Classificação da osteoartrite de joelho em imagens de Raio-X por meio de Ensemble Learning

Ana Carolina Manso Silvério[1], **Alexei Manso Correa Machado**[1,2]

[1]Departamento de Ciência da Computação
Pontifícia Universidade Católica de Minas Gerais (PUC Minas), Belo Horizonte, Brasil
[2]Departamento de Anatomia e Imagem - Faculdade de Medicina
Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Brasil

acmsilverio@sga.pucminas.br, alexeimcmachado@gmail.com

Resumo. *A osteoartrite de joelho (OA) é uma doença inflamatória que pode ocasionar a deformidade da cartilagem articular, entre outros sintomas, e cujo diagnóstico preciso é essencial para predizer sua progressão. Este trabalho propõe um ensemble de redes neurais profundas como a ResNet50, Xception, Inception ResNetV2 e EfficientNet para classificar a OA a partir de imagens de raio-X, conforme os níveis estabelecidos pela escala de Kellgren Lawrence. Uma análise preliminar indica que o modelo de ensemble supera resultados obtidos por redes individuais, classificando de forma mais assertiva os diversos estágios da doença.*

Abstract. *Knee osteoarthritis (OA) is an inflammatory disease that can cause deformity of the articular cartilage, among other symptoms, and whose accurate diagnosis is essential to predict its progression. This work proposes an ensemble of deep neural networks such as ResNet50, Xception, Inception ResNetV2 and EfficientNet to classify OA from X-ray images, according to the levels established by the Kellgren Lawrence scale. A preliminary analysis indicates that the ensemble model surpasses results obtained by individual networks, providing a more assertive classification of the different stages of the disease.*

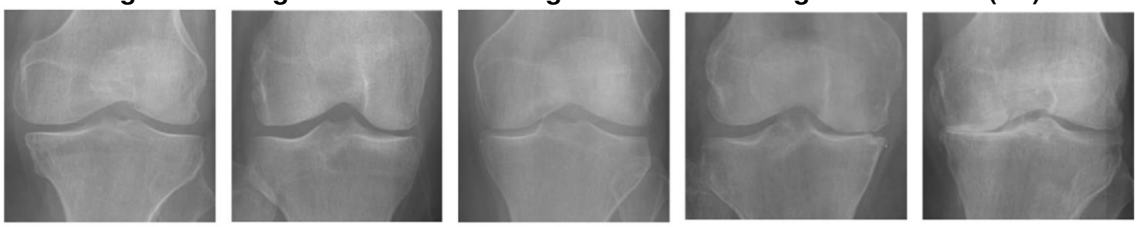
1. Introdução

A osteoartrose de joelho (OA) é uma doença inflamatória que gera o desgaste da cartilagem articular, podendo ocasionar a deformidade da articulação, dores e outros sintomas debilitantes. De acordo com a *Global Burden Disease* (Carga Global de Morbidade), a osteoartrite acomete mais de 12 milhões de brasileiros e 85% da população acima de 65 anos. Essa condição é diagnosticada por meio de raio-X de forma manual por especialistas que necessitam de treinamento específico para tal tarefa. Mesmo entre especialistas, pode haver divergências a respeito do grau de severidade da condição para um mesmo exame. Um diagnóstico precoce pode ajudar a diminuir a velocidade da progressão da doença, melhorando a qualidade de vida do paciente.

A escala Kellgren Lawrence (KL) [Kohn et al. 2016] é tida como referência para a classificação da severidade da doença, classificando-a em 5 estágios: KL 0, sem evidências de osteoartrite; KL 1, possível estreitamento do espaço articular; KL 2, presença de osteófitos definitivos e possível estreitamento do espaço articular; KL 3, múltiplos osteófitos e possível deformidade das bordas ósseas; e KL 4, grandes osteófitos

e acentuada redução do espaço articular. A Fig. 1 adaptada de [Kishore et al. 2023] exemplifica os estágios da OA em um exame de raio-X do joelho.

Figure 1. Estágios da osteoatrite segundo a escala Kellgren Lawrence (KL)



O presente trabalho tem como objetivo propor uma solução baseada em *ensemble* de métodos de visão computacional para auxiliar a detecção da osteoartrite através de imagens de raio-X. Resultados preliminares demonstram a vantagem do uso combinado de técnicas para o aumento da acurácia dos diagnósticos.

2. Trabalhos relacionados

Na última década, a aplicação de métodos baseados em redes neurais profundas tem demonstrado precisão comparável à de especialistas na classificação da osteoartrite, reduzindo o grau de variabilidade no diagnóstico [Schwartz et al. 2020, Brejnebøl et al. 2022]. Nesse contexto, foram propostas várias soluções envolvendo tipos específicos de redes neurais artificiais, cuja maioria utiliza as bases de dados da *Oarthritis Initiative* (OAI) e da *Multicenter Osteoarthritis Study* (MOST), iniciativas voltadas para a obtenção e disponibilização de dados de osteoartrite. Dentre as propostas para classificação do espaço articular femuro-tibial, destacam-se as redes convolucionais profundas (CNNs), que alcançam acurácias de até 0,7 [Chen et al. 2019]. Outras estratégias que combinam o uso da Faster R-CNN atingiram precisão na ordem de 0,8, com sensibilidade e especificidade elevadas [Liu et al. 2020]. [Norman et al. 2019] utilizam os modelos de U-Net e DenseNet para melhorar significativamente os resultados de classificação. Já em [Kwon et al. 2020], a análise biomecânica das passadas dos pacientes demonstrou ser uma valiosa adição à radiografia, elevando a acurácia da classificação KL quando combinada através de *Support Vector Machines* (SVM) e Inception-ResNet-v2. Há também soluções que envolvem a melhora da performance das redes neurais, como em [Sarvamangala and Kulkarni 2021], cuja utilização de Blocos Convolutivos Multiescala (MCBs) e técnicas de transferência de aprendizado resultou em F1-scores superiores a 0,8. Mais recentemente, foram propostas redes neurais especializadas na estimativa da distância entre as bordas dos ossos, também com F1-score na ordem de 0,8 para a detecção de OA utilizando aumento de dados via GANs [Farajzadeh et al. 2023]. Outro trabalho baseado na ResNet-50 apresentou precisão média de 0,85, evidenciando o potencial do aprendizado profundo no diagnóstico precoce da osteoartrite [Neslihan et al. 2022].

3. *Ensemble Learning*

O *Ensemble Learning* é uma técnica de aprendizado de máquina que combina vários modelos para produzir um modelo mais robusto e preciso para alcançar uma performance geral

melhor do que a que os modelos individuais poderiam alcançar. Tal abordagem é particularmente útil em situações onde um único modelo pode ser insuficiente para capturar toda a complexidade dos dados. No contexto em questão, a aplicação de *ensembles* de CNNs busca explorar a diversidade entre diferentes modelos para melhorar o diagnóstico da OA. Dentre os diversos modelos de *ensembles*, o Stacking foi escolhido como a abordagem inicial para este trabalho. Nele, as CNNs são treinadas independentemente e suas saídas são combinadas para se obterem melhores previsões.

4. Materiais e Métodos

Para os experimentos deste estudo, foi utilizada a base "Knee Osteoarthritis Severity Grading Dataset", Mendeley Data (data.mendeley.com) que contém 8260 imagens de raios-X de joelhos anotados pela escala KL. Destas, 70% foram usadas para treinamento, 10% para validação e 20% para teste.

Inicialmente foi feito o recorte da área de interesse através da técnica de correlação cruzada máxima, onde uma imagem da articulação é usada como máscara para se localizarem as demais articulações nas imagens da base de dados. Esse passo elimina partes da imagem que não são úteis para o diagnóstico. Em seguida, foram aplicadas técnicas de aumento de dados por equalização, adição de ruído e espelhamento com o objetivo de balancear as classes e mitigar o *overfitting* dos modelos. O conjunto final ficou composto por 680 imagens de cada classe.

Foram implementadas as redes ResNet50, Inception Resnet v2, Xception e EfficientNet, inicialmente treinadas com a ImageNet. Os modelos são ajustados através de re-treinamento por um número de épocas específico, com ajuste nas taxas de aprendizado. As métricas de desempenho utilizadas foram a acurácia, *recall*, precisão e F1-Score. Para a implementação, foi utilizada a linguagem Python e as bibliotecas Tensorflow, Keras, Opencv e Sklearn no ambiente Google Collaboratory.

Modelo	**Acurácia**	**F1-Score**	**Precisão**	**Recall**
ResNet50	0,62	0,63	0,63	0,63
XCeption	0,62	0,60	0,61	0,61
Inception	0,65	0,64	0,66	0,65
EfficientNet	0,62	0,62	0,62	0,62
Ensemble	0,70	0,65	0,66	0,67

Table 1. Métricas de performance dos modelos

5. Resultados Experimentais

A ResNet50 foi treinada utilizando-se diversas configurações, visando a otimização do desempenho da rede. Foram variados o tamanho de lote, taxa de aprendizado e aplicadas funções de perda tanto personalizadas, quanto as tradicionais Huber, Hinge, e Log Rank, ao longo de 16 épocas de treinamento com os otimizadores RMSprop e Adam. Os melhores resultados foram obtidos com o *batch size* de 256 e entropia cruzada categórica. Foi adicionado o Global Average Pooling 2D, além da introdução de um fator de *dropout* de 0,2 para combater o *overfitting*. O otimizador Adam nesse cenário alcançou acurácia de 0,62 como mostrado na Tabela 1. Observou-se o melhor desempenho para a classe 4, com

uma taxa de acerto de 94%. No entanto, observa-se uma confusão entre as classes 0 e 1, o que é comum mesmo na classificação realizada por especialistas.

O segundo grupo de experimentos utilizou a rede XCeption com *batch size* de 32 ao longo de 25 épocas. Após o teste de diversas configurações, os melhores resultados foram obtidos com Global Average Pooling 2D, fator de *dropout* de 0.5 e otimizador Adam, alcançando uma acurácia semelhante à da ResNet50 (0,62). Uma análise de resultados indica que o modelo Xception também apresenta melhores desempenhos na identificação de estágios mais severos de osteoartrite, mas também para a classe 0, com elevado valor de *recall*. Em contrapartida, apresentou dificuldades na classificação da classe 1, sendo confundida com a classe 0.

A tendência de melhora de resultados com o tamanho de lote de 256, 50 épocas de treinamento e a função de entropia cruzada categórica com otimizador Adam motivou a manutenção desses hiper-parâmetros no teste dos demais modelos. No caso da arquitetura Inception, foi adicionada 1 camada densa de GlobalAvarege Pooling 2D com uma taxa de *dropout* de 0,2. Os resultados obtidos estão representados na Tabela 1. O modelo atingiu acurácia de 0,65 com precisão moderada de 0,64, demonstrando uma capacidade relativamente uniforme de lidar com as diferentes classes de forma equilibrada. Além disso, o modelo se mostrou eficaz na classificação da classe 0, como demonstrado por um *recall* de 0,89, e manteve um alto grau de precisão para a classe 4 com um F1-Score de 0,93.

O último modelo testado individualmente foi o EfficientNet, onde os melhores hiper-parâmetros foram consistentes com os usados nas demais arquiteturas. A acurácia obtida (0,62) foi semelhante à dos demais modelos, conforme mostrado na Tabela 1. Esse valor indica uma capacidade geral de classificação que é comparável à do modelo Inception e ligeiramente abaixo do ResNet50, com a manutenção de um desempenho consistente entre todas as classes, sem viés significativo. Ao se observarem as métricas individuais por classe, o EfficientNet exibe uma precisão particularmente forte na classe 3, com um *recall* de 0,90 e alto F1-Score (0,87) para a classe 4.

Uma comparação entre os modelos testados revela um desempenho médio semelhante, mas pequenas diferenças no tratamento de classes específicas. A ResNet50 destaca-se pelo seu desempenho equilibrado em todas as classes, com um F1-Score que se destaca na classe 4, o que o torna uma escolha adequada para tarefas que requerem alta precisão em condições mais severas ou avançadas da doença. Já o Xception demonstrou maior valor de *recall* para a classe 0, o que é importante para a detecção precoce da doença. O Inception ResNetV2 mostra uma leve vantagem em relação ao equilíbrio entre as classes detectadas sobre os outros modelos. Por fim, o EfficientNet tem um desempenho notável na classe 3, com um *recall* de 0,90.

A grande motivação do uso de *ensembles* é associar métodos que se comportam de forma distinta para classes distintas, de modo que as principais vantagens de cada modelo podem ser associadas. Além disso, a redundância positiva entre os modelos reforça a confiança nas classificações corretas, enquanto as diferenças entre eles podem ser exploradas para se alcançar uma compreensão mais refinada das características de cada estágio da osteoartrite. Neste trabalho, foi proposto um *ensemble* formado pelas quatro redes neurais analisadas — ResNet50, Xception, Inception ResNet V2 e EfficientNet cujos re-

sultados são mostrados na Tabela 1. A proposta alcançou acurácia de 0,70, uma melhoria de até 8 pontos percentuais em relação aos modelos individuais, o que evidencia a vantagem de se combinarem as previsões. A matriz de confusão mostrada na Tabela 2 reflete uma acurácia balanceada entre as classes, à exceção da classe 1, que ainda é confundida com a classe 0. Também é possível verificar que as predições incorretas do modelo de *ensemble* tendem a ser entre classes adjacentes. Tal comportamento demostra-se extremamente adequado para o contexto em questão, no qual um erro para uma classe muito distante poderia significar um erro diagnóstico grave.

	KL0	**KL1**	**KL2**	**KL3**	**KL4**
KL0	0,87	0,13	0	0	0
KL1	0,60	0,19	0,19	0,02	0
KL2	0,20	0,13	0,53	0,13	0
KL3	0	0,02	0,08	0,83	0,06
KL4	0	0	0	0,10	0,90

Table 2. Matriz de confusão Real X Predita para o *ensemble*

Técnicas de imagem comuns, como as radiografias, têm limitações significativas, visto que podem não detectar alterações precoces na cartilagem, que é onde começam as primeiras manifestações da doença, e muitas vezes não refletem com precisão os sintomas do paciente. Tal aspecto se torna ainda mais desafiador quando são aplicadas técnicas de aprendizado de máquina como as CNNs. A semelhança entre as características radiográficas dos níveis KL 0 e 1 pode confundir os modelos de inteligência artificial, afetando a precisão da classificação.

Um fator a se destacar é de que ao se comparar o modelo de *ensemble* com os modelos individuais, nota-se que ele consegue manter ou superar a performance dos modelos individuais em todas as classes. Exemplo disso é que o *ensemble* mostra um *recall* de 0,87 na classe 0, o que é uma melhoria considerável em comparação com modelos como o EfficientNet. Além disso, a precisão do *ensemble* na classe 4 é de 0,94, refletindo um alto grau de precisão em identificar os estágios mais avançados. Dessa forma, o modelo de *ensemble* não só melhora a precisão global, como também minimiza os erros críticos de classificação em estágios distantes da doença, demonstrando ser uma abordagem mais robusta e confiável, além de sua capacidade de combinar as forças dos modelos individuais ter se mostrado um ponto forte para sua escolha como uma ferramenta diagnóstica.

6. Conclusão

Neste trabalho foram apresentados os resultados preliminares de um sistema de diagnóstico assistido da osteoartrite de joelho que utilizou modelos profundos de aprendizado de máquina. Foram analisadas as redes ResNet50, Xception, Inception ResNetV2 e EfficientNet, bem como uma combinação dos seus resultados através de um *ensemble*. Ao final da análise, o modelo de *ensemble* se destacou devido à sua boa precisão média e individual para cada classe, minimizando erros críticos ao evitar confusões entre classes distantes, o que é essencial para um diagnóstico preciso da doença.

A continuação deste trabalho prevê a realização de estudos adicionais que utilizem novos modelos e diferentes técnicas de *ensemble*, com o objetivo de melhorar os resultados preliminares obtidos. Em especial, deseja-se diminuir os erros de classificação

entre as classes 0 e 1. Além disso, a utilização de modelos para o aumento da base de dados deve ser investigada, como o produzido por Redes Generativas Adversárias de modo a mitigar o problema de escassez de dados rotulados e o desbalanceamento entre as classes.

Agradecimentos — AMC Machado agradece o auxílio financeiro do Fundo de Incentivo à Pesquisa FIP-PUCMinas 2025/32467 e da FAPEMIG através dos projetos APQ-02753-24 e APQ-06556-24.

References

Brejnebøl, M., Hansen, P., J., N., Bachmann, R., Ratjen, U., Hansen, I., Lenskjold, A., Axelsen, M., Lundemann, M., and Boesen, M. (2022). External validation of an artificial intelligence tool for radiographic knee osteoarthritis severity classification. *European Journal of Radiology*, 150:110249.

Chen, P., Gao, L., Shi, X., Allen, K., and L., Y. (2019). Fully automatic knee osteoarthritis severity grading using deep neural networks with a novel ordinal loss. *COMPUT MED IMAG GRAP*, 75:84–92.

Farajzadeh, N., Sadeghzadeh, N., and Hashemzadeh, M. (2023). Ijes-oa net: A residual neural network to classify knee osteoarthritis from radiographic images based on the edges of the intra-joint spaces. *MED ENG PHYS*, 113:103957.

Kishore, V., Kalpana, V., and Kumar, H. (2023). Evaluating the efficacy of deep learning models for knee osteoarthritis prediction based on kellgren-lawrence grading system. *e-Prime - Advances in Electrical Engineering, Electronics and Energy*, 5:100266.

Kohn, M., Sassoon, A., and Fernando, D. (2016). Classifications in brief: Kellgren-lawrence classification of osteoarthritis. *CLIN ORTHOP RELAT R*, 474(8):1886–1893.

Kwon, S., Han, H., Lee, M., Kim, H., Ku, Y., and Ro, D. (2020). Machine learning-based automatic classification of knee osteoarthritis severity using gait data and radiographic images. *IEEE ACCESS*, 8:120597–603.

Liu, B., Luo, J., and Huang, H. (2020). Toward automatic quantification of knee osteoarthritis severity using improved faster r-cnn. *INT J COMPUT ASS RAD*, 15:457–466.

Neslihan, B., Miika, N., and Simo, S. (2022). Machine learning based texture analysis of patella from x-rays for detecting patellofemoral osteoarthritis. *INT J MED INFORM*, 157:104627–104627.

Norman, B., Pedoia, V., Noworolski, A., Noworolski, A., Link, T., and Majumdar, S. (2019). Applying densely connected convolutional neural networks for staging osteoarthritis severity from plain radiographs. *J DIGIT IMAGING*, 32:471–477.

Sarvamangala, D. and Kulkarni, V. (2021). Grading of knee osteoarthritis using convolutional neural networks. *NEURAL PROCESS LETT*, 53(4):2985–3009.

Schwartz, A., Clarke, H., Spangehl, M., Bingham, J., Etzioni, D., and Neville, M. (2020). Can a convolutional neural network classify knee osteoarthritis on plain radiographs as accurately as fellowship-trained knee arthroplasty surgeons? *J ARTHROPLASTY*, 35:2423–2428.