Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 23, SEPTEMBER 2025

DOI: 10.37943/23AAOF8219

Beibit Abdikenov
PhD, Director of Science and Innovation Center "Artificial Intelligence"
beibit.abdikenov@astanait.edu.kz, orcid.org/0000-0002-2284-0949
Astana IT University, Kazakhstan

Victor Suvorov
MSc, Researcher at Science and Innovation Center "Artificial Intelligence"
v.suvorov@astanait.edu.kz, orcid.org/0009-0007-1128-8053
Astana IT University, Kazakhstan

CHALLENGES IN GENERALIZING BREAST MRI TUMOR SEGMENTATION ACROSS MULTIPLE DATASETS

Abstract: Accurate segmentation of breast tumors in dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) is essential for precise diagnosis, treatment planning, and quantitative analysis. While deep learning methods have achieved strong performance in controlled research settings, their ability to generalize across diverse clinical datasets remains underexplored and poses a major barrier to clinical adoption. In this study, we evaluate the cross-dataset generalizability of a 3D Residual U-Net model using the multicenter MAMA-MIA benchmark, which consolidates four publicly available breast MRI collections annotated by expert radiologists. A leave-one-out experimental design is employed, with three datasets used for training and validation, and the remaining dataset held-out for independent testing to simulate real-world deployment scenarios. Model performance is assessed using Dice coefficient, Precision, and Recall, alongside quantitative analysis of tumor volume estimation accuracy. The best Dice score achieved by our model was 0.683 when tested on the NACT subset. Results show a consistent degradation in segmentation accuracy when models are applied to unseen datasets, indicating that performance declines significantly outside the distribution of the training data. The most pronounced drop occurs when the DUKE dataset serves as the held-out test set, where the model struggles to adapt to differences in pre-release preprocessing strategies. A targeted qualitative review of 160 representative scans further reveals key factors contributing to both successful and failed segmentations, including variations in image field of view, temporal enhancement patterns, acquisition era, and artifact prevalence. Overall, these findings underscore the importance of accounting for dataset heterogeneity, domain shift, and standardized preprocessing in the development of robust, clinically deployable breast MRI segmentation models capable of generalizing across institutions and imaging protocols.

Keywords: Breast Cancer, Magnetic Resonance Imaging, Tumor Segmentation, Deep Learning, Model Generalizability, Medical Image Analysis, Model Robustness.

Introduction

Breast cancer continues to be its most diagnosed variety and a leading cause of cancer-related deaths among women worldwide. According to the WHO [17], approximately 2.3 million new cases were recorded in 2022 alone, causing 670 000 deaths globally. Prompt detection, correct diagnosis and effective treatment are crucial in the fight against the disease, and medical imaging plays a central role in this process.

Received: 15.08.2025 Accepted: 24.09.2025 Published: 30.09.2025

Magnetic resonance imaging in breast cancer

While mammography and ultrasound remain the standard tools for early detection and evaluation of breast tumors, magnetic resonance imaging (MRI) has emerged as a highly sensitive modality particularly useful for screening high-risk individuals, handling complex diagnostic scenarios and planning treatments. In practice, breast MRI is often performed using a dynamic contrast-enhanced (DCE) protocol, in which a gadolinium-based contrast is administered, and a series of scans is acquired over time. This approach visualizes the temporal enhancement patterns associated with tissue vascularity and permeability which are key indicators of malignancy.

Another advantage of DCE-MRI is its ability to visualize not only the presence of a tumor but also its spatial dimensions, which enables the precise delineation of its boundaries. This process, known as segmentation, involves localizing the tumor and outlining its boundaries within the imaging volume. Accurate segmentation helps quantify tumor size and volume, assess multifocality and make decisions about surgery and therapy, as well as enable the use of computational methods of analysis.

Segmentation can be performed manually, but it is time-consuming, labour-intensive and subject to significant inter-observer variability, especially in ambiguous cases. Automatic segmentation methods, most often based on machine learning or deep learning algorithms, have been developed to address these limitations and produce fast, consistent and reproducible results. These tools not only support clinical decision-making at scale but also enable large-cohort studies and high-throughput analysis, which makes them increasingly valuable in clinical and research settings.

Related work

CNNs are widely used in computer vision, especially for medical image segmentation. Yue et al. [19] enhanced U-Net [14] with residual blocks for breast tumor segmentation. Khaled et al. [9] used an ensemble of three U-Net models trained on different inputs to capture varied lesion features. Rahimpour et al. [13] combined three 3D U-Nets with manual selection by radiologists to better handle outliers and scan variability. Guo et al. [5] improved CNN results using an SVM post-processing step to refine boundaries and reduce noise.

The rise of deep learning in medical imaging has led to adaptable frameworks like nnU-Net, introduced by Isensee et al. [8], which automates preprocessing, hyperparameter tuning, training, and inference using established heuristics. Xu et al. [18] applied nnU-Net to segment triple-negative breast cancer (TNBC) in DCE-MRI scans.

Attention mechanisms [1] and Transformers [16] have gained traction in medical image segmentation for their ability to highlight relevant regions and capture complex patterns. Huang et al. [7] introduced a joint-phase attention approach that integrates pre- and post-contrast MRI features. Zhang et al. [21] and He et al. [6] incorporated multiscale and spatial-temporal attention to enhance tumor detection in dynamic imaging. Meanwhile, Transformers—originally developed for NLP—have shown strong performance in segmentation tasks by modeling long-range dependencies. Qin et al. [12] embedded Transformers into a U-Net-based framework for better feature representation, and Zhang et al. [20] used a spatial-temporal Transformer model to track contrast changes across MRI phases.

Despite their strong reported performance, most of the cited studies share common limitations that constrain their applicability beyond the research setting. Many rely on single-institution or in-house datasets, which restrict the diversity of imaging conditions and patient populations encountered during training. External validation on independent datasets is seldom performed, limiting evidence of robustness in unseen clinical environments. Furthermore, few works provide in-depth analyses of failure cases or explore how variations in data distribution affect performance. These gaps raise critical questions about the generalizability of current breast MRI segmentation approaches to broader, real-world scenarios.

To address these limitations, we conduct a systematic evaluation of model generalizability using the publicly available MAMA-MIA dataset [4], which consolidates four pre-existing breast MRI datasets annotated by a single team of experts. Using a consistent ResUNet architecture similar to [19] and fixed hyperparameters, we adopt a leave-one-out experimental design, where three datasets are used for training and validation, and the remaining dataset is reserved exclusively for testing. Model performance is assessed across Dice, Precision, and Recall metrics on both validation and held-out test sets to reveal differences between in-distribution and out-of-distribution performance. To gain finer insight into the sources of segmentation difficulty, we manually examine 160 randomly selected scans, identifying imaging and lesion characteristics that challenge model fitting and cross-dataset generalization.

This paper makes several contributions:

Conducts a systematic cross-dataset evaluation of breast DCE-MRI tumor segmentation using the multicenter MAMA-MIA benchmark.

Provides an in-depth quantitative and qualitative characterization of domain shift effects in breast MRI segmentation, linking performance drops to concrete factors.

Offers actionable guidelines for improving model generalizability.

The remainder of this paper is organized as follows: Section 2 describes the materials and methods used in this study, including the composition of the dataset, preprocessing steps, model architecture, and training procedure. Section 3 presents the evaluation results, reporting quantitative performance metrics for all experimental configurations. Section 4 provides a detailed discussion of the results, supported by qualitative analysis of selected cases, and examines factors that influence model generalization across datasets.

Methods and Materials

This section describes the materials and methods used in this study, including the composition and characteristics of the dataset, preprocessing steps, model architecture, experimental design, and implementation details. Together, these elements define the framework used to investigate the generalizability of breast MRI segmentation models across heterogeneous data sources.

Dataset

The MAMA-MIA dataset [4] is a large-scale, multicenter benchmark resource designed to address the scarcity of expert-labeled breast DCE-MRI data. It includes 1,506 pre-treatment T1-weighted DCE-MRI cases with expert-verified segmentations of primary tumors and non-mass-enhanced regions.

Data was collected from four publicly available TCIA collections: I-SPY1 [3], I-SPY2 [11], NACT-Pilot [10] and Duke-Breast-Cancer-MRI [15]. Selection focused on pre-treatment cases with available clinical outcome data, especially pathological complete response (pCR) and five-year survival status. As most of the original datasets lacked high-quality segmentation masks, additional annotation was provided by 16 expert radiologists.

The MAMA-MIA dataset exhibits substantial diversity in imaging characteristics, making it particularly valuable for the development and evaluation of robust breast MRI segmentation methods. It includes both bilateral and unilateral scans, acquired using axial and sagittal planes, thereby capturing variability in clinical imaging practices. The MRIs were obtained using scanners with magnetic field strengths of 1.5T (72% of cases) and 3.0T (28%), and the number of DCE-MRI phases per case ranges from 3 to 11, with significant variability in inter-phase timing (e.g., mean time between pre- and first post-contrast: 203 seconds, range: 27–922 seconds). Additionally, the dataset encompasses a wide range of acquisition parameters such as slice thickness, pixel spacing, image matrix size, and scanner manufacturers (GE, Siemens, Philips).

Preprocessing

As a first step in our pipeline, we preprocess the DCE-MRI data to ensure consistency across patients. For each patient, the structural (pre-contrast) and first post-contrast MRI scans are load. Since scans differ in acquisition plane (axial or sagittal) and spatial resolution, we perform anisotropic resampling to standardize voxel spacing across samples. Intensity normalization is carried out using within-sample z-score normalization, where the mean and standard deviation are computed jointly across both phases maintain consistency in intensity scaling. During training, we randomly extract smaller cubic patches from the volumes to reduce GPU memory usage and to handle differences in image dimensions between patients. To further enhance model robustness and reduce overfitting, we apply random flipping as a form of spatial data augmentation.

Model architecture

We employ a 3D Residual U-Net (ResUNet) segmentation network for all experiments, keeping the architecture and hyperparameters fixed across runs to ensure comparability. This choice is motivated by its widespread adoption in medical image segmentation and its close similarity to the U-Net variant implemented in the widely used nnU-Net framework [8], which has consistently demonstrated performance on par with state-of-the-art methods across diverse datasets.

The network takes two input channels (one for the pre-contrast MRI and one for the first post-contrast MRI) and produces a single-channel tumor probability map. The encoder path consists of six downsampling stages. Each stage contains two residual convolutional units, where the output is computed as

$$y = F(x;\, \theta) + x, \tag{1}$$

with x denoting the input feature map and $F(x; \theta)$ representing two consecutive 3D convolution – instance normalization – PReLU operations with $3\times3\times3$ kernels. Explicitly,

$$F(x;\, \theta) = \sigma\Big(\mathrm{IN}\big(W_2 * \sigma\big(\mathrm{IN}(W_1 * x)\big)\big)\Big)\,, \tag{2}$$

where IN is instance normalization, σ is the PReLU activation, and $*$ denotes 3D convolution.

Downsampling is performed at the start of each encoder stage using a strided convolution:

$$x'_d = x *_{s=2} W_d, \tag{3}$$

which reduces the spatial resolution by a factor of two. The number of feature channels increases with depth, starting at 32 and progressing through 64, 128, 256, 380 and again 380 channels. The decoder path mirrors the encoder with four upsampling stages implemented via transpose convolutions:

$$x'_u = x \star_{s=2} W_u, \tag{4}$$

where $*_{s=2}$ denotes a transposed convolution with stride 2. Skip connections link each encoder stage to its corresponding decoder stage, concatenating the features before further processing. Instance normalization and PReLU activation are applied throughout the network. The complete network architecture is shown in Figure 1.

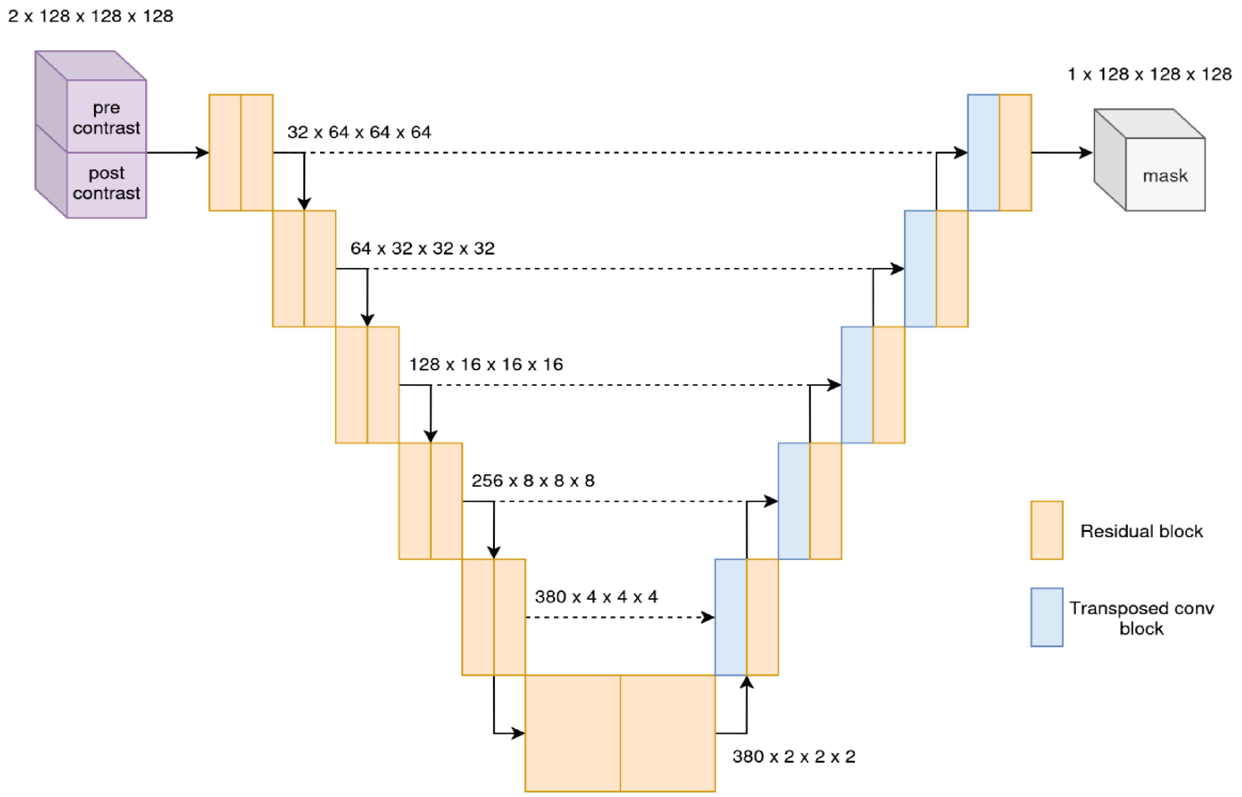

Figure 1. Architecture of the segmentation model used in this study

Methodology

To investigate cross-dataset generalizability, we adopt a leave-one-out experimental design. At each iteration, three of the four constituent MAMA-MIA sub-datasets are used for model development, while the remaining dataset is held out entirely for independent testing. The combined development set is randomly partitioned into 80% for training and 20% for validation, without regard to sub-dataset membership. For example, one configuration trains on I-SPY1, I-SPY2, and NACT-Pilot (split 80−20) and evaluates on Duke-Breast-Cancer-MRI. This process is repeated so that each sub-dataset serves as the held-out test set exactly once, yielding a total of four trained models.

After training, each model is evaluated on both its validation set (in-distribution) and its held-out test set (out-of-distribution). Segmentation performance is quantified using the Dice similarity coefficient, Precision, and Recall. In addition, we compute the absolute tumor volume in mm^3 for both the ground-truth and predicted masks to analyse systematic biases in lesion size estimation.

To obtain a deeper understanding of factors affecting generalization, we perform a targeted manual review of selected cases. For each model and each evaluation set (validation and test), we sample 20 scans according to the following scheme:

- 5 scans from the lowest 10% of Dice scores
- 5 scans from the highest 10% of Dice scores
- 10 scans from the remaining 80%

This procedure results in $4\times2\times20=160$ unique scans for detailed visual inspection. The selected cases are then examined qualitatively to identify imaging characteristics, anatomical variations, and lesion features that may contribute to segmentation success or failure.

DOI: 10.37943/23AAOF8219
© Beibit Abdikenov, Victor Suvorov

Implementation details

All experiments were conducted on a workstation running Ubuntu 22.04, equipped with an NVIDIA GeForce GTX 1080 Ti GPU with 11 GB of VRAM. The neural network was trained with PyTorch, with the model architecture sourced from MONAI's [2] library of pre-implemented neural network classes. Data loading and preprocessing were also handled using MONAI's medical imaging framework.

The training configuration was as follows:

- Batch size: 12
- Patch size: $128 \times 128 \times 128$ voxels
- Optimizer: Adam with an initial learning rate of 0.01
- Learning rate scheduler: PolynomialLR with a power of 1.5
- Maximum number of epochs: 100
- Loss function: Dice loss
- Validation sliding window patch overlap: 50%
- Early stopping tolerance: 20 epochs

These settings were kept constant across all experiments to ensure that performance differences arose solely from dataset composition rather than training variability.

Results

Table 1 summarizes the performance of the four trained models, each identified by the dataset that was excluded from training and used as the held-out test set. For example, the row labeled "DUKE" corresponds to the model trained on ISPY1, ISPY2, and NACT, with DUKE reserved exclusively for testing. For each model, metrics are reported separately for the held-out test set and for the corresponding validation set. Validation results are presented both as an aggregate across all three training sub-datasets and individually for each sub-dataset to capture in-distribution variability. Reported metrics include the mean Dice coefficient, Precision, and Recall, along with the mean ground-truth and predicted tumor volumes (in mm^3), computed per patient. Precision and Recall are calculated voxel-wise, and all reported values represent averages over individual patient scores.

Table 1. Performance of models identified by the held-out test dataset. Metrics are shown for the test set and for validation (aggregate and per sub-dataset), including mean Dice, Precision, Recall, and ground-truth and predicted volumes (mm^3) averaged per patient.

Dataset	val / test	Dice	Precision	Recall	True volume (mm3)	Pred volume (mm3)
DUKE	val overall	0.743	0.751	0.802	24919	24395
	val ISPY1	0.760	0.765	0.812	46545	34639
	val ISPY2	0.745	0.754	0.803	20667	22392
	val NACT	0.670	0.652	0.758	29912	26555
	test	0.260	0.191	0.743	21120	63494
ISPY1	val overall	0.724	0.739	0.795	19616	19755
	val DUKE	0.512	0.483	0.737	12914	23851
	val ISPY2	0.772	0.798	0.802	20943	18440
	val NACT	0.670	0.619	0.889	21585	28239
	test	0.629	0.641	0.727	35257	24802
ISPY2	val overall	0.497	0.520	0.691	32510	35413
	val DUKE	0.382	0.386	0.659	30218	43399
	val ISPY1	0.639	0.647	0.759	28707	29056
	val NACT	0.602	0.737	0.648	51185	19360
	test	0.499	0.624	0.515	22352	15187
NACT	val overall	0.709	0.731	0.772	22520	20119
	val DUKE	0.574	0.564	0.733	17404	21086
	val ISPY1	0.682	0.720	0.764	34686	34111
	val ISPY2	0.759	0.792	0.787	22833	18054
	test	0.683	0.686	0.804	30741	27940

In most configurations, model performance was higher on the validation sets than on the held-out test sets, reflecting a clear drop in accuracy when applied to unseen data. The largest decline was observed when DUKE served as the test set, with the Dice score falling from 0.743 on the aggregated validation set to 0.260 on the test set, suggesting a substantial distribution shift between DUKE and the other datasets. Even within the in-distribution validation data, performance varied notably across individual sub-datasets. For example, when ISPY1 was excluded from training, validation Dice scores ranged from 0.512 on DUKE to 0.772 on ISPY2, indicating heterogeneity in difficulty even among datasets used for training.

Precision–Recall balance also varied by configuration: in some cases, such as DUKE as the test set, models exhibited high Recall but low Precision (0.743 vs. 0.191), consistent with over-segmentation; in others, such as ISPY2 as the test set, higher Precision relative to Recall (0.624 vs. 0.515) suggested under-segmentation. Tumor volume estimates showed similar variability, with the largest overestimation again occurring for DUKE in the test split (21,120 mm^3 ground truth vs. 63,494 mm^3 predicted), while marked underestimation was observed in several validation subsets, such as NACT in the ISPY2 model (51,185 mm^3 ground truth vs. 19,360 mm^3 predicted). Overall, DUKE emerged as the most challenging dataset for cross-dataset generalization, while ISPY1 and NACT were comparatively less affected by the domain shift when used as held-out test sets.

Discussion

To better understand the factors influencing model generalization, we performed a qualitative review of a representative subset of scans from each validation and test set. This visual inspection of individual cases, spanning a range of segmentation performances, revealed several recurring patterns and dataset-specific characteristics that help explain the quantitative results. Key observations from this analysis are presented below.

A key finding of this study is the markedly poor perfomance of models when DUKE was used as the held-out test set, with Dice scores dropping sharply compared to all other configurations. This can be traced to a fundamental difference in the way the source datasets were preprocessed prior to inclusion in MAMA-MIA. The MRI scans in MAMA-MIA do not retain their raw, originally acquired form; instead, they are the result of preprocessing pipelines applied before public release. While ISPY1, ISPY2, and NACT contain images that are often cropped to include only the breast with the tumor (in unilateral cases), DUKE scans consistently include the full MRI field of view. Figure 2 illustrates this difference in coverage.

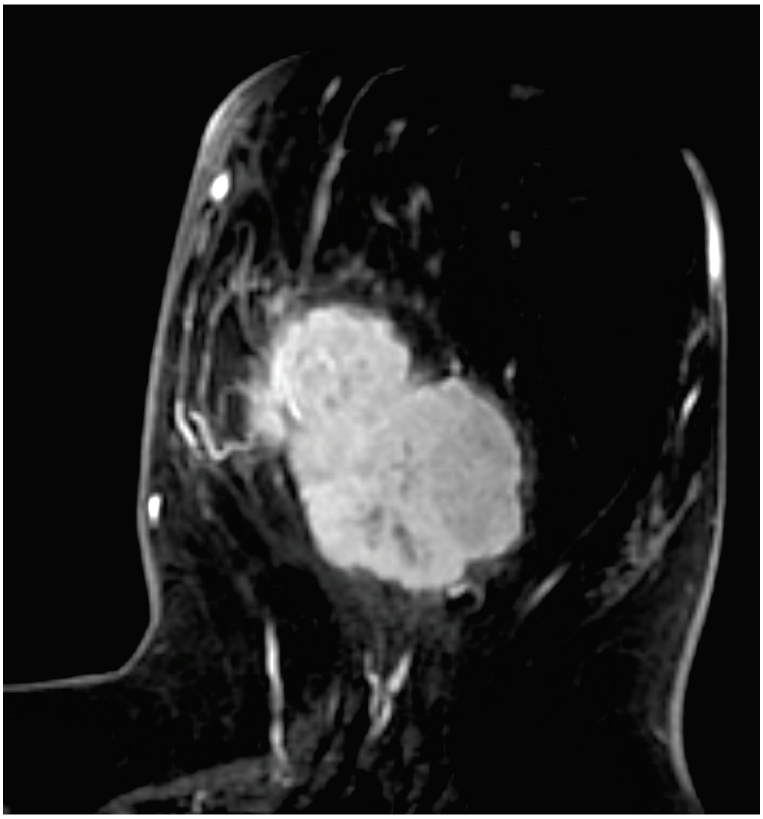

Figure 2. Example from the ISPY2 dataset in its publicly released form, cropped to include only the breast containing the tumor. The cropping was performed prior to release of the dataset and is not part of this study's preprocessing.

This disparity is quantitatively reflected in Table 2, which reports the median total scan volume for each sub-dataset. DUKE's median volume is 21,513 cm^3, while the other three datasets are all around 5,000 cm^3 or less, indicating a high prevalence of cropping in ISPY1, ISPY2, and NACT. As a result, models trained on cropped images encounter a substantial distribution shift when exposed to DUKE: sliding-window inference produces a large number of patches containing only healthy tissue, a scenario rarely seen during training. This mismatch likely drives both the low Dice scores and the systematic overestimation of tumor volumes, as the model tends to generate false-positive segmentations in regions of healthy parenchymal enhancement where it "expects" to find abnormalities. An example of this phenomenon is shown in Figure 3, where widespread false-positive masks appear in otherwise normal tissue. This issue can be addressed by including whole breast segmentation into the preprocessing pipeline.

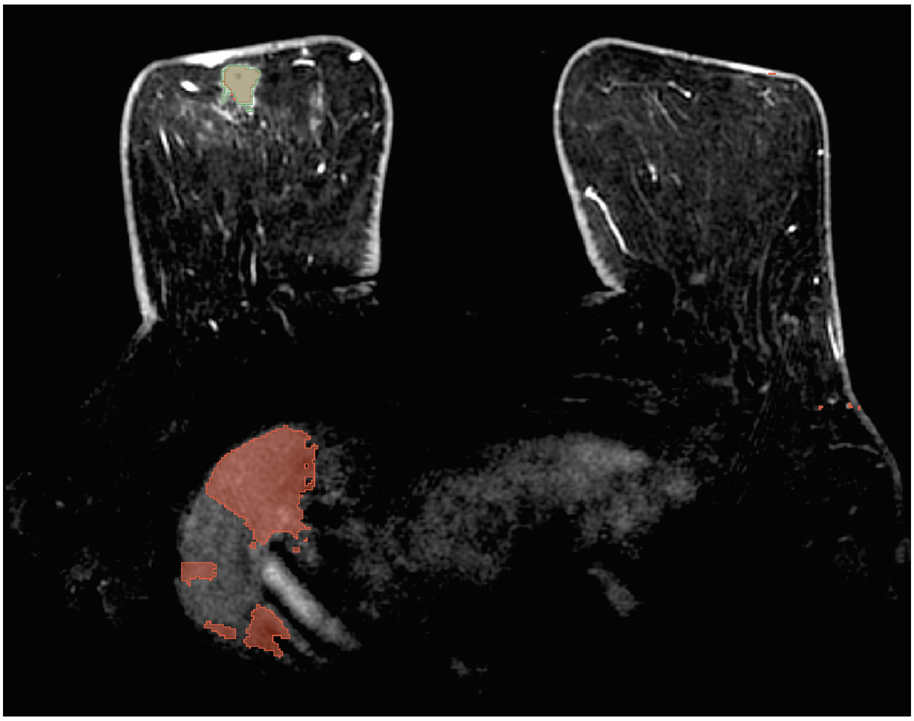

Figure 3. Example from the DUKE dataset showing the predicted mask (red) and the ground-truth mask (green). While the model accurately segments the true tumor, multiple false-positive regions outside the breast substantially reduce the Dice score.

Table 2. Median total scan volume for each sub-dataset in MAMA-MIA.

Dataset	Median volume (cm3)
DUKE	21513
ISPY1	5120
ISPY2	5119
NACT	3888

Bright imaging artifacts and non-tumorous enhancements within the breast are also a frequent source of false positives, as illustrated in Figure 4. This issue could potentially be mitigated by providing more DCE-MRI phases as input. In our design, only the pre-contrast anatomical reference and the first post-contrast scan are used. Access to additional phases would allow reconstruction of the kinetic enhancement curve of the lesion, enabling better differentiation between malignancies, healthy tissue, and consistently bright artifacts. However, incorporating more temporal information is challenging: the number of available phases varies significantly within the dataset, and restricting analysis to the lowest common number discards a substantial amount of potentially informative data. Furthermore, in some cases, later post-contrast phases provide greater lesion enhancement and improved background contrast compared to the first post-contrast scan, as shown in Figure 5.

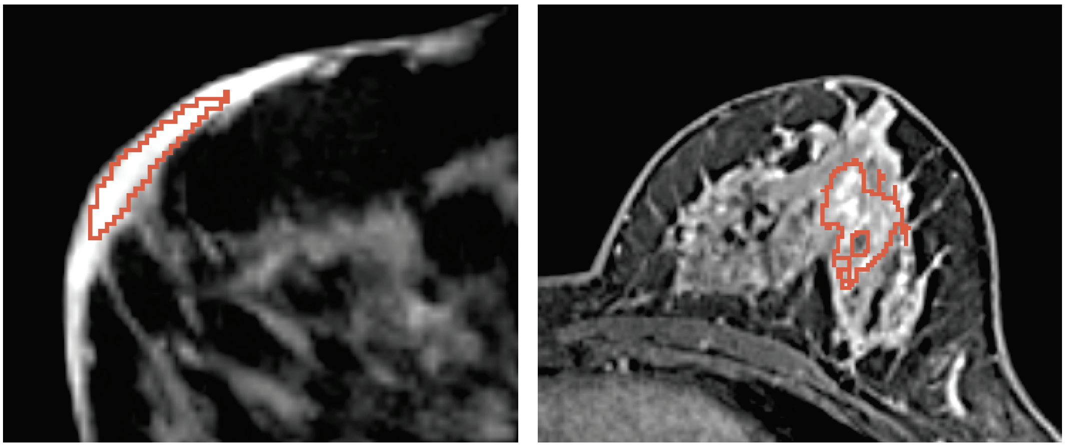

Figure 4. Examples of false positives caused by non-tumorous regions. The left panel shows an MRI artifact along the breast boundary, while the right panel shows an area of benign enhancement within the breast parenchyma.

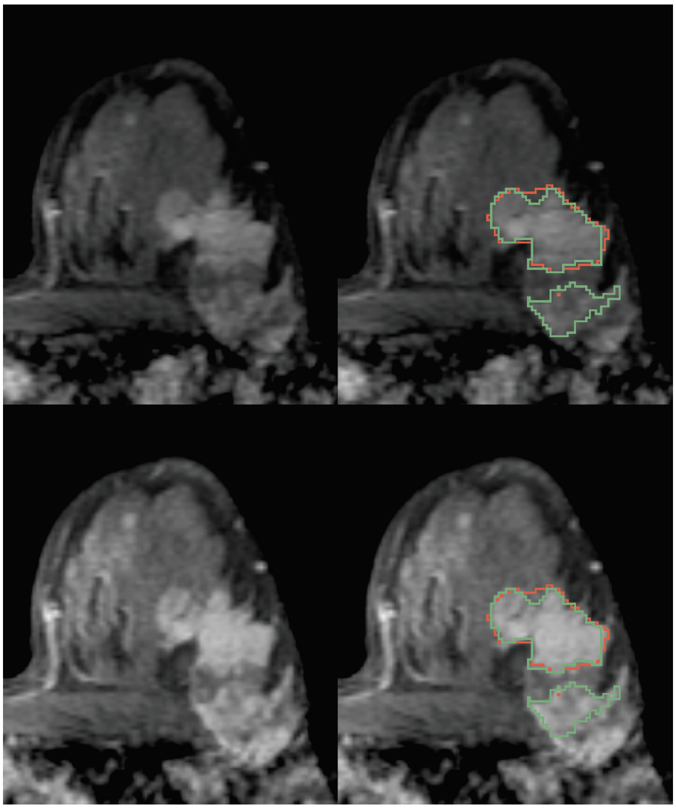

Figure 5. Example illustrating delayed enhancement effects. The top row shows the first post-contrast image, without masks (left) and with predicted (red) and ground-truth (green) masks (right). The bottom row shows the corresponding fourth post-contrast image. In the later phase, the lower section of the tumor—missed by the prediction in the first post-contrast image—exhibits greater enhancement and improved contrast with surrounding tissue.

The drop in performance metrics with ISPY1 as the test set is also noticeable and significant, even if not as severe as the drop observed for DUKE. One contributing factor may be the age of the imaging in ISPY1, as most scans were acquired in the mid-1980s. Figure 6 shows an example scan from 1986 in which the model, trained on the other three datasets, failed to segment a large visible tumor. This failure is likely due to the lower image quality and greater noise levels in these older scans, especially when compared to the more recent imaging found in the other datasets, which primarily includes acquisitions from the late 1990s, 2000s, and later.

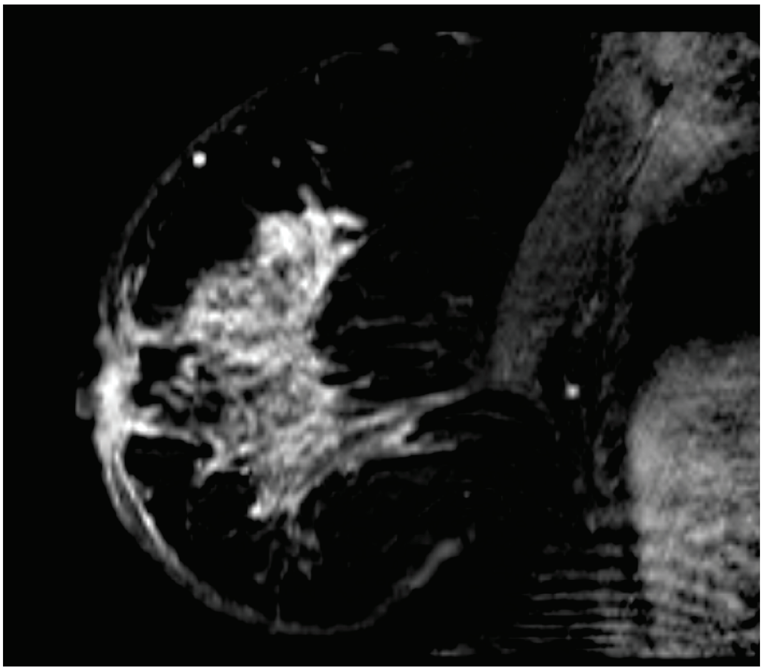

Figure 6. Example from the ISPY1 dataset acquired in 1986. The model trained on the other three datasets failed to segment the large tumor visible in the image, likely due to the lower quality and higher noise levels of older scans compared to the more recent imaging in the other datasets.

Overall, the case-level analysis highlights how differences in dataset preprocessing, imaging protocols, acquisition eras, and inherent anatomical or physiological variability can substantially impact segmentation performance. While some of these limitations may be addressed through harmonization of preprocessing pipelines, richer temporal information, and targeted augmentation strategies, others reflect fundamental shifts in data distribution that are more challenging to overcome. These findings underscore the importance of considering dataset composition and heterogeneity when developing and evaluating models intended for broad clinical deployment.

Conclusion

In summary, this study highlights the challenges of achieving robust cross-dataset generalization in breast DCE-MRI tumor segmentation. By systematically evaluating models across four diverse public datasets and performing targeted case reviews, we identified key sources of performance degradation, including differences in field of view, presence of bright non-tumorous enhancements, temporal enhancement patterns, and historical variations in image quality. These findings emphasize that models can suffer substantial drops in performance when confronted with domain shifts inherent to multi-institutional imaging data. Addressing such challenges will require not only improved modeling strategies, but also greater harmonization of acquisition protocols, preprocessing approaches, and annotation standards across datasets. Ultimately, bridging these gaps is essential for translating automated breast MRI analysis into reliable, clinically deployable tools.

Acknowledgements

This research was funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan, grant number BR24993145.

References

[1] Bahdanau, D., Cho, K., & Bengio, Y. (2016). *Neural machine translation by jointly learning to align and translate*. https://arxiv.org/abs/1409.0473

[2] Cardoso, M. J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., Nath, V., He, Y., Xu, Z., Hatamizadeh, A., Myronenko, A., Zhu, W., Liu, Y., Zheng, M., Tang, Y., . . Feng, A. (2022). *Monai: An open-source framework for deep learning in healthcare*. https://arxiv.org/abs/2211.02701

[3] Chitalia, R., Pati, S., Bhalerao, M., Thakur, S., Jahani, N., Belenky, J. V., McDonald, E. S., Gibbs, J., Newitt, D., Hylton, N., Kontos, D., & Bakas, S. (2021). *Expert tumor annotations and radiomic features for the ISPY1/Acrin 6657 trial data collection*. https://doi.org/10.7937/TCIA.XC7A-QT20

[4] Garrucho, L., Kushibar, K., Reidel, C.-A., Joshi, S., Osuala, R., Tsirikoglou, A., Bobowicz, M., Riego, J. d., Catanese, A., Gwo´zdziewicz, K., Cosaka, M.-L., Abo-Elhoda, P.M., Tantawy, S.W., Sakrana, S.S., Shawky-Abdelfatah, N.O., Salem, A.M.A., Kozana, A., Divjak, E., Ivanac, G., . . . Lekadir, K. (2025). *A large-scale multicenter breast cancer DCE-MRI benchmark dataset with expert segmentations*. Scientific Data, 12 (1), 453. https://doi.org/10.1038/s41597-025-04707-4

[5] Guo, Y. Y., Huang, Y. H., Wang, Y., Huang, J., & … (2022). *Breast MRI tumor automatic segmentation and triple-negative breast cancer discrimination algorithm based on deep learning* [Publisher: Wiley Online Library]. . . . Methods in Medicine. https://doi.org/10.1155/2022/2541358

[6] He, J., Zhao, X., Luo, Z., Su, S., Li, S., & Zhang, G. (2024). *TSESNet: Temporal-spatial enhanced breast tumor segmentation in DCE-MRI using feature perception and separability*. Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, 803–811. https://doi.org/10.24963/ijcai.2024/89

[7] Huang, R., Xu, Z., Xie, Y., Wu, H., Li, Z., Cui, Y., Huo, Y., Han, C., Yang, X., Liu, Z., & Wang, Y. (2023). *Joint-phase attention network for breast cancer segmentation in DCE-MRI* [Publisher: Elsevier BV]. Expert Systems with Applications, 224, 119962. https://doi.org/10.1016/j.eswa.2023.119962

[8] Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2020). *nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation*. Nature Methods, 18 (2), 203–211. https://doi.org/10.1038/s41592-020-01008-z

[9] Khaled, R., Vidal, J., Vilanova, J. C., & Mart´ı, R. (2022). *A u-net ensemble for breast lesion segmentation in DCE MRI*. [Place: United States]. Computers in biology and medicine, 140, 105093. https://doi.org/10.1016/j.compbiomed.2021.105093

[10] Newitt, D., & Hylton, N. (2016). *Single site breast DCE-MRI data and segmentations from patients undergoing neoadjuvant chemotherapy*. https://doi.org/10.7937/K9/TCIA.2016.QHSYHJKY

[11] Newitt, D.C., Partridge, S.C., Zhang, Z., Gibbs, J., Chenevert, T., Rosen, M., Bolan, P., Marques, H., Romanoff, J., Cimino, L., Joe, B.N., Umphrey, H., Ojeda-Fournier, H., Dogan, B., Oh, K.Y., Abe, H., Drukteinis, J., Esserman, L.J., & Hylton, N.M. (2021). *Acrin 6698/I-SPY2 breast DWI*. https://doi.org/10.7937/TCIA.KK02-6D95

[12] Qin, C., Wu, Y., Zeng, J., Tian, L., Zhai, Y., Li, F., & Zhang, X. (2022). *Joint transformer and multi-scale CNN for DCE-MRI breast cancer segmentation* [Publisher: Springer]. Soft Computing, 26 (17), 8317–8334.

[13] Rahimpour, M., Saint Martin, M.-J., Frouin, F., Akl, P., Orlhac, F., Koole, M., & Malhaire, C. (2022). *Visual ensemble selection of deep convolutional neural networks for 3D segmentation of breast tumors on dynamic contrast enhanced MRI* [Publisher: Springer Science and Business Media LLC]. European Radiology, 33 (2), 959–969. https://doi.org/10.1007/s00330-022-09113-7

[14] Ronneberger, O., Fischer, P., & Brox, T. (2015). *U-net: Convolutional networks for biomedical image segmentation*. https://arxiv.org/abs/1505.04597

[15] Saha, A., Harowicz, M.R., Grimm, L.J., Weng, J., Cain, E.H., Kim, C.E., Ghate, S.V., Walsh, R., & Mazurowski, M.A. (2021). *Dynamic contrast-enhanced magnetic resonance images of breast cancer patients with tumor locations.* https://doi.org/10.7937/TCIA.E3SV-RE93

[16] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). *Attention is all you need.* https://arxiv.org/abs/1706.03762

[17] World Health Organization. (2021). *Breast cancer.* Geneva, Switzerland: World Health Organization. Retrieved from https://www.who.int/news-room/fact-sheets/detail/breast-cancer

[18] Xu, Z., Rauch, D.E., Mohamed, R.M., Pashapoor, S., Zhou, Z., & … (2023). *Deep learning for fully automatic tumor segmentation on serially acquired dynamic contrast-enhanced MRI images of triple-negative breast cancer* [Publisher: mdpi.com Type: HTML]. Cancers. https://www.mdpi.com/2072-6694/15/19/4829

[19] Yue, W., Zhang, H., Zhou, J., Li, G., Tang, Z., Sun, Z., & … (2022). *Deep learning-based automatic segmentation for size and volumetric measurement of breast cancer on magnetic resonance imaging* [Publisher: frontiersin.org Type: HTML]. Frontiers in … https://doi.org/10.3389/fonc.2022.984626

[20] Zhang, J., Cui, Z., Shi, Z., Jiang, Y., Zhang, Z., Dai, X., Yang, Z., Gu, Y., Zhou, L., Han, C., Huang, X., Ke, C., Li, S., Xu, Z., Gao, F., Zhou, L., Wang, R., Liu, J., Zhang, J., … Shen, D. (2023). *A robust and efficient AI assistant for breast tumor segmentation from DCE-MRI via a spatial-temporal framework.* [Place: United States]. Patterns (New York, N.Y.), 4 (9), 100826. https://doi.org/10.1016/j.patter.2023.100826

[21] Zhang, Q., Xiao, J., & Zheng, B. (2023). *Image segmentation of triple-negative breast cancer by incorporating multiscale and parallel attention mechanisms* (S. Hussain, Ed.) [Publisher: Wiley]. Scientific Programming, 2023, 1–13. https://doi.org/10.1155/2023/6629189