High-Sensitivity Bone Tumor Classification in Radiographs Using Deep Learning and Ensemble Learning Strategies

Francisco J. dos Santos Reis [1], **Antonio O. de Carvalho Filho**[1], **José D. L. Araújo**[1]

[1]Universidade Federal do Piauí - Programa de pós graduação em ciência da computação
Picos, Brasil

{franciscojose17, antoniooseas, jose.denes}@ufpi.edu.br

Abstract. *Bone tumor classification in radiographic images is a challenging task due to the visual variability of lesions. This work presents a deep learning-based approach for binary classification of bone tumors into benign and malignant classes, using radiographic images from the BTXRD dataset. Six pre-trained deep learning architectures were evaluated, including ConvNeXt, MaxViT, and Swin Transformer models. In addition, ensemble strategies based on soft voting, weighted voting, and stacking were investigated. On the test set, the proposed method achieved an accuracy of 97.87%, an F1-score of 97.92%, and an AUC-ROC of 99.59%. Furthermore, the method obtained a recall of 100.00% for the malignant class.*

Resumo. *A classificação de tumores ósseos em imagens radiográficas é uma tarefa desafiadora devido à variabilidade visual das lesões. Este trabalho apresenta uma abordagem baseada em aprendizado profundo para a classificação binária de tumores ósseos em benignos e malignos, utilizando imagens radiográficas do conjunto de dados BTXRD. Foram avaliadas seis arquiteturas profundas pré-treinadas, incluindo modelos ConvNeXt, MaxViT e Swin Transformer. Além disso, foram investigadas estratégias de ensemble baseadas em soft voting, weighted voting e stacking. No conjunto de teste, o método proposto alcançou acurácia de 97,87%, F1-score de 97,92% e AUC-ROC de 99,59%. Além disso, o método obteve recall de 100,00% para a classe maligna.*

1. Introdução

Os tumores ósseos primários são um grupo raro e heterogêneo de neoplasias que apresentam um amplo espectro de características morfológicas, biológicas, genéticas e radiológicas [Saran and Phulware 2022]. Seu desenvolvimento ocorre principalmente no tecido ósseo, podendo ser classificados como malignos ou benignos. Os tumores ósseos primários malignos são menos frequentes do que os tumores ósseos primários benignos. Embora ambos possam crescer e danificar áreas de osso saudável, os tumores benignos tendem a não se espalhar e raramente oferecem risco à vida.

Os tumores benignos são geralmente achados incidentais, enquanto os tumores malignos são frequentemente diagnosticados em estágios avançados. A investigação por imagem adequada e o diagnóstico histopatológico preciso são cruciais para um tratamento apropriado. [Saran and Phulware 2022]. Os tumores ósseos primários estão incluídos na categoria mais ampla de cânceres chamados sarcomas e vários são seus subtipos, sendo os mais comuns o osteossarcoma, sarcoma de Ewing e o condrossarcoma.

Destaca-se que os tumores ósseos malignos primários são a terceira maior causa de mortalidade relacionada ao câncer em pacientes com menos de 20 anos [He et al. 2020]. Afetando predominantemente pacientes jovens, mas também pode acometer adultos. O condrossarcoma representa 40% das neoplasias nessa faixa etária, seguido pelo osteossarcoma com 28%. Em crianças e adolescentes, o osteossarcoma representa 56% dos casos, sendo o mais comum, enquanto o sarcoma de Ewing representa 34% dos casos [Pullan and Lotfollahzadeh 2025].

Os métodos de diagnóstico empregados na identificação de tumores ósseos primários incluem exames de imagem, análises laboratoriais de sangue e a biópsia do tecido tumoral [Pullan and Lotfollahzadeh 2025]. Entre as técnicas de imagem, a radiografia convencional(RC), que utiliza o raio X como princípio físico de formação da imagem, permanece como o primeiro e principal exame utilizado. Radiografias podem mostrar a localização, o tamanho e a forma de um tumor óssoo [National Cancer Institute 2018], características essenciais para diagnóstico de um tumor ósseo. Portanto, a maioria dos tumores ósseos pode ser diagnosticada, ou pelo menos podemos restringir o diagnóstico diferencial, com base em radiografias convencionais [Saran and Phulware 2022]

Radiografia convencional (RC) é a primeira modalidade de imagem utilizada para avaliar a lesão, por ser amplamente disponível e acessível.[Saran and Phulware 2022]. Contudo, o diagnóstico manual baseado na análise de radiografias continua sendo um desafio para especialistas. A análise manual de imagens médicas introduz variabilidade, subjetividade e também acarreta o risco de diagnóstico incorreto [Yao et al. 2025]. A interpretação radiográfica exige experiência avançada, já que tumores ósseos podem apresentar padrões sutis e sobreposições estruturais, o que aumenta o risco de erros e inconsistências entre avaliadores, tornando o processo vulnerável à fadiga humana e à subjetividade.

Nesse contexto, o surgimento de algoritmos de aprendizado profundo, especialmente redes neurais convolucionais (CNNs), impactou significativamente práticas clínicas como o diagnóstico assistido por computador (CAD) [Wang et al. 2025b]. Dessa forma, CAD se tornou um dos principais temas de pesquisa em diagnóstico por imagem por aliviar significativamente os desafios da saúde em países em desenvolvimento, aproveitando dados de imagem e informações de diagnóstico clínico [Liu et al. 2025]. Tendo entre outras funções: delimitação de áreas suspeitas, identificação de lesões de forma mais precisa e apoiar a triagem inicial, reduzindo erros e ampliando a precisão no diagnóstico.

Assim, este trabalho apresenta uma abordagem para análise de tumores ósseos em imagens de raio-X, visando auxiliar os profissionais de saúde a realizar um diagnóstico mais veloz e exato. Atraves da construção de um sistema baseado em modelos de *Deep Learning*(DL) para a classificação de tumores ósseos utilizando imagens de radiografia. As principais contribuições deste artigo estão especificadas abaixo:

- É realizada uma avaliação comparativa de seis arquiteturas profundas pré-treinadas, incluindo modelos baseados em Transformers e redes convolucionais, tais como MaxViT, Swin Transformer e ConvNeXt.
- É investigado o uso de diferentes estratégias de ensemble, incluindo *soft voting*, *weighted voting* e *stacking*, visando combinar as predições dos modelos individuais.

2. Trabalhos Relacionados

Nos últimos anos, diversas abordagens têm sido desenvolvidas para a classificação e segmentação automática de tumores ósseos primários, utilizando técnicas de pré-processamento e aprendizado profundo. Nesta seção, apresentamos os principais estudos relacionados que empregam estratégias semelhantes à proposta deste trabalho, destacando seus avanços, contribuições e limitações metodológicas.

[Ajay et al. 2026] propõem uma abordagem de deep learning para a classificação de tumores ósseos em radiografias, utilizando uma arquitetura híbrida que combina CNNs e um mecanismo de atenção de tripla fusão, capaz de capturar simultaneamente características locais e globais das imagens. Além do modelo de classificação, o estudo incorpora métodos de IA explicável (XAI), permitindo gerar visualizações que destacam as regiões mais relevantes usadas pela rede na tomada de decisão, aumentando a interpretabilidade clínica. Avaliado no dataset BTXRD, o modelo alcançou 84,35% de acurácia na classificação de tumores em malignos e benignos.

Já [Zhao et al. 2024] em seu trabalho apresentam um sistema automático capaz de segmentar tumores ósseos e, simultaneamente, prever sua origem, com foco especial na identificação de metástases provenientes de câncer de pulmão. A abordagem combina o poder representacional do ConvNeXtv2 com a arquitetura Mask R-CNN, resultando em uma fusão eficiente entre extração robusta de características e segmentação precisa. O modelo busca capturar tanto padrões morfológicos quanto texturais relevantes para distinguir tumores primários de metastáticos. Obtendo um Coeficiente de Similaridade de Dice 84.9%, uma sensibilidade de 91,1% e uma especificidade de 93,1%

[Bokhari and Shafi 2026] Ao desenvolvimento de dois modelos principais. O primeiro é uma combinação de Redes Neurais Convolucionais com Transformers (CNNT) e o segundo é uma ResNet50 aprimorada com o módulo de atenção CBAM, visando melhorar a extração de características e o entendimento contextual das imagens. Os modelos são avaliados no conjunto de dados BTXRD, com aplicação de técnicas de pré-processamento e balanceamento de dados para lidar com ruído e desbalanceamento de classes. Obtendo uma acurácia média de 97,41%, F1-score de 97.59%, uma pontuação AUC-ROC de 98,4% e um coeficiente Kappa de Cohen de 97,18%.

Em seu estudo [Wang et al. 2025a] investigaram o uso de redes neurais profundas para diferenciar tumores ósseos primários de infecções ósseas, duas condições que frequentemente apresentam padrões radiográficos semelhantes e de difícil distinção clínica. Os autores avaliam diferentes arquiteturas de DL aplicadas diretamente a radiografias, explorando técnicas de pré-processamento e otimização para maximizar a acurácia. Utiliza conjunto de dados próprio, dividindo os pacientes em dados externos e internos, atingindo uma área sob a curva de 94,8% e 96,3% e acurácias de 88,15 e 89,5%.

Por fim, [Song et al. 2024] propõe um modelo de DL capaz de classificar tumores ósseos primários utilizando imagens multimodais de radiografias, tomografia computadorizada e ressonância magnética, juntamente com características clínicas. Para classificar tumores ósseos primários como benignos, intermediários ou malignos. Apresentando um modelo de fusão de Redes Transformers para classificação de tumores ósseos em um conjunto de dados próprio, obtendo um AUC de 90% acurácia de 84,3%, sensibilidade de 84,3% e especificidade de 92,1% em pacientes submetidos a radiografia.

Apesar dos avanços na classificação de tumores ósseos por aprendizado profundo, observa-se que parte dos trabalhos concentra-se em arquiteturas individuais ou em tarefas distintas, como segmentação, identificação de metástases e classificação multimodal. Além disso, há poucos estudos que comparam diferentes arquiteturas profundas e investigam estratégias de ensemble para a classificação binária de tumores benignos e malignos em radiografias do conjunto BTXRD.

Nesse contexto, este trabalho busca suprir essa lacuna por meio da avaliação de arquiteturas heterogêneas, incluindo ConvNeXt, MaxViT e Swin Transformer, aplicadas a recortes da região tumoral. Posteriormente, os modelos são combinados por estratégias de ensemble, com destaque para *weighted voting*. Essa abordagem busca reduzir falsos negativos e priorizar a identificação de tumores malignos, aspecto especialmente relevante em aplicações médicas.

3. Materiais e Métodos

Nessa seção serão apresentadas todas as etapas para classificação de tumores ósseos. Os métodos incluem a aquisição da base de dados, pré-processamento dos dados, treinamento dos modelos e, finalmente, avaliação dos resultados, conforme pode ser visto na figura 1.

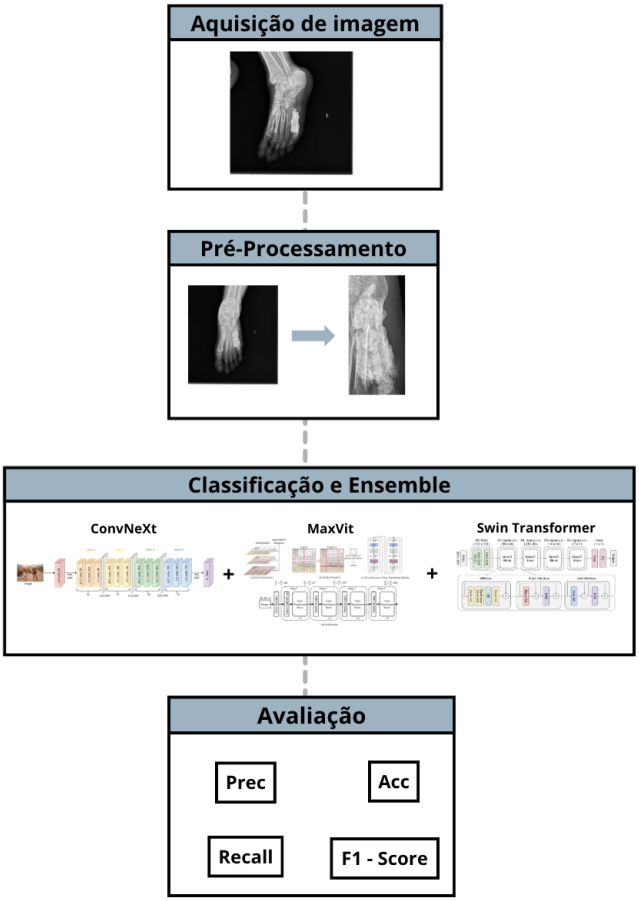


Figura 1. Metodologia proposta. Fonte: Autoria própria

3.1. Aquisição dos dados

Este trabalho utiliza o conjunto de dados de radiografias de tumores ósseos BTXRD proposto por [Yao et al. 2025] que foi desenvolvido com o propósito de apoiar pesquisas voltadas à classificação, localização e segmentação automatizadas de tumores ósseos primários em radiografias. O dataset é composto por 3.746 imagens radiográficas, distribuídas entre 1.879 exames de ossos normais, 1.525 casos de tumores ósseos benignos e 342 casos de tumores ósseos malignos. Na figura 2 são mostrados exemplos de imagens no conjunto de dados.

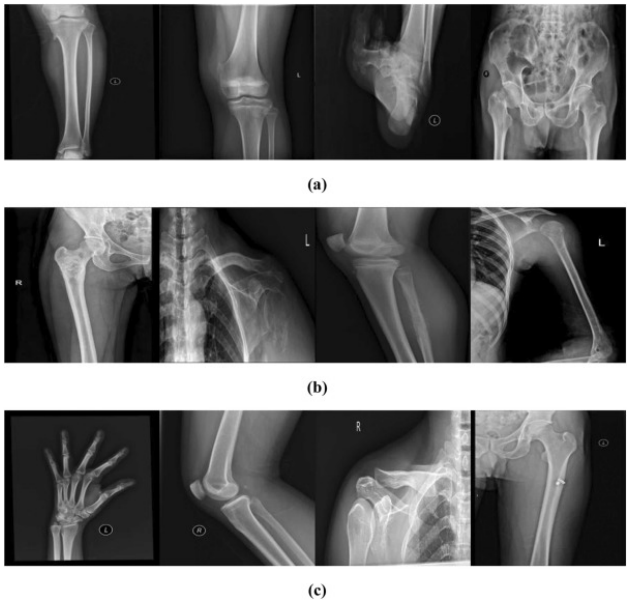


Figura 2. Exemplos de imagens de raios X. Fonte:[Yao et al. 2025]

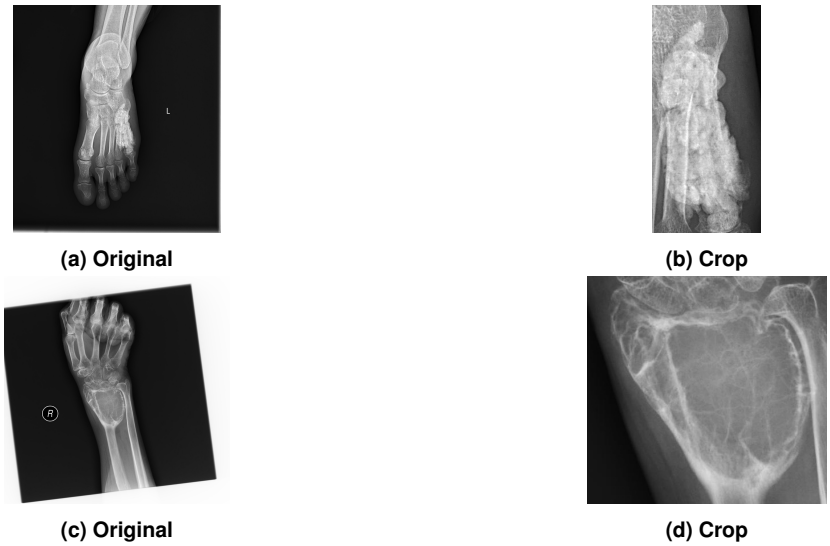
Cada imagem é acompanhada de metadados clínicos relevantes, incluindo sexo, idade, região anatômica examinada e ângulo de aquisição da radiografia. As imagens foram obtidas de três centros médicos diferentes: um é a plataforma online Radiopaedia.org, o banco de dados público MedPix e hospitais na China. O conjunto de dados é distribuído em nove subtipos de tumores ósseos: osteocondroma, osteossarcoma, osteocondromatose múltipla, cisto ósseo simples, tumor de células gigantes, osteofibroma, osteocondroma sinovial, outros bt que são uma coleção de vários tumores benignos, enquanto 'outros mt' é uma coleção de vários tumores malignos.

3.2. Pré-Processamento

Como etapa de pré-processamento, foram utilizadas somente as imagens com anotações dos tumores disponíveis no conjunto de dados para realizar o recorte das regiões contendo os tumores. Desta forma, as imagens normais não foram utilizadas neste trabalho. Essas regiões posteriormente foram ampliadas em 20% em relação à sua área original, com o objetivo de fornecer ao modelo uma maior quantidade de informações contextuais durante o processo de treinamento e avaliação. Esses recortes foram então empregados como entrada para a etapa de classificação, conforme ilustrado na Figura 3.

O objetivo desse pré-processamento é simular o fluxo operacional de um sistema integrado de segmentação e classificação de tumores ósseos futuro. Nesse pipeline, um

modelo segmentador será empregado para detectar e delimitar automaticamente a região tumoral, que é então fornecida como entrada ao modelo de classificação desenvolvido neste trabalho. Assim, o processo experimental aproxima-se das condições de uso esperadas em um ambiente real, possibilitando uma avaliação mais representativa do desempenho do sistema proposto.



(a) Original (b) Crop

(c) Original (d) Crop

Figura 3. Comparação antes e depois do crop

3.3. Aumento dos dados

Conforme visto na figura 4 o conjunto de dados apresenta forte desbalanceamento entre as categorias de tumores ósseos, com classes majoritárias como *osteochondroma* e *multiple osteochondromas* contendo centenas de amostras, enquanto classes minoritárias, *other malignant tumors* e *osteofibroma*, apresentam poucas imagens. Esse desequilíbrio se mostra ainda maior, quando se divide o conjunto de imagens em classes de tumores benignos ou malignos, uma vez que as classes malignas são apenas *osteosarcoma* e *other mt*, enquanto todas as demais são tumores benignos.

Para mitigar esse efeito, foi empregada uma estratégia de aumento de dados em transformações geométricas, incluindo *RandomResizedCrop, RandomHorizontalFlip, RandomRotation, ColorJitter e GaussianBlur*. O aumento de dados engloba um conjunto de técnicas que aumentam o tamanho e a qualidade dos conjuntos de dados de treinamento, permitindo a construção de modelos de DL [Shorten and Khoshgoftaar 2019]. Este aumento de dados foi empregado apenas no conjunto de treinamento, que representa 90% dos dados totais.

3.4. Classificação

Nesta etapa, foram empregados modelos de aprendizado profundo para realizar a classificação das imagens em benigna e maligna. O objetivo foi investigar a capacidade de diferentes arquiteturas em extrair características discriminativas das regiões de interesse e identificação de padrões. Os modelos utilizados foram ConvNeXt, Swin Transformer v2 e MaxViT, que serão descritos abaixo.

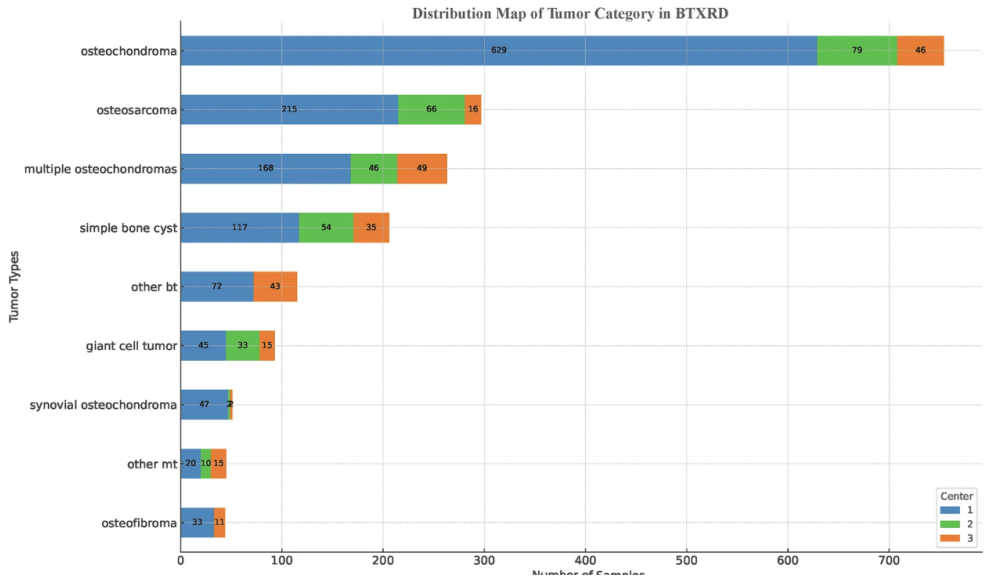


Figura 4. Distribuição das imagens por classe Fonte:[Yao et al. 2025]

ConvNext é uma arquitetura de rede neural convolucional que revisita e aprimora o paradigma das CNNs tradicionais ao incorporar princípios de projeto inspirados nos Vision Transformers (ViTs). Proposta por [Liu et al. 2022a], essa arquitetura demonstra que modelos puramente convolucionais podem alcançar desempenho promissor em reconhecimento de imagens, ao mesmo tempo em que preservam características desejáveis como eficiência computacional e escalabilidade. Dessa forma, o ConvNeXt tem sido amplamente adotado em cenários de aprendizado por transferência, pesquisas em visão computacional e aplicações práticas no mundo real. Dessa forma, o trabalho evidencia que redes convolucionais, quando modernizadas com práticas arquiteturais contemporâneas, permanecem altamente relevantes, oferecendo uma alternativa eficiente, escalável e de fácil implementação para aplicações em visão computacional [Liu et al. 2022a].

O MaxViT (*Multi-Axis Vision Transformer*) é uma arquitetura híbrida que combina operadores convolucionais e mecanismos de autoatenção. Sua principal contribuição é o mecanismo de atenção multi-eixo, composto pela *block attention* e pela *grid attention*. A primeira operação modela relações locais entre regiões próximas da imagem, enquanto a segunda permite a interação entre regiões espacialmente distantes, incorporando informações globais ao processo de representação. Dessa forma, o modelo consegue capturar padrões locais e dependências de longo alcance com menor custo computacional quando comparado à autoatenção global convencional. A arquitetura é organizada hierarquicamente em múltiplos estágios, nos quais blocos MaxViT combinam camadas convolucionais do tipo MBConv com os mecanismos de atenção local e global. Essa característica torna o modelo adequado para tarefas de classificação de imagens médicas, nas quais tanto detalhes morfológicos da lesão quanto o contexto visual da região analisada podem contribuir para a decisão diagnóstica [Tu et al. 2022].

O Swin Transformer V2 é uma evolução do Swin Transformer desenvolvida para permitir o treinamento de modelos com maior capacidade e o processamento de imagens em resoluções mais elevadas. A arquitetura utiliza mecanismos de autoatenção em

janelas locais deslocadas, reduzindo o custo computacional em relação à atenção global convencional. Entre suas principais contribuições estão a atenção baseada em similaridade do cosseno, a normalização pós-residual e a codificação posicional relativa contínua, que aumentam a estabilidade do treinamento e favorecem a generalização para imagens com diferentes resoluções. Dessa forma, o modelo é capaz de extrair características hierárquicas locais e globais, sendo adequado para tarefas de classificação de imagens médicas [Liu et al. 2022b].

3.5. Treinamento

Os modelos de classificação foram implementados com as bibliotecas PyTorch, utilizando pesos pré-treinados e uma camada de saída adaptada para classes benigno e maligno. O treinamento foi realizado com imagens redimensionadas para o tamanho (224x224). O treinamento utilizou o otimizador Adam, taxa de aprendizado de 1×10^{-5}, com no máximo 50 épocas. A cada época, o desempenho foi avaliado por meio da acurácia global e da acurácia por classe. Foi aplicado *early stopping* com paciência de 15 épocas, salvando-se o modelo quando a acurácia global apresentava melhora superior a $0{,}0005$. As imagens foram divididas em 80% para treinamento, 10% para validação e 10% para teste.

Os experimentos foram realizados em um servidor com sistema operacional baseado em Linux, com 125 GB de memória RAM e três unidades de processamento gráfico (GPUs) NVIDIA GeForce RTX 2080 Ti, cada uma com 11 de GB de memória dedicada.

3.6. Ensemble

Como forma de aprimorar o desempenho dos modelos descritos na seção 3.4, foi usada técnica de *ensemble learning*. Esse é um método que integra múltiplos modelos base em uma mesma estrutura para obter um modelo mais robusto que os supera[Mohammed and Kora 2023]. Essa abordagem fundamenta-se no princípio de que múltiplos modelos, ainda que individualmente apresentem desempenho fraco, podem ser combinados para constituir um modelo mais forte, capaz de elevar o desempenho global, reduzir o sobreajuste e melhorar a generalização.

Uma das principais motivações para o uso de métodos de *ensemble* é o equilíbrio entre viés e variância. Modelos com alto viés tendem a simplificar excessivamente os dados, resultando em subajuste, enquanto modelos com alta variância são sensíveis ao ruído, levando ao sobreajuste. A combinação de diferentes modelos permite equilibrar esses efeitos, produzindo um sistema mais estável e eficiente.

Assim, foram usadas técnicas de *soft voting* que combinam as probabilidades preditas pelos modelos individuais por meio de uma média aritmética simples. No *weighted voting*, as probabilidades também são combinadas, mas cada modelo recebe um peso diferente. Assim, modelos mais relevantes para o desempenho do conjunto têm maior influência na decisão final. Por fim, foi usada a estratégia de *stacking* utilizou as probabilidades produzidas pelos modelos base como atributos de entrada para um meta-modelo. Dessa forma, o meta-modelo foi responsável por aprender uma regra de combinação das predições individuais, em vez de utilizar pesos previamente fixados.

3.7. Avaliação

Para validar os resultados de classificação, foram utilizadas métricas de avaliação estatística, como Acurácia (Acc), precisão (Pre), Recall (Rec), F1-Score (F1) e Área sob a

Curva (AUC-ROC). Essas métricas foram calculadas com base na matriz de confusão. A matriz de confusão fornece uma hipótese das medidas eficazes do modelo de classificação, mostrando o número de classificações corretas versus as classificações previstas para cada classe em um determinado conjunto de exemplos.

4. Resultados e Discussão

Nesta seção serão descritos os resultados obtidos com os modelos de classificação. As arquiteturas treinadas foram o modelo ConvNeXt-Small em duas configurações distintas: uma pré-treinada apenas no ImageNet-1K e outra com pré-treinamento estendido no ImageNet-22K. Para as arquiteturas baseadas em Transformers, foram utilizadas duas variantes do Swin Transformer na versão Tiny e v2 Small. No caso do MaxViT, foram avaliadas as versões MaxViT-Small e MaxViT-Base, ambas treinadas no ImageNet-1K.

4.1. Resultados dos experimentos

Para avaliação dos resultados, foram usadas as métricas descritas na seção 3.7. Na Tabela 1 temos os resultados globais para classificação de tumores ósseos. Os resultados obtidos demonstram que os modelos individuais apresentaram desempenho elevado na tarefa, com destaque para o MaxViT-Base, que alcançou a maior acc de 96,81% e o maior F1 96,87%, evidenciando forte capacidade de discriminação entre as classes.

Modelo/Estratégia	**Acc (%)**	**Pre (%)**	**Rec (%)**	**F1 (%)**	**AUC-ROC**
MaxViT-Base	**96,81**	**97,07**	**96,81**	**96,87**	**0,9966**
MaxViT-Small	96,28	96,29	96,28	96,16	0,9617
ConvNeXt-Small 22K	95,21	95,17	95,21	95,19	0,9589
Swin V2 Small	91,49	92,91	91,49	91,88	0,9593
ConvNeXt-Small	94,68	94,68	94,68	94,68	0,9584
Swin Tiny	95,21	95,27	95,21	95,24	0,9539

Tabela 1. Desempenho global dos modelos individuais

Como visto na Tabela 2, a adoção de estratégias de ensemble resultou em melhorias consistentes no desempenho global. O melhor desempenho foi obtido pelo *weighted voting*, atingindo acc de 97,87% e pre de 98,09%. Demonstrou-se que as técnicas de ensemble conseguiram melhorar o desempenho global dos modelos.

Modelo/Estratégia	**Acc (%)**	**Pre (%)**	**Rec (%)**	**F1 (%)**	**AUC-ROC**
Soft Voting	96,28	96,24	96,28	96,26	0,9935
Weighted Voting	**97,87**	**98,09**	**97,87**	**97,92**	**0,9959**
Stacking (OOF)	96,28	96,33	96,28	96,30	0,9955

Tabela 2. Desempenho global dos modelos ensemble

Já na Tabela 3, são apresentados os resultados de precisão e recall por classe. Observa-se que os modelos individuais apresentam variação de desempenho, especialmente na classe maligna, evidenciando diferenças na capacidade de sensibilidade entre as arquiteturas avaliadas. O maior recall para a classe maligna foi obtido pelo MaxViT-Base, com 97,14%, enquanto o MaxViT-Small apresentou a maior precisão para essa classe, atingindo 96,67%.

Modelo/Estratégia	Pre benigno (%)	Rec benigno (%)	Pre maligno (%)	Rec maligno (%)
MaxViT-Base	**99,33**	**96,73**	**87,18**	**97,14**
MaxViT-Small	96,20	99,35	96,67	82,86
ConvNeXt-Small 22K	96,75	97,39	88,24	85,71
Swin V2 Small	97,90	91,50	71,11	91,43
ConvNeXt-Small	96,73	96,73	85,71	85,71
Swin Tiny	97,37	96,73	86,11	88,57

Tabela 3. Precisão e recall por classe dos modelos individuais e ensemble.

A Tabela 4 exibe as estratégias de ensemble por classe. Observa-se uma melhora significativa no desempenho geral, com destaque para a configuração baseada em *weighted voting*, a qual alcançou 100,00% de recall para a classe maligna e 97,39% para a classe benigna. Esse resultado é particularmente relevante em aplicações médicas, pois reduz drasticamente a ocorrência de falsos negativos, garantindo a identificação de todos os casos malignos no conjunto de teste.

Modelo/Estratégia	Pre benigno (%)	Rec benigno (%)	Pre maligno (%)	Rec maligno (%)
Soft Voting	97,40	98,04	91,18	88,57
Weighted Voting	**100,00**	97,39	89,74	**100,00**
Stacking (OOF)	98,03	97,39	88,89	91,43

Tabela 4. Precisão e recall por classe dos modelos individuais e ensemble.

4.2. Comparação com outros trabalhos

Na Tabela 5 temos um comparativo entre resultados obtidos neste trabalho com os encontrados na literatura. O trabalho de [Ajay et al. 2026] utiliza uma arquitetura híbrida complexa CNN + Transformer + atenção, o presente trabalho adota uma abordagem baseada em Ensemble com recorte da região tumoral, obtendo desempenho superior ACC de 97,87% contra 84,35%. Já [Zhao et al. 2024] foca em segmentação com Mask R-CNN e ConvNeXtv2, enquanto este trabalho resolve diretamente a classificação com menor complexidade e resultados competitivos.

O estudo de [Bokhari and Shafi 2026] apresenta métricas com resultados similares, porém com destaque para AUC-ROC de 99,59% alcançada por esse trabalho contra 98,40%. Vale ressaltar que o trabalho de [Bokhari and Shafi 2026] utiliza a mesma base deste artigo. Já trabalhos como [Wang et al. 2025a] e [Song et al. 2024] obtêm resultados inferiores ou dependem de dados multimodais, enquanto o presente estudo, utilizando apenas radiografias, atinge maior acurácia. Assim, o uso de ensemble mostrou-se fundamental, elevando o desempenho global ACC de 97,87% e AUC-ROC de 99,59% e garantindo maior robustez.

5. Conclusão

Neste trabalho, foi proposto um sistema de classificação de tumores ósseos em radiografias baseado em modelos ConvNeXt, MaxViT e Swin Transformer junto ao recorte da região de interesse e ao uso de ensemble. A abordagem mostrou-se eficaz mesmo diante de desafios como o desequilíbrio de classes e a variabilidade das imagens médicas. Os

Trabalho	Acurácia	AUC-ROC	Pre	Rec
Ajay et al. [Ajay et al. 2026]	84,35%	–	–	–
Zhao et al. [Zhao et al. 2024]	–	–	–	91,10%
Bokhari et al. [Bokhari and Shafi 2026]	97,41%	98,40%	–	–
Wang et al. [Wang et al. 2025a]	89,5%	96,30%	–	
Song et al. [Song et al. 2024]	84,30%	90,00%		– 84,30%
Método proposto	**97,87%**	**99,59%**	**98,09** %	**97,87%**

Tabela 5. Comparação entre trabalhos relacionados e o método proposto.

resultados obtidos demonstram alto desempenho, com acurácia de 97,87%, AUC-ROC de 99,59% e F1-score de 97,92%, destacando a robustez do ensemble na redução de erros dos modelos individuais. Além disso, o método apresentou melhor desempenho na classe maligna com rec de 100%, mantendo resultados consistentes na classe benigna rec de 97,39 %.Dessa forma, a abordagem proposta se mostra eficiente e promissora, combinando alto desempenho com menor complexidade, sendo adequada para aplicações em CAD.

Além disso pretende-se futuramente integrar o trabalho proposto a um pipeline completo de localização, segmentação e classificação de tumores ósseos. Assim, o método proposto se destaca pelo equilíbrio entre eficiência e alto desempenho.

Referências

Ajay, A. et al. (2026). Bonevisionnet: A deep learning approach for the classification of bone tumours. *Computerized Medical Imaging and Graphics*.

Bokhari, S. M. and Shafi, M. (2026). Deep learning driven radiographic classification of primary bone tumors using attention augmented hybrid models. *Biomedical Signal Processing and Control*, 113:108888.

He, Y. et al. (2020). Deep learning-based classification of primary bone tumors on radiographs: A preliminary study. *eBioMedicine*, 62.

Liu, Y. et al. (2025). Bone tumor recognition strategy. *Scientific Reports*.

Liu, Z. et al. (2022a). A convnet for the 2020s. *arXiv*.

Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., Wei, F., and Guo, B. (2022b). Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Mohammed, A. and Kora, R. (2023). A comprehensive review on ensemble deep learning: Opportunities and challenges. *Journal of King Saud University - Computer and Information Sciences*, 35(2):757–774.

National Cancer Institute (2018). Primary bone cancer fact sheet.

Pullan, J. E. and Lotfollahzadeh, S. (2025). Primary bone cancer. *StatPearls*.

Saran, S. and Phulware, R. H. (2022). World health organization classification of bone tumors (fifth edition): What a radiologist needs to know? *Indian Journal of Musculoskeletal Radiology*, 4:73–86.

Shorten, C. and Khoshgoftaar, T. (2019). Image data augmentation survey. *Journal of Big Data*.

Song, L. et al. (2024). Deep learning with multimodal images. *Cancer Imaging*.

Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A., and Li, Y. (2022). Maxvit: Multi-axis vision transformer. In *European Conference on Computer Vision*.

Wang, H. et al. (2025a). Deep learning models for bone tumors. *npj Precision Oncology*.

Wang, H., He, Y., Wan, L., et al. (2025b). Deep learning models in classifying primary bone tumors and bone infections based on radiographs. *npj Precision Oncology*, 9:72.

Yao, S. et al. (2025). A radiograph dataset for the classification, localization, and segmentation of primary bone tumors. *Scientific Data*.

Zhao et al. (2024). Convnext-based mask r-cnn for bone tumor segmentation. *Journal of Bone Oncology*.