Slice-Aware Attention for Quality Control of Breast MRI Segmentations

Edwin Ma
Department of Computer Science
Duke University
Durham, NC, USA
edwin.ma@duke.edu

Rachel Gordon
Department of Computer Science
University of Chicago
Chicago, IL, USA
rachelgordon@uchicago.edu

Anna Woodard
UChicago Data Science Institute
University of Chicago
Chicago, IL, USA
annawoodard@uchicago.edu

Ian Foster
Department of Computer Science
University of Chicago
Chicago, IL, USA
foster@cs.uchicago.edu

Kyle Chard
Department of Computer Science
University of Chicago
Chicago, IL, USA
chard@uchicago.edu

2025 IEEE International Conference on eScience (eScience) | 979-8-3315-9145-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/eScience65000.2025.00063

***Abstract*—Quality control (QC) is essential for ensuring the reliability of automated tumor segmentations in breast magnetic resonance imaging (MRI), which are increasingly used in both clinical and research workflows. Manual QC by experts is subjective, time-consuming, and unscalable, highlighting the need for automated solutions. In this work, we propose a 2D convolutional neural network (CNN) with attention-based slice aggregation to classify the quality of tumor segmentations in breast MRI. Leveraging the large-scale, expert-annotated MAMA-MIA dataset, our method processes 2D slices as 2-channel inputs (image and segmentation) and learns to weight their contribution using an attention mechanism for volume-level classification. Our best-performing configuration (attention + augmentation) achieved 64% accuracy, 0.58 F1, and 0.67 AUC, outperforming other settings. These findings show that augmentation and attention are complementary—augmentation increases slice-level diversity, while attention helps identify informative slices among mostly uninformative ones. While performance was moderate, the results establish a baseline for scalable QC in breast MRI, with considerable potential for improvement through enhanced architectures, richer augmentations, and integration with 3D context.**

I. Introduction

MRI is widely used in breast cancer diagnosis and treatment for its high sensitivity, but its growing use creates a bottleneck due to labor-intensive expert review.

Automated tumor segmentation algorithms can reduce workload but are prone to failure in cases of low contrast, artifacts, or anatomical variability. Therefore, quality control (QC) is critical to ensure segmentation reliability.

Manual QC is subjective, time-consuming, and unscalable. To address this, we propose a 2D convolutional neural network (CNN) with attention-based slice aggregation for automated breast segmentation QC, which produces interpretable volume-level predictions.

This work was supported in part by the BigDataX REU program, NSF grant, 2150500. The authors also acknowledge the use of ChatGPT (GPT-4o, OpenAI) to assist in rephrasing and refining portions of the manuscript. All AI-generated text was reviewed, edited, and integrated by the authors.

II. Background and Related Work

While deep learning has advanced breast tumor segmentation [2], segmentation quality varies across scanner types, image quality, and patient anatomy, necessitating robust QC systems.

Early QC approaches used handcrafted features and heuristics, while more recent work employed supervised classifiers on image and mask features, though generalization remains limited [3]. CNN-based QC has shown success in other domains such as brain and cardiac imaging [4], often using 2D or 3D architectures with multiple instance learning (MIL) for slice aggregation. However, automated QC with attention for breast tumor segmentation remains underexplored.

III. Dataset

We use the MAMA-MIA dataset [1], a large-scale, multi-center breast MRI collection with 1,506 pre-treatment T1-weighted DCE-MRI scans and expert-corrected tumor segmentations. The scans span four TCIA cohorts—ISPY-1, ISPY-2, Duke-Breast-Cancer-MRI, and NACT-Pilot—covering diverse imaging protocols and scanner types.

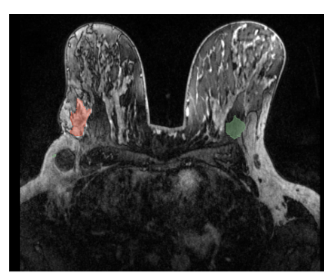

Fig. 1. Example 2D slice with automatic segmentation (red) and expert-corrected segmentation (green).

Each segmentation was double-reviewed by expert radiologists and rated as *Good*, *Acceptable*, *Poor*, or *Missed*. We

2325-3703/25/$31.00 ©2025 IEEE
DOI 10.1109/eScience65000.2025.00063

binarized labels: scans rated *Good* by both reviewers were labeled *good*, all others as *bad*, yielding 411 *good* (50.4%) and 404 *bad* (49.6%) cases.

For splitting, single-breast scans from ISPY-1, ISPY-2, and NACT-Pilot were divided 80/20 into train/test, maintaining a balanced distribution. Duke scans, originally double-breast, were halved by selecting the side containing the automatic mask and treated as single-breast volumes. All inputs were resized and padded to a uniform shape. Data augmentation was applied only to the training set.

Fig. 1 shows an example MRI slice with overlaid automatic and expert segmentations. Full dataset details are available in the original MAMA-MIA publication [1].

IV. Methods

To standardize inputs, we applied N4 bias correction, denoising, resampling, and intensity normalization, and ultimately cropped scans to center the breasts and resized them to $256 \times 256 \times D$ (with D slices). Sagittal scans were automatically excluded.

Our 2D CNN processes MRI–mask slice pairs as 2-channel inputs. A three-layer CNN backbone extracts slice features, aggregated into volume-level predictions via an attention mechanism with entropy regularization to encourage weight diversity. The model is trained end-to-end with binary labels.

Training used Adam (lr $= 1 \times 10^{-4}$) for 50 epochs with patient-wise splits to prevent leakage. BCEWithLogitsLoss with class weighting addressed imbalance, and a grid search tuned dropout, attention dropout, weight decay, and entropy regularization. Data augmentation (applied to images and masks) included horizontal flips ($p = 0.5$), affine transforms ($p = 0.95$), and elastic deformations (magnitude 2–5, spacing 64; $p = 0.95$) (Fig. 2).

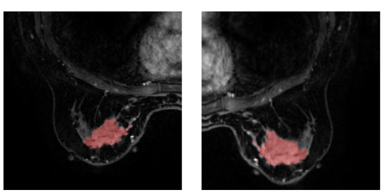

Fig. 2. Original (left) and augmented (right) 2D MRI slice with corresponding mask.

Models were evaluated on accuracy, precision, recall, F1, and AUC, with emphasis on the *bad* class. Implementation used PyTorch on the Purdue Anvil cluster (NVIDIA A100s). Logs, predictions, and attention weights were saved for analysis. Code is available at https://github.com/edwin-ma26/mama-mia-seg-qc.

V. Results

We evaluated three configurations: (1) Attention + Augmentation, (2) Attention only, and (3) Augmentation only. Attention + Augmentation yielded the most stable training and strongest overall performance across accuracy, precision, recall, F1, and AUC. Its advantage stems from attention focusing on the small fraction of slices containing tumors, while augmentation introduced variability to counter intra-scan similarity. Without augmentation, attention models overfit; without attention, augmentation failed to emphasize critical slices.

Table I summarizes validation metrics. The optimal setup depends on task priorities: for minimizing missed tumors, "Augmentation only" offers the highest recall (0.81), while "Attention + Augmentation" achieves better precision and AUC, providing the best balance overall.

TABLE I
Validation performance under different configurations. The combination of attention and augmentation yielded the strongest overall results.

Configuration	**Acc**	**Prec**	**Rec**	**F1**	**AUC**
Attention + Augmentation	0.64	0.68	0.51	0.58	0.67
Attention, No Augmentation	0.56	0.54	0.67	0.60	0.55
No Attention + Augmentation	0.50	0.50	0.81	0.62	0.47

A. Interpretability and Slice Attention

We visualized attention scores across slices to assess the model's focus. As shown in Fig. 3, the model often emphasized slices with segmentation boundaries or visible errors, though attention was inconsistently placed on slices without tumor presence, suggesting noise in slice-level features. Despite these limitations, the results provide a baseline for applying 2D CNNs with attention-based slice aggregation to breast MRI segmentation QC and a foundation for more robust future methods.

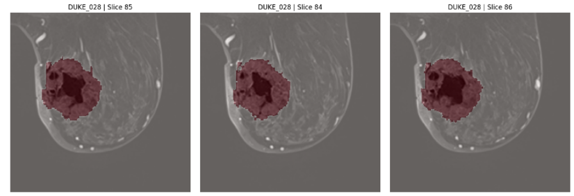

Fig. 3. Top three slices ranked by slice logits for a representative patient. Highlighted slices often correspond to segmentation-relevant regions, though in some cases attention was placed on slices without tumor presence.

References

[1] L. Garrucho *et al.*, "MAMA-MIA: A Large-Scale Multi-Center Breast Cancer DCE-MRI Benchmark Dataset with Expert Segmentations," arXiv:2406.13844v1 [cs.CV], June 2024.:contentReference[oaicite:1]index=1

[2] F. Isensee *et al.*, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, pp. 203–211, 2021.

[3] J. Wasserthal *et al.*, "Automatic quality control for image segmentation using shape, intensity and uncertainty features," in *Proc. MICCAI*, 2021, pp. XX–XX.

[4] R. M. McKinley *et al.*, "Automated detection of cerebral microbleeds in MRI using knowledge distillation," *NeuroImage*, vol. 239, p. 118325, 2021.