



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, 17 de dezembro de 2024

Guilherme de Sousa Santos

Deteccção e Classificação Automática de Osteoartrite de Joelho em Radiografias Utilizando Visão Computacional

Projeto de Graduação apresentado ao Programa de Graduação em Ciência da Computação (área de concentração: Visão Computacional), 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

17 de dezembro de 2024

Resumo

A osteoartrite de joelho (OA) 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. 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 ResNet34, ResNet50, ResNet101, VGG16, VGG19, DenseNet121, DenseNet169, Inception, ViT-B/16, DeiT e Swin Transformer. A análise comparativa considera tanto métricas de performance, após o uso de *transfer learning*, quanto o consumo computacional envolvido no treinamento dos modelos. Após a realização dos experimentos, observou-se que as arquiteturas ResNet-50 e DenseNet-169 obtiveram os melhores desempenhos, com acurácias de 72,48% e 73,19% na classificação da OA de joelho em cinco classes, respectivamente.

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 debilitating joint conditions worldwide, characterized as a progressive disease that primarily affects the knee cartilage. Although there is no cure, early detection is crucial to prevent its progression. Radiography is the main technique used to diagnose OA and classify it based on the Kellgren/Lawrence (KL) scale. However, radiological diagnosis depends on the professional's experience, interpretation, and time, which can lead to inconsistencies or errors. In this context, deep learning techniques offer a faster and more efficient alternative, enabling the automation of knee OA detection and classification. This study proposes a comparison between convolutional neural network (CNN) models and vision transformers (ViTs) for the task of classifying knee OA severity, including the models ResNet34, ResNet50, ResNet101, VGG16, VGG19, DenseNet121, DenseNet169, Inception, ViT-B/16, DeiT, Swin Transformer, and ResNet50-ViT-B/16. The comparative analysis considers both performance metrics, following the application of transfer learning, and the computational resources required to train the models. It is expected that the dense networks (DenseNet121 and DenseNet169), along with the hybrid architecture ResNet50-ViT-B/16, will get the best results.

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

Lista de ilustrações

Figura 1 – Número de imagens em cada classe do conjunto de dados	8
Figura 2 – Metodologia para os vision transformers	12
Figura 3 – Matriz de confusão do modelo ResNet-50.	19

Lista de tabelas

Tabela 1 – Escala de Kellgren/Lawrence para classificação da severidade de osteoartrite.	4
Tabela 2 – Desempenho dos modelos de RNCs e ViTs na classificação da OA de joelho usando a função de perda <i>crossentropy</i>	18
Tabela 3 – Desempenho dos modelos de RNCs e ViTs na classificação da OA de joelho usando a função da perda CORN.	20

Lista de abreviaturas e siglas

OA	Osteoartrite
KL	Kellgren/Lawrence
IA	Inteligência Artificial
RNC	Rede Neural Convolucional
ViT	Vision Transformer
WHO	World Health Organization
OAI	Osteoarthritis Initiative
NIH	National Institutes of Health
CAM	Class Activation Mapping
GAP	Global Average Pooling

Sumário

	Introdução	1
1	FUNDAMENTAÇÃO TEÓRICA	3
1.1	Osteoartrite de Joelhos	3
1.2	Visão Computacional na Saúde	4
1.3	Aprendizado Profundo	5
2	METODOLOGIA	7
2.1	Coleta de dados	7
2.2	Pré-processamento das imagens	8
2.2.1	Normalização	8
2.2.2	Equalização de Histograma	8
2.2.3	Aumento de dados	9
2.3	Arquitetura do modelo de Rede Neural Convolucional	9
2.3.1	ResNet (Residual Network)	10
2.3.2	VGG (Visual Geometry Group Network)	10
2.3.3	DenseNet (Densely Connected Convolutional Networks)	10
2.3.4	Inception (GoogLeNet)	11
2.4	Arquitetura do modelo de Vision Transformer	11
2.4.1	ViT-B/16	11
2.4.2	DeiT (Data-efficient Image Transformer)	12
2.4.3	Swin Transformer (Shifted Window Transformer)	12
2.5	Métricas de avaliação	13
2.5.1	Acurácia	13
2.5.2	Precisão	13
2.5.3	Recall	13
2.5.4	F1-Score	13
2.5.5	Matriz de Confusão	14
2.5.6	AUC-ROC	14
2.6	Método de visualização	14
3	RESULTADOS	17
	REFERÊNCIAS	21

Introdução

A osteoartrite (OA) é uma forma muito comum de doença articular, definida como uma condição degenerativa que se inicia nas articulações e afeta principalmente a cartilagem, o revestimento articular e os ligamentos (1), resultando em sintomas de dor, rigidez e mobilidade articular limitada (2). Tais fatores podem comprometer significativamente a qualidade de vida, especialmente em idosos e indivíduos obesos (3). A OA é altamente prevalente, sendo uma das principais causas de incapacidade no mundo, com grande incidência em articulações como joelhos e quadris, afetando uma em cada sete pessoas globalmente (4). De acordo com um estudo do World Health Organization (WHO), em 2023, estimava-se a prevalência global da OA de joelho em 365 milhões de indivíduos, com maior predominância em pessoas idosas e mulheres, com cerca de 70% e 60%, respectivamente (5).

Terapias farmacêuticas têm sido aplicadas a pacientes diagnosticados com OA de joelho com o objetivo de reduzir os sintomas de dor, uma vez que não existem medicamentos capazes de retardar o seu desenvolvimento. No entanto, a progressão da doença pode ser prevenida com um diagnóstico precoce, ou seja, nos estágios iniciais em que a OA de joelho ainda é reversível (6). A medicina comumente avalia a severidade da doença através dos graus de Kellgren/Lawrence (KL), que categoriza a doença em cinco níveis de progressão: 0 (saúdável), 1 (duvidoso), 2 (mínimo), 3 (moderado) e 4 (severo), dependendo da experiência e cuidado médico na interpretação das radiografias (4). Isso pode levar a inconsistências entre o grau previsto e o grau real da OA de joelho, devido às mínimas diferenças entre os estágios adjacentes da doença (7). Estudos indicam que procedimentos como artroscopia são invasivos e podem causar complicações (8), enquanto técnicas como tomografia computadorizada e ressonância magnética também são usadas, mas o diagnóstico pode ser impreciso por falta de experiência do profissional (9). Esses desafios têm impulsionado estudos sobre sistemas automáticos de detecção e classificação da OA de joelho.

Nos últimos anos, muitas áreas têm visto a introdução de sistemas de inteligência artificial (IA) para executar tarefas que eram realizadas de forma manual, incluindo na área da medicina para o diagnóstico de patologias, por exemplo. Avanços recentes em técnicas de aprendizado de máquina no campo da saúde levaram a uma aceleração no diagnóstico de diversas doenças, incluindo a OA de joelho (7). O uso de modelos de aprendizado profundo baseados em redes neurais convolucionais (RNCs) tem ganhado espaço no que tange tarefas relacionadas a visão computacional (10). Porém, isso só foi possível após a introdução de novas técnicas para treinar redes profundas em paralelo com avanços a nível de hardware (11). Aprendizado por transferência também é amplamente utilizado para reduzir uso

de recursos computacionais para tarefas que já são executadas por modelos existentes, como as redes residuais (ResNet), Visual Geometry Group (VGG) e as redes densamente conectadas (DenseNet) (10). Enquanto o uso de RNCs tem se mostrado útil em soluções de detecção em imagens médicas, a operação de convolução limita o relacionamento entre pixels distantes numa imagem. Para tanto, a habilidade de codificar dependências de longo alcance tem sido possível graças às arquiteturas de aprendizado profundas baseadas em atenção, como o Vision Transformer (ViT). Tais modelos de ViT têm sido empregados para várias tarefas, incluindo classificação e detecção de objetos (12).

A relevância desta pesquisa reside na necessidade de aprimorar o processo de diagnóstico da osteoartrite de joelho, uma doença que afeta milhões de pessoas em todo o mundo e cuja detecção precoce é crucial para retardar sua progressão. O diagnóstico manual, feito por radiologistas, muitas vezes é subjetivo e suscetível a erros, o que pode levar a diagnósticos tardios ou incorretos. A aplicação de RNCs oferece uma solução promissora para automatizar esse processo, proporcionando uma avaliação mais precisa e eficiente a partir de radiografias. Essa automatização pode reduzir a carga dos profissionais de saúde e aumentar a acessibilidade de diagnósticos mais rápidos e confiáveis. Além disso, a comparação entre arquiteturas de RNCs e modelos baseados em transformers é relevante para identificar qual abordagem oferece melhor desempenho na classificação da severidade da osteoartrite.

O objetivo deste trabalho consiste em realizar uma comparação entre as métricas de performance e eficiência computacional de RNCs e modelos de ViTs na tarefa de detecção e classificação da OA de joelho seguindo a escala de Kellgren/Lawrence a partir de radiografias, com o intuito de identificar qual abordagem é mais adequada para uso em diagnósticos clínicos. Para atingir esse objetivo, será necessário estudar as modelos de RNCs e ViTs e propor uma arquitetura capaz de solucionar o problema. Em seguida, deverá ser feito o treinamento dos modelos e, por fim, será realizada uma análise detalhada das métricas de performance, incluindo acurácia, tempo de processamento e consumo de recursos, permitindo uma avaliação comparativa das duas abordagens.

A metodologia adotada envolve o uso de técnicas de pré-processamento de imagens para reduzir ruído das radiografias, melhorar contraste e ampliar o conjunto de dados para melhorar a performance e evitar o problema de *overfitting*. Em seguida, diferentes arquiteturas de RNCs serão treinadas usando a estratégia de *transfer learning*, e seus resultados comparados com os modelos de vision transformer treinados para a mesma tarefa.

1 Fundamentação Teórica

Neste capítulo, são apresentados os conceitos e as definições necessárias para o entendimento deste trabalho. A Seção 1.1 apresenta a osteoartrite de joelhos e suas características clínicas. A seção 1.2 aborda a visão computacional na área da saúde. A Seção 1.3 mostra alguns conceitos fundamentais de arquiteturas de aprendizado profundo, incluindo as redes neurais convolucionais e os vision transformers.

1.1 Osteoartrite de Joelhos

A osteoartrite (OA) é uma doença heterogênea e degenerativa, que afeta as articulações e estruturas ósseas de pacientes (13). A OA é a forma mais comum de doença articular, com uma prevalência global estimada em 365 milhões de indivíduos em 2023, e uma das principais causas de incapacidade no mundo, sendo altamente prevalente em idosos e indivíduos obesos (3). A OA é caracterizada por sintomas de dor, rigidez e mobilidade articular limitada, que podem comprometer significativamente a qualidade de vida dos pacientes. Embora possa afetar várias articulações, como ombros, cotovelos, pulso, coluna, entre outros, a OA é mais comum em joelhos e quadris (2), onde a cartilagem articular é mais suscetível a desgastes causados pela carga do corpo.

A prevalência da OA cresceu 132% nos últimos 30 anos, cuja projeção é de crescimento de 60 a 100% até 2050. É observado também que a prevalência está correlacionada com o status socioeconômico do país, sendo mais comum em países desenvolvidos, como os Estados Unidos, onde quase 10% da população adulta é afetada pela doença. Entre as causas da OA, estão fatores genéticos, idade, sexo, obesidade, trauma articular, entre outros. Entretanto, embora a OA seja uma doença multifatorial, a obesidade é um dos principais fatores de risco, contribuindo com aproximadamente 20% no crescimento dos casos, uma vez que o excesso de peso aumenta a carga nas articulações, acelerando o desgaste da cartilagem (OA Prevalence and Burden) (14).

Kellgren e Lawrence (6) 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 (Tabela 1): 0 (nenhum), 1 (duvidoso), 2 (mínimo), 3 (moderado) e 4 (grave) (4). Tal classificação é comumente feita por radiologistas, que avaliam as radiografias e atribuem um grau de acordo com a experiência e cuidado médico 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 num cenário onde a detecção precoce é crucial para retardar a

progressão da doença, uma vez que não existem medicamentos capazes de retardar o seu desenvolvimento.

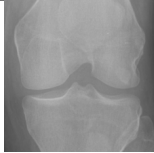
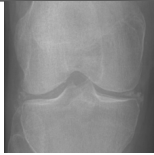
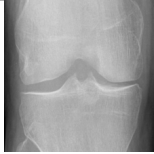
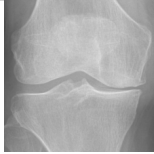
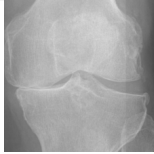
Classe KL	Exemplo de Imagem
0 (saudável)	
1 (duvidoso)	
2 (mínimo)	
3 (moderado)	
4 (severo)	

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

1.2 Visão Computacional na Saúde

A visão computacional é uma subárea da inteligência artificial (IA) que tem como objetivo automatizar a análise de imagens digitais, permitindo que a máquina "veja" e interprete o conteúdo visual de uma imagem. Essa ideia emergiu por volta da década de 1960, quando pioneiros como David Marr e Hans Moravec se questionaram da possibilidade de tornar computadores capazes de enxergar. Desde então, com o desenvolvimento de pesquisas na área de IA e melhorias em hardware, houveram diversos avanços na área, como o surgimento de algoritmos para detecção de bordas, detecção de objetos, segmentação de imagens, entre outros (15).

Na área da saúde, a visão computacional tem sido muito utilizada para melhorar a acurácia de diagnósticos, automatização de tarefas clínicas e tratamentos médicos. Ao analisar imagens médicas, como radiografias, tomografias, ressonâncias magnéticas e ultrassonografias, a máquina pode detectar e classificar patologias com maior precisão e rapidez do que um médico humano. Além disso, a visão computacional pode ser utilizada

para monitorar o progresso de doenças, monitorar a eficácia de tratamentos e até mesmo recomendar tratamentos personalizados para pacientes (16).

1.3 Aprendizado Profundo

O uso de modelos de aprendizado profundo baseados em redes neurais convolucionais (RNCs) tem ganhado espaço em tarefas de visão computacional. Aprendizado por transferência também é amplamente utilizado para reduzir o uso de recursos computacionais para tarefas que já são executadas por modelos existentes, como as redes residuais (ResNet), Visual Geometry Group (VGG) e as redes densamente conectadas (DenseNet) (10). Enquanto o uso de RNCs tem se mostrado útil em soluções de detecção em imagens médicas, a operação de convolução limita o relacionamento entre pixels distantes numa imagem. Para tanto, a habilidade de codificar dependências de longo alcance tem sido possível graças às arquiteturas de aprendizado profundas baseadas em atenção, como o Vision Transformer (ViT). Tais modelos de ViT têm sido empregados para várias tarefas, incluindo classificação e detecção de objetos (12).

2 Metodologia

Esta seção descreve a metodologia proposta para a tarefa de classificação da OA de joelho a partir de radiografias. A principal abordagem desta pesquisa consiste no uso de *transfer learning* para aproveitar o conhecimento já obtido por modelos pré-treinados e melhorar a performance da predição final.

2.1 Coleta de dados

A escolha e coleta dos dados é a primeira tarefa a ser realizada quando o objetivo é treinar um modelo de aprendizado profundo, incluindo redes neurais artificiais e vision transformers. Um conjunto de dados adequado é essencial para que o modelo tenha uma boa performance e seja útil para se tornar uma ferramenta de suporte no diagnóstico de OA de joelho. O conjunto de dados foi obtido a partir da plataforma Kaggle (17), uma fonte amplamente reconhecida por fornecer dados de alta qualidade e de domínio público para estudos acadêmicos e projetos de aprendizado de máquina. O conjunto de dados escolhido é baseado na Osteoarthritis Initiative (OAI), um estudo observacional multicêntrico de dez anos de homens e mulheres, patrocinado pelo National Institutes of Health (NIH), com o objetivo de permitir uma melhor compreensão da prevenção e tratamento da osteoartrite de joelho (18). Este conjunto contém radiografias de joelhos, juntamente com suas respectivas classificações de severidade da OA, conforme o sistema de Kellgren/Lawrence. Este dataset foi selecionado por sua relevância na plataforma, fornecendo uma base sólida para o treinamento dos modelos de RNCs e ViTs propostos nesta pesquisa. A Figura 1 ilustra a distribuição do conjunto de dados entre treino, teste e validação.

O conjunto de dados contém quatro pastas nomeadas “auto_test”, “test”, “train” e “val”, cada uma contendo as subpastas com imagens 224x224 representando cada um dos graus de KL. O dataset foi dividido entre dados de treino, teste e validação, com uma proporção de 7:2:1. O conjunto de treino é usado para treinar os modelos e consiste na maior proporção de imagens. O conjunto de validação é usado para ajustar os hiperparâmetros do modelo e monitorar o seu desempenho, enquanto o conjunto de teste é usado após o treinamento completo do modelo, para medir o desempenho final e verificar sua capacidade de generalização em dados completamente novos.

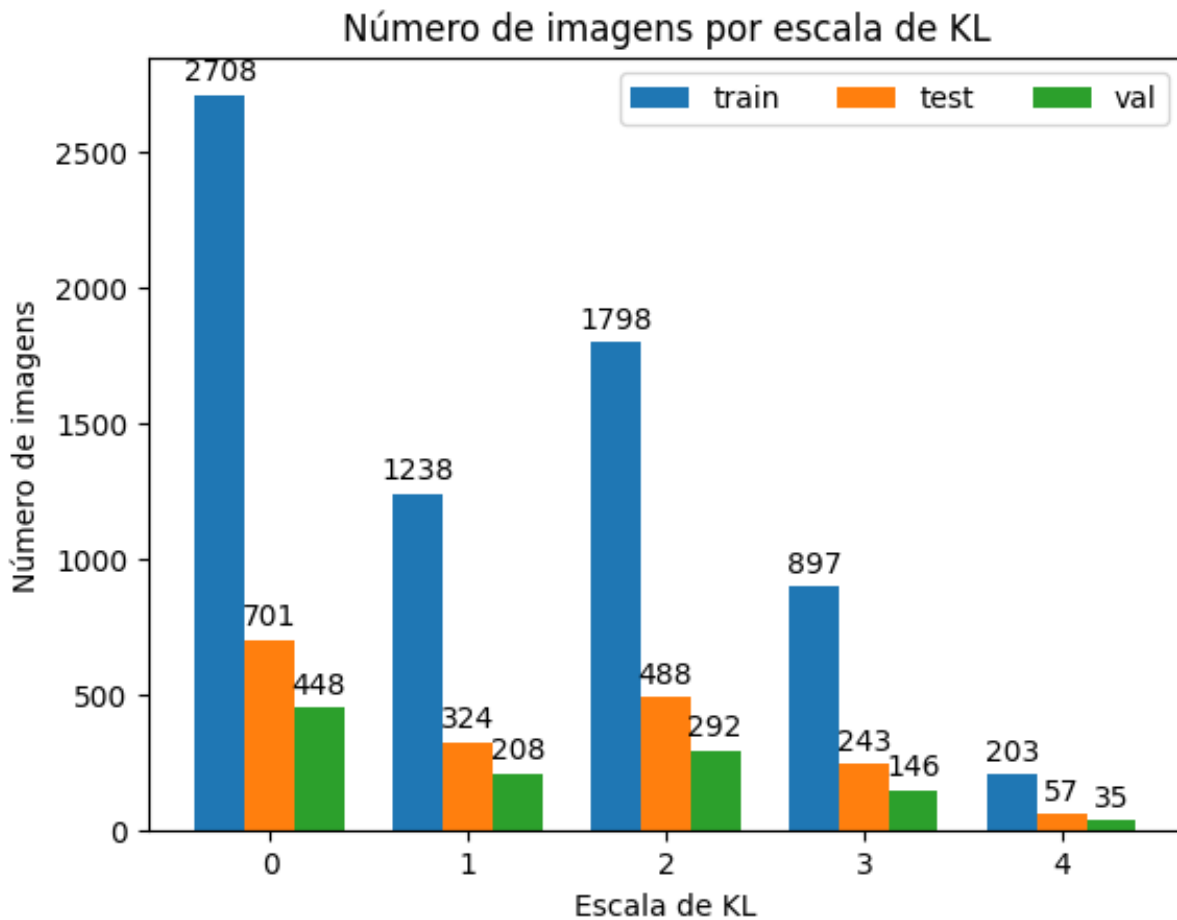


Figura 1 – Número de imagens em cada classe do conjunto de dados

2.2 Pré-processamento das imagens

O pré-processamento de imagens de raio-X é crucial para melhorar a qualidade e facilitar a análise automatizada pelos modelos. Para isso, algumas técnicas devem ser utilizadas, incluindo:

2.2.1 Normalização

A normalização de dados visa ajustar os valores para um intervalo padrão, melhorando a consistência dos dados e a eficiência dos modelos treinados. Para as radiografias, os pixels devem ter seus valores transformados para o intervalo entre 0 e 1.

2.2.2 Equalização de Histograma

A equalização de histograma é um método de processamento de imagem que busca melhorar o contraste e a visibilidade dos detalhes em uma imagem. Para isso, esta técnica redistribui os níveis de cinza da imagem, de forma que a distribuição dos valores de intensidade seja mais uniforme. Isso é feito calculando o histograma acumulado da

imagem original e utilizando-o para redistribuir os valores de cada pixel. Em particular, a equalização de histograma é útil para radiografias, onde a variação de intensidade pode ser sutil e a distinção entre diferentes graus de KL pode ser difícil.

Dada uma imagem $I(x, y)$ com intensidades de pixel $i \in \{0, 1, \dots, L - 1\}$ e número total de pixels N , onde x e y são as coordenadas do pixel, e L é o número de níveis de intensidade (normalmente $L = 256$ para imagens de 8 bits), a probabilidade de ocorrência de cada intensidade i é calculada como:

$$p(i) = \frac{n(i)}{N}$$

onde $n(i)$ é o número de pixels com intensidade i na imagem. O histograma acumulado H é então calculado como:

$$H(i) = \sum_{j=0}^i p(j)$$

O novo valor de intensidade j para um pixel com intensidade i é calculado como:

$$j = (L - 1) \times H(i)$$

onde $L - 1$ garante que o novo valor esteja no intervalo de intensidade da imagem. O resultado da equalização é normalmente arredondado para o valor inteiro mais próximo.

2.2.3 Aumento de dados

A ideia desta técnica é expandir artificialmente o tamanho e a variabilidade de um conjunto de dados, principalmente quando o volume de dados disponível é limitado. Isso torna os modelos mais robustos e genéricos, prevenindo *overfitting* e melhorando o desempenho em dados novos. As técnicas de aumento de dados que serão utilizadas nas radiografias são: rotação e reflexão (espelhamento) horizontal.

2.3 Arquitetura do modelo de Rede Neural Convolutacional

As redes neurais convolucionais possuem um papel muito relevante no contexto de inteligência artificial, especialmente em tarefas de visão computacional devido à sua capacidade de extrair características relevantes de imagens de forma automática, sem qualquer intervenção manual. Sua arquitetura é especialmente eficaz para reconhecer e classificar objetos em imagens complexas, inclusive em radiografias, com o intuito de auxiliar no processo de diagnóstico médico. As RNCs conseguem identificar variações sutis que podem estar associadas a condições patológicas, como é o caso da osteoartrite de

joelho, onde as variações entre os graus de KL reside no espaçamento articular da junção do joelho.

Fazer o treinamento de uma RNC sem nenhum conhecimento prévio do modelo é custoso em termos de quantidade de dados necessário, consumo de recursos computacionais e tempo. Para resolver este problema, o uso de *transfer learning* é essencial, pois permite aproveitar modelos já treinados em grandes conjuntos de dados genéricos, como o ImageNet, e adaptá-los para o conjunto de dados específico para o problema. Ao utilizar o *transfer learning*, as primeiras camadas do modelo, que capturam características gerais da imagem, são congeladas, enquanto as camadas finais são ajustadas para a tarefa específica, tal processo é chamado de *fine-tuning*. Isso economiza tempo e recursos computacionais e aumenta a eficácia do treinamento, resultando em modelos que podem fornecer diagnósticos precisos mesmo com volumes menores de dados disponíveis. Nos últimos anos, algumas arquiteturas performaram muito bem em algumas tarefas, como por exemplo a ResNet, VGG, Inception (GoogLeNet) e DenseNet. A arquitetura para os modelos de RNC pode ser vista na Figura ??.

2.3.1 ResNet (Residual Network)

A ResNet (19) é uma arquitetura amplamente utilizada em tarefas de classificação de imagens devido à sua capacidade de treinar redes profundas sem problemas de desaparecimento de gradiente. A inovação da ResNet está em seus blocos residuais, que introduzem conexões de atalho para permitir que os gradientes fluam melhor durante o treinamento. Isso torna a ResNet altamente eficiente para tarefas de classificação de imagens médicas. Para este trabalho, serão treinados os modelos ResNet34, ResNet50 e ResNet101, que oferecem um bom equilíbrio entre profundidade e performance.

2.3.2 VGG (Visual Geometry Group Network)

O VGG (20) é um modelo mais simples comparado ao ResNet, mas ainda é muito eficaz. Ele se destaca por usar camadas convolucionais de pequenos filtros (3x3) empilhadas seguidas por camadas de pooling. Embora o VGG tenha mais parâmetros que modelos mais modernos, sua estrutura é eficaz para capturar detalhes visuais em imagens médicas. O VGG16 e VGG19 serão utilizados nesta pesquisa.

2.3.3 DenseNet (Densely Connected Convolutional Networks)

O DenseNet (21) utiliza conexões densamente conectadas, onde cada camada recebe entradas de todas as camadas anteriores. Isso promove um fluxo eficiente de gradientes e incentiva o reuso de características aprendidas, o que pode ser muito útil nas radiografias de osteoartrite de joelho, onde detalhes finos precisam ser capturados, especialmente na

diferenciação entre graus de KL adjacentes. Os modelos do DenseNet121 e DenseNet169 serão as opções para este trabalho.

2.3.4 Inception (GoogLeNet)

A rede Inception (22), também chamada de GoogLeNet, é conhecida por seu uso de módulos Inception, que permitem que a rede aprenda de forma mais eficiente ao explorar convoluções de diferentes tamanhos em paralelo. A habilidade da Inception de capturar informações em várias escalas pode ser especialmente útil ao lidar com imagens médicas de diferentes resoluções. O Inception-v3 é uma versão mais moderna e, portanto, será utilizada nesta pesquisa.

2.4 Arquitetura do modelo de Vision Transformer

A arquitetura Vision Transformer tem se destacado como uma abordagem poderosa para tarefas de visão computacional devido à sua capacidade de capturar relações globais em imagens através do mecanismo de atenção (23). Essa abordagem permite que os modelos de ViTs alcancem ótimos resultados e superem as limitações das RNCs, que focam mais em características locais da imagem. Tal capacidade é particularmente relevante para o diagnóstico de patologias em imagens médicas, incluindo radiografias, onde o modelo é capaz de processar toda a imagem simultaneamente, associando partes distantes e próximas com igual relevância. Além disso, os ViTs também se beneficiam do *transfer learning*, permitindo que os modelos sejam treinados de forma eficiente em conjuntos de dados limitados. Para esta pesquisa será feito o *fine-tuning* de alguns modelos de ViT para a tarefa de classificação da OA de joelho, como o ViT-B/16, DeiT (Data-efficient Image Transformer), Swin Transformer (Shifted Window Transformer) e ResNet50-ViT-B/16. A arquitetura para os modelos de ViT pode ser vista na Figura 2.

2.4.1 ViT-B/16

O ViT-B/16 (23) é uma das primeiras variantes da arquitetura Vision Transformer, onde "B" representa o modelo base (base model) e "16" refere-se ao tamanho do *patch* em que a imagem é dividida (16x16 pixels). O ViT-B/16 recebe uma imagem e a divide em *patches*, tratando cada *patch* como um *token*, semelhante ao processamento de palavras em texto nos *transformers* tradicionais. O modelo usa um mecanismo de atenção para processar os *tokens* de maneira global, capturando interdependências entre diferentes regiões da radiografia. Essa abordagem permite que o ViT-B/16 compreenda melhor a estrutura geral da imagem, identificando padrões que podem se estender por grandes áreas da mesma. Este modelo pode ser especialmente eficaz para a tarefa de classificação da

OA de joelho, visto que existe o padrão notável do espaçamento articular que se estende horizontalmente na radiografia.

2.4.2 DeiT (Data-efficient Image Transformer)

O DeiT (24) é uma versão otimizada dos ViTs, projetada para melhorar a eficiência no uso de dados. Enquanto os ViTs originais, como o ViT-B/16, geralmente precisam de grandes quantidades de dados para atingir um bom desempenho, o DeiT foi projetado para ser treinado em conjunto de dados reduzidos. Isso acontece devido à técnica do *distillation token*, que permite ao modelo aprender a partir de um "professor" (modelo mais simples), aumentando a eficiência do treinamento. Este modelo pode ser particularmente útil na tarefa de classificação da OA de joelho, podendo ser um importante fator ao comparar com outros modelos de ViTs e RNCs.

2.4.3 Swin Transformer (Shifted Window Transformer)

O Swin Transformer (25) é uma arquitetura de ViT que introduz uma abordagem nova que utiliza *hierarchical feature maps* e *sliding windows* para aplicar a atenção e melhorar a eficiência e performance do modelo. Em vez de processar toda a imagem como uma sequência de *patches* globalmente, o Swin Transformer aplica a atenção dentro de pequenas janelas locais, de forma hierárquica, permitindo que o modelo mantenha a eficiência computacional e ainda capture detalhes locais e globais. Conforme o modelo avança pelas camadas, as janelas se expandem e se deslocam, permitindo que o modelo agregue contexto global ao longo do processamento. Essa estrutura hierárquica é particularmente eficaz para imagens de alta resolução, como as radiografias, onde há muitos detalhes importantes em diferentes escalas. Além disso, o Swin Transformer pode ser facilmente escalado para diferentes tamanhos de imagens e é altamente eficiente em termos de uso de memória e poder computacional, sendo uma escolha apropriada para a tarefa de classificação da OA de joelho.

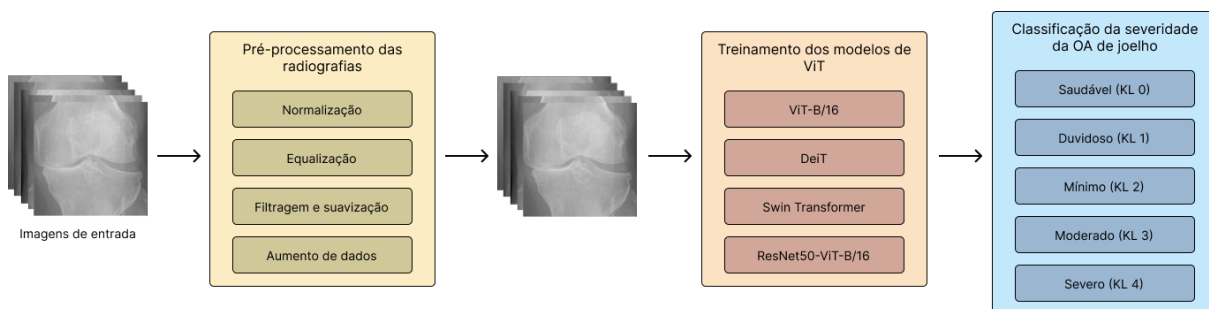


Figura 2 – Metodologia para os vision transformers

2.5 Métricas de avaliação

Para comparar a performance dos modelos treinados na tarefa de classificação da severidade da OA de joelho, serão utilizadas as seguintes métricas de avaliação: acurácia, precisão, revocação, F1-score e matriz de confusão. Essas métricas são amplamente utilizadas em problemas de classificação para medir a qualidade das previsões e o equilíbrio entre os diferentes tipos de erros. Para o cálculo das métricas, os seguintes acrônimos serão utilizados nas 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:

$$\text{Acurácia} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.1)$$

2.5.2 Precisão

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

$$\text{Precisão} = \frac{TP}{TP + FP} \quad (2.2)$$

2.5.3 Recall

O recall mede a capacidade do modelo de identificar corretamente todos os exemplos positivos. É definido como:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (2.3)$$

2.5.4 F1-Score

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

$$F1 = 2 \cdot \frac{\text{Precisão} \cdot \text{Recall}}{\text{Precisão} + \text{Recall}} \quad (2.4)$$

2.5.5 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

2.5.6 AUC-ROC

Para tarefas de classificação binária, será utilizada também a métrica AUC-ROC (Área Sob a Curva da Característica de Operação do Receptor), que 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.

$$\text{AUC-ROC} = \int_0^1 \text{TPR}(FPR) dFPR \quad (2.5)$$

onde TPR é a taxa de verdadeiros positivos e FPR é a taxa de falsos positivos.

2.6 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 Grad-CAM (Gradient-weighted Class Activation Mapping) é uma técnica usada para interpretar e visualizar as decisões feitas por redes neurais convolucionais (RNCs). Em tarefas de classificação, como a avaliação da severidade da OA de joelho a partir de radiografias, entender quais regiões da imagem 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 (26). 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 RNC utilizada 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 RNC 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 , 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 α_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:

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

Para esta pesquisa, a utilização do Grad-CAM permitirá a visualização das regiões das radiografias que o modelo considera mais relevantes para suas decisões de classificação. Isso não só facilita a interpretação dos resultados do modelo, mas também ajuda na validação de sua eficácia ao garantir que o modelo está focando nas áreas corretas da imagem, como o espaço articular do joelho.

3 Resultados

Esta pesquisa explora a transferência de aprendizado utilizando modelos pré-treinados no dataset ImageNet, aplicando ajuste fino para classificar o nível de severidade da osteoartrite de joelho com base na escala de Kellgren/Lawrence. O treinamento dos modelos foi realizado com a linguagem de programação Python, através de notebooks disponibilizados pela plataforma Google Colab, aproveitando os recursos computacionais de uma GPU T4 para acelerar o treino e a experimentação.

Para garantir a consistência entre os modelos, o treinamento foi realizado utilizando os mesmos hiperparâmetros. A classificação da osteoartrite de joelho foi organizada em cinco classes: KL 0, KL 1, KL 2, KL 3 e KL 4. O conjunto de dados foi dividido em 70% para treinamento, 10% para teste e 20% para validação. Para mitigar possíveis vieses, a base de dados foi balanceada por meio de técnicas de *undersampling* e *oversampling*, limitando cada classe a um máximo de 1700 imagens, complementadas por estratégias de *data augmentation*. O treinamento foi configurado com *batches* de 28 imagens, executado ao longo de 30 épocas, com um *early stopping* com 5 épocas de paciência para evitar *overfitting*.

Dois conjuntos de treinamento foram conduzidos: no primeiro, a função de perda utilizada foi a *crossentropy*, enquanto no segundo, foi substituída pela função de perda *Conditional Ordinal Regression for Neural Networks* (CORN), com o objetivo de explorar sua adequação ao problema de classificação ordinal. Em ambos os casos, o otimizador adotado foi o Adam, configurado com uma taxa de aprendizado inicial de 0.0001, ajustada dinamicamente a cada 3 épocas.

A [Tabela 2](#) mostra a acurácia dos modelos de RNCs e ViTs treinados para a classificação da OA de joelho usando a função de perda *crossentropy*. Em relação ao tempo de treinamento, é possível notar que o modelo mais rápido foi o ResNet-50, com um tempo de 11.29 segundos. Por outro lado, o modelo mais lento foi o DeiT, com um tempo de 79.5 segundos. Tais valores não necessariamente indicam que o modelo mais rápido é o pior, ou o contrário, mas é importante considerar o tempo de treinamento como um fator relevante ao escolher um modelo, especialmente se houver restrições de recursos computacionais. O tempo de treinamento mostrado varia, principalmente, com o número de épocas, já que modelos que levaram mais tempo são aqueles que tiveram a parada antecipada mais tarde, ou executaram as 30 épocas completas.

Quanto à acurácia geral (*overall*), todos os modelos apresentaram resultados razoavelmente bons, com valores variando de 0.6723 a 0.7319. Isso indica que todos os modelos foram capazes de aprender padrões relevantes para a classificação da OA de joelho.

Modelo	Tempo (min)	Overall	Classe KL				
			0	1	2	3	4
ResNet-34	14.85	0.7044	0.8146	0.3889	0.6680	0.8519	0.8246
ResNet-50	11.29	0.7248	0.8359	0.4475	0.6988	0.7942	0.8596
ResNet-101	18.81	0.7126	0.8031	0.4414	0.7029	0.7984	0.8596
VGG-16	51.33	0.6723	0.8588	0.2562	0.6783	0.8272	0.8421
VGG-19	23.21	0.6851	0.8217	0.2716	0.6906	0.7984	0.8246
DenseNet-121	24.48	0.7170	0.8616	0.2963	0.707	0.8477	0.8596
DenseNet-169	16.85	0.7319	0.8288	0.4043	0.7377	0.8477	0.8596
Inception-v3	16.45	0.7215	0.8017	0.4136	0.75	0.8148	0.8421
ViT-B	43.07	0.6955	0.8046	0.3241	0.7029	0.823	0.8596
DeiT	79.5	0.6862	0.7718	0.3488	0.6947	0.8354	0.8421
Swin	17.88	0.6977	0.8388	0.3117	0.6824	0.8025	0.8421

Tabela 2 – Desempenho dos modelos de RNCs e ViTs na classificação da OA de joelho usando a função de perda *crossentropy*.

No entanto, é importante notar que o modelo DenseNet-169 obteve a maior acurácia geral, com um valor de 0.7319. Isso sugere que arquiteturas de RNCs densamente conectadas podem ser muito eficazes na extração de características relevantes em imagens médicas como radiografias de joelho. Além disso, os modelos de conexões residuais (ResNet) também apresentaram resultados competitivos, com acurácias gerais variando de 0.7044 a 0.7248, onde o ResNet-50 obteve a maior acurácia dentre eles e com o menor tempo de treinamento, oferecendo um bom equilíbrio entre generalização do modelo e custo computacional.

Por outro lado, os modelos da família VGG (VGG-16 e VGG-19) apresentaram acurácias gerais mais baixas, variando de 0.6723 a 0.6851, o que sugere que essas arquiteturas mais simples podem não ser tão eficazes na extração de características complexas em radiografias de joelho. Embora fosse esperado que esses modelos tivessem desempenho inferior em relação aos modelos ResNet, devido à sua profundidade, os resultados indicam que esses modelos são capazes de aprender padrões relevantes e ter uma menor probabilidade de *overfitting*, como observado no tempo de treinamento do VGG-16, que foi maior que a maioria dos modelos justamente por não ter parada antecipada em virtude da queda do erro no conjunto de validação.

O GoogLeNet, com sua arquitetura Inception (versão 3), permitiu que o modelo tivesse uma acurácia geral de 0.7215, indicando que o modelo pode ser eficaz na extração de características relevantes e superar a maioria dos modelos de RNCs. Esse comportamento pode ser justificado pelo uso de uma técnica chamada de "bottleneck" ou "redução de dimensionalidade", que reduz a quantidade de parâmetros e a complexidade computacional do modelo, sem comprometer significativamente o desempenho.

Os modelos de transformers, por sua vez, apresentaram acurácias gerais variando de 0.6862 a 0.7215, indicando que essas arquiteturas podem ser eficazes, mas talvez não

sejam tão eficientes quanto os modelos de RNCs. O modelo Swin Transformer obteve a maior acurácia geral entre os modelos de transformers, com um valor de 0.6977, sugerindo que a abordagem hierárquica de atenção pode ser eficaz na extração de características relevantes em radiografias de joelho.

Entretanto, é importante notar que a acurácia para a classe KL 1 foi baixa para todos os modelos, variando de 0.2562 e 0.4475. Isso indica que a classificação da OA de joelho no estágio 1 (duvidoso) pode ser mais desafiadora, possivelmente devido à semelhança visual com as classes adjacentes KL 0 e KL 2. Esse resultado pode ser observado na [Figura 3](#), que mostra a matriz de confusão do modelo ResNet-50. A classe KL 1 tem a menor acurácia dentre todas as classes, o que reflete o desafio na classificação dessa classe devido ao nível de detalhe ou até mesmo incoerência na rotulação das imagens do dataset.

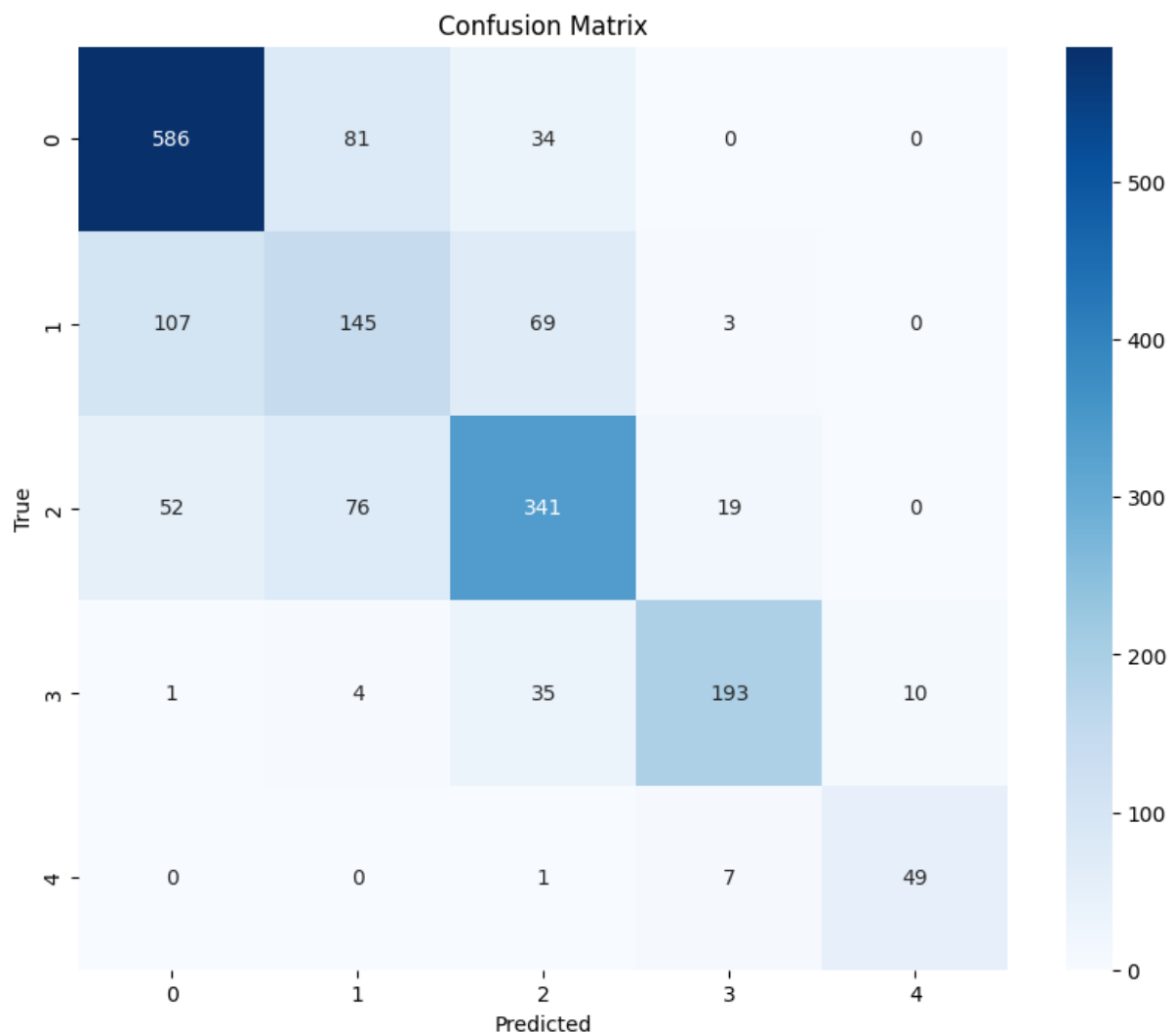


Figura 3 – Matriz de confusão do modelo ResNet-50.

Em resumo, os modelos ResNet-50 e DenseNet-169 se destacaram em termos de tempo de treinamento e acurácia geral, respectivamente. No entanto, é importante considerar as características de cada classe ao escolher um modelo, pois diferentes modelos

Modelo	Tempo	Overall	Classe KL				
			0	1	2	3	4
ResNet-34	14.93	0.6895	0.7518	0.5586	0.6107	0.8107	0.8246
ResNet-50	10.32	0.7181	0.796	0.5031	0.6824	0.823	0.8421
ResNet-101	16.17	0.6994	0.7418	0.4506	0.707	0.8519	0.8772
VGG-16	19.29	0.6762	0.7646	0.358	0.6824	0.7984	0.8246
VGG-19	24.05	0.6669	0.7974	0.3549	0.6066	0.7901	0.8246
DenseNet-121	10.62	0.6911	0.729	0.4444	0.7172	0.8272	0.8246
DenseNet-169	13.75	0.717	0.7874	0.5833	0.6393	0.8148	0.8596
Inception-v3	17.09	0.701	0.7932	0.5093	0.6639	0.7325	0.8421
ViT-B	36.97	0.6817	0.7447	0.4815	0.6393	0.8066	0.8772
DeiT	34.73	0.6602	0.7047	0.4877	0.6209	0.7984	0.8421
Swin	35.58	0.6546	0.7803	0.4658	0.6722	0.8395	0.7894

Tabela 3 – Desempenho dos modelos de RNCs e ViTs na classificação da OA de joelho usando a função da perda CORN.

podem ter desempenhos diferentes para cada classe.

A Tabela 3 apresenta os resultados dos modelos de RNCs e ViTs treinados para a classificação da OA de joelho usando a função de perda CORN. Em relação ao tempo de treinamento, não houve uma mudança significativa comparado com a função de perda *crossentropy*. O modelo mais rápido foi, novamente, o ResNet-50, com um tempo de 10.32 segundos, enquanto o modelo mais lento foi o Swin Transformer, com um tempo de 35.58 segundos. Em relação à acurácia geral, os resultados variaram de 0.6546 a 0.7181, indicando que a função de perda CORN pode ser eficaz na classificação da OA de joelho, mas não necessariamente supera a função de perda *crossentropy*. Isso é justificado pelo fato de que a função de perda CORN é mais adequada quando o modelo faz previsões mais afastadas do rótulo real, o que não foi evidenciado ao observar as matrizes de confusão dos modelos.

No entanto, é importante notar que o modelo ResNet-50 obteve a maior acurácia geral, com um valor de 0.7181, superando os demais modelos, inclusive o modelo DenseNet-169, que obteve a maior acurácia geral com a função de perda *crossentropy*. Isso sugere que a função de perda CORN pode ser eficaz em arquiteturas de RNCs, especialmente aquelas com conexões residuais. Entretanto, o modelo DenseNet-169 foi quem obteve a maior acurácia para a classe KL 1, com um valor de 0.5833, que é a classe mais desafiadora de ser classificada, como observado anteriormente.

Referências

- 1 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. Citado na página 1.
- 2 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. Citado 2 vezes nas páginas 1 e 3.
- 3 PACCA, D. M. et al. Desenvolvimento e aplicação de rede neural convolucional para o diagnóstico de osteoartrite de joelho. *Revista CPAQV - Centro de Pesquisas Avançadas em Qualidade de Vida*, v. 15, 2022. Disponível em: <<https://revista.cpaqv.org/index.php/CPAQV/article/view/1079>>. Citado 2 vezes nas páginas 1 e 3.
- 4 KELLGREN, J. H.; LAWRENCE, J. S. Radiological assessment of osteo-arthritis. *Annals of the rheumatic diseases*, v. 16, 1957. ISSN 00034967. Citado 2 vezes nas páginas 1 e 3.
- 5 ORGANIZATION, W. H. *Osteoarthritis*. 2023. <<https://www.who.int/news-room/fact-sheets/detail/osteoarthritis>>. Acessado em: 15 de agosto de 2024. Citado na página 1.
- 6 KANAMOTO, T. et al. *Significance and definition of early knee osteoarthritis*. 2020. Citado 2 vezes nas páginas 1 e 3.
- 7 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. Citado na página 1.
- 8 SARAIEV, A. V. et al. Arthroscopy for knee osteoarthritis in the xxi century: a systematic review of current high quality researches and guidelines of professional societies. *Traumatology and Orthopedics of Russia*, v. 26, 2020. ISSN 2311-2905. Citado na página 1.
- 9 ALSHAMRANI, H. A. et al. Osteo-net: An automated system for predicting knee osteoarthritis from x-ray images using transfer-learning-based neural networks approach. *Healthcare (Switzerland)*, v. 11, 2023. ISSN 22279032. Citado na página 1.
- 10 TARIQ, T.; SUHAIL, Z.; NAWAZ, Z. Knee osteoarthritis detection and classification using x-rays. *IEEE Access*, v. 11, 2023. ISSN 21693536. Citado 3 vezes nas páginas 1, 2 e 5.
- 11 LITJENS, G. et al. *A survey on deep learning in medical image analysis*. 2017. Citado na página 1.
- 12 SHAMSHAD, F. et al. *Transformers in medical imaging: A survey*. 2023. Citado 2 vezes nas páginas 2 e 5.

- 13 KRAUS, V. B. et al. *Call for standardized definitions of osteoarthritis and risk stratification for clinical trials and clinical use*. 2015. Citado na página 3.
- 14 COURTIES, A. et al. Osteoarthritis year in review 2024: Epidemiology and therapy. *Osteoarthritis Research Society International*, 2024. Citado na página 3.
- 15 TEAM, H. F. *What is Computer Vision?* 2024. Accessed: 2024-12-15. Disponível em: <<https://huggingface.co/learn/computer-vision-course/unit1/chapter1/definition>>. Citado na página 4.
- 16 JAVAID, M. et al. Computer vision to enhance healthcare domain: An overview of features, implementation, and opportunities. *Intelligent Pharmacy*, v. 2, n. 6, p. 792–803, 2024. ISSN 2949-866X. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S2949866X24000662>>. Citado na página 5.
- 17 CHEN, P. *Knee Osteoarthritis Dataset with Severity Grading*. 2018. <<https://www.kaggle.com/datasets/shashwatwork/knee-osteoarthritis-dataset-with-severity>>. Acessado em: 29 de setembro de 2024. Citado na página 7.
- 18 HEALTH, N. I. of. *Osteoarthritis Initiative*. 2024. <<https://www.niams.nih.gov/grants-funding/funded-research/osteoarthritis-initiative>>. Acessado em: 17 de julho de 2024. Citado na página 7.
- 19 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. Citado na página 10.
- 20 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. Citado na página 10.
- 21 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. Citado na página 10.
- 22 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. Citado na página 11.
- 23 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. Citado na página 11.
- 24 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. Citado na página 12.
- 25 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. Citado na página 12.

- 26 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. Citado na página [14](#).