POLITECNICO DI TORINO
Repository ISTITUZIONALE

The MAMA-MIA Challenge: Advancing Generalizability and Fairness in Breast MRI Tumor Segmentation and Treatment Response Prediction

Original
The MAMA-MIA Challenge: Advancing Generalizability and Fairness in Breast MRI Tumor Segmentation and Treatment Response Prediction / Garrucho, L., Joshi, S., Kushibar, K., Osuala, R., Bobowicz, M., Bargalló, X., Jaruševiius, P., Geissler, K., Schäfer, R., Alberb, M., Xu, T., Martel, A., Sleiman, D., Awasthi, N., Awwad, H., Vilanova, J.C., Martí, R., Schouten, D., Hoon Lee, J., Rusu, M., et al.. - In: IEEE TRANSACTIONS ON MEDICAL IMAGING. - ISSN 0278-0062. - (2026). [10.1109/TMI.2026.3733343]

Availability:
This version is available at: 11583/3015449 since: 2026-09-30T14:46:16Z

Publisher:
IEEE

Published
DOI:10.1109/TMI.2026.3733343

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

25 November 2026

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2026.3733343

The MAMA-MIA Challenge: Advancing Generalizability and Fairness in Breast MRI Tumor Segmentation and Treatment Response Prediction

Lidia Garrucho*,†, Smriti Joshi *, Kaisar Kushibar, Richard Osuala, Maciej Bobowicz, Xavier Bargalló, Paulius Jaruševičius, Kai Geissler, Raphael Schäfer, Muhammad Alberb, Tony Xu, Anne Martel, Daniel Sleiman, Navchetan Awasthi, Hadeel Awwad, Joan C. Vilanova, Robert Martí, Daan Schouten, Jeong Hoon Lee, Mirabela Rusu, Eleonora Poeta, Luisa Vargas, Eliana Pastor, Maria A. Zuluaga, Jessica Kächele, Dimitrios Bounias, Alexandra Ertl, Katarzyna Gwoździewicz, Maria-Laura Cosaka, Pasant M. Abo-Elhoda, Sara W. Tantawy, Shorouq S. Sakrana, Norhan O. Shawky-Abdelfatah, Amr Muhammad Abdo-Salem, Androniki Kozana, Eugen Divjak, Gordana Ivanac, Katerina Nikiforaki, Michail E. Klontzas, Rosa García-Dosdá, Meltem Gulsun-Akpinar, Oğuz Lafcı, Carlos Martín-Isla, Oliver Díaz, Laura Igual, and Karim Lekadir

***Abstract*— Breast cancer is the most frequently diagnosed malignancy among women worldwide and a leading cause of cancer-related mortality. Dynamic contrast-enhanced magnetic resonance imaging plays a central role in tumor characterization and treatment monitoring, particularly in patients receiving neoadjuvant chemotherapy. However, existing artificial intelligence models for breast magnetic resonance imaging are typically developed and evaluated using heterogeneous datasets, study populations, and assessment protocols, making direct comparison difficult and limiting understanding of model robustness across institutions and clinically relevant patient subgroups. The MAMA-MIA Challenge was designed to address these challenges by providing a standardized benchmark for the joint evaluation of primary tumor segmentation and prediction of pathologic complete response using pre-treatment magnetic resonance imaging only. The training**

*L. Garrucho and S. Joshi contributed equally to this work.

L. Garrucho, S. Joshi, K. Kushibar, R. Osuala, C. Martín-Isla, O. Díaz, L. Igual, and K. Lekadir are with Barcelona Artificial Intelligence in Medicine Lab (BCN-AIM), Facultat de Matemàtiques i Informàtica, Universitat de Barcelona, Barcelona, Spain.

M. Bobowicz and K. Gwoździewicz are with 2nd Department of Radiology, Medical University of Gdansk, Gdansk, Poland.

X. Bargalló is with Department of Radiology, Hospital Clínic of Barcelona, Barcelona, Spain.

P. Jaruševičius is with Department of Radiology, Lithuanian University of Health Sciences, Kaunas, Lithuania.

K. Geissler and R. Schäfer are with Fraunhofer Institute for Digital Medicine MEVIS, Germany.

M. Alberb and T. Xu are with Department of Medical Biophysics, University of Toronto, Canada.

A. Martel is with Sunnybrook Research Institute, Toronto, Canada.

D. Sleiman is with University of Amsterdam, Amsterdam, The Netherlands.

N. Awasthi is with University of Amsterdam and Amsterdam UMC, Amsterdam, The Netherlands, and with Indian Institute of Technology, Jodhpur, Rajasthan, India.

J. C. Vilanova is with Department of Radiology, Clínica Girona, Institute of Diagnostic Imaging (IDI), Girona, Spain.

H. Awwad and R. Martí are with Computer Vision and Robotics Institute (ViCOROB), University of Girona, Girona, Spain.

D. Schouten, J. H. Lee, and M. Rusu are with Stanford University, Stanford, CA, USA.

E. Poeta and E. Pastor are with Politecnico di Torino, Turin, Italy.

L. Vargas and M. A. Zuluaga are with EURECOM, France.

J. Kächele, D. Bounias, and A. Ertl are with German Cancer Research Center (DKFZ), Division of Medical Image Computing, and the Medical Faculty Heidelberg, Heidelberg University, Germany.

J. Kächele is also with German Cancer Consortium (DKTK), DKFZ, Core Center Heidelberg, Heidelberg, Germany.

M.-L. Cosaka is with Centro Mamario Instituto Alexander Fleming, Buenos Aires, Argentina.

P. M. Abo-Elhoda, S. W. Tantawy, S. S. Sakrana, N. O. Shawky-Abdelfatah, and A. M. Abdo-Salem are with Department of Diagnostic and Interventional Radiology and Molecular Imaging, Faculty of Medicine, Ain Shams University, Cairo, Egypt.

A. Kozana is with Department of Radiology, University Hospital of Heraklion, Heraklion, Greece.

E. Divjak and G. Ivanac are with Department of Diagnostic and Interventional Radiology, University Hospital Dubrava, and the University of Zagreb School of Medicine, Zagreb, Croatia.

K. Nikiforaki is with Computational BioMedicine Laboratory, Institute of Computer Science, Foundation for Research and Technology–Hellas, Heraklion, Greece.

M. E. Klontzas is with Department of Radiology, School of Medicine, University of Crete, Heraklion, Greece.

R. García-Dosdá is with Medical Imaging and Radiology, Universitary and Politechnic Hospital La Fe, Valencia, Spain.

M. Gulsun-Akpinar is with Department of Radiology, Hacettepe University Faculty of Medicine, Ankara, Turkey.

O. Lafcı is with Department of Biomedical Imaging and Image-guided Therapy, Medical University of Vienna, Vienna, Austria.

K. Lekadir is also with Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain.

Benchmark team participation: MIC team (J. Kächele, D. Bounias, A. Ertl); AIH-MAMA team (E. Poeta, L. Vargas, E. Pastor, M. A. Zuluaga); pimed-lab team (D. Schouten, J. H. Lee, M. Rusu); ViCOROB team (H. Awwad, J. C. Vilanova, R. Martí); AI Strollers team (D. Sleiman, N. Awasthi); Martel Lab team (M. Alberb, T. Xu, A. Martel); FME team (K. Geissler, R. Schäfer).

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols were exempt from review board approval.

† Corresponding author: Lidia Garrucho (lgarrucho@ub.edu).

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2026.3733343

cohort comprised 1,506 patients from multiple institutions in the United States, while evaluation was conducted on an external test set of 574 patients from three independent European centers to assess cross-continental and cross-institutional generalization. A unified scoring framework combined predictive performance with subgroup consistency across age, menopausal status, and breast density. Twenty-one international teams participated in the final evaluation phase. Results demonstrate substantial performance variability under a common external evaluation framework and reveal trade-offs between overall accuracy and subgroup fairness. The challenge provides standardized datasets, evaluation protocols, and public resources to promote the development of robust and equitable artificial intelligence systems for breast cancer imaging.

***Index Terms*— benchmark, breast cancer, fairness, magnetic resonance imaging, neoadjuvant chemotherapy, pathologic complete response.**

I. Introduction

Breast cancer is the most frequently diagnosed cancer among women worldwide, with 2.3 million new cases reported in 2022 and persistent disparities in outcomes across regions and populations [1], [2]. Dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) plays a central role in tumor characterization and treatment monitoring, particularly in the neoadjuvant chemotherapy (NAC) setting [3], [4]. Pathological complete response (pCR) following NAC is a well-established surrogate marker of favorable long-term outcomes, motivating increasing interest in imaging-based pCR prediction from pre-treatment MRI [5].

Artificial intelligence (AI) methods have demonstrated strong potential for automating breast tumor segmentation [6], [7] and predicting treatment response from DCE-MRI [8]–[11]. Deep learning approaches based on U-Net architectures and their variants have achieved increasingly accurate tumor delineation, supporting downstream quantitative analysis and treatment monitoring [6], [12]–[14]. However, reported performance varies substantially across datasets, acquisition protocols, and institutions [15], and evaluation practices remain heterogeneous, complicating direct comparison of methods and assessment of real-world robustness [16], [17].

Similarly, substantial progress has been made in predicting treatment response from pre-treatment breast MRI. Earlier studies primarily relied on radiomics-based approaches, while more recent work has gradually moved to deep learning strategies integrating multimodal data [18]–[20]. Recent work has explored a range of increasingly sophisticated strategies including MRI-based quantification of intratumoral heterogeneity [8], heterogeneity-informed nomograms integrating imaging biomarkers [9], and multimodal frameworks combining mammography, ultrasound, MRI for HER2 prediction [10]. Importantly, many of these approaches have moved beyond single-center evaluation by validating their models on independent public cohorts, such as Duke Breast Cancer MRI [21] and I-SPY2 [22], reflecting an increasing emphasis on assessing model generalizability. This demonstrates the growing maturity of AI-based response prediction and the increasing adoption of external validation practices.

Despite these advances, existing studies are typically developed and evaluated using different patient populations, imaging protocols, outcome definitions, and evaluation methodologies. Consequently, reported results are difficult to compare directly, and it remains unclear how competing approaches perform under a common evaluation framework. Moreover, while demographic and biological factors such as age, menopausal status, breast density, and tumor subtype are increasingly reported and analyzed, subgroup-specific performance assessment remains inconsistent across studies, limiting systematic evaluation of robustness and fairness.

Beyond breast MRI, several community benchmarks have been introduced to promote reproducible and generalizable AI development in medical imaging [23]–[25]. At the same time, growing evidence indicates that AI systems may exhibit performance disparities across demographic and clinical subgroups, motivating the incorporation of fairness-oriented analyses alongside conventional accuracy metrics [26]–[29]. However, fairness-aware benchmarking remains uncommon in breast imaging, particularly in conjunction with external multi-institutional validation.

To address these challenges, we introduce the **MAMA-MIA Challenge**, a multi-country benchmark for breast cancer analysis using DCE-MRI. Rather than proposing a single predictive model, the benchmark provides a standardized framework for evaluating and comparing AI methods across institutions and populations. The benchmark jointly evaluates (i) primary tumor segmentation and (ii) prediction of pathological complete response using pre-treatment imaging only. Models are trained on a large multi-institutional United States cohort and evaluated on an external European test set collected from three independent centers in different countries, enabling rigorous assessment of cross-institutional and cross-continental generalization. The evaluation framework further quantifies performance consistency across clinically relevant subgroups, explicitly incorporating fairness alongside predictive accuracy. The benchmark remains active on the CodaBench platform[1].

II. Methodology

A. Benchmark Overview

We define a benchmark framework for advancing computational methods in breast cancer analysis using dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI). Participants were tasked with addressing two complementary objectives: (i) automatic segmentation of the primary breast tumor (*Task 1*) and (ii) prediction of pathologic complete response (pCR) to neoadjuvant chemotherapy (*Task 2*).

Model development and training relied on large-scale, multicenter breast DCE-MRI data collected across 25 institutions in the United States. Submitted methods were evaluated on private validation and test datasets acquired from three independent European clinical centers in Spain, Poland and Lithuania. This design is to enable systematic assessment of cross-institutional generalizability and supports the analysis of performance disparities across clinically relevant demographic subgroups [30].

[1]https://www.codabench.org/competitions/7425/

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2026.3733343

To this end, Table I summarizes the composition and key characteristics of the training, validation, and test cohorts.

1) Training Dataset: The training cohort was based on the publicly available MAMA-MIA dataset [31], which was respectively built for this purpose and encompassess a multi-center resource comprising 1,506 patients with pre-treatment T1-weighted DCE breast MRI examinations. As detailed in [31], the MAMA-MIA dataset aggregates and harmonizes data from multiple TCIA [32] collections, including ISPY-1 [33], ISPY-2 [22], NACT [34], and DUKE [21].

Expert annotations are provided for the training data, including voxel-wise primary tumor segmentation masks and bounding boxes indicating the affected breast. Relevant clinical metadata, such as age, menopausal status, and tumor subtype, are available for training to support subgroup analysis and bias-aware model development. This metadata is withheld at test time to ensure that evaluation relies exclusively on imaging data.

The training dataset further exhibits substantial heterogeneity in acquisition parameters. Data were acquired from multiple vendors (GE, Siemens, Philips) and field strengths (1.5T and 3T), reflecting substantial acquisition heterogeneity.

In addition to the MAMA-MIA dataset, participants were permitted to use any publicly available datasets released under CC-BY-NC or more permissive licenses for model training.

2) Validation and test Datasets: The validation and test datasets are private and consist of 574 fat-suppressed DCE-MRI examinations collected from three European clinical centers: Medical University of Gdansk (GUM, Poland; 30 cases), Kauno Klinikos (KAU, Lithuania; 232 cases), and Hospital Clínic Barcelona (HCB, Spain; 312 cases).

Imaging was predominantly acquired in the axial plane (98.8%), with a small proportion in the sagittal plane (1.2%). Scans were obtained using 1.5T (59.2%) and 3T (40.8%) systems across multiple vendors, including Siemens (26.7%), GE (37.6%), and Philips (35.7%).

The dataset was split into a validation set of 58 cases and a held-out test set of 516 cases. Splitting was stratified by center, age, breast density, and menopausal status to ensure well-balanced representation in test and validation across these clinically relevant subgroups.

To assist teams in designing robust pipelines for both tasks, full cohort-level summaries of both the training and test datasets were made publicly available from the outset. For the classification task, the prevalence of target labels was kept consistent across the training and test splits, allowing participants to focus on model generalization and target-specific challenges (such as data imbalance) without experiencing significant label distribution shifts.

B. Annotation Protocol

1) Training Dataset: For the training data, the primary tumor was annotated on the first post-contrast DCE-MRI volume through a multi-stage workflow. First, an initial automatic segmentation was generated using a baseline nnU-Net model [13]. These automatic segmentations were then clinically reviewed by two expert clinicians, who assigned a quality label (Good, Acceptable, Poor, or Missed). Based on these labels, the cases were distributed among a panel of 16 expert clinicians from 10 institutions across Europe, Asia, and Africa, who manually corrected the segmentations as needed by adding or removing regions to obtain the final annotations. Following collection of the expert-corrected annotations, the authors performed a technical quality-control pass to verify data integrity (e.g., file format consistency and the absence of corrupted or invalid segmentations), without modifying the experts' delineations. No additional clinical approval step was performed after the expert corrections. Further details regarding the annotation guidelines and quality-control workflow are provided in the MAMA-MIA dataset publication [31].

2) Validation and Test Datasets: Primary tumor segmentations for the validation and test datasets were produced by four expert clinicians from the corresponding clinical centers (GUM, KAU, and HCB). To reduce inter-annotator variability, a harmonized annotation protocol was established, in which pathological boundaries were standardized through a pre-annotation calibration workshop [35]. Segmentation was performed using a cloud-based collaborative annotation platform [2], and tumor volumes were initialized with a shared 3D Smart Paint tool [36]. This tool provides interactive, user-guided tumor segmentation: clinicians first define a coarse delineation of the lesion on the slice with the highest conspicuity, after which the algorithm generates an initial 3D tumor region consistent with local image context across adjacent slices. These initial predictions are then iteratively refined through user interaction, where clinicians add or remove regions via a painting-based interface until consensus is reached.

C. Selection of Fairness Variables

To assess potential subgroup disparities in model performance, fairness analyses were conducted across age, menopausal status, and breast density groups. These variables were selected following recent recommendations for fairness evaluation in medical AI, which emphasize subgroup-based assessment across clinically relevant patient populations [37]. Age, menopausal status, and breast density are routinely reported patient-level characteristics in breast imaging and are associated with differences in breast cancer risk, hormonal environment and breast tissue composition [38]–[43]. Importantly, these variables were included as stratification factors for fairness assessment rather than as assumed causal determinants of model performance trained on Breast MRI.

We note that additional variables, such as molecular subtype (e.g., HR/HER2-defined subtypes), may also influence imaging appearance and treatment response. Accordingly, reporting performance stratified by subtype is common practice in the literature. However, unlike age, menopausal status, and breast density, tumor subtype is a disease-specific biological characteristic that is strongly associated with the target outcome, particularly in the pCR prediction task [44], [45]. Therefore, although we report subtype-specific performance to facilitate comparison with prior studies, molecular subtype was not included in the computation of the fairness score.

[2]https://www.cmrad.com/

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2026.3733343

TABLE I: Clinical and imaging variables across the MAMA-MIA benchmark datasets [31]. Four cohorts are used for training (NACT, ISPY1, ISPY2, DUKE) and three for validation/test (GUM, KAU, HCB). Data cells with more than 3 items are split into two rows to maximize readability. Best viewed in color.

MAMA-MIA Dataset	**Training**					**Validation / test**				**Total**
Collection	NACT	ISPY1	ISPY2	DUKE	Sum	GUM	KAU	HCB	Sum	Train+Val+Test
Country	USA	USA	USA	USA		Poland	Lithuania	Spain		
Year	2002–06	2010–16	2000–14	1995–02		2010–20	2010–20	2010–20		
Clinical variables										
pCR										
Yes/No/N.A.	11/53/0	49/118/4	316/664/0	64/216/11	440/1051/15	11/19/–	71/161/–	78/234/–	160/414/–	600/1465/15
Tumor subtype										
Lum / TN / HER2enr	21 / 11 / 8	67 / 45 / 25	536 / 358 / 86	123 / 85 / 50	747 / 499 / 169	11 / 9 / 3	95 / 71 / 39	209 / 43 / 49	315 / 123 / 91	1062 / 622 / 260
HER2pure / N.A.	6 / 18	29 / 5	0 / 0	30 / 3	65 / 26	7 / –	27 / –	11 / –	45 / –	110 / 26
Age group										
<40/40–49/50–59	13/25/18	35/62/56	208/300/317	61/104/74	317/491/465	7/9/6	21/51/54	22/57/86	50/117/146	367/608/611
60–69/>70/N.A.	6/2/0	18/0/0	134/18/3	41/11/0	199/31/3	7/1/–	61/45/–	87/60/–	155/106/–	354/137/3
Menopausal status										
Pre/Post/N.A.	0/0/64	0/0/171	512/296/172	166/124/1	678/420/408	17/13/–	73/159/–	109/203/–	199/375/–	877/795/408
Imaging variables										
Acq. Plane										
Axial/Sagittal	0/64	0/171	980/0	291/0	1271/235	30/0	225/7	312/0	567/7	1838/242
Field strength (T)										
1.5/3.0	64/0	171/0	715/265	136/155	1086/420	30/0	27/205	283/29	340/234	1426/654
Fat suppressed										
Yes/No	64/0	170/1	976/4	290/1	1500/6	30/0	232/0	312/0	574/0	2074/6
Manufacturer										
Siemens/GE										
Philips	0/64/0	44/115/12	252/611/117	115/176/0	411/966/129	30/0/0	27/0/205	96/216/0	153/216/205	564/1182/334
Slice thickness (mm)										
<1/1–1.49/1.5–1.99	0/0/0	0/0/5	11/163/8	0/283/4	11/446/17	0/30/0	27/0/196	0/2/36	27/32/232	38/478/249
2–2.49/≥2.5	64/0	83/83	695/103	3/1	845/187	0/0	9/0	272/2	281/2	1126/189
Pixel spacing (mm)										
<0.5/0.5–0.99										
1–1.49	2/62/0	15/150/6	2/939/39	0/274/17	19/1425/62	0/1/29	8/224/0	0/306/6	8/531/35	27/1956/97
Echo time (ms)										
1–1.99/2–2.99										
≥3	0/0/64	1/1/169	244/338/398	115/176/0	360/515/631	30/0/0	20/205/7	98/208/6	148/413/13	508/928/644
Repetition time (ms)										
3–3.99/4–4.99	0/0	0/0	1/226	32/116	33/342	30/0	0/223	0/269	30/492	63/834
5–5.99/≥6	0/64	1/170	218/535	116/27	335/796	0/0	2/7	34/9	36/16	371/812
Image size [x, y, z]										
Median	[256,256,60]	[256,256,60]	[512,512,80]	[512,512,170]	[512,512,80]	[352,352,160]	[528,528,187]	[512,512,82]	[512,512,112]	[512, 512, 80]
Mean	[264,264,59]	[278,278,63]	[481,481,105]	[479,479,169]	[449,449,111]	[352,352,166]	[525,525,190]	[485,485,88]	[494,494,134]	[461,461,117]
Std	[44,44,2]	[72,72,28]	[69,69,41]	[49,49,22]	[99,99,48]	[0,2,10]	[71,71,20]	[40,40,24]	[66,66,54]	[94,94,51]
Subjects	**64**	**171**	**980**	**291**	**1506**	**30**	**232**	**312**	**574**	**2080**

D. Data Preprocessing

All training, validation, and test cases underwent standardized preprocessing to ensure consistent data handling across sites. Raw DICOM files were converted to NIfTI format using the *pycad* library [46]. Dataset harmonization included quality control to remove scans with artifacts or incomplete sequences, extraction of relevant metadata, and enforcement of a uniform naming convention and folder structure [31].

Image orientation was standardized to preserve anatomical consistency: sagittal acquisitions were reoriented to the posterior–superior–right (PSR) coordinate system, and axial acquisitions to the left–anterior–superior (LAS) system. No additional preprocessing steps, such as intensity normalization, bias field correction, or voxel resampling, were enforced at the benchmark level. This design choice avoids introducing assumptions that may favor specific modeling strategies and allows participants to apply task-specific preprocessing tailored to their methods.

E. Benchmark Setup

All submissions were evaluated automatically using a standardized Python-based evaluation pipeline executed within a containerized environment on the CodaBench platform [47]. Participants were required to submit a single `.zip` archive containing source code and any required model weights, following a predefined submission template to ensure reproducibility and fair comparison.

The MAMA-MIA benchmark challenge was organized in multiple phases. An initial *Sanity-Check* phase using training data was used to validate submission formatting and code execution. This was followed by a validation phase in external data, during which teams were limited to one submission per day, with a maximum of 15 submissions per team, to mitigate overfitting. The benchmark concluded with a final test phase, in which performance was assessed on a held-out test set and only a single submission per team was permitted. To support participants throughout the development process, dedicated communication channels were maintained by the

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2026.3733343

organizers. All evaluations were executed locally on a Linux system equipped with an NVIDIA GeForce RTX 3090 GPU with 24 GB of memory. To ensure computational feasibility during large-scale evaluation and to encourage clinically practical solutions, submissions were required to complete inference within a maximum execution time of 5 minutes per image. Submissions exceeding this limit were automatically terminated and considered invalid.

The benchmark challenge was initially introduced in conjunction with the MICCAI 2025 conference [48], where it was hosted as part of the 2nd Deep-Breath Workshop [49]. The challenge website, documentation, and resources are publicly available at `https://www.ub.edu/mama-mia`.

F. Evaluation

Participants were encouraged to participate in both tasks: primary tumor segmentation (*Task 1*) and pathologic complete response (pCR) prediction (*Task 2*). Although the tasks differ in nature, both were evaluated using a unified scoring framework that jointly accounts for predictive performance and algorithmic fairness across clinically relevant demographic subgroups.

1) Unified Scoring Framework: For each task, submissions were ranked using a combined score that integrates task-specific predictive performance with fairness:

$$\mathcal{S} = (1-\lambda)\,\mathcal{S}_p + \lambda\,\mathcal{S}_f, \tag{1}$$

where $\mathcal{S}_p$ denotes the performance score, $\mathcal{S}_f$ the fairness score, and $\lambda \in [0,1]$ controls the trade-off between the two. In this benchmark, we set $\lambda = 0.5$, assigning equal importance to predictive accuracy and fairness.

Fairness was evaluated across a shared set of demographic variables:

$$V = \{\text{age},\ \text{menopausal status},\ \text{breast density}\}, \tag{2}$$

For each fairness variable $v \in V$, let G_v denote the set of subgroups induced by v.

$$\mathcal{G}_{age} = \{<40,\ 41{-}50,\ 51{-}60,\ 61{-}70,\ >71\}, \tag{3a}$$

$$\mathcal{G}_{menopausal} = \{\text{pre},\ \text{post}\}, \tag{3b}$$

$$\mathcal{G}_{density} = \{\text{A},\ \text{B},\ \text{C},\ \text{D}\}. \tag{3c}$$

This formulation incentivizes models to achieve consistent performance across subgroups of fairness variables rather than optimizing for average accuracy alone. Breast density was categorized according to BI-RADS breast composition categories (A–D: fatty, scattered fibroglandular density, heterogeneously dense, and extremely dense breasts) [50].

2) Task 1: Primary Tumor Segmentation: Segmentation performance was assessed using two complementary metrics: the Dice Similarity Coefficient (DSC) and the normalized Hausdorff Distance (NormHD). The performance score is defined as:

$$\mathcal{S}_p^{\text{seg}} = \frac{1}{2}\left(\text{DSC} + (1 - \text{NormHD})\right), \tag{4a}$$

$$\text{DSC} = \frac{1}{N}\sum_{i=1}^{N}\frac{2|P_i \cap Q_i|}{|P_i| + |Q_i|}, \tag{4b}$$

$$\text{NormHD} = \frac{1}{N}\sum_{i=1}^{N}\frac{HD(P_i, Q_i)}{HD_{\max}}, \tag{4c}$$

where N are the total number of evaluated cases, P_i and Q_i denote the predicted and ground-truth segmentations for case i, respectively, and $HD_{\max} = 150$ mm.

Additionally, for a given performance metric M, we denote by $\overline{M}_{v,g}$ the average value of M computed over all samples belonging to subgroup $g \in \mathcal{G}_v$, defined in Equation 3. The disparity D_v associated with fairness variable $v \in V$ is then computed as,

$$D_v = \frac{1}{2}\sum_{M \in \{\text{DSC},\ \text{NormHD}\}}\left(\max_{g\in\mathcal{G}_v}\overline{M}_{v,g} - \min_{g\in\mathcal{G}_v}\overline{M}_{v,g}\right) \tag{5a}$$

The fairness score $\mathcal{S}_f^{\text{seg}}$ for segmentation over all fairness variables is computed as:

$$\mathcal{S}_f^{\text{seg}} = \frac{1}{|V|}\sum_{v\in V}(1 - D_v) \tag{5b}$$

Empty segmentations were handled by assigning a DSC of 0 and a Hausdorff Distance of $HD_{\max}$, accounting for incorrect or missing predictions in bilateral breast images. The value of $HD_{\max}$ was derived from the maximum distance observed in the automatic segmentation baseline.

3) Task 2: Pathologic Complete Response Prediction: Classification performance was evaluated using Balanced Accuracy (BA), defined as:

$$\text{TPR} = \frac{1}{N_+}\sum_{i\in\mathcal{P}}\mathbf{1}(y_i = \hat{y}_i), \tag{6a}$$

$$\text{TNR} = \frac{1}{N_-}\sum_{i\in\mathcal{N}}\mathbf{1}(y_i = \hat{y}_i), \tag{6b}$$

$$\mathcal{S}_p^{\text{cls}} = \frac{\text{TPR} + \text{TNR}}{2}, \tag{6c}$$

where $\mathcal{P}$ and $\mathcal{N}$ denote the sets of positive and negative samples, respectively. The classification fairness score $\mathcal{S}_f^{\text{cls}}$ was defined using an Equalized Odds–based criterion. For each variable v:

$$\mathcal{D}_v^{\text{TPR}} = \max_{g\in\mathcal{G}_v}\text{TPR}_{v,g} - \min_{g\in\mathcal{G}_v}\text{TPR}_{v,g}, \tag{7a}$$

$$\mathcal{D}_v^{\text{FPR}} = \max_{g\in\mathcal{G}_v}\text{FPR}_{v,g} - \min_{g\in\mathcal{G}_v}\text{FPR}_{v,g}, \tag{7b}$$

$$\mathcal{D}_v^{\text{EQ}} = \mathcal{D}_v^{\text{TPR}} + \mathcal{D}_v^{\text{FPR}}, \tag{7c}$$

$$\mathcal{S}_f^{\text{cls}} = 1 - \frac{1}{|V|}\sum_{v\in V}\mathcal{D}_v^{EQ}. \tag{7d}$$

This formulation rewards classifiers that maintain consistent error rates across demographic subgroups. Submissions producing constant predictions (e.g., all zeros or all ones) were considered invalid and excluded from evaluation.

Although not part of the official challenge evaluation framework, we also computed Recall (sensitivity) and the Area

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2026.3733343

Under the Receiver Operating Characteristic Curve (AUC) to compare the different classification methods in Section III. Recall is defined as Recall $= \frac{\text{TP}}{\text{TP+FN}}$ (where TP are true positives and FN are false negatives). The AUC is mathematically equivalent to the probability that the model ranks a randomly chosen positive instance higher than a randomly chosen negative one, denoted as AUC $= P(f(x^+) > f(x^-))$, where $f(\cdot)$ is the classifier's output score.

G. Baselines

For *Task 1* (primary tumor segmentation), the baseline model was a 3D nnUNet trained on the 1,506 DCE-MRI volumes with expert annotations from the MAMA-MIA dataset [31]. Given nnUNet's widespread adoption in medical image segmentation benchmarks, this baseline provides a strong reference point while encouraging participants to explore novel architectures and learning strategies. Baseline model weights are publicly available in Synapse [3].

For *Task 2* (pCR prediction), the baseline was a random classifier. Specifically, for each test case the classifier draws a prediction score uniformly at random from $[0, 1]$ using `numpy.random.rand()` function and assigns the binary pCR label by thresholding this score at 0.5, without using any imaging, segmentation-derived, or clinical information. The prevalence of pCR was comparable across splits (training: 440/1,491 labeled cases, ≈29.5%; validation/test: 160/574, ≈27.9%; Table I). This baseline corresponds to the `predict_classification` method of the official sample submission provided to all participants, publicly available in the competition CodaBench repository. This model establishes a lower bound on performance and allows participants to quantify the added value of their imaging-driven predictive models.

H. Participating Methods

This section provides a structured overview of the participating methods for each task, with the goal of characterizing the methodological space explored by the community rather than exhaustively describing individual implementations. For each task, we summarize common design choices and methodological trends, followed by concise technical descriptions of the top-ranked approaches. Detailed architectural, preprocessing, and training configurations are reported in Tables II and III.

1) Task 1: Primary Tumor Segmentation: The primary tumor segmentation task attracted a diverse set of approaches that, nevertheless, shared several common design principles. A comparative summary of the top five segmentation methods is provided in Table II. All approaches employ fully three-dimensional architectures operating directly on volumetric DCE-MRI data, reflecting the importance of spatial context for accurate tumor delineation. Input standardization through resampling and intensity normalization was consistently adopted, while voxel spacing, interpolation strategies, and normalization schemes varied across teams.

[3] `https://www.synapse.org/Synapse:syn60868042/wiki/628716`

Most methods incorporated multiple DCE phases as input channels, typically including the pre-contrast and one or more early post-contrast acquisitions, in order to capture tumor enhancement dynamics. Architecturally, the majority of approaches were based on nnU-Net variants, including residual encoder configurations, while one method explored a 3D Vision Transformer backbone. Inference strategies frequently relied on ensembling across cross-validation folds, and several methods applied lightweight post-processing steps to suppress anatomically implausible segmentations.

❶ **MIC**: The MIC team proposed a residual-encoder nnU-Net with large-scale self-supervised pretraining on external DCE-MRI data [56]. A masked autoencoding strategy was used to pretrain on n=4,799 public DCE-MRI volumes, followed by fine-tuning on the MAMA-MIA dataset [31]. Multi-phase input and five-fold ensembling were combined with breast-bounding-box post-processing to improve robustness and subgroup consistency. Further details are reported in [57].

❷ **FME** The FME team adopted a robustness-oriented strategy based on a single subtraction image (first post-contrast minus pre-contrast) to reduce sensitivity to acquisition variability. A five-fold ensemble of 3D residual-encoder nnU-Nets was trained, complemented by axis flipping test time augmentation and largest connected component filtering. This design prioritized generalization and stability across unseen centers.

❸ **ViCOROB** The ViCOROB team, as described in [58], employed a residual-encoder architecture using three representative DCE phases selected to preserve early enhancement kinetics. Breast-level cropping focused on the tumor-containing region, eliminating contralateral interference and improving signal-to-noise ratio. To enhance robustness across multi-center data, percentile-based intensity clipping, isotropic resampling, and z-score normalization were applied to standardize inputs.

❹ **Martel Lab** The Martel Lab introduced a joint learning framework that simultaneously performs tumor segmentation and pCR prediction using a shared encoder. A 3D Vision Transformer [51] pretrained on large-scale medical imaging data [59] was combined with temporal attention modules to model contrast enhancement dynamics across multiple DCE phases within an end-to-end multi-task architecture.

❺ **AIH-MAMA** The AIH-MAMA team applied FairMedSeg [60], a fairness-aware segmentation framework that integrates subgroup performance divergence directly into the training objective. Clinically relevant subgroups were identified using DivExplorer [61], and sample reweighting was used to mitigate performance disparities without requiring subgroup metadata at inference time [62].

2) Task 2: Pathologic Complete Response Prediction: For the pCR prediction task, participating methods explored complementary strategies for modeling treatment response from pre-treatment DCE-MRI. Table III summarizes the key technical characteristics of the top three classification approaches. All methods formulated pCR prediction as a supervised binary classification problem, but differed substantially in how imaging information was represented and exploited.

Two approaches adopted lesion-centered, end-to-end 3D

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2026.3733343

TABLE II: Technical summary of top-performing primary tumor segmentation methods (Task 1). Abbreviations used in the table- nnUNet pre: default nnUNet [13] preprocessing strategy; RAS: Right-Anterior-Superior orientation; p0/p1/p2: pre-contrast, first post-contrast and second post-contrast phases; px: last available post-contrast phase; s1: subtraction of first post-contrast phase with pre-contrast phase; bbox: Bounding Box; CE: Cross Entropy; CV: Cross Validation; n/a: not applicable/available.

	❶ MIC	❷ FME	❸ ViCOROB	❹ Martel Lab	❺ AIH-MAMA
Preprocessing Strategy					
resample	third-order spline $1 \times 1 \times 1$	nnUNet pre $0.7 \times 0.7 \times 2.0$	b-spline $1 \times 1 \times 1$	trilinear $1 \times 1 \times 1$	bilinear $1 \times 1 \times 1$
crop or pad	×	×	breast bbox	breast bbox	non-zero voxels $320 \times 320 \times 128$
reorient	×	×	×	RAS	RAS
intensity clipping	×	×	percentile $[0.1, 99.9]$	percentile $[0.05, 99.95]$	×
normalize	z-score	z-score	z-score	min-max	z-score (non-zero pixels only)
Training Strategy					
DCE phases	p0, p1, p2	s1	p0, p1, p2	p0, p1, px	p0, p1, p2
network	3D nnUNet (ResEncL) [13]	3D nnUNet (ResEncL) [13]	3D nnUNet (ResEncL) [13]	3D ViT [51]	3D UNet
# parameters (M)[a]	383.5	2637.35	102.35	597.58	4.8
pretrained	✓	×	×	✓	×
optimizer	SGD [52]	SGD [52]	SGD [52]	AdamW [53]	AdamW [53]
batch size	3	2	3	8	1
loss	Soft Dice [54]	CE + Dice	CE + Dice	CE + Dice	Focal [55]
epochs / iterations	10K	1000	1000	30K	200
Inference Strategy					
ensemble	5-fold CV	5-fold CV	n/a	2 folds of CV (space limit)	n/a
post-processing	breast bbox	largest connected component	n/a	n/a	breast bbox

[a] For ensemble approaches, the number of parameters is the total parameter count over all the models in the ensemble. Public implementations are available at the repositories provided by the participating teams: `https://github.com/jojoka-234/mama-mia-challenge`, `https://github.com/FraunhoferMEVIS/MAMA-MIA-Challange-FME`, `https://github.com/VICTORIA-project/MAMA-MIA_Challenge`, `https://github.com/Muhammad-Al-Barbary/MAMA-MIA_MartelLab`, `https://github.com/luvargas2/AIH-MAMA`

classification pipelines, extracting fixed-size regions-of-interest around the primary tumor. These methods varied in their handling of temporal information, either treating multiple post-contrast phases as separate input channels or using phase selection and augmentation strategies. In contrast, one approach decoupled segmentation and classification by reusing deep features extracted from a pretrained segmentation network.

❶ **pimed-lab** The pimed-lab team formulated pCR prediction as a lesion-centered 3D classification task and adopted a hybrid supervision strategy combining self-supervised learning through a 3D SimCLR [68] objective early in training, gradually transitioning to supervised learning once robust features had been learned. Two 3D convolutional architectures [63], [64] were trained and ensembled to improve feature learning and generalization under limited labeled data.

❷ **FME** The FME team employed pretrained 3D video classification architectures [65] operating on lesion-centered multi-phase DCE-MRI inputs. Hyperparameters were optimized using Ray Tune [70] via cross-validation, and model selection explicitly considered a composite score balancing predictive performance and subgroup fairness. A large ensemble and test time augmentation were used at inference to improve robustness.

❸ **AI Strollers** The AI Strollers team proposed a modular pipeline in which deep features were extracted from a pretrained segmentation network based on parameter-efficient SegRestNet [14]. These features were subsequently used to train an XGBoost classifier [66], decoupling voxel-level segmentation learning from patient-level response prediction.

III. Results

A total of 21 teams from 14 countries submitted valid entries to the final test phase of the benchmark. 13 teams submitted to both tasks, while 7 teams and 1 team submitted exclusively to Task 1 and Task 2, respectively. This section presents and analyzes the results obtained by the participating teams. As shown in Fig. 1, 12 teams outperformed the baseline in both fairness and performance for Task 1 (Fig. 1a). In Task 2 (Fig. 1b), only three teams demonstrated improvements over the baseline random model; all three improved fairness, while only two also achieved higher-than-baseline performance. The results for each task are presented in the following dedicated subsections.

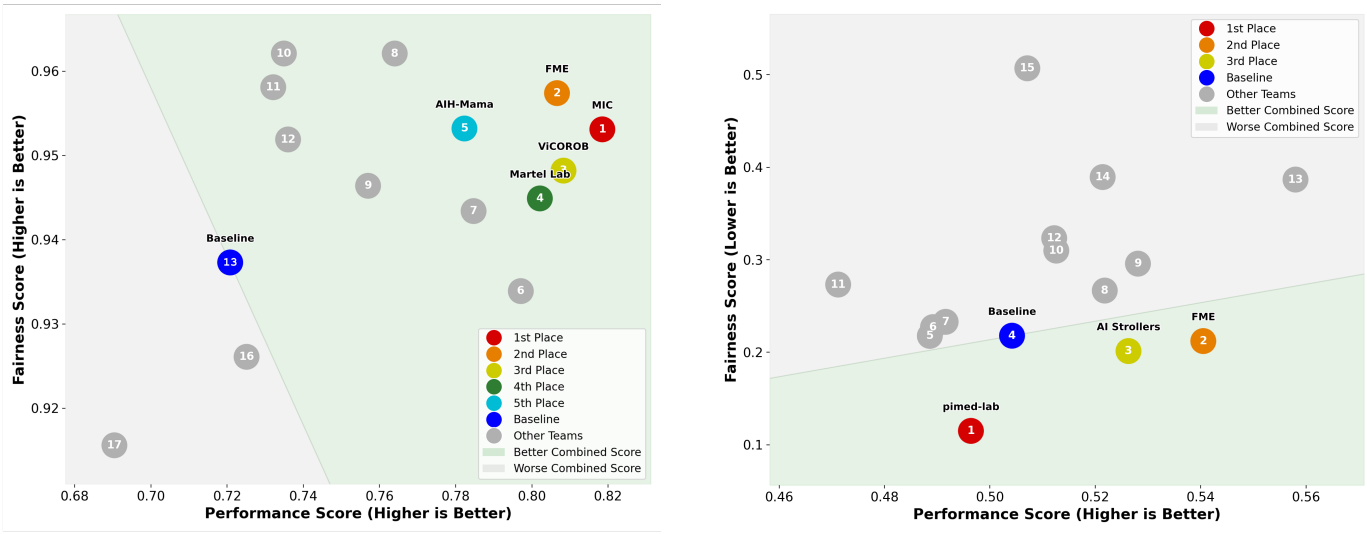
A. Task 1: Primary Tumor Segmentation

Table IV summarizes final rankings. The top five teams improved over the baseline by 0.43–4.89% DSC, with concurrent fairness gains ($\sim 2\%$), indicating consistent performance improvements across subgroups.

Qualitative analysis (Fig. 2) shows strong performance for well-circumscribed mass lesions, while failures are concentrated in small, non-mass, low-contrast tumors and implant-

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2026.3733343



(a) Primary Tumor Segmentation (Task 1). (b) Pathologic Complete Response Prediction (Task 2).

Fig. 1: Fairness plotted against performance across the two MAMA-MIA benchmark tasks. Teams with fairness-performance tradeoffs within the area indicated in green improve upon the baseline tradeoff. Note that the axis ranges are restricted to improve readability; consequently, teams falling outside these bounds are not shown.

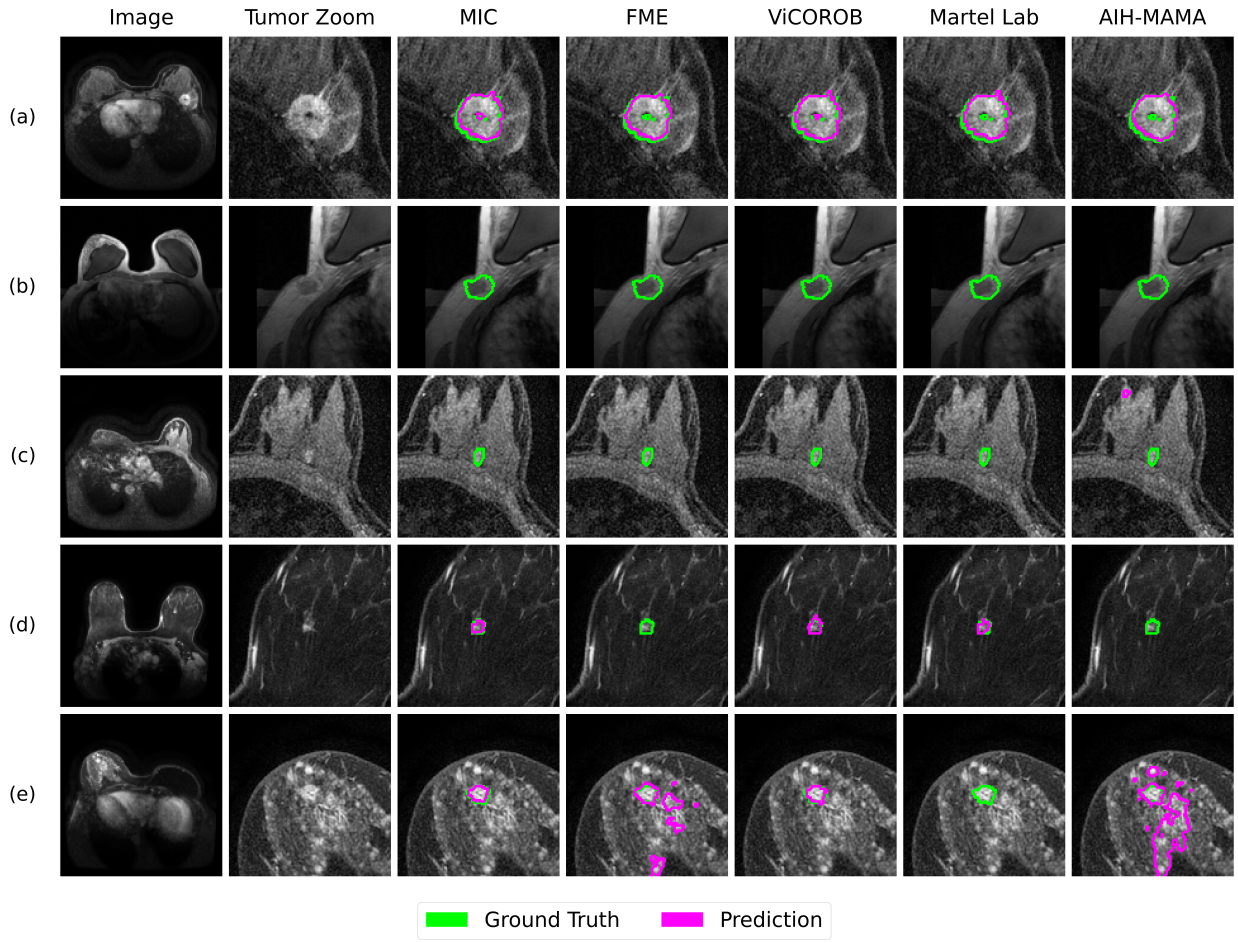


Fig. 2: Qualitative results for challenging cases. The columns correspond to the MRI image, the zoomed-in tumor region (Tumor Zoom), and the top-5 team predictions (predicted segmentations) overlaid with the expert annotated tumor segmentation (ground truth). Results are shown for: (a) clip artifact (successful), (b) breast implant (failure), (c) low contrast (failure), (d) small tumor (high variance between teams), and (e) non-mass enhancement (high variance between teams). Artifacts, low discriminability, and atypical morphology remain key challenges.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2026.3733343

TABLE III: Technical summary of top-performing pathologic complete response (pCR) prediction methods (Task 2). Abbreviations used in the table - nnUNet pre: default nnUNet [13] preprocessing strategy; bbox: Bounding Box; p0/p1/p2: pre-contrast, first post-contrast and second post-contrast phases; px: last available post-contrast phase; BCE: Binary Cross Entropy; CE: Cross Entropy; CV: Cross Validation; n/a: not applicable/available.

	① **pimed-lab**	② **FME**	③ **AI Strollers**
Preprocessing			
resample	$1 \times 1 \times 1$	varying per model	nnUNet pre [13] $0.7 \times 0.7 \times 2.0$
crop / pad	lesion bbox	lesion bbox	×
reorient	×	×	×
clip	×	×	×
normalize	z-score	z-score	z-score
Training Strategy			
DCE phases	p1,..px (train) p1 (inference)	p0, p1, p2	p0, p1, p2
network[a]	3D ResNet-18 [63] mc3_18 [64]	r3d_18, swin3d_t, r2plus1d_18, mc3_18 [65]	SegResNet [14] XGBoost [66]
parameters (M)[b]	44.65	593.6	2.61
pretrained	×	✓	×
optimizer	AdamW [53]	AdamW [53]	Adam [67] n/a
batch size	48	[16, 32]	2 n/a
loss	CE + NT-Xent [68]	BCE	CE + Dice + FRLoss [69] log-loss
epochs	200	25	1000 n/a
Inference Strategy			
ensemble	2 models	5 models × 5-fold CV (25 models)	2 estimators

[a] AI Strollers combined two models: a segmentation model for feature extraction (black) and a pCR prediction model (gray).

[b] For ensemble approaches, the number of parameters is the total parameter count over all the models in the ensemble. Public implementations are available at the repositories provided by the participating teams: `https://github.com/dnschouten/mamamia_code_submission`, `https://github.com/FraunhoferMEVIS/MAMA-MIA-Challange-FME`, `https://github.com/slndaniel/MamaMiaSubmission`

associated artifacts. These cases exhibit high inter-method variability, indicating sensitivity to ambiguous morphology.

This variability is consistent with the analysis shown in Fig. 3, which indicates that smaller tumors achieve lower average DSC compared to medium and large tumors. Notably, the performance gap between the top five and bottom five methods is largest for small tumors, suggesting that top-performing approaches yield the greatest relative improvements in this particularly challenging regime. Tumor size also influences performance across centers. Overall DSC varies by center, with the highest performance observed for GUM, followed by KAU and HCB, as shown in Fig. 4. This trend may be partly explained by differences in tumor size distributions across centers: centers with a higher proportion of small tumors, most notably KAU and HCB, tend to exhibit lower DSC.

The fairness analysis shown in Fig. 5 demonstrates that the top five teams consistently achieved higher average DSC than the bottom five across all subgroups. Performance was largely comparable across subgroups defined by breast density, age, and menopausal status, with no systematic subgroup-specific differences observed. This suggests a high degree of fairness and robustness in the analyzed methods with respect to these patient characteristics. In addition to the fairness variables, Fig. 5d compares segmentation performance across four different breast cancer tumor subtypes. The triple-negative subtype yields the best results, with the top 5 models achieving the highest mean DSC (0.76) and the tightest performance distribution. This suggests that tumors of this specific subtype might have the most distinct or recognizable morphological characteristics in the imaging modality used. Conversely, the luminal subtype proved to be the most challenging overall; it resulted in the lowest mean DSC (0.47) and the widest spread for the weaker models, while also producing the lowest individual outlier (DSC < 0.30) even among the top-performing methods.

We evaluated the stability of the segmentation leaderboard across varying levels of the fairness weighting parameter λ (Fig. 6a). Because the segmentation models achieved relatively high overall performance, the final rankings proved resilient to adjustments in λ. When evaluating only based on performance score ($\lambda = 1$), the the top four teams remain unchanged. Notably, the team AIH-MAMA, which incorporates explicit fairness adjustments into their algorithm, demonstrated significantly stronger fairness metrics than the teams in 5th and 6th position on performance score only ranking. Consequently, this strong fairness profile secures them the 5th position in the final aggregated leaderboard.

B. Task 2: Pathologic Complete Response Prediction

Table V summarizes the final rankings for pCR prediction. The top three teams achieved combined scores of 0.6907, 0.6642, and 0.6625, compared to 0.6431 for the baseline random model. Notably, top three submissions did not substantially outperform the baseline in performance score alone.

Only one team (PM) achieved performance significantly different from random prediction ($p = 0.0140$), while FME showed marginal significance ($p = 0.063$). All other methods were statistically indistinguishable from random performance ($p > 0.1$), underscoring the intrinsic difficulty of pCR prediction from pre-treatment MRI alone.

Clear performance–fairness trade-offs were observed. While PM obtained the highest performance score (0.5580), it exhibited the largest fairness gap (0.3867), resulting in a lower combined score (rank 13). In contrast, the top-ranked teams achieved competitive performance with improved fairness, indicating balanced subgroup behavior. However, further analysis between subgroups is not conducted for this task due to the lack of discriminative performance by the submitted models.

In all submissions, recall for the pCR class was generally low (Fig. 7). For instance, the top-performing method, pimed-lab, achieved a pCR recall of just 0.01 compared to a 'not pCR' recall of 0.98 (Δ = +0.96). Similarly, FME and AI Strollers demonstrated low pCR recalls of 0.26, alongside much higher 'not pCR' recalls of 0.82 and 0.80, respectively (Δ = +0.56 and Δ= +0.54). This suggests conservative prediction strategies and limited sensitivity to complete responders. This likely reflects both class imbalance and the inherent heterogeneity of treatment response patterns.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2026.3733343

TABLE IV: Final leaderboard for Task 1: Primary Tumor Segmentation. Confidence intervals (95%) for all Task 1 metrics were estimated using a non-parametric percentile bootstrap ($B = 1000$, seed=42). DSC, NormHD, and Performance Score CIs were obtained by resampling patients with replacement and recomputing the mean metric on each bootstrap sample. Fairness Score CIs were computed using patient-level bootstrapping with repeated group-level disparity estimation across demographic groups. Combined Score CIs were derived jointly from the same bootstrap resamples for both performance and fairness components.

Rank	Team	Combined Score [95% CI] ↑	Fairness Score [95% CI] ↑	Performance Score [95% CI] ↑	DSC [95% CI] ↑	NormHD [95% CI] ↓
1	MIC	0.8858 [0.8577, 0.8969]	0.9531 [0.9051, 0.9651]	0.8185 [0.8018, 0.8350]	0.7360 [0.7160, 0.7539]	0.0990 [0.0816, 0.1172]
2	FME	0.8820 [0.8464, 0.8913]	0.9574 [0.8976, 0.9659]	0.8066 [0.7873, 0.8248]	0.7125 [0.6895, 0.7330]	0.0993 [0.0815, 0.1182]
3	ViCOROB	0.8782 [0.8485, 0.8891]	0.9482 [0.8963, 0.9643]	0.8083 [0.7913, 0.8248]	0.7182 [0.6976, 0.7381]	0.1017 [0.0859, 0.1189]
4	Martel Lab	0.8735 [0.8471, 0.8856]	0.9449 [0.8999, 0.9619]	0.8021 [0.7863, 0.8182]	0.7121 [0.6918, 0.7315]	0.1078 [0.0915, 0.1244]
5	AIH-Mama	0.8677 [0.8379, 0.8781]	0.9532 [0.8984, 0.9661]	0.7823 [0.7634, 0.8012]	0.6914 [0.6701, 0.7139]*	0.1268 [0.1081, 0.1452]
6	HWT@YCH	0.8655 [0.8365, 0.8789]	0.9339 [0.8843, 0.9514]	0.7971 [0.7798, 0.8143]	0.7080 [0.6872, 0.7273]	0.1138 [0.0967, 0.1324]
7	Flamingo	0.8640 [0.8287, 0.8781]	0.9434 [0.8798, 0.9593]	0.7847 [0.7634, 0.8033]	0.7033 [0.6807, 0.7245]*	0.1338 [0.1120, 0.1576]
8	CALADAN	0.8631 [0.8239, 0.8683]	0.9621 [0.8947, 0.9629]	0.7640 [0.7412, 0.7862]	0.7022 [0.6811, 0.7246]*	0.1742 [0.1487, 0.2018]
9	bigAI	0.8517 [0.8189, 0.8626]	0.9464 [0.8860, 0.9596]	0.7570 [0.7360, 0.7766]	0.6872 [0.6649, 0.7076]*	0.1732 [0.1503, 0.1970]
10	Shangqi,Gao@CAM	0.8485 [0.8157, 0.8551]	0.9621 [0.9025, 0.9655]	0.7349 [0.7144, 0.7555]	0.6101 [0.5836, 0.6359]	0.1404 [0.1223, 0.1584]
11	GK_KI	0.8451 [0.8060, 0.8512]	0.9581 [0.8896, 0.9606]	0.7321 [0.7074, 0.7552]	0.6330 [0.6083, 0.6570]	0.1688 [0.1441, 0.1955]
12	Jeff	0.8439 [0.8071, 0.8529]	0.9519 [0.8886, 0.9570]	0.7360 [0.7107, 0.7588]	0.7025 [0.6819, 0.7221]	0.2305 [0.1979, 0.2636]*
13	*Baseline*	0.8290 [0.7868, 0.8416]	0.9373 [0.8645, 0.9516]	0.7208 [0.6984, 0.7432]	0.6871 [0.6662, 0.7075]	0.2455 [0.2125, 0.2806]
14	Dynamo	0.8290 [0.7868, 0.8416]	0.9373 [0.8645, 0.9516]	0.7208 [0.6984, 0.7432]	0.6871 [0.6662, 0.7075]*	0.2455 [0.2126, 0.2806]*
15	PM	0.8290 [0.7868, 0.8416]	0.9373 [0.8645, 0.9516]	0.7208 [0.6984, 0.7432]	0.6871 [0.6662, 0.7075]*	0.2455 [0.2125, 0.2806]*
16	AEHRC-MIA	0.8256 [0.7876, 0.8399]	0.9261 [0.8605, 0.9405]	0.7251 [0.7010, 0.7490]	0.6781 [0.6562, 0.7004]*	0.2280 [0.1960, 0.2601]*
17	AI Strollers	0.8030 [0.7507, 0.8249]	0.9156 [0.8211, 0.9459]	0.6904 [0.6612, 0.7189]	0.6296 [0.6005, 0.6568]	0.2489 [0.2166, 0.2818]*
18	MedImgLab_Unipa	0.7270 [0.6912, 0.7438]	0.9084 [0.8432, 0.9292]	0.5456 [0.5208, 0.5716]	0.4717 [0.4466, 0.4979]	0.3805 [0.3494, 0.4110]
19	FPixel	0.7270 [0.6912, 0.7438]	0.9084 [0.8432, 0.9292]	0.5456 [0.5208, 0.5716]	0.4717 [0.4466, 0.4979]	0.3805 [0.3494, 0.4110]
20	BWS-KNU	0.7257 [0.6666, 0.7348]	0.9382 [0.8352, 0.9368]	0.5132 [0.4740, 0.5497]	0.4556 [0.4245, 0.4859]	0.4291 [0.3816, 0.4790]
21	CIG@Illinois	0.6593 [0.6122, 0.6802]	0.8931 [0.8086, 0.9213]	0.4256 [0.3968, 0.4550]	0.5195 [0.4936, 0.5456]	0.6683 [0.6265, 0.7096]

* DSC and NormHD not statistically different from the baseline according to a paired t-test. All other values are significantly different ($p < 0.05$).

TABLE V: Final leaderboard for Task 2: Treatment Response Prediction to Neoadjuvant Chemotherapy. Note that AUC is reported as an additional result and is excluded from the ranking scheme. Confidence intervals (95% CI) were estimated using a non-parametric percentile bootstrap ($B = 1000$, seed=42). Balanced Accuracy and AUC were bootstrapped from prediction–label pairs, while Fairness and Combined Scores used patient-level resampling with repeated equalized-odds disparity computation across demographic groups. The Combined Score CI was obtained by repeating the resampling jointly for performance and fairness components.

Rank	Team	Combined Score [95% CI] ↑	Fairness Score [95% CI] ↓	Performance Score [95% CI] ↑	AUC [95% CI] ↑
1	pimed-lab	0.6907 [0.6364, 0.7272]	0.1150 [0.0399, 0.2292]	0.4964 [0.4853, 0.5102]	0.4556 [0.4035, 0.5121]
2	FME	0.6642 [0.4893, 0.6602]	0.2121 [0.2232, 0.5561]	0.5405 [0.4992, 0.5841]+	0.5722 [0.5195, 0.6305]
3	AI Strollers	0.6625 [0.4778, 0.6569]	0.2013 [0.2148, 0.5663]	0.5263 [0.4870, 0.5658]	0.5218 [0.4629, 0.5771]
4	*Baseline*	0.6431 [0.4600, 0.6360]	0.2179 [0.2418, 0.5701]	0.5042 [0.4551, 0.5498]	0.5031 [0.4455, 0.5551]
5	BWS-KNU	0.6355 [0.4792, 0.6325]	0.2177 [0.2205, 0.5324]	0.4886 [0.4478, 0.5314]	0.5075 [0.4513, 0.5656]
6	Martel Lab	0.6310 [0.4802, 0.6316]	0.2272 [0.2397, 0.5164]	0.4892 [0.4477, 0.5316]	0.4740 [0.4195, 0.5330]
7	Flamingo	0.6293 [0.4459, 0.6238]	0.2329 [0.2584, 0.5906]	0.4916 [0.4449, 0.5384]	0.4683 [0.4151, 0.5284]
8	AEHRC-MIA	0.6275 [0.4681, 0.6206]	0.2667 [0.2889, 0.5832]	0.5218 [0.4783, 0.5698]	0.5378 [0.4819, 0.5996]
9	Jeff	0.6162 [0.4345, 0.6273]	0.2958 [0.2815, 0.6459]	0.5281 [0.4793, 0.5748]	0.5075 [0.4519, 0.5633]
10	MIC	0.6013 [0.5046, 0.6348]	0.3099 [0.2448, 0.5035]	0.5126 [0.4716, 0.5530]	0.5000 [0.5000, 0.5000]
11	CALADAN	0.5990 [0.4363, 0.5978]	0.2732 [0.2818, 0.5951]	0.4712 [0.4243, 0.5173]	0.4819 [0.4264, 0.5417]
12	HWT@YCH	0.5945 [0.4282, 0.6186]	0.3233 [0.2790, 0.6543]	0.5122 [0.4643, 0.5591]	0.5193 [0.4650, 0.5704]
13	PM	0.5856 [0.4407, 0.6117]	0.3867 [0.3432, 0.6739]	0.5580 [0.5111, 0.6014]*	0.5702 [0.5160, 0.6221]
14	ViCOROB	0.5660 [0.4608, 0.6101]	0.3894 [0.2989, 0.6001]	0.5214 [0.4783, 0.5647]	0.5595 [0.5036, 0.6151]
15	bigAI	0.5001 [0.3610, 0.5398]	0.5070 [0.4317, 0.7725]	0.5071 [0.4609, 0.5522]	0.5011 [0.4461, 0.5528]

* Balanced accuracy significantly higher than a random classifier based on 1000 random simulations ($p < 0.05$).
+ Marginally higher than random ($0.05 \leq p < 0.10$).
All other methods showed no statistically significant difference from random ($p \geq 0.10$).

Analysis of the calibration curves for the top three methods (Fig. 8) reveals limited separation in the predicted probabilities, which are predominantly clustered around the overall class prevalence, of approximately 0.27. The mean predicted probabilities are narrowly constrained; pimed-lab's predictions range only from 0.12 to 0.45, while FME and AI Strollers show slightly wider dispersions from 0.21 to 0.59 and 0.07 to 0.70, respectively. Furthermore, while the models output varying predicted probabilities, the actual fraction of true positives (pCR cases) fails to scale proportionally. For instance, even when AI Strollers and FME output their most confident predictions (0.70 and 0.59, respectively), the actual fraction of pCR cases in those subsets reaches only 0.33 and 0.38. For pimed-lab, the true positive rate never exceeds 0.35 despite predictions reaching 0.45. Because the actual response rates rarely rise meaningfully above the population average, all curves fall heavily below the reference line at higher thresholds, revealing systematic overconfidence and a shared inability to reliably identify true treatment responders.

We evaluated the stability of the classification leaderboard

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2026.3733343

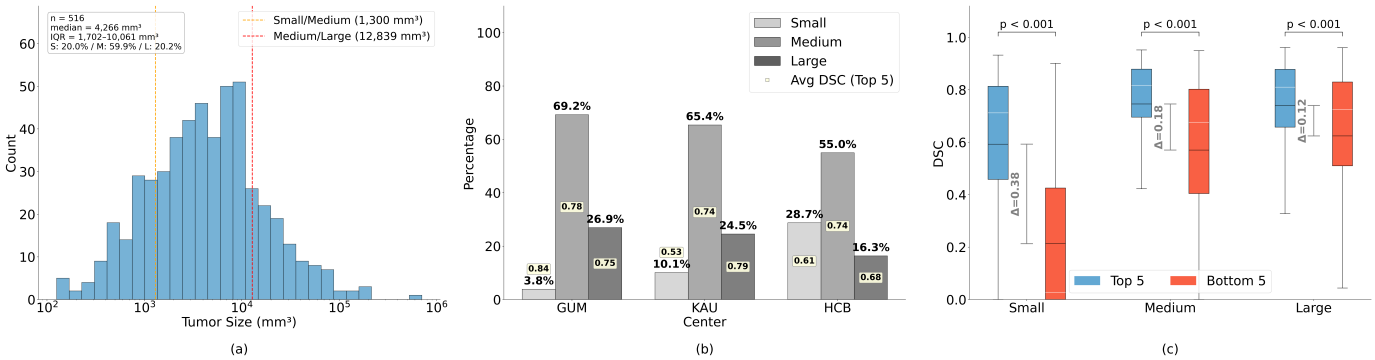


Fig. 3: Impact of tumor size on segmentation performance. (a) Histogram of tumor volumes (mm^3) on a logarithmic scale, with dashed lines indicating the 20th and 80th percentile thresholds used to categorize tumors as Small, Medium, or Large, (b) Percentage distribution of tumor size categories across the three participating centers (GUM, KAU, HCB), with boxed values showing the average Dice Similarity Coefficient (DSC) for the top 5 teams in each category. (c) Comparison of DSC distributions between the top 5 and bottom 5 performing teams across tumor size categories, with gap bars (Δ) indicating the difference in mean DSC. Brackets give two-sided Wilcoxon signed-rank p-values comparing the Top 5 against the Bottom 5 teams. For each patient, the mean DSC across the five top-ranked teams was compared with the mean DSC across the five bottom-ranked teams, so patients are the unit of analysis and the two arms are paired on identical cases.

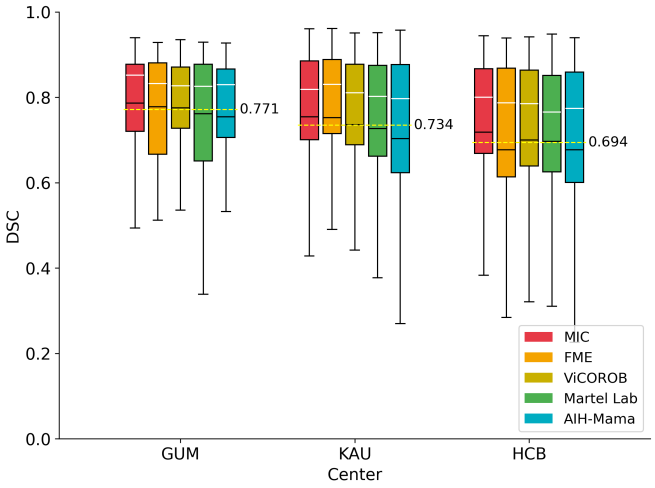


Fig. 4: Distribution of DSC across the three participating centers (GUM, KAU, HCB) for the top 5 performing teams in Task 1. Each boxplot represents the per-case DSC distribution for a team, in order of their ranking. Horizontal dashed lines indicate the average DSC across all top 5 teams for each center.

by analyzing how final rankings changed across varying levels of the fairness weighting parameter λ (Fig. 6b). Because the pCR prediction task yielded lower overall performance scores, the rankings were highly sensitive to changes in λ. Setting $\lambda = 1.0$ (i.e., ignoring fairness) results in participant PM moving from rank 13 out of 15 to rank 1, while only one (FME) of the top three teams at $\lambda = 0.5$ remains in the top tier. Even modest increases in λ lead to substantial reordering: at $\lambda = 0.6$, 6 out of 15 teams change rank, and at $\lambda = 0.8$, 13 out of 15 teams are affected.

IV. Discussion and Future Work

Building upon recent advances in breast MRI AI, including increasingly sophisticated externally validated segmentation and treatment-response prediction models, the MAMA-MIA Challenge was designed as a real-world stress test for AI in breast MRI, explicitly evaluating generalization to unseen clinical sites and consistency across patient subgroups under a common benchmark framework. We synthesize outcomes across primary tumor segmentation and pretreatment pCR prediction, analyzing performance trends, failure modes, and fairness behavior. As MAMA-MIA evolves from a one-time competition into a standardized benchmark with unified accuracy–fairness evaluation, we discuss implications for evaluation, clinical translation, and future research aimed at improving robustness and clinical relevance.

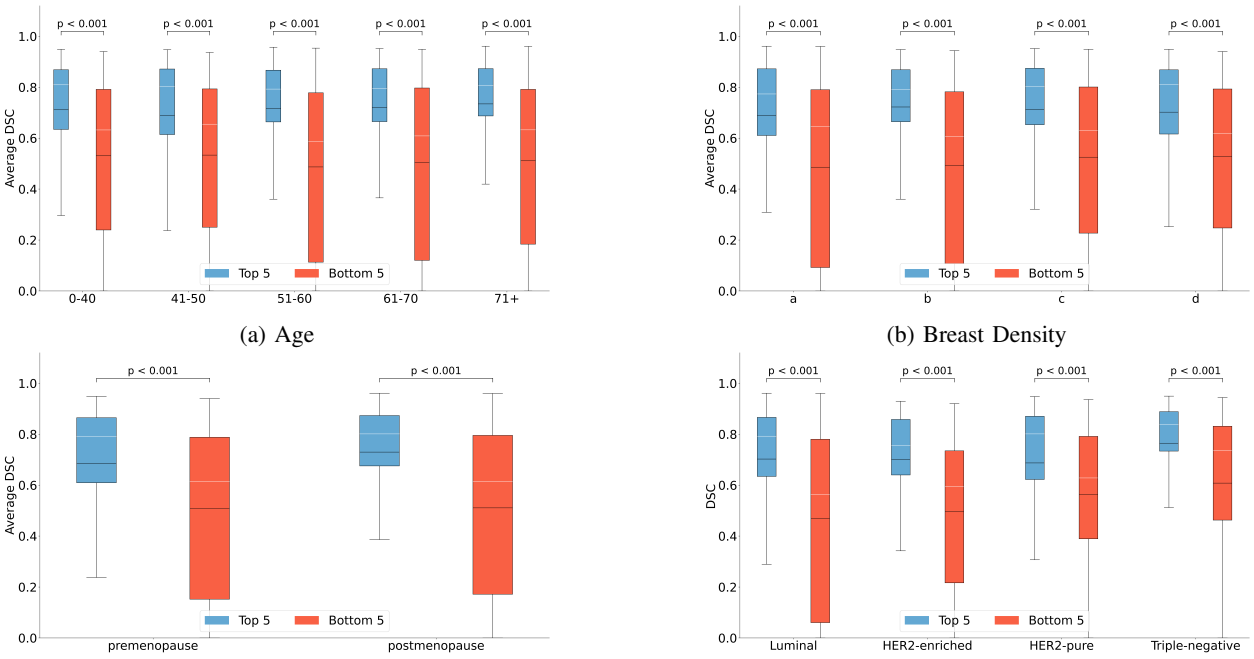
A. Robustness of Primary Tumor Segmentation

Primary tumor segmentation attracted 20 valid submissions to the final test phase, reflecting both its clinical relevance and methodological maturity. Under cross-continental evaluation, most top-performing methods improved over the nnU-Net baseline in both accuracy and subgroup consistency, indicating that contemporary 3D pipelines generalize reasonably well to unseen centers and heterogeneous acquisition protocols.

Methodologically, submissions showed strong convergence toward robustness-oriented design choices. All top five methods relied on fully 3D architectures, predominantly nnU-Net variants, with one exception adopting a 3D Vision Transformer backbone (Table II). Ensemble-based inference was employed by four of the top five teams, together with standardized resampling and intensity normalization, highlighting the importance of variance reduction under cross-center deployment. With respect to input representation, four of the five top-ranked methods incorporated multiple DCE phases to explicitly model contrast enhancement dynamics, whereas only one method (e.g., FME) relied exclusively on a single subtraction image.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

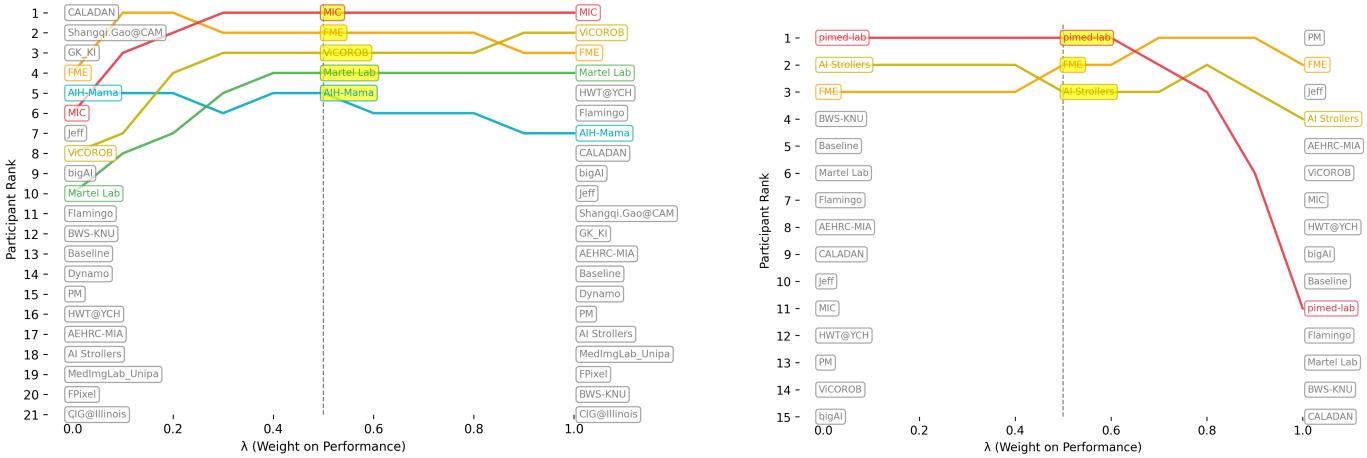
This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2026.3733343



(a) Age (b) Breast Density

(c) Menopausal Status (d) Tumor Subtype

Fig. 5: Subgroup analysis of segmentation performance. Boxplots show DSC distributions for the top 5 (blue) and bottom 5 (red) performing teams, stratified by (a) breast density, (b) age group, (c) menopausal status, and (d) Tumor Subtype. Brackets give two-sided Wilcoxon signed-rank p-values comparing the Top 5 against the Bottom 5 teams. For each patient, the mean DSC across the five top-ranked teams was compared with the mean DSC across the five bottom-ranked teams, so patients are the unit of analysis and the two arms are paired on identical cases. The performance of high-ranking methods consistently improve in comparison to the low-ranking methods. There is no substantial differences between subgroups in the challenge fairness categories. Across tumor subtype categories, the triple_negative subtype has the highest overall median DSC, while luminal category is especially challenging for the weaker models.



(a) Task 1: λ in Primary Tumor Segmentation (b) Task 2: λ in Pathologic Complete Response Prediction

Fig. 6: Fairness-versus-performance evaluation tradeoffs: Effect of λ on the ranking across tasks and participating teams.

While both strategies yielded competitive performance, no input design consistently mitigated failures in small, non-mass enhancement (NME), or low-contrast lesions, suggesting that current gains stem from incremental robustness improvements rather than fundamentally distinct paradigms.

Performance varied with tumor size (Fig. 3), with small tumors remaining challenging even for top-performing methods and accounting for the largest gap between top- and bottom-ranked teams (mean DSC delta 0.38 between averaged top 5 and bottom 5). Stability increased with tumor size both within and across performance tiers (e.g., mean DSC delta 0.12 for large tumors), indicating that leaderboard separation is largely driven by a method's ability to handle small, fragmented, or visually ambiguous lesions rather than larger cases.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2026.3733343

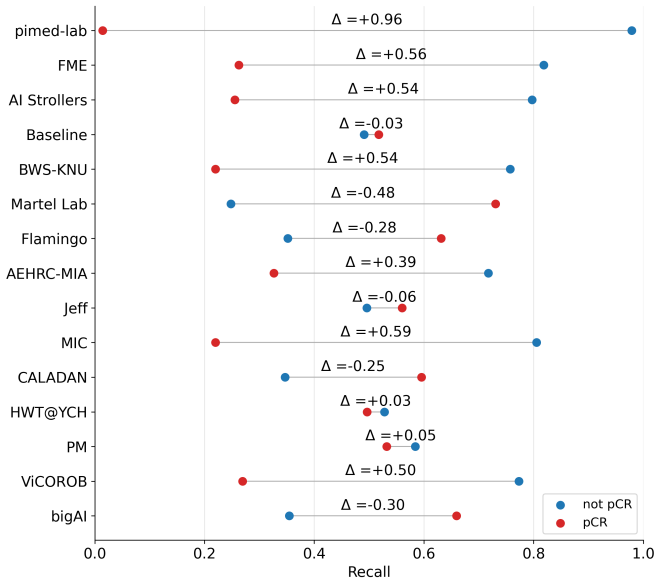


Fig. 7: Recall comparison between pCR (minority) and not pCR (majority) classes across participating teams. Δ denotes the recall gap, reflecting class-wise prediction bias.

Qualitative analysis (Fig. 2) corroborates these failure modes, including NME, weak lesion–background contrast, implant-related artifacts, and small or fragmented lesions. Substantial inter-method variability in such cases highlights model design trade-offs and the limitations of voxel-level overlap metrics in capturing clinically relevant uncertainty. Together, these findings indicate readiness for multi-center deployment in common scenarios, while motivating uncertainty-aware or interactive strategies for challenging lesions.

Stratifying segmentation performance by molecular subtype revealed an interesting pattern (Fig. 5d): triple-negative tumors were segmented most accurately, whereas the luminal subtype was the most challenging, with the performance differences being more pronounced in bottom-5 methods. This ordering can be explained by subtype-specific imaging morphology observed in the literature. Specifically, triple-negative cancers predominantly present as round or oval, circumscribed, rim-enhancing masses [71], producing high-contrast, and closed boundaries that the networks likely have an easier time delineating reliably. Luminal subtype is known to present as irregular tumors with spiculated margins, heterogenous enhancements, and in some cases, multifocal presentation [72], which can make it a more difficult segmentation target. Further comparing these results to existing state-of-the-methods in breast tumor segmentation, a subtype-robustness analyses conducted on the MAMA-MIA dataset [73] reported overall lower dice coefficient values (0.741 – 0.779), with the lowest performance also obtained on the luminal subtype. Another recent multicenter framework by Wang et al. [7] reports dice scores (n=687; subtype-wise DSC 0.820 – 0.857) with similar values as the top-5 methods, however, the worst performance is reported on HER2-enriched subtype. The absence of tumor-wise labels (e.g. mass, non-mass, heterogenous, irregular etc.) prevents us from further analyzing these correlations, but it does make an interesting future direction.

B. Challenges in Pretreatment pCR Prediction

Pretreatment pCR prediction task from baseline DCE-MRI proved substantially more challenging than tumor segmentation in this challenge. Across 14 submissions, performance gains over the random baseline were marginal, rarely statistically significant, and often driven by fairness rather than discriminative accuracy (Table V). Most models struggled with class imbalance and asymmetric recall, frequently defaulting to majority-class predictions while failing to reliably identify true pCR cases. Calibration analysis (Fig. 8) revealed limited probability separation and systematic overconfidence, suggesting reliance on weak imaging cues rather than robust response-related biomarkers.

FME was a consistently strong-performing submission across all evaluation settings, achieving the highest AUC among participating methods. Their approach combined an ensemble of pretrained CNN- and transformer-based models and test-time augmentation. Although the individual contributions of these components cannot be disentangled, the observed results suggest that such strategies may promote more robust representations, contributing not only to their predictive performance but also to the stability of the method across different values of the fairness weighting parameter λ. This approach contrasts with the findings from Martel Lab, who employed a vision transformer pretrained on a large-scale medical database for joint segmentation and pCR prediction. While it reached the top-5 in tumor segmentation, it did not consistently improve pCR prediction (AUC 0.47), indicating that representation scale alone may beinsufficient without complementary strategies for variance reduction.

The failure of highly accurate segmentation models to effectively predict pCR was not unique to Martel Lab. Participating approaches included lesion-centered end-to-end classifiers and pipelines leveraging segmentation-derived features (MIC team top-1 Task 1 [57]), yet no clear relationship emerged between segmentation quality and pCR performance. While one team (AI Strollers) reused Task 1 features within an XGBoost classifier, the top three submissions on Task 2 generally addressed segmentation and response prediction as separate modeling problems, without employing a unified multi-task framework with shared feature representations.

In this setup of predicting pCR from pre-treatment imaging, temporal information was restricted to contrast dynamics within DCE acquisition. All top-ranked methods incorporated multiple contrast phases (Table III), but differed in modeling strategy: two applied end-to-end deep learning to multi-phase inputs, while one used segmentation-derived features followed by a separate classifier, resulting in only implicit temporal modeling.

In regards to the fairness analysis, no clear and reliable trends emerged in Task 2 due to the overall difficulty of predicting pCR. Molecular subtype, which was used as an independent variable to analyze Task 1, is an intrinsic disease characteristic that strongly influences the likelihood of achieving pCR [74]. Consequently, treating subtype as a protected

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2026.3733343

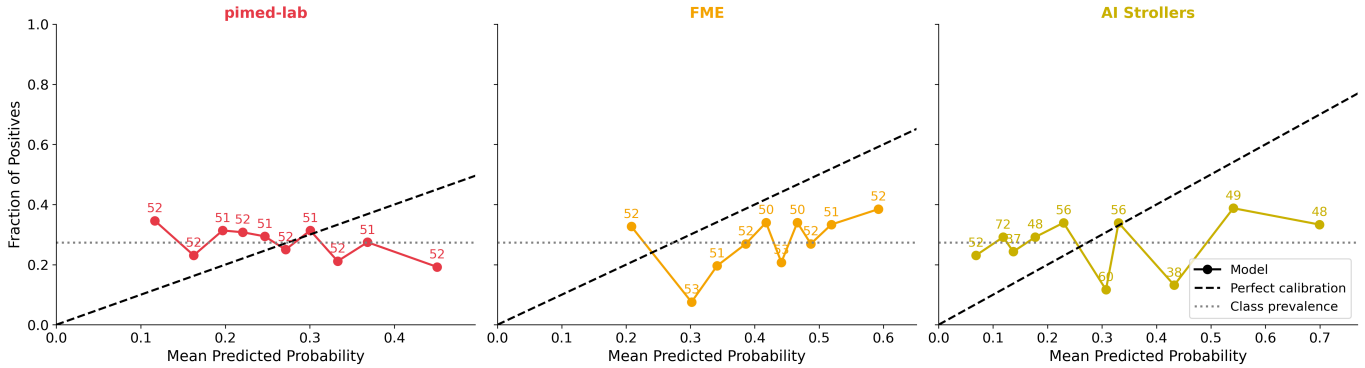


Fig. 8: Reliability diagrams for top 3 teams in pCR prediction. Predicted probabilities are grouped into quantile bins; the fraction of true positives per bin is plotted against mean predicted probability. Dashed line: perfect calibration; dotted line: class prevalence. Annotated values indicate bin sample sizes.

attribute would conflate a genuine biological determinant of the target outcome with an equity concern. Large clinical studies have consistently demonstrated substantial differences in pCR rates across subtypes, with luminal tumors exhibiting the lowest response rates and HER2-positive and triple-negative tumors achieving markedly higher rates [44]. Similarly, previous MRI studies have reported subtype-dependent differences in the diagnostic performance of imaging for treatment response assessment [75], as well as improvement when subtype information is incorporated in the model [76].

Overall, the results show that a substantial gap remains between the promising metrics reported in the pCR prediction literature and the performance observed under a common benchmark setting. Several recent studies have demonstrated encouraging performance using pretreatment MRI with advanced imaging biomarkers and external validation cohorts [8]–[11]. However, other multicenter studies have reached more cautious conclusions, reporting that pretreatment MRI radiomics alone provides limited predictive value, does not consistently outperform clinicopathologic models, or offers little additional benefit when combined with established clinical predictors [76]–[78]. Our findings support this latter body of evidence, suggesting that pretreatment DCE-MRI in isolation may not provide sufficient discriminative information for robust and generalizable pCR prediction across heterogeneous clinical centers.

At the same time, recent work has increasingly shifted toward longitudinal and multimodal strategies, integrating early on-treatment MRI, pathology, or tumor microenvironment information, which consistently report improved predictive performance [79]–[83]. Together, these studies suggest that incorporating temporal and complementary biological information may be a more promising direction than relying solely on pretreatment imaging. We therefore view the MAMA-MIA benchmark not as evidence that pretreatment MRI has reached a definitive performance ceiling, but as a challenging and standardized baseline against which future methodological advances can be rigorously evaluated.

C. On Subgroup Fairness as Challenge Objective

Participating teams varied in how explicitly fairness was addressed during model development, with most submissions optimizing predictive performance and assessing fairness post-hoc through benchmark evaluation. A notable exception was the work from participant AIH-MAMA (FairMedSeg), which explicitly integrated subgroup fairness into their segmentation objective by penalizing performance gaps between subgroups during training. Other teams such as FME and MIC adopted robustness-oriented strategies, such as subtraction-based inputs or large-scale self-supervised pretraining and augmentation, thereby indirectly promoting more consistent performance across subgroups.

Considering the pCR prediction, the results indicate that fairness considerations are meaningful primarily once a minimum level of predictive performance is achieved. As illustrated in Fig. 1b, low-performing models may appear particularly fair due to uniformly low predictions across subgroups, whereas fairness becomes a relevant and discriminative criterion only among models with competitive performance. This suggests that future benchmark editions are to consider adaptive weighting schemes in which fairness is emphasized more strongly the higher the performance of a downstream task model, thus aligning fairness incentives with clinical utility.

In this regard, the impact of the fairness weighting parameter λ further highlights the normative nature of weighting the importance of subgroup fairness, particularly when dealing with tasks of differing clinical maturity. Our choice of an equal baseline weight ($\lambda = 0.5$) serves as a neutral starting heuristic; however, a fixed linear composite score has clear limitations. As observed in the lower-performing classification task, equal weighting risks inadvertently favoring fair but clinically uninformative models when predictive accuracy is close to chance.

These findings emphasize that fairness-aware benchmarking does not yield a single "correct" ranking, but instead exposes trade-offs that depend on institutional values, regulatory expectations, and clinical risk tolerance. Future iterations of such benchmarks should consider alternative formulation paradigms, such as performance-gated weighting (where a

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2026.3733343

model must pass a baseline accuracy threshold to be eligible for fairness evaluation) or adaptive weighting schemes that scale λ dynamically based on baseline task difficulty. By making these trade-offs explicit and adjustable, MAMA-MIA provides a transparent framework for interrogating how fairness objectives interact with performance.

We further note limitations of basing subgroup fairness on the differences between the means of performance metrics, as averaging subgroup performance can obscure relevant variability. For instance, models with similar mean fairness scores can exhibit substantially different, and likely undesirable, variability across subgroups. Moreover, mean-based across subgroups gaps can, in principle, be reduced by degrading performance in the highest-performing subgroup rather than by improving underperforming ones. This motivates future consideration of incorporating maximum-performance-based and variance–based measures to provide complementary insight into the robustness and reliability of reported fairness measures, particularly under heterogeneous subgroup sizes or outcome prevalence.

D. Implications for Future Benchmark Development and Clinical Translation

The MAMA-MIA results highlight several directions for future benchmark design and model development. Most notably, the contrast between the strong performance achieved in primary tumor segmentation and the limited success of pretreatment pCR prediction suggests that current breast MRI AI has reached different levels of maturity across tasks.

While achieving impressive performance, the majority of submissions for primary tumor segmentation relied heavily on established convolutional neural networks (CNNs), particularly the self-configuring nnUNet framework [13] and its variants. There remains substantial scope to investigate how these techniques can be modified, or how novel architectures [84] can be introduced, to address specific vulnerabilities identified in this challenge, such as small tumors and the luminal molecular subtype. Ultimately, optimally tuning these newer approaches for multi-phase 3D DCE-MRI will be necessary to break through the performance ceilings established by the current CNN baselines.

For predicting pCR, participating methods exhibited substantial diversity in modelling architectures, pretraining strategies, and pre- and post-processing pipelines. However, overall performance remained limited, with no method exceeding an AUC of 0.57. This suggests that the challenge setting, characterized by domain shift, remains difficult to address using imaging information alone. The results further indicate that future progress will likely require the integration of treatment-specific variables (e.g., regimen, dosage), molecular subtype, and proliferation markers (e.g., Ki-67 [85]) to contextualize the imaging findings. In particular, the observed performance variability across molecular subtypes and patient subgroups indicates that response prediction cannot be treated solely as an imaging problem, but must account for the underlying biological heterogeneity of breast cancer. Likewise, extending evaluation beyond binary pCR classification toward more continuous measures of treatment response, such as residual disease burden or tumor volume change, may provide richer information and better reflect clinically relevant treatment outcomes. The inclusion of longitudinal imaging acquired during therapy could further help capture treatment-induced changes that are not observable from pretreatment imaging alone [86], [87].

From a benchmarking perspective, the challenge also revealed several opportunities for methodological refinement. The sensitivity of leaderboard rankings to the fairness weighting parameter λ demonstrates the importance of transparently reporting performance–fairness trade-offs and motivates future investigation of adaptive or performance-gated fairness formulations. Similarly, computational efficiency was not explicitly included in the challenge ranking, despite its relevance to clinical usability. While a strict upper limit of 5 minutes per case was defined, individual runtime measurements were not considered sufficiently reproducible for leaderboard ranking, as execution times in containerized benchmark environments can be affected by implementation details and infrastructure-related variability. Future benchmark editions could investigate standardized efficiency metrics, including inference latency, memory consumption, and energy usage, measured under controlled execution settings.

V. Conclusion

This work introduced the MAMA-MIA benchmark, a large-scale, multi-center framework designed to jointly evaluate generalization and subgroup fairness in breast MRI tumor segmentation and pretreatment pCR prediction. While automated tumor segmentation demonstrated robust performance and consistent generalization across unseen clinical sites, pCR prediction from pretreatment DCE-MRI showed only marginal gains over baseline under realistic cross-center evaluation.

These findings suggest that, despite encouraging progress reported in recent imaging-based pCR prediction studies, substantial challenges remain in achieving robust and reproducible performance under a common multi-center evaluation framework. By providing standardized evaluation across institutions, countries, imaging protocols, and clinically relevant patient subgroups, MAMA-MIA offers a complementary perspective to individual model-development studies and helps quantify the gap between experimental performance and real-world generalizability.

More broadly, the challenge demonstrates the value of integrating subgroup fairness alongside predictive performance when assessing medical AI systems. We hope that MAMA-MIA will serve as a lasting benchmark resource for the community, supporting the development of breast MRI models that are not only accurate, but also robust, equitable, and clinically deployable across diverse patient populations.

Acknowledgment

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

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2026.3733343

by the project FUTURE-ES (PID2021-126724OB-I00) and project AIMED (PID2023-146786OB-I00) from the Ministry of Science and Innovation of Spain. K.K. holds a RyC fellowship grant (RYC2024-048256-I).

Team Martel Lab (M. A., T.X., and A.M.) are supported by the Canada Foundation for Innovation (40206), and the Ontario Research Fund. Computational resources were provided in part by the Digital Research Alliance of Canada (alliance-can.ca). Team ViCOROB (H.A., J.C.V. and R.M.) funded by the projects VICTORIA, "PID2021-123390OB-C21" and IMPACT, "PID2024-157201OB-C21" from the Ministerio de Ciencia, Innovación y Universidades of Spain, and the PhD grant IFUdG2024 from the University of Girona. Team pimed-lab (D.S., J.H.L. and M.R.) is supported by the National Cancer Institute of the National Institutes of Health under Award Number R37CA260346. Team AIH-MAMA (E.P., L.V., E.P., M.A.Z.) is partly supported by the ANR-BMBF TRAIN (ANR-22-FAI1-0003-02). We would like to thank all team members who helped during the benchmark: Daniele Falcetta (AIH-MAMA), Vincenzo Marciano' (AIH-MAMA), Tania Cerquitelli (AIH-MAMA), Elena Baralis (AIH-MAMA).

REFERENCES

[1] World Health Organization, "Breast cancer: Global patterns of incidence, mortality, and survival," https://www.who.int/news-room/fact-sheets/detail/breast-cancer, 2024, accessed: 2025-04-27.

[2] H. Sung *et al.*, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[3] R. M. Mann, N. Cho, and L. Moy, "Breast MRI: state of the art," *Radiology*, vol. 292, no. 3, pp. 520–536, 2019.

[4] N. Hylton *et al.*, "Locally advanced breast cancer: MR imaging for prediction of response to neoadjuvant chemotherapy—results from ACRIN 6657/I-SPY trial," *Radiology*, vol. 279, no. 1, pp. 44–55, 2016.

[5] A. Pennisi, T. Kieber-Emmons, I. Makhoul, and L. Hutchins, "Relevance of pathological complete response after neoadjuvant therapy for breast cancer," *Breast cancer: basic and clinical research*, vol. 10, pp. BCBCR–S33 163, 2016.

[6] J. Zhang *et al.*, "A robust and efficient ai assistant for breast tumor segmentation from dce-mri via a spatial-temporal framework," *Patterns*, vol. 4, no. 9, 2023.

[7] X. Wang, X. Hu, C. Wang, H. Yang, Y. Hu, X. Lan *et al.*, "Automatic segmentation and molecular subtype classification of breast cancer using an mri-based deep learning framework," *Radiology: Imaging Cancer*, vol. 7, no. 3, p. e240184, 2025.

[8] Z. Shi, X. Huang, Z. Cheng, Z. Xu, H. Lin, C. Liu, X. Chen, C. Liu, C. Liang, C. Lu *et al.*, "Mri-based quantification of intratumoral heterogeneity for predicting treatment response to neoadjuvant chemotherapy in breast cancer," *Radiology*, vol. 308, no. 1, p. e222830, 2023.

[9] Y. Huang, X. Wang, Y. Cao, X. Lan, X. Hu, F. Mou, H. Chen, X. Gong, L. Li, S. Tang *et al.*, "Nomogram for predicting neoadjuvant chemotherapy response in breast cancer using mri-based intratumoral heterogeneity quantification," *Radiology*, vol. 315, no. 1, p. e241805, 2025.

[10] J. Zhang, Y. Li, Z. Li, Z. Cui, P. Li, J. Li, Z. Li, Y. Xie, K. W. Chan, Q. Zhang *et al.*, "Deep-learning-based her2 status assessment from multimodal breast cancer data predicts neoadjuvant therapy response," *Nature Biomedical Engineering*, pp. 1–12, 2025.

[11] M. Caballo, W. B. Sanderink, L. Han, Y. Gao, A. Athanasiou, and R. M. Mann, "Four-dimensional machine learning radiomics for the pre-treatment assessment of breast cancer pathologic complete response to neoadjuvant chemotherapy in dynamic contrast-enhanced mri," *Journal of Magnetic Resonance Imaging*, vol. 57, no. 1, pp. 97–110, 2023.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[13] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021. [Online]. Available: https://doi.org/10.1038/s41592-020-01008-z

[14] A. Myronenko, "3d mri brain tumor segmentation using autoencoder regularization," in *International MICCAI brainlesion workshop*. Springer, 2018, pp. 311–320.

[15] S. Joshi, R. Osuala, L. Garrucho, K. Kushibar, D. Kessler, O. Diaz, and K. Lekadir, "Single image test-time adaptation via multi-view co-training," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2025, pp. 617–627.

[16] J. Wang *et al.*, "Generalizing to unseen domains: A survey on domain generalization," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2022.

[17] W. Li *et al.*, "Breast multiparametric mri for prediction of neoadjuvant chemotherapy response in breast cancer: the bmmr2 challenge," *Radiology: Imaging Cancer*, vol. 6, no. 1, p. e230033, 2024.

[18] E. J. Sutton *et al.*, "A machine learning model that classifies breast cancer pathologic complete response on mri post-neoadjuvant chemotherapy," *Breast Cancer Research*, vol. 22, pp. 1–11, 2020.

[19] Y. Peng *et al.*, "Pretreatment dce-mri-based deep learning outperforms radiomics analysis in predicting pathologic complete response to neoadjuvant chemotherapy in breast cancer," *Frontiers in oncology*, vol. 12, p. 846775, 2022.

[20] N. Khan, R. Adam, P. Huang, T. Maldjian, and T. Q. Duong, "Deep learning prediction of pathologic complete response in breast cancer using mri and other clinical data: a systematic review," *Tomography*, vol. 8, no. 6, pp. 2784–2795, 2022.

[21] A. Saha *et al.*, "Dynamic contrast-enhanced magnetic resonance images of breast cancer patients with tumor locations [data set]," *The Cancer Imaging Archive*, 2021. [Online]. Available: https://doi.org/10.7937/TCIA.e3sv-re93

[22] W. Li *et al.*, "I-SPY 2 breast dynamic contrast enhanced MRI trial (version 1) [data set]," *The Cancer Imaging Archive*, 2022. [Online]. Available: https://doi.org/10.7937/TCIA.D8Z0-9T85

[23] S. Bakas *et al.*, "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge," *arXiv preprint arXiv:1811.02629*, 2018.

[24] V. Campello *et al.*, "Multi-centre, multi-vendor and multi-disease cardiac segmentation: the m&ms challenge," *IEEE Transactions on Medical Imaging*, vol. 40, no. 12, pp. 3543–3554, 2021.

[25] R. Dorent *et al.*, "Crossmoda 2021 challenge: Benchmark of cross-modality domain adaptation techniques for vestibular schwannoma and cochlea segmentation," *Medical Image Analysis*, vol. 83, p. 102628, 2023.

[26] R. J. Chen *et al.*, "Algorithmic fairness in artificial intelligence for medicine and healthcare," *Nature biomedical engineering*, vol. 7, no. 6, pp. 719–742, 2023.

[27] K. Drukker *et al.*, "Toward fairness in artificial intelligence for medical image analysis: identification and mitigation of potential biases in the roadmap from data collection to model deployment," *Journal of Medical Imaging*, vol. 10, no. 6, pp. 061 104–061 104, 2023.

[28] V. N. Dang *et al.*, "Auditing unfair biases in cnn-based diagnosis of alzheimer's disease," in *Workshop on Clinical Image-Based Procedures*. Springer, 2023, pp. 172–182.

[29] Y. Yang, H. Zhang, J. W. Gichoya, D. Katabi, and M. Ghassemi, "The limits of fair medical imaging ai in real-world generalization," *Nature Medicine*, vol. 30, no. 10, pp. 2838–2848, 2024.

[30] E. T. Sedeta, B. Jobre, and B. Avezbakiyev, "Breast cancer: Global patterns of incidence, mortality, and trends." *Journal of Clinical Oncology*, vol. 41, no. 16_suppl, pp. 10 528–10 528, 2023.

[31] L. Garrucho *et al.*, "A large-scale multicenter breast cancer dce-mri benchmark dataset with expert segmentations," *Scientific Data*, vol. 12, no. 1, p. 453, 2025.

[32] K. Clark *et al.*, "The cancer imaging archive (tcia): maintaining and operating a public information repository," *Journal of digital imaging*, vol. 26, no. 6, pp. 1045–1057, 2013.

[33] D. Newitt *et al.*, "Multicenter breast DCE-MRI data and segmentations from patients in the I-SPY 1/ACRIN 6657 trials," *The Cancer Imaging Archive*, 2016. [Online]. Available: https://doi.org/10.7937/K9/TCIA.2016.HdHpgJLK

[34] D. Newitt and N. Hylton, "Single site breast DCE-MRI data and segmentations from patients undergoing neoadjuvant chemotherapy (version 3) [data set]," *The Cancer Imaging Archive*, 2016. [Online]. Available: https://doi.org/10.7937/K9/TCIA.2016.QHsyhJKy

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2026.3733343

[35] S. Joshi, R. Osuala, L. Garrucho, A. Tsirikoglou, J. del Riego, K. Gwoździewicz, K. Kushibar, O. Diaz, and K. Lekadir, "Leveraging epistemic uncertainty to improve tumour segmentation in breast mri: an exploratory analysis," in *Medical Imaging 2024: Image Processing*, vol. 12926. SPIE, 2024, pp. 292–300.

[36] F. Malmberg, R. Nordenskjöld, R. Strand, and J. Kullberg, "Smartpaint: a tool for interactive segmentation of medical volume images," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 5, no. 1, pp. 36–44, 2017.

[37] K. Lekadir *et al.*, "Future-ai: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare," *bmj*, vol. 388, 2025.

[38] S. Y.-H. Chiu, S. Duffy, A. M.-F. Yen, L. Tabár, R. A. Smith, and H.-H. Chen, "Effect of Baseline Breast Density on Breast Cancer Incidence, Stage, Mortality, and Screening Parameters: 25-Year Follow-up of a Swedish Mammographic Screening," *Cancer Epidemiology, Biomarkers & Prevention*, vol. 19, no. 5, pp. 1219–1228, 2010. [Online]. Available: https://doi.org/10.1158/1055-9965.EPI-09-1028

[39] N. J. Engmann, M. K. Golmakani, D. L. Miglioretti, B. L. Sprague, K. Kerlikowske *et al.*, "Population-Attributable Risk Proportion of Clinical Risk Factors for Breast Cancer," *JAMA Oncology*, vol. 3, no. 9, pp. 1228–1236, 2017. [Online]. Available: https://doi.org/10.1001/jamaoncol.2016.6326

[40] D. van der Waal, T. M. Ripping, A. L. Verbeek, and M. J. Broeders, "Breast cancer screening effect across breast density strata: A case–control study," *International Journal of Cancer*, vol. 140, no. 1, pp. 41–49, 2017. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/ijc.30430

[41] V. King, Y. Gu, J. B. Kaplan, J. D. Brooks, M. C. Pike, and E. A. Morris, "Impact of menopausal status on background parenchymal enhancement and fibroglandular tissue on breast mri," *European radiology*, vol. 22, no. 12, pp. 2641–2647, 2012.

[42] G. von Waldenfels, S. Loibl, J. Furlanetto, A. Machleidt, B. Lederer, C. Denkert, C. Hanusch, S. Kümmel, G. von Minckwitz, A. Schneeweiss *et al.*, "Outcome after neoadjuvant chemotherapy in elderly breast cancer patients–a pooled analysis of individual patient data from eight prospectively randomized controlled trials," *Oncotarget*, vol. 9, no. 20, p. 15168, 2018.

[43] S. Loibl, C. Jackisch, B. Lederer, M. Untch, S. Paepke, S. Kümmel, A. Schneeweiss, J. Huober, J. Hilfrich, C. Hanusch *et al.*, "Outcome after neoadjuvant chemotherapy in young breast cancer patients: a pooled analysis of individual patient data from eight prospectively randomized controlled trials," *Breast cancer research and treatment*, vol. 152, no. 2, pp. 377–387, 2015.

[44] W. Haque, V. Verma, S. Hatch, V. Suzanne Klimberg, E. Brian Butler, and B. S. Teh, "Response rates and pathologic complete response by breast cancer molecular subtype following neoadjuvant chemotherapy," *Breast cancer research and treatment*, vol. 170, no. 3, pp. 559–567, 2018.

[45] M. Ghezzawi, M. H. El Charif, L. Lteif, V. Panossian, A. Haddad, S. Sarkis, Y. Zebian, M. Kheil, M. A. Maktabi, N. Fakhruddin *et al.*, "Patterns of response of breast cancer after neoadjuvant chemotherapy according to molecular subtype," *Frontiers in Oncology*, vol. 15, p. 1670594, 2025.

[46] M. Amine, "Pycad: A python library for medical imaging analysis," https://github.com/amine0110/pycad, 2023, gitHub repository.

[47] Z. Xu *et al.*, "Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform," *Patterns*, vol. 3, no. 7, 2022.

[48] J. C. Gee *et al.*, *Medical Image Computing and Computer Assisted Intervention–MICCAI 2025: 28th International Conference, Daejeon, South Korea, September 23–27, 2025, Proceedings, Part V*. Springer Nature, 2025.

[49] T. Zhang *et al.*, Eds., *Artificial Intelligence and Imaging for Diagnostic and Treatment Challenges in Breast Care*, ser. Lecture Notes in Computer Science, vol. 16142. Cham: Springer, 2026. [Online]. Available: https://link.springer.com/book/10.1007/978-3-032-05559-0

[50] C. D'Orsi, E. Sickles, E. Mendelson, and E. Morris, *ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System (ACR BI-RADS® Magnetic Resonance Imaging)*. Reston, VA: American College of Radiology, 2013.

[51] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[52] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers*. Springer, 2010, pp. 177–186.

[53] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[54] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. Ieee, 2016, pp. 565–571.

[55] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[56] F. Isensee *et al.*, "nnu-net revisited: A call for rigorous validation in 3d medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*. Cham: Springer Nature Switzerland, 2024, pp. 488–498.

[57] J. Kächele, D. Bounias, A. Ertl, and K. Maier-Hein, "On tackling domain shift in breast MRI using only publicly-available data: Reproducible breast cancer segmentation and pCR prediction," in *Lecture Notes in Computer Science*, ser. Lecture notes in computer science. Cham: Springer Nature Switzerland, 2026, pp. 310–319.

[58] H. Awwad, J. C. Vilanova, and R. Martí, "Can we teach ai to understand breast tumour behaviour? our mama-mia challenge journey," in *Artificial Intelligence and Imaging for Diagnostic and Treatment Challenges in Breast Care*, T. Zhang *et al.*, Eds. Cham: Springer Nature Switzerland, 2026, pp. 248–257.

[59] T. Xu *et al.*, "A generalizable 3d framework and model for self-supervised learning in medical imaging," 2025. [Online]. Available: https://arxiv.org/abs/2501.11755

[60] E. Poeta *et al.*, "Divergence-aware training with automatic subgroup mitigation for breast tumor segmentation," in *Deep Breast Workshop on AI and Imaging for Diagnostic and Treatment Challenges in Breast Care*. Springer, 2025, pp. 52–62.

[61] E. Pastor, L. de Alfaro, and E. Baralis, "Looking for trouble: Analyzing classifier behavior via pattern divergence," in *Proceedings of the 2021 International Conference on Management of Data*, ser. SIGMOD '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 1400–1412. [Online]. Available: https://doi.org/10.1145/3448016.3457284

[62] A. Koudounas, E. Pastor, L. de Alfaro, and E. Baralis, "Mitigating subgroup disparities in speech models: A divergence-aware dual strategy," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 883–895, 2025.

[63] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015. [Online]. Available: https://arxiv.org/abs/1512.03385

[64] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," 2018. [Online]. Available: https://arxiv.org/abs/1711.11248

[65] TorchVision maintainers and contributors. (2016) TorchVision: PyTorch's computer vision library. GitHub repository. [Online]. Available: https://github.com/pytorch/vision

[66] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. ACM, Aug. 2016, p. 785–794. [Online]. Available: http://dx.doi.org/10.1145/2939672.2939785

[67] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[68] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," 2020. [Online]. Available: https://arxiv.org/abs/2002.05709

[69] H. Li, W. Song, P. Qin, X. Yuan, and Z. Chen, "Hcma-unet: A hybrid cnn-mamba unet with axial self-attention for efficient breast cancer segmentation," in *2025 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2025, pp. 1–6.

[70] R. Liaw, E. Liang, R. Nishihara, P. Moritz, J. E. Gonzalez, and I. Stoica, "Tune: A research platform for distributed model selection and training," 2018. [Online]. Available: https://arxiv.org/abs/1807.05118

[71] J. H. Youk, E. J. Son, J. Chung, J.-A. Kim, and E.-k. Kim, "Triple-negative invasive breast cancer on dynamic contrast-enhanced and diffusion-weighted MR imaging: comparison with other breast cancer subtypes," *European radiology*, vol. 22, no. 8, pp. 1724–1734, 2012.

[72] C. C. Tajima, F. P. S. G. Arruda, V. C. Mineli, J. M. Ferreira, B. B. Bettim, C. A. B. d. T. Osório *et al.*, "MRI features of breast cancer immunophenotypes with a focus on luminal estrogen receptor low positive invasive carcinomas," *Scientific Reports*, vol. 14, no. 1, p. 19305, 2024.

[73] A. Bernal Rodriguez, B. Remeserío, J. Engelmann, L. Gago, and J. Velasco Rebolledo, "Input simplification impact on robustness for targeted therapy subtypes in breast MRI segmentation ai," in *Deep Breast*

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2026.3733343

Workshop on AI and Imaging for Diagnostic and Treatment Challenges in Breast Care. Springer, 2025, pp. 320–328.

[74] G. Von Minckwitz, M. Untch, J.-U. Blohmer, S. D. Costa, H. Eidtmann, P. A. Fasching *et al.*, "Definition and impact of pathologic complete response on prognosis after neoadjuvant chemotherapy in various intrinsic breast cancer subtypes," *Journal of clinical oncology*, vol. 30, no. 15, pp. 1796–1804, 2012.

[75] J. Kim, B.-K. Han, E. Y. Ko, E. S. Ko, J. S. Choi, and K. W. Park, "Prediction of pathologic complete response on MRI in patients with breast cancer receiving neoadjuvant chemotherapy according to molecular subtypes," *European Radiology*, vol. 32, no. 6, pp. 4056–4066, 2022.

[76] M. H. Janse, L. M. Janssen, E. J. Wolters-van der Ben, M. R. Moman, M. A. Viergever, P. J. van Diest *et al.*, "Deep learning-based radiomics does not improve residual cancer burden prediction post-chemotherapy in LIMA breast MRI trial," *European Radiology*, vol. 36, no. 2, pp. 850–862, 2026.

[77] R. W. Granzier, A. Ibrahim, S. P. Primakov, S. Samiei, T. J. van Nijnatten, M. de Boer *et al.*, "MRI-based radiomics analysis for the pretreatment prediction of pathologic complete tumor response to neoadjuvant systemic therapy in breast cancer patients: a multicenter study," *Cancers*, vol. 13, no. 10, p. 2447, 2021.

[78] J. O'Donnell, S. Gasior, M. Davey, E. O'Malley, A. Lowery, J. McGarry *et al.*, "The accuracy of breast mri radiomic methodologies in predicting pathological complete response to neoadjuvant chemotherapy: A systematic review and network meta-analysis," *European journal of radiology*, vol. 157, p. 110561, 2022.

[79] I. Janíčková, Y. Y. Tan, T. H. Helbich, K. Miloserdov, Z. Bago-Horvath, U. Heber, and G. Langs, "Temporal representation learning of phenotype trajectories for pcr prediction in breast cancer," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2025, pp. 606–615.

[80] W. Tang, C. Jin, Q. Kong, C. Liu, S. Chen, S. Ding *et al.*, "Development and validation of an MRI spatiotemporal interaction model for early noninvasive prediction of neoadjuvant chemotherapy response in breast cancer: a multicentre study," *EClinicalMedicine*, vol. 85, 2025.

[81] X. Huang, Z. Xu, Y. Zhao, Y. Wang, Y. Liu, W. Hu *et al.*, "Longitudinal MRI-based deep learning model for predicting pathological complete response in breast cancer: a multicenter, retrospective cohort study," *NPJ Precision Oncology*, vol. 10, no. 1, p. 57, 2026.

[82] E. Krasniqi, L. Filomeno, T. Arcuri, G. Ferretti, S. Gasparro, A. Fulvi *et al.*, "Multimodal deep learning for predicting neoadjuvant treatment outcomes in breast cancer: a systematic review," *Biology direct*, vol. 20, no. 1, p. 72, 2025.

[83] L. Yan, X. Huang, L. Liu, A. Wu, Y. Luo, H. Li *et al.*, "Predicting response to neoadjuvant therapy in breast cancer using longitudinal dce-mri deep learning integrated with tumor microenvironment data," *Frontiers in Immunology*, vol. 17, p. 1749668, 2026.

[84] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[85] X. Chen *et al.*, "The predictive value of ki-67 before neoadjuvant chemotherapy for breast cancer: a systematic review and meta-analysis," *Future oncology*, vol. 13, no. 9, p. 843, 2017.

[86] W. Li, D. C. Newitt, J. Gibbs, L. J. Wilmes, E. F. Jones, V. A. Arasu, F. Strand, N. Onishi, A. A.-T. Nguyen, J. Kornak *et al.*, "Predicting breast cancer response to neoadjuvant treatment using multi-feature mri: results from the i-spy 2 trial," *NPJ breast cancer*, vol. 6, no. 1, p. 63, 2020.

[87] D. Ma, J. Cao, H. Cheng, D. Zhou, J. Liu, X. Zhang, K. Wu, C. Gu, and X. Guan, "Longitudinal mri-clinical multimodal fusion for pcr prediction in breast cancer," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2025, pp. 323–332.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/