



Comparative Performance of Deep Learning Architectures for Breast Tumor Segmentation on Multicenter DCE-MRI

Loyani Loyani, MS, PhD Student, Emory University; Biniam A. Garomsa, PhD Student; Frank Li, PhD; Hari M. Trivedi, MD; Judy W. Gichoya, MD, MS, FSIIM

Introduction/Background

Automated breast tumor segmentation on dynamic contrast-enhanced MRI supports treatment planning and response assessment, but aggregate metrics do not reveal whether failures recur across models. We compared five deep-learning architectures to evaluate performance, patient-level failure overlap, and characteristics associated with failure.

Methods/Intervention

We evaluated five segmentation architectures: CNN/transformer-based (U-Net, SegResNet, SwinUNETR), nnU-Net with large residual-encoder configuration (ResEnc L), and a foundation model (MedSAM) using MAMA-MIA, a multicenter dataset of 1506 pretreatment T1-weighted DCE-MRI cases with expert segmentations. All models were evaluated on a fixed held-out test set (N=306). Each architecture was preprocessed according to pipeline-specific configuration, including slice-wise MedSAM inference using ground-truth-derived bounding boxes. Segmentation performance was evaluated using Dice Similarity Coefficient (DSC) and normalised 95th-percentile Hausdorff Distance (NormHD), with hard failure defined as $DSC < 0.5$ to determine whether models fail on the same or different cases. Mean DSC differences were analyzed using a paired patient-level bootstrap with 20,000 resamples.

Results/Outcome

In the validation cohort, nnU-Net achieved the highest mean DSC (0.846 ± 0.105), compared with the next-best SegResNet (0.787 ± 0.185) (Table 1). The mean paired DSC difference was 0.060 (95% CI, 0.045–0.076). Small tumors comprised 33.4% of the cohort but 55.0% of hard cases, a 2.4-fold higher failure rate (Figure 1). Failures were largely shared rather than distinct (Figure 2): of 40 cases failing at least one pipeline, all four nnU-Net failures fell within the baseline failures, and nnU-Net succeeded on 24 of U-Net's 27 hard cases. With ground-truth-derived localization and tumor-positive slice selection, MedSAM achieved a DSC of 0.839 ± 0.099 .

Conclusion

Segmentation failures overlapped across architectures. Small tumor volume was an important marker of failure and should be prioritized for quality assurance and future model development. nnU-Net provided the highest validation performance. Localization-conditioned MedSAM may support human-in-the-loop workflows but requires prior lesion localization, limiting fully automatic use.

Statement of Impact

Segmentation failures in breast MRI are concentrated in small tumors and shared across architectures rather

than resolved by model choice. Prioritizing small tumors for quality assurance and model development can improve how segmentation systems are evaluated and monitored.

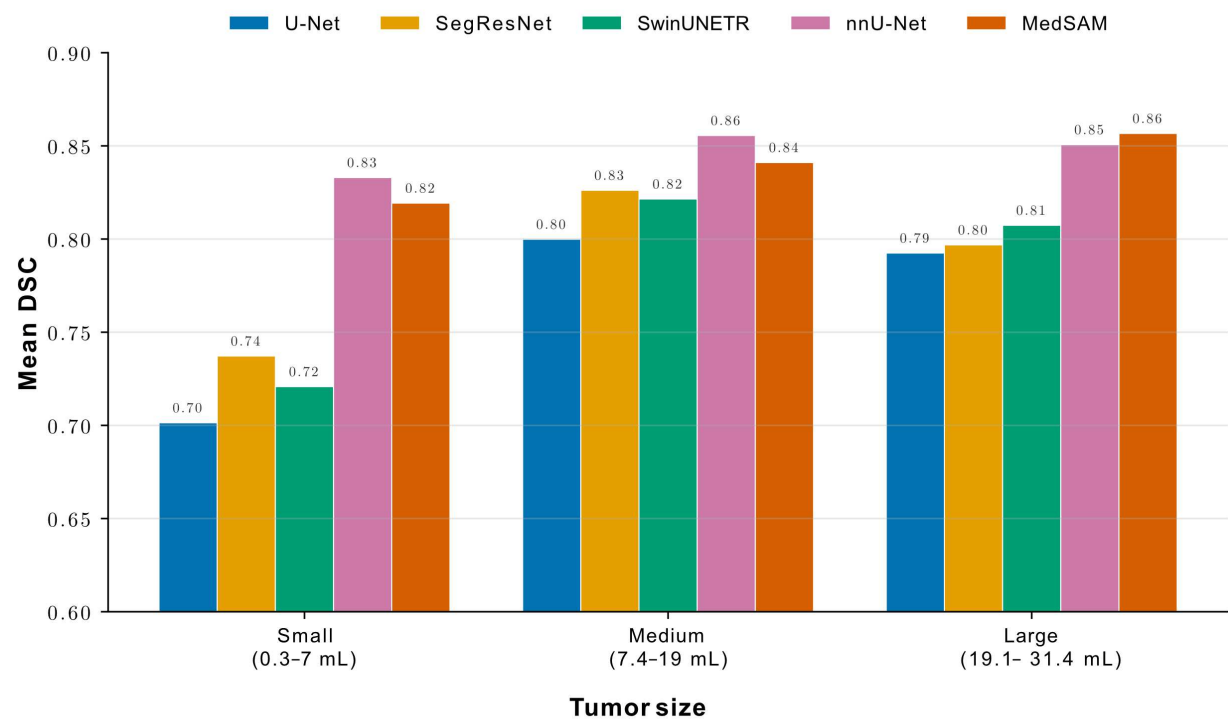


Figure 1. Segmentation accuracy declines with decreasing tumor size, and models differ most on small tumors. Mean Dice similarity coefficient (DSC) by tumor volume (small, medium, large; volume ranges in mL shown on the axis), with tertiles defined by equal case counts within the validation cohort.

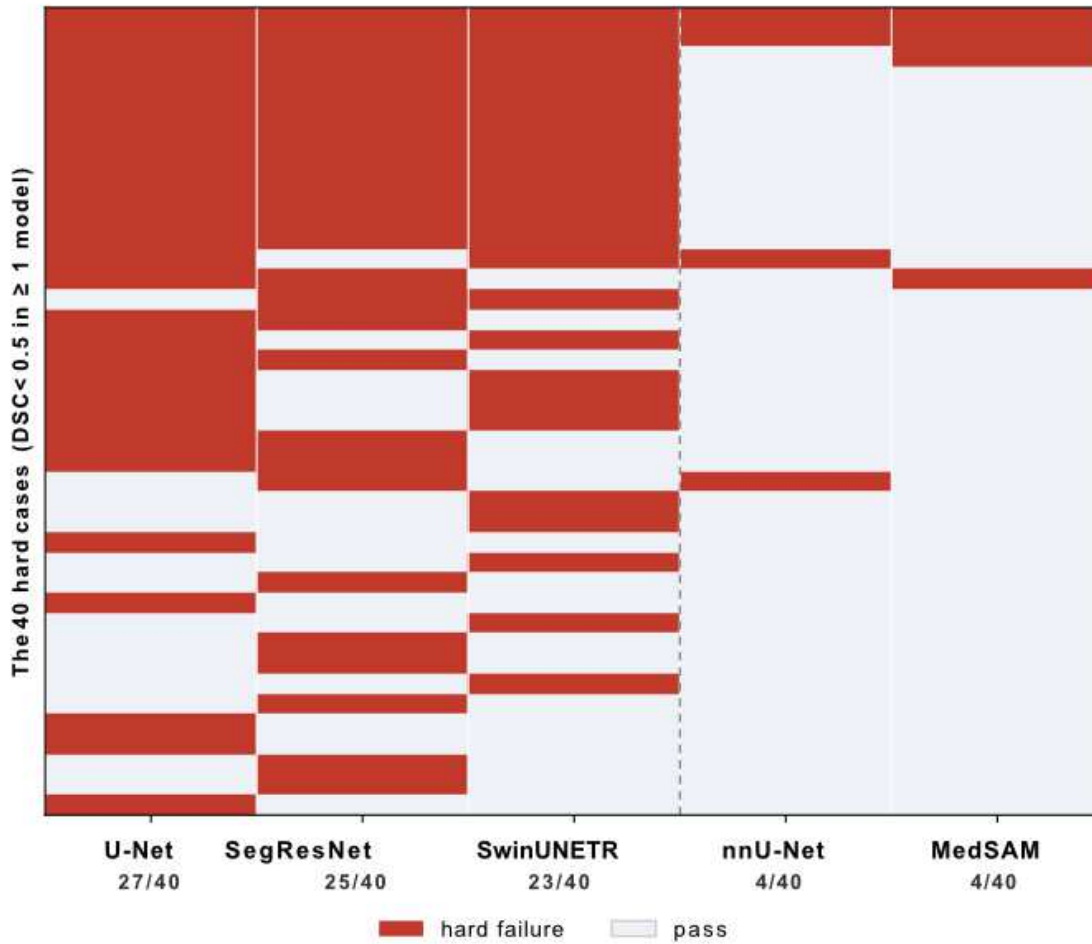


Figure 2. Segmentation failures are shared across architectures rather than model-specific. Each row is one of the 40 cases that failed in at least one of the five pipelines (hard failure defined as DSC < 0.5); red indicates failure and grey indicates a passing segmentation. Columns are ordered by total failures, with per-model counts shown above.

Model	DSC (mean ± SD)	DSC median	NormHD (mean ± SD)	DSC<0.5	DSC<0.3
U-Net	0.76 ± 0.19	0.83	0.100 ± 0.188	8.9%	5.2%
SegResNet	0.79 ± 0.19	0.85	0.091 ± 0.178	8.2%	3.3%
SwinUNETR	0.78 ± 0.18	0.85	0.085 ± 0.170	7.5%	3.3%
nnU-Net ResEnc L	0.85 ± 0.11	0.87	0.016 ± 0.025	1.3%	0.3%
MedSAM	0.84 ± 0.10	0.86	0.018 ± 0.019	1.3%	0.7%

Table 1: Segmentation performance comparison across architectures on the test set.

Keywords

Breast MRI; Deep Learning; Tumor Segmentation; Computer Vision; Model Evaluation

