



OPEN

DATA DESCRIPTOR

A large-scale multicenter breast cancer DCE-MRI benchmark dataset with expert segmentations

Lidia Garrucho *et al.*[#]

Artificial Intelligence (AI) research in breast cancer Magnetic Resonance Imaging (MRI) faces challenges due to limited expert-labeled segmentations. To address this, we present a multicenter dataset of 1506 pre-treatment T1-weighted dynamic contrast-enhanced MRI cases, including expert annotations of primary tumors and non-mass-enhanced regions. The dataset integrates imaging data from four collections in The Cancer Imaging Archive (TCIA), where only 163 cases with expert segmentations were initially available. To facilitate the annotation process, a deep learning model was trained to produce preliminary segmentations for the remaining cases. These were subsequently corrected and verified by 16 breast cancer experts (averaging 9 years of experience), creating a fully annotated dataset. Additionally, the dataset includes 49 harmonized clinical and demographic variables, as well as pre-trained weights for a baseline nnU-Net model trained on the annotated data. This resource addresses a critical gap in publicly available breast cancer datasets, enabling the development, validation, and benchmarking of advanced deep learning models, thus driving progress in breast cancer diagnostics, treatment response prediction, and personalized care.

Background & Summary

Magnetic Resonance Imaging (MRI) is recognized as a highly sensitive imaging modality for breast cancer assessment, particularly in preoperative staging and treatment response evaluation. Breast MRI, specifically T1-weighted dynamic contrast-enhanced imaging (DCE-MRI), uses contrast agents to enhance blood vessels and tissues within the breast, aiding in the localization of tumors, often identified through angiogenesis¹. The precise delineation of the tumor boundary, or tumor segmentation, enables accurate quantitative evaluation of tumor characteristics such as shape, size, and volume, which can help monitor disease progression and treatment effectiveness. In addition to its clinical value, gold-standard segmentations enable a more nuanced analysis of breast cancer characteristics and contribute to the development of AI models for improved diagnosis and prognosis. Radiomics², a method widely employed in machine learning applied to radiology, involves extracting numerous quantitative features from images and is highly dependent on gold-standard segmentations³. Although Radiomics has proven effective in predicting treatment response and survival status in breast cancer research, particularly using breast DCE-MRI images^{4–6}, current studies in public datasets include only a small number of subjects (up to 300) due to the limited availability of expert segmentations⁷. Besides the lack of expert segmentations, open-access DCE-MRI datasets with clinical outcomes are scarce and, currently, all available collections are part of The Cancer Imaging Archive⁸ (TCIA). Within the selected collections, only 163 expert primary tumor segmentations from the I-SPY1/ACRIN 6657 trial⁹ were available in TCIA¹⁰. The TCIA collections lack standardization in terms of folder structure, file naming, and clinical variables. Similarly to the M&Ms benchmark dataset for cardiac imaging^{11,12} and the BraTS dataset for brain imaging¹³, our initiative presents a multicenter breast cancer dataset comprising 1506 pre-treatment DCE-MRI cases with expert segmentations. This dataset is specifically designed to support the benchmarking of advanced medical imaging models leveraging AI.

The main contributions of our work are shown in Fig. 1.

First, we collected pre-treatment T1-weighted DCE-MRI cases from four different collections in TCIA, sourcing a total of 1506 cases. The selection criteria, shown in Fig. 2, included pre-treatment cases where patients

[#]A full list of authors and their affiliations appears at the end of the paper.

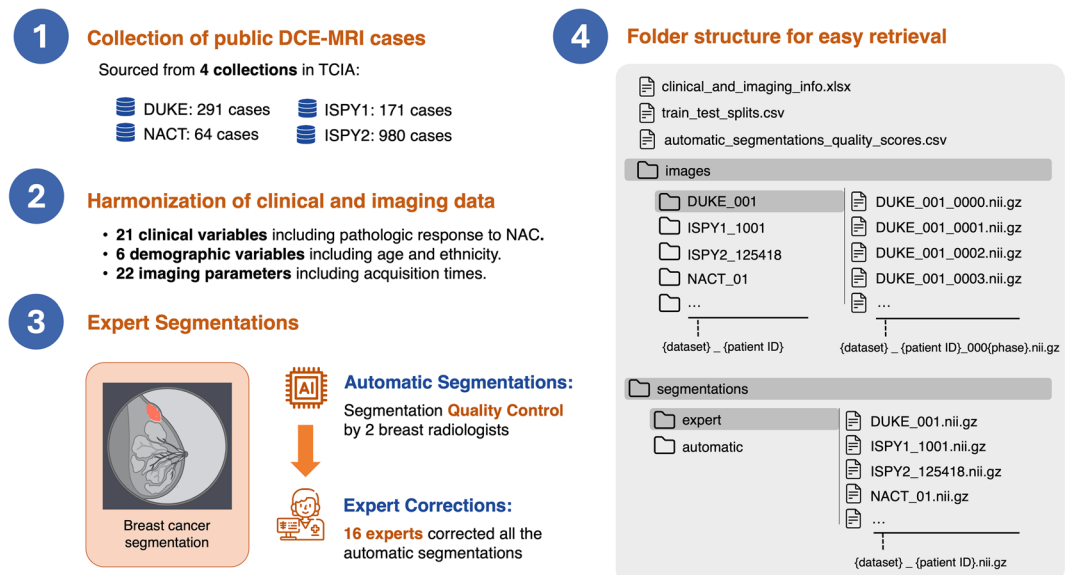


Fig. 1 Summary of the main contributions in the MAMA-MIA dataset. The dataset includes three tables with the harmonized clinical and imaging data, train and test splits for benchmarking, the preliminary automatic segmentations quality scores, the images and the expert and the preliminary automatic segmentations. Each case in the dataset consists of a pre-treatment T1-weighted DCE-MRI sequence with all the phases in a subfolder under the images folder, and two primary tumour segmentations, one expert-corrected and the preliminary automatic segmentation, without expert corrections.

underwent neoadjuvant chemotherapy (NAC) within months of diagnosis and for which corresponding clinical data was available, such as the pathological complete response (pCR) to NAC or five-year survival information.

Second, the clinical and imaging data of the selected cases across four collections were consolidated and harmonized into a single table, which contains 21 clinical variables, 6 demographic variables, and 22 imaging parameters.

Third, a total of 16 experts participated in the segmentation of the primary tumors and non-mass-enhanced areas present in the 1506 T1-weighted DCE-MRI cases. Manual segmentation of breast tumors in 3D MR images is both tedious and time-consuming, making it impractical for large-scale datasets. To address this challenge, we adopted a strategy similar to that used in the BraTS dataset¹³ for brain tumor segmentation, where automatic methods assist in generating preliminary annotations that are then refined by experts. Specifically, we initially trained a standard state-of-the-art deep learning model using private expert segmentations of DCE-MRI. This model produced preliminary segmentations that the 16 experts manually corrected, inspected, and verified, resulting in 1506 expert segmentations. This approach not only reduced annotation time but also ensured consistency across a large dataset, similar to BraTS, where semi-automated tools accelerate initial segmentation, followed by expert refinement. Additionally, two expert clinicians visually assessed all the preliminary automatic segmentations for quality assurance. The quality labels from these preliminary automatic segmentations, along with expert corrections, offer insights for the design of AI-driven models for segmentation quality control.

Forth, the dataset folder structure was standardized for easy retrieval and harmonized to support plug-and-play AI training.

Last, an additional contribution of this work is the pre-trained weights of a baseline nnU-Net¹⁴ tumor segmentation model, trained on the 1506 expert segmentations of the primary tumors segmented in the MAMA-MIA dataset. These weights can be used for inference or to fine-tune models for a wide variety of segmentation tasks involving MRI or other 3D medical image modalities.

We note that our dataset may have potential biases arising from preliminary automatic segmentations and inter-annotator variability among the 16 experts performing manual corrections. Typically, radiologists employ similar software tools (e.g., thresholding) to generate rough approximations of lesions, saving annotation time. Despite these potential biases, our dataset represents the largest collection of expert segmentations in breast cancer MRI to date, paired with harmonized imaging and clinical data. This addresses a significant gap in the availability of gold-standard segmentations in breast MRI, adding substantial value to breast cancer research. It is important to highlight that experts were instructed to segment only the primary lesion in cases of multifocal or multicenter breast cancers. This restriction was due to the clinical information, such as tumor subtype and pathologic complete response (pCR), being available only for the primary lesion and to reduce segmentation time.

In the following paragraphs some of the most important potential applications of the MAMA-MIA dataset are introduced in detail.

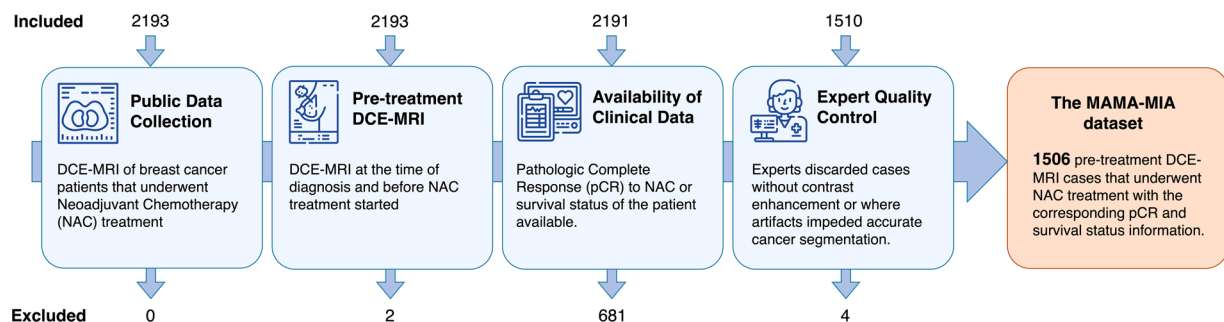


Fig. 2 Selection criteria to build the MAMA-MIA dataset. The DCE-MRI cases are collected from four public collections available on TCIA⁸: *level 2b cohort*²⁴ from I-SPY1/ACRIN 6657 trial (I-SPY1)⁹, I-SPY2/ACRIN 6698 trial^{25,26}, NACT-Pilot²⁷, and Duke-Breast-Cancer-MRI²⁸, referred to as NACT, ISPY1, ISPY2, and DUKE, respectively.

Treatment Response and Survival Prediction. Despite its benefits, NAC has associated side effects, making it desirable to predict patient response before treatment planning. Most deep learning methods predicting pCR to NAC using MRI data have been developed with fewer than 300 samples and are difficult to benchmark due to the lack of a standardized dataset with expert segmentations⁷. By including treatment and survival outcomes alongside other clinical variables, the MAMA-MIA dataset can serve as a benchmark for developing AI models to predict treatment response and patient survival.

Automatic Segmentation of Breast Cancer in MRI. Automated segmentation algorithms can process medical images much faster than manual methods, minimizing inter-observer variability and providing more consistent results. The 1506 expert segmentations in this dataset enable the development of large-scale, generalizable, and robust automatic tumor segmentation models. The pre-trained weights of a baseline nnU-Net segmentation model trained on this dataset are provided to facilitate further improvements and to serve as a baseline for comparative studies.

Segmentation Quality Control. Visual inspection by expert radiologists is the gold standard for quality control, but it is impractical at scale¹⁵. The expert segmentations, combined with evaluations of preliminary automatic segmentations, form the basis for developing robust quality control mechanisms in breast cancer MRI.

Image Synthesis. The synthesis of realistic and diverse 2D MRI slices, as well as full 3D DCE-MRI volumes, can enhance the optimization of image analysis algorithms via data augmentation, domain adaptation, or privacy preservation. This can also support radiologist decision-making, such as simulating treatment response or predicting disease progression^{16–19}. Patient demographics and clinical data in the dataset can be utilized to condition generative models or analyze their impact on generated images, enabling AI fairness analysis and bias mitigation (e.g., age or ethnicity).

Image Standardization. The dataset includes both bilateral and unilateral images, with variations in magnetic field strengths, number of slices, slice thickness, and scanner manufacturers, making it a valuable resource for developing domain generalization and image standardization techniques. Exploring contrast dynamics in tumors and their correlation with acquisition times, as included in the harmonized imaging data, represents a promising avenue for future studies.

Fine-tuning of Foundational Models. Foundation models like MedSAM²⁰, based on SAM²¹, address segmentation tasks across imaging modalities but suffer from imbalances in modality representation in training data and challenges with segmenting vessel-like structures. While SAM has been explored for breast tumor segmentation in ultrasound images²² and mammographic mass segmentation²³, the MAMA-MIA dataset could accelerate the ingestion of 3D medical imaging data for training or fine-tuning foundational models tailored to breast MRI tasks.

Methods

Data Collection and Harmonization. The steps to collect the DCE-MRI cases that form the MAMA-MIA data set are illustrated in Fig. 2. The initial selection criterion was to gather all available open-access DCE-MRI studies of breast cancer patients who underwent NAC treatment. Four collections available on TCIA⁸ met this requirement: the *level 2b cohort*²⁴ from the I-SPY1/ACRIN 6657 trial (I-SPY1)⁹, the I-SPY2/ACRIN 6698 trial^{25,26}, NACT-Pilot²⁷, and Duke-Breast-Cancer-MRI²⁸, referred to as ISPY1, ISPY2, NACT, and DUKE, respectively.

The second criterion was to select the DCE-MRI series acquired before NAC treatment started, often referred to as pre-treatment sequences or timepoint T0. The third criterion was to exclude cases that lacked information on treatment response or survival status. The final criterion involved the quality control by experts during cancer segmentation in the DCE-MRI images. Experts excluded cases without sufficient contrast enhancement or those with artifacts that significantly impeded segmentation.

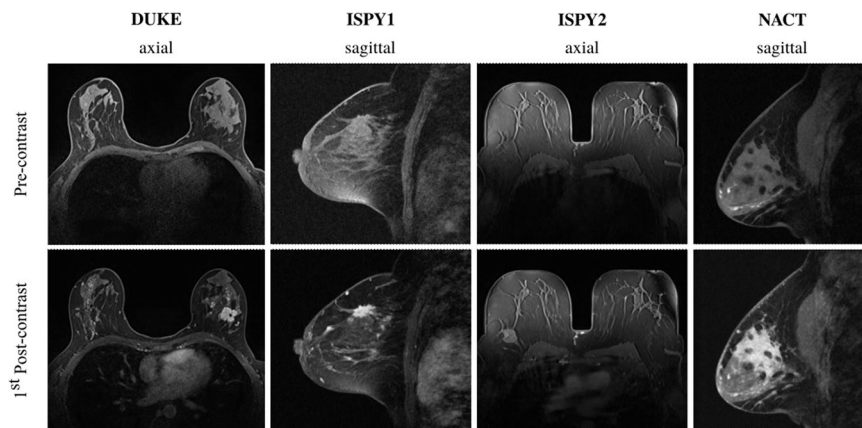


Fig. 3 Pre-treatment DCE-MRI sequences from the four collections forming the dataset. Left to right: images are shown in the acquisition plane (axial or sagittal) from DUKE, ISPY1, ISPY2 and NACT. Only two phases of the DCE-MRI sequence are shown, the pre-contrast phase (first row) and the first post-contrast phase (second row).

The final MAMA-MIA dataset comprises 1506 DCE-MRI cases that meet all the selection criteria. Figure 3 illustrates the pre-contrast and post-contrast phases of one case from each of the four different collections included in the dataset. In the first post-contrast phase, malignant tissues are better visualized due to enhanced contrast after injection. The dataset includes both bilateral and unilateral breast MRIs, as well as images acquired in axial and sagittal planes. This diversity arises from scanner-dependent acquisition protocols at different centers, where some scanners capture MRIs in the axial plane, while others use sagittal protocols. By including this variety, the dataset reflects real-world clinical practices, enhances its generalizability, and supports research into a wide range of imaging scenarios.

The dataset harmonization steps included data curation, image quality control, extraction of clinical and imaging data from DICOM headers, and establishing a standardized naming and folder structure for all sequences in the dataset. In addition, the orientation was standardized to ensure consistency and facilitate usability in computational analysis. Axial MRIs were reoriented to the LAS (left-anterior-superior) coordinate system, while sagittal MRIs were reoriented to the PSR (posterior-superior-right) coordinate system. This decision aligns with common practices in machine learning and deep learning research, where software such as nibabel²⁹ and SimpleITK³⁰ is frequently used to process medical imaging data. This uniformity simplifies downstream tasks such as visualization, annotation, and model development, while preserving anatomical accuracy and clarity for each imaging plane.

No additional preprocessing steps, such as bias field correction, image normalization (e.g., z-score or min-max normalization), or voxel resampling, were applied to the dataset. These methods can significantly alter image resolution and potentially affect model performance. Preprocessing choices depend on the specific requirements of the downstream task³¹. For instance, voxel resampling may influence spatial resolution crucial for tumor segmentation, while normalization strategies might affect intensity-based tasks like radiomics or contrast enhancement studies. Researchers can tailor these procedures to their use cases.

Expert Segmentations. The dataset cohort includes a highly heterogeneous group of locally advanced breast cancers, encompassing cases with single tumors, multiple tumors (multifocal cases), non-mass enhanced areas where the cancer has spread, and bilateral breast cancers. In our dataset, both automatic and expert segmentations were performed within the Volume of Interest (VOI), excluding other cancerous findings outside the VOI. The reason for segmenting only the tissues within the VOI is related to the clinical outcomes and tumor subtype information, which is available only for the primary tumor (delineated VOI) and not for bilateral or multifocal breast cancers.

Selection of the Volume of Interest. The VOI is defined as a 3D rectangular box manually drawn to encompass the entire enhanced region. Its dimensions depend on the tumor morphology, ranging from a few centimeters to the entire breast in cases of more advanced tumors. Accurate selection of the VOI is important for segmentation, as the clinical information regarding tumor subtype and treatment response can only be reliably linked to the volume within the VOI.

In DUKE, the bounding boxes are provided within the clinical information. However, not all datasets include straightforward 3D coordinates of the tumor. For the NACT, ISPY1, and ISPY2 collections, tumor volumetric analysis images are available in most of the DCE-MRI cases. These volumetric analysis images include various annotations of the breast tissue and pixel-level annotations of the peak enhanced region after contrast injection, also known as the Functional Tumor Volume (FTV).

The FTV is determined by filtering the percent enhancement (PE) image and the signal enhancement ratio (SER) image using specific thresholds on their pixel values, following the steps described in⁴. Figure 4 compares the FTV to the expert tumor segmentation. Although the FTV delineates a region likely to encompass malignant

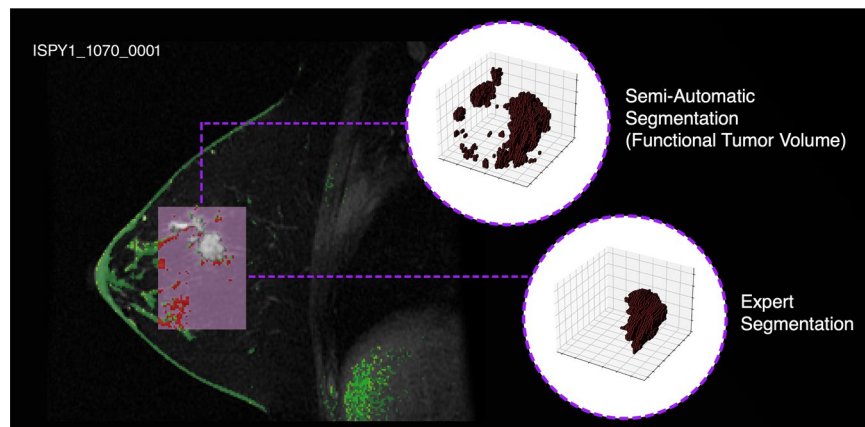


Fig. 4 Volumetric analysis image with the corresponding tumor bounding box (in purple) and the tumor volume extracted using the signal enhancement ratio (SER) method denoted as Functional Tumor Volume (FTV). In comparison to FTV, the expert segmentation of the tumor is more precise and contains only the malignant tissues.

tissue, it frequently fails to accurately represent the precise tumor volume. Consequently, expert refinement of the FTV masks is essential to ensure anatomical accuracy and reliable segmentation for downstream analyses.

In cases where the analysis mask was unavailable, an approximate VOI was created using the same filtering steps applied to the SER and PE images. Using these procedures, we obtained 3D bounding boxes, or VOIs, encapsulating the primary tumor or non-mass-enhanced area for all 1506 cases in the dataset.

Preliminary Automatic Segmentations. In this study, the preliminary automatic tumor segmentations were generated using the popular nnU-Net framework¹⁴. A segmentation model was trained using a total of 331 primary tumor and non-mass enhanced (NME) segmentations from DUKE^{28,32} and the TCGA-BRCA collection³³. The training dataset encompasses 251 axial DCE-MRI cases from DUKE with expert segmentations shared by the authors from a treatment response study^{6,30} and other 80 sagittal DCE-MRI cases with expert-validated automatic tumor segmentations^{34,35} (Chicago Dynamic MRI Explorer 2005 Version) from TCGA-BRCA, increasing the heterogeneity of the training data. This combination of axial and sagittal cases in the training dataset introduced diversity in imaging orientations, which was hypothesized to improve the preliminary segmentation model's robustness for segmenting the MAMA-MIA dataset, as it contains both axial and sagittal cases. However, this hypothesis has not been explicitly tested in this study. At the time this model was developed, the DUKE and TCGA datasets were the only sources of expert tumor segmentations available to the research team.

The 331 expert-validated tumor segmentations were performed on the first post-contrast images. Due to negligible patient movement, these segmentations were applied across all DCE-MRI phases, including pre- and post-contrast images, serving as additional data for training. Preprocessing steps prior to training included cropping the images to the Volume of Interest (VOI) and resampling to $1 \times 1 \times 1 \text{ mm}^3$ isotropic pixel spacing. Data augmentation was performed by adding a 25% pixel margin to the VOI and applying random flipping.

The VOI preliminary automatic segmentations were up-sampled and mapped back to the original image space to generate full-image segmentation masks of the primary lesions. The nnU-Net model achieved a mean validation Dice coefficient of 0.8287 ± 0.0112 in a 5-fold cross-validation setting. As a note, the 80 DCE-MRI cases from TCGA-BRCA collection were not included in the final MAMA-MIA dataset due to the absence of clinical information, such as tumor subtype, treatment response, or survival status.

Visual Quality Control of the Preliminary Automatic Segmentations. Two expert breast radiologists evaluated the quality of the preliminary automatic segmentations using an in-house graphical user interface (GUI). For each case, three 2D slices from the first post-contrast image—spanning the axial, sagittal, and coronal planes—were displayed with segmentation contours highlighted in red. Full image slices were provided to help the experts quickly identify whether the primary tumor or NME region was missed by the segmentation model.

The visual inspection process was guided by specific instructions provided to the expert breast radiologists. Based on the different images displayed, the experts assessed the 1506 preliminary automatic segmentations and categorized them into four quality categories: *Good*, *Acceptable*, *Poor*, or *Missed*. A *Good* segmentation indicated precision with no need for major corrections. An *Acceptable* segmentation captured the tumor but required minor corrections. A *Poor* segmentation lacked precision and contained numerous pixels outside the tumor region. Finally, a segmentation categorized as *Missed* corresponded to an area of the breast unrelated to the tumor.

This categorization was also used to evenly distribute segmentations among the experts for correction, ensuring that the difficulty of cases assigned to each radiologist was balanced. Notably, all cases, including those rated as *Good* during the visual assessment, were sent to the radiologists for manual correction without disclosing the assigned quality category (*Good*, *Acceptable*, *Poor*, or *Missed*). This blinded approach ensured that no bias was

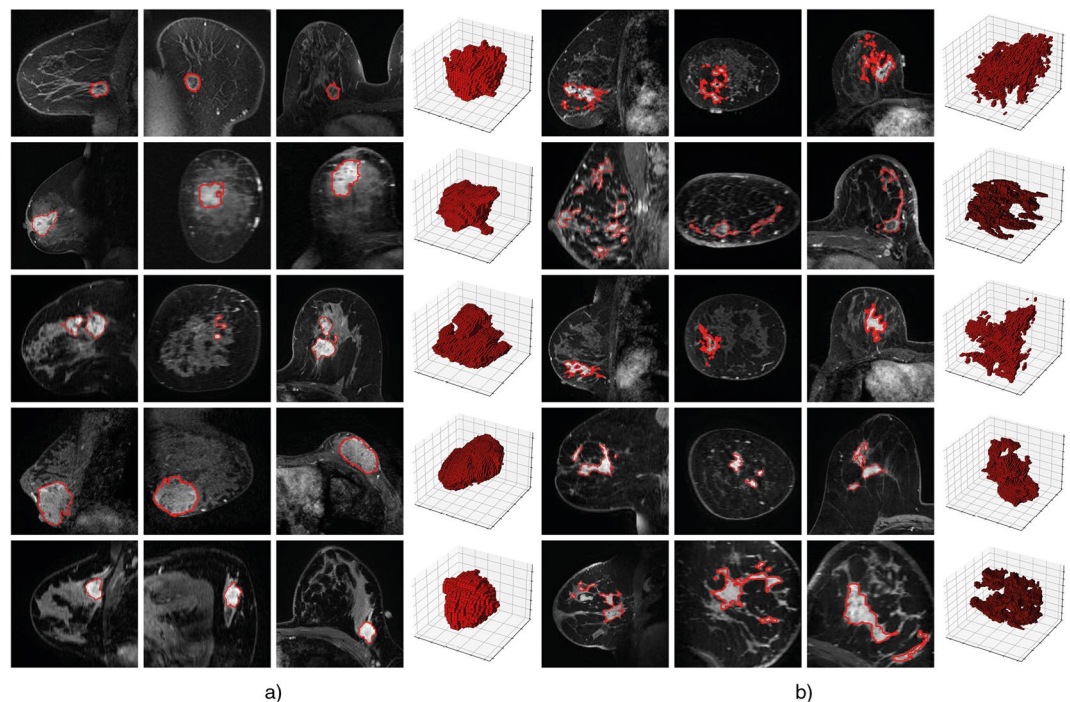


Fig. 5 Manual segmentations were performed for both (a) primary tumors and (b) non-mass enhanced areas. For each case, the middle slice of the manual segmentation is displayed in sagittal, coronal, and axial views, with the segmentation contour highlighted in red. The rightmost columns in (a) and (b) present the corresponding 3D segmentation.

introduced during the correction process, preserving the integrity of the evaluation and ensuring that expert corrections remained consistent across categories.

Expert Manual Corrections. From the 1506 DCE-MRI cases forming the MAMA-MIA dataset only a total of 160 manual segmentations from the ISPY1 collection¹⁰ were available in the TCIA platform. Additionally, the authors from a treatment response study using DUKE dataset^{6,30} shared the expert manual segmentations from an additional 251 cases. Therefore, a total of 411 out of 1506 cases had expert manual segmentations. Our main contribution, together with the dataset harmonization, are the manual segmentations of the missing 1095 cases. A total of 16 experts from nine different institutions from Europe and Africa participated in the manual correction of the 3D segmentations. The group, with an average of 9 years of expertise in breast cancer radiology, was formed by fourteen breast radiologists, one surgical oncologist and one medical physicist. The preliminary automatic segmentation quality scores from one radiologist were used to stratify the cases and assign each expert an evenly distributed set of 70 cases in terms of corrections needed. However, as previously stated, this categorization was blind to the manual annotators. Along with the automatic segmentations, each case consisted of the pre-contrast and the first post-contrast phase. The experts were asked to segment the tumor in the first post-contrast phase but the subtracted image or a later phase could be used as a support in the manual correction process if needed. The Mango viewer³⁶ was the tool selected to correct the automatic segmentations. The guidelines provided to the experts included: 1) segment only the primary tumor if the secondary tumors are not within the FTV volume in multifocal cases, 2) avoid as much as possible the inclusion of healthy tissue in non-mass enhanced cases, 3) exclude tissue markers (or clips) from the segmentations, 4) include tumor necrosis in the segmentation, 5) do not include intra-mammary lymph nodes, 6) verify the segmentation is consistent in all views, not only in the highest resolution view. In Fig. 5 some examples of first post contrast images and the corresponding manually corrected segmentations are shown.

Baseline Segmentation Model using the Expert Segmentations. An additional contribution of this work is the pretrained weights of a vanilla nnU-Net¹⁴ tumor segmentation model, trained using the 1506 expert segmentations. The nnU-Net model was trained with DCE-MRI full-images as input over a total of 1000 epochs in a 5-fold cross-validation setting. The model achieved a mean validation Dice coefficient of 0.7620 ± 0.2113 .

The image preprocessing steps included z-scoring the DCE-MRI images using the mean and standard deviation of all its phases (pre- and post-contrast) and resampling to $1 \times 1 \times 1 \text{ mm}^3$ isotropic pixel spacing. In the training pipeline, all post-contrast phases and the subtraction MRI image (computed by subtracting the pre-contrast image from the first post-contrast image) were included as data augmentation. The model was evaluated only on the first post-contrast phases, which are the images used by the experts to perform the segmentations.

Country		MAMA-MIA Dataset									
		ISPY1		ISPY2		DUKE		NACT		Total	
		United States		United States		United States		United States		United States	
		2002–2006		2010–2016		2000–2014		1995–2002		1995–2016	
Studies time-period		#	(%)	#	(%)	#	(%)	#	(%)	#	(%)
Age	< 40	35	(20.5)	208	(21.2)	61	(21.0)	13	(20.3)	317	(21.0)
	40–49	62	(36.3)	300	(30.6)	104	(35.7)	25	(39.1)	491	(32.6)
	50–59	56	(32.7)	317	(32.3)	74	(25.4)	18	(28.1)	465	(30.9)
	60–69	18	(10.5)	134	(13.7)	41	(14.1)	6	(9.4)	199	(13.2)
	≥70	0	(0.0)	18	(1.8)	11	(3.8)	2	(3.1)	31	(2.1)
	N/A	0	(0.0)	3	(0.3)	0	(0.0)	0	(0.0)	3	(0.2)
Ethnicity	Caucasian	129	(75.4)	777	(79.3)	177	(60.8)	45	(70.3)	1128	(74.9)
	African American	31	(18.1)	116	(11.8)	91	(31.3)	3	(4.7)	241	(16.0)
	Asian	7	(4.1)	68	(6.9)	7	(2.4)	4	(6.2)	86	(5.7)
	Other	2	(1.2)	16	(1.6)	14	(4.8)	3	(4.7)	35	(2.3)
	N/A	2	(1.2)	3	(0.3)	2	(0.7)	9	(14.1)	16	(1.1)
BMI	Underweight	6	(3.5)	17	(1.7)	7	(2.4)	5	(7.8)	35	(2.3)
	Normal	57	(33.3)	301	(30.7)	73	(25.1)	39	(60.9)	470	(31.2)
	Overweight	40	(23.4)	224	(22.9)	71	(24.4)	15	(23.4)	350	(23.2)
	Obesity class I	23	(13.5)	155	(15.8)	50	(17.2)	5	(7.8)	233	(15.5)
	Obesity class II	10	(5.8)	62	(6.3)	19	(6.5)	0	(0.0)	91	(6.0)
	Obesity class III	26	(15.2)	39	(4.0)	13	(4.5)	0	(0.0)	78	(5.2)
	N/A	9	(5.3)	182	(18.6)	58	(19.9)	0	(0.0)	249	(16.5)
Implants	Yes	1	(0.6)	29	(3.0)	0	(0.0)	0	(0.0)	30	(2.0)
	No	170	(99.4)	951	(97.0)	291	(100.0)	64	(100.0)	1476	(98.0)
Bilateral cancer	Yes	3	(1.8)	20	(2.0)	7	(2.4)	0	(0.0)	30	(2.0)
	No	168	(98.2)	960	(98.0)	284	(97.6)	64	(100)	1476	(98.0)
Multifocal cancer	Yes	4	(2.3)	389	(39.7)	139	(47.8)	7	(10.9)	539	(35.8)
	No	7	(4.1)	591	(60.3)	152	(52.2)	57	(89.1)	807	(53.6)
	N/A	160	(93.6)	0	(0.0)	0	(0.0)	0	(0.0)	160	(10.6)
Tumor subtype	Luminal	67	(39.2)	536	(54.7)	123	(42.3)	21	(32.8)	747	(49.6)
	HER2-enriched	25	(14.6)	86	(8.8)	50	(17.2)	8	(12.5)	169	(11.2)
	HER2-pure	29	(17.0)	0	(0.0)	30	(10.3)	6	(9.4)	65	(4.3)
	Triple neg.	45	(26.3)	358	(36.5)	85	(29.2)	11	(17.2)	499	(33.1)
	N/A	5	(2.9)	0	(0.0)	3	(1.0)	18	(28.1)	26	(1.8)
pCR	Yes	49	(28.7)	316	(32.2)	64	(22.0)	11	(17.2)	440	(29.2)
	No	118	(69.0)	664	(67.8)	216	(74.2)	53	(82.8)	1051	(69.8)
	N/A	4	(2.3)	0	(0.0)	11	(3.8)	0	(0.0)	15	(1.0)
Total cases		171		980		291		64		1506	

Table 1. Social (upper half) and clinical (bottom half) variables of the accumulated dataset, MAMA-MIA, consisting of four breast cancer TCIA collections: ISPY1⁹, ISPY2²⁵, DUKE²⁸, and NACT²⁷. Age is measured in years, *Ethnicity* is categorized in Caucasian/White, African American/Black, Asian and Other (Hispanic, American Indian/Alaskan native, Hawaiian/Pacific Islander, Multiple race) groups, while *BMI* (Body Mass Index) is categorized in the indicated groups using the patient weight and height (if patient height was missing, 1.65cm was used as default). *pCR* stands for pathological Complete Response and N/A for *not available*. Last row summarizes the number of cases per dataset and in total.

Data Records

All data records, including the DCE-MRI images, the automatic and expert segmentations for each of the 1506 cases in the MAMA-MIA dataset and the weights of the baseline segmentation model, are available online in the MAMA-MIA Synapse repository³⁷ under CC-BY-NC license at the level of the most restrictive primary dataset (Duke-Breast-Cancer-MRI²⁸). Data records also include three tables, one that contains all the clinical and imaging information, another with the quality scores from the two breast radiologists that evaluated the preliminary automatic segmentations using the GUI, and a table with the train and test split to promote reproducibility in future studies using the dataset. Figure 1 illustrates the file content and folder structure of the MAMA-MIA dataset. Each case identifier consists of the original collection/dataset acronym and the corresponding patient identification number (patient ID). For instance, the *ISPY1_1221* case corresponds to the pre-treatment DCE-MRI sequences of patient ID 1221 from the ISPY-1 collection. The different phases are named using the same case ID plus the corresponding phase number (*ISPY1_1221_000X*). For example, *ISPY1_1221_0000* represents the pre-contrast phase and *ISPY1_1221_0002* represents the second post-contrast phase.

		MAMA-MIA Dataset									
		ISPY1		ISPY2		DUKE		NACT		Total	
		#	(%)	#	(%)	#	(%)	#	(%)	#	(%)
Acquisition plane	Axial	0	(0.0)	980	(100.0)	291	(100.0)	0	(0.0)	1271	(84.4)
	Sagittal	171	(100.0)	0	(0.0)	0	(0.0)	64	(100.0)	235	(15.6)
Magnetic field strength	1.5	171	(100.0)	715	(73.0)	136	(46.7)	64	(100.0)	1086	(72.1)
	3.0	0	(0.0)	265	(27.0)	155	(53.3)	0	(0.0)	420	(27.9)
Fat suppression	Yes	170	(99.4)	976	(99.6)	290	(99.7)	64	(100.0)	1500	(99.6)
	No	1	(0.6)	4	(0.4)	1	(0.3)	0	(0.0)	6	(0.4)
Scanner manufacturer	SIEMENS (S)	44	(25.7)	252	(25.7)	115	(39.5)	0	(0.0)	411	(27.3)
	GE	115	(67.3)	611	(62.3)	176	(60.5)	64	(100.0)	966	(64.1)
	PHILIPS (P)	12	(7.0)	117	(11.9)	0	(0.0)	0	(0.0)	129	(8.6)
Scanner model	Avanto (S)	0	(0.0)	123	(12.6)	70	(24.1)	0	(0.0)	193	(12.8)
	SIGNA HDxt (GE)	0	(0.0)	536	(54.7)	59	(20.3)	0	(0.0)	595	(39.5)
	SIGNA GENESIS (GE)	103	(60.2)	0	(0.0)	0	(0.0)	64	(100)	167	(11.1)
	Other (S, GE, P)	68	(39.8)	321	(32.8)	162	(55.7)	0	(0.0)	551	(36.6)
Bilateral MRI	Yes	3	(1.8)	171	(17.4)	291	(100.0)	0	(0.0)	465	(30.9)
	No	168	(98.2)	809	(82.6)	0	(0.0)	64	(100.0)	1041	(69.1)
Image matrix	[256, 256]	156	(91.2)	33	(3.4)	0	(0.0)	62	(96.9)	251	(16.7)
	[384, 384]	0	(0.0)	149	(15.2)	0	(0.0)	0	(0.0)	149	(9.9)
	[512, 512]	15	(8.8)	721	(73.6)	176	(60.5)	2	(3.1)	914	(60.7)
	Other	0	(0.0)	77	(7.9)	115	(39.5)	0	(0.0)	192	(12.7)
Number of phases	3	143	(83.6)	0	(0)	4	(1.4)	58	(90.6)	205	(13.6)
	4-5	15	(8.8)	9	(0.9)	287	(98.6)	5	(7.8)	316	(21.0)
	6	13	(7.6)	971	(99.1)	0	(0.0)	1	(1.6)	985	(65.4)
	mean [min, max]	3	[3, 6]	6	[4, 6]	4	[3, 5]	3	[3, 6]	5	[3, 6]
Number of slices	< 100	166	(97.1)	593	(60.5)	3	(1.0)	64	(100.0)	826	(54.8)
	100-199	1	(0.6)	369	(37.7)	257	(88.3)	0	(0.0)	627	(41.6)
	>= 200	4	(2.3)	18	(1.8)	31	(10.7)	0	(0.0)	53	(3.5)
	mean [min, max]	64	[44, 256]	106	[52, 256]	169	[60, 256]	60	[46, 64]	111	[44, 256]
Slice thickness	< 2.0	5	(2.9)	183	(18.7)	287	(98.6)	0	(0.0)	475	(31.5)
	2.0-2.9	131	(76.6)	796	(81.2)	4	(1.4)	64	(100.0)	995	(66.1)
	>= 3.0	35	(20.5)	1	(0.1)	0	(0.0)	0	(0.0)	36	(2.4)
	mean [min, max]	2.4	[1.5, 4.0]	2.0	[0.8, 3.0]	1.1	[1.0, 2.5]	2.0	[2.0, 2.4]	1.9	[0.8, 4.0]
Pixel spacing	< 0.5	15	(8.8)	2	(0.2)	0	(0.0)	2	(3.1)	19	(1.3)
	0.5-0.9	150	(87.7)	939	(95.8)	274	(94.2)	62	(96.9)	1425	(94.6)
	>= 1.0	6	(3.5)	39	(4.0)		(5.8)	0	(0.0)		(4.1)
	mean [min, max]	0.8	[0.4, 1.2]	0.7	[0.3, 1.4]	0.7	[0.5, 1.3]	0.7	[0.4, 0.9]	0.7	[0.3, 1.4]
Total cases		171		980		291		64		1506	

Table 2. Image acquisition variables of the accumulated dataset, MAMA-MIA, consisting of four breast cancer TCIA collections: ISPY1⁹, ISPY2²⁵, DUKE²⁸, and NACT²⁷. The upper half shows general acquisition characteristics, while the bottom half shows specifics of the acquired sequences and slices. *Magnetic field strength* is measured in Tesla (T), while *Slice thickness* and *Pixel spacing* in mm. Other *Scanner models* include MAGNETOM Symphony, SymphonyTim, TrioTim, Verio, Skyra, Sonata, Vision, Vision plus, Prisma fit, Espree from Siemens; Signa Excite, Discovery MR750w, Discovery MR750, Signa HDx, Optima from GE; and Gyroscan Intera, Ingenia, Achieva, Intera from Philips. Other *Image matrices* or slice sizes (measured in pixels) include [320, 320], [400, 400], [416, 416], [432, 432], [448, 448], [480, 480], [528, 528], [560, 560], [576, 576], [640, 640], and [1024, 1024]. The *Number of phases* include the pre-contrast and all post-contrast phases. To have a broader overview, the mean value, as well as the minimum and maximum values (in []) are given for a selection of variables. Last row summarizes the number of cases per dataset and in total.

Tables 1 and 2 summarize the most representative dataset demographics, clinical variables, and image acquisition parameters. As described in Table 1, age and ethnicity information is available for more than 98% of the cases, while Body Mass Index (BMI) is available for 83% of cases. The MAMA-MIA dataset comprises 314 cases from women younger than 40 years old, constituting 21% of the total patients, and half of the dataset patients were younger than 50 years old at diagnosis. Therefore, the MAMA-MIA dataset can be considered well-balanced in terms of the young versus older population. Ethnicity distributions in the dataset are reflective of United States demographics³⁸, with 16% African American patients, less than 6% Asian, a 2.3% of other ethnicities, and a majority Caucasian population (74.9%).

	Average Time Intervals									
	pre to 1 st post		1 st to 2 nd post		2 nd to 3 rd post		3 rd to 4 th post		4 th to 5 th post	
ISPY1	390	[27, 915]	284	[20, 531]	324	[24, 899]	142	[121, 162]	142	[121, 162]
ISPY2	145	[77, 761]	92	[59, 206]	92	[58, 206]	92	[58, 217]	91	[58, 206]
DUKE	241	[94, 922]	124	[75, 354]	114	[74, 169]	118	[88, 391]	—	—
NACT	442	[331, 752]	362	[290, 901]	314	[283, 435]	288	[286, 289]	286	[286, 286]
MAMA-MIA	203	[27, 922]	131	[20, 901]	100	[24, 899]	96	[58, 391]	91	[58, 286]

Table 3. Average time intervals, measured in seconds, between the pre-contrast and the subsequent post-contrast phases with the corresponding minimum and maximum values in brackets. Here, details of up to fifth post-contrast phase are shown.

Clinical information available in more than 90% of the cases includes the presence of bilateral cancer at diagnosis, multifocal cancer, tumor subtype, and pathological complete response (pCR) after neoadjuvant chemotherapy (NAC) treatment. Other relevant information included in the dataset, albeit not present in all cases, comprises survival status in over 450 cases, different tumor receptors, days to recurrence or metastasis, agents prescribed during NAC, the necessity of mastectomy after treatment, breast density, and more. A comprehensive list of clinical and imaging variables included in the dataset can be found in the Excel Table as part of the Supplementary Material.

Table 2 presents the most common imaging information included in the dataset: date of original collections, acquisition plane, magnetic field strength (Tesla) used for DCE-MRI acquisition, scanner manufacturers and models, number of bilateral and fat-suppressed DCE-MRIs, mean number of slices, slice thickness, pixel spacing, number of phases, and total number of cases obtained from each original collection. In the DCE-MRI medical imaging modality, the acquisition time interval between contrast administration (pre-contrast MRI) and subsequent post-contrast phases is an important factor. Table 3 summarizes the average time intervals between phases per dataset. It should be noted that older datasets like ISPY1 and NACT have longer acquisition intervals than later collections.

Technical Validation

This technical validation evaluates the visual quality control of preliminary automatic segmentations, emphasizing the role of expert-corrected segmentations as the primary contribution of this study. The goal is to provide actionable guidelines and recommendations for future segmentation quality assessments by analyzing inter-rater agreement between two expert radiologists and correlating their evaluations with established quality metrics, such as the Dice Similarity Coefficient (DSC) and Hausdorff Distance (HD).

Representative examples of expert corrections are shown in Fig. 6, categorized by expert-assigned quality scores of the preliminary automatic segmentations. It is important to note that all cases, including those rated as *Good* during the visual assessment, were sent to the radiologists for manual correction without revealing the quality category (*Good*, *Acceptable*, *Poor*, or *Missed*). This blinded approach ensured that no bias was introduced during the correction process. Green voxels highlight areas added by the experts, while pink voxels indicate areas removed. Minimal corrections were necessary for segmentations rated as *Good*, which typically involved refining tumor boundaries. In contrast, cases categorized as *Missed*, such as *ISPY2_566011*, required the experts to perform full tumor segmentation from scratch. These examples underscore the critical role of robust automatic segmentation methods in minimizing the burden of manual corrections and highlight the value of expert-corrected segmentations.

Figure 7 illustrates the distribution of quality scores assigned by the experts, alongside the DSC and 95 percentile HD values calculated between expert-corrected and preliminary automatic segmentations. Segmentations rated as *Good* were strongly correlated with high DSC values and low HD, confirming the reliability of the visual quality control interface in distinguishing segmentations that meet clinical standards. However, the greatest discrepancies between experts occurred in the *Acceptable* and *Poor* categories, where subjective interpretation played a larger role in the quality assessment.

To quantify inter-rater agreement, we computed Cohen's Kappa under two categorization schemes: the original four categories (*Good*, *Acceptable*, *Poor*, and *Missed*) and a simplified binary scheme (*Good* versus *Corrections Needed*). The four-category scheme yielded a Cohen's Kappa of 0.39, reflecting low to moderate agreement and indicating inconsistencies, particularly in intermediate categories. Simplifying to a binary scheme improved agreement significantly, with a Cohen's Kappa of 0.53, demonstrating that a simplified approach improves consistency between experts while maintaining the primary goal of identifying segmentations that require manual correction.

These findings emphasize the efficiency of adopting a binary categorization scheme in visual quality control workflows. By focusing only on segmentations requiring correction (*Corrections Needed*) versus those considered satisfactory (*Good*), future visual assessments can achieve greater inter-rater agreement, reduce annotation time, and optimize the use of expert resources. The results of this analysis aim to contribute to the development of robust visual quality control guidelines for assessing automatic segmentation models.

Furthermore, the categorization of the preliminary automatic segmentations by the radiologists provides a valuable dataset that can be leveraged to train a classifier capable of mimicking radiologists in assessing segmentation quality. Such a classifier could automatically identify *Good* segmentations or flag

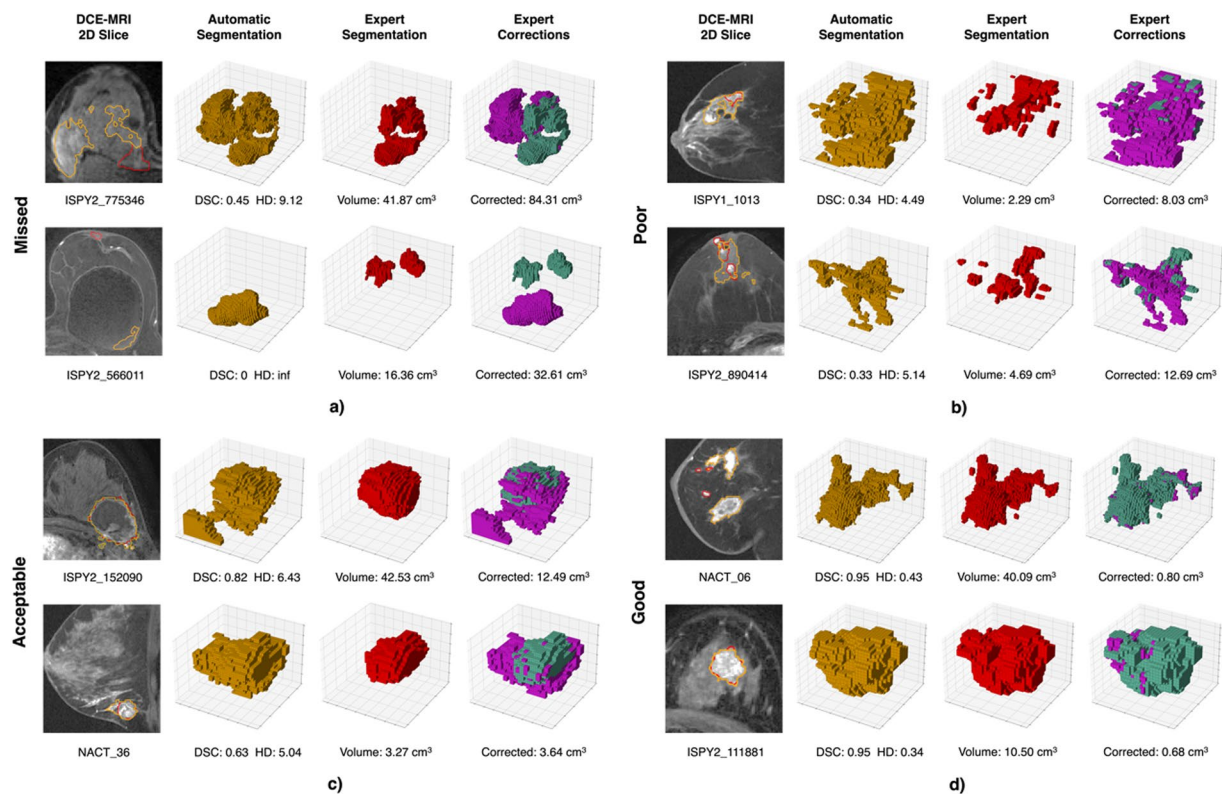


Fig. 6 Examples of preliminary automatic (yellow) and expert segmentations (red) and the corresponding expert corrections (pink if the voxels were removed, and green if they were added). The examples are grouped by the expert-assigned quality scores for automatic segmentations during the visual inspection (**a**) Missed, (**b**) Poor, (**c**) Acceptable, and (**d**) Good. Under the automatic segmentation there is the Dice Similarity Coefficient (DSC) and the 95 percentile Hausdorff Distance (HD) between automatic and expert segmentations. Also, under the manual segmentation, the final volume of the expert segmentation in cm^3 . The total volume corrected by the experts is quantified in cm^3 .

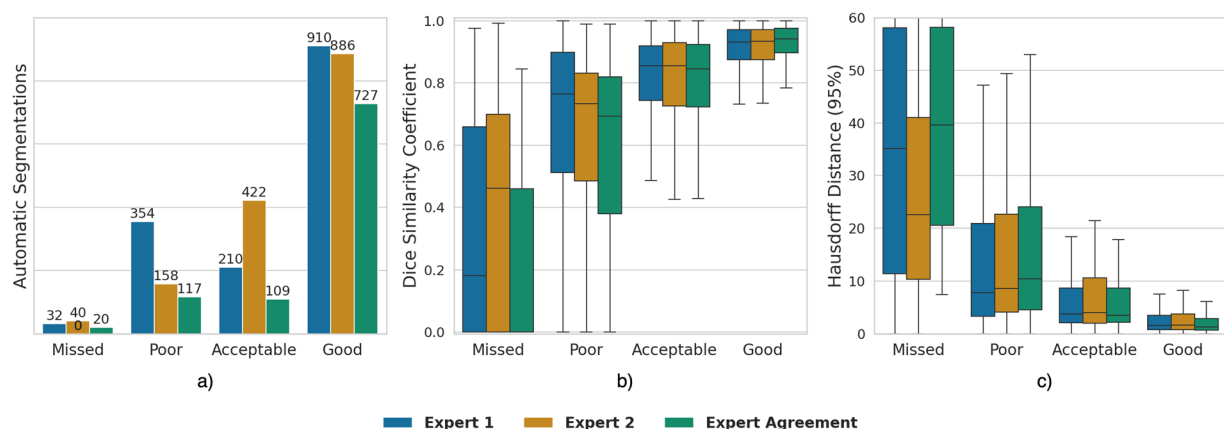


Fig. 7 From left to right: (**a**) distribution of expert-assigned quality scores for preliminary automatic segmentations, (**b**) Dice Similarity Coefficient (DSC) and (**c**) the 95 percentile Hausdorff Distance (HD) between automatic and expert segmentations across different quality categories. Panel (**b**) displays DSC (higher values indicate better agreement), while Panel (**c**) shows HD (lower values indicate better alignment), respectively.

those requiring corrections, further reducing the reliance on manual evaluations and improving scalability. A comprehensive list of expert quality scores and corresponding distance metrics is included in the Data Records as a CSV file, enabling further analysis, benchmarking efforts, and the potential training of automated quality assessment tools.

Code availability

The GitHub repository for the MAMA-MIA dataset is available at: <https://github.com/LidiaGarrucho/MAMA-MIA>. This repository provides detailed resources and instructions to support researchers in integrating the MAMA-MIA dataset into their studies. It includes comprehensive scripts to facilitate working with the dataset and enable customization of workflows for segmentation, detection, and analysis in various radiomic and AI-driven studies. The key components of the repository are as follows:

• **Data Filtering and Visualization Scripts:** Jupyter notebooks are provided to guide users in filtering and visualizing clinical and imaging data using the `pandas` library, along with the clinical and imaging variables included in the dataset's supplementary Excel table. The notebooks include:

- Filtering patients based on clinical outcomes, such as tumor subtype, treatment response, or survival status.
- Sorting and grouping cases by scanner manufacturer, image resolution, or pixel spacing.
- Exploring relevant clinical and imaging metadata from the dataset information table.
- Displaying filtered datasets in tabular or graphical formats for further analysis.
- Plotting MRI images using `matplotlib` for quick visual inspection.
- Displaying VOIs and segmentation contours, along with 3D visualizations of tumor segmentations.

• **Preprocessing Script:** A Jupyter notebook describing common preprocessing steps has been designed to ensure the dataset's compatibility with a wide range of analysis pipelines, while offering flexibility for researchers to adapt workflows to their specific objectives. For a detailed guide, refer to the [notebooks directory](#). The script includes methods to perform:

- **Bias Field Correction:** Corrects for intensity inhomogeneities and biases, which can enhance the accuracy of subsequent analyses.
- **Denoising:** A non-linear filter that averages similar patches throughout the image, allowing it to selectively reduce noise without sacrificing edges and subtle intensity variations.
- **Intensity Clipping:** clipping intensity values to a defined percentile range can help reduce outliers.
- **Normalization (Z-score):** Pixel intensity normalization is performed by subtracting the mean and dividing by the standard deviation, resulting in a distribution with zero mean and unit standard deviation.
- **Histogram Equalization:** Enhances the contrast of the images by redistributing the intensity values to span the full range of the histogram.
- **Resampling:** e.g. Resampling the images to isotropic voxel size (e.g., $1 \times 1 \times 1 \text{ mm}^3$), ensuring uniform resolution across all images.

• **Bounding Box Extraction:** Code to extract 3D bounding box coordinates containing primary lesions to ease the detection and isolation of VOIs for tumor analysis.

• **nnU-Net Segmentation Model Training:** Detailed steps for training an nnU-Net model using the dataset's segmentations. The automatic segmentation models were trained using the `nnU-Net` framework in a 5-fold cross-validation setting. Additionally, the pretrained weights of the segmentation model, trained on all expert segmentations, are shared and hosted in the Synapse repository. This facilitates model development and enables easier benchmarking and comparison by other researchers.

The original DICOM images from TCIA were transformed to NIfTI format using the `pycad` Python library. Metrics such as the Dice coefficient and 95% Hausdorff Distance were computed using the `seg-metrics 1.2.7` Python library³⁹.

Received: 20 June 2024; Accepted: 26 February 2025;

Published online: 19 March 2025

References

1. Mann, R. M., Cho, N. & Moy, L. Breast MRI: state of the art. *Radiology* **292**, 520–536, <https://doi.org/10.1148/radiol.2019182947> (2019).
2. Lambin, P. *et al.* Radiomics: extracting more information from medical images using advanced feature analysis. *Eur. J. Cancer* **48**, 441–446, <https://doi.org/10.1016/j.ejca.2011.11.036> (2012).
3. Poirot, M. G. *et al.* Robustness of radiomics to variations in segmentation methods in multimodal brain MRI. *Sci. Reports* **12**, 16712, <https://doi.org/10.1038/s41598-022-20703-9> (2022).
4. Hylton, N. M. *et al.* Neoadjuvant chemotherapy for breast cancer: functional tumor volume by MR imaging predicts recurrence-free survival-results from the ACRIN 6657/CALGB 150007 I-SPY 1 trial. *Radiology* **279**, 44–55, <https://doi.org/10.1148/radiol.2015150013> (2016).
5. O'Donnell, J. *et al.* The accuracy of breast MRI radiomic methodologies in predicting pathological complete response to neoadjuvant chemotherapy: A systematic review and network meta-analysis. *Eur. J. Radiology* **157**, 110561, <https://doi.org/10.1016/j.ejrad.2022.110561> (2022).
6. Caballo, M. *et al.* Four-dimensional machine learning radiomics for the pretreatment assessment of breast cancer pathologic complete response to neoadjuvant chemotherapy in dynamic contrast-enhanced MRI. *J. Magn. Reson. Imaging* **57**, 97–110, <https://doi.org/10.1002/jmri.28273> (2023).
7. Khan, N., Adam, R., Huang, P., Maldjian, T. & Duong, T. Q. Deep learning prediction of pathologic complete response in breast cancer using MRI and other clinical data: a systematic review. *Tomography* **8**, 2784–2795, <https://doi.org/10.3390/tomography8060232> (2022).

8. Clark, K. *et al.* The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J. Digital Imaging* **26**, 1045–1057, <https://doi.org/10.1007/s10278-013-9622-7> (2013).
9. Newitt, D., Hylton, N. *et al.* multicenter breast DCE-MRI data and segmentations from patients in the I-SPY 1/ACRIN 6657 trials. *The Cancer Imaging Arch.* <https://doi.org/10.7937/K9/TCIA.2016.HdHpgJLK> (2016).
10. Chitalia, R. *et al.* Expert tumor annotations and radiomics for locally advanced breast cancer in DCE-MRI for ACRIN 6657/I-SPY1. *Sci. Data* **9**, 440, <https://doi.org/10.1038/s41597-022-01555-4> (2022).
11. Campello, V. M. *et al.* Multi-centre, multi-vendor and multi-disease cardiac segmentation: the M&Ms challenge. *IEEE Transactions on Med. Imaging* **40**, 3543–3554, <https://doi.org/10.1109/TMI.2021.3090082> (2021).
12. Martin-Isla, C. *et al.* Deep learning segmentation of the right ventricle in cardiac MRI: The M&Ms challenge. *IEEE J. Biomed. Heal. Informatics* <https://doi.org/10.1109/JBHI.2023.3267857> (2023).
13. Bakas, S. *et al.* Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci. Data* **4**, 1–13, <https://doi.org/10.1038/sdata.2017.117> (2017).
14. Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J. & Maier-Hein, K. H. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **18**, 203–211, <https://doi.org/10.1038/s41592-020-01008-z> (2021).
15. Robinson, R. *et al.* Automated quality control in image segmentation: application to the uk biobank cardiovascular magnetic resonance imaging study. *J. Cardiovasc. Magn. Reson.* **21**, 1–14, <https://doi.org/10.1186/s12968-019-0523-x> (2019).
16. Osuala, R. *et al.* Data synthesis and adversarial networks: A review and meta-analysis in cancer imaging. *Med. Image Analysis* **102704**, <https://doi.org/10.1016/j.media.2022.102704> (2022).
17. Osuala, R. *et al.* Pre- to post-contrast breast MRI synthesis for enhanced tumour segmentation. In Colliot, O. & Mitra, J. (eds.) *Medical Imaging 2024: Image Processing*, vol. 12926, 129260Y, <https://doi.org/10.1117/12.3006961>. International Society for Optics and Photonics (SPIE, 2024).
18. Osuala, R. *et al.* Towards learning contrast kinetics with multi-condition latent diffusion models. In *Medical Image Computing and Computer Assisted Intervention 2024*, 713–723, https://doi.org/10.1007/978-3-031-72086-4_67 (Springer Nature Switzerland, 2024).
19. Müller-Franzes, G. *et al.* Using machine learning to reduce the need for contrast agents in breast MRI through synthetic images. *Radiology* **307**, e222211, <https://doi.org/10.1148/radiol.222211> (2023).
20. Ma, J. *et al.* Segment anything in medical images. *Nat. Commun.* **15**, 654, <https://doi.org/10.1038/s41467-024-44824-z> (2024).
21. Kirillov, A. *et al.* Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026 (2023).
22. Hu, M., Li, Y. & Yang, X. BreastSAM: adapting the segmentation anything model for breast tumor segmentation in ultrasound imaging. *Medical Imaging 2024: Ultrasonic Imaging and Tomography*. Vol. 12932. SPIE, <https://doi.org/10.1117/12.3006773> (2024).
23. Xiong, X., Wang, C., Li, W. & Li, G. Mammo-SAM: Adapting foundation segment anything model for automatic breast mass segmentation in whole mammograms. In *International Workshop on Machine Learning in Medical Imaging*, 176–185, https://doi.org/10.1007/978-3-031-45673-2_18 (Springer, 2023).
24. Hylton, N. M. *et al.* Locally advanced breast cancer: MR imaging for prediction of response to neoadjuvant chemotherapy-results from ACRIN 6657/I-SPY trial. *Radiology* **263**, 663–672, <https://doi.org/10.1148/radiol.12110748> (2012).
25. Li, W. *et al.* I-SPY 2 breast dynamic contrast enhanced MRI trial (version 1) [Data set]. *The Cancer Imaging Arch.* <https://doi.org/10.7937/TCIA.D8Z0-9T85> (2022).
26. Newitt, D. C. *et al.* ACRIN 6698/I-SPY2 breast DWI [Data set]. *The Cancer Imaging Arch.* <https://doi.org/10.7937/TCIA.KK02-6D95> (2021).
27. Newitt, D. & Hylton, N. Single site breast DCE-MRI data and segmentations from patients undergoing neoadjuvant chemotherapy (version 3) [Data set]. *The Cancer Imaging Arch.* <https://doi.org/10.7937/K9/TCIA.2016.QHsyhJKy> (2016).
28. Saha, A. *et al.* Dynamic contrast-enhanced magnetic resonance images of breast cancer patients with tumor locations [Data set]. *The Cancer Imaging Arch.* <https://doi.org/10.7937/TCIA.e3sv-re93> (2021).
29. Brett, M. *et al.* nipy/nibabel: 5.3.1. Zenodo, <https://doi.org/10.5281/zenodo.13936989> (2024).
30. Yaniv, Z. *et al.* SimpleITK image-analysis notebooks: a collaborative environment for education and reproducible research. *Journal of Digital Imaging* **31**, 290–303, <https://doi.org/10.1007/s10278-017-0037-8> (2018).
31. Schwarzhans, F. *et al.* Intensity Normalization Techniques and Their Effect on the Robustness and Predictive Power of Breast MRI Radiomics. Preprint at <https://doi.org/10.48550/arXiv.2406.01736> (2024).
32. Saha, A. *et al.* A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 DCE-MRI features. *Br. J. Cancer* **119**, 508–516, <https://doi.org/10.1038/s41416-018-0185-8> (2018).
33. Lingle, W. *et al.* The Cancer Genome Atlas Breast Invasive Carcinoma Collection (TCGA-BRCA), *The Cancer Imaging Archive*, Version 3, <https://doi.org/10.7937/K9/TCIA.2016.AB2NAZRP> (2016).
34. Morris, E. *et al.* Using computer-extracted image phenotypes from tumors on breast MRI to predict stage [Data set]. *The Cancer Imaging Arch.* <https://doi.org/10.7937/K9/TCIA.2014.8SIPIY6G> (2014).
35. Li, H. *et al.* Quantitative MRI radiomics in the prediction of molecular classifications of breast cancer subtypes in the TCGA/TCIA Data set. *npj breast cancer* **2**, 16012, <https://doi.org/10.1038/npjbcancer.2016.12> (2016).
36. Lancaster, J. L. & Martinez, M. J. Mango Viewer. <http://rii.uthscsa.edu/mango/> Accessed: 2024-05-27 (2019).
37. Garrucho, L. *et al.* MAMA-MIA dataset. <https://doi.org/10.7303/syn60868042> Licensed under a Creative Commons CC-BY-NC License (2024).
38. Bureau, U. C. Quickfacts. *Houston Texas/Harris County: United States Census* (2016).
39. Jia, J. Jingnan-jia/segmentation_metrics: V1.2.7, <https://doi.org/10.5281/zenodo.12094185> (2024).

Acknowledgements

The authors would like to express gratitude to all the participants who contributed to this study. Special appreciation is extended to the experts who performed manual segmentations for 1095 cases and the two experts who conducted quality control on the preliminary automatic segmentations. Additionally, our thanks go to Ritse Mann and Marco Caballo for facilitating the 251 manual segmentations from DUKE, as derived from their study⁶. We also acknowledge The Cancer Imaging Archive team for making the imaging and clinical data used in this study publicly available, as well as the additional 160 ISPY1 tumor segmentation data from the University of Chicago lab of Maryellen Giger, whose members participated in the TCGA Breast Phenotype Research Group. The preliminary automatic segmentations in this study were trained using data generated by the TCGA Research Network (<http://cancergenome.nih.gov/>). This project has received funding from the European Union's Horizon 2020 research and innovation programmes under grant agreement No 952103 (EUCanImage) and No 101057699 (RadioVal). Also, this work was partially supported by the project FUTURE-ES (PID2021-126724OB-I00) from the Ministry of Science and Innovation of Spain. The co-author K.K. holds the Juan de la Cierva fellowship with a reference number FJC2021-047659-I.

Author contributions

L.G. designed and led the study. C.R., K.K. and L.G. collected the dataset, standardized the images, curated the clinical and scanner information, verified all the manual segmentations, and contributed to manuscript writing. S.J. and R.O. assisted in curating the manual segmentations from Caballo *et al.*'s study. S.J. assisted in training the automatic segmentation models and contributed to writing the corresponding section in the paper. R.O. led the section on image synthesis. K.K. designed the evaluation tool used to assess the automatic segmentations. A.T. assisted the experts during the manual segmentation process and built the dataset tables in the paper. J. R. and A.C. provided the expert quality scores of the automatic segmentations in this study. R.M. manually annotated 251 from DUKE dataset. F.P. helped with the data managing in TCIA. M.C., P.M.A., S.W.T., S.S.S., N.O.S., A.M.A., A.K., E.D., G.I., K.N., M.E.K., R.G., M.G., O.L. K.G., M.B. were the sixteen experts that performed the manual segmentations of the missing 1095 cases. K.K., A. T., C. M., M.P.S., K.M., F.S., O.D., L.I., and K.L. supervised the study. K.L. is the PI of the project that funded this research. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41597-025-04707-4>.

Correspondence and requests for materials should be addressed to L.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025

Lidia Garrucho^{1,2}✉, Kaisar Kushibar¹, Claire-Anne Reidel¹, Smriti Joshi¹, Richard Osuala^{1,3,4}, Apostolia Tsirikoglou², Maciej Bobowicz⁵, Javier del Riego⁶, Alessandro Catanese⁷, Katarzyna Gwoździwicz⁵, Maria-Laura Cosaka⁸, Pasant M. Abo-Elhoda⁹, Sara W. Tantawy⁹, Shorouq S. Sakrana⁹, Norhan O. Shawky-Abdelfatah⁹, Amr Muhammad Abdo Salem⁹, Androniki Kozana¹⁰, Eugen Divjak^{11,12}, Gordana Ivanac^{11,12}, Katerina Nikiforaki¹³, Michail E. Klontzas¹⁴, Rosa García-Dosdá¹⁵, Meltem Gulsun-Akpınar¹⁶, Oğuz Lafci¹⁷, Ritse Mann¹⁸, Carlos Martín-Isla¹, Fred Prior¹⁹, Kostas Marias^{20,21}, Martijn P. A. Starmans^{22,23}, Fredrik Strand^{2,24}, Oliver Díaz¹, Laura Igual¹ & Karim Lekadir^{1,25}

¹Barcelona Artificial Intelligence in Medicine Lab (BCN-AIM), Facultat de Matemàtiques i Informàtica, Universitat de Barcelona, Gran Via de les Corts Catalanes 585, 08007, Barcelona, Spain. ²Department of Oncology-Pathology, Karolinska Institutet, Stockholm, Sweden. ³Institute of Machine Learning in Biomedical Imaging, Helmholtz Center Munich, Munich, Germany. ⁴School of Computation, Information and Technology, Technical University of Munich, Munich, Germany. ⁵2nd Dept. of Radiology, Medical University of Gdansk, Gdansk, Poland. ⁶Àrea de Radiologia Mamària y Ginecològica (UDIAT CD), Parc Taulí Hospital Universitari, Sabadell, Spain. ⁷Unitat de Diagnòstic per la Imatge de la Mama (UDIM), Hospital Germans Trias i Pujol, Badalona, Spain. ⁸Centro Mamario Instituto Alexander Fleming, Buenos Aires, Argentina. ⁹Department of Diagnostic & Interventional Radiology and Molecular Imaging, Faculty of Medicine, Ain Shams University, Cairo, Egypt. ¹⁰Department of Radiology, University Hospital of Heraklion, Stavrakia, Greece. ¹¹Department of Diagnostic and Interventional Radiology, University Hospital Dubrava, Zagreb, Croatia. ¹²University of Zagreb, School of Medicine, Zagreb, Croatia. ¹³Computational BioMedicine Laboratory, Institute of Computer Science, Foundation for Research and Technology-Hellas, Heraklion, Greece. ¹⁴Department of Radiology, School of Medicine, University of Crete, Heraklion, Greece. ¹⁵Medical Imaging and Radiology, University and Polytechnic Hospital La Fe, Valencia, Spain. ¹⁶Department of Radiology, Hacettepe University Faculty of Medicine Sıhhiye, Ankara, Turkey. ¹⁷Department of Biomedical Imaging and Image-guided Therapy, Medical University of

Vienna, Vienna, Austria. ¹⁸Department of Radiology and Nuclear Medicine, Radboud University Medical Center, Nijmegen, The Netherlands. ¹⁹University of Arkansas for Medical Sciences, Little Rock, AR, US. ²⁰Department of Electrical and Computer Engineering, Hellenic Mediterranean University, Heraklion, Greece. ²¹Computational BioMedicine Laboratory, Institute of Computer Science, Foundation for Research and Technology - Hellas (FORTH), Heraklion, Greece. ²²Department of Radiology and Nuclear Medicine, Erasmus MC Cancer Institute, University Medical Center Rotterdam, Rotterdam, The Netherlands. ²³Department of Pathology, Erasmus MC Cancer Institute, University Medical Center Rotterdam, Rotterdam, The Netherlands. ²⁴Breast Radiology, Karolinska University Hospital, Stockholm, Sweden. ²⁵Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys 23, Barcelona, Spain. ✉e-mail: lgarrucho@ub.edu