# Avaliação de Técnicas de Aumento de Dados para Classificação Binária de Imagens Volumétricas de Ressonância Magnética do Joelho

**Thalles C. Fontainha**[1], **Felipe da R. Henriques**[1,2], **Amaro A. Lima**[1,3], **Gabriel M. Araujo**[4], **Ricardo de S. Tesch**[5]

[1]Programa de Pós-Graduação em Instrumentação e Óptica Aplicada – CEFET/RJ
[2]Programa de Pós-Graduação em Ciência da Computação – CEFET/RJ
[3]Departamento de Telecomunicações – CEFET/RJ, Campus Nova Iguaçu
[4]Programa de Pós-Graduação em Engenharia Elétrica – CEFET/RJ
[5]Departamento de Medicina Regenerativa – UNIFASE

thalles.fontainha@aluno.cefet-rj.br, {felipe.henriques, amaro.lima, gabriel.araujo}@cefet-rj.br, ricardotesch@prof.unifase-rj.edu.br

***Abstract.*** *Knee osteoarthritis leads to pain and reduced function. We assessed data augmentation for binary classification of volumetric knee Magnetic Resonance Imaging (MRI) using double-echo steady-state (DESS) and R3D-18. Ten experiments (EX1–EX10) used 5-fold stratified cross-validation; augmentation was training-only. We compared transfer learning, Automatic Mixed Precision (AMP), and spatial/radiometric transforms. Pretrained initialization (EX1 vs. EX2) had the greatest effect. AMP ablation (EX3) required adjustments due to graphics processing unit (GPU) video memory (VRAM) limits, demonstrating feasibility rather than a controlled comparison. Peak area under the receiver operating characteristic (ROC) curve (AUC) was 0.90 in EX8 and EX10.*

***Resumo.*** *A osteoartrite de joelho causa dor e perda funcional. Avaliamos aumento de dados para classificação binária de volumes de ressonância magnética na sequência double-echo steady-state (DESS) com o modelo R3D-18, em dez experimentos (EX1–EX10) com validação cruzada estratificada 5-fold e aumento de dados apenas no treino. Comparamos aprendizado por transferência (transfer learning), precisão mista automática (AMP) e transformações espaciais/radiométricas. A inicialização pré-treinada (EX1 vs. EX2) teve o maior efeito. A ablação da AMP (EX3) exigiu ajustes por limite de memória de vídeo (VRAM), demonstrando viabilidade, em vez de uma comparação controlada. O pico da AUC (área sob a curva ROC) foi 0,90 em EX8 e EX10.*

## 1. Introdução

O diagnóstico preciso da osteoartrite (OA) do joelho por imagens de ressonância magnética (do inglês, *Magnetic Resonance Imaging* - MRI) é relevante para o manejo clínico e o acompanhamento da progressão da doença [Peterfy et al. 2008]. Nos últimos anos, modelos de aprendizado profundo têm mostrado potencial para automatizar a identificação de padrões associados à OA em exames de imagem [Guida et al. 2021, Yeoh et al. 2023]. Contudo, o treinamento consistente desses modelos permanece desafiador devido à limitação de dados anotados, ao desbalanceamento de classes e à variabilidade inerente aos protocolos e condições de aquisição [Chlap et al. 2021, Shorten and Khoshgoftaar 2019].

Nesse contexto, o aumento de dados (do inglês, *data augmentation*) é uma estratégia amplamente adotada para aumentar o volume, o tamanho, a quantidade de elementos, e a diversidade do conjunto de treino e melhorar a generalização. Em dados volumétricos 3D, entretanto, as transformações devem preservar estruturas anatômicas e manter a plausibilidade clínica. Por isso, transformações geométricas (por exemplo, rotação e translação) e radiométricas (por exemplo, ruído, contraste e gama) são geralmente parametrizadas em intervalos compatíveis com variações realísticas de posicionamento e aquisição [Shorten and Khoshgoftaar 2019, Chlap et al. 2021, Islam et al. 2024]. Apesar do uso disseminado dessas técnicas, há poucos estudos comparativos controlados em imagens tridimensionais de ressonância magnética (MRI 3D) que isolem, de forma reprodutível, o efeito de escolhas de treinamento e do tipo/intensidade do aumento de dados sob validação estrita, sem vazamento entre treino e avaliação (*data leakage*). Estudos dessa natureza tendem a fortalecer as recomendações práticas para a área.

Nesse contexto, este trabalho conduz uma avaliação do uso de aumento de dados para classificação binária de imagens volumétricas de ressonância magnética do joelho na sequência 3D DESS, utilizando dados da *Osteoarthritis Initiative* (OAI) [Peterfy et al. 2008, NIMH Data Archive 2026]. Considerou-se uma tarefa de classificação binária (*negative/positive*) baseada em severidade radiográfica de OA, seguindo a escala de *Kellgren–Lawrence* (KL) em que os rótulos são derivados dos graus de Kellgren–Lawrence (KL) [Kellgren et al. 1957, Kohn et al. 2016] obtidos a partir de radiografias (*X-ray*) correspondentes, de modo que a variável-alvo reflete a severidade radiográfica de OA, enquanto as entradas do modelo são os volumes 3D DESS. As principais contribuições deste trabalho são:

- Protocolo reprodutível com validação cruzada estratificada em 5 *folds* e divisão interna de validação (*inner_val*) para *early stopping*, com separação estrita entre treino, validação e teste. O aumento de dados é aplicado apenas no treino para evitar vazamento; além disso, as predições por *fold* são salvas para permitir a reconstrução consistente da curva ROC e do valor de AUC-ROC.
- Avaliação de transformações espaciais e radiométricas, incluindo uma varredura controlada de intensidade do aumento de dados (EX7–EX10), com evidências quantitativas para apoiar escolhas de *pipeline* [Shorten and Khoshgoftaar 2019, Chlap et al. 2021].

### 1.1. Trabalhos relacionados

A classificação automática de osteoartrite (OA) em volumes de MRI tem sido objeto de investigações recentes, incluindo estudos preliminares dos autores que exploraram o desempenho da arquitetura R3D-18 para essa tarefa [Fontainha et al. 2025]. Além desses resultados iniciais, a literatura apresenta diversas abordagens para a classificação de OA, abrangendo desde representações 2D até modelos 3D que capturam a informação volumétrica integral [Guida et al. 2021, Yeoh et al. 2023]. Guida et al. [Guida et al. 2021], por exemplo, analisaram o emprego de CNNs 3D para identificar padrões estruturais em sequências de MRI, enquanto Yeoh et al. [Yeoh et al. 2023] avaliaram o impacto de técnicas de *transfer learning* nessas arquiteturas, sugerindo que a inicialização de pesos pode ser um fator relevante em contextos com disponibilidade limitada de dados.

Quanto ao aumento de dados, revisões compilam transformações geométricas e fotométricas e discutem seus efeitos em tarefas clínicas, enfatizando a necessidade de manter a plausibilidade anatômica, isto é, aplicar transformações

que preservem a geometria e as relações espaciais das estruturas do joelho, sem introduzir artefatos ou deformações incompatíveis com variações realistas de aquisição [Shorten and Khoshgoftaar 2019, Chlap et al. 2021].

Além das revisões, investigações estruturadas sobre o papel do aumento de dados em imagens clínicas são relevantes para embasar escolhas metodológicas. Destaca-se o estudo de [Krinski et al. 2022], que avaliou 20 técnicas de aumento de dados na segmentação semântica de tomografias computadorizadas (CT) de COVID-19 e, a partir de mais de 3.000 experimentos em cinco *datasets*, indicou que transformações espaciais (como *Elastic Transform*, *Grid Distortion* e *Rotate*) tendem a ser as mais promissoras para redes codificador–decodificador, especialmente em cenários com alto desbalanceamento de classes.

De forma complementar, embora em outro domínio, [Scalercio and Freitas 2023] propõem técnicas de aumento de dados para texto baseadas em transformações sintáticas e realizam uma avaliação intrínseca da qualidade das amostras geradas, destacando a importância de preservar propriedades fundamentais dos dados durante o aumento. Essa preocupação é análoga ao requisito, em imagens médicas, da manutenção da plausibilidade anatômica ao aplicar transformações. Em contraste com [Krinski et al. 2022], voltado a CT (2D) e segmentação, este trabalho avalia o aumento de dados em um cenário distinto: a classificação binária de osteoartrite em volumes de MRI 3D; assim, sob protocolo controlado, investigamos como o tipo e a intensidade do aumento (geométrico e radiométrico) influenciam a generalização em uma arquitetura 3D para classificação.

Apesar desses avanços, ainda é relativamente menos comum encontrar avaliações controladas em dados volumétricos 3D que isolem, de forma reprodutível, fatores de treinamento (como *transfer learning* e AMP) e tipos/intensidades de aumento de dados. Nesse cenário, o presente trabalho contribui com um protocolo estruturado (EX1 até EX10) que usa validação cruzada estratificada, com aumento de dados restrito ao treino, além da persistência de predições por *fold* para reconstrução consistente de ROC/AUC.

## 2. Materiais e Métodos

### 2.1. Base de dados, tarefa e pré-processamento

Os experimentos utilizaram dados da *Osteoarthritis Initiative* (OAI) [Peterfy et al. 2008, NIMH Data Archive 2026], contendo imagens volumétricas de ressonância magnética (MRI) do joelho na sequência 3D DESS adquiridos no tempo basal (*baseline*). Formulou-se uma tarefa de classificação binária (*negative*/*positive*), na qual os rótulos foram derivados dos graus de *Kellgren–Lawrence* (KL) obtidos a partir de radiografias correspondentes do conjunto de raios-X da própria OAI. Dessa forma, a variável-alvo refletiu a severidade radiográfica de OA (via KL), enquanto as entradas do modelo corresponderam aos volumes de MRI 3D DESS.

Os volumes 3D DESS utilizados foram obtidos a partir de uma versão pública organizada no Kaggle [Berrimi 2022], que disponibiliza a extração do OAI em dois arquivos *NumPy*: `normal-3DESS-128-64.npy` e `abnormal-3DESS-128-64.npy`. Neste trabalho, esses arquivos foram mapeados para as classes *negative* e *positive*, respectivamente, e manteve-se a nomenclatura original dos arquivos por compatibilidade com os scripts de carregamento. O conjunto total empregado contém 2.976 volumes (1.659 do tipo *negative* e 1.317 do tipo *positive*), caracterizando desbalanceamento moderado entre classes (proporção aproximada de 1,26:1).

Para reduzir variações de escala entre exames e tornar os experimentos comparáveis, aplicou-se normalização *min-max por volume*, mapeando as intensidades para o intervalo $[0, 1]$. Essa normalização foi aplicada antes de quaisquer operações de aumento de dados e foi mantida idêntica em todos os experimentos, sendo calculada *por volume* (isto é, $\min$ e $\max$ são obtidos do próprio volume no momento do carregamento), sem estimar parâmetros a partir do conjunto completo. Em outras palavras, trata-se de uma normalização auto-contida por amostra, e não de uma normalização global *fit* no treino e depois aplicada às demais partições. Os dados não foram redistribuídos neste trabalho ou no repositório associado devido a restrições de tamanho e licenciamento. Assim, a reprodutibilidade dos experimentos dependeu do *download* da mesma fonte [Berrimi 2022] e da manutenção dos nomes e formatos esperados pelos scripts.

## 2.2. Protocolo de avaliação e controle experimental

Para estimar desempenho de forma robusta e evitar vazamento de dados (*data leakage*), adotou-se validação cruzada estratificada em 5 *folds* ($k = 5$, semente 42). Em cada iteração, o conjunto total ($N = 2.976$ volumes) é particionado em $80\%$ para desenvolvimento (*outer_train*, $n \approx 2.380$) e $20\%$ para teste final isolado (*outer_test*, $n \approx 596$). O *outer_test* permanece estritamente isolado de todas as decisões de hiperparâmetros e treinamento, sendo utilizado apenas na avaliação final.

Em cada iteração da validação cruzada estratificada, o conjunto de desenvolvimento (*outer_train*, $n \approx 2.380$) foi particionado internamente em subconjuntos de treinamento efetivo (*inner_train*, $85\%$, $n \approx 2.023$) e validação interna (*inner_val*, $15\%$, $n \approx 357$). A validação interna foi utilizada exclusivamente para monitorar a convergência do modelo e definir a *best_epoch* por *early stopping*, sem participação direta na atualização dos pesos. Assim, em cada *fold*, manteve-se a separação entre treinamento efetivo, validação interna e teste externo, reduzindo o risco de vazamento de informação.

O aumento de dados foi aplicado apenas ao *inner_train*, de forma *on-line*, por meio da função `aug_fn`. Já os conjuntos *inner_val* e *outer_test* permaneceram inalterados, utilizando um `clean_dataset` sem qualquer transformação, de modo a garantir uma avaliação imparcial do desempenho.

Em termos operacionais, cada *fold* seguiu o fluxo: (i) particionamento estratificado em *outer_train* e *outer_test*; (ii) particionamento interno do *outer_train* em *inner_train* e *inner_val*; (iii) treinamento com aumento aplicado exclusivamente ao *inner_train*; (iv) monitoramento da perda no *inner_val* para definição da *best_epoch*; (v) reinicialização do modelo e novo treinamento no conjunto *outer_train* completo pelo número fixo de épocas definido anteriormente; e (vi) avaliação final no *outer_test*.

Para mitigar o desbalanceamento de classes sem introduzir redundância física de dados, não se utilizou *oversampling* nos experimentos do Projeto 2.0 (`oversample=False` em EX1–EX10). Em vez disso, utilizou-se a função de perda `BCEWithLogitsLoss` com o parâmetro `pos_weight` calculado automaticamente a partir da razão entre classes negativas e positivas do conjunto de treino. Por fim, para rastreabilidade e auditoria, as predições (`y_prob`) e os rótulos verdadeiros (`y_true`) de cada *fold* foram persistidos em arquivos `.npy`, permitindo a reconstrução consistente de métricas e curvas ROC a partir de dados brutos.

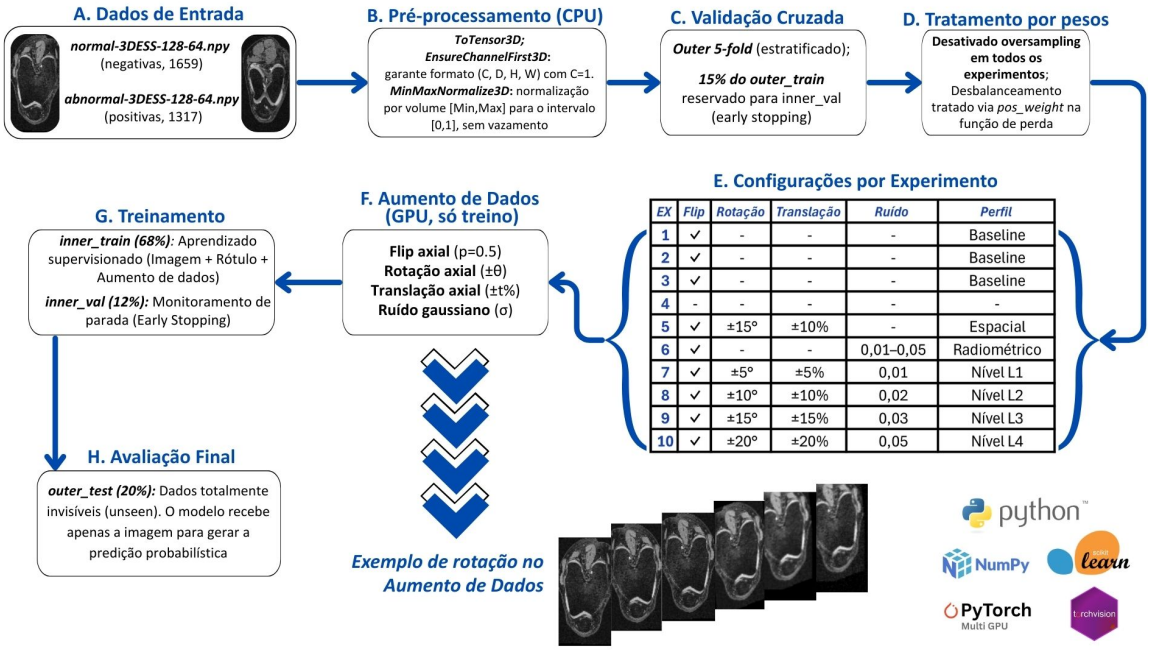


**Figura 1. Fluxograma do pipeline experimental. Fonte: elaboração própria.**

## 2.3. Arquitetura do modelo e configuração de treinamento

A arquitetura empregada foi a R3D-18, com variante pré-treinada no Kinetics-400 quando aplicável [Kay et al. 2017, Hara et al. 2018]. O treinamento utilizou otimizador Adam com taxa de aprendizado de $1 \times 10^{-4}$ [Kingma and Ba 2014] e função de perda `BCEWithLogitsLoss` com `pos_weight` definida automaticamente a partir da razão entre classes. Em todos os cenários, aplicou-se *early stopping* com paciência 15 e limite de 60 épocas (parâmetros definidos no `ExperimentConfig`). O AMP (*Automatic Mixed Precision*) é gerenciado centralmente pelo orquestrador e aplicado tanto no treinamento quanto na inferência por meio das funções `autocast` e `GradScaler`, visando otimizar o uso de memória de vídeo (VRAM) e reduzir o tempo de processamento das épocas. Ele foi ativado em todos os experimentos, com exceção do EX3, que serviu como ablação para avaliar o impacto computacional e numérico do AMP.

## 2.4. Configurações experimentais e aumento de dados

Os experimentos EX1, EX2 e EX3 isolaram fatores de treinamento: (EX1) - R3D-18 pré-treinada com AMP; (EX2) - treinamento do zero (sem *transfer learning*); e (EX3) - pré-treinado sem AMP (treinamento em `float32`). Os experimentos EX4, EX5, EX6 e EX7 a EX10 avaliaram o impacto do uso de aumento de dados: (EX4) - sem aumento de dados; (EX5) - transformações espaciais (incluindo rotação e translação); (EX6) - transformações radiométricas (ruído, contraste e gama); e (EX7 a EX10) - varredura de intensidade em quatro níveis (L1 a L4) usando apenas transformações espaciais e ruído, com parâmetros escalonados. No EX7–EX10, realizou-se uma varredura (*sweep*) de intensidade em quatro níveis (L1, L2, L3 e L4), isto é, os mesmos tipos de transformação foram mantidos (flip, rotação, translação e ruído), e apenas a magnitude dos parâmetros foi escalonada por nível, mantendo-se o mesmo protocolo experimental e variando-se apenas rotação, translação e ruído (sem aplicação de ajuste de contraste e correção gama). Os valores adotados para cada nível (L1: ±5°, ±5%, $\sigma = 0{,}01$; L2: ±10°, ±10%, $\sigma = 0{,}02$;

L3: ±15°, ±15%, $\sigma = 0{,}03$; L4: ±20°, ±20%, $\sigma = 0{,}05$) foram definidos para representar uma escala monotônica de intensidade (leve→muito forte), mantendo fixos os tipos de transformação e variando apenas suas magnitudes. Essa escolha visa quantificar o efeito da *intensidade global* do aumento de dados sob custo computacional controlado, sem realizar uma busca combinatória completa de hiperparâmetros. Os limites foram escolhidos com base em inspeção qualitativa de plausibilidade anatômica e em recomendações gerais da literatura para transformações moderadas em imagens médicas [Shorten and Khoshgoftaar 2019, Chlap et al. 2021].

As transformações espaciais (rotação axial e translação) foram escolhidas por simularem pequenas variações no posicionamento do joelho durante a aquisição, sendo comuns em *pipelines* de aumento de dados para imagens médicas [Ronneberger et al. 2015, Shorten and Khoshgoftaar 2019]. As transformações radiométricas (ruído gaussiano, ajuste de contraste e correção gama) modelaram variações nos equipamentos de MRI e nos protocolos de aquisição, aumentando a estabilidade do modelo a diferentes condições de imagem [Chlap et al. 2021]. Todos os parâmetros (ângulos, deslocamentos e níveis de ruído) foram definidos explicitamente e controlados por experimento, e, no EX7–EX10, a intensidade das transformações foi escalonada em quatro níveis (L1–L4) para quantificar o efeito da magnitude do aumento, garantindo que as imagens aumentadas permanecessem anatomicamente plausíveis. As transformações geométricas foram implementadas via `affine_grid` e `grid_sample` do PyTorch [PyTorch 2026], e executadas na GPU durante o treino para evitar gargalo de CPU.

### 2.5. Delineamento experimental, infraestrutura e reprodutibilidade

Para avaliar de forma abrangente o impacto do treinamento e das estratégias de aumento de dados em imagens volumétricas de ressonância magnética, foram conduzidos dez experimentos independentes em máquinas equipadas com duas GPUs NVIDIA GeForce RTX 2080 Ti (12 GB cada), utilizando Python 3.10.19 e PyTorch 2.0.1 (cu118) com suporte CUDA habilitado (CUDA 11.8) e cuDNN 8.9.5.

Os códigos-fonte correspondentes (`EX1.py` a `EX10.py`) seguiram a mesma arquitetura modular, diferenciando-se exclusivamente pela configuração de inicialização do modelo, pelo uso de AMP (exceto EX3) e pelo *pipeline* de aumento de dados. A configuração é centralizada na classe `ExperimentConfig`, garantindo consistência entre experimentos. A arquitetura empregada foi a R3D-18, com a variante pré-treinada no Kinetics-400 [Kay et al. 2017, Hara et al. 2018] utilizada quando aplicável.

O código-fonte, os logs e os artefatos dos experimentos foram organizados em repositório público do projeto, com registros por *fold* para auditoria. O *dataset* não foi redistribuído por restrições de tamanho/licenciamento, e o treinamento médio variou entre 29 e 122 minutos por *fold*.

## 3. Resultados e Discussão

A Tabela 1 apresenta a matriz experimental e os resultados médios de AUC-ROC para os dez experimentos. Para orientar a interpretação, adotam-se duas referências principais: o *baseline* (EX1), que combina inicialização pré-treinada, AMP e apenas *flip* axial, e a condição sem aumento (EX4), utilizada como controle para quantificar o ganho marginal do *data augmentation*. A partir dessas referências, analisam-se, em sequência, as ablações de treinamento (EX2 e EX3), o efeito do tipo de aumento (EX5 e EX6) e a varredura de intensidade (EX7–EX10).

**Tabela 1. Matriz experimental de aumento de dados estocástico e resultados (AUC-ROC; média em 5 *folds*).**

| **Dimensão** | **EX1** | **EX2** | **EX3** | **EX4** | **EX5** | **EX6** | **EX7** | **EX8** | **EX9** | **EX10** |
|---|---|---|---|---|---|---|---|---|---|---|
| Propósito | Estabelecer baseline com R3D-18 pré-treinado | Medir impacto do transfer learning | Medir impacto computacional do AMP (tempo / VRAM)[1] | Quantificar ganho do aumento de dados | Avaliar aumento espacial (rotação / translação / flip) | Avaliar aumento radiométrico (ruído, contraste, gama) | Iniciar sweep de intensidade leve (L1) | Sweep com aumento moderado (L2) | Sweep com aumento forte (L3) | Sweep com aumento muito forte (L4) |
| Inicialização | Pré-treinado | Aleatória | Pré-treinado | Pré-treinado | Pré-treinado | Pré-treinado | Pré-treinado | Pré-treinado | Pré-treinado | Pré-treinado |
| Aumento de dados | Flip | Flip | Flip | Nenhum | Espacial | Radiométrico | L1[2] | L2[2] | L3[2] | L4[2] |
| Flip (axial) | ✓ | ✓ | ✓ | — | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Rotação axial (graus) | — | — | — | — | $\pm 15°$ | $\pm 15°$ | $\pm 5°$ | $\pm 10°$ | $\pm 15°$ | $\pm 20°$ |
| Translação axial (%) | — | — | — | — | $\pm 10\%$ | $\pm 10\%$ | $\pm 5\%$ | $\pm 10\%$ | $\pm 15\%$ | $\pm 20\%$ |
| Aplicação geométrica ($p$) | — | — | — | — | 0,5 | 0,5 | 0,5 | 0,5 | 0,5 | 0,5 |
| Ruído gaussiano ($\sigma$) | — | — | — | — | — | $[0{,}01, 0{,}05]$ | 0,01 | 0,02 | 0,03 | 0,05 |
| Contraste (fator) | — | — | — | — | — | $[0{,}8, 1{,}2]$ | — | — | — | — |
| Gama (fator) | — | — | — | — | — | $[0{,}7, 1{,}5]$ | — | — | — | — |
| Aplicação radiométrica ($p$) | — | — | — | — | — | 0,5 | — | — | — | — |
| **AUC-ROC (%)** | $88{,}49 \pm 1{,}32$ | $76{,}48 \pm 1{,}76$ | $55{,}20 \pm 14{,}80$ | $89{,}34 \pm 1{,}86$ | $88{,}92 \pm 1{,}55$ | $88{,}43 \pm 1{,}93$ | $88{,}20 \pm 1{,}24$ | $90{,}08 \pm 1{,}26$ | $89{,}47 \pm 1{,}14$ | $89{,}68 \pm 1{,}33$ |
| **IC95%** | [86,86; 90,13] | [74,30; 78,66] | [36,8; 73,6] | [87,03; 91,65] | [86,99; 90,85] | [86,03; 90,83] | [86,66; 89,74] | [88,51; 91,65] | [88,06; 90,88] | [88,03; 91,33] |

**Notas:** [1]No EX3, a execução em FP32 exigiu ajustes por limitação de VRAM (erro *out-of-memory*), com redução do *batch size* (32→2) e desativação do *multi-GPU*. [2](rotação, translação e ruído). Parâmetros em colchetes indicam intervalos de amostragem uniforme; $p$ é a probabilidade de aplicação por amostra. "—" indica componente não utilizado. Os intervalos de confiança (95%) foram calculados assumindo distribuição $t$ de Student com 4 graus de liberdade.

Como os intervalos de confiança (IC95%) das médias apresentados na Tabela 1 podem se sobrepor entre algumas condições, complementa-se a análise com um teste estatístico que avalia diferenças estruturadas entre experimentos ao longo dos mesmos *folds*. Define-se $\Delta$AUC (em pontos percentuais, pp) como $\Delta$AUC=AUC(*primeiro*)$-$AUC(*segundo*); assim, $\Delta$AUC$> 0$ indica vantagem da primeira condição, e $\Delta$AUC$< 0$ indica desvantagem. A Tabela 2 reporta o teste $t$ [Montgomery 2017] pareado bicaudal aplicado às AUCs obtidas em validação cruzada estratificada ($n = 5$ *folds*), comparando pares de condições de interesse. O teste é pareado porque cada comparação utiliza as AUCs calculadas nos *mesmos folds*, avaliando a hipótese nula de diferença média zero entre condições. As comparações foram organizadas em cinco blocos planejados: (1) ablações de treinamento, (2) efeito sequencial do tipo de aumento, (3) varredura de intensidade em relação ao *baseline* (EX1), (4) varredura de intensidade em relação à condição sem aumento (EX4) e (5) contraste entre os níveis extremos da varredura. Essa estrutura permite responder de forma sistemática às principais questões sobre o impacto do *transfer learning*, da presença e do tipo de aumento, e da intensidade das transformações.

**Tabela 2. Teste $t$ pareado bicaudal por *fold* ($n = 5$) comparando AUC-ROC entre condições. $\Delta$AUC (pp) = AUC(primeiro) $-$ AUC(segundo). Valores negativos indicam desvantagem da primeira condição; positivos, vantagem.**

| Comparação | $\Delta$AUC (pp) | $t$ | $p$ |
|---|---|---|---|
| **1. Ablações de treinamento** | | | |
| EX2 *vs.* EX1 (sem pré-treino) | -12,02 | -18,45 | $< 0,0001$ |
| EX3 *vs.* EX1 (sem AMP) | -33,30 | -5,02 | $0,007$ |
| **2. Efeito do tipo de aumento (sequencial)** | | | |
| EX4 *vs.* EX1 (sem aumento *vs.* baseline) | +0,84 | 1,25 | 0,278 |
| EX5 *vs.* EX4 (espacial *vs.* sem aumento) | -0,42 | -0,48 | 0,659 |
| EX6 *vs.* EX5 (radiométrico *vs.* espacial) | -0,49 | -0,43 | 0,692 |
| **3. Varredura de intensidade – comparação com baseline** | | | |
| EX7 *vs.* EX1 (L1 – leve) | -0,30 | -0,41 | 0,702 |
| EX8 *vs.* EX1 (L2 – moderado) | +1,58 | 2,12 | 0,101 |
| EX9 *vs.* EX1 (L3 – forte) | +0,97 | 1,31 | 0,261 |
| EX10 *vs.* EX1 (L4 – muito forte) | +1,18 | 1,65 | 0,174 |
| **4. Varredura de intensidade – comparação com sem aumento (EX4)** | | | |
| EX7 *vs.* EX4 (L1 *vs.* sem aumento) | -1,14 | -1,20 | 0,297 |
| EX8 *vs.* EX4 (L2 *vs.* sem aumento) | +0,74 | +0,90 | 0,419 |
| EX9 *vs.* EX4 (L3 *vs.* sem aumento) | +0,13 | +0,15 | 0,889 |
| EX10 *vs.* EX4 (L4 *vs.* sem aumento) | +0,34 | +0,47 | 0,662 |
| **5. Comparação entre níveis extremos** | | | |
| EX7 *vs.* EX10 (L1 *vs.* L4) | -1,48 | -1,75 | 0,155 |

**Notas:** $\Delta$AUC em pontos percentuais. $t$ e $p$ calculados por diferenças pareadas. $p$-valores não corrigidos para múltiplas comparações (exploratório). Em EX3 (ablação de AMP), a execução em FP32 elevou o uso de VRAM e demandou ajustes de treinamento (redução de *batch size* 32$\rightarrow$2) para viabilizar a execução. Por alterar o regime de otimização, EX3 deve ser interpretado como evidência do impacto computacional da AMP, e não como comparação estritamente controlada de desempenho; a AUC média foi 55,20%. As condições completas estão na Tabela 1.

Em complemento à análise por IC95% (Tabela 1), a Tabela 2 oferece uma visão mais refinada das diferenças de desempenho. Os resultados indicam que as ablações de treinamento (EX2 vs. EX1 e EX3 vs. EX1) produzem diferenças estatisticamente significativas, com perdas expressivas de 12,02 e 33,30 pontos percentuais na AUC-ROC, respectivamente. O resultado do EX3, no entanto, deve ser interpretado com cautela. A execução em precisão total (FP32) sem AMP elevou o consumo de VRAM a ponto de causar erros de memória (*Out of Memory* - OOM) nas duas GPUs de 12GB disponíveis cada. Para viabilizar a execução, foi necessário reduzir o *batch size* (32→2) e desativar o *multi-GPU*, o que altera o regime de otimização e evidencia o impacto prático do AMP na viabilização da configuração padrão.

Assim, a queda drástica de desempenho no EX3 (AUC média de 55,20%) não deve ser atribuída exclusivamente à ausência de AMP em termos numéricos, mas também à necessidade de reduzir o *batch size* de 32 para 2, o que altera substancialmente o regime de otimização e tende a degradar a estimativa do gradiente. Desse modo, EX3 deve ser interpretado como uma ablação de viabilidade computacional, e não como uma comparação estritamente controlada do efeito isolado da AMP sobre o desempenho.

Todas as demais comparações apresentaram $p$-valores superiores a $0,05$, indicando que, para este conjunto de dados e arquitetura, as variações no *pipeline* de aumento de dados não geram ganhos ou perdas consistentes de desempenho. Em particular:

- A ausência de aumento (EX4) não difere significativamente do *baseline* ($p = 0,278$), embora sua média seja ligeiramente superior.
- A adição de transformações espaciais (EX5) não melhora em relação à ausência ($p = 0,659$), e a inclusão de transformações radiométricas (EX6) também não supera o espacial ($p = 0,692$).
- Na varredura de intensidade, nenhum dos níveis (L1 a L4) apresenta diferença significativa em relação ao *baseline* ou à condição sem aumento. Observa-se que EX8 (L2) vs. EX1 apresenta um $p$-valor de $0,101$, sugerindo uma possível tendência, porém não significativa ao nível de $5\%$. A comparação entre os níveis extremos (EX7 vs. EX10) também não é significativa ($p = 0,155$).

A comparação entre EX7 (L1) e EX10 (L4) também ajuda a esclarecer que o *sweep* de intensidade não corresponde apenas a um aumento arbitrário da força das transformações. Entre esses extremos, houve escalonamento simultâneo da magnitude espacial (rotação axial de $\pm 5^\circ$ para $\pm 20^\circ$ e translação de $\pm 5\%$ para $\pm 20\%$) e do ruído gaussiano ($\sigma = 0{,}01$ para $\sigma = 0{,}05$). Ainda assim, a diferença entre EX7 e EX10 não foi estatisticamente significativa ($p = 0{,}155$), sugerindo que, dentro dos limites avaliados, o aumento progressivo da intensidade global do *augmentation* não produziu efeito consistente sobre a AUC-ROC.

Esses resultados, obtidos sob um protocolo estrito que mitiga vazamento de informação, indicam que o *baseline* (apenas flip axial) já proporciona um patamar de desempenho robusto, e que intervenções mais complexas no aumento de dados não trazem benefícios adicionais detectáveis nas condições testadas. Uma possível interpretação para esse comportamento é que a arquitetura R3D-18 com inicialização pré-treinada já possua robustez intrínseca suficiente para absorver pequenas variações espaciais e radiométricas. Alternativamente, o próprio conjunto OAI pode apresentar relativa homogeneidade de aquisição, reduzindo o potencial de ganho marginal proporcionado pelo aumento de dados nas condições testadas. Por fim, ressalta-se que em todos os experimentos o aumento

de dados foi aplicado exclusivamente ao conjunto de treino, e a partição de avaliação de cada *fold* (mantida fora do treino) permaneceu inalterada, garantindo que as variações observadas possam ser atribuídas unicamente às escolhas de inicialização, AMP e aumento de dados, sem evidência de vazamento no nível de volume.

## 4. Limitações e Trabalhos Futuros

Uma limitação importante deste estudo é que a validação cruzada foi conduzida no nível de volume, e não no nível de paciente. Assim, não se garante independência completa entre amostras provenientes do mesmo indivíduo em diferentes partições, o que pode inflar as estimativas de generalização do modelo.

Embora o estudo tenha contemplado transformações clássicas (geométricas e radiométricas), não foram exploradas estratégias mais recentes de aumento de dados, como *Mixup* (que cria amostras virtuais por interpolação linear entre pares de imagens e seus respectivos rótulos) em regime volumétrico [Zhang et al. 2017], *CutMix* adaptado para volumes [Yun et al. 2019] ou *generative augmentation* baseada em modelos de difusão [Kazerouni et al. 2023].

Diferentemente de métodos de *oversampling* como o SMOTE, que atuam no espaço de características de dados tabulares, essas técnicas operam diretamente no domínio dos pixels, preservando a estrutura espacial. No EX7–EX10, a varredura de intensidade concentrou-se em rotação, translação e ruído, mantendo contraste e gama em intervalos fixos; como extensão natural, recomenda-se otimizar conjuntamente esses hiperparâmetros. Recomenda-se ainda avaliar a generalização do *pipeline* em bases externas e realizar validação qualitativa da plausibilidade das imagens aumentadas com especialistas clínicos.

## 5. Conclusão

Investigou-se, de forma controlada, estratégias de aumento de dados para classificação binária de volumes 3D DESS de MRI do joelho, utilizando a base pública da *Osteoarthritis Initiative* (OAI) [Peterfy et al. 2008, NIMH Data Archive 2026]. Para isso, conduziram-se dez experimentos (EX1 a EX10) com validação cruzada estratificada em 5 *folds*, mantendo-se o protocolo fixo e variando-se apenas (i) *transfer learning*, (ii) AMP e (iii) o tipo/intensidade do aumento de dados.

Os resultados indicaram que no cenário avaliado, o *transfer learning* foi o fator determinante: o modelo pré-treinado (EX1, AUC$= 88{,}49 \pm 1{,}32\%$) superou significativamente o treinamento do zero (EX2, AUC$= 76{,}48 \pm 1{,}76\%$), com $p < 0{,}0001$ no teste pareado. Em contraste, as variações no *pipeline* de aumento de dados não produziram diferenças estatisticamente significativas. Em particular, entre EX1 e EX4–EX10 as médias de AUC permaneceram na faixa de aproximadamente $88\%$ a $90\%$ (EX4: $89{,}34 \pm 1{,}86\%$, EX5: $88{,}92 \pm 1{,}55\%$, EX6: $88{,}43 \pm 1{,}93\%$, EX7: $88{,}20 \pm 1{,}24\%$, EX8: $90{,}08 \pm 1{,}26\%$, EX9: $89{,}47 \pm 1{,}14\%$, EX10: $89{,}68 \pm 1{,}33\%$), e os testes $t$ pareados (Tabela 2) não apontaram vantagem consistente da inclusão de transformações espaciais, radiométricas ou da variação de intensidade sobre o *baseline* (apenas flip axial).

O experimento EX3 (ablação de AMP) exigiu ajustes de *batch size* e desativação de multi-GPU por limitação de VRAM, e resultou em AUC média de 55,20%. Como essa condição altera o regime de otimização, EX3 é interpretado principalmente

como evidência do papel do AMP na viabilização computacional do treinamento com parâmetros padrão, e não como comparação estritamente controlada de desempenho.

Em síntese, os experimentos mostram que o ganho mais expressivo vem da inicialização pré-treinada, enquanto o aumento de dados além do flip não trouxe benefício detectável no cenário avaliado. Ressalta-se, porém, que a validação cruzada foi realizada por volume, sem garantia de separação por paciente, o que pode inflar as estimativas de generalização. A contribuição prática é oferecer um protocolo reprodutível e evidências de que um pipeline mais simples pode ser suficiente nesse contexto, reduzindo custos de implementação.

## Referências


Berrimi, M. (2022). OAI MRI 3D DESS dataset (NumPy preprocessed). https://www.kaggle.com/datasets/mohamedberrimi/oaimri3ddess/data. Publicado em: 2022. Acesso em: 10 jan. 2026.

Chlap, P., Min, H., Vandenberg, N., Dowling, J., Holloway, L., and Haworth, A. (2021). A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology*, v. 65, n. 5, p. 545–563.

Fontainha, T. C., Henriques, F. d. R., Lima, A. A., Araujo, G. M., and Tesch, R. d. S. (2025). Classificação binária de imagens de ressonância magnética de osteoartrite com o modelo r3d_18 modificado. In *Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS)*, p. 1017–1022. SBC.

Guida, C., Zhang, M., and Shan, J. (2021). Knee osteoarthritis classification using 3D CNN and MRI. *Applied Sciences*, v. 11, n. 11, p. 5196.

Hara, K., Kataoka, H., and Satoh, Y. (2018). Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 6546–6555, Salt Lake City, USA.

Islam, T., Hafiz, M. S., Jim, J. R., Kabir, M. M., and Mridha, M. (2024). A systematic review of deep learning data augmentation in medical imaging: Recent advances and future research directions. *Healthcare Analytics*, v. 5, p. 100340.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al. (2017). The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Kazerouni, A., Aghdam, E. K., Heidari, M., Azad, R., Fayyaz, M., Hacihaliloglu, I., and Merhof, D. (2023). Diffusion models in medical imaging: A comprehensive survey. *Medical image analysis*, v. 88, p. 102846.

Kellgren, J. H., Lawrence, J., et al. (1957). Radiological assessment of osteo-arthrosis. *Ann Rheum Dis*, v. 16, n. 4, p. 494–502.

Kingma, D. P. and Ba, J. (2014). A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kohn, M. D., Sassoon, A. A., and Fernando, N. D. (2016). Classifications in brief: Kellgren-lawrence classification of osteoarthritis. *Clinical Orthopaedics and Related Research®*, v. 474, n. 8, p. 1886–1893.

Krinski, B., Ruiz, D., and Todt, E. (2022). Light in the black: An evaluation of data augmentation techniques for COVID-19 CT's semantic segmentation. In *Anais do XXII Simpósio Brasileiro de Computação Aplicada à Saúde*, p. 156–167, Porto Alegre, Brasil.

Montgomery, D. C. (2017). *Design and analysis of experiments*. John wiley & sons.

NIMH Data Archive (2026). The osteoarthritis initiative (OAI). https://nda.nih.gov/oai. Acesso em: 11 fev. 2026.

Peterfy, C. G., Schneider, E., and Nevitt, M. (2008). The osteoarthritis initiative: report on the design rationale for the magnetic resonance imaging protocol for the knee. *Osteoarthritis and Cartilage*, v. 16, n. 12, p. 1433–1441.

PyTorch (2026). torch.nn.functional.grid_sample. https://docs.pytorch.org/docs/stable/generated/torch.nn.functional.grid_sample.html. Acesso em: 11 fev. 2026.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, p. 234–241, Munich, Germany.

Scalercio, A. and Freitas, C. (2023). Proposta e avaliação linguística de técnicas de aumento de dados. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, p. 207–223, Porto Alegre, Brasil.

Shorten, C. and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, v. 6, n. 1, p. 1–48.

Yeoh, P. S. Q., Lai, K. W., Goh, S. L., Hasikin, K., Wu, X., and Li, P. (2023). Transfer learning-assisted 3D deep learning models for knee osteoarthritis detection: Data from the osteoarthritis initiative. *Frontiers in Bioengineering and Biotechnology*, v. 11, p. 1164655.

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., and Yoo, Y. (2019). Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, p. 6023–6032, Seoul, South Korea.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. (2017). Mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.