesse nessas redes e ao desenvolvimento de novas arquiteturas, como é o caso da rede neural convolucional (RNC).

As RNCs são modelos de aprendizado profundo projetados para processar dados com estrutura de grade, como imagens. Inspiradas na organização do córtex visual, RNCs são amplamente utilizadas em tarefas de visão computacional, como classificação de imagens, detecção de objetos e segmentação semântica.

A camada de convolução é o componente central das RNCs, responsável por extrair características locais dos dados de entrada. Essa camada utiliza filtros, que são pequenas matrizes de pesos (por exemplo,  $3 \times 3$  ou  $5 \times 5$ ) aplicadas em toda a imagem de entrada para gerar um mapa de características, representando a presença dessas características em diferentes regiões da imagem.

Esses filtros são ajustados durante o treinamento da rede, permitindo que a RNC aprenda a detectar padrões relevantes, como bordas, texturas e formas. Conforme a rede avança pelas camadas, os filtros se tornam mais complexos e capazes de capturar características de alto nível, como objetos inteiros. Após as convoluções, é comum utilizar a função de ativação *Rectified Linear Unit* (ReLU), que substitui valores negativos por zero e introduz não linearidades no modelo, permitindo que ele aprenda representações complexas.

Após as camadas de convolução, as RNCs geralmente incluem camadas de *pooling* para reduzir a dimensionalidade dos mapas de características, enquanto preservam as características mais relevantes. O *pooling* pode ser feito de várias maneiras, como *max pooling* (onde o valor máximo de uma região é mantido) ou *average pooling* (onde a média dos valores é calculada). Esse processo contribui para:

- Reduzir a quantidade de parâmetros e o custo computacional da rede.
- Tornar a rede mais robusta a pequenas variações nos dados de entrada.

Após diversas camadas de convolução e *pooling*, uma ou mais camadas totalmente

conectadas são adicionadas ao final da rede para combinar as características extraídas de camadas anteriores e realizar a tarefa de classificação. Cada neurônio dessas camadas está conectado a todos os neurônios da camada anterior, permitindo decisões baseadas em combinações globais das informações aprendidas. Em tarefas de classificação, essa camada geralmente utiliza a função de ativação *softmax*, que transforma as saídas em probabilidades.

Durante o treinamento, a RNC ajusta os pesos dos filtros por meio do algoritmo de retropropagação, em que o erro de saída é retropropagado pela rede para atualizar os pesos e minimizar a função de perda. Esse processo é repetido por várias épocas, permitindo que a rede aprenda a reconhecer padrões complexos nos dados de entrada.

A Figura 2 ilustra uma rede neural convolucional composta por cinco camadas. O número de camadas, a disposição dessas camadas, o número e tamanho dos filtros, a forma de conexão entre as camadas, entre outros fatores, podem variar dependendo da arquitetura escolhida. Em seguida, serão apresentadas algumas das arquiteturas populares de RNC que foram utilizadas neste trabalho.

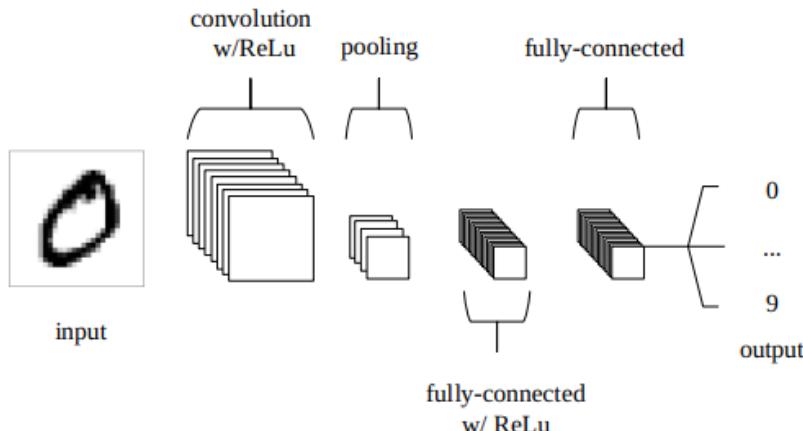


Figura 2 – Uma rede neural convolucional simples, composta por apenas cinco camadas.  
Fonte: Saxena (2022).

### 2.2.1 Visual Geometry Group Network (VGG)

Os modelos VGG foram introduzidos pelo *Visual Geometry Group* da Universidade de Oxford por Simonyan e Zisserman (2015), que depois serviu como base para a competição do ImageNet em 2014, quando conquistaram o primeiro e segundo lugar na época. A arquitetura VGG é conhecida por sua simplicidade e profundidade, utilizando filtros convolucionais pequenos ( $3 \times 3$ ) empilhados em camadas profundas, variando de 11 a 19 camadas. O objetivo dos autores era explorar o impacto da profundidade na performance do modelo, e eles descobriram que redes neurais mais profundas superavam redes mais rasas, desde que treinadas adequadamente.

A arquitetura VGG processa imagens RGB de  $224 \times 224$  pixels, utilizando uma série de camadas convolucionais seguidas por camadas de *pooling*, onde cada camada contém um número crescente de filtros  $3 \times 3$ . O *stride* é fixo em 1 pixel, e o *padding* é utilizado para manter a dimensão da imagem. Após as camadas convolucionais, são aplicadas camadas de *max-pooling* com um tamanho de  $2 \times 2$  e *stride* de 2, reduzindo a dimensão da imagem pela metade. Por fim, são adicionadas três camadas totalmente conectadas, seguidas por uma camada de saída com ativação *softmax* para classificação. Além disso, as camadas ocultas são ativadas por funções ReLU, responsáveis por introduzir a não-linearidade no modelo.

A Tabela 2 apresenta a configuração das arquiteturas VGG-16 e VGG-19, com um total de 16 e 19 camadas, respectivamente. Ambas se destacaram na competição do ImageNet e são amplamente utilizadas devido à sua performance em tarefas de classificação, incluindo o diagnóstico a partir de imagens médicas (SAINI et al., 2023; SITAULA; HOSSAIN, 2021).

VGG-16	VGG-19
16 camadas	19 camadas
imagem RGB de entrada (224 x 224)	
conv3-64	conv3-64
conv3-64	conv3-64
	maxpool
conv3-128	conv3-128
conv3-128	conv3-128
	maxpool
conv3-256	conv3-256
conv3-256	conv3-256
conv3-256	conv3-256
	<b>conv3-256</b>
	maxpool
conv3-512	conv3-512
conv3-512	conv3-512
conv3-512	conv3-512
	<b>conv3-512</b>
	maxpool
conv3-512	conv3-512
conv3-512	conv3-512
conv3-512	conv3-512
	<b>conv3-512</b>
	maxpool
	FC-4096
	FC-4096
	FC-1000
	softmax

Tabela 2 – Configuração dos modelos VGG-16 e VGG-19. Fonte: Simonyan e Zisserman (2015).

### 2.2.2 Residual Network (ResNet)

He et al. (2016) venceram a competição ILSVRC 2015 com a arquitetura ResNet, que introduziu a ideia de blocos residuais e alcançou uma taxa de erro de 3,57% no conjunto de validação do ImageNet com um *ensemble* de seus modelos. Os autores abordaram o problema da degradação de desempenho: conforme a profundidade da rede aumentava, a acurácia saturava e começava a diminuir. Para resolver, eles introduziram a ideia de conexões de atalho entre as camadas, onde o sinal de entrada de uma camada é somado ao sinal de saída de uma camada subsequente (Figura 3).

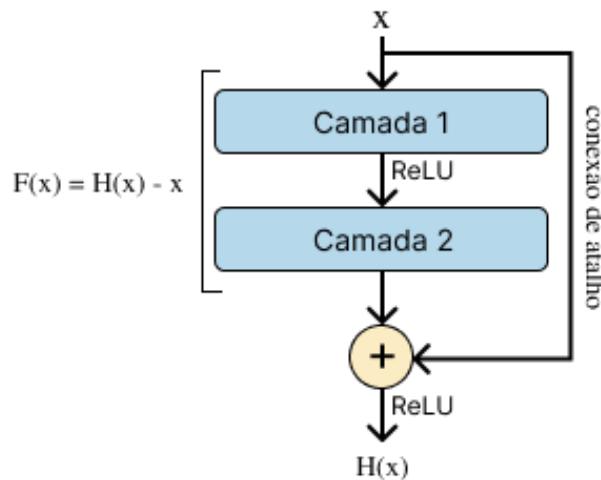


Figura 3 – Aprendizado residual introduzido pela arquitetura ResNet.

Formalmente, considerando que o objetivo de uma rede neural é aprender uma função  $H(x)$ , onde  $x$  é a entrada, a ResNet propõe que a rede aprenda uma função residual  $F(x) = H(x) - x$ , onde a entrada  $x$  é adicionada à saída  $H(x)$ , reformulando a função de aprendizado como  $H(x) = F(x) + x$ . Essa abordagem permite que a rede aprenda funções de identidade mais facilmente, facilitando o treinamento de redes mais profundas sem adicionar complexidade.

A arquitetura ResNet é composta por pilhas de blocos residuais que consistem em duas camadas convolucionais, com uma normalização em lote e uma função de ativação ReLU entre elas. As camadas convolucionais utilizam filtros de tamanho  $3 \times 3$ , com um *stride* de 1 e *padding* de 1, para manter a dimensão da imagem. A saída do bloco residual é então somada à entrada original, permitindo que o modelo aprenda a função residual. A rede termina com uma camada de *average pooling* global e uma camada totalmente conectada com ativação *softmax* para classificação.

A Tabela 3 apresenta a configuração das arquiteturas ResNet-34, ResNet-50 e ResNet-101, que são variantes da ResNet com diferentes profundidades. Essas arquiteturas são amplamente utilizadas devido à sua eficácia em tarefas de classificação de imagens, especialmente em radiografias. Como exemplo, Leung et al. (2020) utilizaram a arquitetura

ResNet-34 para diagnosticar a OA de joelho em pacientes submetidos à TKA e obtiveram resultados que superaram modelos de resultados binários.

Camada	Tamanho da saída	34 camadas	50 camadas	101 camadas
conv1	112×112		7×7, 64, stride 2	
			3×3 max pool, stride 2	
conv2_x	56×56	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 3$	$1 \times 1, 64$ $3 \times 3, 64$ $1 \times 1, 256$ × 3	$1 \times 1, 64$ $3 \times 3, 64$ $1 \times 1, 256$ × 3
conv3_x	28×28	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 4$	$1 \times 1, 128$ $3 \times 3, 128$ $1 \times 1, 512$ × 4	$1 \times 1, 128$ $3 \times 3, 128$ $1 \times 1, 512$ × 4
conv4_x	14×14	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 6$	$1 \times 1, 256$ $3 \times 3, 256$ $1 \times 1, 1024$ × 6	$1 \times 1, 256$ $3 \times 3, 256$ $1 \times 1, 1024$ × 23
conv5_x	7×7	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 3$	$1 \times 1, 512$ $3 \times 3, 512$ $1 \times 1, 2048$ × 3	$1 \times 1, 512$ $3 \times 3, 512$ $1 \times 1, 2048$ × 3
	1×1		average pool, 1000-d fc, softmax	
FLOPs		$3.6 \times 10^9$	$3.8 \times 10^9$	$7.6 \times 10^9$

Tabela 3 – Configuração dos modelos ResNet-34, ResNet-50 e ResNet-101. Fonte: He et al. (2016).

### 2.2.3 Densely Connected Convolutional Networks (DenseNet)

A arquitetura DenseNet introduziu uma nova abordagem para lidar com redes profundas e aliviar o problema de *vanishing gradients*, melhorando a propagação e reuso da informação, além de diminuir o número de parâmetros. A ideia principal foi conectar cada camada a todas as camadas anteriores, formando conexões densas entre elas. Isso significa que cada camada recebe como entrada não apenas a saída da camada anterior, mas também as saídas de todas as camadas anteriores (Figura 4). Essa abordagem permite que o modelo aprenda representações mais ricas e complexas, facilitando a extração de características relevantes para a tarefa de classificação (HUANG et al., 2017).

O componente fundamental da DenseNet é o bloco denso (ou *dense block*, do inglês), que consiste em várias camadas convolucionais conectadas densamente. Cada camada dentro do bloco denso aplica três operações consecutivas: normalização em lote, seguida de uma função de ativação ReLU e, por fim, uma convolução  $3 \times 3$ . Após a aplicação do bloco denso, uma transição é realizada para reduzir a dimensão dos mapas de características usando uma camada de convolução  $1 \times 1$ , seguida por uma camada de *average pooling*  $2 \times 2$ .

Dessa forma, a arquitetura DenseNet é composta por quatro blocos densos, cada um seguido por camadas de transição. A saída final (classificador) é obtida através de uma camada de *global average pooling* e uma camada totalmente conectada com ativação *softmax* para classificação. A Tabela 4 apresenta a configuração dos modelos DenseNet-121 e DenseNet-169, que são variantes da DenseNet com diferentes profundidades e que

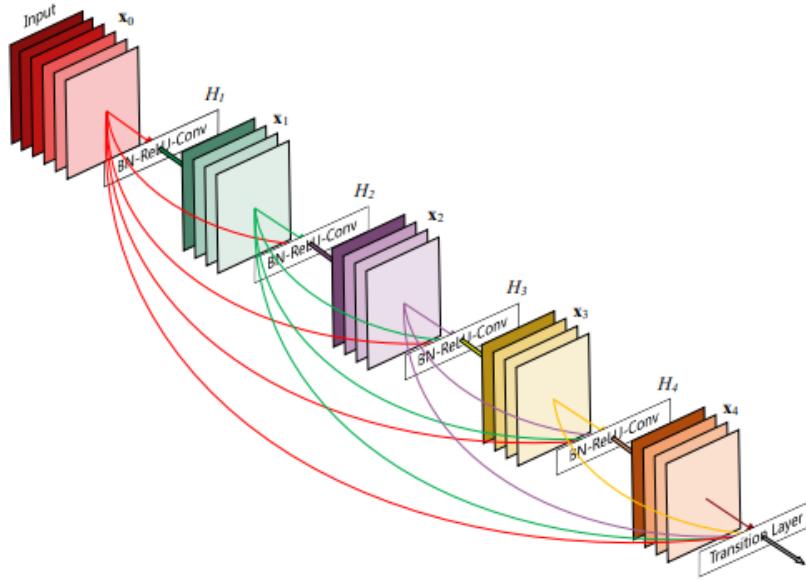


Figura 4 – Um bloco de cinco camadas de uma DenseNet. Cada camada recebe como entrada a saída de todas as camadas anteriores. Fonte: Huang et al. (2017)

fornecem um bom equilíbrio entre complexidade e desempenho em comparação com outras arquiteturas mais profundas.

Nos últimos anos, a arquitetura DenseNet tem sido amplamente utilizada em diversas tarefas de classificação de imagens, incluindo diagnósticos médicos. Por exemplo, (RAJPURKAR et al., 2017) propuseram um modelo chamado CheXNet baseado na arquitetura DenseNet-121 para detectar pneumonia a partir de radiografias torácicas, superando o desempenho médio de radiologistas na métrica F1-score.

Camadas	Tamanho da saída	DenseNet-121	DenseNet-169
Convolution	112×112	7×7 conv, stride 2	
Pooling	56×56	3×3 max pool, stride 2	
Dense Block (1)	56×56	$\begin{bmatrix} 1 \times 3 \\ 3 \times 3 \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 3 \\ 3 \times 3 \end{bmatrix} \times 6$
Transition Layer (1)	56×56	1 × 1 conv	
	28×28	2 × 2 average pool, stride 2	
Dense Block (2)	28×28	$\begin{bmatrix} 1 \times 3 \\ 3 \times 3 \end{bmatrix} \times 12$	$\begin{bmatrix} 1 \times 3 \\ 3 \times 3 \end{bmatrix} \times 12$
Transition Layer (2)	28×28	1 × 1 conv	
	14×14	2 × 2 average pool, stride 2	
Dense Block (3)	14×14	$\begin{bmatrix} 1 \times 3 \\ 3 \times 3 \end{bmatrix} \times 24$	$\begin{bmatrix} 1 \times 3 \\ 3 \times 3 \end{bmatrix} \times 32$
Transition Layer (3)	14×14	1 × 1 conv	
	7×7	2 × 2 average pool, stride 2	
Dense Block (4)	7×7	$\begin{bmatrix} 1 \times 3 \\ 3 \times 3 \end{bmatrix} \times 16$	$\begin{bmatrix} 1 \times 3 \\ 3 \times 3 \end{bmatrix} \times 32$
Classification Layer	1×1	7 × 7 global average pool	1000D fully-connected, softmax

Tabela 4 – Configuração dos modelos DenseNet-121 e DenseNet-169. Fonte: Huang et al. (2017).

## 2.2.4 Inception-v3

A arquitetura Inception, introduzida por Szegedy et al. (2015) no contexto do desafio ILSVRC 2014, representou um avanço significativo na evolução das redes neurais convolucionais. Seu principal diferencial está na proposta de uma estrutura modular — o módulo Inception — que combina convoluções de diferentes tamanhos ( $1 \times 1$ ,  $3 \times 3$ ,  $5 \times 5$ ) e operações de *pooling* em paralelo, promovendo o processamento de informações em múltiplas escalas (Figura 5).

O modelo GoogLeNet, uma instância da arquitetura Inception com 22 camadas profundas, obteve a primeira colocação no ILSVRC 2014 (RUSSAKOVSKY et al., 2015), alcançando um notável desempenho em tarefas de classificação e detecção, mesmo utilizando significativamente menos parâmetros que modelos anteriores, como o VGG.

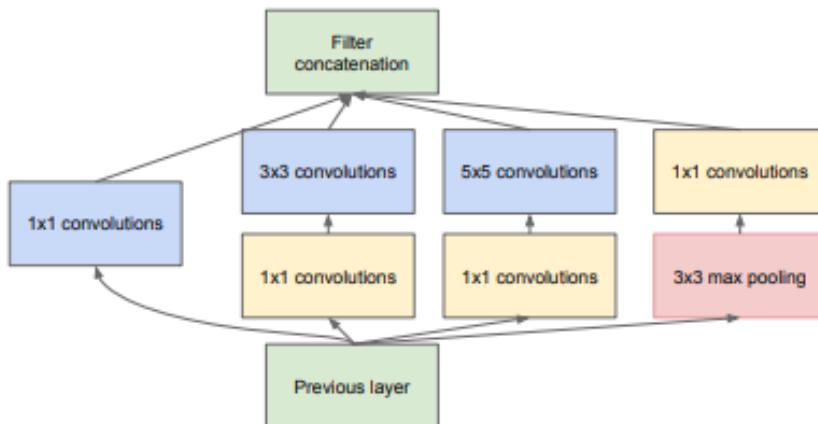


Figura 5 – Um módulo Inception. Fonte: Szegedy et al. (2015)

A arquitetura Inception-v3 (SZEGEDY et al., 2016) representa uma evolução significativa em relação ao modelo original Inception (GoogLeNet), incorporando diversas inovações voltadas à melhoria da eficiência computacional e da acurácia. Entre as principais contribuições estão a fatoração de convoluções em operações menores e assimétricas (Figura 6), o uso mais sistemático da normalização em lote e a adoção da técnica de *label smoothing* como forma de regularização. Tais aprimoramentos resultaram em um modelo mais profundo e preciso, mantendo um custo computacional viável para aplicações práticas.

A Tabela 5 apresenta a configuração da arquitetura Inception-v3, com um total de 42 camadas, que inclui a fatoração de convoluções tradicionais  $7 \times 7$  em convoluções  $3 \times 3$ . A arquitetura substitui o otimizador padrão do *Stochastic Gradient Descent* (SGD) por um otimizador mais avançado, o *Root Mean Square Propagation* (RMSProp), favorecendo a convergência do modelo durante o treinamento, além de utilizar classificadores auxiliares com normalização em lote nas camadas intermediárias, melhorando a propagação do sinal do gradiente e, por consequência, a eficiência do treinamento.

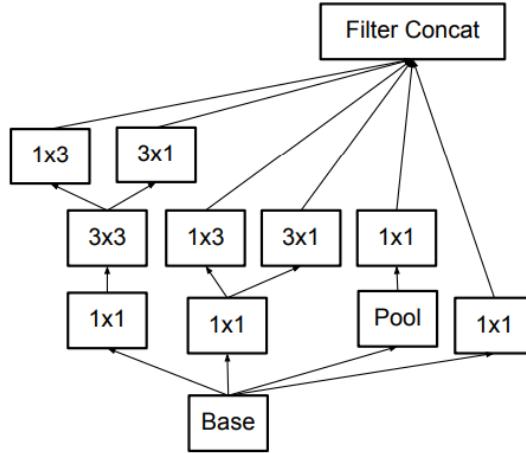


Figura 6 – Um módulo Inception com fatoração de convoluções. Fonte: Szegedy et al. (2015)

Tipo	Tamanho do patch/stride	Tamanho da entrada
conv	$3 \times 3/2$	$299 \times 299 \times 3$
conv	$3 \times 3/1$	$149 \times 149 \times 32$
conv padded	$3 \times 3/1$	$147 \times 147 \times 32$
pool	$3 \times 3/2$	$147 \times 147 \times 64$
conv	$3 \times 3/1$	$73 \times 73 \times 64$
conv	$3 \times 3/2$	$71 \times 71 \times 80$
conv	$3 \times 3/1$	$35 \times 35 \times 192$
$3 \times$ Inception	-	$35 \times 35 \times 288$
$5 \times$ Inception	-	$17 \times 17 \times 768$
$2 \times$ Inception	-	$8 \times 8 \times 1280$
pool	$8 \times 8$	$8 \times 8 \times 2048$
linear	logits	$1 \times 1 \times 2048$
softmax	classifier	$1 \times 1 \times 1000$

Tabela 5 – Configuração do modelo Inception-v3. Fonte Szegedy et al. (2016).

Além de seu excelente desempenho na tarefa de classificação de imagens do ILSVRC 2012 (RUSSAKOVSKY et al., 2015), a arquitetura Inception-v3 tem sido utilizada em outras aplicações, incluindo o diagnóstico médico. Por exemplo, Mujahid et al. (2022) adotaram a arquitetura Inception-v3 para a tarefa de classificação de pneumonia em radiografias e obtiveram resultados promissores, alcançando uma acurácia de 99,29% com um modelo *ensemble*, superando outros modelos, como VGG-16 e ResNet-50.

## 2.2.5 Aprendizado por Transferência

O aprendizado por transferência (ZHUANG et al., 2021) é uma técnica de aprendizado de máquina na qual o conhecimento adquirido por um modelo treinado em uma tarefa é reutilizado para solucionar outra tarefa relacionada, mas diferente. Essa abordagem é especialmente útil para evitar o treinamento de modelos do zero, economizando tempo e recursos computacionais, além de melhorar o desempenho em tarefas com poucos dados disponíveis.

Em redes neurais, o aprendizado por transferência é frequentemente realizado reutilizando pesos de um modelo pré-treinado, cujos estágios iniciais da rede geralmente capturam características genéricas das entradas, como bordas ou texturas, que podem ser úteis para resolver novos problemas. Por exemplo, redes neurais treinadas em grandes conjuntos de dados, como o ImageNet (RUSSAKOVSKY et al., 2015), podem ser reaproveitadas para resolver tarefas específicas, como a classificação de imagens médicas.

Essa estratégia é realizada através do ajuste fino (ou *fine-tuning*, do inglês) do modelo pré-treinado em duas etapas principais. Na primeira, caso seja necessário, as camadas finais do modelo são substituídas por novas camadas adaptadas à tarefa-alvo, como uma camada totalmente conectada com o número de classes correspondente. Na segunda etapa, parte ou toda a rede é treinada com os novos dados. As camadas iniciais geralmente são mantidas inalteradas, enquanto as camadas finais são ajustadas para aprender as características específicas da nova tarefa.

Aplicações de visão computacional e processamento de linguagem natural têm se beneficiado da transferência de aprendizado. Ao reduzir a necessidade de grandes volumes de dados e de poder computacional, essa técnica torna-se uma alternativa viável e eficiente para o desenvolvimento de soluções baseadas em redes neurais profundas.

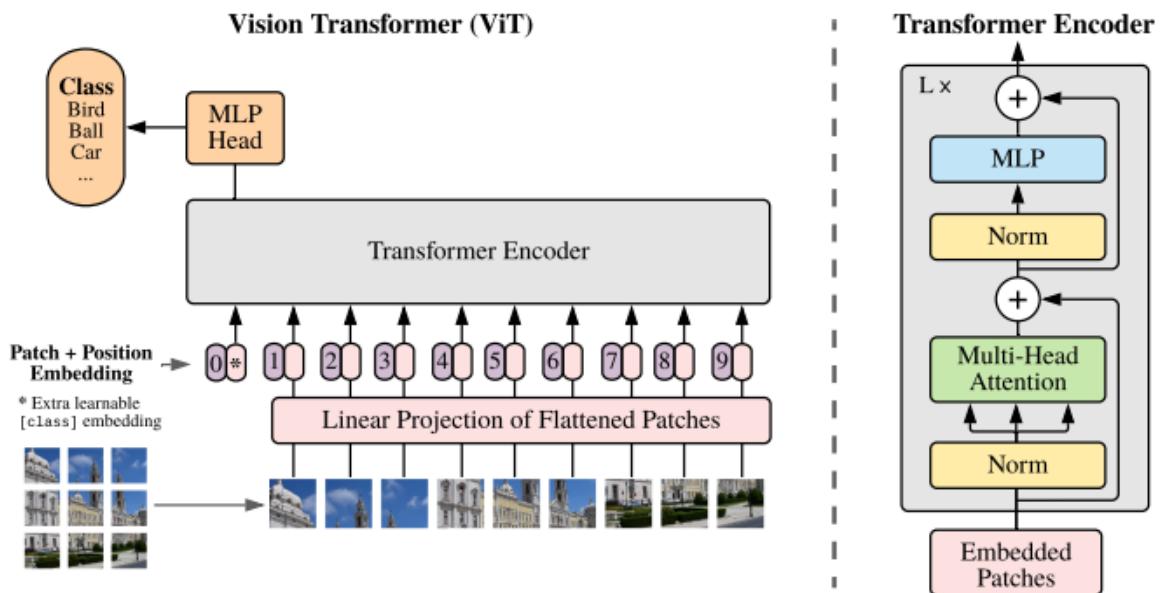
## 2.3 Vision Transformer (ViT)

O vision transformer (ViT) é uma abordagem inovadora de aprendizado profundo que aplica a arquitetura transformer (VASWANI et al., 2023), originalmente desenvolvida para tarefas de Processamento de Linguagem Natural (PLN), ao domínio da visão computacional. Introduzido por Dosovitskiy et al. (2021) no artigo “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”, o ViT demonstrou que os transformers podem ser eficazes para tarefas de classificação de imagens ao obter excelentes resultados quando treinados em grandes conjuntos de dados (14M-300M de imagens), superando modelos tradicionais baseados em RNCs, como o ResNet.

A ideia principal do ViT, ilustrada na Figura 7, é tratar imagens como sequências de blocos com tamanho fixo (por exemplo,  $16 \times 16$  pixels), semelhantes aos tokens em uma sequência de texto. Cada bloco é linearmente projetado em um vetor de dimensão fixa, e esses vetores resultantes são combinados em sequência junto com vetores de posição e de classe, para preservar a informação espacial e representar a classe da tarefa de classificação, respectivamente. Esses vetores são então alimentados em um modelo *encoder*, onde a sequência é processada por camadas de *multi-head self-attention* e *feedforward*, como no transformer tradicional. O mecanismo de auto-atenção permite que o modelo aprenda relações de longo alcance entre diferentes regiões da imagem, sem a necessidade de convoluções locais, oferecendo maior flexibilidade na captura de dependências espaciais.

Ao final do processamento, o token de classificação é utilizado para realizar a predição da tarefa-alvo, como prever o nível de severidade de uma doença.

Em cenários com poucos dados, as RNCs tendem a apresentar melhor desempenho, enquanto os ViTs se destacam no cenário oposto. Isso ocorre porque os transformers não possuem os vieses indutivos herdados pelas redes convolucionais, como a hierarquia espacial, a localidade e a translação equivariante, que são fundamentais para a generalização dos modelos. No entanto, modelos de ViT podem ser adaptados para funcionarem bem com conjuntos de dados reduzidos através do uso de técnicas de pré-treinamento e ajuste fino.



Todas as variantes de ViT compartilham a mesma estrutura básica, que consiste na divisão da imagem em *patches* de tamanho fixo, a projeção linear desses *patches* em vetores de dimensão fixa, a inclusão de vetores de posição e um token de classe, e o processamento desses vetores em um *encoder* de transformer. O ViT-B/16 (DOSOVITSKIY et al., 2021) é uma das primeiras variantes da arquitetura, onde “B” representa o modelo base e “16” refere-se ao tamanho do *patch* em que a imagem é dividida ( $16 \times 16$  pixels). Os modelos que surgiram posteriormente introduziram melhorias e adaptações buscando aumentar a eficiência e/ou reduzir a necessidade de grandes volumes de dados para treinamento. A seguir, são apresentadas as variantes que foram utilizadas nesta pesquisa.

### 2.3.1 Data-efficient image Transformer (DeiT)

A arquitetura DeiT, introduzida por pesquisadores do *Facebook* em 2021 (TOUVRON et al., 2021), representa um avanço significativo na adaptação de transformers. Além de ser uma abordagem livre de convoluções, ela se destaca por não necessitar de grandes

volumes de dados e infraestrutura computacional para alcançar resultados competitivos, ao contrário do que se pressupõe de arquiteturas ViT (DOSOVITSKIY et al., 2021).

O diferencial do DeiT reside na introdução de uma nova estratégia de distilação de conhecimento, adaptada especificamente para a arquitetura transformer. Como ilustrado na Figura 8, um token de distilação é incorporado diretamente à entrada do transformer e atua de maneira similar ao token de classificação: interage com os demais tokens da rede através das camadas de auto-atenção e sua saída é observada após a última camada. Este token é treinado com o objetivo de replicar a predição de um “modelo professor”, estratégia conhecida como *hard-label distillation*:

$$L_{\text{global}}^{\text{hardDistill}} = \frac{1}{2} L_{CE}(\psi(Z_s), y) + \frac{1}{2} L_{CE}(\psi(Z_s), y_t), \quad (2.1)$$

onde  $Z_s$  são os *logits* do “modelo aluno”,  $L_{CE}$  é a entropia cruzada sobre os rótulos corretos ( $y$ ) e os rótulos preditos pelo “modelo professor” ( $y_t = \text{argmax}_c Z_t(c)$ ), sendo  $Z_t$  os seus *logits*, e  $\psi$  é a função *softmax*. Como resultado, ambos os tokens compartilham informação ao longo das camadas e gradualmente convergem para vetores similares, porém ainda distintos. Por fim, seus valores são associados com classificadores lineares para produzir o rótulo da imagem.

Entre suas variantes, o modelo DeiT-B com a estratégia de distilação, que possui arquitetura semelhante ao ViT-B, é o maior modelo em termos de número de parâmetros (87 milhões). Em experimentos com o ImageNet-1K, tal modelo atingiu uma acurácia top-1 de 83,4% (com entrada de  $224 \times 224$  pixels), superando arquiteturas de RNC e inclusive variantes do ViT pré-treinadas com conjuntos de dados significativamente maiores. Adicionalmente, avaliações em tarefas de *transfer learning* em diversos *benchmarks* (CIFAR-10, CIFAR-100, Flowers) demonstram a capacidade de generalização do modelo, onde o DeiT ficou no mesmo nível que RNCs competitivas e superou modelos ViT tradicionais.

Diante desses resultados, o DeiT se mostra como uma alternativa promissora e eficiente aos modelos convolucionais e ViT clássicos para diversas tarefas, incluindo análise de imagens médicas. Por exemplo, Alotaibi et al. (2022) propuseram um modelo *ensemble* com ViT e DeiT (ViT-DeiT) para classificar imagens histopatológicas do câncer de mama em oito classes (benignas e malignas), obtendo um resultado de 98,17% de acurácia.

### 2.3.2 Swin Transformer

A adaptação de arquiteturas transformer para tarefas de visão computacional apresenta desafios únicos, como a grande variação de escala das entidades visuais e a alta resolução das imagens. Em resposta a esses desafios, Liu et al. (2021) propuseram o Swin Transformer, uma nova arquitetura de ViT que serve como uma espinha dorsal de propósito geral para a área. O modelo introduz uma abordagem hierárquica e um

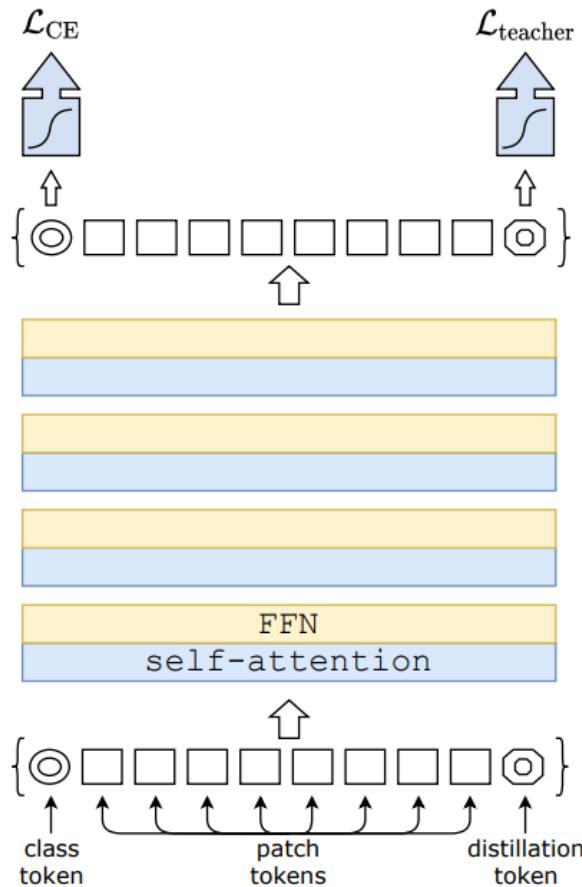


Figura 8 – Estratégia de distilação em transformers através da introdução de um token de distilação. Fonte: Touvron et al. (2021).

mecanismo de auto-atenção baseado em janelas deslocadas, o que lhe confere eficiência e flexibilidade para modelar em múltiplas escalas com complexidade computacional linear em relação ao tamanho da imagem.

A representação hierárquica do Swin Transformer, começando com pequenos *patches* e aumentando gradualmente a resolução (Figura 9), e o esquema de janelas deslocadas são os principais diferenciais do Swin Transformer em relação a outras arquiteturas ViT, limitando o cálculo da auto-atenção a janelas locais e não sobrepostas, ao mesmo tempo que permite conexões cruzadas entre essas janelas em camadas consecutivas. Essa estratégia aumenta significativamente o poder de modelagem sem sacrificar a eficiência.

A arquitetura do Swin Transformer, ilustrada na Figura 10, representa a versão *tiny* do modelo (Swin-T), que é a menor variante. Inicialmente, a imagem é dividida em *patches* (tokens), e um conjunto de blocos Swin Transformer é aplicado sobre esses tokens. Para criar a hierarquia, camadas de fusão de *patches* reduzem a resolução espacial (por um fator de 2x) e aumentam a dimensão dos canais (por 2x) à medida que a rede se aprofunda. Isso permite que o modelo gere mapas de características em múltiplas escalas (por exemplo, 4x, 8x, 16x e 32x), tornando-o compatível com tarefas de predição densa,

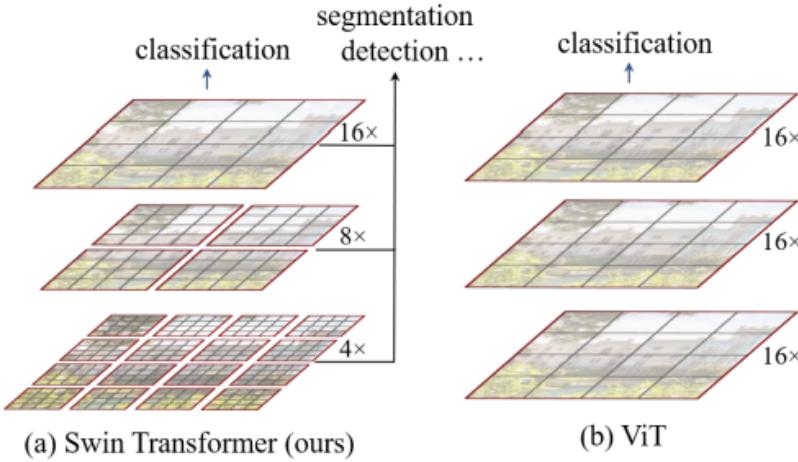


Figura 9 – (a) Mapa de características hierárquico do Swin Transformer. (b) Em contraste, o formato de resolução única dos mapas de características do ViT. Fonte: Liu et al. (2021).

como a detecção de objetos e a segmentação.

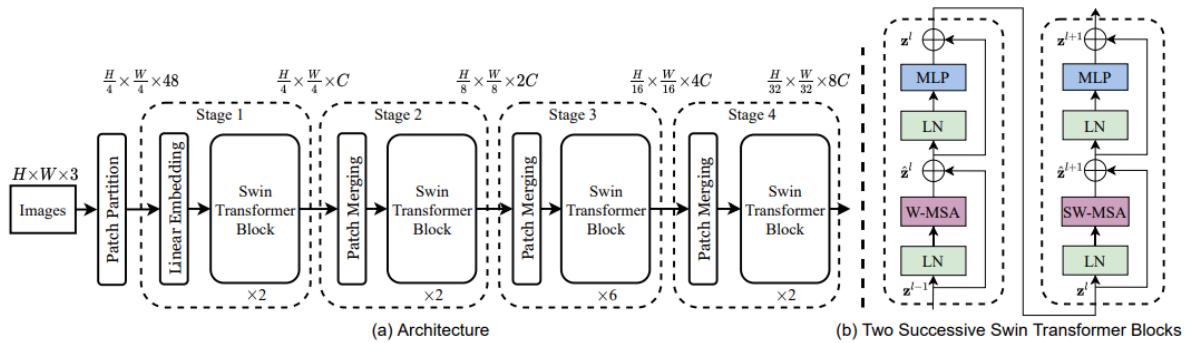


Figura 10 – (a) A arquitetura do Swin Transformer (Swin-T); (b) Dois blocos Swin Transformer sucessivos. Fonte: Liu et al. (2021).

Em camadas consecutivas, os blocos Swin Transformer alternam entre duas configurações de atenção: uma baseada em janelas regulares (W-MSA) e outra em janelas deslocadas (SW-MSA). A formulação de dois blocos sucessivos é dada por:

$$\hat{z}^l = \text{W-MSA}(\text{LN}(z^{l-1})) + z^{l-1} , \quad (2.2)$$

$$z^l = \text{MLP}(\text{LN}(\hat{z}^l)) + \hat{z}^l , \quad (2.3)$$

$$\hat{z}^{l+1} = \text{SW-MSA}(\text{LN}(z^l)) + z^l , \quad (2.4)$$

$$z^{l+1} = \text{MLP}(\text{LN}(\hat{z}^{l+1})) + \hat{z}^{l+1} , \quad (2.5)$$

onde  $\hat{z}$  e  $z$  denotam as saídas dos módulos de atenção e da MLP para um bloco  $l$ , respectivamente. A atenção é sempre calculada com um viés de posição relativa, o que se mostrou crucial para o desempenho do modelo.

O Swin Transformer possui quatro configurações principais: Swin-T, Swin-S, Swin-B e Swin-L, que variam em capacidade. A versão base (Swin-B), possui 88 milhões de parâmetros e alcançou uma acurácia top-1 de 83,5% no ImageNet-1K (com entrada de  $224 \times 224$  pixels), superando os modelos ViT-B/16 (77,91%) e DeiT-B com distilação (83,4%) (DOSOVITSKIY et al., 2021; TOUVRON et al., 2021).

### 2.3.3 Dual Attention Vision Transformers (DaViT)

Com o avanço das arquiteturas de ViT, diversos métodos têm buscado o equilíbrio entre a capacidade de capturar contexto global e a eficiência computacional necessária para lidar com imagens de alta resolução. Nesse contexto, Ding et al. (2022) propuseram uma nova arquitetura de ViT que introduz um mecanismo de atenção dual, combinando janelas espaciais de atenção e grupos de canais de atenção, de forma a integrar representações locais e globais de maneira eficiente e complementar.

O principal diferencial do DaViT está na aplicação do mecanismo de atenção no domínio dos canais. Após transpor o vetor de características gerado pelo mecanismo de auto-atenção em blocos locais, cada canal passa a representar uma visão abstrata global da imagem. A atenção é então aplicada entre os grupos de canais, o que permite o modelo capturar interações globais com complexidade linear. A Figura 11 ilustra a perspectiva ortogonal do DaViT.

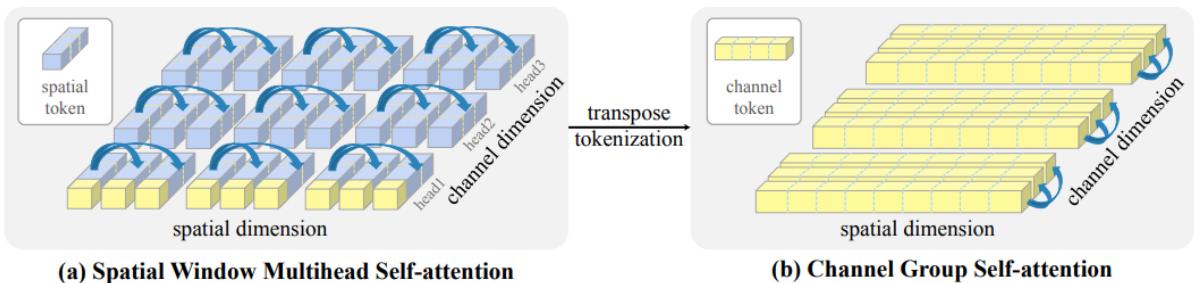


Figura 11 – (a) *Spatial window multihead self-attention* divide a dimensão espacial em janelas locais, onde cada janela contém múltiplos tokens espaciais. (b) *Channel group single-head self-attention* agrupa tokens de canal em múltiplos grupos. Fonte: Ding et al. (2022).

O mecanismo de atenção local, aplicado em janelas espaciais, está ilustrado na Figura 12(b). Ele divide a imagem em janelas não sobrepostas e aplica a atenção apenas entre os tokens espaciais (*patches* da imagem) dentro de cada janela. Supondo  $N_w$  janelas diferentes contendo  $P_w$  *patches* cada, onde  $P = P_w * N_w$ , o mecanismo de atenção local pode ser representado como:

$$A_{\text{window}}(Q, K, V) = \{A(Q_i, K_i, V_i)\}_{i=0}^{N_w}, \quad (2.6)$$

onde  $Q_i$ ,  $K_i$  e  $V_i$  são os vetores de consulta, chave e valor correspondentes a cada janela. Isso reduz significativamente o custo computacional, visto que a complexidade é linear com tamanho espacial  $P$ , embora isso limite a capacidade do modelo de capturar relações de longo alcance.

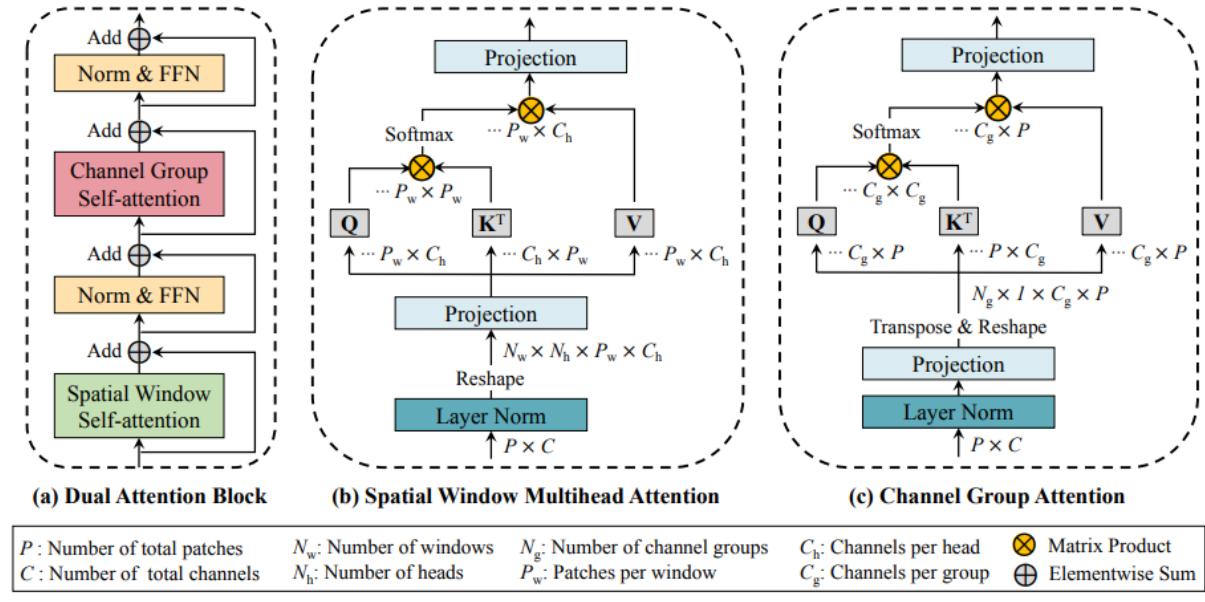


Figura 12 – Arquitetura DaViT do bloco *dual attention*. Fonte: Ding et al. (2022).

Já o mecanismo de atenção global, aplicado em grupos de canais, é ilustrado na Figura 12(c). Ao invés de atuar sobre *patches* espaciais, esta abordagem transpõe o vetor de características e aplica a atenção em tokens de canal. Cada token de canal representa uma visão abstrata global da imagem, pois abrange todos os locais espaciais. Ao computar a atenção entre esses tokens, o modelo consegue naturalmente capturar interações globais com complexidade linear. Formalmente, seja  $N_g$  o número de grupos e  $C_g$  o número de canais em cada grupo, tem-se que  $C = N_g * C_g$ . Assim:

$$A_{\text{channel}}(Q, K, V) = \{A_{\text{group}}(Q_i, K_i, V_i)^T\}_{i=0}^{N_g} \quad (2.7)$$

$$A_{\text{group}}(Q_i, K_i, V_i) = \text{softmax} \left( \frac{Q_i^T K_i}{\sqrt{C_g}} \right) V_i^T, \quad (2.8)$$

onde  $Q_i, K_i, V_i \in \mathbb{R}^{P \times C_g}$  são os vetores de consulta, chave e valor correspondentes a cada grupo de canais.

Existem três configurações diferentes da arquitetura DaViT para classificação de imagens, detecção de objetos e segmentação, que diferem na quantidade de camadas, tamanho do *patch*, número de grupos em cada canal e número de cabeças de atenção. O modelo DaViT-B é a maior configuração, com quase 88 milhões de parâmetros, e obteve

acurácia top-1 de 84,6% no ImageNet-1K (com entrada de  $224 \times 224$  pixels), superando modelos como o DeiT-B com distilação (83,4%) e o Swin-B (83,5%) (TOUVRON et al., 2021; LIU et al., 2021).

### 2.3.4 Multi-Axis Vision Transformer (MaxViT)

A escalabilidade da auto-atenção em transformers para imagens de alta resolução tem sido um desafio significativo, limitando sua aplicação em arquiteturas de visão de ponta. Para superar essa barreira, Tu et al. (2022) propuseram o MaxViT, uma arquitetura que introduz um modelo de atenção eficiente e escalável, denominado auto-atenção multi-eixo (*multi-axis self-attention* - Max-SA). Essa abordagem combina convoluções e um novo módulo de atenção que efetivamente captura interações espaciais locais e globais com complexidade apenas linear, permitindo que o modelo “veja” globalmente em todas as etapas da rede.

A Figura 13 ilustra o conceito fundamental do Max-SA. O mecanismo de atenção em bloco é responsável pelas interações locais. Seja  $X \in \mathbb{R}^{H \times W \times C}$  a entrada de um mapa de características, a ideia é dividi-lo em um vetor na forma  $(\frac{H}{P} \times \frac{W}{P}, P \times P \times C)$ , representando a partição da imagem em janelas não sobrepostas de tamanho  $P \times P$ . A atenção é então aplicada dentro dessas janelas, permitindo que o modelo capture relações locais.

O módulo de atenção em grade, por outro lado, é responsável pelas interações globais do espaço 2D. Em vez de usar janelas de tamanho fixo, ela divide o mapa de características em uma grade uniforme na forma  $(G \times G, \frac{H}{G} \times \frac{W}{G}, C)$  usando um tamanho de grade fixo  $G \times G$ . Isso cria janelas de tamanho adaptativo, e a auto-atenção é aplicada entre os pixels que caem na mesma posição relativa dentro de cada célula da grade. Esse processo corresponde a uma mistura espacial dilatada e global dos tokens, permitindo um campo receptivo global com complexidade também linear.

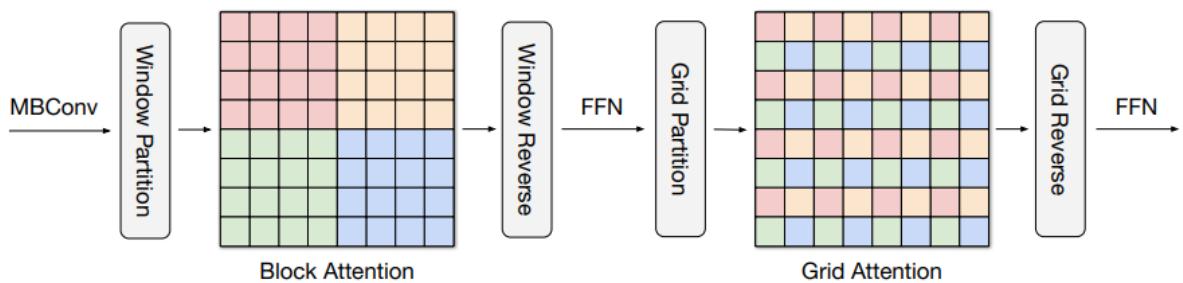


Figura 13 – Módulo de atenção multi-eixo do MaxViT (Max-SA). O módulo *block-attention* aplica atenção dentro das janelas, enquanto o módulo *grid-attention* atua globalmente no espaço 2D. Fonte: Tu et al. (2022).

Esses dois mecanismos de atenção são combinados com uma camada de convolução MBConv para formar o bloco MaxViT, a unidade fundamental da arquitetura, conforme ilustrado na Figura 14. Esses blocos são empilhados para formar a arquitetura MaxViT,

que por sua vez possui algumas variantes, como MaxViT-T, MaxViT-B e MaxViT-L, que aumentam em número de blocos e canais em cada estágio para escalar a capacidade do modelo. O modelo MaxViT-L, por exemplo, estabeleceu um novo estado da arte na classificação do ImageNet-1K, alcançando uma acurácia top-1 de 85,17% (com entrada de  $224 \times 224$  pixels), seguido pelo MaxViT-B com 84,95%, superando também modelos anteriores como o DeiT-B (83,4%), Swin-B (83,5%) e DaViT-B (84,6%).

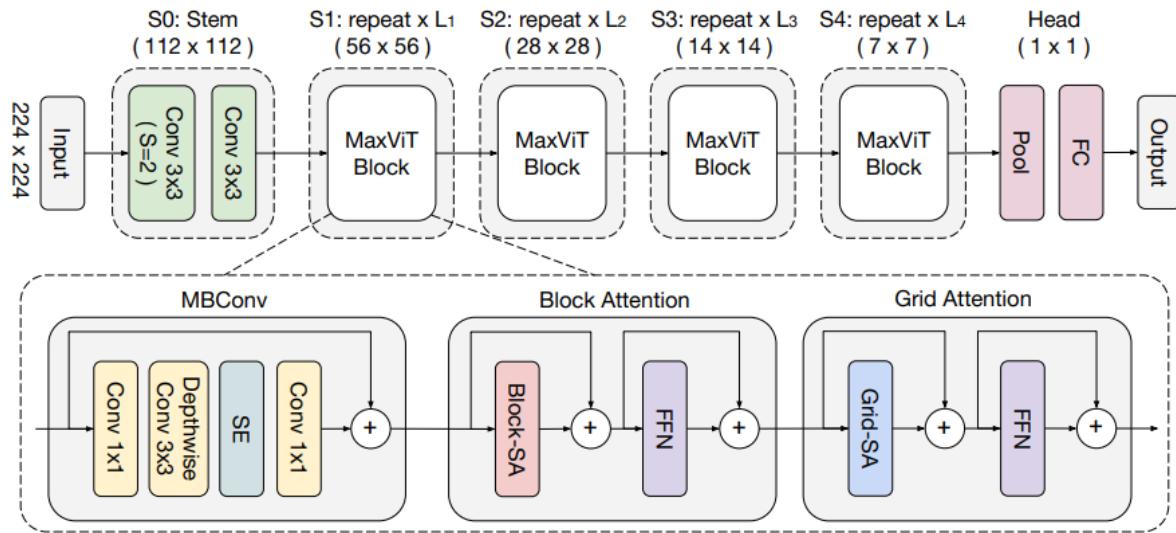


Figura 14 – Arquitetura MaxViT. Fonte: Tu et al. (2022).

### 2.3.5 Global Context Vision Transformer (GCViT)

Em 2022, Hatamizadeh et al. (2022) introduziram o GCViT, uma nova arquitetura que aumenta a eficiência de cômputo e parâmetros ao integrar módulos de auto-atenção de contexto global com a atenção local tradicional, modelando de forma eficaz as interações espaciais de curta e longa distância. Além disso, os autores propuseram o uso de blocos residuais Fused-MBConv modificados, que incorporaram o viés induutivo convolucional na arquitetura.

O GCViT surgiu para resolver as limitações dos modelos ViT anteriores, que apesar do progresso, o campo receptivo limitado das janelas locais restringia a capacidade de capturar informações de longo alcance, e esquemas de deslocamento de janelas apenas cobriam uma pequena fração do contexto global.

O diferencial do GCViT é a sua capacidade de capturar informações globais sem a necessidade de operações custosas, como o deslocamento de janelas. Para isso, a cada estágio da sua arquitetura hierárquica, o modelo utiliza um gerador de consultas para extrair “tokens de consulta globais”. Esses tokens globais, que contêm informações contextuais de diferentes regiões da imagem, são então compartilhados entre todos os módulos de atenção

global para interagir com as representações locais de chave e valor. A Figura 15 ilustra a diferença entre a atenção local e a atenção global com consultas globais.

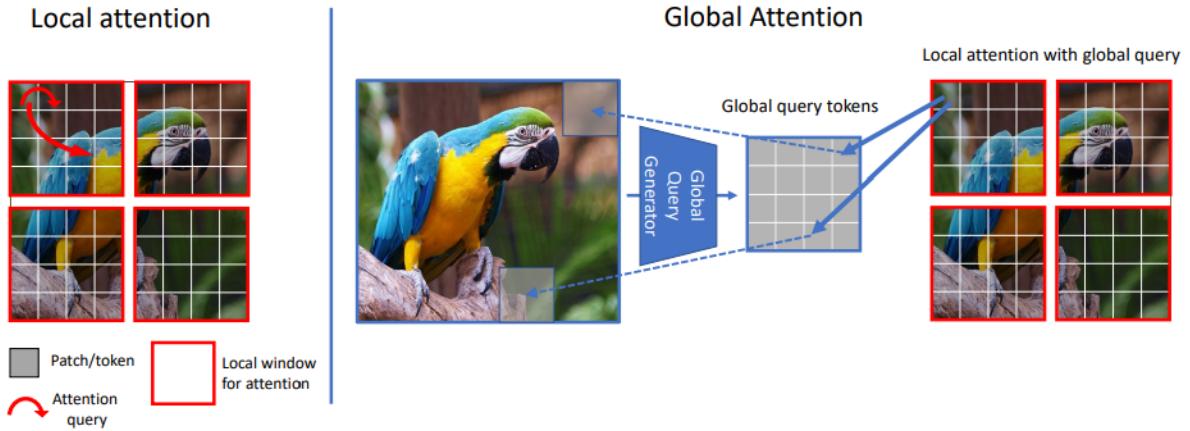


Figura 15 – Formulação da atenção no GCViT. A atenção local (esquerda) é restrita a uma janela local. Na atenção global (direita), um gerador de consultas extrai características de toda a imagem para formar tokens de consulta globais, que então interagem com os tokens de chave e valor locais, permitindo a captura de informações de longo alcance. Fonte: Hatamizadeh et al. (2022).

A arquitetura geral do GCViT é apresentada na Figura 16. A cada estágio, blocos de atenção local e global são aplicados de forma alternada. Enquanto a atenção local modela as informações de curto alcance, a atenção global utiliza os consultas pré-calculados pelo gerador de consultas para interagir com as representações locais de chave e valor dentro de cada janela. A atenção global é formulada como:

$$\text{Attention}(q_g, k, v) = \text{Softmax} \left( \frac{q_g k}{\sqrt{d}} + b \right) v , \quad (2.9)$$

onde  $q_g$  são as consultas globais,  $k$  e  $v$  são as chaves e valores locais,  $d$  é um fator de escala e  $b$  é um viés de posição relativa aprendido. Adicionalmente, o GCViT incorpora blocos Fused-MBConv modificados, tanto no gerador de consultas quanto nos módulos de *downsampling*, para introduzir um viés indutivo convolucional e modelar dependências entre canais.

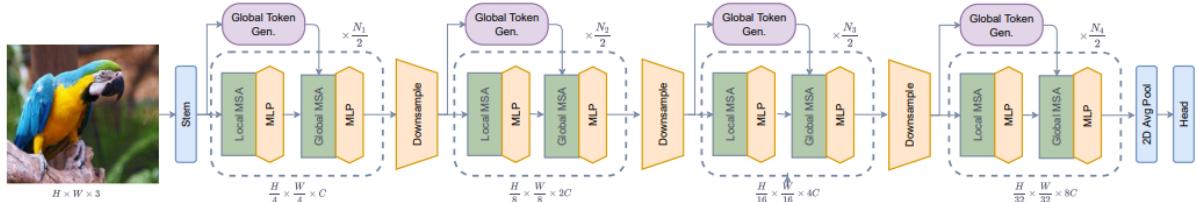


Figura 16 – Arquitetura do GCViT. A cada estágio, um gerador de tokens extrai consultas globais que interagem com as representações locais de chave e valor para capturar contexto de longo alcance. Fonte: Hatamizadeh et al. (2022).

O GCViT é apresentado em diversas configurações, que variam em capacidade. Na classificação no ImageNet-1K, as variantes GCViT-S (51 milhões de parâmetros) e GCViT-B (90 milhões de parâmetros) atingiram acuráncias top-1 de 84,3% e 85,0%, respectivamente, com resolução de  $224 \times 224$  pixels e sem pré-treinamento. Esses resultados superam modelos de tamanho comparável, como o Swin-B (83,5%) e o MaxViT-B (84,9%).

## 2.4 Funções de Perda

A função de perda é um componente essencial no treinamento de modelos, pois guia o ajuste dos pesos da rede neural ao quantificar a diferença entre as previsões e os rótulos reais. Neste trabalho, foram utilizadas duas funções de perda com o objetivo de compará-las: a entropia cruzada (ou *cross-entropy*, do inglês) e a *Conditional Ordinal Regression for Neural Networks* (CORN).

### 2.4.1 Entropia Cruzada

A entropia cruzada é uma opção comum para problemas de classificação, pois mede o quanto bem as previsões do modelo se alinham com os rótulos reais. Ela é definida como:

$$J(\Theta) = -\frac{1}{m} \sum_{i=1}^m \sum_{k=1}^K y_k^{(i)} \log(\hat{p}_k^{(i)}), \quad (2.10)$$

onde  $y_k^{(i)}$  é a probabilidade real da classe  $k$  para o exemplo  $i$ .

Ao penalizar mais fortemente casos em que o modelo apresenta baixa confiança na classe correta, a entropia cruzada, de modo geral, contribui para aumentar a precisão do modelo para tarefas de classificação. No entanto, ela não leva em consideração a natureza ordinal das classes, o que constitui uma limitação em problemas onde a ordem das classes é relevante, como no problema abordado neste trabalho.

### 2.4.2 Conditional Ordinal Regression for Neural Networks (CORN)

Shi, Cao e Raschka (2023) propuseram um framework de regressão ordinal para redes neurais profundas, chamado CORN, que é projetado para lidar com tarefas de classificação ordinal, mantendo a consistência ordinal entre as classes.

Dado um problema de classificação com  $K$  classes e um conjunto de treino  $D = \{(x^{[i]}, y^{[i]})\}_{i=1}^N$ , onde  $x^{[i]}$  é a entrada e  $y^{[i]}$  é o rótulo ordinal, o CORN divide o problema de classificação ordinal em  $K - 1$  tarefas de classificação binária associadas às classes  $r_1, r_2, \dots, r_K$ , onde  $y_k^{[i]} \in \{0,1\}$  indica se o exemplo  $y^{[i]}$  excede a classe  $r_k$  ou não (Figura 17).

A saída da  $k$ -ésima tarefa binária  $f_k(x^{[i]})$  representa a probabilidade condicional de que o exemplo  $x^{[i]}$  exceda a classe  $r_k$ , e é calculada como:

$$f_k(x^{[i]}) = \hat{P}(y^{[i]} > r_k | y^{[i]} > r_{k-1}), \quad (2.11)$$

onde os eventos estão aninhados:  $\{y^{[i]} > r_k\} \subseteq \{y^{[i]} > r_{k-1}\}$ .

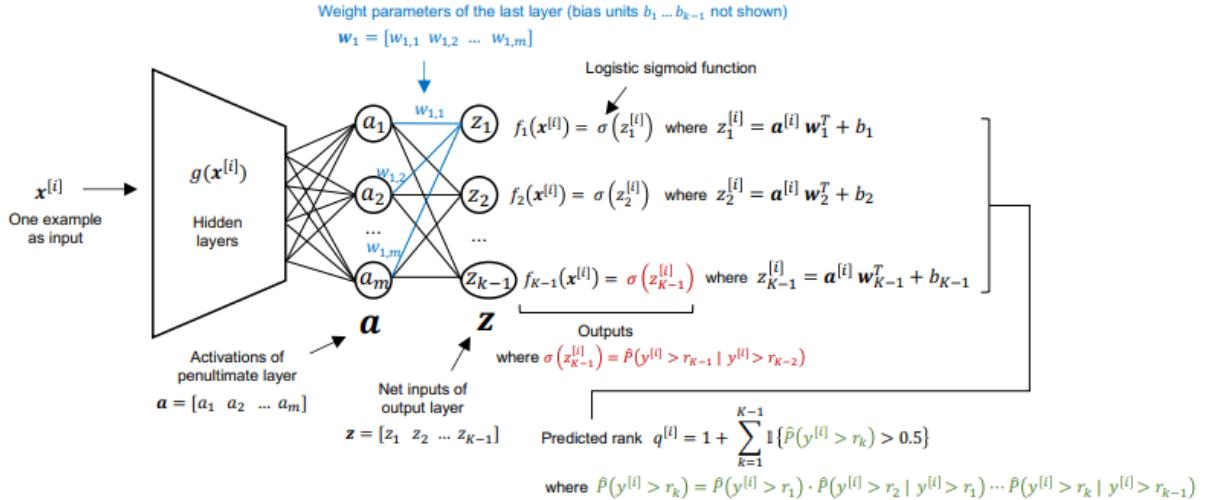


Figura 17 – Arquitetura do framework CORN. Fonte: Shi, Cao e Raschka (2023).

Com o objetivo de estimar  $f_1(x^{[i]})$  e as probabilidades condicionais  $f_2(x^{[i]}), \dots, f_{K-1}(x^{[i]})$ , o modelo CORN utiliza uma rede neural com  $K - 1$  saídas, onde cada saída é treinada para prever a probabilidade de que o rótulo ordinal exceda a classe correspondente. Para isso, são construídos subconjuntos de treino condicionais da seguinte maneira:

$$\begin{aligned} S_1 &: \text{todo } \{(x^{[i]}, y^{[i]})\}, \text{ para } i \in \{1, \dots, N\}, \\ S_2 &: \{(x^{[i]}, y^{[i]}) | y^{[i]} > r_1\}, \\ &\dots \\ S_{K-1} &: \{(x^{[i]}, y^{[i]}) | y^{[i]} > r_{k-2}\}, \end{aligned} \quad (2.12)$$

onde  $N = |S_1| \geq |S_2| \geq |S_3| \geq \dots \geq |S_{K-1}|$ , e  $|S_k|$  é o número de exemplos no subconjunto  $S_k$ .

Para treinar o modelo CORN, seja  $f_j(x^{[i]})$  o valor predito pela rede neural para o  $j$ -ésimo nó da camada de saída, a função de perda a ser minimizada é definida como:

$$\begin{aligned} L(X, y) = -\frac{1}{\sum_{j=1}^{K-1} |S_j|} \sum_{j=1}^{K-1} \sum_{i=1}^{|S_j|} & [\log(f_j(x^{[i]})) \cdot \mathbb{I}(y^{[i]} > r_j) \\ & + \log(1 - f_j(x^{[i]})) \cdot \mathbb{I}(y^{[i]} \leq r_j)], \end{aligned} \quad (2.13)$$

onde  $\mathbb{I}(\cdot)$  é a função indicadora, que retorna 1 se a condição for verdadeira e 0 caso contrário. Essa função de perda penaliza as previsões incorretas de forma proporcional à distância ordinal entre as classes, permitindo que o modelo aprenda a estrutura ordinal dos rótulos. Por fim, para obter o índice da classe predita  $q$  do  $i$ -ésimo exemplo, basta calcular:

$$q^{[i]} = 1 + \sum_{j=1}^{K-1} \mathbb{I}(\hat{P}(y^{[i]} > r_j) > 0.5), \quad (2.14)$$

onde a classe predita será  $r_{q^{[i]}}$ .

## 2.5 Avaliação e métricas de desempenho

Para avaliar o desempenho dos modelos na tarefa de classificação da severidade da OA de joelho, foram empregadas métricas amplamente utilizadas, como acurácia, precisão, revocação, F1-score e *Quadratic Weighted Kappa* (QWK). A matriz de confusão foi utilizada para visualizar a distribuição das previsões corretas e incorretas entre as diferentes classes, enquanto a métrica AUC-ROC (Área Sob a Curva da Característica de Operação do Receptor) avaliou a capacidade do modelo de distinguir uma classe em relação às demais. Essas métricas fornecem uma visão abrangente do desempenho dos modelos. Para o cálculo delas, foram adotados os seguintes acrônimos nas respectivas fórmulas:

- $TP$  é o número de verdadeiros positivos,
- $TN$  é o número de verdadeiros negativos,
- $FP$  é o número de falsos positivos,
- $FN$  é o número de falsos negativos.

### 2.5.1 Acurácia

A acurácia mede a proporção de previsões corretas em relação ao total de exemplos. Ela pode ser calculada pela fórmula:

$$\text{Acurácia} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.15)$$

### 2.5.2 Precisão

A precisão indica a proporção de exemplos classificados como positivos que realmente são positivos. Ela é calculada pela fórmula:

$$\text{Precisão} = \frac{TP}{TP + FP} \quad (2.16)$$

### 2.5.3 Revocação

A revocação (ou *recall*, do inglês) mede a capacidade do modelo de identificar corretamente todos os exemplos positivos. É definida como:

$$\text{Revocação} = \frac{TP}{TP + FN} \quad (2.17)$$

### 2.5.4 F1-Score

O F1-score é a média harmônica entre a precisão e a revocação, e é uma métrica útil quando busca-se um equilíbrio entre os dois. A fórmula do F1-score é:

$$F1 = 2 \cdot \frac{\text{Precisão} \cdot \text{Revocação}}{\text{Precisão} + \text{Revocação}} \quad (2.18)$$

### 2.5.5 Quadratic Weighted Kappa (QWK)

O QWK é uma métrica que avalia a concordância entre as previsões do modelo e os rótulos reais, levando em consideração a característica ordinal das classes. É especialmente útil para este estudo devido à natureza ordinal das classes de severidade da OA de joelho, onde erros maiores são mais penalizados do que erros menores. O QWK é calculado pela seguinte fórmula:

$$QWK = 1 - \frac{\sum_{i,j} w_{ij} O_{ij}}{\sum_{i,j} w_{ij} E_{ij}}, \quad (2.19)$$

onde  $w_{ij}$  é a matriz de pesos que penaliza os erros de classificação,  $O_{ij}$  é a matriz de confusão observada e  $E_{ij}$  é a matriz de confusão esperada. O QWK varia entre -1 e 1, onde 1 indica concordância perfeita, 0 indica concordância aleatória e valores negativos indicam discordância.

### 2.5.6 Matriz de Confusão

A matriz de confusão é uma ferramenta para visualizar o desempenho do modelo de classificação, detalhando as previsões corretas e incorretas em cada classe. Ela apresenta os valores de  $TP$ ,  $TN$ ,  $FP$  e  $FN$  de forma estruturada, permitindo avaliar o desempenho em classes específicas.

	Previsto Positivo	Previsto Negativo
Verdadeiro Positivo	$TP$	$FN$
Verdadeiro Negativo	$FP$	$TN$

### 2.5.7 AUC-ROC

A métrica AUC-ROC (Área Sob a Curva da Característica de Operação do Receptor) é bastante útil, pois mede a capacidade do modelo de separar as classes positivas e negativas. A curva ROC é um gráfico que exibe a taxa de verdadeiros positivos (sensibilidade) em função da taxa de falsos positivos. A AUC, por sua vez, quantifica a área sob essa curva, variando de 0 a 1, onde 0,5 representa um modelo aleatório e 1 representa um modelo perfeito. A AUC-ROC é calculada pela seguinte integral:

$$\text{AUC-ROC} = \int_0^1 \text{TPR}(FPR) \, dFPR, \quad (2.20)$$

onde  $\text{TPR}$  é a taxa de verdadeiros positivos e  $FPR$  é a taxa de falsos positivos.

### 2.5.8 Eficiência computacional

Além da performance em termos de métricas relacionadas à classificação, a eficiência computacional constitui um aspecto fundamental na avaliação de modelos de aprendizado profundo, especialmente em contextos com restrições de tempo ou recursos computacionais. Essa métrica torna-se ainda mais relevante quando se considera a aplicabilidade clínica dos modelos, onde a rapidez na inferência pode ser crucial para a tomada de decisão em tempo real.

Para mensurar a eficiência computacional, foram considerados três aspectos principais: o tempo de treinamento, o tempo de inferência e a quantidade de operações computacionais realizadas por cada modelo. Os tempos de treinamento e inferência foram calculados por meio da diferença entre os instantes de término e início de cada processo:

$$\text{Tempo Total} = \text{Tempo Final} - \text{Tempo Inicial}. \quad (2.21)$$

A terceira métrica adotada foi a quantidade estimada de operações de ponto flutuante, conhecida como *Floating Point Operations* (FLOPs), uma medida amplamente utilizada para quantificar o custo computacional associado à execução de modelos de redes neurais. A quantidade de FLOPs está diretamente relacionada à complexidade arquitetural do modelo, abrangendo as operações realizadas durante as fases de *forward* e *backward*, bem como o número de amostras e épocas de treinamento (LOHN; MUSSER, 2022).

### 2.5.9 Predição Conformal

A predição conformal é uma técnica estatística que fornece intervalos de confiança às previsões de qualquer modelo de aprendizado de máquina. Dada uma probabilidade de erro  $\epsilon$ , o método gera, para cada nova entrada, um conjunto de possíveis rótulos que inclui a predição  $\hat{y}$  do modelo, com garantia teórica de que o rótulo verdadeiro estará nesse conjunto com probabilidade de ao menos  $1 - \epsilon$  (ANGELOPOULOS; BATES, 2021).

Considere um modelo classificador  $\hat{f}$  e um conjunto de imagens classificadas em uma das  $K$  classes possíveis. Para cada imagem  $x$ , o modelo atribui uma distribuição de probabilidades  $\hat{f}(x) \in [0, 1]^K$  sobre as classes, geralmente obtida por meio da função *softmax*. Com base nessas probabilidades, utiliza-se um conjunto de calibração para então encontrar o conjunto de predição. Em resumo, a predição conformal é realizada da seguinte forma:

1. Para cada par de imagem  $(x, y)$  do conjunto de calibração, calcula-se a pontuação de conformidade  $s(x, y)$ :

$$s(x, y) = \sum_{j=1}^k \hat{f}(x)_{\pi_j(x)}, \text{ onde } y = \pi_k(x) \quad (2.22)$$

e  $\pi(x)$  é uma permutação dos rótulos de classe  $\{1, \dots, K\}$ , ordenada de acordo com a probabilidade atribuída pelo modelo, ou seja,  $\hat{f}(x)_{\pi_1(x)} \geq \hat{f}(x)_{\pi_2(x)} \geq \dots \geq \hat{f}(x)_{\pi_k(x)}$ . Em outras palavras, as probabilidades de cada classe são somadas até que se alcance a classe correta  $y$ .

2. Define-se o limiar de confiança  $\hat{q}$  como sendo o quantil  $\lceil (n+1)(1-\epsilon) \rceil / n$  sobre  $s_1, \dots, s_n$ , onde  $\lceil \cdot \rceil$  é a função teto.
3. Para um novo par de imagem de teste  $(x_{\text{test}}, y_{\text{test}})$ , forma-se o conjunto de predição  $\{y : s(x_{\text{test}}, y_{\text{test}}) \leq \hat{q}\}$ :

$$C(x_{\text{test}}) = \{\pi_1(x), \dots, \pi_k(x)\}, \text{ onde } k = \sup \left\{ k' : \sum_{j=1}^{k'} \hat{f}(x_{\text{test}})_{\pi_j(x_{\text{test}})} < \hat{q} \right\} + 1 \quad (2.23)$$

A predição conformal tem sido aplicada em diversas áreas, incluindo ciência forense, biometria e medicina, onde o objetivo é fornecer previsões mais confiáveis sobre a saída do modelo (FONTANA; ZENI; VANTINI, 2023). Por exemplo, Pereira et al. (2020) utilizaram a predição conformal para prever o intervalo de confiança da probabilidade de que pacientes com comprometimento cognitivo leve evoluam para demência.

### 2.5.9.1 Verificação de corretude

A verificação de corretude é uma técnica para testar se a predição conformal atende às garantias teóricas de cobertura, definida pelo Teorema 1. A ideia é verificar se o conjunto de predição  $C(x)$  contém o rótulo verdadeiro  $y$  com probabilidade de pelo menos  $1 - \epsilon$ .

**Teorema 1** (*Garantia de cobertura conformal; Vovk, Gammerman e Saunders (1999)*)  
*Suponha  $(X_i, Y_i)_{i=1,\dots,n}$  e  $(X_{test}, Y_{test})$  são independentes e identicamente distribuídos (i.i.d.) e defina  $\hat{q}$  como o quantil  $\lceil(n+1)(1-\epsilon)\rceil/n$  e  $C(X_{test}) = \{y : s(X_{test}, y) \leq \hat{q}\}$ . Então, segue que:*

$$P(Y_{test} \in C(X_{test})) \geq 1 - \epsilon. \quad (2.24)$$

Para calcular a cobertura  $C$ , é necessário executar o algoritmo de predição conformal em um conjunto de teste. A cobertura é então calculada como a proporção de casos em que o rótulo verdadeiro  $Y_{test}$  está contido no conjunto de predição  $C(X_{test})$ :

$$C = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(Y_i \in C(X_i)), \quad (2.25)$$

onde  $N$  é o número de casos no conjunto de teste e  $\mathbb{I}$  é a função indicadora, que retorna 1 se a condição for verdadeira e 0 caso contrário. A cobertura deve ser comparada com o nível de confiança  $\epsilon$  para verificar se a predição conformal atende às garantias teóricas.

### 2.5.10 Método de visualização

A visualização é uma técnica importante para avaliar quais foram as regiões da imagens que ajudaram o modelo a fazer determinada previsão. O método de visualização *Gradient-weighted Class Activation Mapping* (Grad-CAM) é uma técnica usada para interpretar e visualizar as decisões feitas por RNCs e ViTs. Em tarefas de classificação, como a avaliação da severidade da OA de joelho, entender quais regiões da radiografia contribuíram para a decisão do modelo é crucial para a validação e a confiança nos resultados do modelo.

O Grad-CAM fornece mapas de ativação que mostram quais partes da imagem foram mais influentes para a predição de uma classe específica (SELVARAJU et al., 2016). Para isso, essa técnica utiliza os gradientes da saída da camada final da rede em relação às ativações das camadas intermediárias para gerar uma visualização da importância das regiões da imagem.

Primeiro, é gerado um mapa de localização a partir da rede para classificar a imagem usando a técnica do *Class Activation Mapping* (CAM). O CAM utiliza mapas de características convolucionais, que são globalmente agrupados usando a técnica de *Global Average Pooling* (GAP) e transformados linearmente para produzir uma pontuação  $y_c$  para cada classe  $c$ . Especificamente, se a penúltima camada da rede produz  $K$  mapas de características  $A_k \in \mathbb{R}^{u \times v}$ , esses mapas são agrupados espacialmente e combinados linearmente para gerar a pontuação:

$$y_c = \sum_k w_{ck} \frac{1}{Z} \sum_i \sum_j A_{k_{ij}}$$

Para produzir o mapa de localização  $L_c^{CAM}$  para a classe  $c$ , a CAM calcula a combinação linear dos mapas de características finais usando os pesos aprendidos da camada final:

$$L_c^{CAM} = \sum_k w_{ck} A_k$$

Este mapa é então normalizado para o intervalo entre 0 e 1 para fins de visualização.

Em seguida, os gradientes são então globalmente averiguados (*pooling*) para obter pesos que indicam a importância de cada canal de ativação. Esses pesos são usados para ponderar as ativações da camada convolucional final. A seguinte fórmula representa este cálculo dos pesos:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

O peso  $\alpha_k^c$  representa a linearização parcial da rede e captura a importância de  $k$  para a classe  $c$ . Por fim, o mapa de ativação é obtido ao multiplicar as ativações ponderadas pelos pesos dos gradientes. Esse mapa é então normalizado e sobreposto na imagem original para mostrar as áreas mais influentes na decisão do modelo.

A fórmula para o Grad-CAM pode ser expressa como:

$$\text{Grad-CAM} = \text{ReLU} \left( \sum_k \alpha_k^c A^k \right)$$



### 3 Trabalhos Relacionados

A OA de joelho é uma área de pesquisa ativa na medicina e na ciência da computação, especialmente com o advento de técnicas de visão computacional. Este capítulo revisa alguns trabalhos relevantes que abordam a detecção e classificação da doença, destacando as metodologias e resultados obtidos.

Em 2023, Tariq, Suhail e Nawaz (2023) apresentaram uma abordagem de classificação ordinal (5 classes) baseada em aprendizado profundo utilizando radiografias posteroanteriores de joelhos. O estudo aplicou a estratégia de aprendizado por transferência ao fazer o ajuste fino de modelos pré-treinados, como ResNet-34, VGG-19, DenseNet-121 e DenseNet-169, combinando suas saídas em um modelo de *ensemble* (Figura 18). Usando o CORN como a função de perda, os autores alcançaram uma acurácia geral de 98% e 0,99 de QWK.

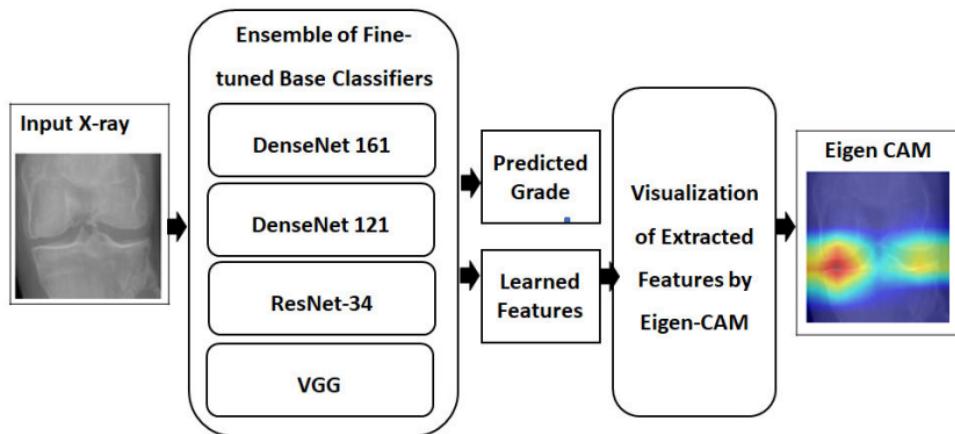


Figura 18 – Metodologia proposta por Tariq, Suhail e Nawaz (2023).

Ainda em 2023, Mohammed et al. (2023) utilizaram seis modelos pré-treinados de RNC (VGG-16, VGG-19, ResNet-101, MobileNetV2, InceptionResNetV2 e DenseNet-121) para diagnosticar a OA de joelho, considerando vários cenários de teste, como a classificação binária e o nível de severidade com três e cinco classes (Figura 19). O destaque do estudo foi a experimentação dos modelos em diferentes cenários, modelando tanto a detecção, quanto a própria classificação da OA, através do agrupamento das radiografias. O modelo ResNet-101 registrou as acurárias máximas com cinco, duas e três classes, sendo 69%, 83% e 89%, respectivamente.

Partindo para a introdução de um modelo customizado, os autores brasileiros Domingues et al. (2023) propuseram um modelo de RNC baseado na arquitetura DenseNet-161, treinado com um conjunto de radiografias obtidas do Estudo Longitudinal de Saúde do Adulto Musculoesquelético (ELSA-Brasil Musculoesquelético), para a classificação

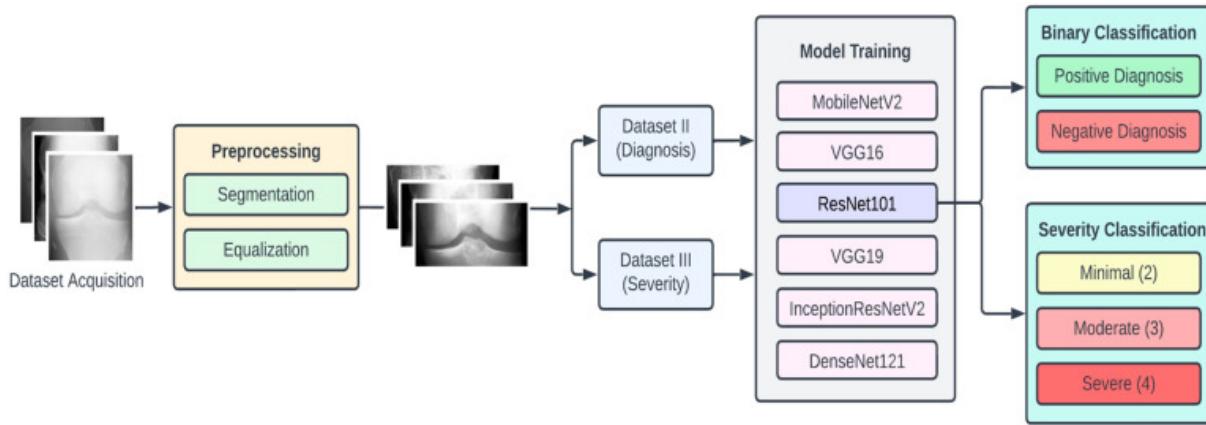


Figura 19 – Metodologia proposta por Mohammed et al. (2023).

binária automática da OA de joelho (Figura 20). Eles aplicaram diversas técnicas de pré-processamento, como rotação, desfoco gaussiano e inversão horizontal, e alcançaram uma AUC de 0,866 (IC 95%: 0,842-0,882), considerando uma média entre os subconjuntos de treino e teste. O modelo também pode ser calibrado por meio do ajuste de limiares para alcançar uma acurácia máxima de 90,7% e uma sensibilidade de 93,8%.

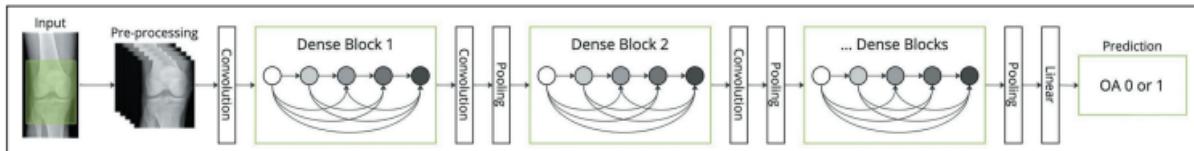


Figura 20 – Metodologia proposta por Domingues et al. (2023).

Cueva et al. (2022) desenvolveram um sistema de diagnóstico por computação assistida (CAD) utilizando a técnica de ajuste fino do modelo ResNet-34 para detectar OA nos dois joelhos simultaneamente (Figura 21). Os autores resolveram o problema de desequilíbrio do conjunto de dados por meio de técnicas de *oversampling* e *data augmentation*, como rotação aleatória e variação de cor. O modelo alcançou uma acurácia média de 61,71% em múltiplas classes, com melhor desempenho para as classes KL-0, KL-3 e KL-4 em comparação com KL-1 e KL-2 devido às sutis diferenças nos estágios intermediários.

Utilizando uma outra abordagem, Yeoh et al. (2023) investigaram o uso de redes neurais convolucionais 3D para a detecção binária de OA de joelho a partir de imagens de ressonância magnética 3D. O estudo também utilizou transferência de aprendizado, transformando pesos de modelos pré-treinados em 2D para 3D. A abordagem permitiu capturar informações espaciais nas três dimensões, resultando em uma acurácia de 87,5% e um F1-score de 0,871 para o melhor modelo, o ResNet-34.

Com a introdução dos ViTs, novas possibilidades surgiram para trabalhar o mesmo problema, oferecendo uma alternativa às RNCs, por vezes superando-as em tarefas de classificação de imagens. Em 2023, Sekhri et al. (2023) introduziram uma abordagem

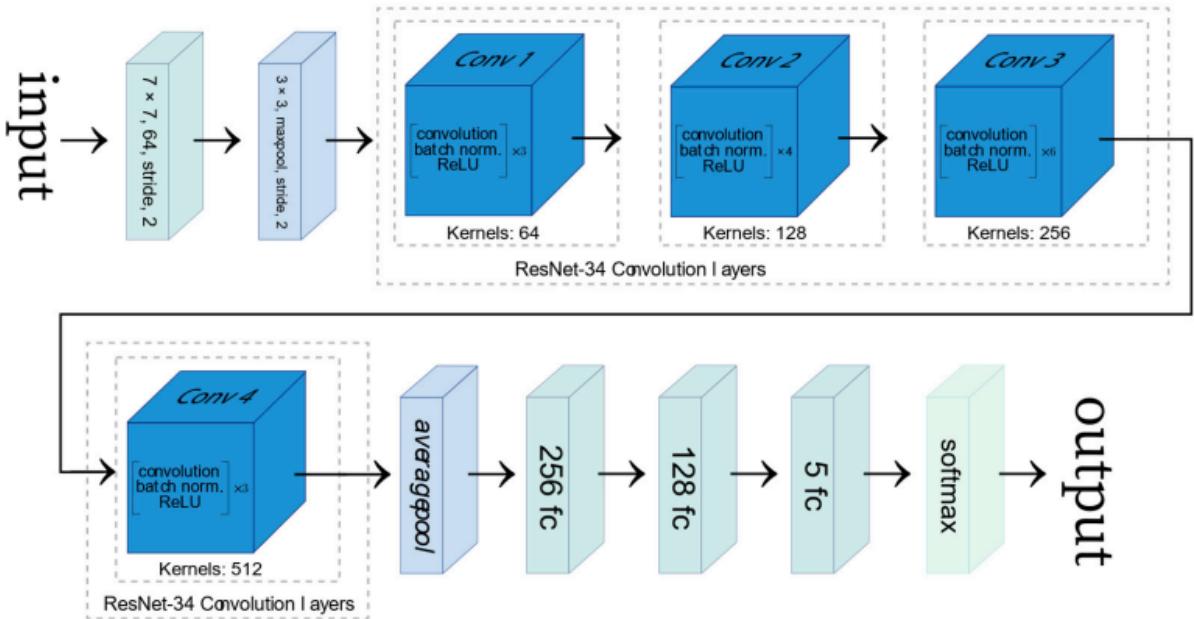


Figura 21 – Metodologia proposta por Cueva et al. (2022).

utilizando o Swin Transformer para previsão da severidade da OA de joelho. Para lidar com a alta similaridade entre os graus adjacentes da escala KL, eles implementaram uma arquitetura de múltiplas previsões composta por cinco redes perceptron multicamadas (MLP), cada uma dedicada a prever um grau específico de KL. Além disso, para reduzir o desvio de dados entre os conjuntos de dados (OAI e MOST), congelaram as camadas MLP após o treinamento inicial em um conjunto de dados e continuaram treinando o extrator de características em outro para alinhar os espaços representacionais latentes. Essa abordagem alcançou acurácia de 70,17% e F1-score de 0,67 no conjunto de dados OAI, superando os métodos existentes do estado da arte.

Wang et al. (2024b) criaram um modelo baseado em ViT para a detecção precoce da OA de joelho, focando na distinção entre o grau KL-0 e KL-2 (Figura 23). A metodologia incorporou três inovações principais:

- *Selective Shuffled Position Embedding* (SSPE): Ao fixar o posicionamento de “patches-chave” (regiões com características de grau KL) e embaralhar os demais, o modelo foi forçado a focar nas áreas críticas afetadas pela OA.
- Estratégia de troca de patches-chave: Como a técnica de aumento de dados, patches-chave de imagens candidatas foram trocados com a imagem alvo para gerar sequências de entrada diversas.
- Função de perda híbrida: Uma combinação de *Label Smoothing Cross-Entropy* (LSCE) para sequências mistas de grau KL e *cross-entropy* (CE) para sequências completas de grau KL foi otimizada para melhorar a generalização do modelo.

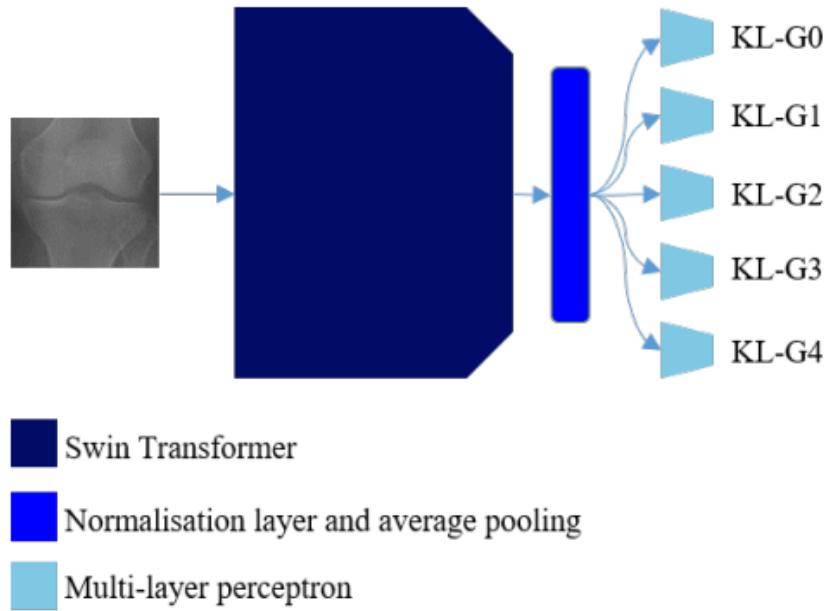


Figura 22 – Metodologia proposta por Sekhri et al. (2023).

Essas estratégias resultaram em uma melhoria notável no desempenho de classificação, com o modelo alcançando uma acurácia de 89,80%.

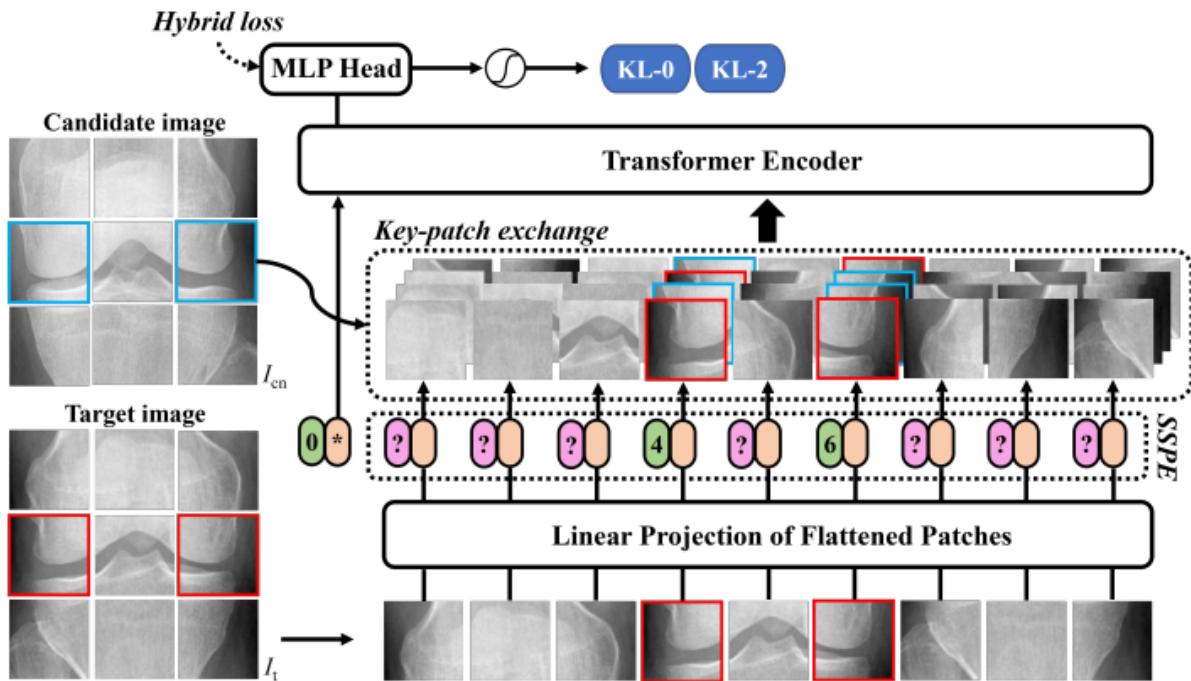


Figura 23 – Metodologia proposta por Wang et al. (2024b).

Seguindo uma linha semelhante a este estudo, Apon et al. (2024) conduziram uma análise comparativa entre modelos ViT pré-existentes (DaViT, GCViT, MaxViT) e RNCs tradicionais (Figura 24). Eles destacaram as forças arquitetônicas do DaViT com auto-atenção dupla, do GCViT com auto-atenção de contexto global e do MaxViT com atenção multi-eixo. Esses modelos ViT se destacaram com as melhores métricas,

alcançando uma acurácia máxima de 66,14%, precisão de 0,703, revocação de 0,614 e AUC superior a 0,835, superando consistentemente as RNCs (com acurácia entre 55-65%).

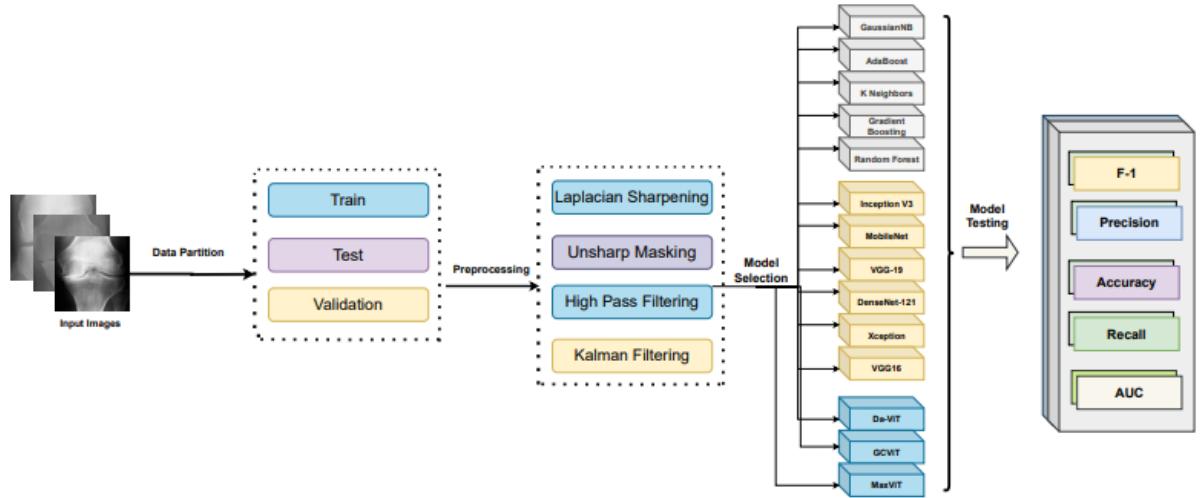


Figura 24 – Metodologia proposta por Apon et al. (2024).



## 4 Metodologia

Esta seção descreve a metodologia desenvolvida para comparar um conjunto de modelos de visão computacional, abrangendo tanto RNCs quanto ViTs, na tarefa de classificação da severidade da OA de joelho a partir de radiografias. A abordagem metodológica se baseia em quatro pilares principais: o uso de um conjunto de dados público, uma *pipeline* de pré-processamento, a aplicação de transferência de aprendizado e uma avaliação robusta que abrange não apenas o desempenho preditivo, mas também a eficiência computacional, a quantificação de incerteza e a interpretabilidade dos modelos. Uma visão geral do processo é apresentada na Figura 25, que ilustra as etapas desde a coleta de dados até a avaliação final dos modelos.

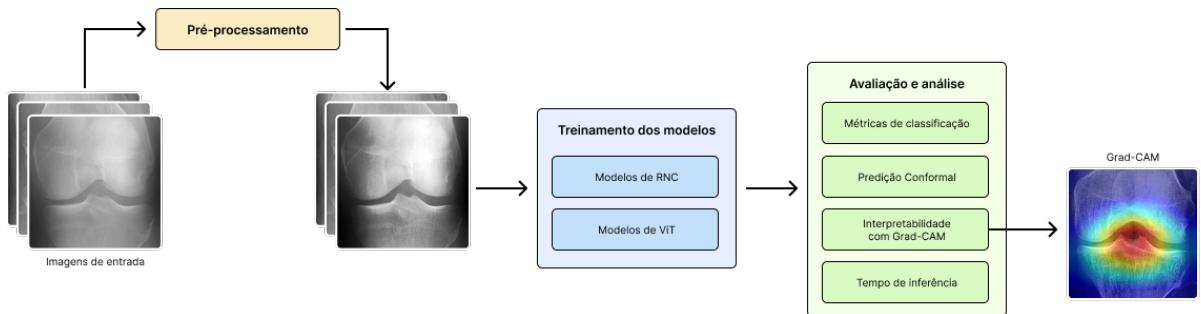


Figura 25 – Visão geral da metodologia adotada neste estudo, desde a coleta de dados até a avaliação dos modelos.

### 4.1 Coleta de dados

A base para qualquer modelo de aprendizado profundo é um conjunto de dados de alta qualidade. Para este estudo, foi utilizado o conjunto de dados de radiografias de joelho da Osteoarthritis Initiative (OAI), obtido através da plataforma Kaggle (CHEN, 2018). Este conjunto de dados é amplamente utilizado na literatura científica (TARIQ; SUHAIL; NAWAZ, 2023; MOHAMMED et al., 2023) e consiste em 9.786 radiografias classificadas por especialistas segundo a escala de Kellgren/Lawrence, cuja distribuição é detalhada na Tabela 6. A sua vasta dimensão e anotações confiáveis fornecem uma base sólida para o treinamento e a validação dos modelos.

Todas as imagens possuem resolução de  $224 \times 224$  pixels no formato PNG. Para garantir uma avaliação robusta, o conjunto de dados foi dividido em quatro subconjuntos distintos: treino (70%), validação (10%), teste (10%) e calibração (10%). Essa separação é uma prática recomendada que permite não apenas o treinamento (treino), o ajuste de hiperparâmetros (validação) e a avaliação final (teste) em dados não vistos, mas também a

Classe KL	Descrição	Total de imagens	% do total
0	saudável	3.857	40%
1	duvidoso	1.770	18%
2	mínimo	2.578	26%
3	moderado	1.286	13%
4	severo	295	3%
<b>Total</b>	-	9.786	100%

Tabela 6 – Número de radiografias por classe KL no conjunto de dados original.

quantificação de incerteza (calibração), como detalhado na subseção 2.5.9. A distribuição das classes em cada subconjunto é apresentada na Figura 26.

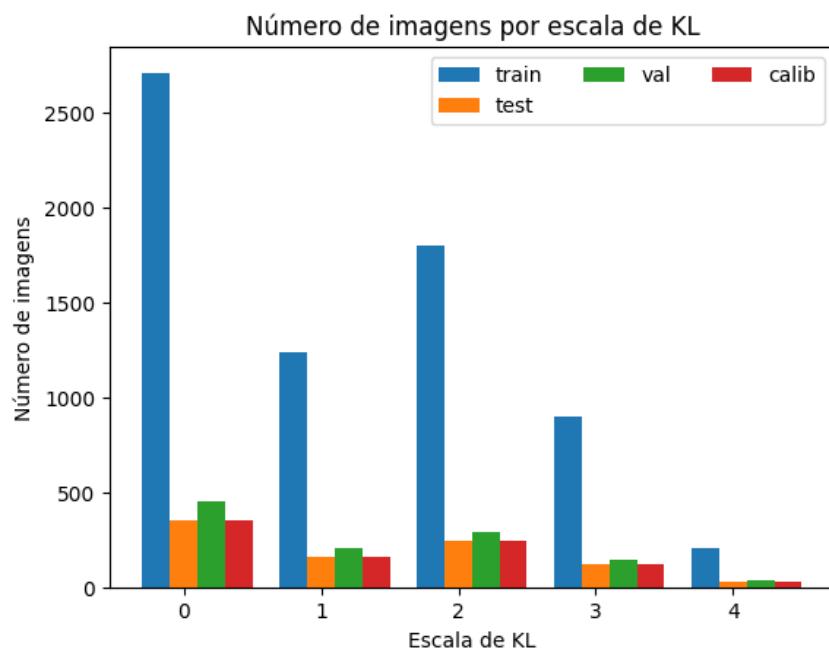


Figura 26 – Distribuição das radiografias por classe KL nos subconjuntos de treino, teste, validação e calibração.

## 4.2 Pré-processamento das imagens

Uma *pipeline* de pré-processamento foi aplicada para padronizar os dados e otimizar o aprendizado dos modelos. As técnicas foram aplicadas sequencialmente, visando aprimorar a qualidade da imagem e a robustez do treinamento.

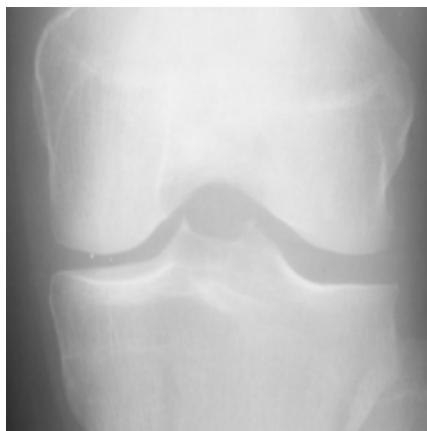
Neste estudo, o pré-processamento das radiografias foi estruturado em duas etapas: uma etapa geral e outra específica para cada modelo. O pré-processamento geral, realizado antes do treinamento, envolveu técnicas simples como a conversão para escala de cinza e a equalização de histograma, com o objetivo de melhorar a qualidade das imagens. Por sua vez, o pré-processamento específico, executado durante o treinamento, consistiu na adaptação das imagens aos requisitos de entrada de cada modelo, incluindo operações

como redimensionamento e normalização dos valores dos pixels. Adicionalmente, técnicas de aumento de dados foram empregadas para ampliar a variabilidade do conjunto de imagens e reduzir os efeitos do desbalanceamento entre as classes.

#### 4.2.1 Equalização de Histograma

Para padronizar o contraste entre as radiografias, que podem ter sido obtidas com diferentes equipamentos e configurações, a equalização de histograma foi aplicada. Essa técnica redistribui a intensidade dos pixels, realçando detalhes sutis nas estruturas ósseas e no espaço articular, que são cruciais para a identificação de características da OA.

A implementação foi realizada com a biblioteca OpenCV (ITSEEZ, 2015) do Python. A Figura 27(a) ilustra uma radiografia original do joelho, enquanto a Figura 27(b) mostra a mesma radiografia após a equalização de histograma. É possível observar que a equalização melhorou o contraste da imagem, tornando as estruturas ósseas mais visíveis. As respectivas distribuições de intensidade dos pixels antes e depois da equalização são apresentadas na Figura 28.



(a) Radiografia original.



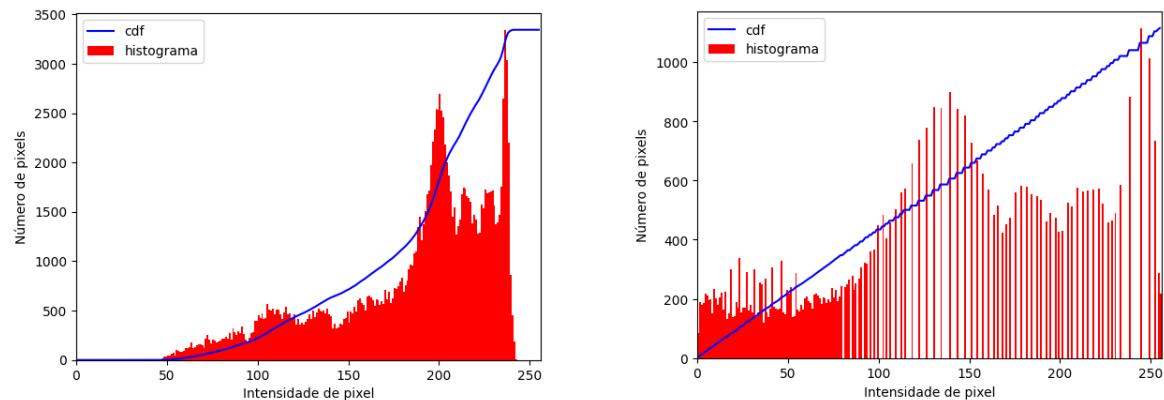
(b) Radiografia após equalização de histograma.

Figura 27 – Exemplo de equalização de histograma aplicada a uma radiografia de joelho.

#### 4.2.2 Normalização

A normalização dos valores dos pixels é um passo crítico para a eficácia da transferência de aprendizado. Os valores de cada canal de cor foram padronizados para uma média de 0 e desvio padrão de 1.

Neste estudo, a normalização foi implementada em todos os subconjuntos de dados com a biblioteca PyTorch (PASZKE et al., 2017), que aplica a normalização em cada canal (RGB), subtraindo a média e dividindo pelo desvio padrão. Para os modelos de RNC, como ResNet e VGG, utilizaram-se os valores convencionais para manter a compatibilidade com o pré-treinamento no ImageNet:



(a) Histograma da radiografia original. (b) Histograma da radiografia após a equalização.

Figura 28 – Distribuições de intensidade dos pixels antes e depois da equalização de histograma.

- Média: 0,485, 0,456 e 0,406
- Desvio padrão: 0,229, 0,224 e 0,225

Para os modelos de ViT, como o DeiT e o Swin Transformer, foram utilizados os valores específicos de normalização fornecidos por seus respectivos pacotes no Hugging Face.

#### 4.2.3 Aumento de dados

Para aumentar a diversidade do conjunto de treinamento e mitigar o risco de *overfitting*, foram aplicadas transformações geométricas aleatórias em tempo de execução. As técnicas incluíram a inversão horizontal (reflexão), com probabilidade de 50%, e rotações leves limitadas a um intervalo de -10 a 10 graus. Essa estratégia expandiu artificialmente o conjunto de dados, expondo o modelo a uma maior variedade de exemplos e melhorando sua capacidade de generalização, especialmente para as classes minoritárias.

Antes das transformações serem aplicadas, no entanto, as imagens foram redimensionadas para o tamanho esperado pelo respectivo modelo, definido como  $224 \times 224$  pixels para todos os modelos, exceto para o modelo Inception-v3, que requer imagens de  $299 \times 299$  pixels.

#### 4.2.4 Subamostragem

O desbalanceamento de classes, evidente na Tabela 6, com a classe 0 (saudável) representando 40% do total de imagens e a classe 4 (severo) apenas 3%, poderia enviesar o modelo em favor das classes majoritárias. Diante disso, a técnica foi aplicada apenas no conjunto de treinamento, de modo a não comprometer a representatividade das distribuições

nos conjuntos de validação, teste e calibração. Ela consistiu na seleção aleatória de um subconjunto das amostras das classes até um limite definido de 1.700 imagens por classe. Esse limite foi escolhido com base na classe 2 (mínima), que possui o maior número de imagens entre as classes com severidade, garantindo que todas as classes fossem representadas de forma mais equilibrada no conjunto de treinamento.

### 4.3 Treinamento dos modelos

A estratégia central de treinamento foi o uso da transferência de aprendizado a partir de modelos pré-treinados no conjunto de dados ImageNet-1K (RUSSAKOVSKY et al., 2015). As arquiteturas, obtidas de repositórios como PyTorch e Hugging Face (Tabela 7), tiveram sua camada final de classificação substituída e adaptada ao número de classes esperado. A natureza do problema consiste em  $K = 5$  classes, correspondendo às categorias de severidade da OA de joelho, mas para a função de perda CORN o número de classes considerado foi  $K - 1$ .

Adotou-se também a abordagem de ajuste fino, na qual todos os pesos da rede foram liberados para treinamento, permitindo que o modelo se adaptasse completamente às especificidades das imagens radiográficas.

Modelo	Fonte	Parâmetros (M)	FLOPs (GMac)
ResNet-34	PyTorch	21,29	3,68
ResNet-50	PyTorch	23,52	4,13
ResNet-101	PyTorch	42,51	7,86
VGG-16	PyTorch	138,36	19,63
VGG-19	PyTorch	139,64	19,69
DenseNet-121	PyTorch	6,96	2,90
DenseNet-169	PyTorch	12,49	3,44
Inception-v3	PyTorch	25,12	2,85
DeiT-Distilled-Base	Hugging Face	85,80	16,95
DaViT-Base	PyTorch	86,93	57,00
MaxViT-Tiny	PyTorch	30,41	5,48
GCViT-Base	PyTorch	89,30	13,89
Swin-Base	PyTorch	86,75	10,55

Tabela 7 – Lista dos modelos utilizados neste estudo, com a fonte, o número de parâmetros e FLOPs estimados.

A estimativa de FLOPs foi realizada com o auxílio da biblioteca *FLOPs Counter PyTorch* (SOVRASOV, 2018-2024), que permite a análise do custo computacional por meio da instrumentação do modelo em PyTorch. Essa análise visa fornecer uma perspectiva complementar à avaliação de desempenho, destacando modelos que, além de eficazes, também são eficientes em termos de recursos computacionais, o que é especialmente

relevante para implementação em ambientes com capacidade limitada, como dispositivos embarcados ou sistemas hospitalares com restrições de hardware.

### 4.3.1 Hiperparâmetros

O treinamento foi realizado com os seguintes hiperparâmetros: otimizador Adam com taxa de aprendizado inicial de  $1 \times 10^{-4}$ , tamanho do lote de 28 imagens, e um total de 60 épocas. Um agendador reduziu a taxa de aprendizado por um fator de 10 a cada 5 épocas para refinar o ajuste dos pesos. Para evitar o *overfitting*, foi implementado um mecanismo de parada antecipada que monitorava a perda no conjunto de validação, com uma paciência de 5 épocas.

Reconhecendo a natureza inherentemente ordinal da escala KL, uma limitação frequentemente negligenciada em estudos de classificação, este trabalho adotou não apenas a função de perda entropia cruzada como linha de base, mas também a função CORN. Esta última reformula o problema de  $K$  classes em  $K - 1$  tarefas de classificação binária (por exemplo, “a classe é maior do que 0?”, “a classe é maior do que 1”, etc.), penalizando erros de forma proporcional à sua distância ordinal, o que é conceitualmente mais adequado para a classificação de severidade. Essa escolha metodológica permitiu uma avaliação mais justa e clinicamente relevante do erro do modelo.

O modelo com a melhor acurácia de validação foi salvo para a avaliação final. Por sua vez, ela foi conduzida no conjunto de teste, produzindo o relatório de métricas de classificação descrita na seção 2.5, além de gerar as matrizes de confusão e as curvas AUC-ROC para cada modelo. A complexidade computacional de cada arquitetura, em termos de FLOPs e quantidade de parâmetros, foi estimada com auxílio da biblioteca *ptflops* (SOVRASOV, 2018-2024). Todos os resultados, incluindo métricas, tempos de execução e medidas de complexidade, foram armazenados em formato JSON para análise posterior.

### 4.3.2 Ambiente de execução

Todos os modelos foram executados na plataforma Google Colab, utilizando uma NVIDIA T4 GPU com 16 GB de memória, adequada para a tarefa de ajuste fino em modelos pequenos. A escolha dessa plataforma foi motivada pela sua acessibilidade e capacidade de fornecer recursos computacionais adequados a um custo reduzido.

## 4.4 Avaliação e Análise Complementar

Para ir além das métricas de desempenho tradicionais e abordar a necessidade crítica de confiança e transparência em sistemas de IA para medicina, a metodologia de

avaliação foi enriquecida com três análises complementares: quantificação da incerteza das previsões, avaliação do tempo de inferência e interpretação das decisões dos modelos. Essas análises foram fundamentais para entender não apenas o desempenho preditivo, mas também a aplicabilidade prática em cenários clínicos, diferenciando este estudo de outros trabalhos semelhantes.

#### 4.4.1 Predição Conformal

Para quantificar a incerteza, foi aplicada a predição conformal. Em vez de uma única classe, essa técnica produz um conjunto de predição (por exemplo,  $\{1, 2\}$ ) que é garantido conter a classe verdadeira com uma probabilidade especificada (neste caso, 95%). Isso fornece uma medida de confiabilidade essencial para aplicações clínicas.

O conjunto de calibração foi usado para determinar os limiares de confiança para cada modelo. Para aqueles treinados com a entropia cruzada, o limiar de confiança  $\hat{q}$  foi calculado seguindo a abordagem descrita na subseção 2.5.9, onde os *scores* foram obtidos a partir das previsões do modelo.

No entanto, para modelos treinados com o CORN foi adotada uma abordagem alternativa, tratando a saída da última camada como um conjunto de tarefas binárias ordenadas, que é exatamente como o CORN opera. Assim, para um problema com  $K$  classes KL, a predição conformal sobre o CORN foi aplicada da seguinte forma:

- Para cada exemplo do conjunto de calibração e cada  $k \in \{0, \dots, K - 2\}$  com probabilidade prevista  $p_k$ :
  - Se  $y_{\text{true}} > k$ , score =  $1 - p_k$ .
  - Se  $y_{\text{true}} \leq k$ , score =  $p_k$ .
- O limiar de confiança  $\hat{q}_k$  foi calculado da mesma forma que na abordagem tradicional. Durante a inferência, para cada  $k$ , a classe  $k + 1$  foi adicionada ao intervalo de predição se  $1 - p_k \leq \hat{q}_k$ . Por exemplo,  $\hat{y} \in [0, 2]$  se apenas  $k_0$  e  $k_1$  estiverem acima do limiar de confiança.

#### 4.4.2 Análise do Tempo de Inferência

A viabilidade de um modelo em um ambiente clínico depende não só da sua acurácia, mas também da sua velocidade. Para medir a eficiência prática, o tempo de inferência de cada modelo foi cronometrado. O processo foi repetido 50 vezes para cada arquitetura, e a média foi registrada, fornecendo uma métrica robusta para comparar a velocidade de predição em um cenário com hardware realista.

#### 4.4.3 Análise de Interpretabilidade com Grad-CAM

Para promover a transparência e a confiança nos modelos, a técnica de visualização Grad-CAM foi utilizada. Ela gera mapas de calor que sobrepõem a radiografia, indicando quais regiões foram mais importantes para a decisão do modelo. Essas visualizações foram geradas para todos os modelos para permitir uma análise qualitativa, verificando se as arquiteturas estavam focando em áreas clinicamente relevantes (como o espaço articular) e comparando as diferenças entre os modelos RNC e ViT.

A implementação foi realizada com a biblioteca *pytorch-grad-cam* (GILDENBLAT; CONTRIBUTORS, 2021), que facilita a aplicação do Grad-CAM em modelos treinados com PyTorch. As camadas escolhidas para a geração dos mapas de calor variaram de acordo com a arquitetura do modelo, sendo as últimas camadas convolucionais para os RNCs e a última camada de atenção para os ViTs. A Tabela 8 resume as camadas selecionadas para cada arquitetura.

<b>Arquitetura</b>	<b>Camada escolhida</b>
ResNet	model.layer4[-1]
VGG	model.features[-1]
DenseNet	model.features[-1]
Inception-v3	model.Mixed_7c
DeiT-Distilled-Base	model.model.deit.encoder.layer[-1].layernorm_before
DaViT-Base	model.stages[3].blocks[0][1].norm1
MaxViT-Tiny	model.blocks[-1].layers[-1].layers[-1].attn_layer[0]
GCViT-Base	model.stages[-1].blocks[-1].norm2
Swin-Base	model.features[-1][-1].norm1

Tabela 8 – Lista das camadas escolhidas para a geração dos mapas de calor Grad-CAM.

# 5 Resultados

Este capítulo apresenta os resultados obtidos e discute as implicações do estudo comparativo. Os resultados são apresentados em tabelas e gráficos, seguidos de uma análise detalhada do que foi observado.

## 5.1 Métricas Gerais de Desempenho

Os resultados das métricas gerais de desempenho dos modelos de RNC e ViT são apresentados na Tabela 9. Em suporte às hipóteses 1 (desempenho base) e 2 (comparação entre arquiteturas), observa-se que os modelos superaram significativamente o desempenho aleatório, onde a acurácia geral variou de 0,6894 (DeiT-Distilled-B) a 0,7885 (DenseNet-169), com o modelo DenseNet-169 alcançando a maior acurácia com uso da entropia cruzada. No entanto, considerando a característica ordinal da classificação, a métrica QWK oferece uma visão mais precisa do desempenho dos modelos. O modelo DenseNet-121 obteve o melhor QWK de 0,8878, seguido pelo GCViT-B com QWK de 0,8832, ambos com uso da entropia cruzada. Como esperado, esses resultados sugerem que os modelos da família DenseNet foram particularmente úteis para capturar as características visuais da OA de joelho, principalmente devido à sua arquitetura densa que permite que as camadas mais profundas acessem diretamente as características de baixo nível, sem precisar reaprendê-las (HUANG et al., 2017).

A avaliação comparativa das funções de perda testou diretamente a hipótese 3 (consistência ordinal) e revelou uma clara compensação entre a acurácia categórica e a correção ordinal. Na maioria dos cenários, os modelos treinados com a entropia cruzada apresentaram um desempenho superior em termos de acurácia, com uma média de 1,58% a mais em relação aos seus equivalentes treinados com a função CORN. Isso sugere que a entropia cruzada é mais eficaz para maximizar o número de classificações exatamente corretas.

Em contrapartida, a função CORN demonstrou sua superioridade em métricas que avaliam a natureza ordinal do problema. Observou-se um aumento médio de 1,06% no QWK, como exemplificado pelo modelo Inception-v3, cujo QWK aumentou de 0,8571 para 0,8813. Adicionalmente, houve uma redução de 0,89% no Erro Absoluto Médio (MAE), confirmando a eficácia do CORN em minimizar a magnitude dos erros de predição.

A escolha da função de perda está muito ligada ao objetivo da aplicação. Para maximizar a precisão categórica, a entropia cruzada é a abordagem preferível. No entanto, em um contexto de suporte à decisão clínica, onde um erro ordinal pequeno (por exemplo,

Modelo	Função de perda	Acurácia	QWK	MAE
ResNet-34	Entropia Cruzada	0,7258	0,8475	0,3282
ResNet-34	CORN	0,7203	0,8568	0,3194
ResNet-50	Entropia Cruzada	0,7478	0,8509	0,3095
ResNet-50	CORN	0,7379	0,8779	0,2874
ResNet-101	Entropia Cruzada	0,7445	0,8556	0,3040
ResNet-101	CORN	0,7214	0,8633	0,3128
VGG-16	Entropia Cruzada	0,7159	0,8534	0,3293
VGG-16	CORN	0,7115	0,8614	0,3216
VGG-19	Entropia Cruzada	0,7048	0,8522	0,3370
VGG-19	CORN	0,7037	0,8596	0,3260
DenseNet-121	Entropia Cruzada	0,7709	<b>0,8878</b>	0,2599
DenseNet-121	CORN	0,7357	0,8830	0,2852
DenseNet-169	Entropia Cruzada	<b>0,7885</b>	0,8811	<b>0,2522</b>
DenseNet-169	CORN	0,7324	0,8767	0,2919
Inception-v3	Entropia Cruzada	0,7247	0,8571	0,3172
Inception-v3	CORN	0,7533	0,8813	0,2742
DeiT-Distilled-B	Entropia Cruzada	0,6938	0,8321	0,3634
DeiT-Distilled-B	CORN	0,6960	0,8514	0,3381
DaViT-B	Entropia Cruzada	0,7709	0,8758	0,2687
DaViT-B	CORN	0,7357	0,8700	0,2974
MaxViT-T	Entropia Cruzada	0,7467	0,8778	0,2841
MaxViT-T	CORN	0,7456	0,8800	0,2819
GCViT-B	Entropia Cruzada	0,7555	0,8832	0,2742
GCViT-B	CORN	0,7335	0,8804	0,2896
Swin-B	Entropia Cruzada	0,7059	0,8463	0,3425
Swin-B	CORN	0,7026	0,8617	0,3293

Tabela 9 – Métricas de desempenho de cada modelo na tarefa de classificar a OA de joelho em cinco classes de severidade.

prever KL-2 para um KL-3 real) é significativamente menos crítico do que um erro grande (por exemplo, prever KL-0), a função CORN é mais adequada. Isso se deve à sua capacidade de gerar modelos que, mesmo quando erram, produzem predições mais próximas do rótulo verdadeiro, alinhando-se melhor à relevância clínica dos erros.

Adicionalmente, para avaliar a capacidade discriminativa dos modelos, foram calculadas as curvas AUC-ROC. A Tabela 10 apresenta as curvas e os valores de AUC para os principais modelos, demonstrando sua robustez na separação entre as diferentes classes KL. O melhor modelo, DenseNet-169, alcançou um AUC de 0,938, reforçando seu desempenho superior.

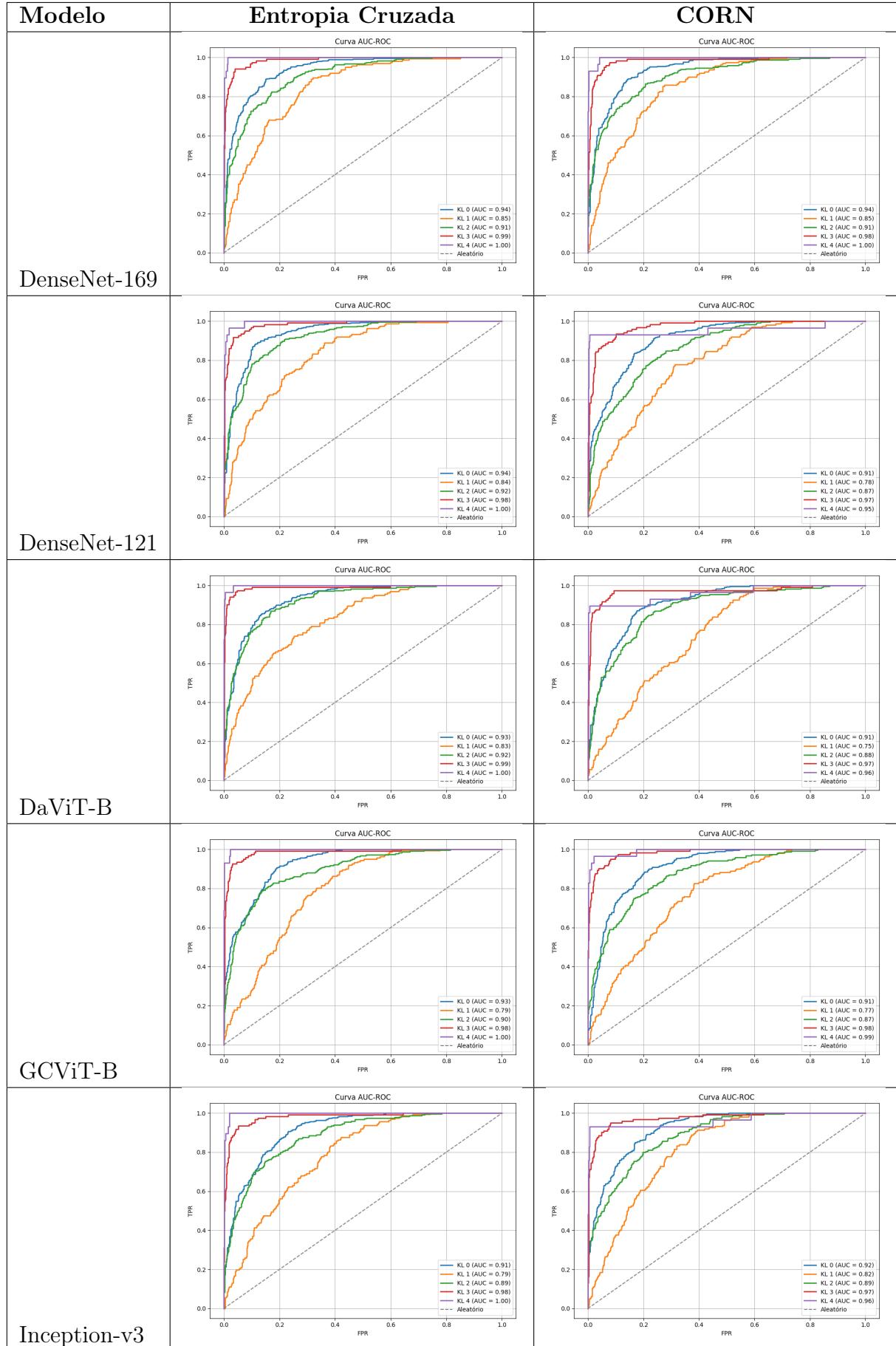


Tabela 10 – Curvas AUC-ROC para os cinco principais modelos.

## 5.2 Métricas de Desempenho por Classe

A análise detalhada dos F1-scores por classe, apresentada na Tabela 11, revela padrões de desempenho que vão além da simples identificação de um modelo superior. Os resultados expõem a complexidade intrínseca da classificação da escala KL e destacam como diferentes arquiteturas respondem a cada estágio da doença.

A tabela revela a observação central onde o desempenho dos modelos é consistentemente baixo na classe KL-1 (duvidoso). Nenhum modelo, independentemente da arquitetura ou da função de perda, conseguiu superar um F1-score de 0,6000 para esta classe, com o melhor sendo o DenseNet-169 (0,5970) e o pior sendo o VGG-19 (0,3750). Esse resultado sugere que a classe KL-1 é inherentemente ambígua e inconsistente, assim como diz Spector e Cooper (1993). Ela representa um estágio da OA onde os achados radiológicos, como um possível estreitamento do espaço articular ou a formação de osteófitos, são muito sutis. Como consequência, essa classe sofre com a sobreposição de características entre as classes adjacentes, tornando a classificação pelos modelos mais desafiadora. Esse padrão é visualmente confirmado nas matrizes de confusão apresentadas na Tabela 12, onde observa-se um número elevado de classificações incorretas entre as classes KL-1 e as classes adjacentes KL-0 e KL-2.

Em forte contraste com a classe KL-1, as classes nos extremos da escala, KL-0 (saudável) e KL-4 (severo), apresentam F1-scores consistentemente altos na maioria dos modelos. Para a classe KL-0, modelos como DenseNet-121 (0,8454) e GCViT-B (0,8409) demonstraram uma boa capacidade de identificar corretamente um joelho saudável. Já para a classe KL-4, os resultados são ainda mais expressivos, com modelos como DaViT-B e MaxViT-T alcançando F1-scores muito elevados, ambos com 0,9310.

As classes KL-2 (mínimo) e KL-3 (moderado) representam estágios onde a doença já está presente e o desempenho dos modelos foi mais equilibrado. Modelos como DaViT-B (0,9212 para KL-3) e DenseNet-169 (0,9016 para KL-3) mostraram uma capacidade notável de distinguir os estágios intermediários da doença.

Essa análise por classe não apenas valida a decisão de realizar experimentos excluindo a classe KL-1, mas também confirma que o desempenho dos modelos está fortemente alinhado à realidade clínica da OA de joelho, onde existe uma alta certeza nos casos extremos e dificuldade na zona de transição.

## 5.3 Comparações com Trabalhos Relacionados

O desempenho do melhor modelo deste estudo (DenseNet-169) foi comparado com trabalhos recentes da literatura que abordam a mesma tarefa de classificação da severidade da OA de joelho. A Tabela 13 resume essa comparação. É importante ressaltar que uma

Modelo	Função de perda	Média Macro	KL-0	KL-1	KL-2	KL-3	KL-4
ResNet-34	Entropia Cruzada	0.7384	0.7932	0.4669	0.7187	0.8465	0.8667
ResNet-34	CORN	0.7431	0.8034	0.5117	0.6837	0.8354	0.8814
ResNet-50	Entropia Cruzada	0.7722	0.7983	0.5257	0.7364	0.8852	0.9153
ResNet-50	CORN	0.7564	0.8239	0.5158	0.7244	0.8326	0.8852
ResNet-101	Entropia Cruzada	0.7726	0.7983	0.5683	0.7277	0.8534	0.9153
ResNet-101	CORN	0.7359	0.8142	0.4932	0.6920	0.8412	0.8387
VGG-16	Entropia Cruzada	0.7276	0.8063	0.4201	0.6912	0.8537	0.8667
VGG-16	CORN	0.7384	0.7935	0.4358	0.7042	0.8583	0.9000
VGG-19	Entropia Cruzada	0.7066	0.7898	0.3750	0.7146	0.8468	0.8070
VGG-19	CORN	0.7268	0.7935	0.4216	0.6949	0.8667	0.8571
DenseNet-121	Entropia Cruzada	0.7777	<b>0.8454</b>	0.5378	0.7537	0.8807	0.8710
DenseNet-121	CORN	0.7563	0.8192	0.4890	0.7292	0.8439	0.9000
DenseNet-169	Entropia Cruzada	<b>0.8061</b>	0.8384	<b>0.5970</b>	0.7780	0.9016	0.9153
DenseNet-169	CORN	0.7583	0.8066	0.5433	0.7097	0.8608	0.8710
Inception-v3	Entropia Cruzada	0.7487	0.7959	0.4734	0.7166	0.8455	0.9123
Inception-v3	CORN	0.7811	0.8067	0.5464	0.7589	0.8780	0.9153
DeiT-Distilled-B	Entropia Cruzada	0.7206	0.7670	0.3938	0.6790	0.8631	0.9000
DeiT-Distilled-B	CORN	0.7378	0.7527	0.4230	0.7157	0.8852	0.9123
DaViT-B	Entropia Cruzada	0.7968	0.8111	0.5401	<b>0.7807</b>	<b>0.9212</b>	<b>0.9310</b>
DaViT-B	CORN	0.7622	0.7912	0.4756	0.7510	0.8807	0.9123
MaxViT-T	Entropia Cruzada	0.7649	0.8329	0.4986	0.7143	0.8787	0.9000
MaxViT-T	CORN	0.7728	0.8333	0.4945	0.7100	0.8952	<b>0.9310</b>
GCViT-B	Entropia Cruzada	0.7720	0.8409	0.5128	0.7136	0.8926	0.9000
GCViT-B	CORN	0.7459	0.8093	0.4501	0.7463	0.8760	0.8475
Swin-B	Entropia Cruzada	0.7237	0.7944	0.4037	0.6795	0.8595	0.8814
Swin-B	CORN	0.7261	0.7994	0.4261	0.6681	0.8405	0.8966

Tabela 11 – Métrica F1-score para cada uma das cinco classes e modelo, considerando as funções de perda Entropia Cruzada e CORN.

Modelo	Entropia Cruzada	CORN																																																																								
DenseNet-169	<p>Matriz de Confusão</p> <table border="1"> <tr><td>0</td><td>298</td><td>44</td><td>19</td><td>0</td><td>0</td></tr> <tr><td>1</td><td>32</td><td>100</td><td>30</td><td>0</td><td>0</td></tr> <tr><td>2</td><td>15</td><td>28</td><td>191</td><td>10</td><td>0</td></tr> <tr><td>3</td><td>1</td><td>1</td><td>7</td><td>110</td><td>3</td></tr> <tr><td>4</td><td>0</td><td>0</td><td>0</td><td>2</td><td>27</td></tr> <tr><td>Predição</td><td>0</td><td>1</td><td>2</td><td>3</td><td>4</td></tr> </table>	0	298	44	19	0	0	1	32	100	30	0	0	2	15	28	191	10	0	3	1	1	7	110	3	4	0	0	0	2	27	Predição	0	1	2	3	4	<p>Matriz de Confusão</p> <table border="1"> <tr><td>0</td><td>269</td><td>75</td><td>7</td><td>0</td><td>0</td></tr> <tr><td>1</td><td>33</td><td>113</td><td>16</td><td>0</td><td>0</td></tr> <tr><td>2</td><td>14</td><td>65</td><td>154</td><td>11</td><td>0</td></tr> <tr><td>3</td><td>0</td><td>1</td><td>13</td><td>102</td><td>6</td></tr> <tr><td>4</td><td>0</td><td>0</td><td>0</td><td>2</td><td>27</td></tr> <tr><td>Predição</td><td>0</td><td>1</td><td>2</td><td>3</td><td>4</td></tr> </table>	0	269	75	7	0	0	1	33	113	16	0	0	2	14	65	154	11	0	3	0	1	13	102	6	4	0	0	0	2	27	Predição	0	1	2	3	4
0	298	44	19	0	0																																																																					
1	32	100	30	0	0																																																																					
2	15	28	191	10	0																																																																					
3	1	1	7	110	3																																																																					
4	0	0	0	2	27																																																																					
Predição	0	1	2	3	4																																																																					
0	269	75	7	0	0																																																																					
1	33	113	16	0	0																																																																					
2	14	65	154	11	0																																																																					
3	0	1	13	102	6																																																																					
4	0	0	0	2	27																																																																					
Predição	0	1	2	3	4																																																																					
DenseNet-121	<p>Matriz de Confusão</p> <table border="1"> <tr><td>0</td><td>298</td><td>42</td><td>11</td><td>0</td><td>0</td></tr> <tr><td>1</td><td>40</td><td>89</td><td>33</td><td>0</td><td>0</td></tr> <tr><td>2</td><td>16</td><td>37</td><td>179</td><td>12</td><td>0</td></tr> <tr><td>3</td><td>0</td><td>1</td><td>8</td><td>107</td><td>6</td></tr> <tr><td>4</td><td>0</td><td>0</td><td>0</td><td>2</td><td>27</td></tr> <tr><td>Predição</td><td>0</td><td>1</td><td>2</td><td>3</td><td>4</td></tr> </table>	0	298	42	11	0	0	1	40	89	33	0	0	2	16	37	179	12	0	3	0	1	8	107	6	4	0	0	0	2	27	Predição	0	1	2	3	4	<p>Matriz de Confusão</p> <table border="1"> <tr><td>0</td><td>281</td><td>66</td><td>4</td><td>0</td><td>0</td></tr> <tr><td>1</td><td>40</td><td>89</td><td>33</td><td>0</td><td>0</td></tr> <tr><td>2</td><td>14</td><td>46</td><td>171</td><td>13</td><td>0</td></tr> <tr><td>3</td><td>0</td><td>1</td><td>17</td><td>100</td><td>4</td></tr> <tr><td>4</td><td>0</td><td>0</td><td>0</td><td>2</td><td>27</td></tr> <tr><td>Predição</td><td>0</td><td>1</td><td>2</td><td>3</td><td>4</td></tr> </table>	0	281	66	4	0	0	1	40	89	33	0	0	2	14	46	171	13	0	3	0	1	17	100	4	4	0	0	0	2	27	Predição	0	1	2	3	4
0	298	42	11	0	0																																																																					
1	40	89	33	0	0																																																																					
2	16	37	179	12	0																																																																					
3	0	1	8	107	6																																																																					
4	0	0	0	2	27																																																																					
Predição	0	1	2	3	4																																																																					
0	281	66	4	0	0																																																																					
1	40	89	33	0	0																																																																					
2	14	46	171	13	0																																																																					
3	0	1	17	100	4																																																																					
4	0	0	0	2	27																																																																					
Predição	0	1	2	3	4																																																																					
DaViT-B	<p>Matriz de Confusão</p> <table border="1"> <tr><td>0</td><td>277</td><td>54</td><td>20</td><td>0</td><td>0</td></tr> <tr><td>1</td><td>40</td><td>91</td><td>31</td><td>0</td><td>0</td></tr> <tr><td>2</td><td>15</td><td>29</td><td>194</td><td>6</td><td>0</td></tr> <tr><td>3</td><td>0</td><td>1</td><td>8</td><td>111</td><td>2</td></tr> <tr><td>4</td><td>0</td><td>0</td><td>0</td><td>2</td><td>27</td></tr> <tr><td>Predição</td><td>0</td><td>1</td><td>2</td><td>3</td><td>4</td></tr> </table>	0	277	54	20	0	0	1	40	91	31	0	0	2	15	29	194	6	0	3	0	1	8	111	2	4	0	0	0	2	27	Predição	0	1	2	3	4	<p>Matriz de Confusão</p> <table border="1"> <tr><td>0</td><td>271</td><td>67</td><td>13</td><td>0</td><td>0</td></tr> <tr><td>1</td><td>47</td><td>83</td><td>32</td><td>0</td><td>0</td></tr> <tr><td>2</td><td>16</td><td>36</td><td>181</td><td>11</td><td>0</td></tr> <tr><td>3</td><td>0</td><td>1</td><td>12</td><td>107</td><td>2</td></tr> <tr><td>4</td><td>0</td><td>0</td><td>0</td><td>3</td><td>26</td></tr> <tr><td>Predição</td><td>0</td><td>1</td><td>2</td><td>3</td><td>4</td></tr> </table>	0	271	67	13	0	0	1	47	83	32	0	0	2	16	36	181	11	0	3	0	1	12	107	2	4	0	0	0	3	26	Predição	0	1	2	3	4
0	277	54	20	0	0																																																																					
1	40	91	31	0	0																																																																					
2	15	29	194	6	0																																																																					
3	0	1	8	111	2																																																																					
4	0	0	0	2	27																																																																					
Predição	0	1	2	3	4																																																																					
0	271	67	13	0	0																																																																					
1	47	83	32	0	0																																																																					
2	16	36	181	11	0																																																																					
3	0	1	12	107	2																																																																					
4	0	0	0	3	26																																																																					
Predição	0	1	2	3	4																																																																					
GCViT-B	<p>Matriz de Confusão</p> <table border="1"> <tr><td>0</td><td>304</td><td>41</td><td>6</td><td>0</td><td>0</td></tr> <tr><td>1</td><td>48</td><td>90</td><td>24</td><td>0</td><td>0</td></tr> <tr><td>2</td><td>20</td><td>57</td><td>157</td><td>10</td><td>0</td></tr> <tr><td>3</td><td>0</td><td>1</td><td>9</td><td>108</td><td>4</td></tr> <tr><td>4</td><td>0</td><td>0</td><td>0</td><td>2</td><td>27</td></tr> <tr><td>Predição</td><td>0</td><td>1</td><td>2</td><td>3</td><td>4</td></tr> </table>	0	304	41	6	0	0	1	48	90	24	0	0	2	20	57	157	10	0	3	0	1	9	108	4	4	0	0	0	2	27	Predição	0	1	2	3	4	<p>Matriz de Confusão</p> <table border="1"> <tr><td>0</td><td>278</td><td>67</td><td>6</td><td>0</td><td>0</td></tr> <tr><td>1</td><td>44</td><td>79</td><td>39</td><td>0</td><td>0</td></tr> <tr><td>2</td><td>14</td><td>42</td><td>178</td><td>10</td><td>0</td></tr> <tr><td>3</td><td>0</td><td>1</td><td>10</td><td>106</td><td>5</td></tr> <tr><td>4</td><td>0</td><td>0</td><td>0</td><td>4</td><td>25</td></tr> <tr><td>Predição</td><td>0</td><td>1</td><td>2</td><td>3</td><td>4</td></tr> </table>	0	278	67	6	0	0	1	44	79	39	0	0	2	14	42	178	10	0	3	0	1	10	106	5	4	0	0	0	4	25	Predição	0	1	2	3	4
0	304	41	6	0	0																																																																					
1	48	90	24	0	0																																																																					
2	20	57	157	10	0																																																																					
3	0	1	9	108	4																																																																					
4	0	0	0	2	27																																																																					
Predição	0	1	2	3	4																																																																					
0	278	67	6	0	0																																																																					
1	44	79	39	0	0																																																																					
2	14	42	178	10	0																																																																					
3	0	1	10	106	5																																																																					
4	0	0	0	4	25																																																																					
Predição	0	1	2	3	4																																																																					
Inception-v3	<p>Matriz de Confusão</p> <table border="1"> <tr><td>0</td><td>271</td><td>59</td><td>21</td><td>0</td><td>0</td></tr> <tr><td>1</td><td>44</td><td>80</td><td>36</td><td>2</td><td>0</td></tr> <tr><td>2</td><td>15</td><td>37</td><td>177</td><td>15</td><td>0</td></tr> <tr><td>3</td><td>0</td><td>0</td><td>16</td><td>104</td><td>2</td></tr> <tr><td>4</td><td>0</td><td>0</td><td>0</td><td>3</td><td>26</td></tr> <tr><td>Predição</td><td>0</td><td>1</td><td>2</td><td>3</td><td>4</td></tr> </table>	0	271	59	21	0	0	1	44	80	36	2	0	2	15	37	177	15	0	3	0	0	16	104	2	4	0	0	0	3	26	Predição	0	1	2	3	4	<p>Matriz de Confusão</p> <table border="1"> <tr><td>0</td><td>265</td><td>72</td><td>14</td><td>0</td><td>0</td></tr> <tr><td>1</td><td>31</td><td>103</td><td>28</td><td>0</td><td>0</td></tr> <tr><td>2</td><td>10</td><td>39</td><td>181</td><td>14</td><td>0</td></tr> <tr><td>3</td><td>0</td><td>1</td><td>10</td><td>108</td><td>3</td></tr> <tr><td>4</td><td>0</td><td>0</td><td>0</td><td>2</td><td>27</td></tr> <tr><td>Predição</td><td>0</td><td>1</td><td>2</td><td>3</td><td>4</td></tr> </table>	0	265	72	14	0	0	1	31	103	28	0	0	2	10	39	181	14	0	3	0	1	10	108	3	4	0	0	0	2	27	Predição	0	1	2	3	4
0	271	59	21	0	0																																																																					
1	44	80	36	2	0																																																																					
2	15	37	177	15	0																																																																					
3	0	0	16	104	2																																																																					
4	0	0	0	3	26																																																																					
Predição	0	1	2	3	4																																																																					
0	265	72	14	0	0																																																																					
1	31	103	28	0	0																																																																					
2	10	39	181	14	0																																																																					
3	0	1	10	108	3																																																																					
4	0	0	0	2	27																																																																					
Predição	0	1	2	3	4																																																																					

Tabela 12 – Matriz de confusão para os cinco principais modelos.

comparação direta entre estudos pode ser desafiadora devido a variações em protocolos experimentais, como o conjunto de dados utilizado, as técnicas de pré-processamento e o hardware utilizado. Apesar dessas variações, a análise indica que os resultados deste estudo são altamente competitivos e, em métricas cruciais, superiores à maioria dos trabalhos recentes, especialmente quando se considera uma avaliação holística do desempenho.

Referência	Classes	Acurácia	QWK	F1-score	AUC
Tariq, Suhail e Nawaz (2023)	5	98,00%	0,990	0,980	0,970
Mohammed et al. (2023)	5	69,00%	-	0,670	-
Domingues et al. (2023)	2	90,70%	-	0,553	0,866
Cueva et al. (2022)	5	61,71%	-	-	-
Yeoh et al. (2023)	2	87,50%	-	0,871	0,945
Sekhri et al. (2023)	5	70,17%	-	0,671	-
Wang et al. (2024b)	2	89,90%	-	0,877	-
Apon et al. (2024)	5	66,14%	-	0,618	0,860
Este estudo	5	78,85%	0,888	0,806	0,938

Tabela 13 – Comparação entre os resultados de diferentes estudos que abordam o problema da OA de joelho.

Ao analisar os trabalhos que consideram as cinco classes KL, observa-se que o presente estudo, com 78,85% de acurácia, supera significativamente os resultados de Mohammed et al. (2023) (69,00%), Cueva et al. (2022) (61,71%), Sekhri et al. (2023) (70,17%) e Apon et al. (2024) (66,14%). A superioridade é particularmente evidente no F1-score (0,806), que indica um melhor equilíbrio entre precisão e revocação entre todas as cinco classes em comparação com os valores reportados por Mohammed et al. (2023) (0,670) e Apon et al. (2024) (0,618). O trabalho de Tariq, Suhail e Nawaz (2023) apresenta resultados excepcionalmente elevados (98,00% de acurácia), que foram obtidos a partir de um modelo *ensemble*, além de o resultado poder ser atribuído a um subconjunto de dados menos desafiador ou protocolos de validação que podem não garantir uma separação estrita entre treino e teste.

Um diferencial significativo deste estudo é a avaliação através de métricas mais alinhadas à natureza do problema. O QWK de 0,888 demonstra não apenas que o modelo é acurado, mas que seus erros são, em sua maioria, clinicamente menos graves (próximos da classe real), uma análise de robustez que está ausente na maioria dos trabalhos comparados. Da mesma forma, o valor de AUC de 0,938 indica uma excelente capacidade de separação entre as classes, superando os valores reportados por Domingues et al. (2023) (0,866) e Apon et al. (2024) (0,860).

Portanto, embora alguns estudos possam apresentar valores de acurácia isoladamente mais altos, os resultados aqui obtidos demonstram um desempenho mais robusto, equilibrado e rigorosamente avaliado. O modelo deste estudo não apenas supera a maioria dos trabalhos em métricas padrão, mas também se destaca em métricas ordinais e de sepa-

rabilidade de classes, estabelecendo um *benchmark* sólido e confiável para a classificação da OA de joelho em cinco classes.

## 5.4 Eficiência Computacional

A análise dos tempos de treinamento e inferência, apresentada na Tabela 14, revela distinções cruciais entre os modelos e funções de perda, oferecendo um respaldo da aplicação prática, complementando as métricas de desempenho preditivo e validando a hipótese 2 (comparação entre arquiteturas).

O tempo de inferência por imagem é uma métrica crítica para a aplicação clínica, e as RNCs demonstram uma vantagem sobre os ViTs. Modelos como ResNet-34 (1,55 ms/imagem), ResNet-50 (2,87 ms/imagem) e DenseNet/121 (3,11 ms/imagem) são os mais rápidos, estabelecendo uma linha de base de alta eficiência. Suas arquiteturas são bem otimizadas para o hardware de GPU, o que permite um processamento rápido.

Em contraste, os ViTs apresentaram tempos de inferência significativamente mais altos. O mais rápido entre eles, MaxViT-T (6,92 ms/imagem), já é mais de duas vezes mais lento do que as RNCs eficientes. Essa disparidade pode ser justificada pela natureza das operações. Embora ViTs tenham alcançado complexidade computacional linear em relação ao tamanho de imagem, como citam Ding et al. (2022) e Tu et al. (2022), seus blocos de atenção ainda envolvem operações de multiplicação de matrizes em larga escala que são mais custosas do que as operações de convolução. Isso também pode ser visualizado na Tabela 7, na coluna de FLOPs (GMac).

A literatura sobre modelos como MaxViT e DaViT foca em tornar a atenção global mais eficiente e de fato o fazem, mas os resultados práticos da Tabela 14 demonstram que, na GPU utilizada, a sobrecarga computacional da auto-atenção ainda é um obstáculo significativo para a inferência quando comparada às RNCs.

## 5.5 Análise Quantitativa

A predição conformal foi aplicada para quantificar a incerteza dos modelos, gerando conjuntos de predição com um nível de confiança estatisticamente garantido de 95%. A análise dos resultados, apresentados na Tabela 15, sugere uma compensação entre a robustez da garantia de cobertura e a utilidade prática dos conjuntos de predição, com diferenças expressivas entre a abordagem padrão e a ordinal.

Como descrito na subseção 2.5.9, a cobertura mede a porcentagem de vezes que a classe verdadeira esteve contida no conjunto de predição conformal a partir das imagens de teste. Modelos treinados com entropia cruzada demonstraram uma cobertura muito alta, consistentemente alcançando valores de 100%, como é o caso do ResNet-50, VGG-16,

Modelo	Função de perda	Treinamento (min)	Inferência média/batch (ms)	Inferência média/imagem (ms)
ResNet-34	Entropia Cruzada	33,46	43,53	1,55
ResNet-34	CORN	89,29	43,71	1,56
ResNet-50	Entropia Cruzada	14,55	80,45	2,87
ResNet-50	CORN	9,87	80,83	2,89
ResNet-101	Entropia Cruzada	22,02	137,90	4,93
ResNet-101	CORN	16,94	139,76	4,99
VGG-16	Entropia Cruzada	37,70	128,08	4,57
VGG-16	CORN	28,45	126,44	4,52
VGG-19	Entropia Cruzada	39,32	153,06	5,47
VGG-19	CORN	34,68	154,96	5,53
DenseNet-121	Entropia Cruzada	12,74	87,02	3,11
DenseNet-121	CORN	<b>9,22</b>	87,26	3,12
DenseNet-169	Entropia Cruzada	15,06	108,95	3,89
DenseNet-169	CORN	16,72	108,48	3,87
Inception-v3	Entropia Cruzada	12,52	108,58	3,88
Inception-v3	CORN	14,39	108,86	3,89
DeiT-Distilled-B	Entropia Cruzada	77,31	331,55	11,84
DeiT-Distilled-B	CORN	40,20	330,48	11,80
DaViT-B	Entropia Cruzada	57,83	335,72	11,99
DaViT-B	CORN	27,08	339,09	12,11
MaxViT-T	Entropia Cruzada	28,03	193,82	6,92
MaxViT-T	CORN	29,36	194,79	6,96
GCViT-B	Entropia Cruzada	49,57	408,07	14,57
GCViT-B	CORN	34,65	409,04	14,61
Swin-B	Entropia Cruzada	46,61	344,37	12,30
Swin-B	CORN	36,02	348,65	12,45

Tabela 14 – Tempos de treinamento e inferência de cada modelo.

Modelo	Função de perda	Cobertura
ResNet-34	Entropia Cruzada	0,9989
ResNet-34	CORN	0,8822
ResNet-50	Entropia Cruzada	<b>1,0000</b>
ResNet-50	CORN	0,9361
ResNet-101	Entropia Cruzada	0,9956
ResNet-101	CORN	0,8844
VGG-16	Entropia Cruzada	<b>1,0000</b>
VGG-16	CORN	0,9273
VGG-19	Entropia Cruzada	0,9989
VGG-19	CORN	0,9262
DenseNet-121	Entropia Cruzada	<b>1,0000</b>
DenseNet-121	CORN	0,8667
DenseNet-169	Entropia Cruzada	0,9989
DenseNet-169	CORN	0,9163
Inception-v3	Entropia Cruzada	0,9989
Inception-v3	CORN	0,9416
DeiT-Distilled-B	Entropia Cruzada	<b>1,0000</b>
DeiT-Distilled-B	CORN	0,9163
DaViT-B	Entropia Cruzada	0,9989
DaViT-B	CORN	0,8844
MaxViT-T	Entropia Cruzada	0,9989
MaxViT-T	CORN	0,9031
GCViT-B	Entropia Cruzada	0,9989
GCViT-B	CORN	0,8954
Swin-B	Entropia Cruzada	<b>1,0000</b>
Swin-B	CORN	0,8822

Tabela 15 – Valores de cobertura da predição conformal para cada modelo e função de perda.

DenseNet-121, DeiT-Distilled-B e Swin-B. Embora isso satisfaça a garantia teórica de cobrir pelo menos 95% dos casos, valores altos podem indicar que o método é excessivamente conservador, ou seja, que o limiar de confiança ( $\hat{q}$ ) calculado é muito alto, levando os modelos a incluírem mais classes para garantir a cobertura.

Já a cobertura para os modelos treinados com CORN foi mais variável, geralmente situando-se mais próxima ao objetivo de 95%, exemplo do ResNet-50 (93,61%) e Inception-v3 (94,16%). No entanto, alguns modelos apresentaram uma cobertura abaixo do esperado, como o DenseNet-121, com 86,67%, indicando que a abordagem ordinal, embora mais “justa” em sua calibração, pode ser menos estável em garantir a cobertura em todos os casos.

Analizando agora o tamanho dos conjuntos de predição gerados, a utilidade do método da predição conformal é inversamente proporcional ao tamanho desses conjuntos, isto é, conjuntos menores indicam maior confiança e são mais informativos. Nesse sentido,

a abordagem ordinal com CORN demonstrou uma superioridade prática.

Como pode ser visto na Tabela 16, os histogramas para os modelos treinados com entropia cruzada mostram uma certa tendência de incerteza, pois raramente expressam confiança suficiente para gerar uma predição única. Por exemplo, o Swin-B não gerou nenhum conjunto de tamanho um, enquanto o ResNet-50 gerou apenas cinco. A grande maioria das predições resultou em conjuntos de tamanho três ou quatro, tornando a saída pouco útil na prática. No modelo DaViT-B, 828 das 908 predições (mais de 91%) continham três ou mais classes possíveis. Para um radiologista, um resultado como “Grau 1, 2 ou 3” oferece pouca orientação para diagnosticar o nível de severidade da OA.

Curiosamente, a abordagem ordinal com uso do CORN produziu conjuntos de predição muito mais informativos e úteis. A principal vantagem foi o número de conjuntos de tamanho um, indicando alta confiança do modelo em muitos casos. Por exemplo, os modelos Inception-v3 e DenseNet-169 geraram 198 e 181 predições únicas, respectivamente. Isso representa aproximadamente 20% do conjunto de teste recebendo uma resposta precisa. Os histogramas mostram uma distribuição mais deslocada para a esquerda, ou seja, conjuntos menores.

Portanto, para uma ferramenta de suporte à decisão clínica, a abordagem baseada em CORN é superior. Um radiologista se beneficiaria muito mais de um sistema que frequentemente fornece uma resposta única e acionável, mesmo que a garantia de cobertura seja ligeiramente menos rígida, do que de um sistema que quase sempre responde com um conjunto de três a quatro possibilidades. A abordagem ordinal traduz melhor a confiança do modelo em resultados práticos e clinicamente relevantes.

## 5.6 Interpretabilidade dos Modelos

Para garantir que o desempenho quantitativo dos modelos seja acompanhado de um processo de decisão clinicamente relevante, foi realizada uma análise qualitativa utilizando a técnica de Grad-CAM. A análise revelou padrões consistentes e distintos entre as arquiteturas, reforçando a confiabilidade dos modelos de melhor desempenho. A Tabela 17 ilustra as Grad-CAMs para os modelos que tiveram o melhor desempenho preditivo.

Uma observação fundamental foi que os modelos de modo geral focaram sua atenção em regiões cruciais para o diagnóstico da OA, como o espaço articular. Por exemplo, como ilustrado na Tabela 17, para uma classificação correta de KL-3 (moderado), o mapa de calor evidencia uma forte ativação sobre o espaço articular com o menor estreitamento (lado esquerdo da radiografia). Esse comportamento foi observado na grande maioria das predições corretas, indicando que os modelos não apenas acertaram a classe, mas o fizeram com base nos mesmos indicadores visuais que um radiologista utilizaria, aumentando a

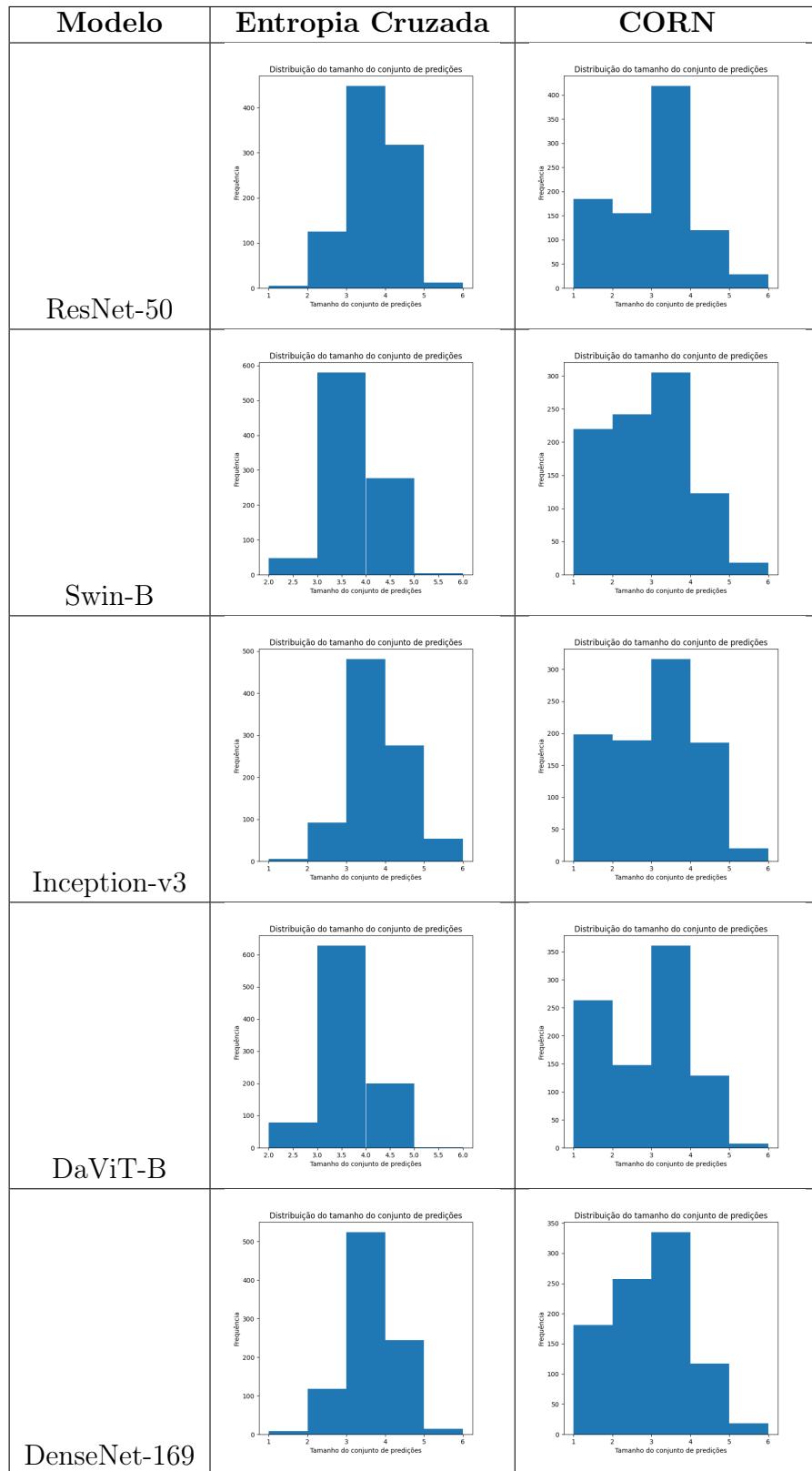


Tabela 16 – Histograma do tamanho dos conjuntos de predição para cinco modelos relevantes.

confiança em sua validade clínica.

Ainda assim, o Grad-CAM revelou estratégias visuais distintas entre as RNCs e os ViTs. Para a mesma radiografia de entrada, as RNCs tenderam a produzir mapas de calor mais focados e localizados. Sua atenção se concentrou onde existem bordas de alta frequência, ou seja, nos espaços articulares. Isso reflete o forte viés indutivo das convoluções para a detecção desses padrões locais.

Por outro lado, os ViTs geraram mapas de calor mais difusos e contextuais. Em vez de focar em um único ponto, a atenção frequentemente se espalhou por uma região mais ampla, como é o caso dos modelos DaViT-B e GCViT-B. Isso também está alinhado com a capacidade dos transformers de modelar relações de longo alcance em uma imagem.

Em síntese, a análise de interpretabilidade com Grad-CAM não só validou a hipótese 4 (interpretabilidade clínica), onde os modelos de alto desempenho aprenderam a identificar características patológicas clinicamente relevantes da OA de joelho, mas também ofereceu uma análise sobre as diferentes estratégias de análise visual das arquiteturas. Tanto RNCs quanto ViTs se mostraram eficazes para a classificação correta com base em características visuais adequadas e podem ser utilizadas como ferramenta de suporte na análise clínica para diagnosticar a severidade da OA de joelho.

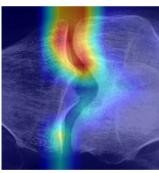
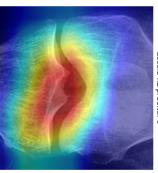
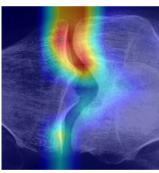
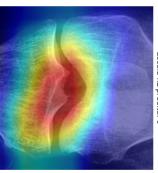
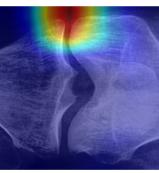
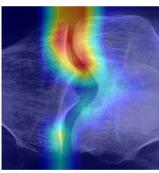
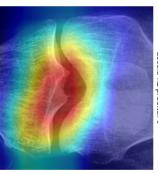
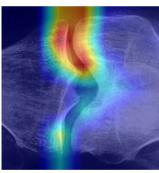
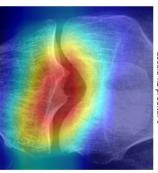
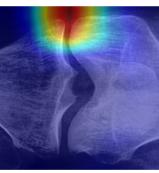
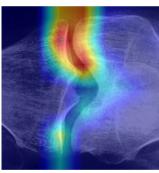
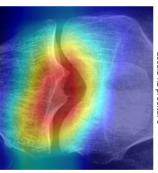
Modelo	KL-0	KL-1	KL-2	KL-3	KL-4
DenseNet-169					
DenseNet-121					
DaViT-B					
GCViT-B					
Inception-v3					

Tabela 17 – Visualização Grad-CAM para os cinco principais modelos.

## 6 Conclusão

Este trabalho se propôs a realizar uma análise comparativa abrangente entre arquiteturas de redes neurais convolucionais (RNCs) e vision transformers (ViTs) para a tarefa de classificação da severidade da osteoartrite (OA) de joelho, utilizando a escala de Kellgren/Lawrence. Diante da subjetividade e do tempo demandado pelo diagnóstico manual, o objetivo central foi identificar os modelos mais robustos, eficientes e confiáveis, empregando uma avaliação com diferentes variáveis que incluiu desempenho preditivo, custo computacional, quantificação de incerteza e interpretabilidade.

A investigação sistemática das treze arquiteturas revelou conclusões claras e significativas. Os modelos da família DenseNet se destacaram como os de melhor desempenho geral. O DenseNet-169 alcançou a maior acurácia (78,85%), enquanto o DenseNet-121 obteve o maior QWK (0,8878), demonstrando a melhor performance preditiva na tarefa de classificação ordinal.

A comparação entre as funções de perda de entropia cruzada e CORN evidenciou uma compensação fundamental. Enquanto a entropia cruzada maximizou a acurácia, a função CORN consistentemente melhorou o QWK, minimizando a gravidade dos erros de classificação. Adicionalmente, a análise com a predição conformal mostrou que a abordagem com CORN gera conjuntos de predição drasticamente mais informativos e úteis para a prática clínica.

A análise da eficiência computacional apresentou uma expressiva vantagem de eficiência das RNCs sobre os ViTs. Modelos como DenseNet-121 e ResNet-50 foram de quatro a cinco vezes mais rápidos em inferência do que os ViTs de alto desempenho, como DaViT-B e Swin-B. Esse resultado sublinha um desafio prático para a implantação de ViTs em ambientes clínicos que demandam baixo custo e alta velocidade.

A análise com Grad-CAM confirmou que os modelos de melhor desempenho basearam suas decisões em marcadores patológicos clinicamente relevantes, como o espaço articular e osteófitos. Também foram identificadas “estratégias visuais” distintas, com as RNCs produzindo ativações mais focadas e os ViTs, mais contextuais.

Apesar do rigor metodológico, este trabalho possui algumas limitações que devem ser reconhecidas. O estudo foi conduzido num único conjunto de dados público com um desbalanceamento claro. Além disso, a análise se baseou apenas em imagens estáticas, sem qualquer informação extra sobre os pacientes. Trabalhos futuros podem explorar outros conjuntos de dados para garantir a generalização dos resultados, integrar com um estudo longitudinal que acompanhasse a progressão da OA ao longo do tempo, além de otimizar as arquiteturas com estratégias de *grid-searching* e *ensemble*, combinando a saída dos

melhores modelos para determinar a classe KL.

Este trabalho demonstrou com sucesso que modelos de aprendizado profundo, em particular as RNCs e ViTs, são capazes de classificar a severidade da OA de joelho com alta acurácia e confiabilidade. Mais importante, o estudo evidenciou que uma avaliação completa deve ir além da acurácia, considerando a relevância clínica do erro, a eficiência computacional e a interpretabilidade. Assim, o estudo fornece um *benchmark* sólido para o desenvolvimento de futuras ferramentas de IA que possam auxiliar profissionais da saúde no diagnóstico e manejo da OA de joelho.

## Referências

- ALOTAIBI, A. et al. *ViT-DeiT: An Ensemble Model for Breast Cancer Histopathological Images Classification*. 2022. 20
- ANDERSON, A. S.; LOESER, R. F. *Why is osteoarthritis an age-related disease?* 2010. 5
- ANGELOPOULOS, C.; BATES, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021. 33
- APON, T. S. et al. *Transforming Precision: A Comparative Analysis of Vision Transformers, CNNs, and Traditional ML for Knee Osteoarthritis Severity Diagnosis*. 2024. 10, 40, 41, 57
- BRASÍLIA, F. *Metade dos adultos brasileiros com obesidade em 20 anos*. 2024. 9
- BRAUN, H. J.; GOLD, G. E. Diagnosis of osteoarthritis: Imaging. *Bone*, v. 51, 2012. ISSN 87563282. 9
- CHEN, P. *Knee Osteoarthritis Dataset with Severity Grading*. 2018. <<https://www.kaggle.com/datasets/shashwatwork/knee-osteoarthritis-dataset-with-severity>>. 43
- COURTIES, A. et al. Osteoarthritis year in review 2024: Epidemiology and therapy. *Osteoarthritis and Cartilage*, v. 32, n. 11, p. 1397–1404, 2024. ISSN 1063-4584. 1, 9
- CUEVA, J. H. et al. Detection and classification of knee osteoarthritis. *Diagnostics*, v. 12, 2022. ISSN 20754418. 10, 38, 39, 57
- DAI, L. et al. A deep learning system for detecting diabetic retinopathy across the disease spectrum. *Nature Communications*, v. 12, 2021. ISSN 20411723. 1
- DESMEULES, F. et al. Waiting for total knee replacement surgery: Factors associated with pain, stiffness, function and quality of life. *BMC Musculoskeletal Disorders*, v. 10, 2009. ISSN 14712474. 7
- DING, M. et al. *DaViT: Dual Attention Vision Transformers*. 2022. 9, 23, 24, 58
- DOMINGUES, J. G. et al. Development of a convolutional neural network for diagnosing osteoarthritis, trained with knee radiographs from the elsa-brasil musculoskeletal. *Radiologia Brasileira*, Publicação do Colégio Brasileiro de Radiologia e Diagnóstico por Imagem, v. 56, n. 5, p. 248–254, Sep 2023. ISSN 0100-3984. 10, 37, 38, 57
- DOSOVITSKIY, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. In: *ICLR 2021 - 9th International Conference on Learning Representations*. [S.l.: s.n.], 2021. 9, 2, 18, 19, 20, 23
- FERREL, B. A. *Pain management in elderly people*. 1992. 7
- FONTANA, M.; ZENI, G.; VANTINI, S. Conformal prediction: A unified review of theory and new challenges. *Bernoulli*, v. 29, 2023. ISSN 13507265. 33

- GILDENBLAT, J.; CONTRIBUTORS. *PyTorch library for CAM methods*. [S.l.]: GitHub, 2021. <<https://github.com/jacobgil/pytorch-grad-cam>>. 50
- GOLDRING, M. B.; MARCU, K. B. *Cartilage homeostasis in health and rheumatic diseases*. 2009. 6
- HATAMIZADEH, A. et al. *Global Context Vision Transformers*. 2022. 10, 26, 27
- HE, K. et al. Deep residual learning for image recognition. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2016. v. 2016-December. ISSN 10636919. 11, 13, 14
- HINTON, G. E.; OSINDERO, S.; TEH, Y. W. A fast learning algorithm for deep belief nets. *Neural Computation*, v. 18, 2006. ISSN 08997667. 10
- HOOGEBOOM, T. J. et al. Longitudinal impact of joint pain comorbidity on quality of life and activity levels in knee osteoarthritis: Data from the osteoarthritis initiative. *Rheumatology (United Kingdom)*, v. 52, 2013. ISSN 14620324. 7
- HUANG, G. et al. Densely connected convolutional networks. In: *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*. [S.l.: s.n.], 2017. v. 2017-January. 9, 11, 14, 15, 51
- ITSEEZ. *Open Source Computer Vision Library*. 2015. Disponível em: <<https://github.com/itseez/opencv>>. 45
- KANAMOTO, T. et al. *Significance and definition of early knee osteoarthritis*. 2020. 1, 6
- KAPETANAKIS, S. Evaluation of improvement in quality of life and physical activity after total knee arthroplasty in greek elderly women. *The Open Orthopaedics Journal*, v. 5, 2011. ISSN 18743250. 8
- KAWANO, M. M. et al. Assessment of quality of life in patients with knee osteoarthritis. *Acta Ortopedica Brasileira*, v. 23, 2015. ISSN 14137852. 8
- KELLGREN, J. H.; LAWRENCE, J. S. Radiological assessment of osteo-arthrosis. *Annals of the rheumatic diseases*, v. 16, 1957. ISSN 00034967. 1, 9
- KRAAN, P. M. van der; BERG, W. B. van den. *Osteophytes: relevance and biology*. 2007. 6
- KRAUS, V. B. et al. *Call for standardized definitions of osteoarthritis and risk stratification for clinical trials and clinical use*. 2015. 1, 9
- LEUNG, K. et al. Prediction of total knee replacement and diagnosis of osteoarthritis by using deep learning on knee radiographs: Data from the osteoarthritis initiative. *Radiology*, v. 296, 2020. ISSN 15271315. 13
- LIN, D. H. et al. Efficacy of 2 non-weight-bearing interventions, proprioception training versus strength training, for patients with knee osteoarthritis: A randomized clinical trial. *Journal of Orthopaedic and Sports Physical Therapy*, v. 39, 2009. ISSN 01906011. 1
- LIU, Z. et al. Swin transformer: Hierarchical vision transformer using shifted windows. In: *Proceedings of the IEEE International Conference on Computer Vision*. [S.l.: s.n.], 2021. ISSN 15505499. 9, 20, 22, 25

- LOESER, R. F. et al. *Osteoarthritis: A disease of the joint as an organ.* 2012. 9, 5, 6, 7
- LOHN, A.; MUSSER, M. Ai and compute. *Blog Open AI*, 2022. 32
- MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, v. 5, 1943. ISSN 00074985. 10
- MOHAMMED, A. S. et al. Knee osteoarthritis detection and severity classification using residual neural networks on preprocessed x-ray images. *Diagnostics*, v. 13, 2023. ISSN 20754418. 10, 1, 37, 38, 43, 57
- MUJAHID, M. et al. Pneumonia classification from x-ray images with inception-v3 and convolutional neural network. *Diagnostics*, v. 12, 2022. ISSN 20754418. 17
- ORGANIZATION, W. H. *WHOQOL: Measuring Quality of Life*. 2012. 7
- PACCA, D. M. et al. Prevalência de dor articular e osteoartrite na população obesa brasileira. *ABCD. Arquivos Brasileiros de Cirurgia Digestiva (São Paulo)*, v. 31, 2018. ISSN 2317-6326. 1, 5
- PASZKE, A. et al. Automatic differentiation in pytorch. *Advances in Neural Information Processing Systems*, 2017. 45
- PEREIRA, T. et al. Targeting the uncertainty of predictions at patient-level using an ensemble of classifiers coupled with calibration methods, venn-abers, and conformal predictors: A case study in ad. *Journal of Biomedical Informatics*, v. 101, 2020. ISSN 15320464. 33
- PESSLER, F. et al. The synovitis of "non-inflammatory"orthopaedic arthropathies: A quantitative histological and immunohistochemical analysis. *Annals of the Rheumatic Diseases*, v. 67, 2008. ISSN 00034967. 6
- RAJPURKAR, P. et al. *CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning*. 2017. 15
- RUSSAKOVSKY, O. et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, v. 115, 2015. ISSN 15731405. 16, 17, 18, 47
- SAINI, D. et al. Automated knee osteoarthritis severity classification using three-stage preprocessing method and vgg16 architecture. *International Journal of Imaging Systems and Technology*, v. 33, 2023. ISSN 10981098. 12
- SARDIM, A. C.; PRADO, R. P.; PINFILDI, C. E. Efeito da fotobiomodulação associada a exercícios na dor e na funcionalidade de pacientes com osteoartrite de joelho: estudo-piloto. *Fisioterapia e Pesquisa*, v. 27, 2020. ISSN 1809-2950. 1
- SAXENA, A. An introduction to convolutional neural networks. *International Journal for Research in Applied Science and Engineering Technology*, v. 10, 2022. 9, 11
- SEKHRI, A. et al. *Automatic diagnosis of knee osteoarthritis severity using Swin transformer*. 2023. 10, 38, 40, 57
- SELVARAJU, R. R. et al. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization*, v. 17, 2016. ISSN 00418781. 34

- SENNA Érika R. et al. Prevalence of rheumatic diseases in brazil: A study using the copcord approach. *Journal of Rheumatology*, v. 31, 2004. ISSN 0315162X. 9
- SHAMSHAD, F. et al. *Transformers in medical imaging: A survey*. 2023. 2
- SHI, X.; CAO, W.; RASCHKA, S. Deep neural networks for rank-consistent ordinal regression based on conditional probabilities. *Pattern Analysis and Applications*, Springer Science and Business Media LLC, v. 26, n. 3, p. 941–955, jun. 2023. ISSN 1433-755X. 10, 28, 29
- SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. In: *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. [S.l.: s.n.], 2015. 11, 12
- SITAULA, C.; HOSSAIN, M. B. Attention-based vgg-16 model for covid-19 chest x-ray image classification. *Applied Intelligence*, v. 51, 2021. ISSN 15737497. 12
- SNIDER, M. G.; MACDONALD, S. J.; POTOTSCHNIK, R. *Waiting times and patient perspectives for total hip and knee arthroplasty in rural and urban Ontario*. 2005. 8
- SOVRASOV, V. *ptflops: a flops counting tool for neural networks in pytorch framework*. 2018–2024. <<https://github.com/sovrasov/flops-counter.pytorch>>. 47, 48
- SPECTOR, T. D.; COOPER, C. *Radiographic assessment of osteoarthritis in population studies: whither kellgren and lawrence?* 1993. 54
- SPECTOR, T. D.; MACGREGOR, A. J. Risk factors for osteoarthritis: Genetics. *Osteoarthritis and Cartilage*, v. 12, 2004. ISSN 10634584. 5
- SUTBEYAZ, S. T. et al. Influence of knee osteoarthritis on exercise capacity and quality of life in obese adults. *Obesity*, v. 15, 2007. ISSN 19307381. 8
- SZEGEDY, C. et al. Going deeper with convolutions. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2015. v. 07-12-June-2015. ISSN 10636919. 9, 16, 17
- SZEGEDY, C. et al. Rethinking the inception architecture for computer vision. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2016. v. 2016-December. ISSN 10636919. 11, 16, 17
- TARIQ, T.; SUHAIL, Z.; NAWAZ, Z. Knee osteoarthritis detection and classification using x-rays. *IEEE Access*, v. 11, 2023. ISSN 21693536. 10, 37, 43, 57
- TEKADE, R.; RAJESWARI, K. Lung cancer detection and classification using deep learning. In: *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*. [S.l.: s.n.], 2018. p. 1–5. 1
- TILVE, A. et al. Pneumonia detection using deep learning approaches. In: *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*. [S.l.: s.n.], 2020. p. 1–8. 1
- TOUVRON, H. et al. Training data-efficient image transformers and distillation through attention. In: *Proceedings of Machine Learning Research*. [S.l.: s.n.], 2021. v. 139. 9, 19, 21, 23, 25

- TSCHON, M. et al. *Gender and sex are key determinants in osteoarthritis not only confounding variables. A systematic review of clinical data.* 2021. 5
- TU, Z. et al. *MaxViT: Multi-Axis Vision Transformer.* 2022. 9, 25, 26, 58
- VASWANI, A. et al. *Attention Is All You Need.* 2023. 18
- VOVK, V.; GAMMERMAN, A.; SAUNDERS, C. Machine-learning applications of algorithmic randomness. *Sixteenth International Conference on Machine Learning,* 1999. 34
- WANG, H. et al. A comprehensive survey on deep active learning in medical image analysis. *Medical Image Analysis,* v. 95, p. 103201, 2024. ISSN 1361-8415. 1
- WANG, Z. et al. Transformer with selective shuffled position embedding and key-patch exchange strategy for early detection of knee osteoarthritis. *Expert Systems with Applications,* Elsevier BV, v. 255, p. 124614, 2024. ISSN 0957-4174. 10, 39, 40, 57
- WARE, J. E.; SHERBOURNE, C. D. The mos 36-item short-form health survey (sf-36): I. conceptual framework and item selection. *Medical Care,* v. 30, 1992. ISSN 15371948. 8
- YEOH, P. S. Q. et al. Transfer learning-assisted 3d deep learning models for knee osteoarthritis detection: Data from the osteoarthritis initiative. *Frontiers in Bioengineering and Biotechnology, Volume 11 - 2023,* 2023. ISSN 2296-4185. 38, 57
- ZHUANG, F. et al. *A Comprehensive Survey on Transfer Learning.* 2021. 17



## Apêndices



# APÊNDICE A – Resultados Suplementares

Como complemento à análise principal, foram conduzidos experimentos para avaliar o desempenho dos modelos na tarefa de classificar a OA de joelho em 4, 3 e 2 classes, através da manipulação das imagens do conjunto de dados.

## A.1 Classificação em 4 Classes

Inicialmente, foi realizado um experimento para a tarefa de classificação em 4 classes, excluindo a classe KL-1 (“duvidoso”). O objetivo foi investigar o impacto da remoção desta classe, identificada como ambígua no desempenho geral das arquiteturas.

A exclusão da classe KL-1 resultou em um aumento substancial e generalizado no desempenho de todos os modelos, validando a hipótese de que a ambiguidade desta classe era um dos principais desafios para a tarefa de classificação. A Tabela 18 resume as métricas de desempenho para os modelos mais relevantes neste cenário.

O resultado mais notável foi o aumento significativo na performance. O modelo GCViT-B, treinado com a função de perda CORN, alcançou uma acurácia de 90,08%, um QWK de 0,9311 e um MAE de 0,1635, estabelecendo-se como o modelo de melhor desempenho neste cenário. Outros modelos, como o DaViT-B, também apresentaram resultados expressivos, com 89,28% de acurácia e 0,9267 de QWK.

Diferentemente do cenário de 5 classes, onde os modelos da família DenseNet se destacaram, a remoção da classe ambígua permitiu que os ViTs, em particular o GCViT-B e o DaViT-B, demonstrassem seu pleno potencial, superando as RNCs em acurácia e QWK.

Modelo	Função de perda	Acurácia	QWK	MAE
ResNet-50	CORN	87,67%	0,9146	0,2105
DenseNet-121	CORN	88,20%	0,9153	0,1957
DenseNet-169	CORN	88,34%	0,9192	0,2051
Inception-v3	Entropia Cruzada	88,61%	0,9225	0,1917
DaViT-B	Entropia Cruzada	89,28%	0,9267	0,1783
GCViT-B	CORN	<b>90,08%</b>	<b>0,9311</b>	<b>0,1635</b>

Tabela 18 – Resumo das métricas gerais de desempenho para modelos selecionados no cenário de 4 classes.

## A.2 Classificação em 3 Classes

Um segundo experimento foi conduzido para avaliar a capacidade dos modelos em distinguir exclusivamente entre os diferentes estágios de severidade da OA, uma vez que a doença já está estabelecida. Para este fim, as classes KL-0 (saudável) e KL-1 (duvidoso) foram removidas, resultando em um problema de classificação com 3 classes: KL-2 (mínimo), KL-3 (moderado) e KL-4 (severo).

Essa simplificação do problema, focando apenas nos estágios da doença, resultou em um desempenho preditivo alto em todas as arquiteturas avaliadas. A Tabela 19 resume as métricas de desempenho para os modelos mais relevantes neste cenário. O resultado mais interessante foi a alta acurácia e concordância ordinal alcançadas. O modelo Inception-v3, treinado com a entropia cruzada, atingiu a maior acurácia de 94,43% e o maior QWK de 0,9285. Ainda assim, os demais modelos tiveram uma ótima performance.

Estes resultados indicam que os modelos avaliados possuem uma alta capacidade não apenas para detectar a OA, mas também para diferenciar seus estágios de severidade, uma tarefa fundamental para o planejamento de tratamentos e o acompanhamento da progressão da doença.

Modelo	Função de perda	Acurácia	QWK	MAE
ResNet-101	Entropia Cruzada	92,91%	0,9082	0,0709
DenseNet-121	CORN	93,42%	0,9136	0,0658
DaViT-B	Entropia Cruzada	93,92%	0,9218	0,0608
DenseNet-169	CORN	93,92%	0,9220	0,0608
Inception-v3	Entropia Cruzada	<b>94,43%</b>	<b>0,9285</b>	<b>0,0557</b>
GCViT-B	Entropia Cruzada	93,16%	0,9110	0,0684

Tabela 19 – Resumo das métricas gerais de desempenho para modelos selecionados no cenário de 3 classes.

## A.3 Classificação em 2 Classes

Finalmente, um terceiro experimento foi realizado para avaliar a capacidade dos modelos em uma tarefa de detecção binária, que possui grande relevância clínica para a triagem inicial de pacientes. Neste cenário, as classes KL-0 e KL-1 foram agrupadas para representar a “ausência de OA”, enquanto as classes KL-2, KL-3 e KL-4 foram consolidadas para representar a “presença de OA”.

Os resultados, resumidos na Tabela 20, indicam que todos os modelos foram capazes de realizar a detecção binária com um desempenho robusto, alcançando acurácias na faixa de 84% a 87%. Nesta formulação do problema, os ViTs demonstraram uma ligeira vantagem. O modelo GCViT-B (entropia cruzada) destacou-se com a maior acurácia, atingindo 87,10%, além do maior QWK de 0,7374 e MAE de 0,1290, indicando a melhor concordância

ajustada ao acaso. Logo em seguida, o MaxViT-T (CORN) também apresentou um excelente resultado, com 86,77% de acurácia.

As diferenças de desempenho entre as funções de perda foram mínimas e inconsistentes neste cenário, o que é esperado, uma vez que para um problema de duas classes a formulação ordinal do CORN se aproxima da logística binária padrão.

Modelo	Função de perda	Acurácia	QWK	MAE
ResNet-50	CORN	84,90%	0,6956	0,1510
DenseNet-121	CORN	85,67%	0,6861	0,1433
DenseNet-169	CORN	85,56%	0,7055	0,1444
DaViT-B	CORN	86,55%	0,7268	0,1345
MaxViT-T	CORN	86,77%	0,7300	0,1323
GCViT-B	Entropia Cruzada	<b>87,10%</b>	<b>0,7374</b>	<b>0,1290</b>

Tabela 20 – Resumo das métricas gerais de desempenho para modelos selecionados no cenário de classificação binária.