



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 Redes Neurais Convolucionais

Guilhere de Sousa Santos

Santo André - SP, 12 de setembro de 2024

Guilhere de Sousa Santos

Deteccção e Classificação Automática de Osteoartrite de Joelho em Radiografias Utilizando Redes Neurais Convolucionais

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

12 de setembro 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, Swin Transformer e ResNet50-ViT-B/16. A análise comparativa considera tanto métricas de performance, após o uso de *transfer learning*, quanto o consumo computacional envolvido no treinamento dos modelos. Espera-se que as redes densas (DenseNet121 e DenseNet169), juntamente com a arquitetura híbrida ResNet50-ViT-B/16, apresentem os melhores resultados.

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 – Metodologia para as redes neurais convolucionais	7
Figura 2 – Metodologia para os vision transformers	8

Lista de tabelas

Tabela 1 – Quantidade de imagens em cada classe Kellgren/Lawrence do conjunto de dados escolhido.	4
---	---

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	METODOLOGIA	3
1.1	Coleta de dados	3
1.2	Pré-processamento das imagens	3
1.2.1	Normalização	4
1.2.2	Equalização	4
1.2.3	Filtragem e suavização	5
1.2.4	Aumento de dados	5
1.3	Arquitetura do modelo de Rede Neural Convolucional	5
1.3.1	ResNet (Residual Network)	6
1.3.2	VGG (Visual Geometry Group Network)	6
1.3.3	DenseNet (Densely Connected Convolutional Networks)	6
1.3.4	Inception (GoogLeNet)	6
1.4	Arquitetura do modelo de Vision Transformer	6
1.4.1	ViT-B/16	7
1.4.2	DeiT (Data-efficient Image Transformer)	7
1.4.3	Swin Transformer (Shifted Window Transformer)	8
1.4.4	ResNet50-ViT-B/16	8
1.5	Métricas de avaliação	9
1.5.1	Acurácia	9
1.5.2	Precisão	9
1.5.3	Recall	9
1.5.4	F1-Score	9
1.5.5	Matriz de Confusão	10
1.5.6	AUC-ROC	10
1.6	Método de visualização	10
2	RESULTADOS ESPERADOS	13
	REFERÊNCIAS	15

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 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.

1.1 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 (13), 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 (14). 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 Tabela 1 ilustra os detalhes do conjunto de dados.

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.

1.2 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

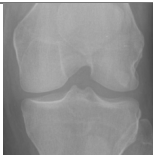




Classe KL	Número de Imagens	Exemplo de Imagem
0 (saudável)	3857	
1 (duvidoso)	1770	
2 (mínimo)	2578	
3 (moderado)	1286	
4 (severo)	295	

Tabela 1 – Quantidade de imagens em cada classe Kellgren/Lawrence do conjunto de dados escolhido.

utilizadas, incluindo:

1.2.1 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.

1.2.2 Equalização

A equalização busca melhorar o contraste e a visibilidade dos detalhes em uma imagem. O objetivo da equalização é redistribuir os níveis de cinza para que todos os valores de intensidade apareçam com uma frequência mais uniforme. Isso é realizado calculando o histograma acumulado da radiografia original e utilizando-o para redistribuir os valores de pixel.

1.2.3 Filtragem e suavização

As técnicas de filtragem e suavização visam reduzir o ruído e melhorar a qualidade visual das imagens. Para isso, será utilizado o filtro gaussiano, onde cada pixel tem o seu valor substituído pelo valor ponderado da média dos pixels vizinhos, onde os pesos são determinados pela função gaussiana.

1.2.4 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 melhorar o desempenho em dados novos. As técnicas de aumento de dados que serão utilizadas nas radiografias são: rotação, escalonamento e reflexão (espelhamento) horizontal.

1.3 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 1.

1.3.1 ResNet (Residual Network)

A ResNet (15) 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.

1.3.2 VGG (Visual Geometry Group Network)

O VGG (16) 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.

1.3.3 DenseNet (Densely Connected Convolutional Networks)

O DenseNet (17) 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.

1.3.4 Inception (GoogLeNet)

A rede Inception (18), 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.

1.4 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 (19). Essa abordagem permite que os

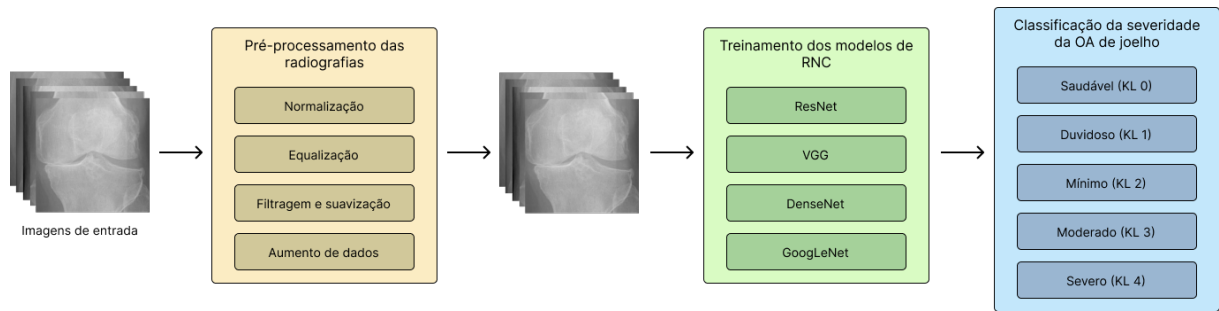


Figura 1 – Metodologia para as redes neurais convolucionais

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.

1.4.1 ViT-B/16

O ViT-B/16 (19) 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.

1.4.2 DeiT (Data-efficient Image Transformer)

O DeiT (20) 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.

1.4.3 Swin Transformer (Shifted Window Transformer)

O Swin Transformer (21) 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.

1.4.4 ResNet50-ViT-B/16

Uma abordagem híbrida, combinando RNC, como o ResNet, e ViT, pode ser aplicada de maneira eficaz na classificação da severidade da osteoartrite de joelho, aproveitando as vantagens de ambas as arquiteturas para obter uma melhor predição das classificações de KL (22, 23). Essa mescla pode ser observada no modelo ResNet50-ViT-B/16, onde o ResNet50 atua como um extrator de características iniciais, processando as radiografias e capturando padrões como texturas e bordas, e o modelo ViT-B/16 utiliza seus mecanismos de atenção para permitir uma análise mais contextualizada e eficiente. A combinação se faz promissora para melhorar a precisão na classificação da OA de joelho, equilibrando eficiência computacional e qualidade do modelo final.

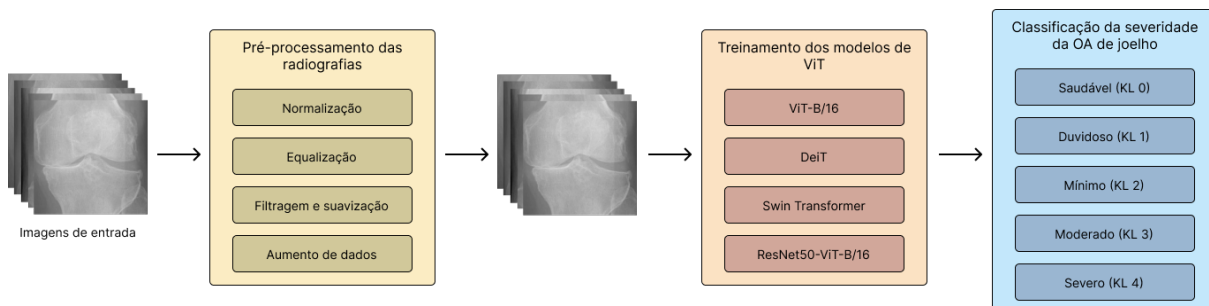


Figura 2 – Metodologia para os vision transformers

1.5 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.

1.5.1 Acurácia

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

$$\text{Acurácia} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1.1)$$

1.5.2 Precisão

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

$$\text{Precisão} = \frac{TP}{TP + FP} \quad (1.2)$$

1.5.3 Recall

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

$$\text{Recall} = \frac{TP}{TP + FN} \quad (1.3)$$

1.5.4 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:

$$F1 = 2 \cdot \frac{\text{Precisão} \cdot \text{Recall}}{\text{Precisão} + \text{Recall}} \quad (1.4)$$

1.5.5 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

1.5.6 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.

$$\text{AUC-ROC} = \int_0^1 \text{TPR}(FPR) dFPR \quad (1.5)$$

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

1.6 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 (24). 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.

2 Resultados esperados

Esta pesquisa utiliza transfer learning de modelos pré-treinados e faz um ajuste fino com o objetivo de classificar o nível de severidade da osteoartrite de joelho usando a escala de Kellgren/Lawrence. Para isso, os treinamentos serão feitos usando a linguagem de programação Python em um ambiente de notebooks disponíveis na plataforma do Google Colab. Caso o treinamento exija um maior poder computacional, o super computador da Universidade Federal do ABC poderá ser utilizado para esta pesquisa.

Nesta seção será discutido os resultados esperados quanto à performance dos modelos treinados usando a metodologia proposta. Espera-se que existam variações significativas de performance entre os modelos, dado que cada arquitetura possui diferentes estratégias para a extração e processamento das características contidas nas radiografias de joelho.

Os modelos ResNet (ResNet34, ResNet50 e ResNet101) são eficazes para extrair características complexas em imagens, incluindo imagens médicas como radiografias de joelho. Sabe-se que quanto mais profunda é a rede, maior é sua capacidade de capturar padrões complexos. Nesse sentido, o modelo ResNet101 (101 camadas) tende a ser o modelo com maior acurácia, mas pode sofrer com *overfitting* se o conjunto de dados for pequeno, como acontece para a classe KL 4, com apenas 295 imagens no total. Além disso, é esperado que o consumo de recursos computacionais seja maior para redes mais profundas. O modelo ResNet50 pode oferecer um bom equilíbrio entre generalização do modelo e custo computacional.

Os modelos VGG (VGG16 e VGG19), apesar de profundos, possuem uma arquitetura mais simples em comparação com o ResNet e são menos eficientes em termos de uso de parâmetros e, portanto, identificação de características complexas das radiografias. Embora seja esperado que estes modelos apresentem performance inferior em relação aos modelos ResNet considerando suas profundidades, pelo fato deles serem modelos com arquitetura mais simples e direta, consistindo principalmente de camadas convolucionais empilhadas seguidas por camadas totalmente conectadas, isso pode levar a uma melhor generalização do modelo e menor probabilidade de *overfitting*, trazendo uma performance melhor.

Os modelos DenseNet (DenseNet121 e DenseNet169) possuem conexões diretas entre todas as camadas, o que facilita o fluxo de informação e melhora a eficiência do aprendizado de padrões em imagens. Esses modelos podem ser especialmente úteis em capturar detalhes sutis nas radiografias, como pequenas degradações no espaço articular. Logo, espera-se que estes modelos tenham uma performance muito competitiva e superem os modelos ResNet e VGG, principalmente o DenseNet169.

O GoogLeNet, com sua arquitetura Inception, permite que o modelo capture diferentes tamanhos de características simultaneamente. Essa flexibilidade pode ser benéfica em radiografias ao extrair padrões importantes em diferentes imagens ou até conjuntos de dados. É esperado que este modelo tenha um bom desempenho, mas não supere as arquiteturas anteriores.

Com relação aos modelos de transformer, o ViT-B/16 tem a capacidade de extrair informações globais da imagem usando um mecanismo de atenção sem a necessidade de convolução. Para o conjunto de dados desta pesquisa, espera-se que ele se destaque, capturando relações complexas e interdependências entre diferentes regiões da imagem, o que pode ser bom na classificação da osteoartrite de joelho. Entretanto, para a classe KL 4, pode ser que o modelo não tenha um desempenho tão bom, dado que os ViTs dependem de muitos dados para treinar efetivamente.

O DeiT é uma abordagem mais eficiente em termos de dados de treinamento, pois ele foi projetado para ser robusto em conjuntos de dados menores. Espera-se que ele tenha um bom equilíbrio entre performance e eficiência computacional, além de apresentar resultados competitivos em relação aos modelos RNCs.

O Swin Transformer é projetado para capturar características locais e globais através de uma abordagem hierárquica de atenção, o que o torna uma boa opção para tarefas que exigem análise de características em várias escalas, como em radiografias. Devido à sua capacidade de trabalhar em diferentes níveis de granularidade, é esperado que o Swin Transformer tenha um desempenho muito competitivo, podendo até mesmo ser superior aos outros modelos de ViT e RNC.

Por fim, o ResNet50-ViT-B/16 junta as forças de RNCs e ViTs para criar uma arquitetura promissora na tarefa de classificação da OA de joelho. A combinação de extração de características locais detalhadas pelo modelo ResNet50 com a capacidade dos transformers de capturar dependências globais na imagem oferece um equilíbrio vantajoso entre precisão e generalização. Espera-se que este modelo apresente resultados superiores em comparação com modelos puramente convolucionais e modelos de transformers isolados.

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 na página 1.
- 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 na página 1.
- 4 KELLGREN, J. H.; LAWRENCE, J. S. Radiological assessment of osteo-arthritis. *Annals of the rheumatic diseases*, v. 16, 1957. ISSN 00034967. Citado na página 1.
- 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 na página 1.
- 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 2 vezes nas páginas 1 e 2.
- 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 na página 2.
- 13 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 3.

- 14 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 3.
- 15 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 6.
- 16 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 6.
- 17 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 6.
- 18 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 6.
- 19 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 2 vezes nas páginas 6 e 7.
- 20 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 7.
- 21 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 8.
- 22 PARK, S. et al. Multi-task vision transformer using low-level chest x-ray feature corpus for covid-19 diagnosis and severity quantification. *Medical Image Analysis*, v. 75, 2022. ISSN 13618423. Citado na página 8.
- 23 WANG, Y. et al. An automatic knee osteoarthritis diagnosis method based on deep learning: Data from the osteoarthritis initiative. *Journal of Healthcare Engineering*, v. 2021, 2021. ISSN 20402309. Citado na página 8.
- 24 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 10.