

Uma Abordagem Sistemática para Seleção de Encoders aplicada à Segmentação de Tumores com TransUNet

Iulle M. G. Neves¹, Mauro A. G. Clark¹, José D. L. Araújo²
Antonio O. de Carvalho Filho¹

¹ Universidade Federal do Maranhão (UFMA) – Programa de Pós-Graduação
Doutorado em Ciência da Computação - Associação (UFMA/UFPI)

² Universidade Federal do Piauí
Curso de Sistemas de Informação - Picos, Brasil (UFPI)

iulle@live.com, mauroclark@ctu.uespi.br,
{jose.denes, antoniooseas}@ufpi.edu.br

Abstract. *The choice of encoder in hybrid segmentation architectures directly impacts model performance. This study evaluates five encoders integrated into a TransUNet for bone tumor segmentation, in addition to comparing the proposed approach with reference architectures. The results reveal relevant differences among the models, with ConvNeXt-Tiny achieving the best overall performance, ResNet34 the highest precision, and EfficientNet-B0 the highest sensitivity. Qualitative analysis highlights distinct segmentation patterns, providing practical guidelines for encoder selection according to the application objective.*

Resumo. *A escolha do encoder em arquiteturas híbridas de segmentação impacta diretamente o desempenho do modelo. Este estudo avalia cinco encoders integrados à uma TransUNet para segmentação de tumores ósseos, além de compará-la com arquiteturas de referência. Os resultados mostram diferenças relevantes entre os modelos, com melhor desempenho global para o ConvNeXt-Tiny, maior precisão para a ResNet34 e maior sensibilidade para a EfficientNet-B0. A análise qualitativa evidencia padrões distintos de segmentação, fornecendo diretrizes práticas para a escolha do encoder conforme o objetivo da aplicação.*

1. Introdução

Os tumores ósseos primários representam um desafio crítico na medicina devido à sua heterogeneidade e à tendência de se desenvolverem em regiões anatomicamente complexas [Li et al. 2026]. A identificação precisa da extensão dessas lesões em imagens radiográficas é o primeiro passo essencial para orientar intervenções multidisciplinares, que podem incluir desde a ressecção cirúrgica com o uso de *scaffolds*¹ até protocolos de radioterapia avançada [Bădilă et al. 2021, Li et al. 2026]. Nesse cenário, o processamento digital de imagens é vital para delimitar a área afetada e garantir a preservação de tecidos saudáveis adjacentes [Kaliraj et al. 2026].

Recentemente, o uso de *Deep Learning* tem transformado a análise radiológica. Embora modelos como o YOLO tenham demonstrado eficácia na detecção de lesões

¹*Scaffolds* são estruturas tridimensionais utilizadas na engenharia de tecidos para servir como suporte ao crescimento celular e à regeneração tecidual.

em radiografias de campo total [Li et al. 2023], a prática clínica exige um detalhamento superior para o planejamento de terapias localizadas. A segmentação de tumores surge como a solução necessária para definir os contornos exatos da massa lesional. No entanto, este processo é dificultado pela textura heterogênea do osso, pela sobreposição de estruturas anatômicas e pelas fronteiras pouco claras das lesões, o que torna a delimitação manual altamente dependente da experiência do radiologista e suscetível a variabilidades [Kalejahi et al. 2026, Kaliraj et al. 2026].

Para superar as limitações das Redes Neurais Convolucionais (CNNs) tradicionais na captura de relações contextuais globais, a arquitetura TransUNet introduziu uma abordagem híbrida, fundindo a capacidade de localização das convoluções com o mecanismo de atenção dos *Transformers*. Um componente determinante para o sucesso desta abordagem é o *encoder* ou *backbone*, responsável pela extração inicial de características fundamentais da imagem. A escolha de *backbones* modernos, que equilibrem a riqueza semântica dos mapas de características com a eficiência computacional, é vital para que estes sistemas possam ser integrados em fluxos de trabalhos clínicos reais [Kalejahi et al. 2026].

Neste contexto, este trabalho investiga sistematicamente o impacto da escolha de diferentes *encoders* (ResNet34, ResNet50, DenseNet121, EfficientNet-B0 e ConvNeXt-Tiny) integrados à arquitetura híbrida *TransUNetWithGUA*, cujo mecanismo de fusão *GUA* foi previamente proposto em trabalho anterior. Diferentemente de estudos anteriores focados predominantemente em detecção [Li et al. 2023] ou classificação multi-classe [Kalejahi et al. 2026], esta pesquisa conduz uma análise controlada do papel do *backbone* na segmentação de tumores ósseos primários, mantendo constantes os demais componentes da arquitetura.

As principais contribuições deste trabalho são: (i) a análise do impacto de diferentes *encoders* em uma arquitetura híbrida de segmentação; (ii) uma comparação controlada entre arquiteturas convolucionais clássicas e modernas; e (iii) uma avaliação quantitativa e qualitativa do desempenho dos modelos. Os resultados evidenciam que a escolha do *backbone* influencia diretamente o desempenho da segmentação.

2. Trabalhos Relacionados

A literatura recente tem avançado no desenvolvimento de ferramentas de auxílio ao diagnóstico de tumores ósseos, impulsionada pela complexidade da interpretação radiológica. [Yao et al. 2025] destacam a escassez de dados públicos como principal limitação histórica, propondo o dataset BTXRD para treinamento de modelos de classificação e segmentação. No entanto, a exploração de arquiteturas mais eficientes para extração de características em tecidos heterogêneos permanece em aberto.

No contexto arquitetural, [Bloier et al. 2022] utilizaram *transfer learning* e aumento de dados para segmentação em radiografias limitadas, evidenciando dificuldades de modelos puramente convolucionais em capturar a complexidade morfológica das lesões. De forma semelhante, [Xie et al. 2022] propuseram a *Merged U-Net*, que melhora a fusão multi-escala, mas ainda apresenta limitações na modelagem de contexto global. Já [Rhyou et al. 2025] combinaram YOLO e U-Net para segmentação, com bons resultados em tumores benignos, porém com limitações em cenários mais complexos.

Apesar desses avanços, a maioria das abordagens ainda utiliza *encoders* convencionais, com limitada exploração do impacto de arquiteturas mais modernas na qualidade da

segmentação. Neste contexto, este trabalho analisa o impacto de diferentes *encoders* em uma arquitetura TransUNet aplicada à segmentação de tumores ósseos.

3. Materiais e Método

Nesta seção, são descritos os procedimentos que compõem o método de segmentação de tumores ósseos proposto. As etapas adotadas no processo estão ilustradas na Figura 1.

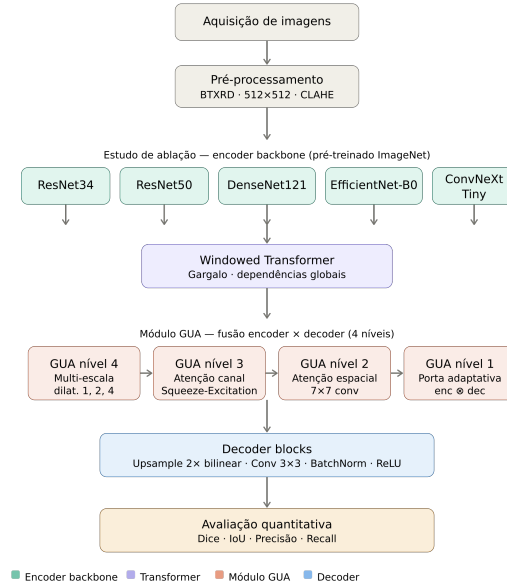


Figura 1. Fluxo metodológico proposto.

3.1. Aquisição de imagens

Os experimentos foram realizados com o conjunto de dados público BTXRD [Yao et al. 2025], que reúne 3746 radiografias ósseas, distribuídas entre 1879 amostras de ossos normais, 1525 casos de tumores benignos e 342 de tumores malignos. Além das imagens, o dataset disponibiliza anotações detalhadas no formato de máscaras de segmentação pixel a pixel, o que permite a avaliação direta de métodos de segmentação semântica. Dessa forma, torna-se possível ir além de abordagens puramente classificatórias, incorporando informações relevantes sobre a localização e a morfologia das lesões. Adicionalmente, o conjunto de dados inclui metadados clínicos associados, como idade, sexo, localização anatômica e subtipo tumoral.

Com o intuito de assegurar a reprodutibilidade dos experimentos, os dados foram particionados de maneira determinística em três subconjuntos distintos: treinamento, validação e teste, garantindo total independência entre eles. Foi adotada a proporção de 80%, 10% e 10% para cada subconjunto, respectivamente. As listas de amostras correspondentes a cada divisão foram previamente definidas e mantidas inalteradas ao longo de todos os experimentos. Exemplos representativos das imagens presentes no dataset podem ser visualizados na Figura 2.

3.2. Pré-processamento das imagens

As imagens foram previamente redimensionadas para uma resolução de 512×512 pixels, convertidas em tensores com valores no intervalo $[0, 1]$ e, em seguida, normalizadas

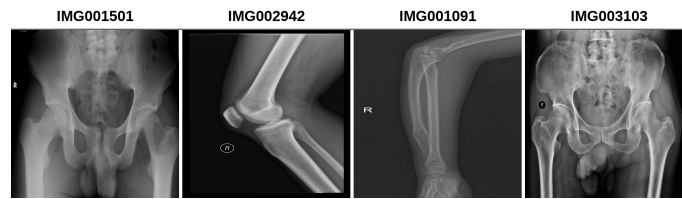


Figura 2. Exemplos de imagens normais e com lesão do conjunto de dados.

por canal utilizando média 0.5 e desvio padrão 0.5, resultando em valores finais no intervalo $[-1, 1]$. Devido às limitações típicas das radiografias convencionais, como baixo contraste e sobreposição de estruturas anatômicas, foi aplicada uma etapa adicional de pré-processamento voltada ao realce de contraste. Especificamente, utilizou-se o método *Contrast Limited Adaptive Histogram Equalization* (CLAHE) [Zuiderveld 1994] antes da normalização, com o intuito de destacar regiões relevantes e favorecer a extração de características discriminativas pelo modelo. A Figura 3 apresenta exemplos comparativos de imagens antes e após a aplicação do pré-processamento.

As máscaras de anotação originais foram transformadas para um formato binário, distinguindo explicitamente as classes normal e tumor. Essa adaptação viabiliza o treinamento supervisionado do modelo no contexto de segmentação semântica binária.



Figura 3. Exemplos de imagens sem e com pré-processamento.

3.3. Encoder

O *encoder* é um componente fundamental da arquitetura proposta, responsável pela extração hierárquica de características da imagem de entrada. A escolha do *encoder* determina diretamente a qualidade das representações aprendidas e, consequentemente, o desempenho final do sistema de segmentação. Para esta investigação, foram selecionados cinco encoders representativos, abrangendo desde arquiteturas clássicas até *designs* modernos inspirados em *Transformers*, cada um com características distintas que justificam sua inclusão no estudo comparativo.

A ResNet34, proposta por He et al. [He et al. 2016], introduziu o conceito de conexões residuais, que revolucionaram o treinamento de redes profundas. Essa arquitetura consolidou-se como *baseline* na literatura de segmentação médica devido à sua excelente relação custo-desempenho. A estrutura residual permite treinamento estável mesmo em cenários com dados limitados, característica particularmente relevante em aplicações médicas, onde a disponibilidade de dados anotados é restrita. A ResNet34 foi adotada como referência por sua ampla validação em tarefas de segmentação e por representar o estado da arte das arquiteturas convolucionais clássicas.

A ResNet50, também proposta por He et al. [He et al. 2016], amplia a profundidade da arquitetura residual para 50 camadas, proporcionando maior capacidade representacional em relação à ResNet34. Essa maior profundidade favorece a captura de padrões tumorais mais complexos e texturas irregulares, frequentemente observadas em lesões ósseas [Das and Majumder 2020]. Sua inclusão neste estudo tem como objetivo avaliar se o aumento da profundidade do encoder, mantendo a mesma filosofia arquitetural, resulta em melhorias significativas no desempenho da segmentação.

A DenseNet121, proposta por Huang et al. [Huang et al. 2017], introduz conexões densas entre camadas, nas quais cada camada recebe como entrada os mapas de características de todas as camadas anteriores. Essa estratégia promove forte reutilização de características e melhora o fluxo de gradiente, resultando em treinamento mais estável e convergência mais eficiente [Jégou et al. 2017]. Essas propriedades são particularmente relevantes em problemas médicos, onde a preservação de detalhes finos é essencial. A DenseNet121 foi selecionada por sua eficiência paramétrica e por representar uma abordagem alternativa baseada em conexões densas.

A EfficientNet-B0, proposta por Tan e Le [Tan and Le 2019], apresenta uma estratégia sistemática de escalonamento de redes neurais, equilibrando profundidade, largura e resolução de entrada. Essa abordagem permite atingir alta eficiência paramétrica sem comprometer a capacidade de modelagem. A arquitetura tem sido amplamente adotada em aplicações médicas [Naseer et al. 2025], especialmente em cenários com restrições computacionais, devido ao seu baixo custo em termos de parâmetros e operações.

O ConvNeXt-Tiny, proposto por Liu et al. [Liu et al. 2022], representa uma evolução das redes convolucionais modernas ao incorporar princípios de design introduzidos pelos *Vision Transformers* (ViTs). A arquitetura utiliza convoluções *depthwise* com kernels de maior dimensão (7×7), normalização por camada (*LayerNorm*) e substitui funções de ativação tradicionais por *GELU* [Liu et al. 2022]. Essas modificações ampliam o campo receptivo e melhoram a estabilidade do aprendizado em tarefas densas, permitindo uma extração de características com alta fidelidade espacial. O ConvNeXt-Tiny foi selecionado por representar o estado da arte em CNNs puras, oferecendo um desempenho competitivo aos Transformers, mas com a eficiência indutiva intrínseca das convoluções.

A escolha desses cinco encoders permite uma análise abrangente do impacto do *backbone* na arquitetura proposta, contemplando diferentes filosofias de projeto: conexões residuais (ResNet34 e ResNet50), conexões densas (DenseNet121), eficiência paramétrica (EfficientNet-B0) e arquiteturas modernas inspiradas em *Transformers* (ConvNeXt-Tiny). Essa diversidade possibilita isolar o efeito do *encoder* no desempenho final, mantendo constantes os demais componentes do modelo.

3.4. Simple Windowed Transformer

Com o objetivo de enriquecer as representações no *bottleneck*, foi incorporado um *Simple Windowed Transformer*, aplicado exclusivamente sobre mapas de características reduzidos, o que permite capturar dependências globais com menor custo computacional. A tokenização é realizada por meio da divisão dos mapas em *patches* fixos não sobrepostos, seguida da aplicação de *self-attention* global entre os tokens gerados. Como o *bottleneck* já contém representações semanticamente ricas e espacialmente compactadas pelas camadas convolucionais anteriores, torna-se mais relevante modelar relações contextuais globais do

que reforçar redundâncias espaciais locais.

A inclusão desse módulo permite que o *GUA-Fusion* receba informações previamente refinadas por uma compreensão global da imagem, favorecendo a reconstrução progressiva da máscara de segmentação com melhor delimitação de bordas e maior contextualização anatômica da lesão.

3.5. *GUA-Fusion*: Fusão Multi-escala com Atenção e Gating Adaptativo

Como principal componente arquitetural, utiliza-se o bloco *GUA-Fusion* (*Grouped Unified Attention Fusion*), previamente proposto em trabalho anterior, com o objetivo de aprimorar as *skip connections* em arquiteturas do tipo U-Net. Diferentemente da concatenação direta entre *encoder* e *decoder*, o *GUA-Fusion* realiza uma integração seletiva das informações, combinando fusão multi-escala, mecanismos de atenção e controle adaptativo do fluxo de informação.

Inicialmente, as características do *encoder* passam por uma etapa de fusão multi-escala baseada em convoluções dilatadas, ampliando o campo receptivo e permitindo capturar simultaneamente detalhes locais e contexto global. Em seguida, aplica-se atenção unificada, composta por atenção de canal e espacial. A atenção de canal recalibra os mapas de características por meio de *global average pooling* e camadas totalmente conectadas, enquanto a atenção espacial gera um mapa bidimensional que destaca regiões relevantes. Diferentemente de abordagens como o CBAM [Woo et al. 2018], esses mecanismos são integrados diretamente à *skip connection*, condicionando a informação transmitida ao *decoder*.

Por fim, um mecanismo de *gating* adaptativo regula dinamicamente a contribuição relativa das features do *encoder* e do *decoder*. Inspirado na *Attention U-Net* [Oktay et al. 2018], esse processo reduz a propagação de ruído e prioriza regiões semanticamente relevantes. Essa combinação permite uma fusão mais robusta, favorecendo a segmentação em radiografias com alta variabilidade estrutural e baixo contraste.

3.6. *Decoder*

O *decoder* é responsável pela reconstrução progressiva da resolução espacial por meio de operações de *upsampling*, combinando as representações semânticas de alto nível com detalhes espaciais de alta resolução provenientes do *encoder*. Essa integração é realizada por meio de *skip connections* aprimoradas pelo mecanismo *GUA-Fusion*, permitindo uma fusão mais eficiente e seletiva das informações ao longo do processo de reconstrução.

3.7. Configuração Experimental

Os experimentos foram implementados em *PyTorch*, utilizando uma GPU NVIDIA Tesla T4 com 16 GB de memória e suporte a precisão mista automática (*AMP*). Para garantir reprodutibilidade, foi adotada semente fixa igual a 42. As imagens foram redimensionadas para 512×512 pixels, normalizadas com os parâmetros do ImageNet e submetidas a pré-processamento com CLAHE para realce de contraste. Durante o treinamento, foram aplicadas técnicas de *data augmentation*, incluindo espelhamento horizontal, rotações de até $\pm 10^\circ$ e ajustes aleatórios de brilho e contraste.

O treinamento foi conduzido por até 150 épocas, com *batch size* igual a 4, utilizando o otimizador AdamW (taxa de aprendizado inicial de 10^{-4} e *weight decay* de 10^{-4}), com

ajuste da taxa de aprendizado via *Cosine Annealing* [Loshchilov and Hutter 2017]. Para estabilidade numérica, aplicou-se *gradient clipping* com norma máxima de 1.0. Como função de perda, utilizou-se uma combinação entre *Cross-Entropy Loss* ponderada e *Dice Loss*, adotando ponderação empírica para reduzir os efeitos do desbalanceamento presente no conjunto de dados.

3.8. Avaliação

A avaliação do modelo foi conduzida utilizando um conjunto de teste independente e previamente definido, garantindo a imparcialidade dos resultados. Foram empregadas métricas amplamente utilizadas em segmentação de imagens médicas, incluindo Dice, *Intersection over Union* (IoU), precisão, *recall*. Adicionalmente, foram realizadas análises quantitativas comparativas com resultados reportados na literatura, permitindo posicionar o desempenho da arquitetura proposta em relação a *benchmarks* consolidados.

4. Resultados e Discussão

Esta seção apresenta os resultados obtidos pela arquitetura proposta *TransUNetWithGUA*, bem como sua comparação com arquiteturas de referência amplamente utilizadas na literatura, com ênfase na avaliação do impacto da escolha do *encoder* na tarefa de segmentação binária de tumores ósseos em radiografias.

4.1. Análise quantitativa

Os resultados apresentados na Tabela 1 evidenciam que a arquitetura proposta supera os modelos de referência U-Net, U-Net++ e YOLOv8-seg em métricas de sobreposição, ao mesmo tempo em que demonstram que a escolha do *encoder* exerce influência significativa sobre a qualidade da segmentação. Entre os *encoders* avaliados, o modelo *baseline* com ResNet34 apresentou desempenho consistente, com bom equilíbrio entre precisão e *recall*, servindo como referência para comparação. A comparação com arquiteturas de re-

Tabela 1. Comparação entre arquiteturas da literatura e os *encoders* avaliados na arquitetura proposta.

Arquitetura	Dice	IoU	Precisão	Recall
U-Net	0.6090	0.4510	0.6490	0.6040
U-Net++	0.5510	0.4000	0.6590	0.5120
YOLOv8-seg [Yao et al. 2025]	–	–	0.8650	0.4220
ResNet34	0.7369	0.6798	0.8924	0.7370
ResNet50	0.6850	0.6219	0.8495	0.6978
DenseNet121	0.7075	0.6445	0.8270	0.7494
EfficientNet-B0	0.6959	0.6295	0.7979	0.7627
ConvNeXt-Tiny	0.7478	0.6835	0.8691	0.7514

ferência evidencia ganhos expressivos da abordagem proposta em métricas de sobreposição. Enquanto a U-Net e a U-Net++ apresentaram valores de Dice de 0.6090 e 0.5510, respectivamente, a arquitetura proposta superou ambas em todas as configurações avaliadas, alcançando seu melhor desempenho com o *encoder* ConvNeXt-Tiny (Dice = 0.7478; IoU = 0.6835). Em relação ao YOLOv8-seg [Yao et al. 2025], embora tenha sido observada elevada precisão (0.8650), o baixo *recall* (0.4220) sugere menor sensibilidade na identificação completa das regiões tumorais, característica crítica em cenários de segmentação médica.

A substituição pelo *encoder* ResNet50, apesar de sua maior profundidade e capacidade representacional, resultou em degradação de desempenho em relação ao baseline ResNet34. Esse comportamento sugere que o aumento da complexidade não se traduz necessariamente em melhor generalização no contexto de radiografias ósseas. O desempenho inferior da ResNet50 pode estar associado ao aumento da capacidade representacional em relação ao tamanho efetivo do dataset, favorecendo *overfitting*. Em bases médicas com número limitado de amostras anotadas, arquiteturas excessivamente profundas podem apresentar menor capacidade de generalização.

O *encoder* DenseNet121 apresentou melhora moderada em relação ao ResNet50, indicando que a reutilização intensiva de características favorece a preservação de estruturas relevantes. Entretanto, quando comparado ao baseline ResNet34, apresentou desempenho inferior em Dice, IoU e precisão, superando-o apenas em *recall*, o que sugere maior sensibilidade às regiões tumorais às custas de maior incidência de falsos positivos.

O EfficientNet-B0 apresentou comportamento distinto, alcançando o maior valor de *recall* entre os *encoders* avaliados (0.7627). Esse resultado indica maior sensibilidade na detecção de regiões tumorais, embora acompanhado de redução na precisão, sugerindo maior incidência de falsos positivos. Tal comportamento pode ser potencialmente útil em aplicações médicas onde minimizar falsos negativos seja prioritário.

O melhor desempenho global foi obtido com o ConvNeXt-Tiny, que alcançou os maiores valores de Dice e IoU. Esse resultado pode ser explicado por sua capacidade de combinar o viés indutivo das arquiteturas convolucionais com princípios modernos inspirados em Vision Transformers, como convoluções com *kernels* maiores, normalização mais estável e maior campo receptivo efetivo. Essas características favorecem a modelagem de padrões espaciais complexos e a segmentação de bordas tumorais irregulares, particularmente relevantes em radiografias ósseas. Além disso, a compatibilidade entre o *encoder* ConvNeXt-Tiny e os módulos propostos — especialmente o Transformer no *bottleneck* e o mecanismo GUA-Fusion — sugere uma cooperação eficaz entre modelagem local refinada e integração contextual global.

De forma geral, os resultados demonstram que o desempenho da arquitetura proposta não depende apenas dos mecanismos de atenção e fusão introduzidos, mas também da capacidade do *encoder* em fornecer representações discriminativas adequadas. Esses achados reforçam a importância da escolha do *backbone* em arquiteturas híbridas e indicam o ConvNeXt-Tiny como a opção mais adequada dentre as avaliadas para a segmentação de tumores ósseos em radiografias.

4.2. Análise Qualitativa

A análise qualitativa foi conduzida com o objetivo de avaliar o comportamento da arquitetura TransUNetWithGUA em diferentes cenários desafiadores, evidenciando o impacto da escolha do *encoder* na qualidade da segmentação. Embora arquiteturas de referência tenham sido incluídas na análise quantitativa para fins de posicionamento comparativo, a análise qualitativa concentra-se na comparação entre os *encoders* da arquitetura proposta, por representar o foco principal desta investigação. As Figuras 4–7 apresentam exemplos representativos, nos quais cada coluna corresponde a um *encoder* distinto. A máscara predita é representada por contorno amarelo, enquanto a anotação de referência (*ground truth*) é indicada em branco. As regiões corretamente segmentadas são destacadas em verde

(verdadeiros positivos), ao passo que regiões em vermelho representam falsos positivos e regiões em azul indicam falsos negativos.

Na Figura 4, que ilustra um caso com bordas tumorais relativamente bem definidas, observa-se que todos os *encoders* conseguem identificar corretamente a região tumoral principal. No entanto, diferenças sutis podem ser observadas na delimitação das extremidades da lesão. *Encoders* como ResNet34 e ResNet50 apresentam pequenas regiões de falso negativo nas bordas, enquanto DenseNet121 e EfficientNet-B0 demonstram maior variabilidade na aderência ao contorno. O ConvNeXt-Tiny apresenta melhor alinhamento global, com menor incidência de erros localizados.

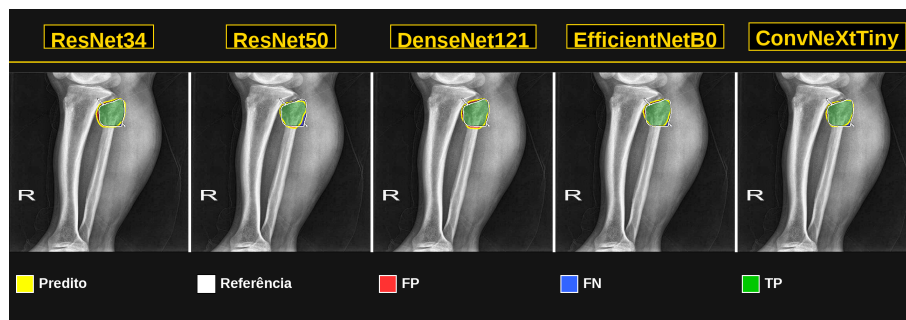


Figura 4. Exemplo 1 de segmentação de tumor ósseo em radiografia obtida por cada *encoder*.

A Figura 5 apresenta um cenário mais desafiador, caracterizado por baixo contraste e presença de artefatos metálicos. Nesse contexto, as diferenças entre os *encoders* tornam-se mais evidentes. Enquanto a ResNet50 introduz um falso positivo em região adjacente, DenseNet121 e EfficientNet-B0 apresentam subsegmentação mais acentuada, evidenciada por regiões de falso negativo. O ConvNeXt-Tiny mantém um comportamento mais equilibrado, conseguindo capturar a maior parte da lesão com menor incidência de erros.

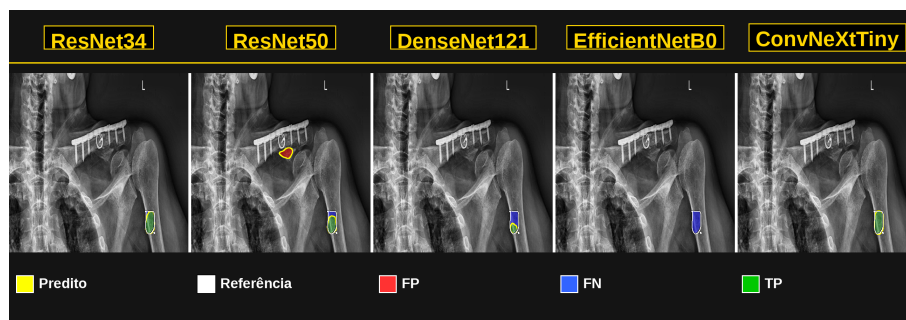


Figura 5. Exemplo 2 de segmentação de tumor ósseo em radiografia obtida por cada *encoder*.

Na Figura 6, é apresentado um caso de lesão tumoral de pequena dimensão. Nesse cenário, pequenas discrepâncias tornam-se mais relevantes devido à reduzida área da lesão. Todos os *encoders* conseguem localizar a região tumoral, porém com variações na precisão da delimitação. DenseNet121 e EfficientNet-B0 apresentam maior fragmentação e perda parcial da região, enquanto ResNet34 e ResNet50 mantêm desempenho consistente. O

ConvNeXt-Tiny novamente se destaca pela melhor preservação das estruturas finas e maior alinhamento com a máscara de referência.

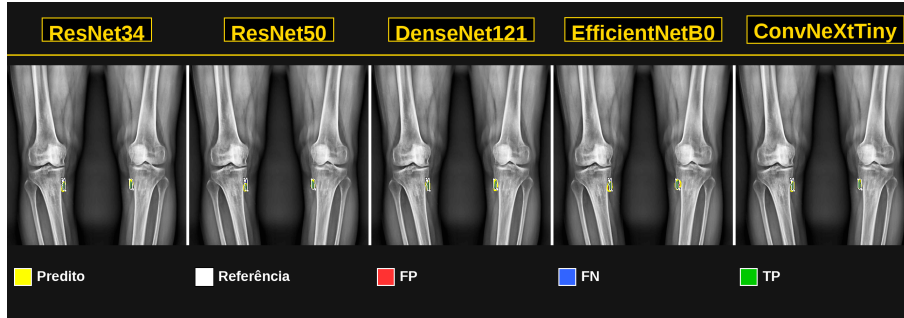


Figura 6. Exemplo 3 de segmentação de tumor ósseo em radiografia obtida por cada *encoder*.

Por fim, a Figura 7 ilustra um caso com presença de falso positivo significativo. Observa-se que DenseNet121 e EfficientNet-B0 ativam regiões extensas fora da área tumoral, indicando dificuldade na discriminação de padrões visuais semelhantes. Em contraste, ResNet34 e ResNet50 apresentam maior controle dessas ativações. O ConvNeXt-Tiny demonstra maior robustez, mantendo a segmentação concentrada na região tumoral e evitando ativações indevidas.

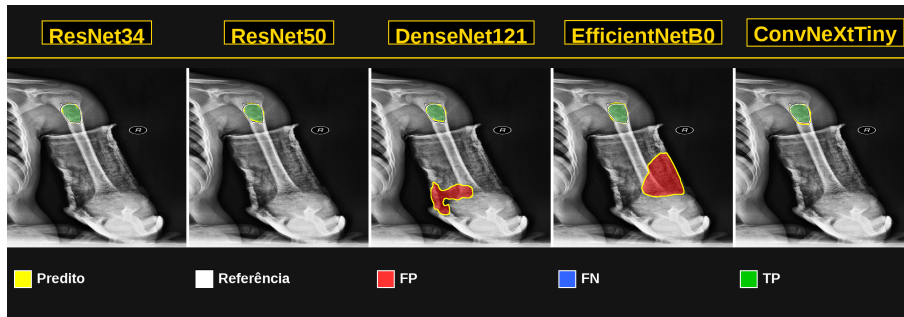


Figura 7. Exemplo 4 de segmentação de tumor ósseo em radiografia obtida por cada *encoder*.

De maneira geral, os resultados qualitativos evidenciam que a escolha do *encoder* impacta não apenas a capacidade de detecção da lesão, mas também o equilíbrio entre falsos positivos e falsos negativos, bem como a precisão na delimitação das bordas tumorais. *Encoders* mais modernos, como o ConvNeXt-Tiny, apresentam maior robustez em cenários desafiadores, contribuindo para uma segmentação mais confiável. Esses achados corroboram os resultados quantitativos, reforçando a importância da escolha do *encoder* na arquitetura proposta.

Com relação aos trabalhos apresentados na Seção 2, como Bloier et al. (2022) com arquiteturas convolucionais, e Xie et al. (2022) com a Merged U-Net, apresentam limitações na captura de contexto global, segmentação de bordas difusas e modelagem de dependências de longo alcance. A abordagem híbrida de Rhyou et al. (2025) (YOLO + U-Net) melhora a detecção inicial, mas ainda falha em segmentações precisas em cenários complexos (baixo contraste, artefatos). Além disso, diferentemente da maioria

dos trabalhos correlatos, este estudo demonstra que a escolha do *encoder* exerce impacto significativo no desempenho, influenciando diretamente o equilíbrio entre precisão e sensibilidade e destacando a importância das representações iniciais no resultado final.

5. Conclusão

A análise de ablação demonstrou que o desempenho da arquitetura depende diretamente do *encoder*, sendo que modelos mais modernos — com melhor modelagem espacial e fluxo de informações — lidam de forma mais eficaz com a variabilidade dos tumores ósseos. Contudo, em cenários com dados limitados, o aumento da profundidade não garante maior generalização, o que torna a escolha do *encoder* e sua integração com a rede fatores críticos. Como limitação, restrições computacionais impuseram o uso de uma divisão fixa dos dados e uma execução única por configuração, impossibilitando análises estatísticas de variabilidade. Para trabalhos futuros, sugere-se investigar a interação entre os componentes do sistema, incorporar validações estatísticas robustas e explorar *encoders* híbridos combinados a estratégias de pré-treinamento específicas para imagens médicas.

Agradecimentos

O presente trabalho foi realizado com apoio da UFPI/UFMA e da Fundação de Amparo à Pesquisa do Estado do Piauí (FAPEPI) e da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior- Brasil (CAPES).

Referências

- Bădilă, A. E., Rădulescu, D. M., Niculescu, A. G., Grumezescu, A. M., Rădulescu, M., and Rădulescu, A. R. (2021). Recent advances in the treatment of bone metastases and primary bone tumors: an up-to-date review. *Cancers*, 13(16):4229.
- Bloier, M., Hinterwimmer, F., Breden, S., Consalvo, S., Neumann, J., Wilhelm, N., and Burgkart, R. (2022). Detection and segmentation of heterogeneous bone tumours in limited radiographs. *Current Directions in Biomedical Engineering*, 8(2):69–72.
- Das, S. and Majumder, S. (2020). Lung cancer detection using deep learning network: A comparative analysis. In *Proceedings of the Fifth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*, pages 30–35. IEEE.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.
- Huang, G., Liu, Z., Maaten, L. V. D., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708.
- Jégou, S., Drozdal, M., Vazquez, D., Romero, A., and Bengio, Y. (2017). The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 11–19.
- Kalejahi, B. K., Khan, S., and Zakirov, R. (2026). Bone-cnn: A lightweight deep learning architecture for multi-class classification of primary bone tumours in radiographs. *Biomedicines*, 14(2):299.

- Kaliraj, V., Mangaleswaran, M., Usha, N. S., and Dharaniya, R. (2026). Automated artificial intelligent approach for enhancing bone cancer detection through hybrid feature extraction and adaptive elman recurrent neural network. *Radiation Physics and Chemistry*, page 113732.
- Li, J., Li, S., Li, X., Miao, S., Dong, C., Gao, C., and Cui, J. (2023). Primary bone tumor detection and classification in full-field bone radiographs via yolo deep learning model. *European Radiology*, 33(6):4237–4248.
- Li, L., Tao, W., Eid, J. E., Gupta, S., Zheng, S., Taswell, C. S., and Trent, J. C. (2026). Advances in radiation therapy for primary bone malignancies. *Technology in Cancer Research & Treatment*, 25:15330338261425345.
- Liu, Z., Mao, H., Wu, C. Y., Feichtenhofer, C., Darrell, T., and Xie, S. (2022). A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11976–11986.
- Loshchilov, I. and Hutter, F. (2017). Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)*.
- Naseer, M. M., Veerabhadraiah, V., Ramji, B. R., Lokeshappa, S. K. K., Nagendra, S. C., and Ramesh, Y. G. (2025). An optimized efficientnet-b0 framework for multi-class brain tumour detection and classification from mri images. *Biomedical and Pharmacology Journal*, 18. October Special Edition.
- Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., Glocker, B., and Rueckert, D. (2018). Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*.
- Rhyou, S. Y., Bang, C., Cho, Y. J., Bae, H., Ha, Y. J., Baek, S. Y., and Moon, J. E. (2025). Tracking temporal progression of benign bone tumors through x-ray based detection and segmentation. *Scientific Reports*, 15(1):39491.
- Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 6105–6114. PMLR.
- Woo, S., Park, J., Lee, J., and Kweon, I. S. (2018). Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19.
- Xie, Z., Zhao, K., Yan, X., Wu, S., Mei, J., and Lu, H. (2022). Merged u-net for bone tumors x-ray images segmentation. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 1276–1280. IEEE.
- Yao, S., Huang, Y., Wang, X., Zhang, Y., Paixao, I. C., Wang, Z., and Song, J. (2025). A radiograph dataset for the classification, localization, and segmentation of primary bone tumors. *Scientific Data*, 12(1):88.
- Zuiderveld, K. J. (1994). Contrast limited adaptive histogram equalization. In Heckbert, P. S., editor, *Graphics Gems*, pages 474–485. Elsevier.