

EXTERNALLY VALIDATED MULTI-TASK LEARNING VIA CONSISTENCY REGULARIZATION USING DIFFERENTIABLE BI-RADS FEATURES FOR BREAST ULTRASOUND TUMOR SEGMENTATION

Jingru Zhang[†] Saed Moradi[‡] Ashirbani Saha^{‡*}

Correspondence: {zhanj390, morads6, sahaa16}@mcmaster.ca

[†] School of Biomedical Engineering, McMaster University, Hamilton, ON, Canada

[‡]Department of Oncology, McMaster University, Hamilton, ON, Canada

*Centre for Data Science and Digital Health, Hamilton Health Sciences, Hamilton, ON, Canada

ABSTRACT

Multi-task learning can suffer from destructive task interference, where jointly trained models underperform single-task baselines and limit generalization. To improve generalization performance in breast ultrasound-based tumor segmentation via multi-task learning, we propose a novel consistency regularization approach that mitigates destructive interference between segmentation and classification. The consistency regularization approach is composed of differentiable BI-RADS-inspired morphological features. We validated this approach by training all models on the BrEaST dataset (Poland) and evaluating them on three external datasets: UDIAT (Spain), BUSI (Egypt), and BUS-UCLM (Spain). Our comprehensive analysis demonstrates statistically significant ($p < 0.001$) improvements in generalization for segmentation task of the proposed multi-task approach vs. the baseline one: UDIAT, BUSI, BUS-UCLM (Dice coefficient=0.81 vs 0.59, 0.66 vs 0.56, 0.69 vs 0.49, resp.). The proposed approach also achieves state-of-the-art segmentation performance under rigorous external validation on the UDIAT dataset.

Index Terms— Breast ultrasound, consistency regularization, tumor segmentation, multi-task learning, BI-RADS, external validation

1. INTRODUCTION

Breast cancer remains a leading cause of mortality among women worldwide, with early detection being critical for improving patient outcomes [1]. Breast ultrasound complements mammography, particularly for dense breast tissue [2], and computer-aided diagnosis systems for tumor segmentation and malignancy classification can assist radiologists [3].

In breast ultrasound analysis, recent multi-task learning approaches [4, 5, 6, 7] have demonstrated promising results for segmentation and malignancy classification. All these approaches relied on internal validation or single-center evaluation, leaving their external generalization capabilities

which is a critical requirement for clinical validation unexplored. However, models trained on single-center data frequently exhibit poor generalization due to domain shift from different ultrasound systems, imaging protocols, and patient populations [8]. Beyond this, traditional multi-task models can suffer from destructive task interference, where gradients from different task heads conflict during training, leading to degraded segmentation performance compared to single-task baselines [9].

To improve the generalizability and mitigate the destructive interference of the multi-task models for the task of breast tumor segmentation, we propose a unified consistency regularization approach. Our main contributions are: (1) A novel consistency regularization loss function that enforces agreement between morphology-derived and predicted malignancy scores to resolve destructive task interference, (2) A formation of a differentiable composite prior based on BI-RADS morphological descriptors computed from soft segmentation masks to support the consistency regularization, (3) State-of-the-art segmentation performance {Dice coefficient (DC)=0.81 on UDIAT under rigorous external validation protocol} with statistically significant improvements over multi-task baselines across three external datasets.

The rest of this work is organized as follows: section 2 presents our method, section 3 presents results, section 4 discusses findings, and section 5 concludes the paper.

2. PROPOSED METHOD

Our method employs an encoder-decoder architecture with a shared backbone for simultaneous segmentation and malignancy classification. The key innovation is consistency regularization through differentiable BI-RADS-inspired morphological features. Figure 1 shows the architecture overview.

Network Architecture. We use an EfficientNet-B1 [10] backbone pretrained on ImageNet as the encoder, with a U-Net [11] decoder incorporating Atrous Spatial Pyramid Pooling (ASPP) [12] modules and attention gates [13]. ASPP

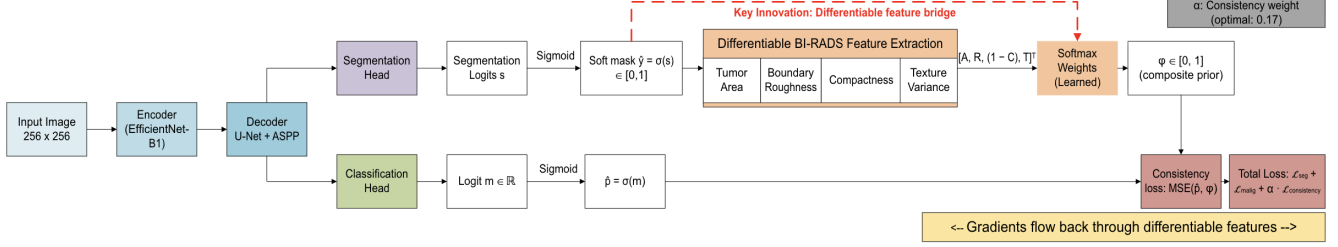


Fig. 1. Encoder–decoder (EfficientNet-B1 + U-Net-ASPP) with differentiable BI-RADS features. Segmentation logits s (no range) yield soft mask $\hat{y} = \sigma(s) \in [0, 1]$. Learned softmax feature weights combine features into composite prior $\phi \in [0, 1]$. Consistency loss enforces agreement: $\text{MSE}(\hat{p}, \phi)$ where $\hat{p} = \sigma(m)$.

enables multi-scale feature extraction to handle objects of varying sizes, while attention gates suppress irrelevant background regions and focus the network on object boundaries. The network accepts dual-stream input where 3-channel images feed the encoder while raw grayscale images are used for texture feature computation (T). It produces two outputs: segmentation logits s yielding soft masks $\hat{y} = \sigma(s)$, and malignancy logit m yielding probability $\hat{p} = \sigma(m)$.

Differentiable BI-RADS Features. We compute four morphological features from soft segmentation masks $\hat{y} = \sigma(s)$ as shown in Table 1. These features are inspired by BI-RADS descriptors [14] but are fully differentiable, enabling end-to-end training.

Composite Malignancy Prior. We compute a morphological malignancy score ϕ by combining the four BI-RADS-inspired features (Area A , Boundary Roughness R , Compactness C , Texture T) through learned softmax weights:

$$\phi = \mathbf{w}^T \mathbf{f}, \quad \phi \in [0, 1], \quad (1)$$

where $\mathbf{f} = [A, R, (1-C), T]^T$ and $\mathbf{w} = [w_A, w_R, w_C, w_T]^T$. The weight vector $\mathbf{w}$ is obtained via softmax normalization of learnable parameters $\mathbf{u}$, i.e., $\mathbf{w} = \text{softmax}(\mathbf{u})$. The weights are initialized to encode BI-RADS priors, then learned via gradient descent with L2 regularization on $\mathbf{u}$ to stabilize learning.

Training Objective. The total loss combines task-specific and consistency terms:

$$\mathcal{L} = w_{\text{seg}} \mathcal{L}_{\text{seg}} + w_{\text{cls}} \mathcal{L}_{\text{cls}} + \alpha \cdot \text{MSE}(\hat{p}, \phi) + \alpha \lambda_{\text{nt}} \text{NTP} + \beta \|\mathbf{u}\|_2^2 \quad (2)$$

where w_{seg} and w_{cls} are task weights for segmentation and classification respectively, α controls consistency strength, λ_{nt} weights the no-tumor penalty (normal regions in the images that are incorrectly identified as mask, i.e., false positives), and β controls L2 regularization. Both $\mathcal{L}_{\text{seg}}$ and $\mathcal{L}_{\text{cls}}$ use Binary Cross-Entropy (BCE) with logits. The consistency term $\text{MSE}(\hat{p}, \phi)$ penalizes disagreement between predicted malignancy $\hat{p} = \sigma(m)$ and morphology score ϕ . The no-tumor penalty NTP discourages false positives (computed

as $\text{mean}(\hat{y})$ for images with no tumors, 0 otherwise). L2 regularization $\beta \|\mathbf{u}\|_2^2$ is applied to weight parameters.

This consistency loss is the key mechanism for resolving task interference: it creates an explicit bridge between segmentation and classification tasks, encouraging the model to produce segmentation masks whose morphological characteristics align with the predicted malignancy assessment. Unlike standard multi-task learning that relies solely on shared representations, this constraint enables beneficial task synergy.

Our multi-task learning strategy aims to leverage supervisory information from two independent sources. While segmentation masks are derived directly from visual attributes of the images, malignancy labels can incorporate information from biopsy tests, making classification a more complex diagnostic task despite being less spatially fine-grained than segmentation. Therefore, we assign higher weight to the segmentation loss ($w_{\text{seg}} > w_{\text{cls}}$) in our training objective. The model selection process is based on the best performance in the more complex task (classification) while maintaining strong segmentation accuracy. Therefore, the internal validation for tuning α value is based on the Area-under-the-Receiver Operating Characteristic (AUC) for classification performance on the validation set.

3. EXPERIMENTS AND RESULTS

3.1. Training Dataset

We trained our models on the BrEaST dataset [16], comprising 252 breast ultrasound images (154 benign, 98 malignant) collected in Poland. All images were resized to 256×256 pixels and split into training (85%) and internal validation (15%) sets. We used data augmentation including horizontal flips, affine transformations (shift, scale, rotation), Gaussian noise, and brightness/contrast adjustments.

3.2. Training Protocol

We used AdamW [17] optimizer with learning rate 9.2×10^{-4} , weight decay 10^{-4} , and batch size 32, employing

Table 1. Differentiable BI-RADS-inspired morphological features computed from soft segmentation masks $\hat{y} = \sigma(s)$.

Feature	Clinical Meaning	Formula & Computation
Area (A)	Tumor size; larger masses may have higher malignancy risk	$A = \text{mean}(\hat{y})$ where mean is computed over all pixels. Provides global scale information about the tumor.
Boundary Roughness (R)	Irregular, spiculated margins suggest malignancy; smooth borders suggest benign	Edge magnitude via Sobel filters; perimeter $P = \sum_{u,v} E_{uv}$; roughness $R = P / \sqrt{\sum_{u,v} \hat{y}_{uv}}$ normalized to $[0, 1]$ via EMA min-max (momentum 0.99).
Compactness (C)	Round/oval shapes (high C) typically benign; irregular shapes (low C) suggest malignancy	$C = \text{clip}\left(\frac{4\pi \sum_{u,v} \hat{y}_{uv}}{P^2 + \varepsilon}, 0, 1\right)$ where P is perimeter from above and $\varepsilon = 10^{-6}$ ensures numerical stability. The clip function bounds values to $[0, 1]$. Measures circularity of the tumor region.
Texture (T)	Heterogeneous internal echogenicity suggests malignancy; homogeneous suggests benign	$T = \text{var}_{\hat{y}}(g)$ where $\text{var}_{\hat{y}}$ is the mask-weighted variance of raw grayscale values g within the tumor region. Normalized to $[0, 1]$ via EMA.

All features are differentiable w.r.t. segmentation logits s , enabling end-to-end training. R and T are normalized to $[0, 1]$ using exponential moving average (EMA) min-max statistics (momentum 0.99), where the EMA update mechanism follows batch normalization [15], computed over training batches.

cosine annealing schedule with early stopping based on internal validation loss. Task weights were set to $(w_{\text{seg}}, w_{\text{cls}}) = (0.9, 0.1)$ to prioritize segmentation. Consistency strength was selected via a nested internal validation sweep involving both classification and segmentation (from 0.1 to 0.9 with different increments). The no-tumor penalty weight (λ_{nt}) was 0.5. Feature weights were initialized to $(0.15, 0.35, 0.25, 0.25)$ for (w_A, w_R, w_C, w_T) to encode BI-RADS prior importance, then learned during training. L2 regularization was applied with $\beta = 0.001$.

3.3. Experimental Setup

To rigorously evaluate external generalization, we tested our method on three datasets from different clinical centers. The BUSI dataset [18] contains 647 images with 437 benign and 210 malignant lesions (normal images excluded). The UDIAT dataset [19] includes 163 images with 110 benign and 53 malignant lesions. The BUS-UCLM dataset [20] provides 117 images with 81 benign and 36 malignant lesions (normal images, images with additional features excluded). These external datasets were used for testing without any fine-tuning, ensuring a true assessment of cross-center generalization.

Our model selection followed a pre-defined protocol choosing two models: (a) *Proposed-best-cls* which uses $\alpha = 0.17$ for highest internal validation AUC of 0.97 and (b) *Proposed-best-seg* which uses $\alpha = 0.18$ for highest internal DC of 0.86. We present results for both consistency strength values: *Proposed-best-cls* represents the model with expected optimal balanced performance, while *Proposed-best-seg* achieves the highest internal segmentation performance.

3.4. Results for External Validation

Table 2 presents results across internal validation and three external datasets. Baseline multi-task learning (seg_malig) experiences destructive task interference where it underperforms single-task (seg_only) on segmentation (DC: 0.72 vs.

0.80 internal, 0.56–0.59 vs. 0.64–0.78 external). Our consistency regularization mitigates this interference, enabling multi-task models to match/exceed single-task performance.

Our *Proposed-best-cls* achieves statistically significant improvements over the multi-task baseline: BUSI DC=0.66 vs. 0.56 (+18%, $p<0.001$), UDIAT DC=0.81 vs. 0.59 (+37%, $p<0.001$), BUS-UCLM DC=0.69 vs. 0.49 (+41%, $p<0.001$). Statistical significance was assessed using paired Wilcoxon signed-rank tests [21] on per-image DC. For classification, our method maintains strong performance (AUC=0.97 internal, 0.74–0.79 external).

Table 2. Internal and external validation results. Best results in **bold**.

Dataset	Model	DC	AUC
<i>BrEaST Internal Validation</i> (N=38; 21 benign, 17 malignant)	malig_only (single-task)	—	0.83
	seg_only (single-task)	0.80	—
	seg_malig (multi-task baseline)	0.72	0.78
	<i>Proposed-best-seg</i>	0.86	0.86
	<i>Proposed-best-cls</i>	0.83	0.97
<i>BUSI External Test</i> (N=647; 437 benign, 210 malignant)	malig_only	—	0.69
	seg_only	0.65	—
	seg_malig	0.56	0.77
	<i>Proposed-best-seg</i>	0.66	0.76
	<i>Proposed-best-cls</i>	0.66	0.79
<i>UDIAT External Test</i> (N=163; 110 benign, 53 malignant)	malig_only	—	0.62
	seg_only	0.78	—
	seg_malig	0.59	0.66
	<i>Proposed-best-seg</i>	0.77	0.59
	<i>Proposed-best-cls</i>	0.81	0.74
<i>BUS-UCLM External Test</i> (N=117; 81 benign, 36 malignant)	malig_only	—	0.56
	seg_only	0.64	—
	seg_malig	0.49	0.69
	<i>Proposed-best-seg</i>	0.68	0.74
	<i>Proposed-best-cls</i>	0.69	0.79

DC: Dice Coefficient, AUC: Area-Under-the-ROC-Curve.

3.5. Comparison with State-of-the-Art Methods

Table 3 compares our method against state-of-the-art approaches on UDIAT under rigorous external validation (train BrEaST, test UDIAT). Our multi-task model achieves DC=0.81, matching the best performing methods while additionally providing malignancy classification. For BUSI, direct comparison to the recent work (involving multi-task learning

Table 3. Segmentation performance of *Proposed-best-cls* in UDIAT against state-of-the-art methods

Method	Training	DC
UNet [22]	Fully-labeled BUSI dataset	0.75
UNet++ [22]	Fully-labeled BUSI dataset	0.75
DeepLabv3+ [22]	Fully-labeled BUSI dataset	0.78
TransUNet [22]	Fully-labeled BUSI dataset	0.72
SwinUNet [22]	Fully-labeled BUSI dataset	0.73
SAMLR (Rank=6) [22]	20%-labeled BUSI dataset	0.81
Ours (<i>Proposed-best-cls</i>)	Fully-labeled BrEaST dataset	0.81

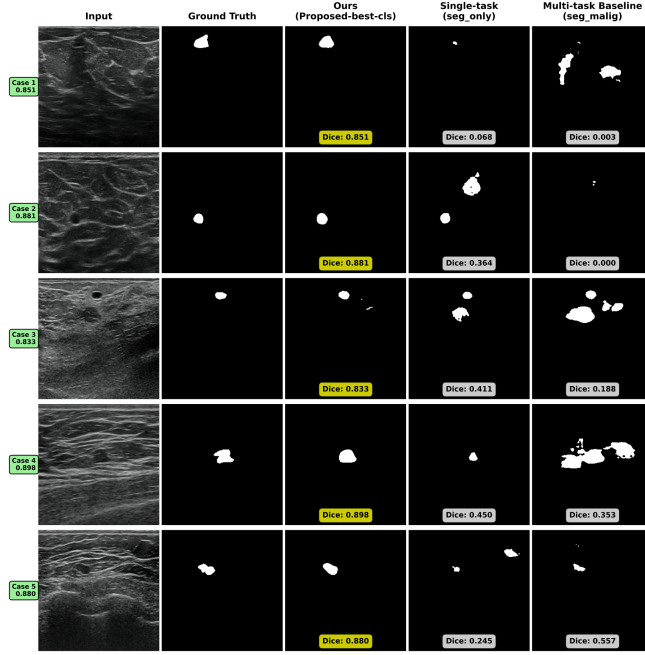


Fig. 2. Examples of segmentation comparing baselines and our *Proposed-best-cls*. Our method produces cleaner boundaries and better handles challenging cases.

for breast ultrasound) of Aumente-Maestro et al. [4] is not protocol-equivalent (their 4-fold cross-validation on curated subset vs. our external test on 647 images).

4. DISCUSSION

Our proposed method achieves substantial external improvements (+37% UDIAT, +18% BUSI, +41% BUS-UCLM DC over baseline multi-tasking of segmentation and classification, all $p < 0.001$), demonstrating state-of-the-art segmentation performance (DC=0.81 on UDIAT) that matches specialized single-task methods including DeepLabv3+ (0.78) and transformer-based architectures (TransUNet 0.72, SwinUNet 0.73). These improvements suggest morphological priors act as domain-robust regularization: while imaging protocols vary across datasets, the relationship between tumor morphology and malignancy remains consistent with BI-RADS

criteria, explaining superior generalization.

While our primary focus was segmentation improvement, the proposed method maintains strong classification performance (AUC=0.74-0.79 external), demonstrating that consistency regularization enables beneficial task synergy for both objectives rather than forcing a performance trade-off. Our consistency mechanism leverages classification as an auxiliary task to improve segmentation generalization while maintaining competitive classification performance. Our aim prioritized segmentation while keeping classification competitive; classifier-optimized variants are future work.

Our model selection strategy, which emphasizes internal validation classification performance (AUC) rather than segmentation performance alone, is validated by external test results. As shown in Table 2, *Proposed-best-cls* consistently matches/outperforms *Proposed-best-seg* on external datasets in both segmentation and classification. This pattern suggests that selecting models based solely on internal segmentation performance may not yield optimal external generalization.

5. CONCLUSION

We presented a consistency regularization approach for multi-task breast ultrasound analysis using differentiable BI-RADS-inspired morphological features. By mitigating destructive interference between segmentation and classification tasks, our method achieves significant external improvements across three datasets from different centers and ultrasound systems. This work demonstrates that consistency regularization with domain-informed priors is a viable strategy for effective multi-task learning in medical imaging, enabling beneficial task synergy while maintaining strong external generalization.

6. COMPLIANCE AND ACKNOWLEDGEMENT

Compliance with Ethics: This study received approval (Project ID#18054) from the Hamilton Integrated Research Ethics Board. *Conflicts of Interest:* The authors have no conflicts of interest to declare. This research has been supported by Hamilton Health Sciences Foundation Grant (RD-239) and a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada. *Acknowledgement:* The authors acknowledge the computational resources provided by The Centre for Data Science and Digital Health (CREATE) and Population Health Research Institute (PHRI), Hamilton Health Sciences.

7. REFERENCES

- [1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a*

- cancer journal for clinicians*, vol. 71, no. 3, pp. 209–249, 2021.
- [2] R. J. Hooley, K. L. Greenberg, R. M. Stackhouse, J. L. Geisel, R. S. Butler, and L. E. Philpotts, “Screening US in patients with mammographically dense breasts: initial experience with Connecticut Public Act 09-41,” *Radiology*, vol. 265, no. 1, pp. 59–69, 2012.
 - [3] H.-D. Cheng, J. Shan, W. Ju, Y. Guo, and L. Zhang, “Automated breast cancer detection and classification using ultrasound images: A survey,” *Pattern recognition*, vol. 43, no. 1, pp. 299–317, 2010.
 - [4] C. Aumente-Maestro, J. Díez, and B. Remeseiro, “A multi-task framework for breast cancer segmentation and classification in ultrasound imaging,” *Computer methods and programs in biomedicine*, vol. 260, p. 108540, 2025.
 - [5] Q. He, Q. Yang, H. Su, and Y. Wang, “Multi-task learning for segmentation and classification of breast tumors from ultrasound images,” *Computers in Biology and Medicine*, vol. 173, p. 108319, 2024.
 - [6] G. Zhang, K. Zhao, Y. Hong, X. Qiu, K. Zhang, and B. Wei, “SHA-MTL: soft and hard attention multi-task learning for automated breast cancer ultrasound image segmentation and classification,” *International journal of computer assisted radiology and surgery*, vol. 16, no. 10, pp. 1719–1725, 2021.
 - [7] J. Chowdary, P. Yogarajah, P. Chaurasia, and V. Gurusviah, “A multi-task learning framework for automated segmentation and classification of breast tumors from ultrasound images,” *Ultrasonic imaging*, vol. 44, no. 1, pp. 3–12, 2022.
 - [8] H. Guan and M. Liu, “Domain adaptation for medical image analysis: a survey,” *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 3, pp. 1173–1185, 2022.
 - [9] X. Zhao, H. Li, X. Shen, X. Liang, and Y. Wu, “A modulation module for multi-task learning with applications in image retrieval,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 401–416, 2018.
 - [10] M. Tan and Q. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in *International Conference on Machine Learning*, pp. 6105–6114, PMLR, 2019.
 - [11] O. Ronneberger, P. Fischer, and T. Brox, “U-NET: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.
 - [12] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.
 - [13] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, “Attention gated networks: Learning to leverage salient regions in medical images,” *Medical image analysis*, vol. 53, pp. 197–207, 2019.
 - [14] D. A. Spak, J. S. Plaxco, L. Santiago, M. J. Dryden, and B. E. Dogan, “BI-RADS® fifth edition: A summary of changes,” *Diagnostic and Interventional Imaging*, vol. 98, no. 3, pp. 179–190, 2017.
 - [15] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *International conference on machine learning*, pp. 448–456, pmlr, 2015.
 - [16] A. Pawłowska, A. Ćwierz-Pieńkowska, A. Domalik, D. Jaguś, P. Kasprzak, R. Matkowski, L. Fura, A. Nowicki, and N. Żołek, “Curated benchmark dataset for ultrasound based breast lesion analysis,” *Scientific Data*, vol. 11, no. 1, p. 148, 2024.
 - [17] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” *arXiv preprint arXiv:1711.05101*, 2017.
 - [18] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, “Dataset of breast ultrasound images,” *Data in brief*, vol. 28, p. 104863, 2020.
 - [19] M. H. Yap, G. Pons, J. Martí, S. Ganau, M. Sentís, R. Zwiggelaar, A. K. Davison, and R. Martí, “Automated breast ultrasound lesions detection using convolutional neural networks,” *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 4, pp. 1218–1226, 2018.
 - [20] N. Vallez, G. Bueno, O. Deniz, M. A. Rienda, and C. Pastor, “BUS-UCLM: Breast ultrasound lesion segmentation dataset,” *Scientific Data*, vol. 12, no. 1, p. 242, 2025.
 - [21] F. Wilcoxon, “Individual comparisons by ranking methods,” *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.
 - [22] T. Jiang, Y. Li, Y. Li, W. Xing, M. Yu, F. Xie, and D. Ta, “A segmentation knowledge-based global-local attention network for tumor classification in breast ultrasound images,” *Pattern Recognition*, vol. 171, p. 112152, 2026.