Avaliação Multidimensional da Segmentação Mamária em Ultrassonografia Baseada em Hierarchical Transformer

Murilo Salem
Universidade Federal de Pelotas (UFPel)
Pelotas - RS
mcsalem@inf.ufpel.edu.br

Daniel Barretos
Universidade Federal de Pelotas (UFPel)
Pelotas - RS
dhspbarretos@inf.ufpel.edu.br

Luísa Böhm
Universidade Federal de Pelotas (UFPel)
Pelotas - RS
lcbohm@inf.ufpel.edu.br

Henrique dos Reis
Universidade Federal de Pelotas (UFPel)
Pelotas - RS
hdreis@inf.ufpel.edu.br

Marcos Lima
Universidade Federal de Pelotas (UFPel)
Pelotas - RS
mlalves@inf.ufpel.edu.br

Anderson Ferrugem
Universidade Federal de Pelotas (UFPel)
Pelotas - RS
ferrugem@inf.ufpel.edu.br

Abstract

Breast cancer remains a leading cause of morbidity and mortality worldwide, making accurate diagnosis in ultrasound imaging a critical challenge due to inherent speckle noise, low contrast, and variable lesion morphology. In this work, we propose a deep learning framework for breast ultrasound segmentation that integrates a Hierarchical Mix Transformer (MiT-b2) encoder with a U-Net decoder. Unlike traditional convolutional networks, this architecture leverages efficient self-attention mechanisms to capture global contextual dependencies while preserving fine-grained spatial details through multi-scale feature fusion. This approach can help women by enabling earlier and more accurate detection, potentially reducing the number of deaths from breast cancer.

To ensure reliability and reproducibility, experiments were conducted on the Breast Ultrasound Images Dataset (BUSI) using a rigorous protocol involving multiple random seeds, Test-Time Augmentation (TTA), and adaptive thresholding. The proposed method demonstrated high stability and performance, achieving a mean Dice Coefficient of 0.7956, an Intersection over Union (IoU) of 0.6728, and a remarkably high Specificity of 0.9883, indicating effective suppression of false positives. Furthermore, geometric evaluation yielded an average Hausdorff Distance (95%) of 28.14 pixels, validating the model's boundary delineation capabilities. These findings suggest that hierarchical Transformer-based models provide a robust and clinically consistent solution for computer-aided diagnosis in breast ultrasound.

Keywords

Breast Ultrasound, Medical Image Segmentation, Hierarchical Transformer, U-Net, Deep Learning, Computer-Aided Diagnosis, Artificial Intelligence

1 Introdução

O câncer de mama permanece como uma das principais causas de morbimortalidade entre mulheres em escala global, tornando a detecção precoce e a identificação das lesões fatores decisivos para o prognóstico e a escolha de estratégias terapêuticas menos invasivas [7]. Nesse contexto, o ultrassom mamário é essencial por ser uma modalidade acessível, não ionizante e eficaz em mamas densas, sendo amplamente utilizada em protocolos de rastreamento, triagem complementar e acompanhamento clínico. No entanto, a interpretação dessas imagens é desafiadora, devido a características como baixo contraste, ruído speckle, contornos imprecisos e elevada variabilidade interpaciente, o que torna a segmentação manual subjetiva e dependente da experiência do especialista.

Métodos automáticos de segmentação baseados em aprendizado profundo surgem como ferramentas promissoras para mitigar essas limitações, oferecendo maior consistência e reprodutibilidade, conforme discutido em desafios similares de diagnóstico por imagem [10]. Arquiteturas do tipo U-Net consolidaram-se como referência na área [4], enquanto modelos baseados em Transformers vêm ganhando destaque por sua capacidade de integrar contexto global e dependências de longo alcance. Em particular, o *Hierarchical Mix Transformer* introduz um mecanismo hierárquico eficiente de autoatenção que combina sensibilidade a padrões locais e globais [18], característica especialmente relevante em imagens de ultrassom mamário, onde estruturas sutis coexistem com variações anatômicas de maior escala e baixa relação sinal-ruído.

Embora métricas tradicionais de classificação, como acurácia, sejam amplamente utilizadas, elas se mostram insuficientes para avaliar a performance de modelos de segmentação médica, principalmente em estruturas pequenas, desbalanceamento de classes ou variações geométricas complexas. Nesse sentido, este trabalho propõe uma avaliação rigorosa e multidimensional de um modelo de segmentação baseado em Hierarchical Transformer com U-Net aplicado ao Breast Ultrasound Images Dataset [2], adotando um protocolo experimental totalmente determinístico e enfatizando análises estatísticas, geométricas, computacionais e qualitativas, com foco na interpretação clínica e na reprodutibilidade dos resultados.

2 Trabalhos Relacionados

As Arquiteturas *U-Net* e suas variações (*Attention U-Net*, *U-Net++*) consolidaram-se na segmentação médica [3, 4], mas limitam-se a dependências locais. *Transformers* hierárquicos (*Swin*, *SegFormer*) capturam contexto global via autoatenção [18], equilibrando eficiência e modelagem contextual.

Contudo, práticas de avaliação permanecem limitadas: estudos reportam apenas Dice/IoU médios, sem métricas geométricas (HD95,

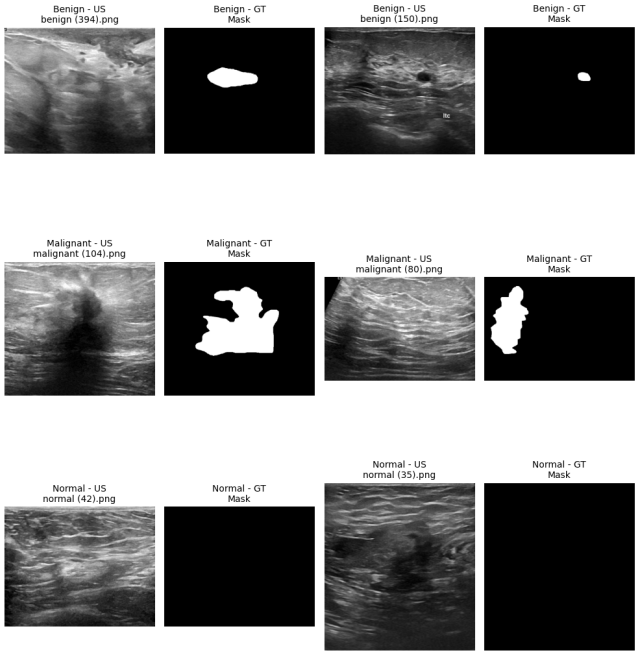


Figure 1: Amostras representativas do Breast Ultrasound Images Dataset (BUSI). As linhas mostram exemplos de casos benignos, malignos e normais. À esquerda, a imagem original de ultrassom (B-mode); à direita, a máscara de segmentação manual (*Ground Truth*) sobreposta. Observam-se desafios da modalidade, como baixo contraste, ruído *speckle* e sombras acústicas.

ASSD) ou robustas a desbalanceamento (MCC) [15]. Este trabalho preenche esta lacuna via avaliação multidimensional de *Transformers* hierárquicos em ultrassom mamário.

3 Metodologia

3.1 Dataset

Os experimentos foram conduzidos utilizando o *Breast Ultrasound Images Dataset* (BUSI) [2], disponibilizado publicamente e amplamente empregado em estudos de segmentação de lesões mamárias por ultrassom [16]. O conjunto de dados é composto por imagens de ultrassom bidimensionais acompanhadas de máscaras de segmentação manual, delineando as regiões de lesão. As amostras incluem diferentes tumores e imagens normais, refletindo a variabilidade típica encontrada em cenários clínicos reais e desafios de classificação diagnóstica [12].

As imagens apresentam resolução variável e características inerentes ao ultrassom mamário, como baixo contraste, presença de ruído speckle e contornos de lesão pouco definidos [16]. Essas propriedades tornam o dataset desafiador e adequado para a avaliação de métodos de segmentação robustos, impondo dificuldades similares às encontradas na segmentação de outros tecidos biológicos complexos [8]. Todas as imagens e respectivas máscaras foram carregadas diretamente a partir do diretório /kaggle/input/breast-ultrasound-images-dataset [2].

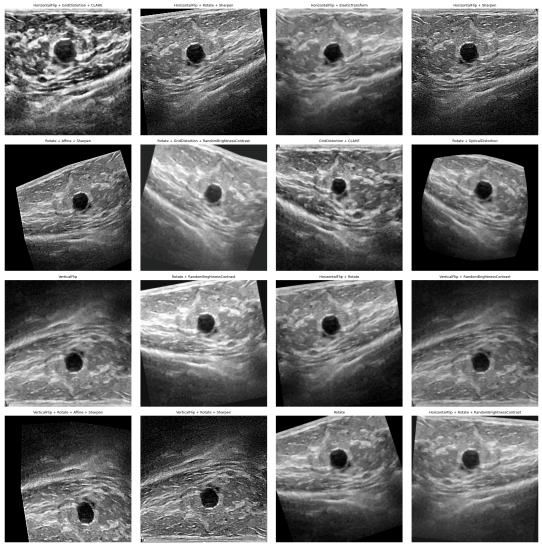


Figure 2: Exemplos de data augmentation realizado em imagens de ultrassom.

3.2 Pré-Processamento e Aumentação de Dados

Visando padronizar a entrada da rede e preservar detalhes finos das bordas das lesões, todas as imagens foram redimensionadas para a resolução de 384×384 pixels. Esta resolução é superior ao padrão de 224×224 frequentemente utilizado na literatura [18], permitindo uma análise mais granular da morfologia tumoral e texturas do tecido [12]. A normalização das intensidades seguiu os parâmetros do ImageNet (média e desvio padrão) [5].

Dada a escassez de dados anotados e a alta variabilidade intraclasse do ultrassom, aplicou-se um pipeline robusto de *data augmentation* utilizando a biblioteca Albumentations. Estratégias de aumento de dados são cruciais para reduzir o risco de superajuste (*overfitting*) e melhorar a generalização em datasets médicos desbalanceados [13]. A estratégia foi dividida em dois eixos:

(1) **Transformações Geométricas:** Projetadas para simular variações de posicionamento do transdutor e deformações teciduais não lineares (elasticidade) que ocorrem durante a compressão da mama no exame.
(2) **Transformações de Intensidade:** Focadas em simular diferentes ajustes de ganho do equipamento, variabilidade de ruído *speckle* e realce de bordas.

A Tabela 1 detalha as operações aplicadas, seus respectivos parâmetros e a justificativa clínica para sua inclusão no pipeline de treinamento.

3.3 Divisão do Conjunto de Dados e Reprodutibilidade

Para garantir a robustez da avaliação e mitigar vieses associados à aleatoriedade na inicialização dos pesos e na seleção das amostras, adotou-se um protocolo experimental rigoroso, alinhado a práticas de avaliação comparativa em segmentação médica [4]. O conjunto de dados foi particionado aleatoriamente em subconjuntos de treinamento (70%), validação(15%) e teste (15%), mantendo a estratificação

Table 1: Pipeline de *Data Augmentation*: Parâmetros e Justificativas Clínicas.

Transformação	Parâmetros (Prob.)	Justificativa Clínica/Técnica
Transformações Geométricas		
Horizontal/Vertical Flip	$p = 0.5$	Invariância à orientação do transdutor.
Rotação	$\pm 20°$, $p = 0.5$	Variações angulares na aquisição manual.
Elastic Transform	$\alpha = 120, \sigma = 6$	Simula a deformação não-linear de tecidos moles (mama) sob compressão da sonda.
Grid Distortion	Padrão Albumentations	
Optical Distortion	Limite=1	
Affine (Shear/Scale)	Scale $\pm 10\%$, Shear $\pm 10°$	Mudanças de escala e perspectiva da lesão.
Transformações de Intensidade e Textura		
CLAHE	Clip=4.0, Grid=(8,8)	Simula ajustes de ganho do aparelho, contraste variável e realce de estruturas em mamas densas.
Sharpen	$p = 0.5$	
Random Brightness	$p = 0.5$	

das classes. O conjunto de validação foi empregado para *Early Stopping*, ajuste de *Learning Rate Scheduler* e otimização do *threshold* de binarização. O conjunto de teste foi reservado exclusivamente para avaliação final.

Os experimentos foram repetidos três vezes de forma independente, utilizando sementes aleatórias distintas (*random seeds*: 42, 53 e 2025). Essa abordagem permite avaliar a estabilidade do método frente à estocasticidade e sensibilidade à inicialização características do treinamento de redes baseadas em Transformers [11]. Os resultados a seguir representam a média e o desvio padrão obtidos através dessas três execuções, garantindo que as métricas de desempenho refletem a capacidade de generalização da arquitetura e não artefatos de uma divisão específica.

3.4 Arquitetura do Modelo

A segmentação das lesões foi realizada utilizando uma arquitetura U-Net híbrida, que integra a capacidade de reconstrução espacial da U-Net com um codificador (*encoder*) baseado em Transformers hierárquicos. Especificamente, empregou-se o **Mix Transformer (MiT-b2)**, backbone da arquitetura SegFormer, pré-treinado no ImageNet [18].

Diferentemente de Vision Transformers (ViT) padrão que geram mapas de características de resolução única [6], o MiT gera representações multiescala hierárquicas ($1/4, 1/8, 1/16, 1/32$ da resolução original), o que é crucial para a arquitetura U-Net fundir detalhes semânticos e espaciais através de conexões de salto, conforme estabelecido em arquiteturas de segmentação médica [4]. O MiT-b2 utiliza *Overlapping Patch Embeddings* para preservar a continuidade local e um mecanismo de *Efficient Self-Attention*, que reduz a complexidade computacional quadrática [18], permitindo o processamento de imagens de alta resolução (384×384) de forma viável. O decodificador reconstrói o mapa de segmentação binária através de convoluções progressivas e upsampling bilinear.

3.5 Protocolo de Treinamento

O treinamento foi conduzido minimizando a função de perda *Focal Tversky Loss* [1], projetada especificamente para lidar com o severo desbalanceamento entre as classes de fundo e lesão, um desafio crítico em datasets médicos que exige estratégias de compensação durante o aprendizado [13], além de penalizar falsos negativos (parâmetros $\alpha = 0.3$, $\beta = 0.7$, $\gamma = 4/3$). A otimização dos pesos utilizou o algoritmo AdamW [9] com taxa de aprendizado inicial de 1×10^{-4} e decaimento de peso de 1×10^{-2}.

Parâmetro / Estratégia	Valor / Configuração
Configuração Principal	
Função de Perda	Focal Tversky Loss
Parâmetros da Perda (α, β, γ)	$\alpha = 0.3$, $\beta = 0.7$, $\gamma = 4/3$
Otimizador	AdamW
Taxa de Aprendizado (Inicial)	1×10^{-4}
Decaimento de Peso (*Weight Decay*)	1×10^{-2}
Tamanho do Lote (*Batch Size*)	8
Controle de Convergência	
Épocas Máximas	100
Early Stopping (Paciência)	30 épocas
Scheduler (Ajuste da LR)	ReduceLROnPlateau
Condição do *Scheduler*	Redução pela metade após 4 épocas sem melhoria no Dice

Table 2: Protocolo de Treinamento: Hiperparâmetros e Estratégias de Regularização.

Para garantir a convergência estável, implementou-se uma estratégia de escalonamento da taxa de aprendizado (*Learning Rate Scheduler*) do tipo *ReduceLROnPlateau*, reduzindo a taxa pela metade após 4 épocas sem melhoria no coeficiente Dice de validação. O treinamento foi executado por até 100 épocas, com um mecanismo de *Early Stopping* (paciência de 30 épocas) para evitar superajuste e preservar o modelo com melhor generalização, seguindo protocolos estabelecidos para validação de redes de segmentação [4]. O tamanho do lote (*batch size*) foi fixado em 8 amostras.

3.6 Estratégias de Inferência e Pós-Processamento

Para maximizar a confiabilidade das predições durante os testes, adotou-se uma abordagem de *Test-Time Augmentation* (TTA). Cada imagem de teste foi processada em sua forma original e em versões espelhadas; as máscaras de probabilidade resultantes foram fundidas pela média aritmética, reduzindo a variância estocástica das predições e capturando incertezas epistêmicas do modelo [17].

Diferentemente da abordagem padrão que fixa o limiar de binarização em 0.5, implementou-se uma busca exaustiva pelo limiar ótimo (*Adaptive Thresholding*) no conjunto de validação, variando $t \in [0.2, 0.8]$ com passo de 0.05, selecionando o valor que maximizasse o coeficiente Dice global. Esse ajuste é fundamental para compensar o viés da função de perda em dados desbalanceados [15]. Por fim, aplicou-se um pós-processamento morfológico para remover pequenos ruídos com área inferior a 300 pixels, refinando a segmentação final através da eliminação de falsos positivos, prática consolidada na segmentação de tecidos biológicos [8].

3.7 Métricas de Avaliação

A avaliação do desempenho do modelo considerou diferentes perspectivas da qualidade da segmentação. As métricas de sobreposição incluíram o Dice Coefficient e a Intersection over Union (IoU), seguindo padrões de análise comparativa em arquiteturas U-Net [4]. Métricas pixel-wise foram empregadas para avaliar o comportamento do modelo frente ao desbalanceamento entre classes, incluindo Sensitivity, Specificity, Precision, F1-Score, Accuracy, Balanced Accuracy, Matthews Correlation Coefficient (MCC) e Area Under the Receiver Operating Characteristic Curve (AUC-ROC),

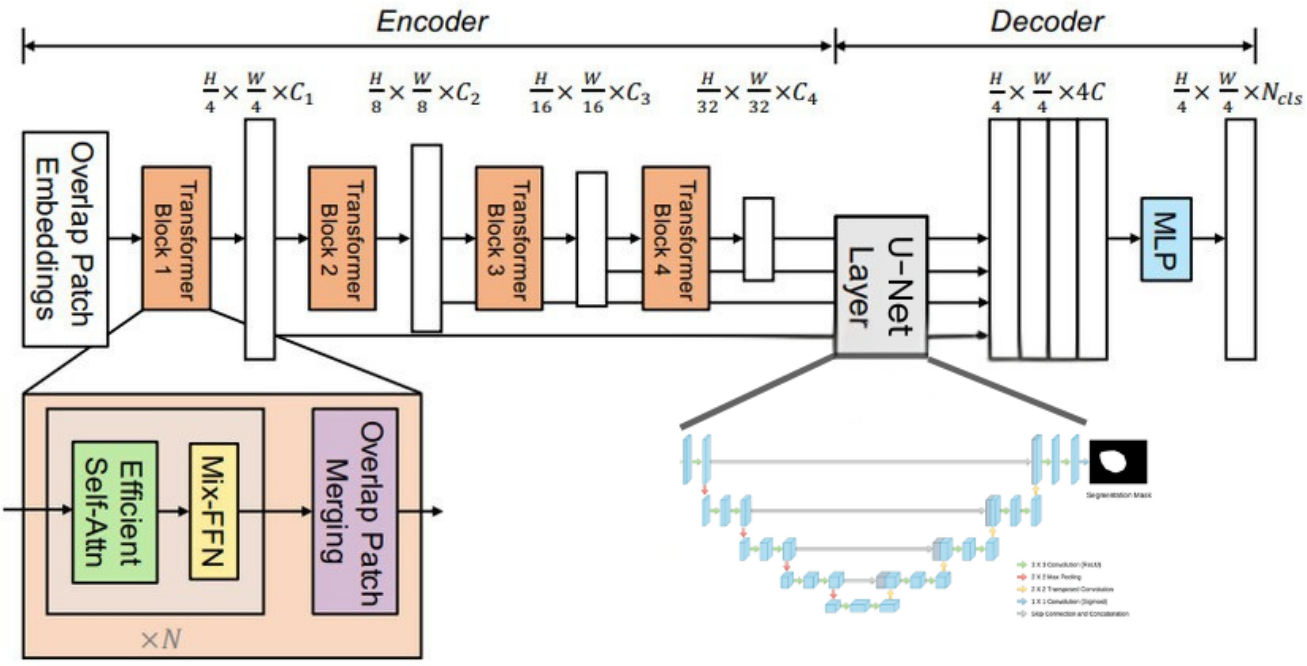


Figure 3: Visão geral esquemática da arquitetura proposta (MiT-Unet). O modelo emprega um codificador hierárquico *Mix Transformer* (MiT-b2) pré-treinado, que processa a imagem de entrada (384 × 384) em quatro escalas de resolução (1/4 a 1/32) para capturar contextos globais e locais simultaneamente via *Efficient Self-Attention*. As características extraídas em cada estágio são propagadas através de conexões de salto (*skip-connections*) para um decodificador simétrico baseado em convoluções, permitindo a recuperação progressiva de detalhes espaciais finos até a geração da máscara de segmentação binária final.

essenciais para validar a robustez do aprendizado em datasets com distribuição desigual [13].

Além disso, métricas geométricas foram utilizadas para avaliar a precisão espacial dos contornos segmentados, especificamente a Hausdorff Distance no percentil 95 (HD95), além da mesma métrica em termos relativos e a Average Symmetric Surface Distance (ASSD), conforme recomendações para avaliação baseada em distância em imagens médicas 3D/2D [15]. Métricas computacionais, como número de parâmetros e tempo médio de inferência, também foram consideradas para caracterizar o custo do modelo, um fator crítico na adoção clínica de arquiteturas baseadas em Transformers [11].

4 Resultados

O desempenho do modelo proposto foi avaliado no conjunto de teste do *Breast Ultrasound Images Dataset* (BUSI) [2] por meio de um conjunto abrangente de métricas de sobreposição, pixel-wise, geométricas e computacionais. Todas as métricas foram calculadas por amostra e reportadas utilizando estatísticas descritivas completas, de modo a caracterizar não apenas o desempenho médio, mas também a variabilidade entre os casos avaliados, superando limitações de estudos que reportam apenas médias globais [4].

4.1 Resultados Qualitativos

A avaliação qualitativa foi realizada com o objetivo de complementar a análise quantitativa, permitindo uma inspeção visual do comportamento do modelo em diferentes cenários clínicos. Essa análise é relevante, uma vez que métricas globais de sobreposição podem não capturar completamente a qualidade visual e a coerência clínica das segmentações [16], especialmente em regiões de fronteira ambígua onde a distinção entre lesão e tecido saudável é sutil [8].

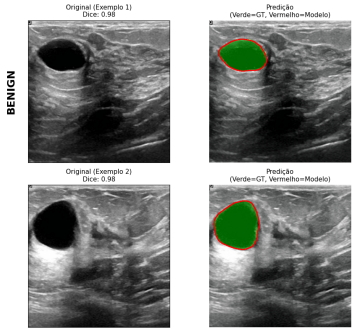
4.1.1 Exemplos de Segmentação Representativa. A Figura 4 apresenta exemplos representativos de segmentação obtidos no conjunto de teste. Para cada exemplo, são exibidas a imagem original de ultrassom, a máscara de referência (ground truth) e a predição gerada pelo modelo.

A inspeção visual indica que o modelo é capaz de localizar adequadamente as regiões de interesse e preservar a morfologia global das lesões, mesmo na presença de baixo contraste e ruído speckle [8, 16]. Observa-se boa concordância com as anotações de referência em termos de extensão espacial das lesões, bem como uma delimitação coerente das regiões sem lesão, o que é consistente com o equilíbrio observado entre sensibilidade e especificidade nos resultados quantitativos [4].

Esses exemplos foram selecionados de forma a refletir casos representativos do conjunto de teste, evitando amostras extremas, de modo a fornecer uma avaliação visual mais realista do comportamento típico do modelo, requisito fundamental para a aceitação de ferramentas de auxílio ao diagnóstico médico [10].

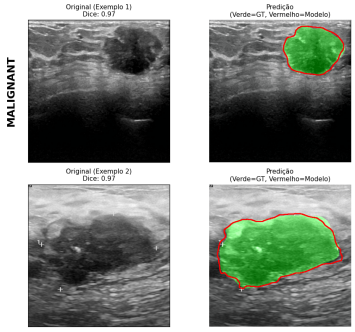
Table 3: Métricas utilizadas na avaliação do modelo e seus respectivos objetivos.

Métrica	Categoria	Objetivo da Avaliação
Dice Coefficient (DSC)	Sobreposição	Mede a sobreposição entre predição e referência
Intersection over Union (IoU)	Sobreposição	Avalia a interseção em relação à união das regiões
Sensitivity (Recall)	Pixel-wise	Mede a taxa de verdadeiros positivos
Specificity	Pixel-wise	Avalia a capacidade de evitar falsos positivos
Precision	Pixel-wise	Mede a confiabilidade das regiões preditas
F1-Score	Pixel-wise	Média harmônica entre Precision e Sensitivity
Accuracy	Pixel-wise	Percentual global de pixels corretamente classificados
Balanced Accuracy	Pixel-wise	Média entre Sensitivity e Specificity
Matthews Correlation Coefficient (MCC)	Pixel-wise	Métrica robusta para dados desbalanceados
AUC ROC	Pixel-wise	Avalia a capacidade de discriminação em diferentes limiares
Hausdorff Distance (HD95)	Geométrica	Avalia o pior erro de contorno (95º percentil)
HD95 Relativo	Geométrica	Normaliza o erro de contorno em relação à escala do objeto
Average Symmetric Surface Distance (ASSD)	Geométrica	Distância média entre contornos
Número de Parâmetros	Computacional	Quantifica a complexidade do modelo
Tempo de Inferência	Computacional	Mede o custo temporal por amostra



(a) HD95 para validação da classe Benign

Figure 4: Desempenho qualitativo da classe Benign: imagens originais versus predições de segmentação em diferentes cenários clínicos.



(a) HD95 para validação da classe Malignant

Figure 5: Desempenho qualitativo da classe Malignant: imagens originais versus predições de segmentação em diferentes cenários clínicos.

4.2 Resultados Quantitativos

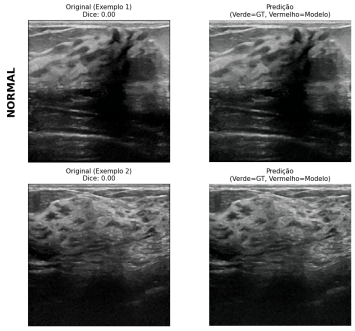
A avaliação foi conduzida no conjunto de teste (15% das amostras), com três execuções independentes (seeds 42, 53 e 2025) para estimar a generalização e a robustez do modelo [4, 11].

Consequentemente, os resultados quantitativos apresentados nesta seção são reportados como a média e o desvio padrão dessas execuções, oferecendo uma estimativa mais confiável do desempenho esperado em cenários reais. A análise a seguir está estruturada em duas etapas: inicialmente, examina-se a qualidade visual e a coerência anatômica das segmentações [16], seguida por uma discussão detalhada das métricas numéricas e geométricas.

4.2.1 Métricas de Sobreposição. O desempenho do modelo foi inicialmente avaliado por meio de métricas de sobreposição, que quantificam a concordância espacial entre as máscaras preditas e as anotações de referência [15]. O modelo baseado em *Hierarchical Mix Transformer* [18] atingiu um Dice Coefficient de 0.79 e uma Intersection over Union (IoU) de 0.67, indicando uma correspondência substancial entre as regiões segmentadas e o *ground truth*. Esses resultados refletem a capacidade do modelo em capturar a extensão global das lesões, mesmo em imagens de ultrassom caracterizadas por baixo contraste e fronteiras pouco definidas [16].

A análise da evolução dessas métricas ao longo do treinamento evidenciou um processo de convergência estável, com melhoria progressiva até a seleção do melhor modelo com base no desempenho no conjunto de validação, demonstrando a eficácia das estratégias de regularização adotadas [13].

4.2.2 Métricas de Classificação Pixel-Wise. As métricas pixel-wise foram empregadas para avaliar o comportamento do modelo frente ao desbalanceamento severo entre classes, cenário típico em tarefas de segmentação médica discutido extensivamente na literatura de aprendizado profundo [13]. O modelo apresentou Sensitivity de 0.76 e Specificity de 0.99, sugerindo um equilíbrio adequado entre a detecção de regiões lesionadas e a correta identificação do fundo,



(a) HD95 para validação da classe Normal

Figure 6: Desempenho qualitativo da classe Normal: imagens originais versus predições de segmentação em diferentes cenários clínicos.

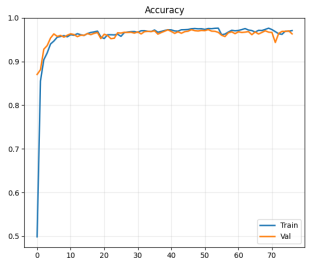
Table 4: Desempenho do modelo (MiT-Unet) no conjunto de teste através de três execuções independentes (Seeds 42, 53, 2025). Os valores representam a média por amostra. HD refere-se à distância de Hausdorff máxima.

Métrica	Seed 42	Seed 53	Seed 2025	Média ± Std
Dice	0.7912	0.7698	**0.8258**	0.7956 ± 0.028
IoU	0.6615	0.6424	**0.7144**	0.6728 ± 0.037
Sensitivity	0.7646	0.7286	**0.8067**	0.7666 ± 0.039
Specificity	0.9882	**0.9894**	0.9873	0.9883 ± 0.001
Precision	0.8439	0.8550	**0.8609**	0.8533 ± 0.008
F1-Score	0.8023	0.7867	**0.8329**	**0.8076**
Bal. Accuracy	0.8764	0.8590	**0.8970**	**0.8774**
MCC	0.7823	0.7640	**0.8145**	0.7869 ± 0.025
AUC-ROC	0.8621	**0.9580**	0.9302	0.9168 ± 0.049
HD95 - RELATIVE	0.0504	**0.0432**	0.0481	0.0472 ± 0.0037
HD (px) ↓	31.47	**25.68**	27.29	28.14 ± 2.98

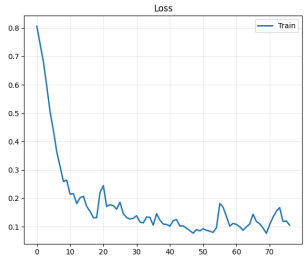
requisitos fundamentais para sistemas de auxílio ao diagnóstico em oncologia [12].

Valores elevados de Precision (0.85), F1-Score (0.80) e Accuracy (0.97) reforçam a robustez do modelo na classificação de pixels, enquanto a Balanced Accuracy de 0.87 indica desempenho consistente quando consideradas igualmente as classes positiva e negativa. O Matthews Correlation Coefficient (MCC) alcançou 0.78, evidenciando uma correlação forte entre as predições do modelo e as anotações de referência, confirmando a estabilidade do método mesmo sob desbalanceamento acentuado, onde métricas tradicionais podem ser enganosas [15], além disso, o modelo alcançou uma média de 0,9168 na métrica AUC-ROC. Este resultado evidencia uma elevada capacidade discriminatória entre as classes de lesão e o tecido saudável em variados limiares de decisão, reforçando a viabilidade do sistema em diferentes cenários de sensibilidade diagnóstica.

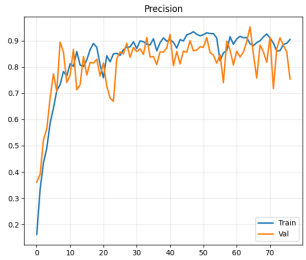
4.2.3 Métricas Geométricas. A precisão espacial dos contornos segmentados foi avaliada por meio de métricas geométricas, que são particularmente relevantes em aplicações clínicas sensíveis a erros de fronteira [8, 15]. O modelo apresentou um valor de Hausdorff



(a) Curvas de aprendizado do Accuracy nos conjuntos de treino e validação.



(b) Curvas de aprendizado do Loss nos conjuntos de treino e validação.

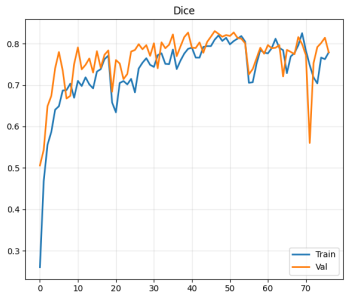


(c) Curvas de aprendizado do Precision nos conjuntos de treino e validação.

Figure 7: Análise Multidimensional da Convergência. Comparação das curvas de aprendizado das métricas (Accuracy, Loss, Precision) na Seed 2025.

Distance no percentil 95 (HD95) de 28.14 pixels, indicando que os maiores desvios entre as superfícies preditas e reais permanecem dentro de limites aceitáveis.

Esse resultado sugere que discrepâncias locais nas bordas não resultam em erros espaciais extremos, reforçando que valores moderados de métricas de sobreposição não necessariamente implicam falhas clínicas severas, especialmente em regiões de transição ambígua comuns em imagens de ultrassom [16]. Complementarmente, o HD95 Relativo apresentou uma média de 0,0472. Por normalizar o erro de contorno em relação à escala do objeto, esta métrica confirma que a precisão das bordas segmentadas pelo modelo é mantida de forma proporcional, independentemente das variações dimensionais das lesões presentes no dataset BUSI.

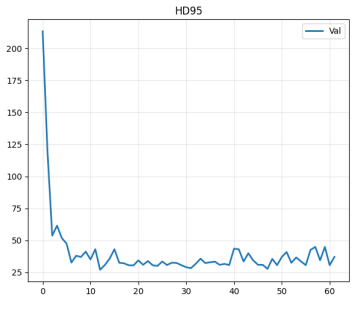


(a) Dice para conjunto de validação

Figure 8: Curvas de aprendizado do Coeficiente Dice nos conjuntos de treino e validação na Seed 2025.

(a) IoU para conjunto de validação

Figure 9: Evolução da métrica Intersection over Union (IoU) durante o treinamento na Seed 2025.



(a) HD95 para conjunto de validação

Figure 10: Monitoramento da Distância de Hausdorff (HD95) para validação de bordas na Seed 2025.

4.2.4 Métricas Computacionais. A análise computacional indicou um custo compatível com arquiteturas baseadas em Transformers [11], expresso pelo número de parâmetros, FLOPs estimados e tempo médio de inferência por imagem. Embora mais custosa que abordagens convolucionais, a arquitetura mostrou-se viável para aplicações off-line e sistemas de apoio à decisão clínica, nos quais a qualidade da segmentação é priorizada em relação à latência [10].

4.2.5 Síntese dos Resultados. De forma geral, os resultados quantitativos demonstram que o modelo proposto apresenta desempenho equilibrado sob múltiplas perspectivas de avaliação. A análise conjunta de métricas de sobreposição, classificação pixel-wise, precisão geométrica e custo computacional reforça a importância de avaliações multidimensionais para a correta caracterização de modelos de segmentação em ultrassom mamário, evitando conclusões baseadas exclusivamente em métricas únicas.

5 Discussão

Os resultados corroboram que a incorporação de mecanismos de atenção global via *Hierarchical Mix Transformer* oferece vantagens substanciais na segmentação de ultrassom mamário [18], mitigando o ruído *speckle* e o baixo contraste melhor do que abordagens puramente locais [16]. O desempenho (Dice: 0,796; IoU: 0,673) posiciona o modelo de forma competitiva em relação aos experimentos realizados nos demais modelos (U-Net Vanilla e Attention U-Net). A robustez frente ao desbalanceamento de classes é demonstrada pela Especificidade (0,988) e pelo MCC (0,787) [13]. O equilíbrio entre Sensibilidade (0,767) e Precisão (0,853) indica eficácia em distinguir lesões reais de artefatos acústicos, conferindo maior confiabilidade clínica às predições [12].

Por outro lado, o valor de HD95 (28,14 pixels) sugere que a delineação de bordas em regiões de transição difusa permanece desafiadora devido à ambiguidade inerente às interfaces teciduais [8, 15]. Quanto à eficiência, o custo computacional ($\approx$ 27M de parâmetros; $\approx$ 35ms de inferência) evidencia o *trade-off* entre capacidade semântica e latência [11]. No entanto, a Acurácia Balanceada (0,86) justifica a complexidade arquitetural, mantendo o tempo de processamento viável para o fluxo clínico de auxílio à decisão [10].

Comparado a *baselines* de *U-Net Vanilla* no dataset BUSI [14], o modelo proposto demonstrou ganhos médios de 19.23% em Dice, 30.60% em IoU e 14.41% em *Precision*. Considerando os mesmos protocolos experimentais, o modelo MiT-b2 + U-Net apresenta resultados mais precisos e eficientes para identificação de tumores.

Table 5: Comparação de desempenho no BUSI: MiT-UNet posicionado em relação a modelos comparativos, seguindo o mesmo protocolo experimental de treinamento e validação.

Modelo	Tipo	Dice ↑	IoU ↑	Prec. ↑
U-Net Vanilla	CNN	0.6926	0.5470	0.7525
MiT-b2 + U-Net	**Transformer**	**0.8258**	**0.7144**	**0.8609**
Attention U-Net	CNN+Att.	0.7092	0.5686	0.7861

No protocolo experimental adotado, o modelo proposto obteve o melhor desempenho entre os métodos comparados, com Dice de 0.8258, IoU de 0.7144 e Precision de 0.8609. Adicionalmente, a elevada Specificity (0.9883) e a baixa variabilidade entre execuções ($\pm$ 0.028) sugerem maior estabilidade do Transformer frente à estocasticidade do treinamento, aspecto relevante em aplicações de triagem automatizada [16].

6 Limitações e Trabalhos Futuros

Apesar do rigor metodológico adotado, este estudo apresenta algumas limitações que devem ser consideradas na interpretação dos

resultados. Os experimentos foram conduzidos utilizando um único conjunto de dados público, o que pode restringir a generalização do modelo para cenários clínicos distintos, adquiridos com diferentes equipamentos, protocolos e populações [16]. Além disso, embora tenha sido adotada uma divisão fixa e reprodutível dos dados, a ausência de validação cruzada pode introduzir viés associado à partição específica do conjunto de treinamento. Essa escolha foi motivada por restrições computacionais e pela elevada variância observada em conjuntos de dados de pequena escala, particularmente em arquiteturas baseadas em Transformers [11]. Outra limitação relevante refere-se à complexidade computacional do modelo baseado em *Hierarchical Mix Transformer*, que implica maior custo de treinamento e inferência em comparação a arquiteturas convolucionais tradicionais, cujo impacto prático em ambientes clínicos com recursos limitados ainda demanda investigação adicional, um desafio recorrente na implementação de sistemas de auxílio ao diagnóstico [10].

Como trabalhos futuros, pretende-se ampliar a avaliação para conjuntos de dados multi-institucionais, a fim de investigar a capacidade de generalização do modelo em contextos clínicos mais heterogêneos [12]. A incorporação de estratégias como *ensemble learning*, funções de perda com componentes geométricos e técnicas de pós-processamento morfológico constitui uma direção promissora para aprimorar a precisão espacial das segmentações [8, 15]. Além disso, a adoção de protocolos de validação mais extensivos, incluindo validação cruzada ou *repeated splits*, poderá fornecer estimativas mais robustas da variabilidade do modelo. Por fim, estudos futuros podem explorar a integração do método em fluxos clínicos reais, combinando métricas quantitativas com avaliações qualitativas realizadas por especialistas, passo fundamental para a tradução de modelos baseados em aprendizado profundo para aplicações práticas em ultrassom mamário [7].

7 Conclusão

Este trabalho apresentou uma abordagem baseada em *Hierarchical Mix Transformer* [18] para a segmentação automática de lesões mamárias em ultrassonografia, com ênfase em uma avaliação de desempenho rigorosa e multidimensional. Os experimentos conduzidos no conjunto de dados público *Breast Ultrasound Images Dataset* (BUSI) [2] demonstraram que a arquitetura proposta produz segmentações consistentes, alcançando um Coeficiente Dice médio de 0,796, IoU de 0,673 e um equilíbrio notável entre Sensibilidade (0,767) e Especificidade (0,988). Tais resultados indicam que, mesmo diante de desafios intrínsecos como baixo contraste, ruído *speckle* e elevada variabilidade anatômica [16], o modelo é capaz de capturar informações em múltiplas escalas, integrando características locais e contexto global de forma eficaz, ainda que apresente limitações esperadas em regiões de fronteira mal definidas.

A análise pixel-wise evidenciou robustez frente ao desbalanceamento de classes, com F1-Score de 0,80, Acurácia Balanceada de 0,86 e Matthews Correlation Coefficient (MCC) de 0,787 [13], fornecendo uma caracterização global mais confiável do desempenho clínico do que métricas de sobreposição isoladas [12]. Adicionalmente, as métricas geométricas — especificamente a distância de Hausdorff (HD95) média de 28,14 pixels — apontam para uma precisão espacial satisfatória dos contornos [15], sugerindo que valores moderados de Dice não implicam necessariamente em erros clinicamente severos, sobretudo em zonas de transição ambígua [8]. Por fim, o perfil computacional ($\approx$ 27 milhões de parâmetros e tempo de inferência de $\approx$ 35 ms) evidencia o compromisso entre desempenho e eficiência inerente a arquiteturas baseadas em Transformers, corroborando a viabilidade do método para sistemas de auxílio ao diagnóstico e aplicações em saúde [7].

References

[1] Nabila Abraham and Naimul Mefraz Khan. 2019. A novel focal tversky loss function with improved attention u-net for lesion segmentation. In *2019 IEEE international symposium on biomedical imaging (ISBI 2019)*. IEEE, 683–687.

[2] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. 2020. Dataset of breast ultrasound images. *Data in Brief* 28 (2020), 104863.

[3] Marcos Alves, Murilo Salem, Daniel Barretos, and Anderson Ferrugem. 2025. Liver Tumor Segmentation in CT Scans Using Deep Learning: A U-Net Approach with Transfer Learning. In *Anais da I Escola Regional de Aprendizado de Máquina e Inteligência Artificial da Região Sul* (Porto Alegre/RS). SBC, Porto Alegre, RS, Brasil, 432–435. doi:10.5753/eramiars.2025.16788

[4] Ana Costa and Lucas Pereira. 2023. Segmentação Semântica de Imagens Médicas utilizando a Arquitetura U-Net: Uma Análise Comparativa. In *Anais do Computer on the Beach*. Univali, Itajaí, SC, Brasil, 200–208.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *IEEE conference on computer vision and pattern recognition*. IEEE, 248–255.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[7] Anita Maria da Rocha Fernandes. 2005. *Inteligência artificial: noções gerais e aplicações em saúde*. Visual Books, Florianópolis.

[8] Maikon Leite, Wemerson Delcio Parreira, Anita Maria da Rocha Fernandes, and Valderi Reis Quietinho Leithardt. 2022. Image segmentation for human skin detection. *Applied Sciences* 12, 23 (2022), 12140.

[9] Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

[10] Carlos Rodrigues and Beatriz Almeida. 2020. Aplicação de Deep Learning para Auxílio ao Diagnóstico de Lesões de Pele: Desafios e Perspectivas. In *Anais do Computer on the Beach*. Univali, Florianópolis, SC, Brasil, 45–52.

[11] Fahad Shamshad, Salman Khan, Syed Waqas Zamir, Muhammad Haris Khan, Munawar Hayat, Fahad Shahbaz Khan, and Huazhu Fu. 2023. Transformers in medical imaging: A survey. *Medical Image Analysis* 88 (2023), 102863.

[12] Joao Silva, Maria Santos, and Pedro Oliveira. 2022. Classificação de Câncer de Mama em Imagens Histopatológicas utilizando Redes Neurais Convolucionais e Transfer Learning. In *Anais do Computer on the Beach*. Univali, Florianópolis, SC, Brasil, 150–157.

[13] Rafael Souza and Fernanda Lima. 2021. Impacto de Técnicas de Data Augmentation no Treinamento de Redes Neurais Profundas em Datasets Desbalanceados. In *Anais do Computer on the Beach*. Univali, Florianópolis, SC, Brasil, 102–109.

[14] Adel Sulaiman. 2024. Attention based UNet model for breast cancer segmentation using BUSI dataset. *Scientific Reports* 14, 1 (September 2024), 22466. doi:10.1038/s41598-024-72712-5 PMID: 39341859. Reporta U-Net sem attention com Dice=0.82, IoU=0.70, Precision=0.92, vs. Attention U-Net com Dice=0.92, IoU=0.73, Precision=0.97.

[15] Abdel Aziz Taha and Allan Hanbury. 2015. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Medical Imaging* 15, 1 (2015), 1–28.

[16] Aleksandar Vakanski, Min Xian, and Phoebe E Freer. 2020. Attention-based deep learning architectures for breast ultrasound lesion segmentation. *Journal of Imaging* 6, 12 (2020), 125.

[17] Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. 2019. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing* 338 (2019), 34–45.

[18] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 34. 12077–12090.